DHQ: Digital Humanities Quarterly
2018
Volume 12 Number 2

Manuscript Study in Digital Spaces: The State of the Field and New Ways Forward

Bridget Almas <balmas_at_gmail_dot_com>, The Alpheios Project, Ltd.
Emad Khazraee <emad_at_kent_dot_edu>, School of Information, Kent State University
Matthew Thomas Miller <mtmiller_at_umd_dot_edu>, Roshan Institute for Persian Studies, University of Maryland College Park
Joshua Westgard <westgard_at_umd_dot_edu>, University Libraries, University of Maryland College Park

Abstract

In the last decade tremendous advances have been made in the tools and platforms available for the digital study of manuscripts. Much work, however, remains to be done in order to address the wide range of pedagogical, cataloging, preservation, scholarly (individual and collaborative), and citizen science (crowdsourcing) workflows and use cases in a user-friendly manner. This study (1) summarizes the feedback of dozens of technologists, manuscript experts, and curators obtained through survey data and workshop focus groups; (2) provides a "state of the field" report which assesses the current tools available and their limitations; and, (3) outlines principles to help guide future development. The authors in particular emphasize the importance of producing tool-independent data, fostering intellectual "trading zones" between technologists, scholars, librarians, and curators, utilizing a code base with an active community of users, and re-conceptualizing tool-creation as a collaborative form of humanistic intellectual labor.

I. Introduction

On 4 November 2016 the Roshan Institute for Persian Studies at the University of Maryland (UMD), College Park 1
hosted the "Manuscripts in the Digital Age Workshop," which was co-sponsored by Tufts University's Perseids Project, UMD's Arts and Humanities Center for Synergy, UMD's School of Languages, Literatures, and Cultures (SLLC), and Kent State University's School of Library and Information Science. The two-day workshop was born out of a concern that the existing digital infrastructure and tools for manuscript studies are failing to address the wide range of workflows, use cases, and research and pedagogical needs of scholars and curators in the field. Some of these issues arise from technological barriers. For example, many of the best solutions currently available require a level of technical knowledge that the vast majority of scholars and curators do not possess, and some solutions even require developer time to be properly set up (a financial hurdle that further stymies utilization of these tools by many at smaller institutions). Other problems, however, are the result of language-specific issues (e.g., poor display of non-Latin script and right-to-left languages) and persistent gaps in and atomization of functions and workflows.

The organizers had four primary objectives for the workshop: 2

1. to assess the current state of tools, services and infrastructure for the creation and preservation of digital editions and annotation of manuscripts, images, and related data objects;
2. to determine the extent to which those tools, services and infrastructure that come closest to meeting the needs of our respective projects can be reused and linked together;
3. to identify any apparent obstacles to their reuse and gaps in the functionality that they provide;
4. to create a "trading zone"[1] to foster dialogue between researchers, technologists, and librarians in the university, gallery, library, archive, and museum (GLAM) contexts regarding the functionality they would

ideally like to see in an integrated image workspace.

Building on the insights of the distinguished group of scholars, curators, librarians, and technologists who participated in this two-day workshop[2], this work first will provide a general overview of the state of the field and then draft principles to help guide future development efforts.

II. Problems and Desiderata

General Overview

While there has been tremendous advancement in the development of digital tools and platforms for the display and study of image data over the last decade, there is still no end-to-end solution that meets the myriad needs of scholars, curators, librarians, and students.[3] This is not a criticism of the pioneering efforts of the developers of the existing tools and platforms. It is the result of the fact that there are an incredible number of diverse use cases, each of which demands different functionality and workflows. Users are also working with a wide variety of different sized collections and linguistic traditions that are not all well supported by current tools and platforms. 3

This has led to a proliferation of tools and platforms that are specific to a particular collection or perform a discrete task (e.g., display, cataloging, transcription) — or, in some of the more well-developed tools, a *set* of discrete tasks — for the user. The tools and platforms listed in Table 1 are just some of the more popular tools and platforms that manuscript scholars report utilizing in their research. 4

Tool/Platform	Type
oXygen	Desktop software application
Microsoft Access	Desktop software application
PDFExpert	Desktop software application
XMLMind	Desktop software application
Kitodo	Software platform
IIIF (International Image Interoperability Framework)	API Specifications
Mirador	Web application
Omeka	Web application (hosted or local installation)
e-ktobe	Online dataset/collection
OPenn	Online dataset/collection
The Vatican Library Platform	Online dataset/collection
Pinakes	Online dataset/collection
Perdita Project	Online dataset/collection
Brotherton Manuscripts	Online dataset/collection
DEx: A Database of Dramatic Extracts	Online dataset/collection
Shelley-Godwin Archive	Online dataset/collection
Blake Archive	Online dataset/collection
CELM: Catalogue of English Literary Manuscripts	Online dataset/collection
Camena	Online dataset/collection
British Literary Manuscripts Online	Online dataset/collection and tools
Gallica (Bibliothèque Nationale de France)	Online dataset/collection and tools
The New Testament Virtual Manuscript Room	Online dataset/collection and tools
Coptic Old Testament Virtual Manuscript Room	Online dataset/collection and tools
vHMML	Online dataset/collection and tools
Coptic Scriptorium	Online dataset/collection and tools
TAPAS	Online platform, repository and tools
Transkribus	Online platform plus installable tools
Papyri.info	Online platform, dataset/collection and tools
Perseids	Online platform and tools

Table 1. Tools/Platforms used by scholars participated in the survery

All of these platforms and tools have certain virtues and do certain tasks well. Ultimately, however, their content, functionality, and workflows are too often siloed/atomized so that the user must either learn numerous different tools and platforms or have the technological skills necessary to link them together or export and manipulate their data post-analysis. This points to one of the biggest problems for users: none of the current tools or platforms reproduces the scholarly, curatorial, or pedagogical workflow in a single, integrated, digital space. As such, they are not always time saving in the same way that other more widely adopted digital technologies are. 5

Requirements of Scholars, Curators, Librarians, and Technologists

Prior to the workshop we distributed a survey to colleagues in the Digital Humanities, Classics, Middle Eastern/Islamic Studies, and library science communities via the popular listservs associated with each.[4] Twenty-seven respondents completed the entire survey. Admittedly, this was not a scientific or comprehensive survey, but its results are instructive and were echoed by many of the participants in the workshop itself. Before delving into the specific requirements of 6 scholars, curators, and technologists, there are a couple general insights that emerged from this survey.

First, the sizes of the manuscript collections with which users are working vary widely. As can be seen in Figure 1, the respondents to our survey are relatively evenly divided into the categories of small (< 100 mss), medium (100-999 mss), and large-scale (> 1,000 mss) users. The diversity in collection size is not surprising given the range of users — from scholars interested in studying a dozen manuscripts of one work to librarians and curators at large institutions with thousands of digital manuscripts — however, it should serve as a cautionary note for those looking to develop a universal, end-to-end solution for the diverse forms of manuscript study, display, and curation, which each have their own use cases and requisite workflows. 7

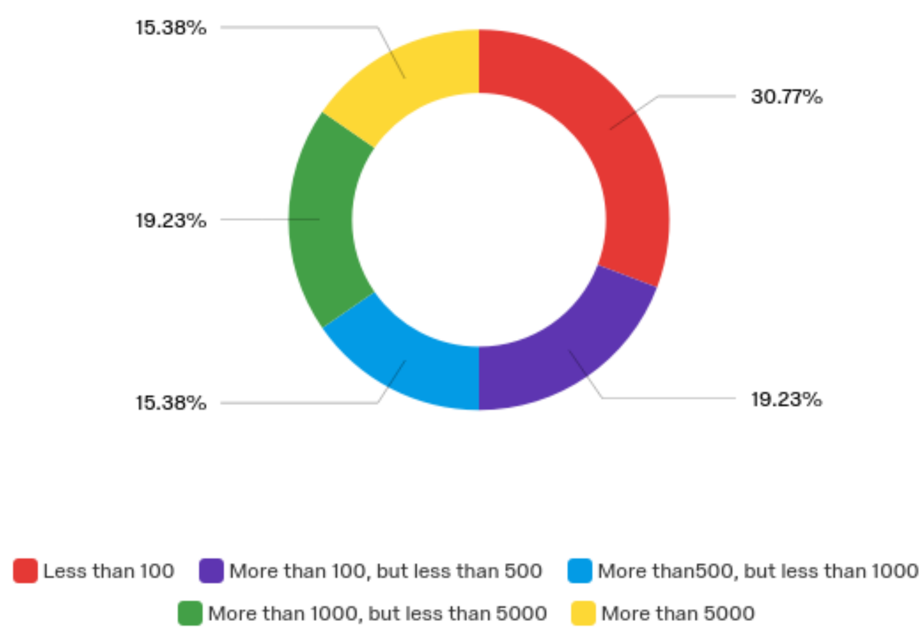

Figure 1. Response to survey question #9 regarding the approximate size of the manuscript collections with which the respondent works.

Secondly, users do not feel they personally have the technical expertise or access to technical staff required for installing and running tools or platforms. Almost 83 percent of survey respondents indicated that they prefer an online, fully hosted site that does not require them to seek assistance from technical staff at their universities in order to install and run. This is an important consideration for future development efforts, especially those aiming to cultivate a large and active community of users. 8

Beyond these general observations, there was a wide array of other specific insights that emerged from both the survey responses and the workshop. Below is a synthesis of the answers provided by both survey participants and workshop attendees (we have elected not to publish the survey responses verbatim due to privacy concerns). Our division of these responses into requirements of scholars, curators/librarians, and technologists is purely heuristic. 9

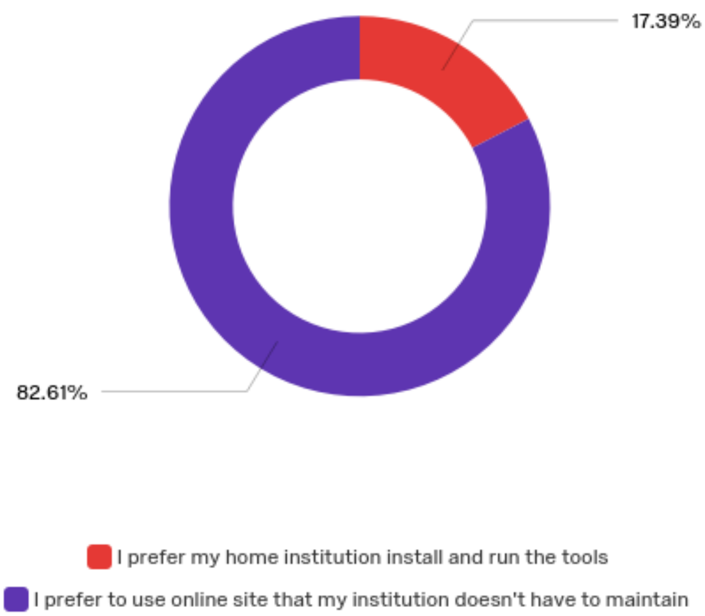

Figure 2. Response to survey question #17 regarding whether the respondent has access to technical staff or services at his or her home institution that could install and run the tools or prefer to use an online site that his or her institution does not have to maintain.

Requirements of scholars:

- A tool/workspace that aggregates previous scholarly comments on manuscripts of particular texts and authors;
- A collaborative space to draw on expertise of catalogers and content specialists of different areas (e.g., manuscript experts, textual scholars, art historians). As we heard repeatedly at the workshop, "no one can do it alone";
- Collaborative scholarly workflows (e.g., functionality and workflows that enable in digital space the scholarly workshop model of multiple scholars working together on one manuscript) and similar workflows for citizen science, knowledge-sourcing, or crowd-sourcing initiatives;
- Pedagogical workflows for use with students and "citizen scientists";
- Ability to access manuscript images from multiple digital repositories and use digital tools on them (i.e., a client application model);
- Integration of text and image data — as one can see in Figure 3, over 65 percent of respondents indicated they need to be able to work with both manuscript text and image data in their work;

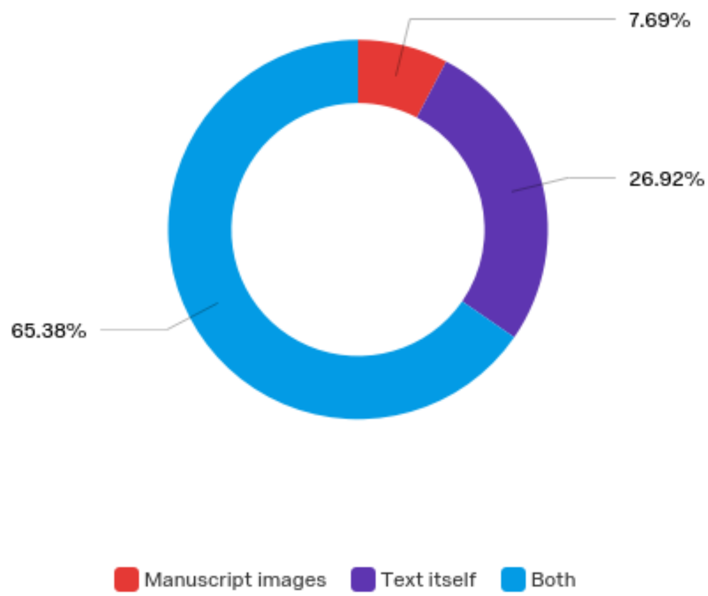

Figure 3. Response to survey question #10 regarding whether the manuscript text, image, or both is the primary focus of the respondent in their work.

- Ability to annotate text and image data. The ability to annotate both types of data was reported as a requirement by 72 percent of respondents (see Figure 4); 11

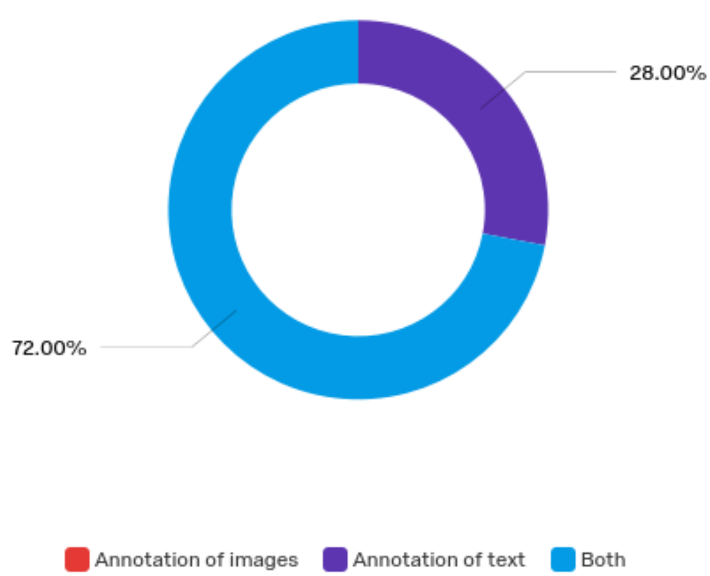

Figure 4. Response to survey question #11 regarding what kind of annotation support the respondent needs in their work.

- Educational materials/tutorials on how to setup and use new digital tools and workspaces; 12
- Institutional mechanisms and support for creating joint projects with digital librarians and technologists;

- A common gateway (because there are too many separate digital manuscript collections with unconnected repositories that cannot be universally searched), linking of similar texts in multiple repositories, and information about the differences between those copies;
- Tools and workflows to create and publish digital scholarly editions of works (e.g., digital critical editions, multi-text editions), including peer-review mechanisms;
- Better non-Latin script/right-to-left language support;
- Support for multimedia archives;
- Geographic tagging (for images of architecture, for example) to enable visualizations of diachronic transformations of place, buildings, etc.;
- Linking of sound recordings to geographic spaces;
- Ability to work, side by side, with multiple versions of the same manuscript — a requirement reported by over 85 percent of survey respondents (Figure 5) — and ability to annotate text variants and compare and classify large collections of manuscripts into manuscript families.

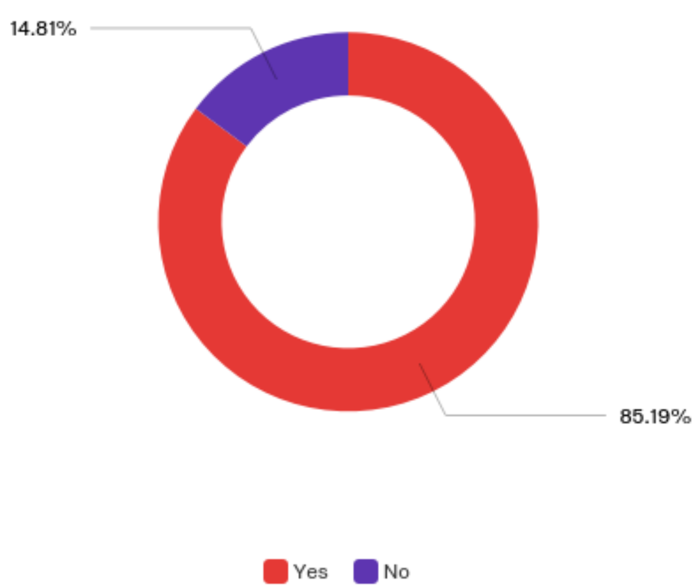

Figure 5. Response to survey question #12 regarding whether the respondent works with multiple manuscripts of the same text in their work.

Requirements of curators and librarians:

- Cataloging workflows (both for lone cataloger and collaborative cataloging projects);
- User-friendly sharing;
- Inclusion of conservation reports;
- More attention to materiality of the object and structure of books (e.g., bindings), including metadata on the whole object;
- Mechanisms that allow for restricted user access (although many collections are moving to open-access models, not all collections are willing to make images completely open access: in the survey results a sizable minority (37 percent) still reported that they need to restrict access to their image or textual data due to licensing/copyright reasons — see Figure 6);

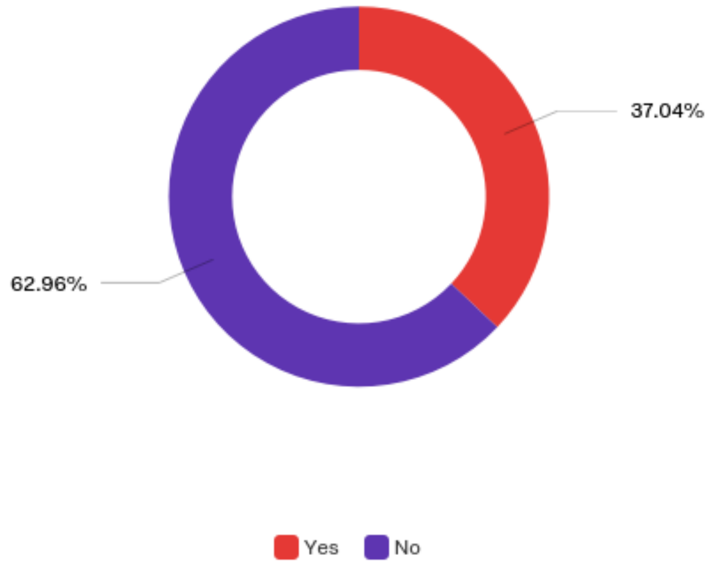

Figure 6. Response to survey question #14 regarding whether respondent needs to restrict access to the images or textual content of their manuscripts for licensing/copyright reasons.

- Tools that can be used by local teams (especially in poor or conflict areas) that are computationally light and standards-compliant (thus helping to facilitate collaboration with wider scholarly communities);
- Standards for dealing with multiple manuscripts in a single codex;
- Digital reunification of dispersed codices;
- Digital repatriation of objects removed from their country of origin;
- Expanded metadata for features such as type of paper, color of ink, type of leather;
- Inclusion of preservation metadata;
- Better support for dealing with alternative (i.e., non-traditional manuscript) image data, such as wedding certificates, audio visual materials, textiles, and 3D objects.

Requirements of technologists:

- Interoperability;
- Standards compliance;
- Persistent identifiers and permanent citation of manuscript data — a critical need, which over 92 percent of survey respondents indicated was required for their work (see Figure 7);

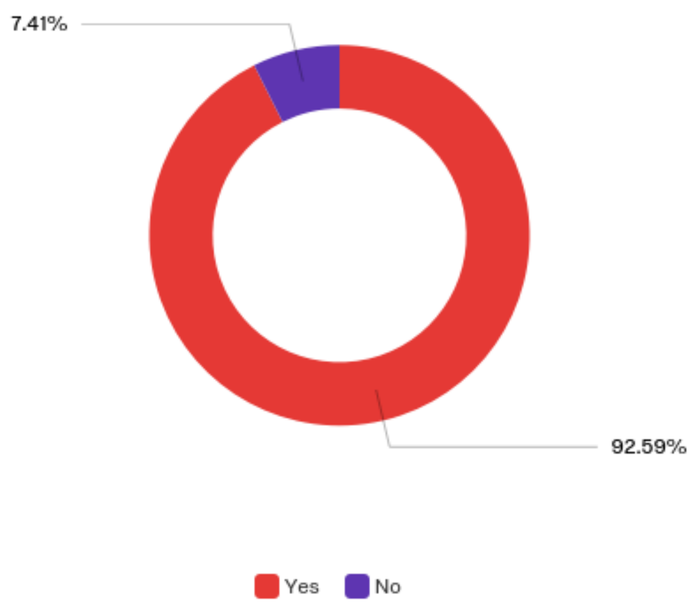

Figure 7. Response to survey question #13 regarding whether enabling citation of manuscript data (text or image) is important for the respondent's work.

- Named entity recognition;
- Ability to display documents together with known metadata;
- Ability to add and correct existing metadata;
- Stable edition numbers for texts published online (i.e., versioning);
- Version control (i.e., detailed tracking of textual changes through time);
- Citation via tag or git commit hash;
- Support for reading in multiple directions;
- Usability of tools/platforms in low bandwidth locations;
- Rejection of digital repository silo models;
- Controlled vocabularies that are built into cataloging/metadata;
- Ability to export data in multiple standard formats, especially TEI XML;
- Online hosted solution (preferable) or downloadable out-of-the-box program, since most users report that they do not have the requisite technical knowledge to setup tools/platforms and they do not have in house support from technologists to do it for them (see Figure 2).

Tool-Specific Issues

Much hard work has been devoted over the past decade to building digital infrastructure and tools for manuscript display and study. The tools and platforms we reviewed represent some of the best of the field currently. There are, however, user requirements that are still not being met. 13

We recognized from the outset that each of the tools selected for review in the workshop was designed with its own use case in mind, and coding for general purpose and reuse very likely was not explicitly in scope for many, or even any, of them. The goals and guidelines we used for the evaluations (presented in full in Appendix II) were developed with the aim of helping us surface some key characteristics that would help us better understand the potential of each tool as a starting point for further work. In particular, we tried to answer these questions for each tool: 14

1. What use cases and workflows is it *intended to* support?
2. What additional use cases and workflows is it *capable of* supporting?

3. Could gaps in functionality be filled by combining it with another tool?
4. Could gaps in functionality be filled by extending it with new development?
5. What skills are necessary to be able to use it effectively?
6. What skills are necessary to be able to extend it with new functionality?
7. How well does it support data management best practices?

Our review in the workshop — reproduced in summary form below — was not comprehensive; it focused on the principal uses and characteristics of each tool, to the extent we were able to discern them in the very limited time we had available to us. In preparation for the workshop we had each tool pre-installed in a virtual environment that workshop participants then used collectively to evaluate the tools. It is important to note that preparing a fully optimized configuration of each tool was not possible, so while our results are certainly of value in identifying some important features and gaps and in prioritizing what should happen next, they should only be considered as a jumping off point and not taken as fully informed analyses or judgments of any of these tools and platforms. While some tools we analyzed have a well-documented code base, in this study we did not do a thorough inspection of the code and documentation of each platform. Future studies will need to investigate these platforms in this regard, considering specific workflows and use cases to evaluate the extensibility of each platform and to discuss development strategies and necessary next steps. 15

Scriptorium was designed specifically for the task of annotation and transcription of manuscript images. Preservation and curation are not among its intended use cases. It supports import of images via externally referenced IIIF manifests [IIIF 2017]. It uses custom metadata fields to describe manuscripts, and its metadata is limited to the text — i.e., physical manuscript details, such as bindings, are not included. Curation of metadata is not supported. The user interface was designed specifically for scholarly and pedagogical workflows. Collaborative features are in place, although not fully fleshed out. The code relies on the Meteor javascript framework [Meteor 2017] and was not developed with the intent of being reused for other purposes. Both of these factors have implications for the long-term sustainability of the codebase. Of all the tools we surveyed, Scriptorium’s user interface had the most appeal to the scholars in the group. The user interface design is a candidate for reuse, even if the underlying code itself might not be. 16

Omeka was the most mature and well supported of the platforms we looked at, with a large community of open source developers contributing plugins and enhancements to the code base. It was also the most general purpose. It is primarily intended for display and publication of modest collections of digital materials. Support for real-time collaboration or pedagogical workflows is not inherent in the platform, although there may be plugins to provide some of this functionality. Omeka does offer excellent support for metadata standards, providing built-in templates using Dublin Core but allowing for easy extension with other controlled vocabularies. It also has good data management support, offering the ability to supply your own persistent identifiers (of any format, including Handles, URNs, DOIs, etc.) for collection items, not restricting you to database specific node identifiers. An existing IIIF plugin is not well integrated, but a current effort to develop new support for IIIF, led by University of Toronto [University of Toronto 2017] is focused specifically on working with manuscripts. The Toronto work supports import of images via IIIF Manifest URLs and treats IIIF annotations on those images as items in an Omeka collection. There is not currently native support for working extensively with manuscript text, outside of annotations on images, and this would need to be added for many use cases. The use of PHP and Javascript as the primary programming languages, and the level of documentation and support available for developers makes it a strong candidate for extension and reuse. It should be noted, though, that Omeka is not intended for long term preservation of data. Omeka is also available as a hosted online platform through a subscription model at omeka.net. Opportunities for project-specific customizations are limited in the hosted version, though, and depend upon the plan chosen. 17

Getty Scholar’s Workspace is a set of extensions to the Drupal Content Management System [Drupal 2017] which aims to provide a research environment for art historians. It adds tools to Drupal for creating bibliographies, grouping and comparing images, and for creating transcriptions, essays and annotations. It has some support for collaboration, based upon Drupal’s role-based permissions, but no pedagogical workflows. Standards supported include Dublin Core for taxonomies, but IIIF was not supported. Digital publication is explicitly excluded from the scope of the application, although it does support export of data in JSON, XML or CSV, which could then be used to create a static website from 18 the data via the Getty Publication’s static web resource generator. By leveraging the underlying flexibility and configurability of Drupal, together with the Getty tools, it is possible to create a fairly sophisticated digital publication adhering to good data citation practices (such as assigning external Persistent Identifiers to the data items and providing citation guidance to users), but the toolset itself does not provide any guidance on how to do this. In our exploration, for example, we could not easily determine exactly how the collection of tools was configured and used to produce the demonstration site, “Pietro Mellini’s Inventory in Verse, 1681.” [Getty Research Institute 2016] PHP programming skills would be needed to customize the toolset for specific use cases. There is good user documentation provided on the installation and use of the workspace’s features, and a Docker image to get up and running easily, so it does have some potential for reuse. We could envision, for example, building themes which provide preset defaults for data management, organization and eventual publication. There are already other existing plugins that provide IIIF support to Drupal, so these could potentially be added to a Getty workspace installation to round out the image support. The primary advantage over Omeka might be that Drupal, serving such a wide multidisciplinary user base, has a large number of plugins available, ranging from the very general to the very specific, but this could also be a drawback for less technical users because taking advantage of them can require significant programming expertise. Resources such as *Drupal for Humanists* [Drupal 2017] could prove useful here.

vHMML is a virtual reading room and cataloging environment designed specifically for the Hill Museum and Manuscript Library, with modules for display of reference material, images and other digital resources. It supports a cataloging workflow and metadata standard that is specific to its core use case. It supports IIIF for image viewing and annotation, via the Mirador viewer and the OpenSeaDragon library. The core application itself is written in Java and deployed on Tomcat with a MySQL database, and uses Elastic Search for indexed searching. It uses w3id.org to provide stable links to the resources published on it. It does not include support for collaborative scholarly or pedagogical workflows. The code, while available in an open source GitHub repository, is not currently intended to be collaboratively maintained or developed. It is, however, available for forking and repurposing, and the technologies used are fairly standard. One important thing to evaluate when looking at the code with an eye towards reuse would be the extent to which the vHMML-specific workflows, data formats, and use cases are tightly coupled to the design of the code. If they are baked into the design, it would likely not be worth the effort to try to reuse, but if done generally it might be possible. 19

Collective Access is a web-based suite of applications that provides a framework for management, description, and discovery of complex physical and digital collections. It is designed as a collection management system with a front end that can be used for public exhibitions. Archives and museums are the intended users of the system, especially ones who need to support both physical and digital collections. Creating and publishing collection catalogs and exhibitions are its core use cases. It is clear that archivists and librarians are the primary intended consumers of the tool and it is not designed for the casual user who does not have an understanding of metadata and cataloguing activities. It does support a wide variety of metadata standards preconfigured in the installation, such as Dublin Core, EAD (Encoding Archival Description), VRA Core (Visual Resources Association metadata standard), and CDWA-Lite (Categories for the Description of Works of Art). The metadata scheme is also easily extendable through the user interface (UI). The system provides control over the exposure of different metadata elements to the end user. A series of controlled vocabularies such as LCSH (Library of Congress Subject Headings) and Getty vocabularies are already integrated into the system. It provides a full tracking system for collections such as insurance valuation, location tracking, and provenance. Extensive logs for changes to each object/collection are also accessible through UI. Collective Access comes with a series of built-in workflows for batch import, export, and editing. There is no support for real-time collaboration or pedagogical workflows in the platform; however, there may be plugins to provide some of this functionality. The platform also provides granular levels of search such as objects, collections, events, and exhibitions. Collective access can use different viewers for access to media. Its default media viewer can allow access to still images, audio, video, and multi-page documents in PDF format. It has a built-in annotation tool for single images and audio/video but not for multipage documents. At the time of our workshop, the development team was in the process of modularizing the viewer and adding support for IIIF image server and multi-page document viewers such as Mirador and Universal Viewer [Digirati 2017]. The code is written in PHP and there is both user and developer documentation available on the project wiki, as well as consulting services offered by the Collective Access development team. The use of PHP as the primary programming language, and the level of documentation and support available for developers 20 make it a strong candidate for extension and reuse. Collective Access is also available as a hosted online platform through a subscription model at collectiveaccess.org. Along with Omeka, these are the only options of the tools we looked at which are publicly offered via this model. The heavy focus on the physical collection and cataloguing use case, and the complexity of the interface, led us to the initial conclusion that it was not a likely candidate for reuse for the manuscripts-centered use cases. It can, however, be used by archivists interested in managing their physical and digital collections through a single platform.

Mirador is a client-side javascript image viewer designed around and for the IIIF standard. It can run as an embedded 21
library in an HTML page, and is designed to be integrated into larger applications. It was used in a number of the platforms discussed above. There is technical documentation available, but using and deploying it does require minimal competence with Javascript and HTML. It is an obvious candidate for reuse. When combined with an IIIF image server and an annotation server it could provide a complete environment for image display and annotation, but it does not aim to cover any other use cases (e.g. such as transcription or cataloging). Tables 2-3 provide a summary of the comparison of the platforms.

Platform/ Tool	Main functionality	Curation and preservation workflows	Support external IIIF manifests	Metadata extensibility	Collaborative workspace features	Code base and framework for development
Scriptorium	Annotation and transcription of manuscripts	No	Yes	No; Idiosyncratic schema	Yes	JavaScript; Meteor Framework
Omeka	Display and publication of digital collections	Some functionalities	In development	Good	No	PHP and JavaScript
Getty Scholar's Workspace	Research environment for art historian	No	No (but a Drupal plugin exists)	Only Dublin Core	Yes	Developed as extensions to Drupal; PHP
vHMML	Virtual reading room and cataloging environment	No	Yes	No; Idiosyncratic schema	No	Java
Collective Access	Physical and digital collection management	Some functionalities	In development	Good	No	PHP
Mirador	Client-side image viewer	No	Yes	NA	NA	JavaScript

Table 2. Summary comparison of platforms (part 1)

Platform/ Tool	User Experience for working on manuscripts	Quality of documentation	Open source	Development community	Extensibility	Manuscript annotation support
Scriptorium	Good	Poor		Small team; limited to the project	NA	Yes
Omeka	NA	Good	Yes	Large active community	Yes	No
Getty Scholar's Workspace	Limited	Good	Yes	Small team; Not active	Yes	Yes (some features implemented)
vHMML	NA	Poor	Yes	Small team; limited to the project	Somewhat	Yes
Collective Access	Limited	Good	Yes	Active community	Yes	Only for single images
Mirador	NA	Good	Yes	Active community	Yes	Yes

Table 3. Summary comparison of platforms (part 2)

III. Observations and Recommendations for the Field

The technology and standards needed to support digital manuscript studies and publication are in place. While we may still be lacking some standardization around highly specific metadata, standards such as IIIF for images, TEI (Text Encoding Initiative) for text, Dublin Core for metadata, and the newly approved W3C Web Annotation Data Model [W3C 2017] cover a large majority of needs. While freely available tools and services for open source development, such as GitHub and various cloud infrastructures, have removed some infrastructure barriers and reduced the overhead required for development and deployment, the depth and diversity of scholarly and pedagogical workflows, the fast pace of technological change, and the shortage of software development expertise are challenges that still need to be overcome. A one-size-fits-all solution for digital manuscript studies remains difficult to envision, but modular infrastructures which allow for tools to be combined in different ways are possible and show a great deal of potential. 22

The following are our concluding insights and recommendations for future development efforts. 23

Insights

1. There is a diversity of workflows that need to be addressed. 24
    - There are different workflows needed for pedagogical, cataloging, preservation, individual scholarly, collaborative scholarly, and citizen science (crowdsourcing) work. Therefore, some tools may use a shared base platform (e.g., Omeka or Drupal) and add different workflows as modules or plugins for the base platform. This approach makes it easier to reuse and repurpose existing platforms.
2. The tool's code base and level of community involvement is critical for sustainability and expansion.
    - The code base significantly impacts future usability of tool. For example, high quality code base reduces the overall cost of recruiting new contributors.
    - An active community of users and developers is critical in sustaining a tool. It is worth studying the best practices on this point.
    - Following development best practices, such as producing comprehensive unit tests and

developer documentation, is essential to reuse.

3. There are no developer-free, end-to-end solutions for scholars and students of manuscripts.
    - There are no end-to-end solutions that are developer-free.
    - The field still lacks basic infrastructure, guidance and sustainable solutions for individual scholars and small institutions to manage and preserve their manuscript images and data.
4. Iteration is inescapable.
    - There is no one tool or platform that is going to be the panacea, so we need to think in terms of iterative development and collaborative development building upon existing tools and platforms.
5. Tool/platform creation is a collaborative form of intellectual labor.
    - We need to bring scholars, curators, and librarians together with technologists to create new tools and platforms and work to reconceptualize this collaborative work as a form of intellectual labor that is recognized by university administration as such in existing hiring, promotion, and tenure review processes.

Recommendations

1. Data curation and tool creation should be treated as separate projects. 25
    - We need to think about and plan for data separately from its manifestation in a particular tool. For example, high resolution images can be stored and preserved in an image repository while tools that aim to facilitate collaborative work on manuscripts can use IIIF manifestation of those images without the need to store and preserve original images.
    - We need to plan for enabling persistent identification and citations, in order to have clearly defined connection points and use APIs that can outlast the tools.
    - Data must be easy to export and transform.
2. We need to create more “trading zones” [Galison 1997] for technologists, librarians, curators, and scholars of manuscripts.
    - We need to establish communities and spaces (i.e., “trading zones”) composed of technologists, librarians, and scholars and develop “interlanguages” that facilitate collaboration. In the course of our workshop round tables with users, for example, it became clear that there are a number of types of functionality that users want but are unable to find in existing solutions and have not even been thought about before. Developers of such platforms will benefit from such trading zones through learning about the needs of their intended community of users, and thereby improve the overall design and functionality of the platform.
3. We need to better document the workflows and use cases our tools are meant to address.
    - Workflows and ideal use cases need to be identified and documented for each tool (in the workshop it became clear that even experts in the field were not always able to easily ascertain the workflows and use cases that a particular tool aimed to support). Well-documented workflows and use cases make extension and repurposing of the tools easier.
4. Avoid starting from scratch.
    - Since there is no platform which satisfies the wide-ranges of needs and workflows, it might be tempting to think about building a platform from scratch. We believe it is better to resist such temptations. There are many platforms (e.g., Omeka and Drupal) which can serve as the base infrastructure for development of tools. Designers and developers can use those platforms and add modules and plugins to extend their functionalities to serve particular workflows and use cases.

These insights and recommendations emerged from a collective effort of scholars, curators, and technologists to survey 26
the state of digital tools available for manuscript study and curation. Through such “trading zones” we can come to a

better understanding of the different emerging needs of these communities and increase their awareness about the status of digital infrastructure and the existing and emerging technologies. Such attempts can reduce redundancies and increase technological convergence. We hope these observations and recommendations encourage further collaboration between these communities and foster efforts aimed at co-creating shared extensible infrastructure.

Appendix I: Survey Information

Survey Questions

Q1 - If you currently study manuscripts, do you employ any digital tools or platforms?

Q2 - If you currently study manuscripts, but do not employ any digital tools or platforms, why do you not make use of any of the existing tools and platforms?

Q3 - If you currently do use platforms or tools when studying manuscripts, which ones do you use?

Q4 - Why do you use the platform you mentioned in the question above?

Q5 - Which tools or platforms would you point to as the best in the field currently?

Q6 - What features and functionality of these tools or platforms are most important to you?

Q7 - What features and functionality of these tools and platforms would you like to see improved and how?

Q8 - What features or functionality would you like to see in a manuscript workspace that you do not currently see in any of the existing tools platforms?

Q9 - What are the approximate sizes of the collections of manuscripts that you typically work with in your capacity as a scholar or curator?

- Less than 100
- More than 100, but less than 500
- More than 500, but less than 1000
- More than 1000, but less than 5000
- More than 5000

Q10 - Which of the following options are the primary focus of your digitization effort?

- Manuscript images
- Text itself
- Both

Q11 - Do you need to support annotation?

- Annotation of images
- Annotation of text
- Both

Q12 - Are you working with multiple manuscripts of the same text?

Q13 - Is enabling citation of your manuscript data (text or image) important for you?

Q14 - Do you need to restrict access to images or textual content for licensing/copyright reasons?

Q15 - Are there particular file formats (for example, TEI XML) that are required for storing the output of your research? Please provide the file formats.

Q17 - Do you have technical staff or services at your home institution that could install and run the tools for you, or do you prefer to use an online site that your institution does not have to maintain?

- I prefer my home institution install and run the tools
- I prefer to use an online site that my institution does not have to maintain

Survey Distributed to the Following Scholarly Listservs:

- *ArabicLitScholars*
- *Adabiyat*
- *IslamAAR*
- *Digital Classicist*
- *French DH*
- *The Digital Humanities Summer Institute*
- *Dublin Core Metadata Data Initiative - Cultural Heritage Task Group*
- *Association for Iranian Studies*

Appendix II: Goals for Tool Evaluations

Our basic goals for the workshop are to evaluate a selection of platforms for use in digital manuscript studies. The use cases are diverse and the platforms we've selected to review each have different objectives and core constituencies. 27

Primary Objectives

The primary objective for the technical portion of the workshop was to obtain answers to the following questions for each tool or platform: 28

1. What use cases and workflows is it *intended to* support?
2. What additional use cases and workflows is it *capable of* supporting?
3. Could gaps in functionality be filled by combining it with another tool?
4. Could gaps in functionality be filled by extending it with new development?
5. What skills are necessary to be able to use it effectively?
6. What skills are necessary to be able to extend it with new functionality?
7. How well does it support data management best practices? (Consider features such as persistent identification of resources, versioning, data import/export, data transformations, standard data formats, ontologies, etc.)

Below were some suggested questions for deeper investigation into each of these topics. These were intended to be a jumping off point only. Participants were encouraged to explore other questions and topics as appropriate for their use cases. 29

Use Cases and Workflows

1. Which of these high-level use cases does it support with regard to digital manuscript content:
    a. Creation
    b. Curation
    c. Publication
    d. Preservation
    e. Collaboration
    f. Pedagogy
    g. Analysis
    h. Other
2. Which of these content types does it support:
    a. Metadata
    b. Text
    c. Images
    d. Annotations
    e. Other

Data Management

- *Persistent Identifiers*
    1. Does it provide stable identifiers to your data objects?
    2. What type? (URLs, DOIs, Handles, ARKs, Database Identifiers, etc.)
    3. Are they globally unique or unique only to the instance of the tool/platform?
    4. What means does it provide to make these identifiers persistent and resolvable outside the context of the tool/platform?
    5. At what level of granularity are they available? (The object, a fragment of an object, annotations on an object, etc.)
    6. Can you supply your own PIDs for your data objects in addition to or instead of those assigned by the platform?
    7. Does it support versioning of identifiers?

Other Data Management Topics

1. Are there any restrictions on what is considered a data object?
2. Does it offer a means to provide formalized, machine-actionable descriptions of the data objects?
3. Does it support versioning of data?
4. Can you export your data?
5. Does it provide an API for access to its data?
6. What data type formats does it support?
7. Can you add/define your own data types and formats?
8. What support does it provide for publishing your data as linked data?
9. What support does it provide for ingesting or referencing external linked data sources?
10. Can you group your data into collections?
11. Can you have multiple collections of different data types?
12. Can you define relationships between items across collections?
13. What cataloging functionality does it provide?

Metadata

1. Does it support OAI/PMH?
    a. For export?
    b. For ingest/harvesting?
2. What metadata vocabularies does it support?
    a. Can you define or supply your own vocabulary?

Text

1. What text formats does it support?
    a. Plain Text, HTML, HOCR, Markdown, XML, PDF, etc.?
2. Is there an interface to upload textual content?
    a. From the file system?
    b. From urls?
    c. Does it support batch mode?
3. Does it provide support for linking text to other objects (images, external sites, annotations, etc.)?

Images

1. What image formats does it support?
    a. JPG, PNG, TIFF, etc.
2. Does it support a IIIF API endpoint?
3. Is there an interface to upload images?
    a. From the file system?
    b. From urls?
    c. Does it support batch mode?
    d. Does it support 3D visualization features?
4. Does it provide support for linking images to other objects (texts, external sites, annotations, etc.)?

Annotation

1. Can you create annotations?
    a. On text?
    b. On images?
    c. On pdfs?
    d. On other?
2. How are annotations stored?
    a. Are they assigned identifiers distinct from the items they are attached to?

Collaboration

1. Does it support collaboration on a shared object?
2. Does it support real time collaboration (multiple users working on the same object at the same time)?
3. Does it provide a user model?
4. What authentication options are offered?
    a. OAuth2 and Social Identity Providers?
    b. Shibboleth?
    c. Username/password?
    d. Other?
5. Does it support group features?
6. Does it support user roles?
    a. At what level of granularity?
        - Individual objects?
        - Application wide?
        - Project?

Extensibility

1. Does it support plugins for new functionality?
2. Does it provide documentation for how to extend?
3. Does it provide APIs for application integration?
4. What programming language(s) are required to extend it?
    a. What skill level is required to extend it?
5. Does it support custom themes/stylesheets for presentation?
6. Does it support mobile devices?

7. Is there an active developer community?
8. Is the code documented?

Usability

1. How easy is to use?
2. Is there user help/documentation?
3. Are there tutorials?

Notes

[1] On the concept and importance of "trading zones," please see: [Galison 1997]. Galison suggests a model for cross-boundary collaboration through the development of new shared cultures and languages. He argues that different subcultures can collaborate and exchange knowledge at the local level, without global agreement, by developing an interlanguage that groups share and use to communicate.

[2] In addition to the authors of the present work, the following individuals were also in attendance: John Abrahams (John Hopkins University), Alberto Campagnolo (Library of Congress), Hugh Cayless (Duke University/Digital Latin Library), Elijah Cooke (Roshan Institute for Persian Studies, University of Maryland), Hirad Dinavari (Library of Congress), Doug Emery (Schoenberg Institute for Manuscript Studies, University of Pennsylvania), Mahmood Gharavi (Columbia University), Ahmet T. Karamustafa (Roshan Institute for Persian Studies, University of Maryland), Fatemeh Keshavarz (Roshan Institute for Persian Studies, University of Maryland), Ida Meftahi (Roshan Institute for Persian Studies, University of Maryland), Patrick Murray-John (George Mason University/Omeka), Mark Patton (Johns Hopkins University), Simon Rettig (Freer Gallery of Art and Arthur M. Sackler Gallery), Raffaele Viglianti (Maryland Institute for Technology in the Humanities, University of Maryland), Joan Weeks (Library of Congress), and Jeffrey Witt (Loyola University Maryland). Daniel Gullo, Chad LaVigne, and William Straub from the Hill Museum and Manuscript Library also joined us for part of Friday afternoon to discuss the vHMML Reading Room.

[3] Henceforth, when we want to collectively refer to "scholars, curators, librarians, and students" we will use the term "users." We will only use the more specific terms "scholars," "curators," "librarians," or "students" when we want to differentiate one class of this broader category of "users" from the others.

[4] For full list of the survey questions, see Appendix I.

Works Cited

Digirati 2017 Digirati. "UV." Accessed March 29, 2017, http://universalviewer.io/.

Drupal 2017 Drupal. "Drupal for Humanists." Accessed March 29, 2017. http://drupal.forhumanists.org/.

Drupal Association 2017 Drupal Association. "Drupal." Accessed March 29, 2017. https://www.drupal.org/

Galison 1997 Galison, Peter. *Image and Logic: A Material Culture of Microphysics*. Chicago: University of Chicago Press, 1997.

Getty Research Institute 2016 Getty Research Institute, "Pietro Mellini's Inventory in Verse, 1681." Accessed March 29, 2017. http://www.getty.edu/research/mellini/.

IIIF 2017 IIIF, "International Image Interoperability Framework." Accessed March 29, 2017. http://iiif.io/.

Meteor 2017 Meteor. "The Fastest Way to Build Javascript Apps." Accessed March 29, 2017. https://www.meteor.com/.

University of Toronto 2017 University of Toronto. "Digital Tools for Manuscript Study." Accessed March 29, 2017. https://digitaltoolsmss.library.utoronto.ca/.

W3C 2017 W3C. "Web Annotation Data Model." Accessed March 29, 2017. https://www.w3.org/TR/annotation-model/