Holistically Modelling the Medieval Book: Towards a Digital Contribution

Peter Anthony Stokes

Peter A. Stokes*

Holistically Modelling the Medieval Book: Towards a Digital Contribution

https://doi.org/10.1515/ang-2021-0002

Abstract: The book has long played an important role in medieval and indeed modern culture, being at the same time a carrier of texts and images, a sign potentially of wealth and/or education, a site of enquiry for modern scholarship for literature, history, linguistics, palaeography, codicology, art history, and more. The 'archaeology of the book' can tell us about its history (or biography) as well as the cultures that produced and used it, right up to its present ownership. This multidimensionality of the object has long been known, but it has also proven a challenge to digital approaches which (like all representations) are by their nature models that involve conscious or unconscious selection of particular aspects, and that have been more successful in some aspects than others. This then raises the question to what degree these different viewpoints can be brought together into something approaching a holistic view, while always allowing for the tension between standardisation and innovation, and while remembering that a 'complete model' is a tautology, neither possible nor desirable.

Key terms: Digital Humanities, Manuscript Studies, palaeography, codicology, textual criticism, modelling, Linked Open Data, ontologies

1 Introduction: On Models and Completeness

As this special issue attests, digital and computational methods have become increasingly important and are now largely indispensable for Book History. Even scholars who do not explicitly acknowledge digital methods still normally rely on digitised images of books, as well as websites, reference databases, and other electronic sources, and indeed this has been the case for some years now. Furthermore, as this special issue demonstrates, books have long been the subject of a

***Corresponding author: Peter A. Stokes,** École Pratique des Hautes Études – Université Paris Sciences et Lettres, Archéologie et Philologie d'Orient et d'Occident (UMR 8546), Paris
E-Mail: peter.stokes@ephe.psl.eu

great deal of scholarly attention, whether for publication in print, digital or both, with even relatively specialised terms such as "digital palaeography" now being in use for over fifteen years (e.g. Ciula 2005 with discussion by Stokes 2009: 319–323). As a result of this, many different approaches have been taken to applying digital methods to Book History, along with a good deal of discussion about their strengths, weaknesses, limitations and potential. In this respect, Book History is a good example and test case of the Digital Humanities more generally, insofar as digital methods have been particularly successful when applied to books, partly because two of the more important aspects of books are their textual content and their physical appearance, both of which are particularly well suited to digital computers. Books of course comprise very many different aspects, or "dimensions" to use the terminology of Elena Pierazzo (2015: 41–64), and any given study is necessarily limited in the number of different dimensions that it can encompass, for simple practical reasons of time, resources, expertise and so on. If our goal is something approaching a 'complete' holistic model of the book, then nevertheless the complexity of each aspect is (one would imagine) too much for any given person, in terms both of expertise and also simple time and effort in representing this. Perhaps the closest known to this author is a monograph on the Old English poem known as 'Cædmon's Hymn' (O'Donnell 2005). In this work, the author studies the poem in enormous detail, looking at the text in all surviving medieval copies, considering the manuscript context, the palaeography, the philology, the literary context and more. It is published in both print and digital formats, including a dynamic facsimile edition, which allows close comparison of the different manuscripts and texts. Nevertheless, even this does not consider all aspects – there is no discussion of the material sciences, for instance – but nor is it reasonable to expect this. Even here, however, the author notes that the book was the result of over a decade of work, and this for a poem which is barely more than forty words in length.

Despite some claims to the contrary, then, no approach is or can ever be 'complete' in any sense, and nor should it be. Digital analyses of books are sometimes criticised for their 'incompleteness', for the fact that they lack some aspect of the original object,[1] and although this accusation of incompleteness is undoubtedly correct, it seems somewhat unreasonable to single out the digital in this respect. A printed article or monograph comprises a selective re-presentation

1 E.g. Treharne (2013: 477): "It is in studying the materiality of the book *in its completeness*, in embracing *all* elements of the participative experience [as opposed to the digital form], that we can begin to sense the incarnated nature of the book" (my emphasis).

of its subject just as an online study does, and a printed study also embeds a particular point of view and a particular disciplinary viewpoint (or set of viewpoints). Rather than emphasising incompleteness as a fault, a more useful approach might be to recognise that *any* analysis or representation of a book presupposes a model of that book: that is, it necessarily involves a simplification that results from a process of selection, whereby one makes choices about what aspects of the original object should be considered and what must be omitted. Rather than (only) a weakness, this is (also) its value, as a well-constructed model allows us to focus on particular aspects of the original, to see those aspects more clearly and in different ways, which should then allow new insights that can be added to our overall understanding of the original object. As Willard McCarty has written, models are "dangerous to us only if we miss the lesson of modeling and mistake the artificial for the real" (2008: 400). The question to be addressed here, then, is rather which different models have been prevalent in digital approaches to Book History, what are their underlying points of view, and what are the challenges and benefits of a holistic approach that tries to bring these different models together.

2 The Book as Text

One of the earliest models of the book, one that is still very prevalent, is that of 'text-bearing object'. This approach places less emphasis on the material object – often ignoring it entirely – and focusses instead on representing the text in its complexity. This approach has been used almost since the beginning of electronic digital computers, in part because computers are extremely well suited to the storage, transmission and treatment of texts. Digital treatment of texts has been central to the founding myths of the Digital Humanities, including well-known and highly influential projects such as the 'Index Thomisticus' of Father Busa (Busa 1980), numerous electronic corpora (e.g. Burton 1981a and Burton 1981b) and dictionaries (e.g. Frank and Cameron 1973), Project Gutenberg and many others. In these early cases, texts have generally been modelled as a sequence of characters, perhaps in part because this is close to the internal structure of computer memory and so is comparatively easy and therefore 'natural'. However, it is clear that texts also contain structure, be it presentational (bold, italic, and so on) or semantic (titles, paragraphs, etc.), and this has led to more complex models which include structuring information in some form of markup. The most influential of these for the Digital Humanities is that of the Text Encoding Initiative (TEI) which has been developing a model of text since the 1980s and which remains today the only viable standard for the scholarly encoding of texts in a manner that

is sustainable, (relatively) transparent and open for reuse, such that there are now literally millions of texts available in TEI XML.[2]

Historically, the Text Encoding Initiative has focussed on text, as one might reasonably expect, and has therefore had little concern for the material carrier. Texts have therefore been modelled as (for instance) chapters containing paragraphs containing words, poems containing stanzas containing lines of verse, and all text more generally as an "ordered hierarchy of content objects" (DeRose et al. 1990; Renear et al. 1996; Renear 2004: 224–225). The principle here is that texts can and in general should be represented as sequences of 'objects containing other objects', where the decision which 'objects' depends on the context and the object of study (paragraphs, chapters, and so on). The 'objects' form a hierarchy (some types of 'object' are inside other types and so on), and they are also ordered in the sense that changing the order of 'objects' is significant (reversing the order of paragraphs produces a different text, and so on). One important point here is that the TEI's focus on the text leads to the explicit decision that the 'objects' that are described should be textual rather than material. In other words, texts should *not* in general be represented as pages containing lines of prose, for instance, because pages and lines of prose are reflections of the material manifestation rather than 'essential' elements of the abstract text. This view has changed more recently with the recognition that at least some texts cannot be removed from their material context (Pierazzo and Stokes 2010; Burnard et al. 2010), and so the TEI now also allows the so-called 'documentary' view for these cases (TEI 2020, Ch. 11). A second and much more challenging case that the TEI model does not easily allow is the possibility that not all 'objects' are neatly contained in other 'objects' but often overlap in complex ways. For instance, grammatical structures or literary motifs may cross lines of poetry, but representing this is very difficult in the TEI because it is not part of that model for text. This incompleteness has been cited as a reason to abandon the TEI and indeed XML entirely (e.g. Schmidt 2010: 343–348). This approach seems short-sighted since, as we have already discussed, no model can ever be complete, and so any replacement model will necessarily also have limits and risks being worse than the TEI which benefits from over 30 years of collaborative scholarship. Certainly challenges remain then, and many interesting discussions are still to be had, but the basic model for textual content of books seems at this date to be relatively well established and to function well for the vast majority of cases where one's scholarly interest is in the textual dimension.

2 The Oxford Text Archive alone currently lists almost 64,000 items of type 'text' marked up in TEI, of which the entire British National Corpus is just one. See further <https://ota.bodleian.ox.ac.uk/>.

3 The Book as Images

Another prominent model of the book in digital contexts is as a set of images, or more precisely an ordered sequence of images since, as with text, the order of images is significant to the model. This approach is evident in Google Books, the Internet Archive, and indeed in many e-readers and websites of digitisation projects in libraries and archives, as well as resources such as Early English Books Online (EEBO), the British Library's Turning the Pages and many others. It is also central to one of the more important international standards in manuscript digitisation and the dissemination of images in cultural heritage, namely the International Image Interoperability Framework, or IIIF. Indeed, the relative accessibility of high-quality digital images and their very wide dissemination through the Web has contributed significantly to the prevalence of this model of the book. Perhaps a further reason for its prevalence, however, is that digital facsimiles seem to create a powerful illusion of perfect reproduction, at least for our image-centric culture which tends to place the burden of verisimilitude on accurate visual reproduction as opposed to any of the many other possible points of view, many of which are presented below.[3] Indeed, Apple explicitly designed its e-book reader on the principle of skeuomorphism, that is, in trying to emulate or in some sense reproduce an older technology in a newer one: here, the e-book was designed to emulate the printed book, and this emulation resided in the representation as a sequence of images, as well as the use of physical gestures to turn the virtual 'pages' (Pierazzo 2015: 159–160). The dominance of this view is evident in current terminology, where 'to digitise' a book typically means to take photographs of it and make these photographs available online. Indeed, it is not unusual to describe this process as 'putting a book online', as if the digital images were somehow the book itself rather than a visual representation of its pages. This slippage has reached extremes where websites have been described as "a little like waking up in the British Library after closing time" (Schmitt 2003: 5, cited by Kichuk 2007: 296), or the Library's own claim that one can "use our award-winning 'Turning the Pages™' software to leaf through our great books", formulations that again elide the difference between digital representation and material object.[4] Such slippage has rightly been criticised, and many reminders have been published of the difference between the object and image (Kichuk 2007; compare also Treharne 2013: esp. 476–477, among others), but the ease

3 I am grateful to Ségolène Tarte for this point.

4 <http://www.bl.uk/onlinegallery/virtualbooks/> [accessed 4 August 2020].

with which we equate identity with visual similitude remains pervasive and difficult to shake.

Despite these risks of mistaking the representation for the object, the fact remains that digital images of books are extremely useful in practice, as they are valuable for studying many aspects such as the text with its original spelling and punctuation, decoration, *mise en page*, palaeography and sometimes even aspects of the parchment, paper or other support on which the text is written. It is worth noting, however, that these aspects are not part of the digital model: a trained expert can obtain the text or study the palaeography from a digital image, for instance, but the computer itself has no 'knowledge' of text or palaeography from the image alone. One can add textual information, however, and this is often done in practice. One approach to this is taken by Google Books and the Internet Archive, where the text is not directly visible to the user but is instead used by the computer to enable the user to search. Here the user types in text and the site shows a corresponding part of the image; it is important to understand, however, that it is not the image that is searched but the text. When a user types in a word for searching, the computer looks for that word in the text. It also has information about the region of the image that presents each character, and so it can go from the text to show the corresponding part of the image. This approach is useful but is also rather misleading. It gives the impression that the image is being searched, and it also hides the text itself, meaning that it is difficult or even impossible to verify the quality of the text or to understand the principles by which it was created. This may seem a small detail but in fact it can be crucial: first, it is easy to assume that the text is a perfectly accurate reproduction of the original, but in practice this is by no means necessarily the case. Indeed, the text in Google Books was at least initially created by automatic OCR and not necessarily ever corrected, meaning that the text of some books contains tens or even a hundred or more errors per page. These errors are not random, either, but often relate to assumptions and biases about the underlying data, such as the assumption that the letter **s** has and always had the same form, namely **s**, as opposed to other forms such as **f**, and this in turn often leads to consistent misreading of **f** for **s**. One can see from the image if the result is a 'false positive', that is, if the system returned an incorrect result, but there is no way even of estimating how many 'false negatives' there are, that is, how many valid occurrences of the word were never found.[5]

5 OCR errors are just one source of bias in Google Books; other discussions can be found very easily, including Pechenick et al. (2015) to name just one.

A more rigorous approach is to display the image and text side-by-side, so that one can see the text and judge its quality and principles of production. This is more common for digital scholarly editions, and particularly for editions of manuscripts.[6] In modern printed books, there is normally a fairly straightforward correspondence between the original text and modern computer type. However, in a manuscript, this is by no means the case, and there are therefore many different editorial principles in terms of the degree to which one does or does not attempt to reproduce aspects of the original (for which see especially Pierazzo 2011 and Pierazzo 2015). Indeed, as has often been noted, and despite some claims to the contrary, there is no clear distinction between transcription and editing but the two lie on a spectrum and any transcription necessarily entails editorial decisions. For instance, it is common practice in digital scholarly editions to present the image of the page alongside different transcriptions, ranging from relatively diplomatic in the sense of reproducing more of the original features such as spelling, line breaks and abbreviations to more edited in the sense of normalising spelling, line breaks and so on. There are also examples where users can select such features individually according to their interests (an example of which is Thorn 2018). An interesting question has been raised about the value of such sites, specifically why scholars have gone to the very significant effort of reproducing details of the original book in the form of transcriptions when the image is there for anyone to see: Kevin Kiernan, for instance, has suggested that "the image-based scholarly edition subsumes the purpose of a diplomatic edition and removes the fruitless frustration of trying to preserve the exact layout, illumination, and physical appearance of a manuscript in print form" (2006: 266). One answer to this question is again that any transcription is an act of interpretation that results from an expert consideration of the book: transcriptions and indeed editions are themselves models of the book which embed a great deal of information and which communicate a scholarly interpretation that results from many expert decisions. Even a highly diplomatic transcription that supposedly reproduces the original is, in fact, a more or less explicitly encoded analysis, and presenting it is important also for transparency and aiding others to understand the principles by which the edition was created (Pierazzo 2011: 472–473; Sutherland and Pierazzo 2012: 207–208).

6 There are dozens of examples here, but some better-known ones include Jane Austen's Fiction Manuscripts (<https://janeausten.ac.uk/>) and those produced by the Edition Vizualisation Technology (EVT: <http://evt.labcd.unipi.it>).

4 The Book as Script, Layout and Decoration

A further model of the book as image is the burgeoning field often referred to as Document Analysis and Recognition. This branch of informatics involves the computational analysis of images of document pages in order to extract information about the object, particularly by analysing very large numbers of such images (potentially hundreds of thousands if not millions). Here the order of images may not be significant, but instead the emphasis is on the page of text or script, with common areas of focus including Handwritten Text Recognition (HTR: essentially the automatic transcription of handwritten text), automatic script identification (identifying, for example, manuscripts written in Gothic *textura* or *hybrida*), writer/scribe identification, and layout analysis, as well as other areas such as automatic dating and localisation, or automatically searching very large corpora of fragments to find examples that were likely once part of the same book. The International Conference on Frontiers in Handwriting Recognition (ICFHR), the International Conference in Document Analysis and Recognition (ICDAR), and the *International Journal of Document Analysis and Recognition* (*IJDAR*) are large and important events in computer science and demonstrate increasing interest in historical material, so much so that ICDAR is now routinely accompanied by the International Workshop on Historical Document Imaging and Processing (HIP@ICDAR).[7] This work draws on methods in computer vision and increasingly on machine learning, particularly deep learning, in order to analyse very large volumes of digitised material, for which researchers and engineers have developed and applied methods in machine vision and artificial intelligence to their analysis. The key principle is often not that the computer is necessarily correct in its identifications, but rather that it can propose possible identifications in, for instance, a corpus of hundreds of thousands of fragments, and these proposals can then be checked by a human expert thereby making tractable a problem that would otherwise be infeasible. In these cases, the manuscript (or corpus of manuscripts) is modelled with advanced statistics, in some cases even with a model loosely based on biological brains in the form of deep neural networks. These models can be very powerful and effective for specific questions such as automatic transcription or writer identification. To give just two examples, the Kraken engine can be trained to automatically transcribe manuscripts, inscriptions or other writing in many different scripts (Kiessling 2019), while

7 Once again the literature here is too vast to include here, but some useful starting-points include the *International Journal of Document Analysis and Recognition* itself, as well as the proceedings of ICDAR, HIP and ICFHR and articles such as Kestemont et al. (2017) and especially Cordell (2020).

other systems have been developed that can automatically classify images of Latin handwriting into Caroline, Gothic *textura*, *semi-hybrida* and so on (Kestemont et al. 2017). However, it is famously difficult for a human observer to understand precisely why a given result is obtained, or indeed to identify biases in training or algorithms that are used, so much so that explainable AI and algorithmic accountability are themselves now important areas of research with implications extending well beyond books and reaching increasingly into our society as a whole.[8]

Alongside this heavily computational work are other approaches which rely less on statistics and computation and more on explicitly modelling and visualising expert knowledge. One example of this is IconClass, which attempts to provide a hierarchical code for describing the iconographic content of manuscript (and other) artwork: so 11F is the code for "the Virgin Mary", 11F5 is "Madonna (i.e. Mary with the Christ-child) in the air, or on the clouds", 11F5(+31) is "Madonna (i.e. Mary with the Christ-child) in the air, or on the clouds (+ angels floating in the air)", and so on. A more symbolic approach was also developed for the Archetype framework used by DigiPal and its successors which focussed on structured descriptions of handwriting which researchers could enter into software. This involves manually drawing annotations on images of handwriting and entering descriptions of the letters. Expert users have already entered their own project-specific model and vocabulary for the script, including information about the components or essential elements of letters (for instance that **b**, **h** and **l** all contain ascenders); it therefore becomes relatively easy for researchers to find and compare forms in ways that are palaeographically meaningful, such as searching for examples of ascenders by a particular scribe or from a particular region. The emphasis here is firmly on knowledge creation through experimentation, exploration and visualisation, as well as the communication of evidence to support scholarly argument (Brookes et al. 2015; Stokes 2017). It also relates directly to larger questions in Digital Humanities and beyond about how one represents expert knowledge in systems that are tractable to the computer, connecting to areas and technologies such as ontologies, formal modelling, Linked Open Data and the Semantic Web, as discussed further below. In all this, then, it provides a further

8 For discussion in the context of palaeography, see Kestemont et al. (2017: S105–S109), Hassner et al. (2013), and Stokes (2009: 323); for libraries, see Cordell (2020: 12–16); and more generally Sculley and Pasanek (2008). Examples of these same issues in society more broadly include much recent discussion on such biases in AI for internet searches and even for sentencing criminals, for which see further (for example) Donohue (2019).

model of the book which is still based on the image but is again very different from those already discussed.

5 The Book as Codicological Structure

Although the emphasis when representing books has overwhelmingly focussed on text and image, there are of course other dimensions of books, and one of these is codicological structure. The emphasis here is on the book in its physical form comprising pages bound together, often (but not necessarily) enclosed in a protective cover. The topic of study includes the ways in which the individual pages are attached and arranged, the means by which the pages are held together, the structure of the binding, and the materials used in their production. Such a study can tell us a great deal about the original production and subsequent use and reuse of the book. Differences in the ways in which pages were created and bound together, and in details of the binding and so on, can allow us to identify the cultural milieu in which it was produced. Furthermore, most early modern and almost all medieval books went through various stages of rebinding, often taking parts of one book and binding them with parts of another, and close codicological analysis can help us piece together these stages in the object's biography. Similarly, close analysis of the structural units of the manuscript can give us further clues about production (for which see especially Andrist et al. 2013). If a text, scribe, and method of ruling all change at the boundary of a distinct physical unit of the book then it is more likely that these units were once separate and were only joined later; if a change falls within a physical unit, for instance in the middle of a page or across two pages that are part of the same sheet of parchment, then the conclusions are clearly different. Such analysis can also help to identify when the order of pages or sections of the book have been rearranged, as seems often to have happened in historical record books, for example, some of which may not have been bound until many years after their production.

Perhaps surprisingly given the current emphasis on the book as image, digital codicology has a longer history than digital palaeography. Some of the earliest databases on books have had a codicological focus, perhaps the best-known of which is SfarData which has been running in various forms for nearly half a century (Beit-Arié 1994). Similar to this is the increasing importance placed on manuscript fragments, including their relationship to one another and their transmission in other books for instance as binding fragments. Studies of this sort have resulted in numerous online projects and databases, such as Books within Books for Hebrew manuscripts and Fragmentarium, to name just

two.[9] Despite this, however, databases and related approaches to codicology seem at the time of writing to be rather less developed than those for palaeography or textual study, insofar as codicology does not seem to have received the same increase in interest that palaeographical or textual treatment of images has seen. Similarly, scholarly editions including those with the Text Encoding Initiative (TEI) typically include some codicological information in their introduction or discussion. However, TEI editions (like almost all editions, print and digital) tend to focus on the text rather than the physical structure of the book, and recording accurate and machine-processable details of the physical structure is much less standardised than for text, with the TEI Guidelines having very little to say on the subject. Nevertheless, there have been important innovations in digital methods for codicology, and some individual projects have begun to develop ways of encoding this information and of connecting it to the text. This is beginning to allow for automatic analysis of manuscript structures, such as identifying points of structural disjunction between parts of the manuscript which may suggest production at different times, or allowing people to test different hypotheses about changes in the order of pages in the book (Stokes and Noël 2019). Perhaps the most important of these is the VisColl project, which includes models for detailed codicological descriptions, and accompanying tools for automatically representing and visualising codicological structures (Porter et al. 2017; Campagnolo et al. 2020). This can be particularly useful for instance in helping to understand manuscripts that have particularly complex histories of production and reuse, as a result of which the codicological structure is highly irregular and also very informative as to the production of the manuscript (examples include Stokes and Noël 2019 and Stokes 2020).

6 The Book as Material and Three-Dimensional Object

Although part of the codicology, another dimension of the book which can potentially be modelled is its physicality, meaning for instance the materials from which it is made, as well as its form, shape and more affective aspects such as weight and even smell. Examples here include the use of enhanced imaging meth-

9 One might be surprised that a database record is considered here a representation of a book just as much as an image or text is, and granted the representation is less immediate and intuitive than an image or text, but it is nevertheless the same in principle: a database, like an image, comprises a selection and re-presentation of a specific subset of information that relates to the object.

ods to represent these aspects, such as Reflectance Transformation Imaging, 3D scanning or even Virtual Reality (Endres 2019). Even relatively standard technology can be helpful here, such as simply including images of the book from different angles, including close-up images of details of binding and so on.[10] Another simple but very effective approach uses video to show someone leafing through a manuscript in order to give a sense of the size, weight, and physicality, as well as explaining details such as the parchment.[11] Studies of paper structure and watermarks may also be modelled, including one approach which measured the level of dirt and wear on pages as a way of inferring how the book was used and which parts of the book received greater attention from readers (Rudy 2010). Material sciences can also play a role, with examples including Raman laser analysis, X-ray fluorescence, and even X-ray absorption spectroscopy for analysis of pigments, inks and other materials.[12] For instance, scholars of the Lindisfarne Gospels (London, British Library, Cotton Nero D.iv) had thought that the blue pigment was produced with lapis lazuli, but Raman laser analysis has more recently revealed the material to be indigo or woad (Brown 2003: 280–282; Brown and Clark 2004). Perhaps seemingly a small point, this detail is in fact important as indigo was available locally in northern England at the time the book was produced, whereas lapis lazuli had to be imported from modern-day Afghanistan, and therefore its presence in the manuscript would be important evidence for ninth-century trade routes as well as for the economic investment required to produce the book.

Some of these approaches can be conducted with mobile devices that can be carried into libraries, whereas others require a high-energy light source, or synchrotron, the core of which comprises a large ring potentially kilometres in diameter. Other uses for such equipment include using synchrotrons or smaller-scale X-ray fluorescence for seeing inside books, for instance producing images of bindings which can reveal the text on binding fragments without damaging the existing structure (Duivenvoorden et al. 2017). Scrolls of papyrus from Herculaneum that were badly burnt in the eruption of Mount Vesuvius have been virtually 'unrolled' and their inks analysed by using synchrotron light sources nor-

10 A good example here is the Datenbank zu Pracht- und Luxuseinbänden of the Bayerische Staatsbibliothek, which combines searchable metadata with a variety of different types of image delivered through IIIF. I thank the reviewers of this article for bringing this site to my attention.

11 Examples include the "Parchment Videos" and "Manuscript Orientations" from the Schoenberg Institute for Manuscript Studies at the University of Pennsylvania. The videos are available at <https://www.youtube.com/user/SchoenbergInstitute/videos> [accessed 22 July 2020].

12 For a sample of the techniques discussed in this section, see Brockmann et al. (2014) and Brockmann et al. (2018).

mally used for advanced research in physics and chemistry, with results including not only the reading of texts but also a significant re-evaluation of the use of metallic inks in manuscripts (Brun et al. 2016). Analysis of techniques of manufacturing and ruling pages can similarly give clues as to when the pages were first prepared for writing, perhaps indicating that different parts of the book were prepared at different times and/or places. More recently, non-destructive peptide and DNA analysis are beginning to reveal the species of animal that were used to produce the parchment pages, perhaps also indicating the date and place of production, a process which is proving highly significant for our understanding of medieval parchment production and indeed of animal husbandry (Fiddyment et al. 2015).

On the face of it, these models rely on physics, chemistry and biochemistry more than Digital Humanities. However, they are nevertheless a form of digitisation, insofar as they involve taking aspects of the original object and storing and representing these in digital form. Indeed, these methods largely depend on databases of comparable data to enable comparison. Identifications of ink might require using Raman laser spectroscopy to measure the interaction between the laser and the pigment. The resulting measurements are then compared against a database of responses from known materials, and this can allow identification of the unknown material in the pigment. This process is clearly scientific and quantitative, but it does also imply a digital model of the object as well as judgement and experience in factors such as which form of spectroscopy to use, alertness to other materials that can interfere with the results, such as the underlying parchment or later treatment of the pigments for instance as part of conservation. It therefore implies the need for standards and centralised databases in order that results can be used, interchanged and compared by different groups, a point to which we will return shortly.

7 The Book as (Mobile) Possession

Another important dimension of the book relates little or not at all to the characteristics of the object as such, but rather that which is done to it: its history of ownership, movement, collection, and association.[13] Tracking which books were popular, valued, in demand at different times and places tells a great deal about

13 For a sample of the extensive literature on this subject, see e.g. De Ricci (1930); Page (1993); Tite (1994); and Pearson (2008, esp. 93–140), as well as other works cited in this paragraph.

different societies and cultures and their values and influences. One example is the importation of books from Rome into Northumbria in northern England in the late seventh and early eighth century, which has been interpreted as emphasising the Roman orthodoxy of this church near the edge of early medieval Christendom (Brown 2003: 63–64). Another example is the very wide range of languages and scripts found in materials in the Dunhuang caves which bears witness to the mixing of cultures along the Silk Road. Taking more modern examples, one can trace recent movements of manuscripts as part of the book trade, with related inferences regarding cultural and indeed financial capital. One example of many is the movement of significant numbers of manuscripts from Europe to the United States, Australia and elsewhere in the so-called 'new' world, particularly at the end of the nineteenth and early twentieth century. This can usually be attributed directly to a number of interconnected factors: a period of relative wealth due to gold rushes and industrialisation, along with the sudden availability of a very large number of manuscripts on the market due to the sale of the collection of Thomas Phillipps, as well as the desire of the so-called 'new' rich to increase their cultural capital through the acquisition of manuscripts and other related heritage from the 'old' world (Cleaver 2018; Hubber 1993). This movement lends itself well to digital visualisations which can show animations and change over time, and indeed some important recent work has focussed on precisely this question. Perhaps the best-known is the Schoenberg Database of Manuscripts (SDBM 2020), which in fact – or at least in origin – is perhaps best thought of as a database of catalogues, and particularly sales catalogues, for premodern and principally Western manuscripts. This has proven a very rich source of data, not least for its information about the changing ownership of manuscripts, who acquired books from whom, when the transaction took place, and how much was paid, and the information has been central to some very significant research projects including the trans-Atlantic Mapping Manuscript Migrations project led by Toby Burrows, and the European Research Council project led by Laura Cleaver, among others.[14]

14 See <http://blog.mappingmanuscriptmigrations.org/about/> and <https://www.ies.sas.ac.uk/research-projects-archives/cultivate-mss-project>, respectively [both accessed 4 August 2020], as well as Burrows (2018).

8 Towards a Holistic Modelling: Linking and Interchange

At this point then, we have seen a range of different dimensions, points of view, models and representations of the book in digital form. This diversity is valuable in allowing very different viewpoints and disciplines, but it has also led to challenges and problems, the most obvious being the fragmentation of information into different disconnected silos of digital and analogue data. An obvious question, then, is that rather than trying to compile all this information into a single project, is it possible instead to interconnect existing digital content and, therefore, enable researchers at least to find information which has already been collected. The principle here is that different groups and individuals from different disciplinary backgrounds are already studying the same objects: a given manuscript has often been studied by palaeographers, art historians, linguists, philologists, biochemists, physicists, conservators, curators, data scientists, image scientists and many others. But how can the palaeographer even know where to look to find the results of the image scientist and, once these results are found, how can the palaeographer then correctly interpret them? In principle this is straightforward as one can simply search online bibliographies for a given shelfmark, but the practice is very much more difficult. One reason is the degree of ambiguity and implied knowledge in scholarship. There are presumably dozens – perhaps hundreds – of scholarly studies which refer to 'the Winchester Psalter', for instance. But there must also surely be dozens if not hundreds of different psalters which have been attributed to Winchester, so which one is 'the' Psalter? Normally this would be clear in the context of a specific scholarly discussion, but if one searches for all information about 'the Winchester Psalter' across all of scholarly literature, then the result will be an unusable mess. A simplistic response is to insist on unambiguous shelfmarks for all manuscripts, as indeed libraries try to do, but this is contrary to practice in some disciplines, and even then goes only part of the way to resolving the problem. Despite the best efforts of librarians, scholars often use different formats for referencing manuscripts in their publications, and while it may be clear to us that 'BL Add. 15350' and 'London, British Library, Additional MS 15350' refer to the same object, this is by no means obvious to a computer, a problem that complicates cross-searching of databases and catalogues. The problem also multiplies quickly. 'Winchester' may seem clear in the context of early medieval England, though even then we may wonder if it refers to the cathedral, the Old or New minster, or another institution in the same town. In a broader context, though, such as that of modern provenance studies, places called 'Winchester' are also found in the United States and

elsewhere, let alone a name such as 'Stanford'. Similar challenges also apply to people, such as the name 'King Henry' (or even 'King Henry II'), which can be highly ambiguous. Indeed, the more we broaden the scope of our data the more the ambiguities multiply, be that scope disciplinary, geographical, chronological or otherwise.

One proposed answer to this problem of ambiguities is the principle of authority lists and Linked Open Data. The principle here is that entities such as places and people can be given unique identifiers which are public, unambiguous, and easy to find. Examples include Pleiades and Geonames for places, the Virtual International Authority File (VIAF) and DBPedia for people and places, and so on. If I wish to refer to Winchester, for instance, I can turn to Geonames and search for the place that I want. Doing so, I find over nine hundred different records for places called Winchester, in countries including the United Kingdom, the United States, Jamaica and Brazil. The principle is that I can then decide which Winchester I mean, take the identifier for that particular Winchester, and include it in my data. This in turn means that someone else can come, search my data automatically for the identifier, and then be confident that we are both indeed referring to the same place. In this way, we can start to build up collections of information that are all interlinked, as well as being fully open and publicly accessible, and then one can relatively easily navigate across datasets, websites, editions, databases and so on, collecting the many different analyses of the object or objects under consideration. To date there is not yet a universal identifier for manuscripts, but work is already underway to provide this through the International Standard Manuscript Identifier project. Medieval Greek manuscripts in major libraries have already received identifiers through the Pinakes project, and significant progress has also been made in other contexts such as the Biblissima project in France. In principle, then, the route is clear and it is simply a matter of implementation.

This, at least, is the ideal, and a great deal of work has gone towards making it the reality. In practice, however, there remain some important questions and challenges which need to be addressed before these methods can become widely useful for 'real' scholarly research. Part of the problem is that ambiguity seems to be inevitable, at least in practice. A concept such as 'Winchester' may seem clear at first, but where *exactly* are its boundaries? The answer again depends very much on the context: the modern definition of municipal Winchester is one possibility, but the boundaries of Winchester today are clearly very different from that of the sixteenth century which are different in turn to that of the sixth century. Once again, none of these answers is inherently wrong, but each is appropriate only in certain contexts. Chronologies are also surprisingly difficult to treat. One question that has already been discussed in this context is when *exactly* the

'late eleventh century' begins and ends (Stokes 2015b). Is it before or after 'fourth quarter of the eleventh century', and is that before or after 's. xi ex.'? Indeed, when did the eleventh century itself begin? The first of January 1001? Lady Day 1001? The first of January 1000? According to which calendar? Even the definition of a single manuscript is by no means clear. As already discussed, manuscripts are mutable objects, and the object that we find in a library today is in many respects different from that which was first produced. So what *exactly* do we mean by 'BL Add. 15350': do we mean the manuscript as we find it now, or the manuscript that was first produced in the second quarter of the twelfth century, or the object in one of its many different intermediate states? Perhaps more importantly, how can you and I be sure that we mean the same thing? To be truly unambiguous, then, we have to be clear not only which manuscript we mean, but also which state in its lifetime. In practice this information may well often be unnecessary, but in other cases it will be critical. Was a given ink sample taken before or after treatment for conservation? Does a particular palaeographical analysis refer to the original script or the later additions? Does the catalogue description refer to the object before or after it was split up and rebound with two other books? And so on. Perhaps these details were not important to the person who created this hypothetical website, and so perhaps they were not recorded (one cannot record everything, as we know). But perhaps these same details are critical for someone else who wants to use the data. Linked Open Data does seem very promising, then, and some important projects on modelling manuscripts and documents that are relevant here include projects like ORIFLAMMS (Stutzmann 2013) but also Biblissima, IIIF, and work by Zhitomirsky-Geffet and Prebor (2016), among others. Indeed, we are already beginning to see benefits, for instance, through the Biblissima portal for western medieval manuscripts which provides access to over a dozen different repositories, databases, image-sets and more. However, the fact remains that this process of interlinking and collecting information cannot fully be automated but still very much requires intervention, interpretation and understanding from those wanting to use the data, and it seems likely that this will remain the case for the foreseeable future.

A question, then, is how we might approach this problem, and to a certain extent at least one possible answer may be through ontologies. Rather than providing links and identifiers, ontologies seek to define concepts as precisely as possible, as well as defining relationships between concepts: in short, they are a way of expressing a formal model of a domain in such a way that it can be understood by a person and also treated by a computer. Perhaps the most relevant ontology for our discussion is known as FRBRoo, which is in fact a combination of two models. The first, Functional Requirements for Bibliographical Records (FRBR), is developed by the International Federation of Library Associations

(IFLA) and is intended to model the relationships between entities in library catalogues such as the work in the abstract sense, different manifestations of that work such as different editions of the text, and the physical books (items) that one finds on the shelf. The second model is CIDOC-CRM, namely the Conceptual Reference Model (CRM) for the International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM). CIDOC-CRM is a large and complex model containing many entities, but, summarising very crudely, its emphasis is on objects and events that affect them such as a 'human-made object' being modified by an event and then acquired by an institution.

Turning to the combined FRBRoo model, then, the concept that we refer to as 'manuscript' normally corresponds to the entity 'F4 Manifestation Singleton'. This entity is defined as (or, more correctly, has the scope of) the following:

> [P]hysical objects that each carry an instance of F2 Expression, and that were produced as unique objects, with no siblings intended in the course of their production. (Bekiari et al. 2015: 57)

An F2 Expression, in turn, comprises

> the intellectual or artistic realisations of works in the form of identifiable immaterial objects, such as texts, poems, jokes, musical or choreographic notations, movement pattern, sound pattern, images, multimedia objects, or any combination of such forms that have objectively recognisable structures. The substance of F2 Expression is signs. Expressions cannot exist without a physical carrier, but do not depend on a specific physical carrier and can exist on one or more carriers simultaneously. Carriers may include human memory. (Bekiari et al. 2015: 55)

Put more simply, then, an 'F4 Manifestation Singleton' is a physical object that carries some sort of artistic realisation, for instance text. Furthermore, this object specifically does not have other identical copies ("has no siblings intended"): for instance, it is not the result of a process of mass production. A manuscript is therefore a type of F4 Manifestation Singleton, but a printed book is not. For our purposes, the 'F2 Expression' usually comprises the text in the manuscript, but it can also refer to any images, decoration, music notation and so on.

As well as defining entities, ontologies also specify relationships between them. For instance, the relationship between the physical object (F4) and its text (F2) is given by *R42 is representative manifestation singleton for*:

> This property identifies an instance of [F4] Manifestation Singleton that has been declared as the unique representative for an instance of F2 Expression by some bibliographic agency. [...] The musical text of Stanislas Champein's opera 'Vichnou' [an F2 Expression] [...] *R42 has representative manifestation singleton* the manuscript identified by shelfmark 'MS-8282' within the collections of the National Library of France, Department for Music (F4). (Bekiari et al. 2015: 107)

As we can see even from this simple example, ontologies are normally very abstract and often represent concepts very differently from the ways that we normally think of them. They also require a great deal of precision and attention, such as understanding the difference between *R17 created (was created by)*, *R18 created (was created by)*, *R21 created (was created through)*, and *R24 created (was created through)*, among others. Clearly this level of precision is not for the faint of heart, and as these relatively simple examples show, such ontologies require a great deal of careful thought to identify and distinguish the many different dimensions of an object as complex as a book. One may reasonably question the feasibility of sustaining such precision across a very large corpus, and the scope for error is very large since understanding the distinctions between different concepts typically requires a very deep knowledge of both the ontology and the subject domain in question. In principle, however, ontologies can allow for the sort of (nearly) automated knowledge collection described above, even reaching automated inference where a computer can deduce new information based on a dataset and the principles described by the ontology (for an example of which see Zhitomirsky-Geffet and Prebor 2016).

As we have seen, the ontology specifies concepts and relationships between them, but this begs the question how to apply it in practice to actual data. To do this we need to return to Linked Open Data and standard identifiers. Let us consider a hypothetical research question such as "find me details of all pigment analyses carried out on medieval manuscripts containing the *Roman de la Rose* of Guillaume de Lorris". This may seem clear and precise to us, but to a computer it is ambiguous: for instance, what *precisely* is meant by "the *Roman de la Rose* of Guillaume de Lorris"? Do we mean the work in general, or a specific version or edition? Is it the same as "*Romaunt of the Rose*", or "*Roman de la rose (extraits)*"? These questions are not abstract hair-splitting but in fact are critical, since all of these forms and more can be found in library catalogues and the computer must know which ones are relevant to the user's search. The idea here is that, as for places above, we can turn to a centralised authority's definitive list of works, and we can use this precise identifier in place of the imprecise title. For instance, VIAF lists the work known as *Roman de la Rose* by Guillaume de Lorris with ID 7464152140002811100009.[15] This then helps me disambiguate my question, since I can now refer to this particular ID, and I can also be confident that other

15 In fact, at the time of writing a search of VIAF for the Work "Roman de la Rose" returns twenty-two distinct results, some of which are by different authors, some of which comprise extracts, but some appear to be direct duplicates of the same work (compare, for instance IDs 74641521400 02811100009, 292531633 and 112015438089563029015 4). This point further demonstrates the practical difficulties of maintaining large-scale combined datasets for complex historical material.

databases which contain the same ID are referring to the same thing. Combining this with the concepts and relationships defined by FRBRoo allows us to navigate from the work to specific versions of it which would give us a further list of VIAF IDs, and from there we can find the records of physical manuscripts which carry those versions.[16] This presupposes a standard international identifier for all manuscripts, and such an identifier does not yet exist as mentioned above, but the International Standard Manuscript Identifier (ISMI) project has been created for precisely this purpose. An authority list such as ISMI could then give us identifiers of the relevant manuscripts, and a hypothetical future ontology would then provide the relationships from the object records to the relevant scientific analyses that we want from our query.

Of course such ontologies and standard identifiers are just the start: they would allow complex searches and even automated reasoning across many different catalogues and datasets, but it still requires human expertise to interpret those results in the context of specific research questions in order to construct an argument. Given the inevitable errors and ambiguities in all data including identifiers, as discussed above, it also seems certain that the results would need to be checked very carefully to ensure that the entities and concepts genuinely do correspond to what was intended. This, in turn, requires a high level of expertise across a range of different disciplines, and one may well question the degree to which one person should or even could have that expertise in practice. Nevertheless training schemes are already in place to help, such as the joint French-German training programme ManuSciences, which runs for a week in alternate years and provides theoretical and hands-on classes for early-career manuscript scholars in topics such as X-ray and Infrared fluorescence, multispectral imaging, and digital image analysis, as well as palaeography, codicology, the manufacture of paper, papyrus, parchment and inks, and so on.[17] Clearly one week of training is a long way from professional expertise, but the hope is that this can at least provide some awareness of the possibilities, limits and pitfalls when dealing with such data. As people are increasingly working together, and increasingly understanding the interdisciplinary issues around these different models, the possibility draws slowly nearer for a relatively complex meta-model that might allow the effective connecting of existing and future representations of the book.

16 In FRBR terms, then, the F1 Work with VIAF ID 7464152140002811100009 *R3 is realized in* F22 Self-Contained Expressions such as VIAF ID 179510163 which *R5i is a component of* F2 Expression which *R42 has representative manifestation singleton* F4 Manifestation Singleton (one or more specific manuscripts).

17 ManuSciences '19 Franco-German Summer School: <http://humanum.ephe.fr/en/node/80> [accessed 22 July 2020].

9 Conclusions: The Challenges and Illusions of Standardisation and Completeness

It should be clear at this point that many different dimensions of the book can be represented and analysed very effectively in a digital context, as long as due caution is exercised as it should always be in any critical analysis. It should also be clear that many other aspects of the book cannot yet be effectively captured or represented digitally, particularly those that are centred on the experiential dimension (the experience of handling a medieval object, its smell, its feel, its fetishization), as well as the related dimension of the book as cultural construct (its social value, its cultural capital, the book as symbol of status, learning, high versus low culture), and so on.[18] Nevertheless, it is becoming increasingly possible to imagine a world where many different types of information about books could be interconnected digitally, not to provide a single view but rather to enable one to find different views from different perspectives and to build analyses and arguments from these. Sustaining these multiple views remains a significant and perhaps under-recognised challenge: as Armando Petrucci (2001: 70) has pointed out in the context of palaeographical terminology, it is essential that we do not enforce a single point of view but instead that we allow different approaches to flourish, with the only requirement being that each view is rigorously founded on valid principles. In the same way, the digital standards that we need for interlinking and interchange must still leave space for these different viewpoints. However, this need is directly contrary to the purpose of a standard, which seeks to impose a single view in order to achieve meaningful interlinking and interchange, and managing this tension has received relatively little attention in practice. Research and innovation is often based largely or entirely in the outliers, the gaps, the parts that do not fit the usual template, and too strict a process of standardisation risks erasing these exceptions, whether literally by omitting them entirely or by forcing them to conform with the perceived norm: the richness and complexity of human culture, or at least that part of it which is interesting to most researchers in the humanities, is almost by definition not 'standard'. On the other hand, the more a standard is flexible and open, the more it becomes unwieldy in practice and the less useful it is for its core purposes of consistency, interoperability and interchange. This challenge has existed for decades and indeed has underlined

18 Discussions that approach this, some more critically than others, include De Hamel (2016), van Lit (2020: 292–310), and Treharne (2013). A useful, curated database containing many more digital representations of books and other aspects of medieval studies can be found in the Medieval Academy of America's Medieval Digital Resources at <http://mdr-maa.org> [accessed 5 August 2020].

much of the history of the TEI, but resolving it in a way that is practical and useful remains one of the core challenges of the Digital Humanities, both for Book History and for the field in general.

Works Cited

Andrist, Patrick, Paul Canart and Marilena Maniaci. 2013. *La syntaxe du codex: essai de codicologie structurale*. Turnhout: Brepols.

Beit-Arié, Malachi. 1994. "SFARDATA: The Henri Schiller Codicological Database of the Hebrew Palaeography Project, Jerusalem". *Gazette du Livre Médiéval* 25: 24–29. DOI: 10.3406/galim.1994.1280.

Bekiari, Chryssoula, Martin Doerr, Patrick Le Bœuf and Pat Riva (eds.). 2015. *Definition of FRBRoo: A Conceptual Model for Bibliographic Information in Object-Oriented Formalism*. Version 2.4. Den Haag: International Federation of Library Associations and Institutions. <http://www.cidoc-crm.org/frbroo/sites/default/files/FRBRoo_V2.4.pdf> [accessed 6 August 2020].

Brockmann, Christian, Michael Friedrich, Oliver Hahn, Bernd Neumann and Ira Rabin (eds.). 2014. *Proceedings of the Conference on Natural Sciences and Technology in Manuscript Analysis at the University of Hamburg*. <https://www.manuscript-cultures.uni-hamburg.de/MC/manuscript_cultures_no_7.pdf> [accessed 6 August 2020].

Brockmann, Christian, Oliver Hahn, Volker Märgner, Ira Rabin and H. Siegfried Stiehl (eds.). 2018. *Proceedings of the Second International Conference on Natural Sciences and Technology in Manuscript Analysis at the University of Hamburg*. <https://www.manuscript-cultures.uni-hamburg.de/MC/manuscript_cultures_no_11.pdf> [accessed 6 August 2020].

Brookes, Stewart, Peter A. Stokes, Matilda Watson and Debora Marques de Matos. 2015. "The DigiPal Project for European Scripts and Decorations". In: Aidan Conti, Orietta Da Rold and Philip A. Shaw (eds.). *Writing Europe 500–1450: Texts and Contexts*. Essays and Studies. Cambridge: Brewer. 25–58.

Brown, Katherine L. and Robin J. H. Clark. 2004. "The Lindisfarne Gospels and Two other 8th Century Anglo-Saxon/Insular Manuscripts: Pigment Identification by Raman Microscopy". *Journal of Raman Spectroscopy* 35: 4–12. DOI: 10.1002/jrs.1110.

Brown, Michelle P. 2003. *The Lindisfarne Gospels: Society, Spirituality and the Scribe*. London: British Library.

Brun, Emmanuel, Marine Cotte, Jonathan Wright, Marie Ruat, Pieter Tack, Laszlo Vincze, Claudio Ferrero, Daniel Delattre and Vito Mocella. 2016. "Revealing Metallic Ink in Herculaneum Papyri". *Proceedings of the National Academy of Sciences of the United States of America* 113: 3751–3754. DOI: 10.1073/pnas.1519958113.

Burnard, Lou, Fotis Jannidis, Elena Pierazzo and Malte Rehbein. 2010. "An Encoding Model for Genetic Editions". Revised Edition. Text Encoding Initiative. <http://www.tei-c.org/Activities/Council/Working/tcw19.html> [accessed 22 July 2020].

Burrows, Toby. 2018. "Digital Representations of the Provenance of Medieval Manuscripts". In: Matthew Evan Davis, Tamsyn Mahoney-Steel and Ece Turnator (eds.). *Meeting the Medieval in a Digital World*. Leeds: Arc. 203–222.

Burton, D. M. 1981a. "Automated Concordances and Word Indexes: The Fifties". *Computers and the Humanities* 15: 1–14. DOI: 10.1007/BF02404370.

Burton, D. M. 1981b. "Automated Concordances and Word Indexes: The Early Sixties and the Early Centers". *Computers and the Humanities* 15: 83–100. DOI: 10.1007/BF02404202.
Busa, R. 1980. "The Annals of Humanities Computing: The Index Thomisticus". *Computers and the Humanities* 14: 83–90. DOI: 10.1007/BF02403798.
Campagnolo, Alberto, and contributors. 2020. *Book Conservation and Digitization: The Challenges of Dialogue and Collaboration*. Leeds: Arc.
Ciula, Arianna. 2005. "Digital Palaeography: Using the Digital Representation of Medieval Script to Support Palaeographic Analysis". *Digital Medievalist* 1. DOI: 10.16995/dm.4.
Cleaver, Laura. 2018. "The Western Manuscript Collection of Alfred Chester Beatty (ca. 1915–1930)". *Manuscript Studies* 2: 445–482. DOI: 10.1353/mns.2017.0021.
Cordell, Ryan. 2020. *Machine Learning + Libraries: A Report on the State of the Field*. Washington D.C.: Library of Congress. <https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf> [accessed 23 July 2020]
De Hamel, Christopher. 2016. *Meetings with Remarkable Manuscripts*. New York: Penguin.
De Ricci, Seymour. 1930. *English Collectors of Books and Manuscripts (1530–1930) and their Marks of Ownership*. Cambridge: Cambridge University Press.
DeRose, Steven J., David G. Durand, Elli Mylonas and Allen H. Renear. 1990. "What Is Text, Really?". *Journal of Computing in Higher Education* 1: 3–26.
Donohue, Michael E. 2019. "A Replacement for Justitia's Scales? Machine Learning's Role in Sentencing". *Harvard Journal of Law and Technology* 32: 657–678.
Duivenvoorden, Jorien R., Anna Käyhkö, Erik Kwakkel and Joris Dik. 2017. "Hidden Library: Visualizing Fragments of Medieval Manuscripts in Early-Modern Book Bookbindings with Mobile Macro-XRF Scanner". *Heritage Science* 5. DOI: 10.1186/s40494-017-0117–6.
Endres, Bill. 2019. *Digitizing Medieval Manuscripts: The St Chad Gospels, Materiality, Recoveries, and Representation in 2D & 3D*. Leeds: Arc.
Fiddyment, Sarah, Bruce Holsinger, Chiara Ruzzier, Alexander Devine, Annelise Binois, Umberto Albarella, Roman Fischer, Emma Nichols, Antoinette Curtis, Edward Cheese, Matthew D. Teasdale, Caroline Checkley-Scott, Stephen J. Milner, Kathryn M. Rudy, Eric J. Johnson, Jiří Vnouček, Mary Garrison, Simon McGrory, Daniel G. Bradley and Matthew J. Collins. 2015. "Animal Origin of 13th-Century Uterine Vellum Revealed Using Noninvasive Peptide Fingerprinting". *Proceedings of the Natural Academy of Sciences* 112: 15066–15071. DOI: 10.1073/pnas.1512264112.
Frank, Roberta and Angus Cameron (eds.). 1973. *A Plan for the Dictionary of Old English*. Toronto: University of Toronto Press, in association with the Centre for Medieval Studies, University of Toronto.
Hassner, Tal, Malte Rehbein, Peter A. Stokes and Lior Wolf (eds.). 2013. "Computation and Palaeography: Potentials and Limits". *Dagstuhl Manifestos* 2: 14–35. DOI: 10.4230/DagMan.2.1.14.
Hubber, Brian. 1993. "'Of the Numerous Opportunities': The Origins of the Collection of Medieval Manuscripts at the State Library of Victoria". *La Trobe Library Journal* 13: 3–11.
Kestemont, Mike, Vincent Christlein and Dominique Stutzmann. 2017. "Artificial Paleography: Computational Approaches to Identifying Script Types in Medieval Manuscripts". *Speculum* 92: S86–S109. DOI: 10.1086/694112.
Kichuk, Diana. 2007. "Metamorphosis: Remediation in Early English Books Online (EEBO)". *Literary and Linguistic Computing* 22: 291–303. DOI: 10.1093/llc/fqm018.
Kiernan, Kevin. 2006. "Digital Facsimile in Editing". In: Lou Burnard, Katherine O'Brien O'Keeffe and John Unsworth (eds.). *Electronic Textual Editing*. New York: MLA. 262–268.

Kiessling, Benjamin. 2019. "Kraken: An Universal Text Recognizer for the Humanities". *Digital Humanities 2019 Book of Abstracts*. Utrecht: Utrecht University. <https://dev.clariah.nl/files/dh2019/boa/0673.html>.

McCarty, Willard. 2008. "Knowing: Modeling in Literary Studies". In: Susan Schreibman and Ray Siemens (eds.). *A Companion to Digital Literary Studies*. Oxford: Blackwell. 391–401.

O'Donnell, Daniel Paul. 2005. *Cædmon's Hymn: A Multi-Media Study, Edition and Archive*. Cambridge: Brewer.

Page, Raymond I. 1993. *Matthew Parker and his Books: Sandars Lectures in Bibliography Delivered on 14, 16 and 18 May 1990 at the University of Cambridge*. Kalamazoo, MI: Medieval Institute.

Pearson, David. 2008. *Books as History: The Importance of Books beyond their Texts*. London: British Library.

Pechenick, E. A., C. M. Danforth and P. S. Dodds. 2015. "Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution". *PLoS ONE* 10: e0137041. DOI: 10.1371/journal.pone.0137041.

Petrucci, Armando. 2001. *La descrizione del manoscritto: Storia, problemi, modelli*. Second ed. Rome: Carocci.

Pierazzo, Elena. 2011. "A Rationale of Digital Documentary Editions". *Literary and Linguistic Computing* 26: 463–477. DOI: 10.1093/llc/fqr033.

Pierazzo, Elena. 2015. *Digital Scholarly Editing: Theories, Models and Methods*. Farnham: Ashgate.

Pierazzo, Elena and Peter A. Stokes. 2010. "Putting the Text Back into Context: A Codicological Approach to Manuscript Transcription". In: Franz Fischer, Christiane Fritze and Georg Vogeler (eds.). *Kodikologie und Paläographie im Digitalen Zeitalter 2 – Codicology and Palaeography in the Digital Age 2*. Norderstedt: Books on Demand. 397–430.

Porter, Dot, Alberto Campagnolo and Erin Connelly. 2017. "VisColl: A New Collation Tool for Manuscript Studies". In: Hannah Busch, Franz Fischer and Patrick Sahle (eds.). *Kodikologie und Paläographie im Digitalen Zeitalter 4 – Codicology and Palaeography in the Digital Age 4*. Norderstedt: Books on Demand. 81–100.

Renear, Allen. 2004. "Text Encoding". In: Susan Schreibman, Ray Siemens and John Unsworth (eds.). *A Companion to Digital Humanities*. Oxford: Blackwell. 218–239.

Renear, Allen, Elli Mylonas and David Durand. 1996. "Refining our Notion of What Text Really is: The Problem of Overlapping Hierarchies". In: Nancy Ide and Susan Hockey (eds.). *Research in Humanities Computing 4: Selected Papers from the 1992 ALLC/ACH Conference*. Oxford: Oxford University Press. 263–280. Preprint version available at <http://hdl.handle.net/2142/9407>.

Rudy, Kathryn. 2010. "Dirty Books: Quantifying Patterns of Use in Medieval Manuscripts Using a Densitometer". *Journal of Historians of Netherlandish Art* 2. DOI: 10.5092/jhna.2010.2.1.1.

Schmidt, Desmond. 2010. "The Inadequacy of Embedded Markup for Cultural Heritage Texts". *Literary and Linguistic Computing* 25: 337–356. DOI: 10.1093/llc/fqq007.

Schmitt, John P. 2003. "Early English Books Online". *The Charleston Advisor* 4: 5–8.

Sculley, D. and Bradley M. Pasanek. 2008. "Meaning and Mining: The Impact of Implicit Assumptions in Data Mining for the Humanities". *Literary and Linguistic Computing* 23: 409–424. DOI: 10.1093/llc/fqn019.

Stokes, Peter A. 2009. "Computer-Aided Palaeography: Present and Future". In: Malte Rehbein, Patrick Sahle and Torsten Schaßan (eds.). *Kodikologie und Paläographie im Digitalen Zeitalter 1 – Codicology and Palaeography in the Digital Age 1*. Norderstedt: Books on Demand. 309–338.

Stokes, Peter A. 2015. "Digital Approaches to Palaeography and Book History: Some Challenges, Present and Future". *Frontiers in Digital Humanities* 2. DOI: 10.3389/fdigh.2015.00005.

Stokes, Peter A. 2015b. "The Problem of Digital Dating: A Model for Uncertainty in Medieval Documents". *Digital Humanities 2015: Book of Abstracts*. Sydney: University of Sydney.

Stokes, Peter A. 2017. "Scribal Attribution across Multiple Scripts: A Digitally-Aided Approach". *Speculum* 92: S65–S85. DOI: 10.1086/693968.

Stokes, Peter A. 2020. "Manuscript 367: A Study in (Digital) Codicology". In: Benjamin Albritton, Georgia Henley and Elaine Treharne (eds.). *Medieval Manuscripts in the Digital Age*. London: Routledge. 64–73. DOI: 10.4324/9781003003441–7.

Stokes, Peter A. and Geoffroy Noël. 2019. "Exon Domesday : Méthodes numériques appliquées à la codicologie pour l'étude d'un manuscrit anglo-normand." *Tabularia: Sources écrites des mondes normands médiévaux* [s.n.]. DOI: 10.4000/tabularia.4118.

Stutzmann, Dominique. 2013. "Ontologie des formes et encodage des textes manuscrits médiévaux: Le projet ORIFLAMMS". *Document Numérique* 16: 81–95. DOI: 0.3166/DN.16.3.69–79.

Sutherland, Kathryn and Elena Pierazzo. 2012. "The Author's Hand: From Page to Screen". In: Marilyn Deegan and Willard McCarty (eds.). *Collaborative Research in the Digital Humanities: A Volume in Honour of Harold Short, on the Occasion of his 65th Birthday and his Retirement, September 2010*. Farnham: Ashgate. 191–212.

TEI = The Text Encoding Initiative. 2020. *Guidelines for Electronic Text Encoding and Interchange*. Version 4.0.0. <https://www.tei-c.org/release/doc/tei-p5-doc/en/html/> [accessed 3 July 2020].

Thorn, Frank (ed.). 2018. "Exon Domesday Book: The Latin Text". In: Peter A. Stokes (ed.). *Exon: The Domesday Survey of South-West England*. Studies in Domesday, gen. ed. J. Crick. London: King's College. <http://www.exondomesday.ac.uk/digipal/manuscripts/1/texts/view/> [accessed 12 July 2020].

Tite, Colin C. 1994. *The Manuscript Library of Sir Robert Cotton*. The Panizzi Lectures. London: British Library.

Treharne, Elaine. 2013. "Fleshing Out the Text: The Transcendent Manuscript in the Digital Age". *postmedieval: a journal of medieval cultural studies* 4: 465–478. DOI: 10.1057/pmed.2013.36.

Van Lit, L. W. C. 2020. *Among Digitized Manuscripts: Philology, Codicology, Paleography in a Digital World*. Leiden: Brill.

Zhitomirsky-Geffet, Maayan and Gila Prebor. 2016. "Toward an Ontopedia for Historical Hebrew Manuscripts". *Frontiers in Digital Humanities* 3. DOI: 10.3389/fdigh.2016.00003.

Cited Online Projects

The following lists specialist online projects cited in the text, with the exception of very well-known cases such as Google Books, Project Gutenberg and the Internet Archive. All URLs were verified on 5 August 2020 and so were last accessed at that date.

Archetype. London: King's College. <http://www.archetype.ink>.
Biblissima: L'observatoire du patrimoine écrit du Moyen Âge et de la Renaissance. Paris: Equipex Biblissima. <https://biblissima.fr>.
BWB = Books within Books: Hebrew Fragments in European Libraries. 2020. Paris: École Pratique des Hautes Études. <http://hebrewmanuscript.com>.
Datenbank zu Pracht- und Luxuseinbänden. Bayerische Staatsbibliothek. <https://einbaende.digitale-sammlungen.de>.
DBPedia. <https://wiki.dbpedia.org>.
DigiPal = Digital Resource and Database of Palaeography, Manuscript Studies and Diplomatic. 2011–2014. London: King's College. <http://www.digipal.eu>.
EVT = Edition Visualization Technology. Pisa: University of Pisa. <http://evt.labcd.unipi.it>.
EEBO = Early English Books Online. ProQuest. <https://www.proquest.com/eebo>.
Fragmentarium: Laboratory for Medieval Manuscript Fragments. 2020. Fribourg: University of Fribourg. <https://www.fragmentarium.ms>.
Geonames. <https://www.geonames.org>.
IconClass: A Multilingual Classification System for Cultural Content. <http://iconclass.org>.
IIIF = International Image Interoperability Framework. <https://iiif.io>
Jane Austen's Fiction Manuscripts. London: King's College. <https://janeausten.ac.uk/>.
Kraken. <http://kraken.re>.
Mapping Manuscript Migrations: Navigating the Network of Connections between People, Institutions and Places within European Medieval and Renaissance Manuscripts. <https://mappingmanuscriptmigrations.org/>.
Medieval Digital Resources: A Curated Guide and Database. 2020. The Medieval Academy of America. <http://mdr-maa.org>.
Oxford Text Archive. Oxford: University of Oxford. <https://ota.bodleian.ox.ac.uk/>.
Pinakes – Πίνακες: Textes et manuscrits grecs. 2016. Paris: Institut de Recherche et d'Histoire des Textes. <https://pinakes.irht.cnrs.fr>.
Pleiades. New York: Institute for the Study of the Ancient World. <https://pleiades.stoa.org>.
SDBM = Schoenberg Database of Manuscripts. 2020. Philadelphia, PA: University of Pennsylvania Libraries. <https://sdbm.library.upenn.edu>.
SfarData = Beit-Arié, Malachi et al. [n.d.]. SfarData: The Codicological Database of the Hebrew Palaeography Project. Jerusalem: The National Library of Israel. <http://sfardata.nli.org.il/>.
Turning the Pages. London: The British Library. <http://www.bl.uk/turning-the-pages/>.
VIAF = Virtual International Authority File. 2010–2019. OCLC. <http://viaf.org>.
VisColl = Modeling and Visualizing the Physical Construction of Codex Manuscripts. <https://viscoll.org>.