Schriften des Instituts für Dokumentologie und Editorik — Band 11

Kodikologie und Paläographie im digitalen Zeitalter 4

Codicology and Palaeography in the Digital Age 4

herausgegeben von | edited by

Hannah Busch, Franz Fischer, Patrick Sahle

unter Mitarbeit von | in collaboration with

Bernhard Assmann, Philipp Hegel, Celia Krause

2017
BoD, Norderstedt

Bibliografische Information der Deutschen Nationalbibliothek:
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de/ abrufbar.

Digitale Parallelfassung der gedruckten Publikation zur Archivierung im Kölner Universitäts-Publikations-Server (KUPS). Stand 4. September 2017.

SPONSORED BY THE

Federal Ministry
of Education
and Research

Diese Publikation wurde im Rahmen des Projektes eCodicology (Förderkennzeichen 01UG1350A-C) mit Mitteln des Bundesministeriums für Bildung und Forschung (BMBF) gefördert.
Publication realised within the project eCodicology (funding code 01UG1350A-C) with financial resources of the German Federal Ministry of Research and Education (BMBF).

2017

Herstellung und Verlag: Books on Demand GmbH, Norderstedt
ISBN: 978-3-7448-3877-1
Einbandgestaltung: Julia Sorouri, basierend auf Vorarbeiten von Johanna Puhl und Katharina Weber; Coverbild nach einer Vorlage von Swati Chandna.
Satz: LuaTEX und Bernhard Assmann

VisColl: A New Collation Tool for Manuscript Studies

Dot Porter, Alberto Campagnolo, Erin Connelly

Abstract

The principal physical feature of the book in codex format, the gathering structure, is usually not visualized within digitization projects. If this information is recorded at all, it is generally done with the use of collation formulas. There is not a standard schema for manuscript collation formulas and not all practices are able to record accurately the structure of books. There have been some attempts in the past to describe gathering structures in more formalised ways. VisColl is building on past experiences and strives to describe, visualize, and communicate the gathering structure of books. Successful applications of the new tool are presented as examples. Future versions will add functionality to link physical details of a manuscript with additional information about the content, which will enable a complete mapping of a physical manuscript.

Zusammenfassung

Das Hauptmerkmal eines Buchs im Kodexformat, seine Lagenstruktur, wird in Digitalisierungsprojekten gewöhnlich nicht visualisiert. Wenn diesbezügliche Informationen überhaupt festgehalten werden, so geschieht dies in aller Regel unter Verwendung fomalisierter Lagenbeschreibungen, für die es bisher kein allgemein anerkanntes Standardformat gibt. Auch eignen sich vorherrschende Beschreibungspraktiken nicht immer für eine detailgenaue Erfassung der Lagenstruktur. In der Vergangenheit gab es einige Versuche, Lagenbeschreibungen stärker zu fomalisieren. VisColl knüpft an diese Erfahrungen an und ist bestrebt, Lagenstrukturen von Büchern zu beschreiben, zu visualisieren und zu vermitteln. In diesem Artikel soll anhand einiger Beispiele veranschaulicht werden, wie das neue Tool bereits erfolgreich angewendet wird. In Zukunft sollen Funktionalitäten hinzugefügt werden, über die sich Angaben zum materiellen Zustand einer Handschrift mit inhaltlichen Informationen verbinden lassen, um auf diese Weise ein umfassendes Verzeichnen des physischen Objekts zu ermöglichen.

1 Introduction

VisColl is a digital tool designed to help scholars to visualize the physical construction of medieval codex manuscripts, also known as *collation*. Manuscript codices, like

Kodikologie und Paläographie im Digitalen Zeitalter 4 – Codicology and Palaeography in the Digital Age 4. Hrsg. Hannah Busch, Franz Fischer und Patrick Sahle, unter Mitarbeit von Bernhard Assmann, Philipp Hegel und Celia Krause. Schriften des Instituts für Dokumentologie und Editorik 11. Norderstedt: Books on Demand, 2017. 81–100.

modern books, consist of a series of pages, however the pages are physically connected in ways that are not always clear to the reader. Manuscripts are built of *quires*, which are normally three to six sheets of parchment or paper (or both), stacked and then folded in half, and then (usually) sewn together in the fold. The folded sheets are called *bifolia* (literally "two folios"), and the pages are called *folios* or *leaves*. Thus the first leaf in a quire is literally half of a bifolia, while the last leaf in a quire is the other half. We say that these two leaves are *conjoined*. Quires are sewn together to create codex books. In addition to sets of bifolia, a quire may have leaves cut out, or added either during the writing process or later. It is these details of physicality—quires, bifolia, added and removed leaves—that the current version of VisColl seeks to describe and visualize. Future versions will add functionality to link physical details of a manuscript with additional information about the content, which will enable a complete mapping of a physical manuscript.

VisColl was conceived in the mid-2000s by Dot Porter during her work at the Collaboratory for Research in Computing for Humanities at the University of Kentucky (UKY). Porter developed the tool in order to address issues she encountered in effectively visualizing standard descriptions of manuscripts in scholarly works. For instance, in *Beowulf and the Beowulf Manuscript* Kevin Kiernan (1981) uses the physical construction of the manuscript to make arguments about the dating of the text (separate from the dating of the manuscript itself). In addition, Ben Withers (of UKY), in *The Illustrated Old English Hexateuch, Cotton MS. Claudius B.IV: the Frontier of Seeing and Reading in Anglo-Saxon England* (2007) similarly used a detailed collation statement of the manuscript as the backbone for his investigation of the construction of the manuscript. There are numerous examples of scholarly works that build an argument about the dating and construction of manuscripts based on the collation of the physical object. In consulting such works, Porter saw an opportunity to enable readers to better visualize the structure of the object beyond the limitations of traditional formulas, diagrams, and collation statements.

1.1 Collation formulas

Traditionally, information on the gathering structure of books is recorded in highly dense expressions, referred to as *collation formulas*. These describe the sequence of bifolia (and singletons) within book gatherings. All formulas contain the same basic information, but this may be presented in a variety of ways, and their decoding in relation to the physical appearance of the object that they describe can prove challenging.

The following examples show different styles of collation formulas:

[1] i, 1-9 (8), 10 (6), 11-20 (8), 21 (7), i
[2] I-III^{8}, IV^{10}, V-IX^{8}

[3] IV(32), IV-1(40), 9 IV(120), IV-4
[4] $1\text{-}4^{8}$, 5^{2}, $6^{4\text{-}1}$, $7\text{-}10^{10}$
[5] 2°: $^{\pi}A^{6}$($^{\pi}$A1+1, $^{\pi}$A5+1.2), A-$2B^{6}$, $2C^{2}$, a-g^{6}, $^{\chi}2g^{8}$, h-v^{6}, x^{4}, “gg3.4”(±”gg3”), ¶-$2¶^{6}$, 3¶1, 2a-$2f^{6}$, $2g^{2}$, “Gg^{6}”, $2h^{6}$, 2k-$3b^{6}$

Of these, the first four illustrate different patterns of collation formulas utilized for manuscripts, whilst the latter shows a bibliographical description of the gathering assembly of a printed book.

Formulas to describe manuscripts and printed books aim at the same scope: representing the gathering structure of a book in codex format; there are, however, some fundamental differences between the two schools. In manuscript studies collation formulas represent book structures exactly as they are, whilst bibliographical formulas represent the ideal copy of the printed book, and not the state of specific exemplars. In addition, manuscript studies—unlike the case of printed books and their bibliographical description—lack a standard for drafting collation formulas that is approved and employed by all scholars. As it can be seen in the examples above—[1] to [4]—some schemas use Roman numerals to signal the sequence of gatherings, whilst others prefer Arabic numerals; some use superscripts, and some show the number of pages in a group. Without being familiar with specific schemas, the interpretation of manuscript collation formulas can be problematic. Nonetheless, for the most part, both bibliographical collation formulas and the various styles of those employed in manuscript studies share a set of information units that are necessary to describe the arrangement of the sheets within textblocks.

Zappella (1996) and Andrist et al. (2013) provide a comprehensive overview of the state of the art of collation formulas in bibliography and manuscript studies.

There have been some attempts to formulate collation schemas, and to model the gathering structure of books, in a way that such information could be easily parsed by computers. Gerardy (1972) describes a numerical system[1] to encode gathering structures of manuscripts. This collation format works like a decimal cataloguing system, and assigns numbers to gatherings (GG), bifolia (BB), and folios (f) or sides (s) of the leaves (i.e. recto and verso) according to a specific template:[2] GG.BB.f|s. This system assigns a unique numerical code to each element and accommodates for irregular structures by allocating special codes to stubs (7) and missing leaves (0). In this manner, also the frequent—and difficult to model—case of quires within quires, can be encoded. However, the unique numerical IDs in themselves do not communicate the relationships that exist amongst the bifolia within a gathering, i.e., looking at the example in figure 1, knowing that a folio ID is 03.03.01 and that of another is 03.04.01 does not convey the fact that bifolio 03 is an example of a quire

[1] See ‘Pagination décimale’ in Muzerelle 1985.
[2] The full template also accommodates for stubs and missing leaves, and not just full folios and their sides.

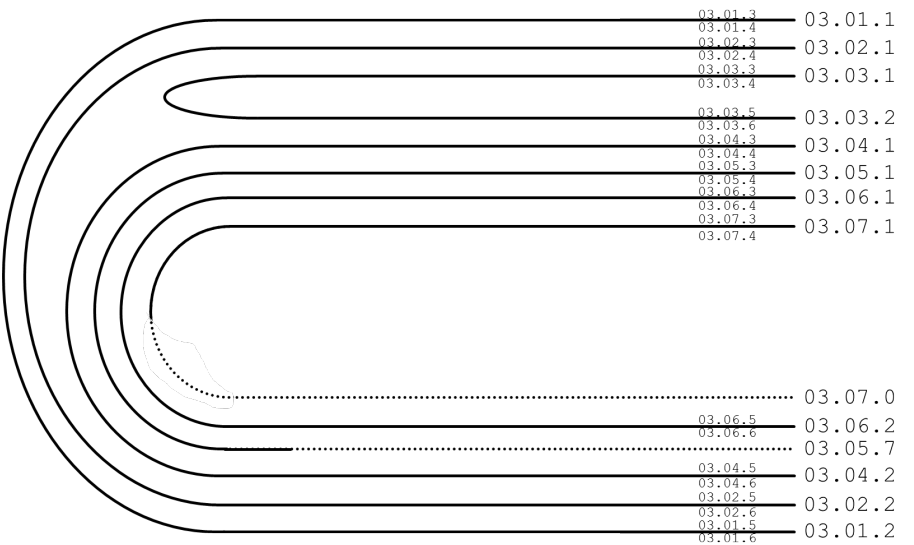

Figure 1: Example of decimal pagination for a complex gathering 3 in a manuscript (after Gruijs 1974, 254, schema 2; and Gerardy 1980, 45).

within quire: only the diagram or the full array of the gathering's IDs yield this important piece of information, and this is a significant flaw of the system.

In 2004, the TEI Physical Bibliography Workgroup put together a proposal to expand the collation recording capabilities of the TEI-MS model (TEI Workgroup on Physical Bibliography 2004). Considering that the physical structure of a book can be conceptualized as a series of hierarchically-organized objects, such as gatherings which contain leaves, and pages which contain lines of text, the working group advanced two distinct models to be integrated within a `<collation>` element in the `<msDescription>` or `<bookDescription>` of the TEI header.

On the one hand, in `<collationFormula>`, the typical layout of collation formula schemas was transposed within a hierarchical structure containing the elements that make up a full bibliographic description of gathering structures: a list of gatherings, an indication of the total number of leaves, pagination statements, etc.

On the other hand, the working group modelled a complex series of elements—i.e. `<gathering>`, `<leaf>`, `<page>`—to directly describe the physical structure of books in codex format.

This module did not become part of TEI P5 (TEI 2016b), and as a result, the standard way of recording collation information within TEI-based descriptions is to insert,

within a <collation> element, using informal prose, or other notational conventions, a description of a book's current and original arrangement of leaves and gatherings (TEI 2016a). The guidelines do not, therefore, prescribe any specific collation notation, but typical collation formulas can be included in a <formula> element as text. The ideas brought forward by the Physical Bibliography working group were, however, valuable, and, as it will be seen, they are being integrated in our own modelling of the gathering structures.

```
<gathering>
 <leaf xml:id="leaf1" conjunct="leaf8">
  <page xml:id="p1" sheetSide="1" cutFromN="p8" W="p16"/>
  <page xml:id="p2" sheetSide="2" cutFromN="p7" E="p15"/>
 </leaf>
 <leaf xml:id="leaf2" conjunct="leaf7">
  <page xml:id="p3" sheetSide="2" cutFromN="p6" W="p14"/>
  <page xml:id="p4" sheetSide="1" cutFromN="p5" E="p13"/>
 </leaf>
 <leaf xml:id="leaf3" conjunct="leaf6">
  <page xml:id="p5" sheetSide="1" cutFromN="p4" W="p12"/>
  <page xml:id="p6" sheetSide="2" cutFromN="p3" E="p11"/>
 </leaf>
 <leaf xml:id="leaf4" conjunct="leaf5">
  <page xml:id="p7" sheetSide="2" cutFromN="p2" W="p10"/>
  <page xml:id="p8" sheetSide="1" cutFromN="p1" E="p9"/>
 </leaf>
 <leaf xml:id="leaf5" conjunct="leaf4">
  <page xml:id="p9" sheetSide="1" cutFromN="p16" cutFromE="p12" W="p8"/>
  <page xml:id="p10" sheetSide="2" cutFromN="p15" cutFromW="p11" E="p7"/>
 </leaf>
 <!-- [...] -->
</gathering>
```

Listing 1: Example of encoding according to the 2004 Physical Bibliography proposal.

1.2 Viewing digitized manuscripts

Digitized medieval manuscripts are typically viewed through single-page or facing-page interfaces, which lack the physical cues present in a physical book, i.e., the size of the book, its thickness, details of the parchment or paper, etc. Indeed, even facing-page interfaces do not usually show a picture of book openings at all, but rather they are composites made with two images: one of the left-side page and another of the right-side page. These images would have been taken at different times. Typically all images of one side pages are taken first, e.g. all the rectos, then of the other side, and then file names or structural metadata are used to order the files correctly in post processing. Most digital libraries provide some information on the pages depicted, and views other than single-page or facing-page: all provide information on the folio number and the side (recto or verso) shown; some indicate the quire number (e.g. The British Library et al. 2016), and some offer a variety of viewing modes, including pages of thumbnails (e.g. *E-Codices* 2016) or thumbnails presented filmstrip-style

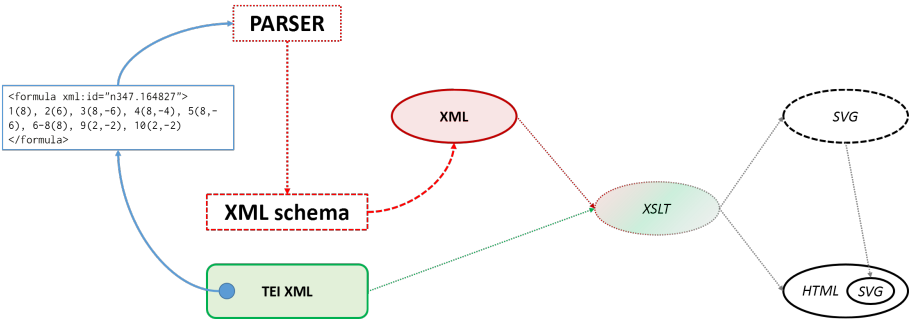

Figure 2: Diagram showing the pipeline of the prototype system.

across the bottom of a page (e.g. *Vitae Sanctorum* 2016). However, again, for the most part, the focus of these resources is on the page, rather than on the physical object. Even the Turning the Pages™ software (*Turning the Pages™* 2016), conceived by the British Library in 1996—and developed by Armadillo Systems (*Armadillo Systems* 2009) since 2001—, which, since version 2.0 (2006), has produced realistic three-dimensional books (including the ability to mimic the different movement of paper and parchment pages as these are turned), lacks any modelling of the gathering structure. To present knowledge, there is no institutional digital library that describes the physicality of manuscripts outside of the standard Physical Description section of the manuscript records and collation formulas.

In VisColl, we first model the collation of manuscripts in an XML format and then process that model in various ways, currently providing both diagrams and formulas, but potentially in other novel ways as well. For instance, in addition to visualizing the physical structure of a manuscript, the Beta Version of VisColl currently under development enables users to create taxonomies describing the content of the manuscript, and other elements, and then the system links those taxonomies to the physical structure, which produces a more robust and descriptive visualization than is possible in the current system.

This paper will document the stages of the development of VisColl, from its conception to its current instantiation, highlighting the steps taken and the reasoning behind each new actualization of the project. The current state of development can be found at the VisColl's GitHub page (Porter 2016a), which documents each new build, and from which the project's code can be downloaded.

2 Proof of concept

In July 2013, work started on the proof of concept for VisColl (cf. Porter 2013). This was established by taking an existing collation formula schema—i.e. that was devised

by William Noel (2011) for the Digital Walters project[3]—and processing it into two separate visualizations: quire diagrams (showing how leaves pair into bifolia) and what the project calls *Bifolia View*, where images of each page are viewed alongside the other half of the sheet as bifolia (useful in cases where it is not clear whether the sheets were written/illustrated before or after they were gathered into quires). In practice these two visualizations were presented together, with a quire diagram on the left side of a page and bifolia view presented to the right. Outside of digital practice, this perspective is only achieved by disbinding a manuscript. Figure 2 shows a diagram of the prototype pipeline: the collation formula (presented as text content of `<formula>`) was extracted from the TEI XML and parsed into XML. This collation XML was then processed along with the TEI XML, and the collation XML was converted into SVG diagrams while the image files, listed in the `<facsimile>` section of the TEI XML, were collected and arranged into bifolia. The bifolia are displayed on an HTML page with the SVG diagrams embedded alongside. Each quire was presented on its own web page using a combination of HTML for the page wrapper and SVG for the quire diagram. Each bifolia was presented in a row with the "active" bifolia highlighted in the diagram. The images were presented alongside the diagram: first, with the "inside" of the sheet facing up and then with the "outside" (as though the sheet were turned over; see fig. 3).

The great benefit of the proof of concept approach is that it enabled the batch processing of several manuscripts at once. At one point, Porter created visualizations in a single afternoon for all the manuscripts on the Digital Walters website that had associated collation formulas. There were, however, several downsides to this approach. The main problem is that it was entirely dependent on a specific collation formula schema. Unlike printed books, there is no single standard for manuscript collation formulas. Although all formulas will contain the same basic information, it may be presented in various ways.

3 Alpha version

The alpha version of VisColl had two main aims. The first, derived from the proof of concept, was to move from a formula-based approach to a model-based approach. The second was to build a system that would be publicly available and easily accessible. This second aim was a weakness of the proof of concept version. Although the proof of concept scripts were available on GitHub, with basic documentation on how to run them, it was difficult for users to run them correctly, especially as the scripts were not able to process any but the most basically constructed manuscripts. With this in mind, the alpha version system was built in two parts, with a third step that a user

[3] A website making available the digital images and metadata of the manuscripts held at the Walters Art Museum (*The Digital Walters* 2016).

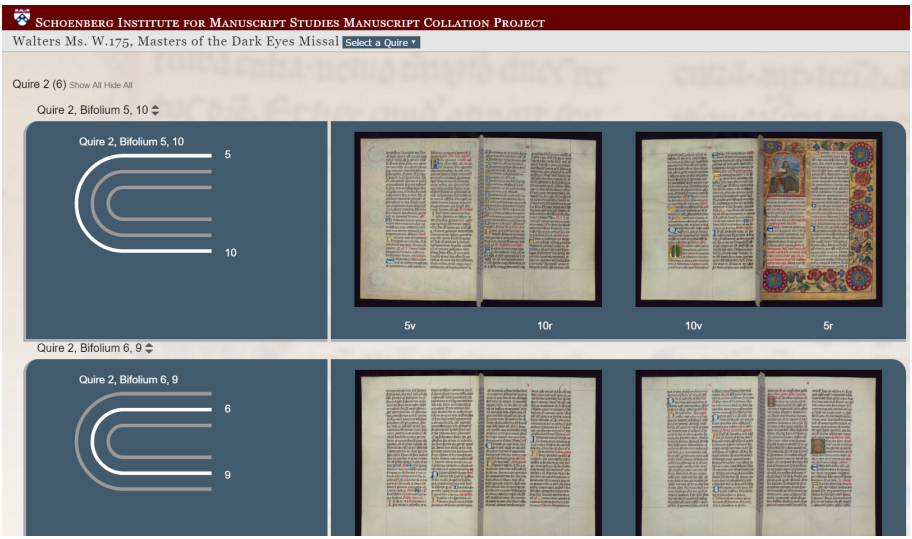

Figure 3: Screenshot of Baltimore, Walters Ms.175, which was visualized with the proof of concept prototype. Note the two views (inside/outside) for each bifolium, and the collation diagrams, highlighting which set of leaves are being visualized.

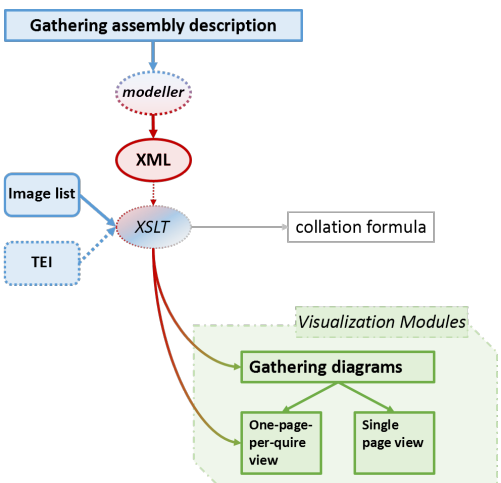

Figure 4: Diagram of the alpha version pipeline with its three steps (Collation Modeler, Image List, and Visualization Generation) and four XSLT outputs (collation formula, collation diagrams, one-page-per-quire visualization, single-page visualization).

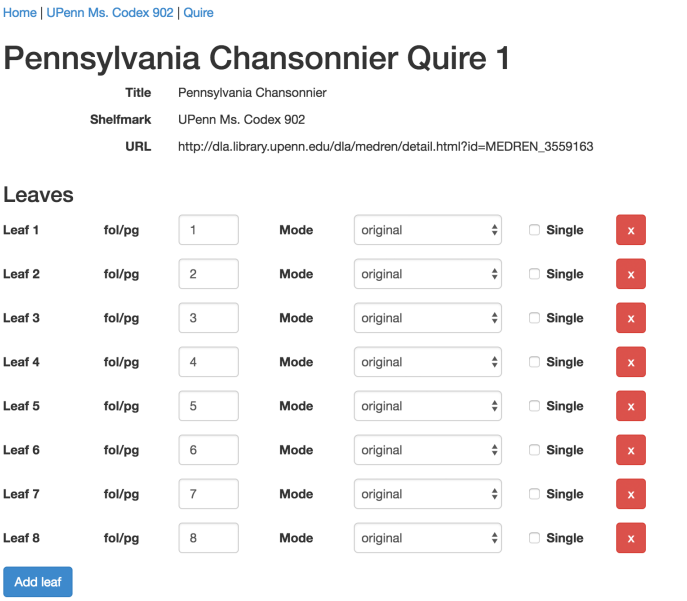

Figure 5: A screenshot of the Collation Modeler showing the complete construction of a quire for UPenn Ms. Codex 902.

would need to perform on their own. These three steps are: Collation Modeler, Image List, and Visualization Generation (see fig. 4).

3.1 Collation Modeler

The Collation Modeler,[4] built in Ruby on Rails by Doug Emery at the University of Pennsylvania, enables a user to construct and export a collation model, which is specifically formatted to be input into the Visualization Generation tool. In the current version of the Collation Modeler, using a form-based interface a user builds a number of quires and then identifies each leaf in the quire as original (to the manuscript), added (to the manuscript), missing (from the manuscript) or replaced (the original leaf having been removed and replaced with another leaf containing the same text as the original).

[4] The publicly accessible Collation Modeler (*Collation Modeler* 2016) and the Collation Modeler code on GitHub (Emery 2016).

An XML file containing the collation model is generated from the Collation Modeler to be used to create visualizations. In the current version, visualizations can't be generated directly from the Collation Modeler, which we recognize as a barrier for use.

```
<?xml version="1.0"?>
<manuscript>
 <url>http://dla.library.upenn.edu/dla/medren/detail.html?id=MEDREN_3559163</url>
 <title>Pennsylvania Chansonnier</title>
 <shelfmark>UPenn Ms. Codex 902</shelfmark>
 <quire n="1">
  <leaf n="1" mode="original" single="false" folio_number="1" conjoin="8"
       position="1" opposite="8"/>
  <leaf n="2" mode="original" single="false" folio_number="2" conjoin="7"
       position="2" opposite="7"/>
  <leaf n="3" mode="original" single="false" folio_number="3" conjoin="6"
       position="3" opposite="6"/>
  <leaf n="4" mode="original" single="false" folio_number="4" conjoin="5"
       position="4" opposite="5"/>
  <leaf n="5" mode="original" single="false" folio_number="5" conjoin="4"
       position="5" opposite="4"/>
  <leaf n="6" mode="original" single="false" folio_number="6" conjoin="3"
       position="6" opposite="3"/>
  <leaf n="7" mode="original" single="false" folio_number="7" conjoin="2"
       position="7" opposite="2"/>
  <leaf n="8" mode="original" single="false" folio_number="8" conjoin="1"
       position="8" opposite="1"/>
 </quire>
 <quire n="2">
  <leaf n="1" mode="original" single="false" folio_number="9" conjoin="8"
       position="1" opposite="8"/>
  <leaf n="2" mode="original" single="false" folio_number="10" conjoin="7"
       position="2" opposite="7"/>
  <leaf n="3" mode="original" single="false" folio_number="11" conjoin="6"
       position="3" opposite="6"/>
  <leaf n="4" mode="original" single="false" folio_number="12" conjoin="5"
       position="4" opposite="5"/>
  <leaf n="5" mode="original" single="false" folio_number="13" conjoin="4"
       position="5" opposite="4"/>
  <leaf n="6" mode="original" single="false" folio_number="14" conjoin="3"
       position="6" opposite="3"/>
  <leaf n="7" mode="original" single="false" folio_number="15" conjoin="2"
       position="7" opposite="2"/>
  <leaf n="8" mode="original" single="false" folio_number="16" conjoin="1"
       position="8" opposite="1"/>
 </quire>
 <quire n="3">
  <leaf n="1" mode="original" single="false" folio_number="17" conjoin="8"
       position="1" opposite="8"/>
  <leaf n="2" mode="original" single="false" folio_number="18" conjoin="7"
       position="2" opposite="7"/>
  <leaf n="3" mode="original" single="false" folio_number="19" conjoin="6"
       position="3" opposite="6"/>
  <leaf n="4" mode="original" single="false" folio_number="20" conjoin="5"
       position="4" opposite="5"/>
  <leaf n="5" mode="original" single="false" folio_number="21" conjoin="4"
       position="5" opposite="4"/>
  <leaf n="6" mode="original" single="false" folio_number="22" conjoin="3"
```

```
      position="6" opposite="3"/>
  <leaf n="7" mode="original" single="false" folio_number="23" conjoin="2"
      position="7" opposite="2"/>
  <leaf n="8" mode="original" single="false" folio_number="24" conjoin="1"
      position="8" opposite="1"/>
 </quire>
 <!-- [...] -->
</manuscript>
```

Listing 2: Example XML code of the collation model for UPenn Ms. Codex 902.

3.2 Image list

The image list is a file required by the Visualization Generation tool. If the user wants a bifolia view, the image list must include folio numbers or page numbers along with URLs to the corresponding image file. The system does not import these images, rather the HTML output points to the image files wherever they reside on the web. If the user does not need a bifolia view an image list file still needs to be uploaded to the Visualization Generation tool, but it may be an empty file.

In the alpha version, the image list needs to be built in an Excel spreadsheet with page/quire numbers in the first column and file URLs in the second column. The file is saved as an XML spreadsheet, and this file is fed into the Visualization Generation tool along with the collation model. The Beta Version will enable input in a TEI facsimile format, which would make it easier for someone working with TEI files.

3.3 Visualization Generation tool

The Visualization Generation tool is a web front-end built on top of an XSLT pipeline that uses XProc-Z, developed by Conal Tuohy (2016). The XSLT scripts are relatively

	A	B
1	1r	http://openn.library.upenn.edu/Data/0001/ljs101/data/web/0241_0006_web.jpg
2	1v	http://openn.library.upenn.edu/Data/0001/ljs101/data/web/0241_0007_web.jpg
3	2r	http://openn.library.upenn.edu/Data/0001/ljs101/data/web/0241_0008_web.jpg
4	2v	http://openn.library.upenn.edu/Data/0001/ljs101/data/web/0241_0009_web.jpg
5	3r	http://openn.library.upenn.edu/Data/0001/ljs101/data/web/0241_0010_web.jpg
6	3v	http://openn.library.upenn.edu/Data/0001/ljs101/data/web/0241_0011_web.jpg
7	4r	http://openn.library.upenn.edu/Data/0001/ljs101/data/web/0241_0012_web.jpg
8	4v	http://openn.library.upenn.edu/Data/0001/ljs101/data/web/0241_0013_web.jpg
9	5r	http://openn.library.upenn.edu/Data/0001/ljs101/data/web/0241_0014_web.jpg
10	5v	http://openn.library.upenn.edu/Data/0001/ljs101/data/web/0241_0015_web.jpg
11	6r	http://openn.library.upenn.edu/Data/0001/ljs101/data/web/0241_0016_web.jpg
12	6v	http://openn.library.upenn.edu/Data/0001/ljs101/data/web/0241_0017_web.jpg

Figure 6: Example image list in Excel.

unchanged from the proof of concept version, except that the first few scripts (which parsed the collation formula into the proto-collation model) have been removed, as the processing now begins with the collation model exported from the Collation Modeler.[5] The final output script has also been changed, as it now outputs four different views instead of the single one-web-page-per quire view from the proof of concept. In addition to the one-page-per quire view, it is now possible to generate the following: a single page for the whole manuscript (quires can be viewed and hidden at will), a diagrams-only view without the bifolia view, and a collation formula. In order to use the Visualization Generation tool, the user must upload both a collation model and an image list. The system depends on the folio or page numbers in the image list and the collation model to match. In a few minutes, the system outputs a zip file containing all four visualizations.[6]

Even in its imperfect alpha version, VisColl is being used in the community of manuscript scholars. Most notably, Lisa Fagin Davis is using VisColl in her project to reconstruct the physical construction of the *Beauvais Missal*, a late thirteenth-century liturgical book that was dismembered in 1942, when individual leaves were sold to institutions and individuals throughout the USA. As of October 2016, Fagin Davis has successfully reconstructed four quires of this manuscript (Fagin Davis 2016). Furthermore, Dot Porter and Will Noel at the University of Pennsylvania have used VisColl in their class for the Rare Book School, "The Medieval Manuscript in the 21st Century", and their students have in some cases made new findings with assistance from the tool (McDowell 2015).

4 Beta version

We are currently working on the beta version of VisColl, with the collaboration of Alexandra Gillespie and her team at the Old Books, New Science (OBNS) Lab at the University of Toronto (Gilllespie and Mitchell 2016). In the beta version we will do three things. First, we will extend the model to include the definition of sets of terms (i.e., *taxonomies*) that users can use to describe both physical and textual aspects of manuscripts. Second, we will add a facility that enables users to link these terms to the physical components of the manuscript. Third, we are changing the physical model itself to be more flexible, and to enable more complex physical structures. The first two changes will be accomplished by creating two new sections in the model: a Taxonomies section, where vocabularies are defined and selected, and a Maps section,

[5] The alpha scripts are still available on GitHub (Porter 2015).

[6] Although the Visualization Generation tool does not allow for bulk processing, the scripts that run the tool are on GitHub and could be used to bulk process multiple collation models (Porter 2016b).

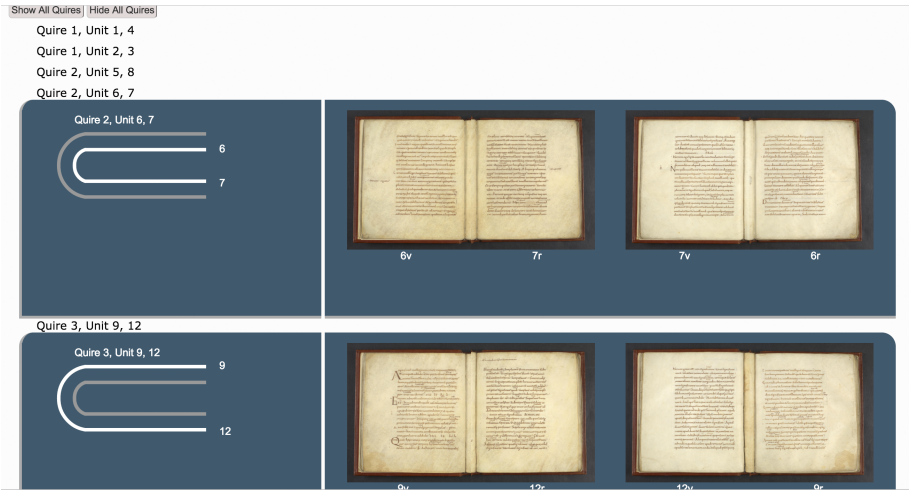

Figure 7: Screenshot of single-page view of University of Pennsylvania LJS 101, *Periermenias Aristotelis*. Note that all quires are on a single HTML page and the quires may be shown or hidden individually.

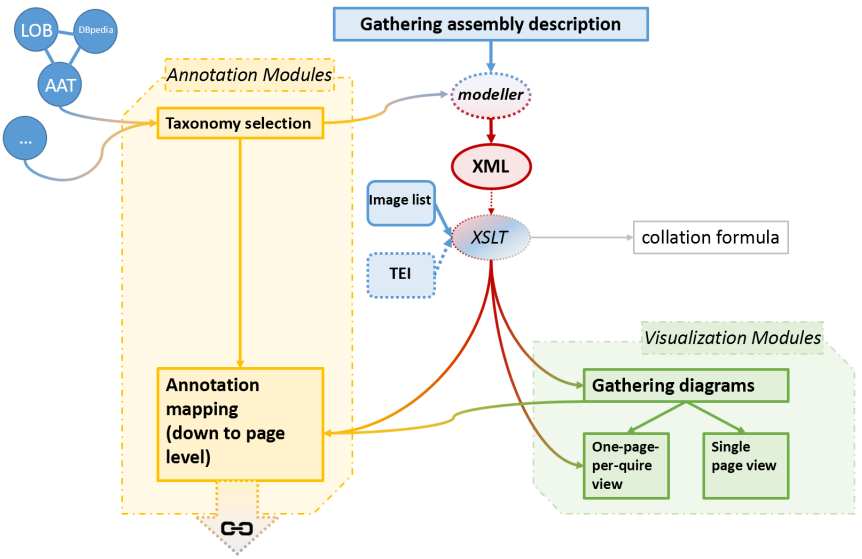

Figure 8: Diagram showing the pipeline of VisColl beta version. Note the integration of the Annotation modules and the possible links with external taxonomies and datasets.

where terms defined in the taxonomies section are linked to physical pieces of the manuscript.

4.1 Taxonomies section

In the Taxonomies section, users define lists of terms that describe important physical or textual aspects of the manuscript, and can then be associated with the physical components of the manuscript (sides of leaves, whole leaves, quires, and the entire manuscript). For example, if a manuscript is made of both parchment and paper, the user can define terms "paper" and "parchment", then in the model they can label each leaf with either term as appropriate. Taxonomies can include both defined by the project (e.g., the five stages of finish on the illustrations in the *Illustrated Hexateuch*, cf. Johnson 2000) and defined by external authorities (such as the Getty Art & Architecture Thesaurus, see The Getty Research Institute 2016; or the Language of Bindings Thesaurus, see Ligatus Research Centre 2016), opening the project to integration with Linked Data activities (Heath 2016). Any number of taxonomies can be defined in this section. Additionally, there are no taxonomies that are native to or required by the project. This is particularly important, as it allows for maximum flexibility on the side of the user, i.e., by selecting suitable taxonomy concepts, the user is able to describe anything in the model without restriction. In the example code below, the taxonomies section does not include values for semantic tags that are characteristic of the object, such as specific catchwords or signatures. However this is actively being addressed for inclusion in the final beta version of the model.

```
<viscoll>
 <taxonomy xml:id="b" xmlns="http://schoenberginstitute.org/schema/taxonomy">
  <!-- [...] -->
  <term xml:id="b5">Deuteronomy</term>
  <term xml:id="b6">Joshua</term>
 </taxonomy>
 <taxonomy xml:id="c">
  <label>Page contents</label>
  <term xml:id="c1">Illustration</term>
  <term xml:id="c2">Text</term>
 </taxonomy>
 <taxonomy xml:id="c"
           ref="http://www.getty.edu/research/tools/vocabularies/aat/">
  <label>Getty Art and Architecture Thesaurus</label>
  <term xml:id="c1" ref="http://vocab.getty.edu/aat/300011851">Parchment</term>
  <term xml:id="c2" ref="http://vocab.getty.edu/aat/300014179">Paper</term>
 </taxonomy>
 <taxonomy xml:id="d" ref="https://www.bl.uk/catalogues/illuminatedmanuscripts/
     glossary.asp">
  <label>Michelle P. Brown, Understanding Illuminated Manuscripts: A Guide to
       Technical Terms (J. Paul Getty Museum: Malibu and British Library: London,
        1994), online on the British Library website</label>
  <term xml:id="d1" ref="https://www.bl.uk/catalogues/illuminatedmanuscripts/
       GlossH.asp#hairside">Hair side</term>
```

```
  <term xml:id="d2" ref="https://www.bl.uk/catalogues/illuminatedmanuscripts/
       GlossF.asp#fleshside">Flesh side</term>
 </taxonomy>
 <taxonomy xml:id="e">
  <label>State of Finish (defined by Withers 2007)</label>
  <term xml:id="e1">Stage 1</term>
  <term xml:id="e2">Stage 2</term>
  <term xml:id="e3">Stage 3</term>
  <term xml:id="e4">Stage 4</term>
  <term xml:id="e5">Stage 5</term>
 </taxonomy>
 <!-- [...] -->
</viscoll>
```

Listing 3: Taxonomy section. Note that taxonomies are the responsibility of the user. They may be created by the user ("Page contents", "State of Finish") or may be drawn from formal schemas ("Getty Art & Architecture Thesaurus", "Understanding Illuminated Manuscripts").

4.2 Mapping section

The Mapping section links terms defined in the Taxonomy section to the physical components of the manuscript: sides of leaves, whole leaves, quires, or the entire manuscript. This creates reference links between semantic tags and physical components of the manuscripts. In the working version (see listing 4) the map identifies leaves by quire number and leaf in the quire (i.e., the third leaf of quire one is identified as `@leaf="1.3"`) and the side is indicated by `@side="r"` or `@side="v"`. Moving forward we will replace this physical identification with pointers to unique identifiers in the collation model, and thus the map will simply be a space for linking together physical components and terms, rather than defining the physical components in any way itself.

```
<mapping>
 <map leaf="1.2" side="r">
  <term target="#c2 #b1 #e5"/>
 </map>
 <map leaf="1.2" side="v">
  <term target="#c2 #e5 #b1"/>
 </map>
 <map leaf="1.3" side="r">
  <term target="#c1 #e5 #b1 #d1"/>
 </map>
 <map leaf="1.3" side="v">
  <term target="#c1 #c2 #e5 #b1 #d1"/>
 </map>
 <map leaf="1.4" side="r">
  <term target="#c1 #c2 #e5 #b1 #d2"/>
 </map>
 <map leaf="1.4" side="v">
  <term target="#c1 #c2 #e5 #b1 #d1"/>
 </map>
 <map leaf="1.5" side="r">
  <term target="#c1 #c2 #e5 #b1 #d1"/>
```

```
 </map>
 <map leaf="1.5" side="v">
  <term target="#c1 #c2 #e5 #b1 #d1"/>
 </map>
 <map leaf="1.6" side="r">
  <term target="#c1 #c2 #e5 #b1 #d2"/>
 </map>
 <map leaf="1.6" side="v">
  <term target="#c1 #c2 #e5 #b1 #d3"/>
 </map>
 <!-- [...] -->
</mapping>
```

Listing 4: Mapping section links the taxonomies (the values of `@target`) to quires, leaves, and pages. The next version of the collation map will assign unique identifiers to leaves and these ids will be used in the map.

The taxonomy and mapping modules allow for the expansion of VisColl beyond the presentation of information, and permit the end user to add knowledge in a way that is directly linked with the physicality of manuscripts. This kind of annotation on a page-by-page basis is not novel per se in manuscript studies (cf. Németh 2015, 309-12, table 6, and Corbach 2013, 27-33, table 1), but for the first time, with VisColl, annotations can be added electronically and then consistently linked with the appropriate parts of manuscripts. Additionally, allowing the use of externally defined Link Data taxonomies fosters collaboration and opens data for further research beyond specific manuscripts and repositories, breaking data free of information silos.

4.3 Collation model

Additionally, at the time of writing, the XML schema behind the Collation Modeler is being totally restructured. Moving away from quires as basic units, the new model considers leaves and stubs as atomic elements—which together form folios, bifolia, quires, and bookblocks—in order to accommodate for those complex structures, a sign of the complicated lives, often found in manuscripts; structures such as that depicted in figure 9, with quires within quires and pasted singletons, are rarely (if ever) encoded in collation formulas. Another element that is not encoded in formulaic quire assembly descriptions, but that is indispensable to understand non-standard and complex quire structures, is the leaf attachment method. Leaves can either be sewn or pasted/glued together to form quires and bookblocks. The new model provides means to indicate the attachment method of each leaf (sewn being the default option), and this will in turn allow the end user to describe and visualize unambiguously exceptionally complex structures.

Finally, in the near future, it is hoped that the collation model within VisColl and its visualization and annotation modules might be integrated with the International Image Interoperability Framework (IIIF - 2016), since such a partnership would be

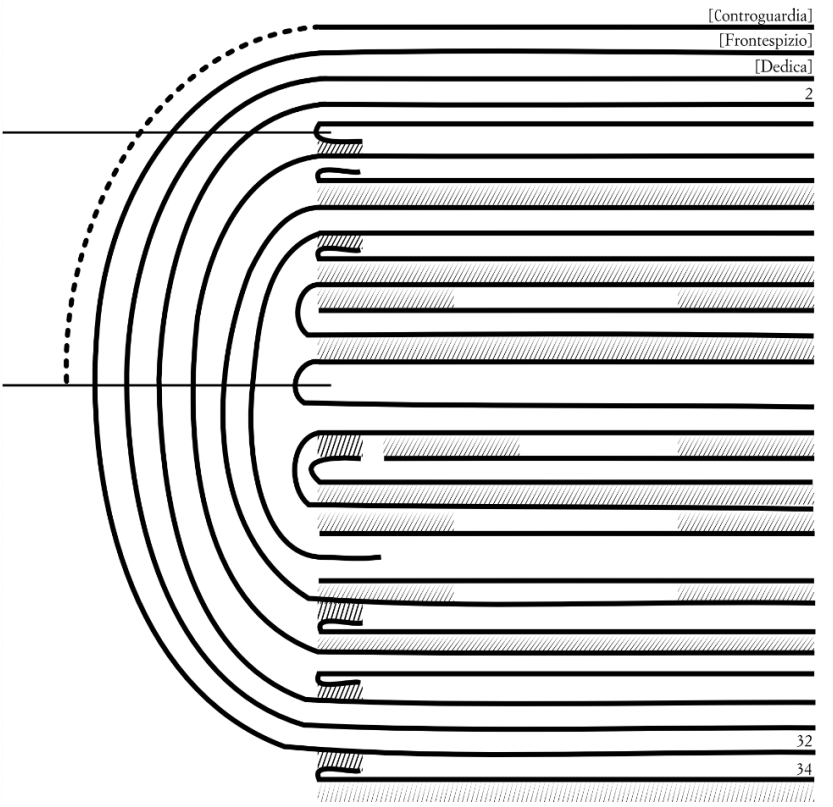

Figure 9: Example of complex manuscript quire structure (Vatican Library, Ferr.208, quire 1). Sewn leaves are indicated by a line representing the sewing thread; shaded areas indicate pasted leaves.

beneficial to both projects. Currently, the IIIF presentation API (Appleby et al. 2012) leverages the Shared Canvas Model (Sanderson and Albritton 2013) and the Web Annotation Data Model (Sanderson et al. 2017), and this accommodates annotation practices, which, by virtue of integrating seamlessly the principles of Linked Data and the Architecture of the Web (Jacobs and Walsh 2004), are perhaps more robust than VisColl annotations alone. The Shared Canvas Model, however, is incapable of representing the connections between different canvases (i.e. different pages of a manuscript) beyond being in a sequence. Integrating this representation model with the VisColl collation model would preserve both IIIF's robust annotation procedures and VisColl's effective representation of the actual structure of codices.

5 Conclusions

Since its initial conception in the mid-2000s through its implementation in 2013 up to current work on the beta version, VisColl, with its conceptual design and front-end usability, has been developed primarily for scholars who work with manuscripts. The project has brought together manuscript scholars, librarians and curators, conservators, and software developers, and serves as an example of the synergistic outcomes possible with interdisciplinary collaboration. Collaboration has increasingly brought flexibility into the project, widening its scopes to accommodate a diverse range of activities typical of specific disciplines that have the study of manuscripts at their core. This should not come as a surprise since the quire assembly is central to the production of codices, and its study is therefore fundamental for all disciplines within manuscript studies (and beyond). This collaborative effort will continue as we finalize the back-end design and modeling challenges and through the development of more ways to effectively visualize the new data brought into the beta version.

Bibliography

Andrist, Patrick, Paul Canart, and Marilena Maniaci. *La syntaxe du codex: essai de codicologie structurale.* Bibliologia 34. Turnhout: Brepols, 2013.

Appleby, Michael et al. (eds.) *IIIF Presentation API 2.1.* IIIF Consortium. 2012. <http://iiif.io/api/presentation/2.1>.

Armadillo Systems. London: Armadillo New Media Communications Ltd. 2009. <http://www.armadillosystems.com>.

Collation Modeler. 2016. <https://protected-island-3361.herokuapp.com>.

Corbach, Almuth. "Der Bernward-Psalter im Wandel der Zeiten. Eine Studie zu Ausstattung und Funktion." In Müller, Monika E. (ed.). *Der Bernward-Psalter im Wandel der Zeiten: Eine Studie zu Ausstattung und Funktion.* (=Wolfenbütteler Mittelalter-Studien 23). Wiesbaden: Harrassowitz, 2013. 263–382.

E-Codices - Virtual Manuscript Library of Switzerland. Fribourg: University of Fribourg, 2016. <http://www.e-codices.unifr.ch/en>.

Emery, Doug. *Collation Modeling.* GitHub 2016. <https://github.com/demery/collation-modeling>.

Fagin Davis, Lisa. "Quire Visualizations." In *Reconstructing the Beauvais Missal.* Cambridge (MA): The Medieval Academy, 2016. <https://brokenbooks2.omeka.net/exhibits/show/quire-visualizations>.

Ferrajoli 208. Rome: Vatican Library.

Gerardy, Theo. "Die Beschreibung der Wasserzeichen in Manuskripten und Drucken." In International Association of Paper Historians (eds.). *XIe Congrès International, Arnhem (Hollande) 4–9 Juin 1972.* Haarlem: Stitching Papiergeschiedenis, 1972. 1–9.

Gerardy, Theo. "Die Beschreibung des in Manuskripten und Drucken vorkommenden Papiers." In Gruys, Albert and Johan Peter Gumbert (eds.). *Les Matériaux Du Livre Manuscrit.* Codicologica 5. Leiden: Brill, 1980. 37–51.

Gilllespie, Alexandra and Laura Mitchell. *Old Books New Science (OBNS) Lab.* Toronto: Centre for Medieval Studies, 2016. <https://oldbooksnewscience.com>.

Gruijs, Albert. "Le Protocole de Restauration et La Description Des Cahiers et Bifolia." In Glénisson, Jean and Louis Hay (eds.). *Les Techniques de Laboratoire Dans L'ètude Des Manuscrits: [Actes Du Colloque International] Paris, 13–15 Septembre 1972.* Colloques Internationaux Du Centre National de La Recherche Scientifique 548. Paris: Centre national de la recherche scientifique, 1974. 253–255.

Heath, Tom. *Linked Data - Connect Distributed Data across the Web.* Linked Data community, 2016. <http://linkeddata.org>.

International Image Interoperability Framework. IIIF Consortium. 2016. <http://iiif.io>.

Jacobs, Ian and Norman Walsh (eds.). *Architecture of the World Wide Web, Volume One.* W3C, 2004. <https://www.w3.org/TR/webarch>.

Johnson, David. "A Program of Illumination in the Old English Illustrated Hexateuch: *Visual Typology.*" In Barnhouse Rebecca, and Benjamin C. Withers (eds.). *The Old English Hexateuch: aspects and approaches.* Kalamazoo (MI): Medieval Institute Publications, Western Michigan University, 2000. 165–200.

Kiernan, Kevin S. *Beowulf and the Beowulf Manuscript.* New Brunswick (NJ): Rutgers University Press, 1981.

Ligatus Research Centre. *Language of Bindings.* London: University of the Arts London, 2016. <http://www.ligatus.org.uk/lob>.

McDowell, Jesse. *An Ideal Collation of LJS 101.* November 16 2015. <http://schoenberginstitute.org/2015/11/16/an-ideal-collation-of-ljs-101>.

Muzerelle, Denis. *Vocabulaire codicologique: répertoire méthodique des termes français relatifs aux manuscrits.* (=Rubricae: Histoire du livre et des textes 1). Paris: Éditions CEMI, 1985.

Németh, András. "Layers of Restorations: Vat. Gr. 73 Transformed in the Tenth, Fourteenth, and Nineteenth Centuries". In *Miscellanea Bibliothecae Apostolicae Vaticanae* XXI, 2015. 281–330.

Noel, William. "Collation." In *The Digital Walters: Describing Manuscripts with TEI.* Baltimore (MD): Walters Art Museum, 2011. <http://thedigitalwalters.org/Supplemental/

ManuscriptDescription.html#collation>.
Porter, Dot. [2013.] "Visualizations of TEI Ms Descriptions." *tei-l@listserv.brown.edu*. 2013. <https://listserv.brown.edu/archives/cgi-bin/wa?A2=tei-l;775d4091.1307>.
— [2015.] "XSLT Alpha." In *Visualizing Physical Manuscript Collation*. GitHub. 2015. <https://github.com/leoba/VisColl/tree/master/xsl/xslts-alpha>.
— [2016a.] *Visualizing Physical Manuscript Collation*. GitHub. 2016. <https://github.com/leoba/VisColl>.
— [2016b.] "XSLTs." In *Visualizing Physical Manuscript Collation*. GitHub. 2016. <https://github.com/leoba/VisColl/tree/master/xsl/xslts>.
Sanderson, Robert, and Benjamin Albritton (eds.). *Shared Canvas Data Model 1.0*. IIIF Consortium. 2013 <http://iiif.io/model/shared-canvas/1.0>.
Sanderson, Robert, Paolo Ciccarese, and Benjamin Young (eds.). *Web Annotation Data Model*. W3C. 2017. <https://www.w3.org/TR/annotation-model>.
TEI. [2016a.] "10.7.1 Object Description." In *Text Encoding Initiative P5 (v. 3.0.0): Guidelines for Electronic Text Encoding and Interchange*, P5 revised and re-edited edition. Oxford, Providence (RI), Charlottesville (VA), Nancy (KY): Text Encoding Initiative Consortium, 2016. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/MS.html#msph1>.
— [2016b.] *Text Encoding Initiative P5 (v. 3.1.0): Guidelines for Electronic Text Encoding and Interchange*. P5 revised and re-Edited edition. Oxford, Providence (RI), Charlottesville (VA), Nancy (KY): Text Encoding Initiative Consortium, 2016. <http://www.tei-c.org/Guidelines/P5/>.
TEI Workgroup on Physical Bibliography. *Physical Bibliography - Draft for P5*. 2004. <http://www.tei-c.org/Activities/Workgroups/PB/PB-draft.xml>.
The British Library, National Library of Russia, St. Catherine's Monastery, and Leipzig University Library. *Codex Sinaiticus: See the Manuscript*. London: The British Library, 2016. <http://www.codexsinaiticus.org/en/manuscript.aspx>.
The Digital Walters. Baltimore (MD): Walters Art Museum, 2016. <http://thedigitalwalters.org>.
The Getty Research Institute. *Art & Architecture Thesaurus® Online*. Los Angeles (CA): The J. Paul Getty Trust, 2016. <http://www.getty.edu/research/tools/vocabularies/aat>.
Tuohy, Conal. *XPro-Z. A Platform for Running XProc Pipelines as Web Applications in a Java Servlet Container*. GitHub. 2016. <https://github.com/Conal-Tuohy/XProc-Z>.
Turning the Pages™. London: Armadillo Systems, 2016. <http://ttp.onlineculture.co.uk>.
"Vitae Sanctorum." In *Beinecke Digital Collections*. New Haven (CT): Yale University Library, 2016. <http://brbl-dl.library.yale.edu/vufind/Record/3592236>.
Withers, Benjamin C. *The Illustrated Old English Hexateuch, Cotton Claudius B.iv: The Frontier of Seeing and Reading in Anglo-Saxon England*. London, Toronto, Buffalo: The British Library ; University of Toronto Press, 2007.
Zappella, Giuseppina. *Manuale del libro antico*. Milano: Editrice Bibliografica, 1996.