Digital Humanities Advancement Grant

White Paper Cover Sheet

Grant Number: HAA-255994

Project Title: Ensuring Access to Endangered and Inaccessible Manuscripts

Project Director: Columba Stewart

Assistant Project Director: Daniel K. Gullo

Grantee Institution: Saint John's University, Collegeville, MN

Original Project Description: NEH funding will make it possible to create vHMML 3.0, with greatly increased discoverability of manuscripts and metadata, and much richer data sharing with other digital humanities projects. vHMML 3.0 will add features requested by partner projects and researchers, and both outright and matching funds will sustain the human resources needed to guarantee best practice administration and continued development of vHMML.

1. Project Activities

Our project followed the three phases outlined in the original proposal.

Phase 1 (September 2017–Mid-January 2018) was dedicated to business analysis to outline the elements of the project in detail, including step-by-step descriptions of functionality, user scenarios, and mock-ups. This was done before engaging directly with the software developers (SDG: Solution Design Group). In the previous stages of vHMML we learned the importance of establishing as much detail as possible before the clock starts on developer hours.

Phase 2 (Mid-January–August 2018) was the intense development phase, which proceeded in 17 two-week sprints. The tasks accomplished within each sprint are listed in Appendix A.

During this period there were essential system upgrades, the development of new features to enrich the user experience, and improvements to the cataloging workflow. Full details are provided in the interim reports, but some merit highlighting in this report, particularly the upgrade of the Elasticsearch software to version 2.4.6 and of the Mirador IIIF image viewer to version 2.6. These are the core components of our system, driving access to manuscripts. We also made the loading of image thumbnails much faster by creating a persistent image directory for thumbnails using Digilib.

We added automatic approval of reader registrations so that users do not have to wait for approval by HMML staff. The accounts can subsequently be locked if there seems to be an issue. Note: access to all manuscript images is

free of charge, but some libraries that have allowed their manuscripts to be displayed in vHMML require that users register their personal details with HMML and agree to terms of use respecting publication and commercial rights. These details are not shared with the libraries but are used by HMML staff in case abuses are discovered.

The user interface throughout vHMML was redesigned. Reading Room receiving the most attention, improving the visualization and navigation of the information and options available on the landing page. We also added a "My List" feature to allow users to create ta custom collection and a News feature to highlight new uploads to Reading Room and important updates.

vHMML 3.0 provides hotlink searching from an individual record to all records containing the same author, title, associated names, and subjects. Users can now print or email a record, as well as export the object data as JSON. We developed new mechanisms for data management and export, allowing users to curate their own data sets for their research use in other DH projects. These are presented in the Data Portal module created for vHMML 3.0, allowing users to download their data sets as well as the complete vHMML metadata schema. An associated website, vHMML DH, provides tools for using the exported datasets and sample DH projects demonstrating applications of the data. Note: vHMML DH has recently been moved into the Initiatives/Digital Humanities section on the new www.hmml.org website.

Finally, all project resources were uploaded to Github.

Phase 3 (September 2018–August 2019) was devoted to the public launch of vHMML 3.0, post-production documentation, and continued uploading of records. The complete new version was announced on Twitter and other social media platforms on September 20, 2018, along with a series of instructional videos created and hosted on YouTube. Some features had been rolled out earlier, starting in April 2018; each was announced on social media and accompanied by a video. A list of announcements about vHMML 3.0 and/or tutorials on specific new features can be found in Appendix B along with a list of conference presentations highlighting vHMML 3.0. HMML's social media postings regularly include links to manuscripts in vHMML Reading Room.

Post-launch, the project team decided to remove vHMML School from the main vHMML application and create a new, more flexible, environment for School on a separate website.

Fundraising for the matching component of the grant was achieved through a combination of restricted gifts from individual and dedication of a portion of unrestricted general HMML fundraising to the match.

2. Accomplishments

Our goals were to increase and improve access to digitized manuscripts. We are proud of what was accomplished for both goals.

When the project began (September 2017), there were:
- 21,115 manuscript records were already online in vHMML Reading Room (including both microfilm and digital formats)
- 14,736 featured full digital image sets
- 900 registered users of vHMML Reading Room

When the project ended (June 2020), there were:
- 75,363 manuscript records were already online in vHMML Reading Room (including both microfilm and digital formats)
- 38,409 featured full digital image sets
- 2500 registered users of vHMML Reading Room

3. Audiences

As noted above, by June 2020 there were 2500 active users of vHMML Reading Room (i.e., registered and active within the past 12 months). The vast majority of registered users are scholars at various stages of their education/career. We do not have specific data about gender or age. Users were based in 86 countries (see Appendix C). Especially notable are users outside of North America or western Europe accessing their own manuscript traditions or those of religious/cultural interest to them.

Although most users are individual scholars, we have used vHMML Reading Room for courses at HMML such as the Dumbarton Oaks-HMML Summer Schools (Syriac, Armenian, Coptic language instruction) and the Minnesota Manuscript Research Laboratory (co-sponsored with the Center for Medieval Studies at the University of Minnesota). HMML staff have also used the site to teach courses at the University of Malta.

4. Evaluation

The rapid growth in registered users of vHMML Reading Room is the best indication of the project's success.

There has also been external recognition of the project by professional societies and numerous invitations to present vHMML at conferences and consultations.

vHMML team members Daniel Gullo, Eileen Smith, and David Calabro received the Minnesota Academic Innovators Award from the Minnesota Library Association in April 2018. The project was awarded the Digital Humanities and Multimedia Studies Prize from the Medieval Academy of America in 2020.

An external review of HMML's digital platforms in November 2019 by Donald J. Waters, formerly head of Scholarly Communications at the Andrew W. Mellon Foundation, highlighted the value of vHMML's rigorous authority control and controlled vocabularies for manuscript description.

The project team has identified three areas in which we did not, however, fully accomplish our goals:

- The current Java version in vHMML did not permit upgrading of much of the software to the latest version. For example, we could not fully upgrade ElasticSearch to the latest version. The Java upgrade will be a key part of vHMML 5.0 development.
- Our hopes to make progress on SEO (Search Engine Optimization) for the records in Reading Room were not realized as we had hoped, though there has been a recent uptick in crawled pages (see Appendix D).
- We were not able to refactor the database to move the controlled vocabulary of scripts and writing systems into discreet tables. This would have provided further support for the tools we have developed using controlled vocabularies to simplify editing and updating of metadata.

5. Continuation of the Project

The project will continue as the principal means of sharing HMML's digitization efforts with the global scholarly community. The centrality of vHMML to HMML's mission ensures that it will remain a permanent part of HMML's digital ecosystem and will continue to be developed in the coming years to reflect advances in technology and user expectations. Planning for vHMML 5.0 is already underway.[1]

vHMML has been chosen as the digital home for manuscript images created by the Zaydi Manuscript Tradition project at the Institute for Advanced Study in Princeton, NJ (a recipient of Humanities Collections and Reference

[1] Note that the current version is actually vHMML 4.0, following the addition of a new component, Museum, in April 2019.

Resources grant from the NEH; the DREAMSEA project on Southeast Asian manuscripts; and a digitization project in the Libyan World Heritage Site of Ghadames sponsored by the American Schools of Oriental Research. We expect such partnerships to continue as other projects recognize that they do not need to devote resources to creating their own platform for access to digital images.

Looking ahead, in October 2020, we moved away from the name "vHMML," referring instead to "HMML's Digital Tools." This recognizes that for the vast majority of HMML's audiences, the digital tools, and especially Reading Room, are HMML.

6. Long-term Impact

We can state with confidence that making these digitized manuscripts available to the world is transforming scholarship, especially for understudied Eastern Christian and Islamic traditions. The quality of the viewing experience, the depth and breadth of the digital holdings, and the rigorous cataloging standards have brought HMML to the attention of much wider audience of scholars.

Several funders—Arcadia, Whiting, Mellon—have pointed to vHMML as the "gold standard" for online manuscript access and required other digitization projects they support to use Reading Room as the means of access to their manuscripts. We have also been in fruitful conversation with funders about the need to build support for vHMML into digitization grants to ensure long-term access to the data.

One of the major impacts—unforeseen at the time vHMML 3.0 was developed—is its contribution to building the new HMML Authority File (HAF) database, a project supported by a recent major grant from the NEH. This project will take the authority creation being done during the cataloging work for vHMML and make it easily accessible to the broadest possible audience.

7. Grant Products

As noted above, all software is available on HMML's GitHub site. The real product, of course, is the data and metadata accessible in Reading Room.

Appendices

Appendix A: vHMML 3.0 Requirements Matrix

Appendix B: List of project announcements and conference presentations

Appendix C: vHMML Registered Users by Country

Appendix D: SEO report

Appendix E: Sample publication about the project (Stewart, "Giving Voice to Ancient Texts")

Appendix A: vHMML 3.0 Development Matrix: January-October 2018

Feature	Concept	Mockup/Logic/ User Scenarios	Walkthrough	Done	Schedule
January-April Task List					
January 23-Feb 6					
Update Redhat Enterprise OS	X	X	X	X	Sprint 42
Update MySQL software	X	X	X	X	Sprint 42
Update JAVA	X	X	X	X	Sprint 42
Update vHMML https VIAF links	X	X	X	X	Sprint 42
User print metadata from RR description JIRA 335	X	X	X	X	Sprint 42
Add IIIF Image checkbox RR JIRA 2354	X	X	X	X	Sprint 42
Add IIIF FOLIO JIRA 2523	X	X	X	X	Sprint 42
User email full metadata to user RR JIRA 1172	X	X	X	X	Sprint 42
Confirmation modals for reindex, import, export buttons JIRA 2546	X	X	X	X	Sprint 42
Update Jenkins software for John	X	X	X	X	Sprint 42
Improved JSON object and Link export JIRA 2372	X	X	X	X	Sprint 42
Authority tooling additions and improvements	X	X	X	X	Sprint 42
Administration additions and improvements	X	X	X	X	Sprint 42
Update Eclipse software for John	X	X	X	X	Sprint 42
Release candidate on February 20					
Feb 6 - Feb 20					
Update DigiLib software on server JIRA 2506	X	X	X	X	Sprint 43
Update Landing Page Images/Links JIRA 2454 and 2504	X	X	X	X	Sprint 43

Update Bootstrap software	X	X	X	X	Sprint 43
Add CSV export to Authorities	X	X	X	X	Sprint 43
User Registration Page Redesign JIRA 2549	X	X	X	X	Sprint 43
Registration bugs and improvements	X	X	X	X	Sprint 43
Database bugs and improvements	X	X	X	X	Sprint 43
All above reported to NEH on February 28, 2018					
Release candidate on March 6					
Feb 20- March 6					
Search user role bug JIRA 2508	X	X	X	X	Sprint 44
User increase session log in times JIRA 2481	X	X	X	X	Sprint 44
User save records to cache "My List" JIRA 2547	X	X	X	X	Sprint 44
User save records to "My List" in account database JIRA 2557	X	X	X	X	Sprint 44
Search add highlights to name and title JIRA 2432	X	X	X	X	Sprint 44
Database bugs and improvements	X	X	X	X	Sprint 44
Release candidate on March 20					
March 6 - March 20					
Update Elastic Search software to version 2.4.6 JIRA 2505	X	X	X	X	Sprint 45
Create single reindex option JIRA 2662	X	X	X	X	Sprint 45
Add locked/unlocked user to ADMIN JIRA 2594	X	X	X	X	Sprint 45
Database bugs and improvements	X	X	X	X	Sprint 45
March 20- April 3					
Registration make new registered users automatically approved JIRA 2652	X	X	X	X	Sprint 46
Update Elastic Search software to version 2.4.6 JIRA 2505	X	X	X	X	Sprint 46

Search collection, series, subseries advanced search JIRA 2691	X	X	X	X	Sprint 46
Search Hotlinks for author, uniform title JIRA 1170	X	X	X	X	Sprint 46
Improve Admin reports by removing ADMIN users JIRA 2545	X	X	X	X	Sprint 46
Improve Save, Cancel, and Delete options in Lexicon JIRA 2675	X	X	X	X	Sprint 46
Search bugs and improvements	X	X	X	X	Sprint 46
Bug fixes and improvements	X	X	X	X	Sprint 46
Release candidate on April 17					
April 3-April 17					
TAG STARTS VERSION 3.0					
Update Elastic Search software to version 2.4.6 JIRA 2505	X	X	X	X	Sprint 47
SEO create sitemap features for vHMML JIRA 2176	X	X	X	X	Sprint 47
SEO JSON-LD to Reading Room description page JIRA 2659	X	X	X	X	Sprint 47
SEO JSON-LD to Folio JIRA 2664	X	X	X	X	Sprint 47
SEO JSON-LD to Lexicon JIRA 2485	X	X	X	X	Sprint 47
Search implement custom stoplist JIRA 2695	X	X	X	X	Sprint 47
Search make name, title, an incipit field " " search JIRA 2693	X	X	X	X	Sprint 47
Admin create logging admin screen JIRA 2235	X	X	X	X	Sprint 47
Search bugs and improvements	X	X	X	X	Sprint 47
Release candidate on April 23					
April 17-May 1					
Mirador 2.6 and plugins JIRA 2367	X	X	X	X	Sprint 48
Reference 3.0 Redesign JIRA 2497	X	X	X	X	Sprint 48
Search hotlinks collection and series sub-series JIRA 2678	X	X	X	X	Sprint 48

Search bugs and improvements	X	X	X	X	Sprint 48
May 1-May 15					
Mirador 2.0 and plugins JIRA 2367 continued	X	X	X	X	Sprint 49
Internal iiif public URLs for vHMML Manifests JIRA 2743	X	X	X	X	Sprint 49
Mirador IIIF sharing bugs JIRA 2285 and 2282	X	X	X	X	Sprint 49
Mirador Autocenter JIRA 2435	X	X	X	X	Sprint 49
Server Persistent Image Directory Cache Thumbnails JIRA 2755	X	X	X	X	Sprint 49
Internal hotlink search associated names JIRA 2710	X	X	X	X	Sprint 49
Release candidate on May 29					
May 15-May 29					
Server Ehcaching JIRA 2596	X	X	X	X	Sprint 50
Sitemap Robots.txt JIRA 2775	X	X	X	X	Sprint 50
Improve Logfiles on vHMML JIRA 2465	X	X	X	X	Sprint 50
SEO JSON-LD to Reference JIRA 2665	X	X	X	X	Sprint 50
Release candidate on June 12					
May 29-June 26					
Database Improve Former Owner field JIRA 2368	X	X	X	X	Sprint 51
Database Improve Artist field JIRA 2520	X	X	X	X	Sprint 51
Database Archival Data import fixes JIRA 2437 and 2438	X	X	X	X	Sprint 51
vHMML Data Portal - curated data download page (API Dynamic Accessible Data) JIRA 2744	X	X	X	X	Sprint 51
Create GITHUB use scenarios for DATA page JIRA 2745	X	X	X	X	Sprint 51
Bug fixes to displaying data and auto deleting data	X	X	X	X	Sprint 51
June 26-July 10					

Batch data process JIRA 2843	X	X	X	X	Sprint 52
JSON download option to individual records JIRA 2660	X	X	X	X	Sprint 52
Database audit remove extraneous fields JIRA 2542	X	X	X	X	Sprint 52
Database Archival Database fixes	X	X	X	X	Sprint 52
Bug fixes to displaying data and auto deleting data	X	X	X	X	Sprint 52
Release candidate on July 10					
July 10-24					
UI add back to top JIRA 2456	X	X	X	X	Sprint 53
vHMML DH page JIRA 2871	X	X	X	X	Sprint 53
vHMML DH project Malta Parish Archives JIRA 2872	X	X	X	X	Sprint 53
vHMML DH project Giovanni Andrea JIRA 2873	X	X	X	X	Sprint 53
Search Create hotlinks for subject JIRA 2671	X	X	X	X	Sprint 53
UI Landing Page Redesign and content manager JIRA 2865	X	X	X	X	Sprint 53
UI News Page JIRA 2866	X	X	X	X	Sprint 53
Add microfilm to popularity reports JIRA 2544	X	X	X	X	Sprint 53
July 24-August 21					
UI Landing Page Redesign and content manager JIRA 2865	X	X	X	X	Sprint 53
UI new header	X	X	X	X	Sprint 53
UI new footer	X	X	X	X	Sprint 53
UI News Page JIRA 2866	X	X	X	X	Sprint 53
August 21-September 4					
UI vHMML Reading Room and content manager JIRA	X	X	X	X	Sprint 54
UI Reference 3.0 Redesign	X	X	X	X	Sprint 54

UI Lexicon 3.0 Redesign	X	X	X	X	Sprint 54
Release candidate September 4					
September 4-September 18					
UI Folio 3.0 Redesign	X	X	X	X	Sprint 55
UI Data Portal 3.0 Redesign	X	X	X	X	Sprint 55
Increase HEAP on image server	X	X	X	X	Sprint 55
Bug fixes	X	X	X	X	Sprint 55
All above reported to NEH on September, 2018					
September 18-October 2					
Server Persistent Image Directory Cache Thumbnails JIRA 2795	X	X	X	X	Sprint 56
Bug fixes and locking code	X	X	X	X	Sprint 56
SDG documenation	X	X	X	X	Sprint 56
Add vHMML to Github Jira 2965	X	X	X	X	Sprint 56
FINAL RELEASE CANDIDATE vHMML 3 October 2					

Appendix B: Marketing and Community Involvement

Social media with vHMML announcements and/or tutorials on specific features

Twitter (@visithmml):

- 9/12/2018 - https://twitter.com/visitHMML/status/1039940273462235142
- 9/20/2018 - https://twitter.com/visitHMML/status/1042798560184393728
- 9/20/2018 - https://twitter.com/visitHMML/status/1042792695213764613
- 9/21/2018 - https://twitter.com/visitHMML/status/1043133077055258624
- 11/28/2018 - https://twitter.com/visitHMML/status/1067823011959586816
- 12/14/2018 - https://twitter.com/visitHMML/status/1073631262508752896
- 2/21/2019 - https://twitter.com/visitHMML/status/1098682026918625281

Facebook (visitHMML):

- Posted 9/8/2018, 9/9/2018, 9/12/2018, 9/18/2018, 9/19/2018, 9/20/2018, 9/21/2018, 11/28/2018, 12/14/2018, 2/21/19.

YouTube (https://www.youtube.com/user/HMMLPreservation):

- 9/11/2018 - https://youtu.be/-VnpnuLp9ew
- 9/20/2018 - https://youtu.be/6xSVvm3fEeU
- 1/18/2018 - https://youtu.be/e7idZTwjTH8
- 11/28/2019 - https://youtu.be/EnGqfvI0MMI

Conference Presentations Featuring vHMML 3.0

Columba Stewart (Project Director):

- 5/9/18: "vHMML Reading Room: Metadata Standards, Authority Control, and Initial Steps Toward Linked Data in a Comparative and Comprehensive Cataloguing Environment," Center for the Study of Manuscript Cultures, University of Hamburg, Germany.
- 9/26/18: "Digital Preservation of Endangered Manuscript Heritage, Global Access through *v*HMML," New York Public Library, New York, New York, conference on "Engaging Cultural Heritage" co-sponsored by the NYPL and the Bibliothèque Nationale de France.
- 10/8/18: "A Modern Monk's Work: Preserving the Manuscript Heritage of Endangered Christianity in the Middle East," Annual Corish Lecture, Saint Patrick's College, Maynooth, Ireland.
- 10/23/18: "The Global Mission of HMML: Digital Preservation of Endangered Documentary Heritage/Online Access," Building Bridges conference of the Cultural Heritage Coordinating Committee, Smithsonian Institution, Washington, DC.

Daniel K. Gullo (Project Assistant Director):

- 4/27/2018: "vHMML Reading Room: Building a Bibliographic Database for Digital Archives, Manuscripts, and Books." Minnesota Library Association Academic and Research Library Division (ARLD) Day 2018, Chaska, MN.
- 10/27/2018: "The Confraternity of Charity, Slavery, and Archives on the Margin," Fall Mediterranean Seminar "On the Margins of the Mediterranean," University of Michigan, Ann Arbor. The talk featured the use of digital technology to unite dispersed archives.
- 11/10/2018: "Curated Metadata and Digital Humanities: Using JSON Exports to Reimagine Data Access for Online Catalogs in vHMML," LITA Forum of the American Library Association Minneapolis, MN, 2018 (with Bryan Lor from SDG, and John Meyerhofer, HMML Systems Librarian).
- 2/25-26/19: Meeting on the future of Digital Scriptorium, Yale University, New Haven, CT.

Melissa Moreton (HMML Assistant Director for Strategic Initiatives):

- 10/31/18: Minzu University, Beijing, vHMML workshop for philology graduate students.
- 11/6/18: Occidental College, Los Angeles, CA, lecture on digital preservation of manuscripts and demo of vHMML.
- 1/30/19: Columbia University, NYC, Medieval and Renaissance Studies Program, vHMML workshop for MedRen graduate students.
- 1/31/19: Zohrab Information Center, Eastern Diocese of the Armenian Apostolic Church, New York, NY, presentation of digital preservation of Armenian manuscripts with demo of vHMML.

Valeria Vanesio (Archivist for the Malta Study Center, HMML):

- 10/18/2018: "I gerosolimitani in Toscana e lungo la Via Francigena. Ospedali, commende e fortificazioni," FORTMED Congress, Turin, Italy (with Valentina Burgassi of the Sorbonne). Use of digitized manuscripts in vHMML to document the construction of commanderies in Tuscany.
- 1/17/2019: "The Archives of the Order of St John. The Case of the Priory of Pisa," Medici Archive Project Archival Seminar Series, Florence, Italy. Demonstrated the use of vHMML and standardized metadata for cataloging the proofs of nobility for the Order of Malta.
- 1/10/2019: "L'albergia della Lingua d'Italia a Malta. L'avventurosa storia di un palazzo e delle sue carte (secoli XVI-XIX)," guest lecture in Archival Studies Program (*Archivistica generale*), Sapienza University, Rome, Italy (with Valentina Burgassi of the Sorbonne). Use of digitized manuscripts in vHMML to document the building of the Auberge d'Italia in Valletta, Malta.

Reading Room Registrations by Country (through September 2020)

US United States	1333		PK Pakistan	9
GB Great Britain	252		DK Denmark	8
MT Malta	249		JO Jordan	8
DE Germany	227		PT Portugal	8
IT Italy	165		BG Bulgaria	7
FR France	164		FI Finland	7
TR Turkey	119		MX Mexico	7
EG Egypt	115		PS Palestine	7
AU Australia	101		ZA South Africa	7
ET Ethiopia	94		TN Tunisia	6
LB Lebanon	60		KR Korea	5
RU Russia	58		QA Qatar	5
SA Saudi Arabia	58		RS Serbia	5
BE Belgium	44		SG Singapore	5
ES Spain	44		BA Bosnia	4
NL Netherlands	42		BH Bahrain	4
SE Sweden	36		GE Georgia	4
CA Canada	33		LV Latvia	4
IQ Iraq	31		NZ New Zealand	4
AT Austria	30		OM Oman	4
CH Switzerland	30		HR Croatia	3
IL Israel and East Jerusale	30		PA Panama	3
IR Iran	30		PH Philippines	3
ID Indonesia	29		SI Slovenia	3
YE Yemen	28		UA Ukraine	3
AM Armenia	24		MR Mauritania	2
DZ Algeria	27		PE Peru	2
CN China	23		UG Uganda	2
HU Hungary	23		VA Vatican City	2
BR Brazil	22		CL Chile	1
IN India	22		CY Cyprus	1
RO Romania	22		KE Kenya	1
SY Syria	21		KG Kyrgyzstan	1
PL Poland	20		LK Sri Lanka	1
AR Argentina	16		MK Macedonia	1
GR Greece	15		ML Mali	1
JP Japan	15		MO Macao	1
IE Ireland	13		MY Malaysia	1
MA Morocco	13		NG Nigeria	1
AE United Arab Emirates	10		NI Nicaragua	1
CZ Czech Republic	9		SK Slovakia	1
KW Kuwait	9		SV El Savador	1
NO Norway	9		TW Taiwan	1

SEO Issues with vHMML

Background

Prior to vHMML 3.0, vHMML wasn't optimized for Google (and other) Search engines to index our content. As part of the 2018 NEH grant, we were tasked with improving this and getting our metadata searchable by web search engines.

NEH Grant Part 1

As part of the initial SEO work, we added functionality to generate a sitemap (or a list) of all the pages within vHMML. This is important because Google and other search engines use this list as an index of pages that they should crawl. Using this list, they examine the pages, then based on their algorithms, they crawl and index each page's content.

The second thing we added to vHMML was the ability for the application to generate and include JSON-LD for each object in Reading Room, Reference, Lexicon, and Folio (vHMML's 4 subsystems). Using the data within each object's page, the JSON-LD structured data is added to the page so that the search crawler can see and interpret it.

NEH Grant Part 2

After noticing that the previous SEO work wasn't having the effects we had hoped, we had an opportunity at the end of the grant to attempt some other changes to hopefully get some other SEO "quick wins". These were chosen after analysis by SDG and weighing the time to complete them. As part of this, we also asked SDG if they had an SEO resource on staff to do more analysis but unfortunately they didn't.

The first change was to add a unique title to each object's page within the 4 vHMML subsystems. Previously all titles were "Virtual Hill Museum & Manuscript Library". Now the permanent URL (PURL) titles are similar to below and unique to each object:

ZFRN 00010 - Turkey - Mardin - Dayr Za'farán (Monastery) - MS 10 - Reading Room - vHMML

The second change was to add unique <H1> HTML tag to each object's page within the 4 vHMML subsystems.

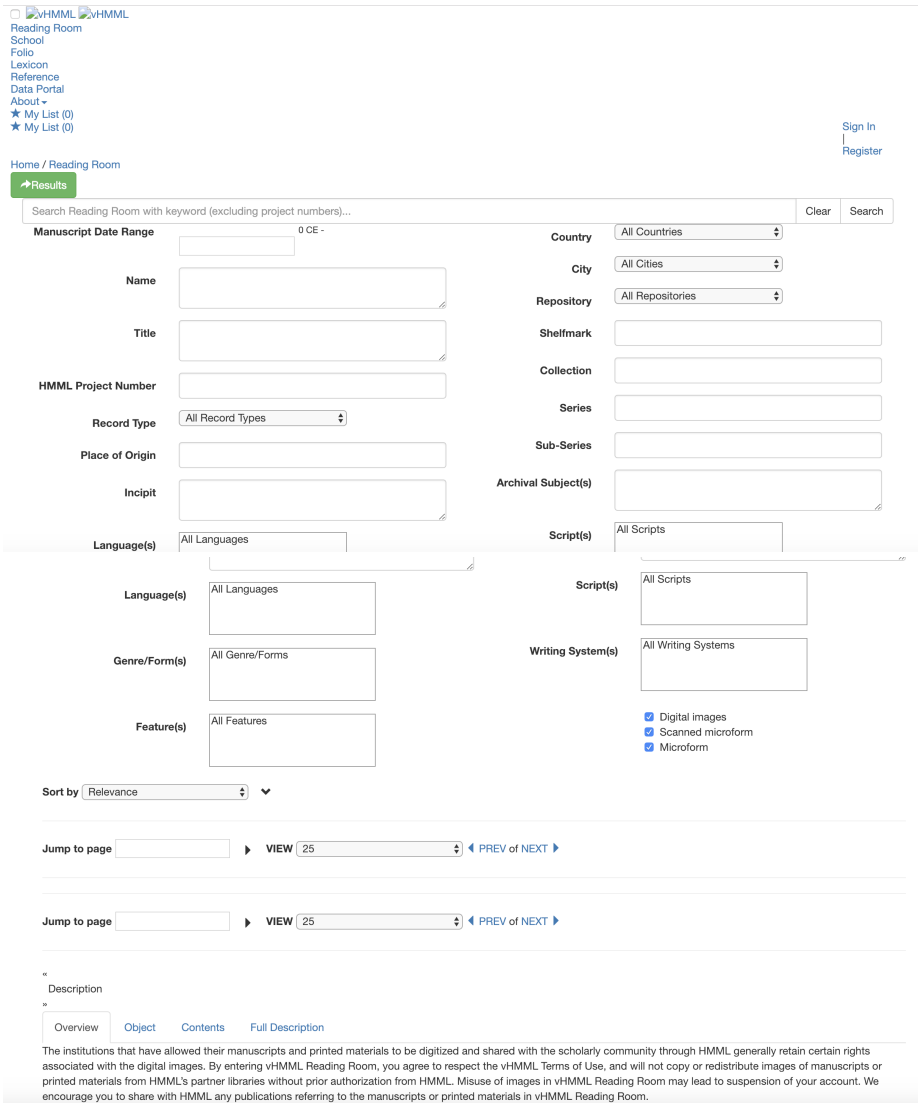
Problems Encountered

After part 1, using the Google and Bing search consoles (tools to see how a search engine is indexing our site), we noticed that the crawlers were seeing our content as the same or nearly the same for most of the objects. For example, when Google would crawl AG 1 and AG 2, even though there is separate metadata for each object and different images, when Google looks at the underlining HTML to analyze the page the two pages look nearly the same. This is inherently due to how vHMML was designed. Specifically, when using Reading Room, the application allows users to search, see results, and then view an object's detail page. Even though to the user these are separate actions that should result in separate, distinct pages (and therefore separate URLs), the application actually never changes its URL - it is always https://www.vhmml.org/readingRoom/. The system is essentially an enclosed system, a Single Page Application (SPA), coded within a JAVA application. So, to the browser it sees an application or page that is changing - but within itself - and not changing its URL.

The next issue is that all the URLs have the same meta tags. These are used to differentiate content by search engines.

Another issue within vHMML is the use of a templating system and JavaScript. The templating system allows vHMML to be dynamic. We have one base detail page and the application swaps or displays the appropriate data for the object a user wants to see. However, this is happening using Javascript so the rendering, or the changing of to this new page, happens after the search engine bot has looked at the page. For example, if we open
https://www.vhmml.org/readingRoom/view/4872 and
https://www.vhmml.org/readingRoom/view/60890

We as users see two different pages. But if we look at how the search engine crawler sees them (https://render-tron.appspot.com/render/https://w3id.org/vhmml/readingRoom/view/511255):



You'll notice that the page is very generic and the content hasn't been populated yet.

Underneath, in the code, this is part of what Google sees:

```
<script id="content-template" type="text/x-handlebars-template">
    <table class="indent">
        {{#if itemNumber}}
            <tr>
                <td><label class="heading">Item {{itemNumber}}</label><td>
            </tr>
        {{/if}}

        {{render 'meta-data-field-template' label="Item Location" value=itemLocation}}

        {{#each authors}}
            {{render 'meta-data-field-search-link-template' label="Author" linkText=contributor.populatedName
searchProp="name" lcUri=contributor.authorityUriLC viafUri=contributor.authorityUriVIAF}}
            {{render 'meta-data-field-template' label="Author NS" value=nameNs valueClass=../languageList}}
        {{/each}}

        {{render 'meta-data-field-template' label="Title" value=provisionalTitle}}
        {{render 'meta-data-field-template' label="Supplied Title" value=suppliedTitle}}
        {{render 'meta-data-field-template' label="Title NS" value=titleNs valueClass=languageList}}
```

The templating system is setup in the page but hasn't rendered the data yet, so the search engine can't see the data in its appropriate location on the page. If you look at the 5th line you see "{{itemNumber}}". This is telling the system to put the value of the item number here and then display it as part of the content. Unfortunately the search crawler bot won't see this data.

We can also see that the data is in the page but the value(s) are currently stored in a JavaScript variable. Ready to be populated into our template system but the JavaScript just hasn't been run yet:


```
        selectedObjectJson = {"id":4872,"summary":null,"processedBy":"HMML","shelfMark":"Neukloster Cod. B
6","commonName":null,"conditionNotes":null,"collation":null,"binding":null,"bindingDimensionsImported":null
e":null,"bibliography":"Eugen Bill, \"Verzeichniss der Handschriften in der Bibliothek des mit der Abtei He
Handschriften-Verzeichnisse der Cistercienser-Stifte, Xenia Bernardina 2.1
(1891)","notes":"Folio","colophon":null,"medium":null,"foliation":"","hmmlProjectNumber":"4944","inputter":
Finc","inputDate":null,"lastUpdate":1522267583000,"lastUpdatedBy":null,"reproductionNotes":null,"acknowledg
ion":{"name":"ON_SITE_ONLY_ORDER_SCAN","displayName":"On-site only or order a scan"},"viewableOnline":false
[],"iconName":null,"citeAs":null,"dataSource":"/var/cache/tomcat7/temp/20180328d_Weiner Neustadt
Subset.mdb","imagesLocation":null,"support":null,"active":true,"locked":false,"lockedBy":null,"publicManife
{"id":69,"name":"Austria","authorityUriLC":"http://id.loc.gov/authorities/names/n79040121","authorityUriVIA
Neustadt","authorityUriLC":"http://id.loc.gov/authorities/names/n80040800","authorityUriVIAF":"https://viaf
{"id":10,"name":"Abtei Heiligenkreuz"},"repository":{"id":528,"name":"Neukloster.
Bibliothek","authorityUriLC":"http://id.loc.gov/authorities/names/n88637961","authorityUriVIAF":"https://vi
[{"id":4869,"type":"MANUSCRIPT","partNumber":1,"partLocation":"fol. 1r-
178v","supportDimensionsImported":null,"supportImported":"Parchment","supportDimensionsWidth":null,"support
```


As a result of the template and JavaScript issues, even after part 2 we still didn't see Google or Bing start to index our content en masse. This lead to a further examination of vHMML's structure and SEO.

Javascript SEO

Currently most search bots/crawlers have problems with dynamically generated content or content that is populated via JavaScript - like vHMML. This has led to a new area of SEO analysis: JavaScript SEO.

From SDG analysis:

The Reading Room, Museum, and Folio searches are run dynamically using JavaScript to give that seamless feeling of a web application which is one of the few ways a user can access an object's details page.

One way we can handle the dynamically generated content is through the use of JavaScript SEO. JavaScript SEO is a subset of SEO, and focuses primarily on web applications built heavy on JavaScript such as jQuery, React, Angular, VueJS, etc.

Googlebot is smart, but it's not smart enough yet. One way to help is through dynamic rendering.

Reference for dynamic rendering:
https://developers.google.com/search/docs/guides/dynamic-rendering

Implementing dynamic renderers however require environment changes to the servers to accommodate the technology used. Resources would also need to be increased if the dynamic renderer is hosted on the same web server.

Using this information we can plan for, and discuss the future of vHMML and its SEO. Current estimates to add this functionality put this work at 1 month of SDG analysis and programming and ~$2000 a year of additional cost to pay for a service provider, such as dynamic renderers like Puppeteer, Rendertron, and prerender.io.

Other Options

As a short-term way to circumvent or help search engines find our content, we could build a finding aid separately from vHMML. This would ideally be a website that would be indexed by Google and point to the vHMML content. Thus making our content findable via a search engine in a roundabout way.

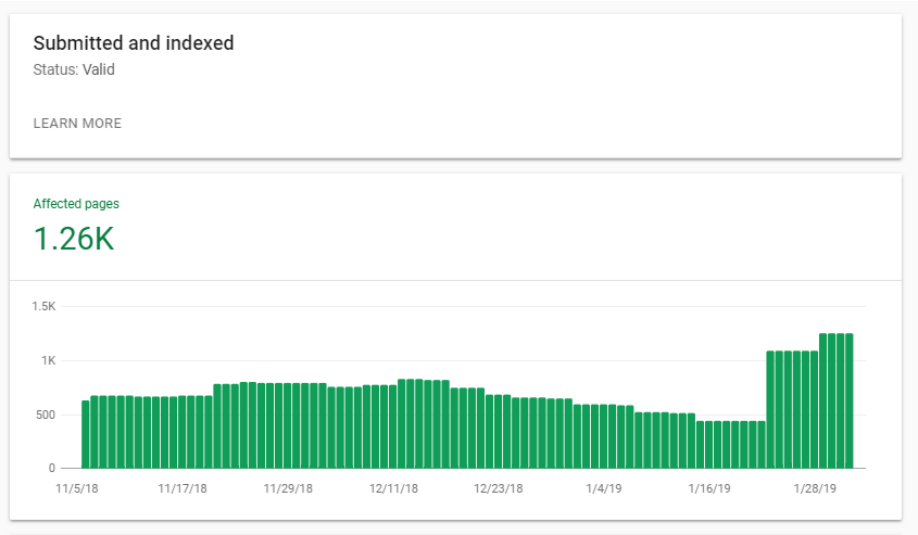
Here are some ideas on ways to implement this:

1) Publish PDFs of minimal amounts of data to Archive.org. This makes the text ORC discoverable. This would be similar to how the HMML checklists are made public: https://ia800201.us.archive.org/7/items/HMMLSpainChecklist/HMMLSpainChecklist.pdf
2) Publish minimal amounts of data as a WordPress site. Again, the site would be searchable and have a list of our content.
3) Publish data as a Static Website.

These options could be quicker and simpler to implement.

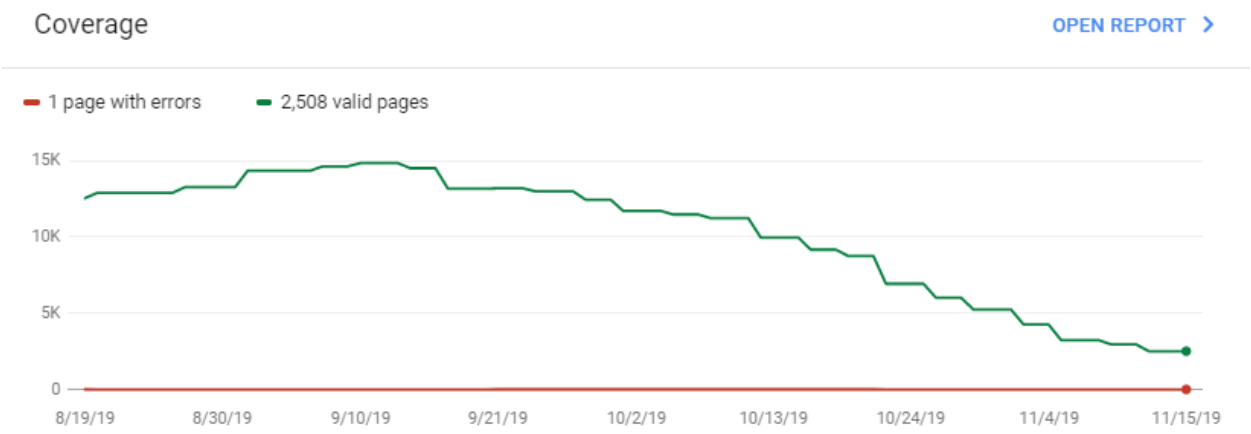
Update 2/4/2019

On January 23rd, there was a big jump in pages searched and indexed. Can we hope this will continue?



Update 11/15/2019

Since mid-September, we've seen a considerable drop in the Google search coverage of vHMML. From nearly 15 thousand vHMML pages indexed to under 3 thousand.



Why the drop? This is more easily asked than answered. Google SEO and Google Search has always been a bit of a black box. Some reasons are listed in the earlier part of this document and a recent change to mobile first indexing might be another part of the reason. JavaScript SEO, the templating system or dynamic generation of the site, and URL issues are all still issues which may have contributed to the fall in indexed pages. After more research, a couple other factors arose as concerns and areas for improvement.

Linked pages - Google recommends that all pages on the site be reached by a link from another findable page. The referring link should include either text or, for images, an alt attribute, that is relevant to the target page. Crawlable links are <a> tags with an href attribute. This is something that vHMML does not do. They only method to get to an object detail page is through (internal) search. Users can't click "HMML" to see the HMML collection, then click into "HMML 00012" to see that object's details with the resulting URL:
http://vhmml.org/readingRoom/HMML/HMML%2000012/

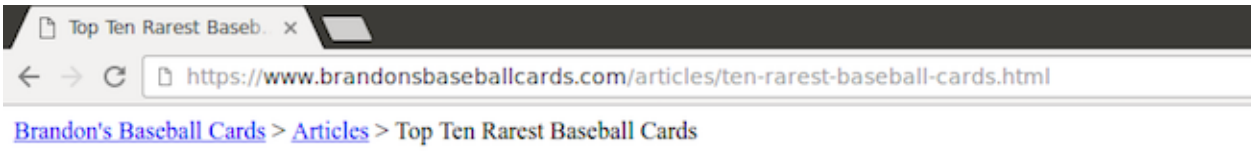
With vHMML's PURL method and page structure, the PURL for that same HMML 00012 object is never linked to from any place on vHMML (https://www.vhmml.org/readingRoom/view/511273).

Meta Tags - With a PURL, Google should see a different title and JSON-LD information for each page that they index. However, when they look at other elements in the page's code, they see the same information across the pages. For example the Meta Description tag is the same for every single object in vHMML:

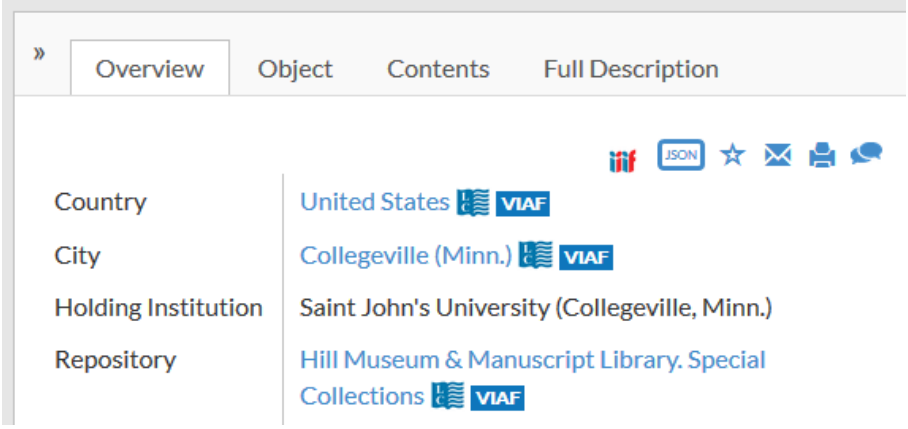
> <meta name="description" content="Digitized manuscripts archives and rare books from Europe the near East and Northern Africa">

This leads Google to think the pages are the same and not different enough to be indexed.

BreadCrumbs - None of the vHMML object pages have breadcrumbs. A breadcrumb is a row of internal links at the top or bottom of the page that allows visitors to quickly navigate back to a previous section or the root page. Many breadcrumbs have the most general page (usually the root page) as the first, leftmost link and list the more specific sections out to the right. Google also recommends using breadcrumb structured data markup when showing breadcrumbs. Something which vHMML does not do.



Tabs - Google also wants our content to be visible by default. Google is able to crawl HTML content hidden inside navigational elements such as tabs or expanding sections, however they consider this content less accessible to users, and believe that we should make the most important information visible in the default page view. On the object detail page, vHMML is utilizing tabs which may be another cause of our recent SEO indexing drop.



Options to improve SEO

As detailed above, we have a few options with different levels of time and cost.

Do nothing - make no changes and accept our level of content indexed by Google.

Add Dynamic Renderer to vHMML - This means that vHMML would be smart enough to detect crawlers and give them static html without any JavaScript vs. the dynamically generated content users see. SDG would install software on the vHMML servers, or add code into vHMML, that would route traffic to vHMML for a normal user OR to a pre-rendered vHMML page that lives on a paid service, or SJU server, if it is search-bot (Googlebot). A rough estimate of time required by SDG would be 2-4 weeks. Additionally, there may be some off-site hosting costs which could be anywhere from hundreds of dollars to $1000 a year. In this case, we might

be dependent on the off-site hosting service and their sustainability. An SDG estimate would be necessary to properly access this option.

Create sitemap page - Create a page like vhmml.org/sitemap that has a list of all of the objects and/or collections using their PURL links. This would hopefully facilitate Google crawling and indexing all of the distinct pages on vHMML.

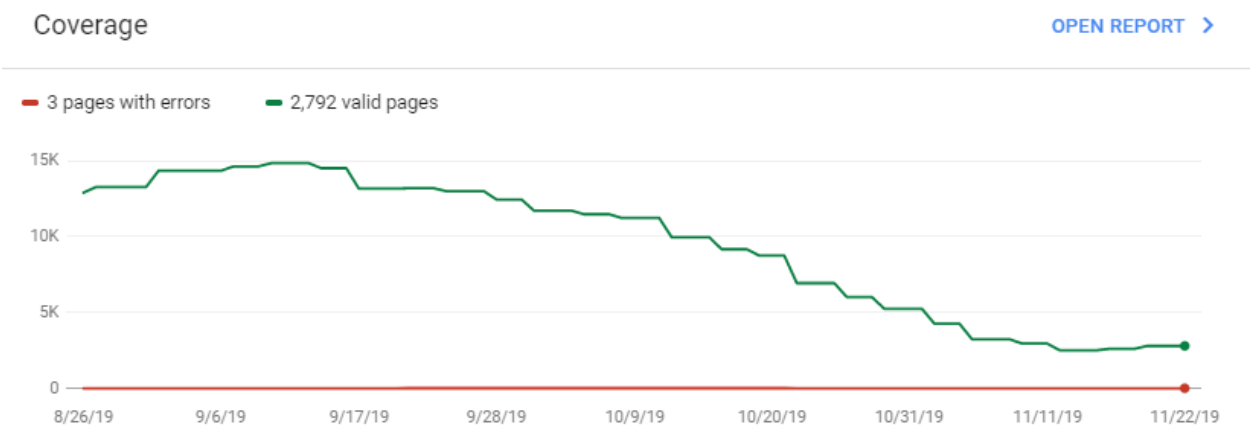
Separate finding aid - As a way to circumvent or help search engines find our content, we could build a finding aid separately from vHMML. This would ideally be a website that could be properly indexed by Google and point to the vHMML content. Thus making our content findable via a search engine in a roundabout way. All three options below would be free or a minimal amount of money. The downside with these options is that our content is duplicated and users may find this other website/content first via Google then have to find their way to the actual vHMML. These options would also require periodic updating as records in vHMML are changed.

Here are some ideas on ways to implement this:

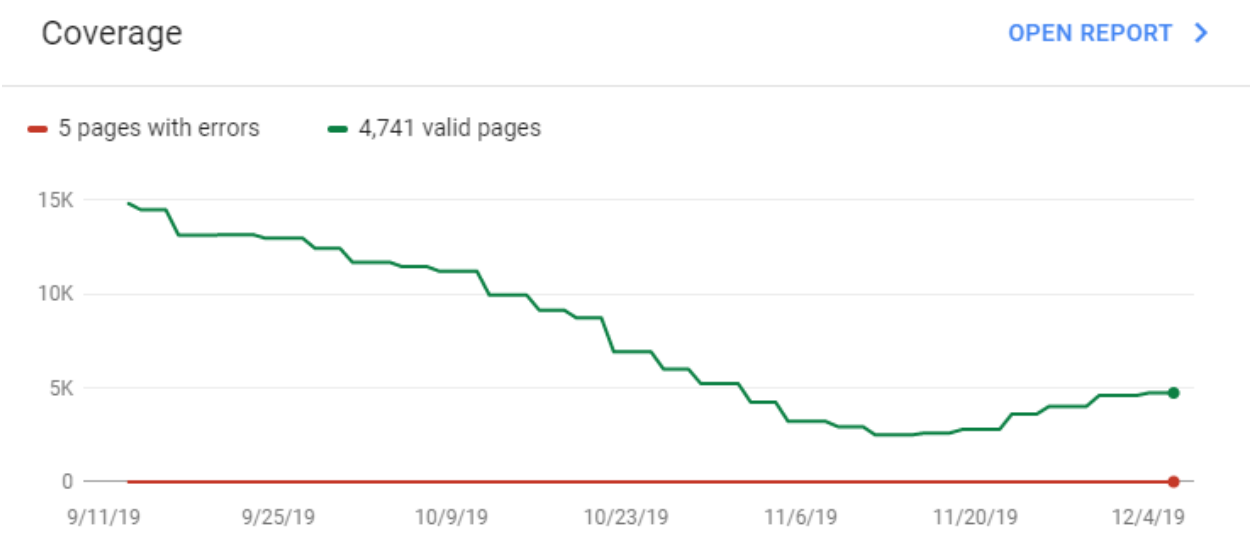
- Publish PDFs of minimal amounts of data to Archive.org. This makes the text ORC discoverable. This would be similar to how the HMML checklists are made public: https://ia800201.us.archive.org/7/items/HMMLSpainChecklist/HMMLSpainChecklist.pdf
- Publish minimal amounts of data as a WordPress site. Again, the site would be searchable and have a list of our content.
- Publish data as a Static Website. This could be a seperate site that points to vHMML content or possibly part of vHMML.
- Wikidata - objects.
- Collection level description pages on Wikidata and ArchiveGrid.

I want to stress that there is no guarantee that the Google SEO for vHMML will be fixed or improved with any of these methods.

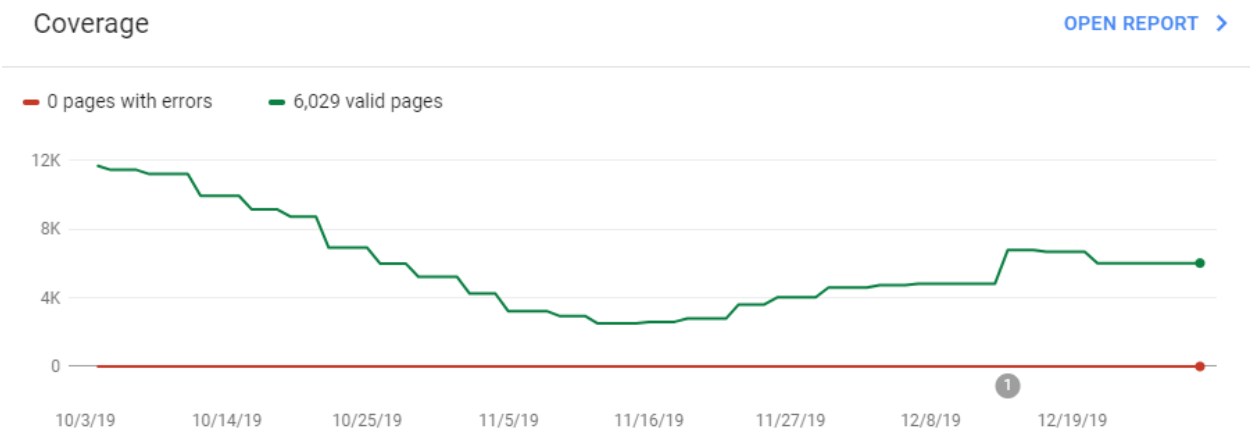
Update 11/25/2019



Update 12/11/2019



Update 1/2/2020



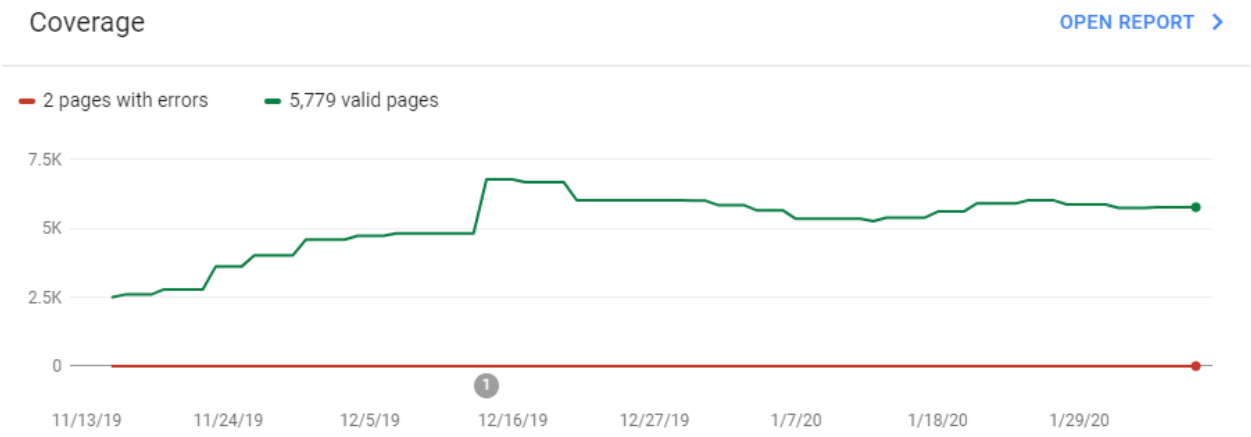
Update 1/7/2020

Other services:

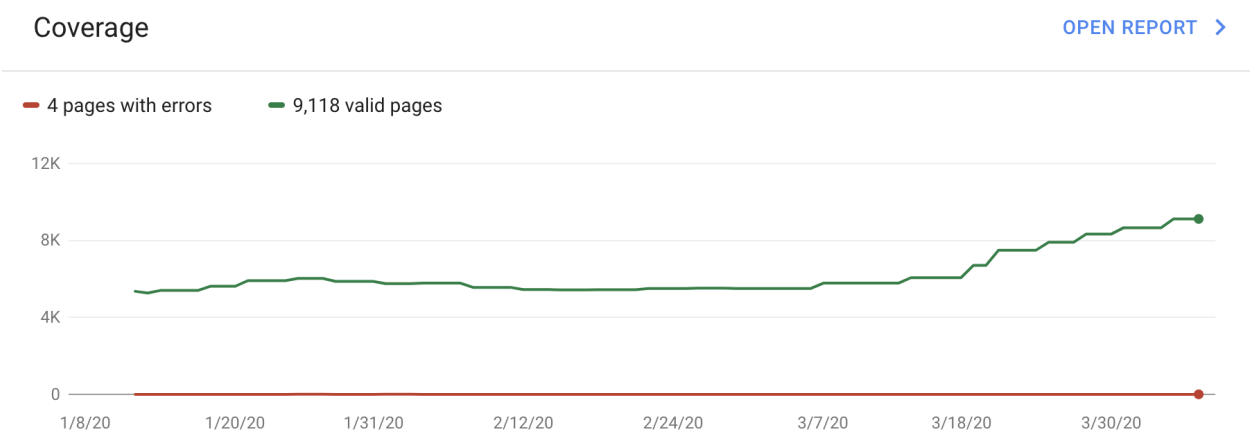
Prerender.io
Brombone
SEO4Ajax
Prerender.cloud
Prerender.io has an open-source version of their service that you can self-host.

https://docs.netlify.com/site-deploys/post-processing/prerendering/#set-up-prerendering

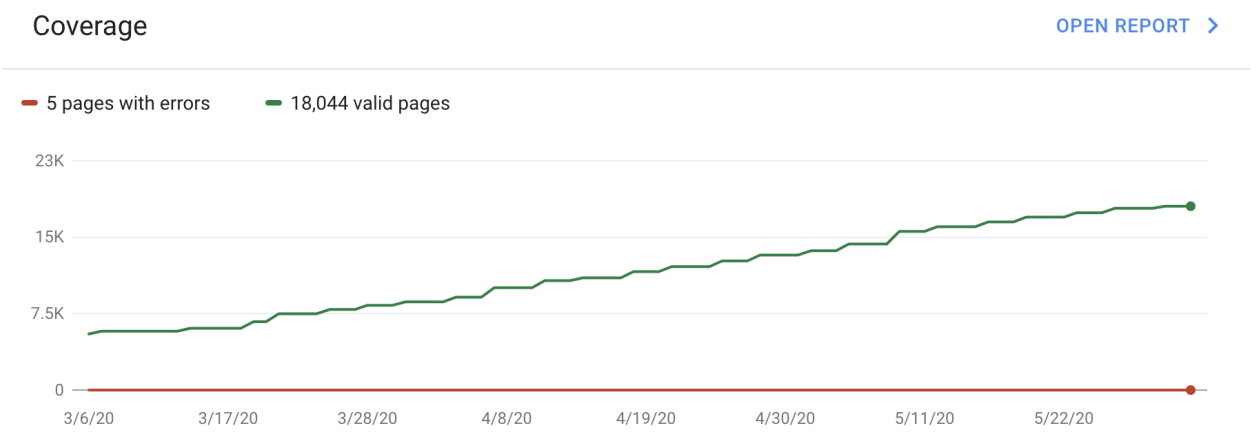
Update 2/12/2020



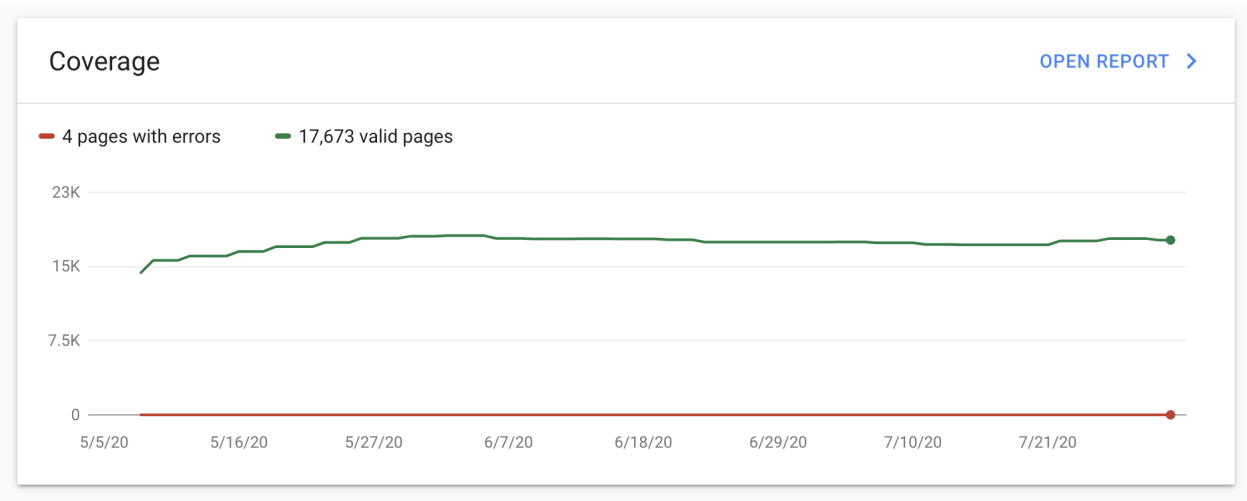
Update 4/8/2020



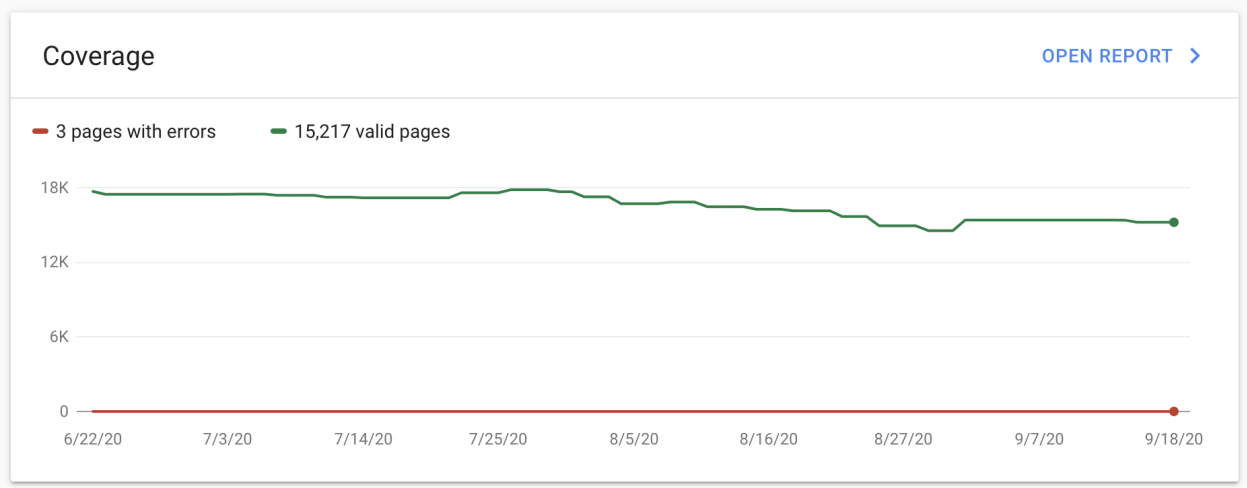
Update 6/5/2020



Update 8/4/2020



Update 9/21/2020



CrossMark

Giving Voice to Ancient Texts: Manuscript Scholarship in the Digital Era

Columba Stewart
Hill Museum & Manuscript Library, Saint John's University, Collegeville, Minn.; e-mail: cstewart@hmml.org

doi:10.1017/S002074381700099X

The study of manuscripts was traditionally the preserve of scholar-curators in research libraries who devoted their lives to the exhaustive study and description of collections that were often gathered from several sources. The expected result of their solitary labor was a printed catalog that might describe at most a few hundred manuscripts from a particular linguistic or religious culture. That model has been challenged by the advent of large-scale digital projects that aggregate thousands or even tens of thousands of manuscripts from multiple libraries in a single database or portal.

This contribution is based on my experience with such a project at the Hill Museum & Manuscript Library (HMML) at Saint John's University in Collegeville, Minnesota. I have been Executive Director since 2003, leading a move from the largely Euro-centric work of HMML's first decades to a new focus on manuscript cultures in the Middle East, Africa, and India, encompassing both Christian and Islamic traditions. HMML had already undertaken an extraordinary microfilming project in Ethiopia in the 1970s (the Ethiopian Manuscript Microfilm Library [EMML]). While the results were groundbreaking for Ethiopian studies, the 8,000 manuscripts were less than 10 percent of HMML's microfilm holdings. The current total of microfilmed and digital manuscripts approaches 200,000 items. Most are codices, with the remainder being archival materials. Thousands more digitized manuscripts are being added every year from digitization projects across the world.

In the microfilm era, HMML published catalogs of many collections, most significantly the ten volumes for the EMML. Some were simply inventories or handlists. Others, like the many volumes of the EMML catalog, provided detailed, text-level descriptions. Copies of microfilms were made upon request, typically requiring several weeks to process because of the need to seek permission from the owning institution and then to send the film to a processor for duplication. As the digital era dawned in the 1990s, HMML was part of a consortium funded by the Andrew W. Mellon Foundation to develop standards for encoding, storing, and presenting catalog descriptions of medieval and Renaissance manuscripts in electronic form. The project, Electronic Access to Medieval Manuscripts (EAMMS), created a complex metadata scheme consisting of nested tables in a relational database to represent: the complete manuscript object; its parts (if the object consisted of more than one text block bound together into a single object); and individual texts. Because the metadata scheme was designed by scholars of premodern western manuscripts, it was inevitably oriented toward the manuscript types familiar to them. The EAMMS scheme would become the basis for the Digital Scriptorium[1] initiative and the European Manuscript Access through Standards for Electronic Records (MASTER) project.[2] HMML's implementation of EAMMS went online in 1999 and continues to exist as a legacy catalog while its contents are being

Downloaded from https://www.cambridge.org/core. IP address: 12.14.15.194, on 09 Feb 2018 at 16:31:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S002074381700099X

conformed to new standards for inclusion in the *Virtual* HMML (*v*HMML) Reading Room (described below).

HMML's 2003 initiative in the Middle East introduced color digital imaging and a new workflow. Photographic work at the many field sites is done by local teams trained and paid by HMML. Their equipment is provided by HMML. Regional HMML field directors provide technical support and manage the flow of data from each site to Minnesota, where images and metadata are archived both on- and off-site. In the early digital years, HMML continued to catalog manuscripts using the EAMMS scheme with some adjustments to support description of nonwestern manuscripts. By 2009 the inadequacy of both the scheme and the underlying technology had become clear. An obvious need was stricter authority control for toponyms, institutional and personal names, languages, scripts, and other features of manuscripts to provide a more consistent search experience. New funding from the Andrew W. Mellon Foundation supported some adjustments to the cataloging procedures and the hiring of more catalogers. An adjunct online system allowed faster, Google-like, searching of the EAMMS-based catalog. For born-digital objects the online catalog now included links to sample images delivered via HMML's implementation of ContentDM, known as *Vivarium*. Catalogers at HMML used an MS Access form to add records to the database. External catalogers entered metadata using templates created in MS Excel. These were then reviewed by HMML curators and exported to a format that could be aggregated to the database. Manuscript images were provided to scholars on CDs upon payment of a service charge or, starting in 2012, free of charge via password-protected galleries in *Vivarium*.

In the early 2010s, HMML began to develop other online tools to support manuscript studies across a wide range of languages and cultures, with an initial focus on students and scholars just starting to use manuscripts in their research. The resulting *v*HMML platform was launched in 2015 with paleography lessons in Latin and Syriac, densely annotated sample images (an online paleographical album), and reference tools.[3] It drew on the pedagogical expertise of HMML staff and outside consultants to complement, or in many cases, substitute for, classroom instruction on manuscript studies.

As the original *v*HMML platform was in development, HMML staff were at work on the next phase of the project, *v*HMML Reading Room, which would provide access to complete image sets of all digitized manuscripts (both born-digital and scanned microfilm) in HMML's collections. A grant from the Henry Luce Foundation supported planning and software development for Reading Room. The issue of authority control for names and features resurfaced with new urgency. Besides improving the search experience for users, rigorous consistency would make data more easily shareable with (and usable by) other Digital Humanities projects. Despite good intentions, HMML's previous efforts at enforcing strict metadata rules had proven inadequate, partly because of flaws in the metadata scheme. A thoroughly revised metadata scheme would be needed for *v*HMML Reading Room. Devising it required both traditional scholarly expertise and familiarity with developments in the Digital Humanities. The latter was not viewed as a mere adjunct to subject knowledge or as an auxiliary technology, but rather as the form of much present-day study of the humanities.

Downloaded from https://www.cambridge.org/core. IP address: 12.14.15.194, on 09 Feb 2018 at 16:31:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S002074381700099X

The new metadata scheme for *v*HMML Reading Room fully accommodates non-western manuscripts, and additionally provides alternative schemes for printed objects, hybrid manuscript-printed objects, and archival materials. Strict application of LC and VIAF authority control for toponyms, institutional names, personal names, and titles (where they exist) is presumed throughout.[4] HMML's curators developed controlled vocabularies for languages, scripts, and various descriptive fields according to current standards. An online cataloging interface was created to support metadata entry by both HMML staff and external catalogers, with a range of permission levels designed to provide quality control for cataloging submitted by off-site catalogers. *v*HMML Reading Room uses the MIRADOR image viewer, a platform compliant with the International Image Interoperability Framework (IIIF), and chosen for its superior viewing experience, potential for sharing with other Digital Humanities projects, and robust user community.[5]

*v*HMML Reading Room was launched in August 2016 as part of *v*HMML 2.0. All metadata is freely searchable by anyone visiting the site. Viewing of complete image sets generally requires free, one-time registration. The implementation of Elasticsearch[6] technology provides greatly improved searching and faceted search options. Each object now has a permalink for reference purposes. *v*HMML Reading Room does not yet support downloading of images, though for most collections this would not be available anyway because the owners would not permit it.

Imposing strict control of names and descriptions of features for objects in *v*HMML Reading Room unavoidably slows the process of metadata creation. To mitigate this problem, some collections are being initially uploaded to *v*HMML Reading Room with minimal metadata sets to make them available without delay. This serves scholars who know which manuscript they need (on the basis of previous catalogs or scholarly references) or who wish to browse uncataloged, or even unknown, collections. Such "stub" records contain basic and stable identifiers (city, repository, shelfmark, HMML project number) as well as the permalink. Over time the records can be enriched with additional metadata conformed to the standards of *v*HMML Reading Room.

*v*HMML currently supports data sharing with other projects via exports of the entire catalog database in JSON format and of the permalinks for all objects in CSV format. *v*HMML 3.0, currently in development with funding from the National Endowment for the Humanities, will allow export of metadata in encoded form (e.g., METS, OAI, and EAD). The new development phase includes exploration of technologies supportive of linked data. The Mirador viewing environment supports easy invoking of images from other projects via IIIF manifest URLs generated for each object. The IIIF manifest URLs are not generally exposed in *v*HMML Reading Room because of restrictions imposed by the owners of the manuscripts. They do appear in the metadata display for items from HMML's own collections and for scans of the EMML microfilms. As our partners grow more confident about the benefits of image sharing, they may choose to allow exposure of the IIIF manifest URLs. Upon request HMML can supply trusted projects with the IIIF manifest URLs of currently restricted manuscripts or collections to permit viewing and annotation in other IIIF environments.

A major Digital Humanities project like *v*HMML offers several kinds of intellectual labor, whether for HMML staff, external catalogers, or the growing number of partner projects using *v*HMML Reading Room as the environment for cataloging and

Downloaded from https://www.cambridge.org/core. IP address: 12.14.15.194, on 09 Feb 2018 at 16:31:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S002074381700099X

displaying their digital objects. There is the obvious task of describing the manuscripts, which requires a different approach from traditional cataloging oriented toward creation of printed reference catalogs. In the online environment, emphasis is placed on discoverability, both of texts and of the material aspects of manuscripts. The latter is especially important for those interested in the production and circulation of manuscripts. The presence of several hundred manuscript libraries in *v*HMML Reading Room permits analysis of collections both as distinct entities and as components of larger data sets. Digitization of several libraries belonging to different communities in the same city or region (such as Aleppo or Jerusalem) allows mapping of intellectual microclimates and the exchanges between them.

Although the metadata fields in *v*HMML Reading Room allow inclusion of highly detailed descriptions, limited resources and the urgency of making as many manuscripts available as quickly as possible mean that, for the present, such traditional exhaustive cataloging is reserved for highly important manuscripts. Given HMML's priorities of digital preservation and access, development of richer descriptions will perhaps best be done by partner projects focused on particular manuscript cultures or textual traditions, or through contributions by individual scholars willing to share their findings with the user community. *v*HMML 2.0 provides basic support for user-contributed corrections or suggested additional metadata; one of the goals for *v*HMML 3.0 is an improved suite of tools for contributions and corrections of metadata.

One of the major imperatives for *v*HMML and similar projects is awareness of evolving expectations for projects in the Digital Humanities for content (both quantity and quality) and for the ease and richness of the user experience. Project teams must pay close attention to other initiatives while having a clear understanding of the purpose and focus of their own project. Any shifts or expansions of scope must be coherent and sustainable. The range of content in *v*HMML Reading Room is already vast, with some 23,000 objects in more than forty languages in the database as of September 2017, most of them with full image sets. More are added weekly. For this and other projects, ambitions for further development of the software and improvement of the user experience are necessarily constrained by limitations of human and financial resources. Wise decisions about choosing which features to add when a new development cycle begins require counsel from an external advisory board and regular communication with users. The dependence of most Digital Humanities initiatives on external grants for initial development can obscure the challenge of sustainability, a central concern for projects like HMML's that promise long-term access to digital assets.

In sum, *v*HMML and its Reading Room call for new kinds of humanists, equally at home in traditional textual study and in the newer forms of scholarship enabled by direct access to tens of thousands of manuscripts representing many cultures. This is surely intellectual labor of a high order.

NOTES

[1]Digital Scriptorium, accessed 29 September 2017, http://bancroft.berkeley.edu/digitalscriptorium/.

[2]Manuscript Access through Standards for Electronic Records, accessed 29 September 2017, http://master.dmu.ac.uk/.

[3]Paleography lessons for Arabic and Armenian manuscripts will be released in 2018 as part of *v*HMML 3.0.

Downloaded from https://www.cambridge.org/core. IP address: 12.14.15.194, on 09 Feb 2018 at 16:31:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S002074381700099X

[4]HMML is contributing new names and titles to the Library of Congress to build up the nonwestern authority records.

[5]For Mirador, see http://projectmirador.org/, accessed 29 September 2017, and on IIIF, see http://iiif.io/, accessed 29 September 2017.

[6]For more information on Elastic Search, see https://www.elastic.co/products/elasticsearch, accessed 29 September 2017.

Downloaded from https://www.cambridge.org/core. IP address: 12.14.15.194, on 09 Feb 2018 at 16:31:37, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S002074381700099X