

A NEW TOOL FOR COMPUTER ASSISTED PALEOGRAPHY: THE DIGITAL ANALYSIS OF SYRIAC HANDWRITING PROJECT

MICHAEL PENN

DEPARTMENT RELIGIOUS STUDIES, STANFORD UNIVERSITY

VIJOY ABRAHAM

CIDR DEVELOPMENT TEAM, STANFORD UNIVERSITY LIBRARY

SCOTT BAILEY

NORTH CAROLINA STATE UNIVERSITY LIBRARIES

PETER BROADWELL

CIDR DEVELOPMENT TEAM, STANFORD UNIVERSITY LIBRARY

R. JORDAN CROUSER

DEPARTMENT OF COMPUTER SCIENCE, SMITH COLLEGE

JAVIER DE LA ROSA

DIGITAL HUMANITIES INNOVATION LAB, UNED UNIVERSITY

NICHOLAS HOWE

DEPARTMENT OF COMPUTER SCIENCE, SMITH COLLEGE

AND SIMON WILES

CIDR DEVELOPMENT TEAM, STANFORD UNIVERSITY LIBRARY

ABSTRACT

Over the last nine years, a forty-seven-person digital humanities project has explored the feasibility of computer assisted paleography for Syriac. That is, could one use big data, visual analytics, and recent advances in the digital analysis of handwriting to better understand Syriac manuscripts and Syriac manuscript culture? Last June we launched the public-facing part of our project titled the “Digital Analysis of Syriac Handwriting” or DASH (dash.Stanford.edu). DASH consists of digital page images from 90% of the world’s surviving Syriac manuscripts securely dated to before the twelfth century, as well as 90,000 individually trimmed letters from these manuscripts. In addition to introducing this tool to scholars of Syriac, this article presents a series of novel visualization tools that can also serve as a starting point for digital paleography projects in other linguistic traditions.

Thanks in large part to the libraries of Deir al-Surian and Saint Catherine’s monasteries, Syriac has a surprisingly high number of very early manuscripts. Syriac scribes also included dated colophons in their works much more frequently than scribes from other traditions.¹ As a result, there is a truly impressive set of early, securely dated Syriac manuscripts.² But in Syriac

¹ Sebastian Brock, "Without Mushê of Nisibis, Where Would We Be?" *Journal of Eastern Christian Studies* 56 (2004): 15–24.

² For example, 14% of the 615 British Library manuscripts thought to have been written in the first millennium include a securely dated colophon. When extended to all fifth- through nineteenth-century Syriac manuscripts, the proportion of those with securely dated colophons increases to approximately 30% (Sebastian P Brock and David G.K. Taylor, *The Hidden Pearl: The Syrian Orthodox Church and Its Ancient Aramaic Heritage*, v.2. [Rome: Trans World Film Italia, 2001], 245). By comparison,

Studies there remains a disconnect between this plethora of early manuscripts and the lack of paleographic resources to help modern scholars analyze them. Several recent articles and an unpublished dissertation have addressed some aspects of Syriac script.³ So, too, the Hill Museum and Manuscript Library at Saint John's University has produced an on-line tutorial of various Syriac scripts and their proper transcription.⁴

among Hebrew manuscripts the percentage with dated colophons is just 3.4%, that is almost ten times less frequent than in Syriac manuscripts (Malachi Beit-Arié, *Hebrew Codicology: Historical and Comparative Typology of Hebrew Medieval Codices Based on the Documentation of the Extant Dated Manuscripts from a Quantitative Approach Pre-Publication, Internet Version 0.1*, 2012). The comparative frequency by which Syriac works include a dated colophon results in a total of 148 extant Syriac codices having a secure date before 1001. In comparison, Armenian—although with a total manuscript corpus about 50% larger than Syriac—has just fifteen manuscripts securely dated to the first millennium (Michael E. Stone, Dickran Kouymjian, and Henning Lehmann. *Album of Armenian Paleography* [Aarhus, Denmark: Aarhus University Press, 2002], 512). Hebrew with a total corpus approximately five times larger than Syriac has eleven (Colette Sirat and N. R. M. De Lange. *Hebrew Manuscripts of the Middle Ages* [Cambridge: Cambridge University Press, 2002] 11).

³ E.g. Ayda Kaplan, "The Shape of the Letters and the Dynamics of Composition in Syriac Manuscripts (Fifth to Tenth Century)," in D. Stutzmann, S. Barret, and G. Vogeler (eds.), *Ruling the Script in the Middle Ages: Formal Aspects of Written Communication (Books, Charters, and Inscriptions)* (Turnhout: Brepols, 2016) pp. 379-98; idem, "La Paléographie Syriaque: Proposition d'une Méthode d'expertise," in Fraçoise Briquel Chatonnet and Muriel Debié (eds.), *Manuscripta Syriaca: Des Sources de Première Main* (Paris: Geuthner, 2015) pp. 307-19; idem, "Syriac Paleography: The Development of a Method of Expertise on the Basis of the Syriac Manuscripts of the British Library (Vth-Xth C.)" (PhD Diss., Université Catholique de Louvain, Louvain-La-Nueve, 2008); Andrew Palmer, "The Syriac Letter-Forms of Tür Abdin and Environs," *Oriens Christianus* 73 (1989), pp. 68-89; Sebastian Brock and Lucas Van Rompay, *Catalogue of the Syriac Manuscripts and Fragments in the Library of Deir-Al Surian, Wadi Al-Natrun (Egypt)* (Leuven: Peeters, 2014) pp. XXI-XXII.

⁴ See, www.vhmmhschool.org/syriac.

Nevertheless, the only published, book-length resource for Syriac paleography remains William Hatch's 1946 *Album of Dated Syriac Manuscripts*.⁵ Hatch's *Album* reproduces a black and white photograph of a single page from each of 200 Syriac manuscripts securely dated between 411 and 1586 CE.

Given the difficulties of obtaining manuscript photographs in the 1930s and 1940s, Hatch's *Album* is a marvel. But it contains only 116 or 63% of the 183 extant Syriac manuscripts that are securely dated to before the twelfth century and its coverage of later manuscripts is much, much spottier.⁶ Even more limiting, however, is the very format of an album. When modern scholars use Hatch's book, they are constrained to the unit of an individual manuscript page. For example, when confronted with the task of estimating the composition date for a manuscript of interest, one often flips through Hatch's *Album* searching for a similar looking manuscript page among Hatch's

⁵ William Henry Paine Hatch, *An Album of Dated Syriac Manuscripts* (Boston: The American Academy of Arts and Sciences, 1946).

⁶ For an invaluable check list of early, securely dated Syriac manuscripts see Sebastian Brock, "A Tentative Checklist of Dated Syriac Manuscripts up to 1300," *Hugoye: Journal of Syriac Studies* 15,1 (2012): 21-48. Our own figures are based on a slight modification of Brock's list and include the following emendations: Added: *Saint Marks* 7; *Sinai Syriac* M51N. Because in CPA script removed *Vatican Syriac* 19. In an abundance of caution, removed those manuscripts that had either a pre- or circa date (i.e. *British Library Additional* 14,526; *British Library Additional* 14,567; *British Library Additional* 14,605; *Damascus Patriarch* 12/25) or a missing number in the colophon (i.e. *British Library Additional* 7158; *Dolabani* 145). Removed *British Library Additional* 14,645 as the colophon date refers to when the text was translated not when it was produced. Removed manuscripts that others have identified as having an incorrect date in their colophon (i.e. *Chester Beatty* 701; *Harvard Syriac* 176; *Paris Syriac* 169). Removed *St. Petersburg N/S* 24 as the folio is too damaged to ensure that the date that appears without much context in the colophon is a composition date. Removed *Mingana Syriac* 106G as the stray leaves Mingana identified are no longer locatable in the University of Birmingham collection.

samples. But, often the desired unit of comparison is not an entire manuscript page, rather individual letter forms. In essence, to use Hatch's *Album* one has to compare the *olaphs* in each manuscript with those of the document in question. Then it is time for the *beths*, at which point one forgets most of the *olaphs*, and so on. Even if one successfully determines the most analogous page, the same procedure needs repeating as soon as one is interested in a different manuscript.

The field of Syriac Studies has advanced greatly in the last seventy-five years. It makes little sense for our knowledge of Syriac paleography to remain almost entirely dependent on Hatch's 1946 *Album* or—given the inherent limitations—any print publication. Thanks to support from the Andrew Mellon Foundation, the American Council of Learned Societies, and the Stanford University Library, we assembled a team of Syriac scholars, paleographers, computer scientists, software developers, and dozens of research assistants to help address this issue through the creation of an on-line tool we entitled the Digital Analysis of Syriac Handwriting or DASH.

We began the DASH project by compiling the world's largest image collection of securely dated Syriac manuscripts. Currently, DASH contains digital images from 156 of the 183 extant Syriac manuscripts securely dated to before 1101 (that is 85%), along with permissions for their public display.⁷ We then created a back- and a front-end interface to help visualize this data. The public facing, front-end interface—along with its

⁷ One can supplement these digital images with published images for which we were not able to obtain image permissions to produce a collection representing 96% of early securely dated Syriac manuscripts. Most of these remaining images can be found in Sebastian Brock and Lucas Van Rompay, *Catalogue of the Syriac Manuscripts and Fragments in the Library of Deir-Al Surian, Wadi Al-Natrun (Egypt)* (Leuven: Peeters, 2014) and Philothée du Sinaï's *Nouveaux Manuscrits Syriaques du Sinaï* (Athènes: Foundation du Mont Sinaï: 2008).

collection of manuscript and letter images—is freely accessible at dash.Stanford.edu. Those interested in the source code for the entire project can find it in Github at <https://github.com/sul-cidr/scriptchart-backend> and <https://github.com/sul-cidr/scriptchart/>.

THE BACK-END INTERFACE: IDENTIFYING INDIVIDUAL SYRIAC LETTERS

Paleographers are often interested not simply in entire manuscript pages but also individual letters. When designing DASH, we found it essential—unlike in Hatch’s *Album*—to include images not only of entire manuscript pages but also of individual letters. Despite recent advances in the optical character recognition of Syriac, especially of printed text, the OCR of Syriac manuscript handwriting is still in its infancy. As a result, our collection of individual Syriac letter images stems from manual identification. To compile this dataset, our team created in Python a “behind the scenes” back-end DASH interface that allows research assistants to rapidly identify multiple examples of a given Syriac letter.

As shown in *Figure 1*, a research assistant uses a pull-down menu to select a specific manuscript page. They then choose a given letter to identify. They next drag selection boxes around examples of that letter. As they do this, the system records the coordinates of each selection box along with appropriate metadata into a myseql database. After the research assistant has identified several examples, they move on to another letter form, another page, or another manuscript.

In our case, we hired almost two dozen research assistants who used the back-end interface to place selection boxes around 132,476 Syriac letters. Employing the boxes’ coordinates, the system later extracted each letter. Then, using custom designed algorithms, one of our computer scientists binarized

the images.⁸ That is, he converted them into a black-and-white version particularly easy for humans and computers to read. In order to produce a particularly clear letter image, research assistants then used a simple graphics editor (GIMP) to remove any remnants of adjacent letters or other stray marks. They then loaded the single-letter image back into the database. Finally, scholars of Syriac proofed these letter images—multiple times—eliminating misidentifications, mistrimings, or any images that were poorly binarized. The result was a particularly clean dataset of 88,075 usable letter images.

⁸ Nicholas Howe, Minyue Dai, and Michael Penn, “Isolated Character Forms from Dated Syriac Manuscripts,” *Proceedings of Historical Image Processing* (November, 2017): 7-12.

The screenshot displays the DASH interface. At the top right is a 'DOWNLOAD' button. Below it is a 'Letters' dropdown menu with options like 'None' and 'All'. A checked checkbox next to 'Alaph' indicates it is the selected letter. To the right is a list of Syriac letters with their names in English: Alaph, Alaph (Angular), Alaph (Final), Alaph (Round), Ayin, Ayin (Final, closed), Ayin (Final, if different), Ayin (Final, open), Beth, Cohen Area, Dalath, Dalath (Angular), Dalath (Round), Gamal, He, He (Angular), He (Round), Heth, Kapth, Lamad, Lamad (Final), Lamad (Final, closed), Lamad (Final, if different), Lamad (Final, open), Mim. Below this is a 'SAVE' button.

The main area shows a photograph of a manuscript page with several green selection boxes highlighting specific letters. Below the image is a table with four rows, each representing a different letter example:

M-add.12150.Alaph.(Angular) [3], 35v/71 (606,1871)	Alaph (Angular)	608	1871	35	71		Currently: https://images.syriac.reclaim.hosting/binary/BL.Add.12150.196v_Alaph_(Angular).1871_606_71_35.png Change: https://images.syriac.reclaim.hosting/binary/BL.Add.12150.196v_
M-add.12150.Alaph.(Angular) [3], 53v/48 (606,1288)	Alaph (Angular)	661	1288	53	60		Currently: https://images.syriac.reclaim.hosting/binary/BL.Add.12150.196v_Alaph_(Angular).1288_661_60_53.png Change: https://images.syriac.reclaim.hosting/binary/BL.Add.12150.196v_
M-add.12150.Alaph.(Angular) [3], 52v/48 (606,1510)	Alaph (Angular)	1510	406	52	60		Currently: https://images.syriac.reclaim.hosting/binary/BL.Add.12150.196v_Alaph_(Angular).1510_406_52_60.png Change: https://images.syriac.reclaim.hosting/binary/BL.Add.12150.196v_
M-add.12150.Alaph.(Angular) [3], 68v/44 (606,288)	Alaph (Angular)	1863	288	68	66		Currently: https://images.syriac.reclaim.hosting/binary/BL.Add.12150.196v_Alaph_(Angular).288_1863_68_66.png Change: https://images.syriac.reclaim.hosting/binary/BL.Add.12150.196v_

Figure 1. The DASH Interface Back-End: Letter Identifier (above) and Letter Database (below). A back-end interface accessible to team members allows research assistants to rapidly identify multiple examples of Syriac letters. After the research assistant uses a selection box to identify a given letter, the letter coordinates and associated metadata are stored in a mysql database (below). © The British Library Board.

THE FRONT-END INTERFACE: VISUALIZING SECURELY DATED MANUSCRIPT AND LETTER IMAGES

As part of a larger digital paleography endeavor, our team collected letter data from over a thousand Syriac codices. For the publicly facing side of the project, however, we prioritized images coming from those manuscripts securely dated before

the twelfth century. Even this smaller sample is a large set of visual data and raised the question how best to display these images for researchers. We designed DASH primarily for scholars of Syriac. However, we also wanted to create an interface that could serve as a model for digital paleography in other linguistic traditions. As a result, when building the front-end interface we kept several design goals in mind.

When a human expert examines a physical manuscript, they do so at different levels. For example, at times a paleographer concentrates on the level of one or more complete manuscript pages. At other times they go to the level of a single example of a given letter. Often, they focus on intermediate levels, such as how two individual letters connect to each other or the slight differences between multiple samples of a scribe writing the same letter form. Taking this as our model, we desired an interface that would allow the digital scholar to relatively seamlessly navigate between various levels of analysis ranging from the macro (e.g. looking at multiple manuscript pages), to the micro (e.g. a single letter image), to several levels in between. We also wanted to give the researcher control over the complexity of the visualization, enabling them to at first view particularly easy-to-see images and then to add greater context, or to move in the other direction from complexity to simplicity. In data visualization research, this paradigm is known as the Shneiderman Mantra: “Overview First / Zoom & Filter / Details-on-Demand.”⁹

Our interface design also emphasized the comparative. At whatever level the scholar was viewing the data, they should be able to compare what they were seeing with other data. That is, if they were viewing a specific manuscript page, we wanted the scholar to be able to simultaneously see pages from other

⁹ Ben Shneiderman, “The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations,” *Proceedings 1996 IEEE Symposium on Visual Languages* (IEEE, 1996).

manuscripts. If viewing an individual letter form, we wanted the scholar to be able to view multiple examples of that letter or to view letters from different manuscripts next to each other. So, too, the interface should facilitate explorations of how a given letter or set of letters developed over time.

Finally, we wanted an interface that was easy-to-use. It was to have an intuitive, compact design. It ought to work not only on desktop machines and laptops but also on tablets and phones. It needed to have basic help features and to follow best principals of accessibility. It should have options allowing one to easily share their discoveries with other researchers or embed them in a publication.

At times these goals competed with each other. For example, keeping the interface simple meant limiting some of the display options. At other times, budget constraints meant the interface still has areas that could be expanded. As one example, currently the interface provides very little metadata about a given manuscript and the scholar instead has to consult the standard manuscript catalogs. In the future we would like to link our project to others such as syriaca.org's forthcoming digital version of William Wright's catalog of the British Library's Syriac manuscripts. Most of all, we want to revise the interface in accord with suggestions scholars have sent us using the DASH website's contact form. Despite such limitations, it still felt appropriate to publicize the project, both to enable researchers to begin using this on-line tool and to solicit additional recommendations for its improvement.

A TOUR OF THE DIGITAL ANALYSIS OF SYRIAC HANDWRITING PROJECT

The most effective way to explore DASH is, of course, on one's own. But a quick tour limited to still images may help motivate direct exploration of the site. Typing DASH.Stanford.edu into a browser brings up the landing page. This provides basic

information about the project, a list of project publications, a user guide, and—most importantly—the means to provide feedback about the project. Most, however, will begin simply by clicking on the big “Get Started” button.

Now one sees the publicly accessible interface with options on a collapsible palette to the left and a tab on the right allowing one to move between the view at the level of manuscript page or at the level of individual letters. As a small case study, one can explore *British Library Additional 12,149* which is a liturgical manuscript copied in 1006 CE by the scribe Yeshua son of Andrew. After selecting the manuscript shelfmark on the left palette and choosing to view the manuscript from the right tab, *British Library Additional 12,149* opens in a Stanford developed manuscript viewer called Mirador (*Figure 2*). Here one can toggle to full screen, zoom in and out of the manuscript image, move around the manuscript page by simply dragging or using control arrows. One can also open up a series of basic image tools to adjust brightness, contrast, or color saturation, switch to greyscale, invert the color, or revert to the original image. If a library has publicly accessible images in the IIIF format that is becoming the worldwide standard for manuscript display (e.g. most Vatican Library manuscripts), one can see all its pages. For the remainder, we have permissions to display a few pages; that is one cannot read the book cover to cover but will certainly have enough pages for paleographic analysis.

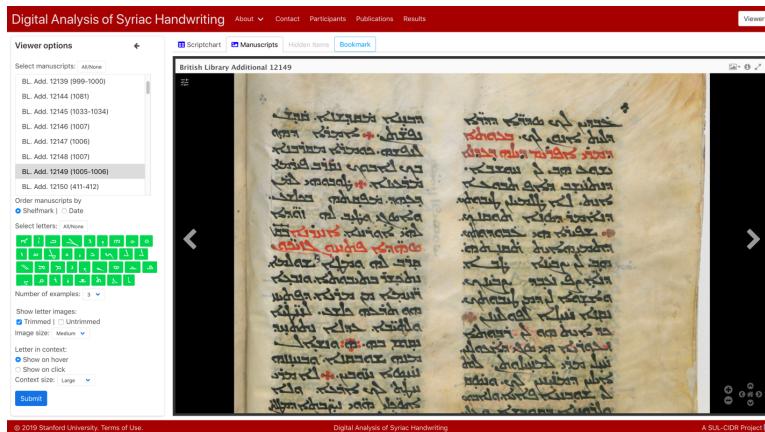


Figure 2. British Library Additional 12,149 in the MIRADOR Manuscript Viewer. Scholars can view each of the project's 156 manuscripts securely dated to before 1101 CE in a Stanford designed manuscript viewer. This enables them to zoom in and out, to flip between manuscript pages, and to adjust brightness, contrast, and color. In addition to seeing the manuscript in the context of the larger interface, scholars can also choose to view the manuscript in a full-screen mode. © The British Library Board.

Up to this point, the project seems to be a glorified, digital version of Hatch's *Album*, albeit with almost 50% more manuscripts, multiple pages, better images, and the ability to zoom and modify the images. Nevertheless, the overarching focus on a manuscript page remains. But if one instead chooses the "script chart" tab the perspective quickly changes. As shown in *Figure 3*, the project instantly makes a custom designed script chart. Here one can specify which letters to show, how many examples, and the size of the resulting image. At first, one might view the trimmed, binarized images as these make for a very clean chart. But, at the same time, this very simplicity excludes information. As shown in *Figure 4*, to help regain information one can also include the original letter images prior to trimming and binarization. This gives more detail and is still fairly clean. But it does not yet have much context. But, as *Figure 5*

illustrates, hovering over any letter in the chart allows one to see that same letter example, clearly marked in fluorescent green. But now the letter includes the letters and lines that originally surrounded that letter. If one wants more context, one can always go back to viewing the entire manuscript page by clicking on the manuscripts tab.

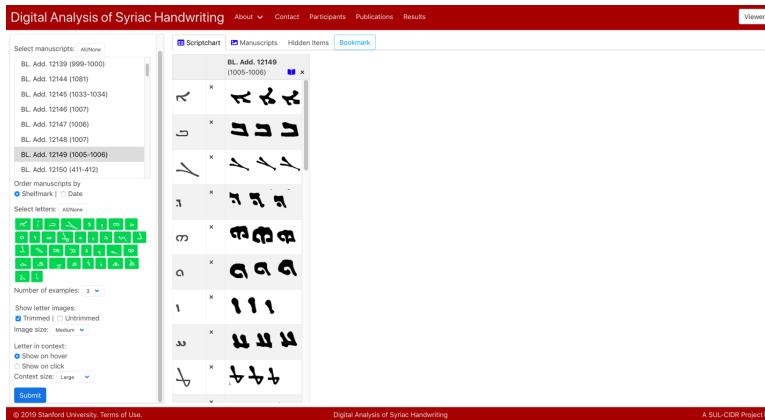


Figure 3. Script Chart of British Library Additional 12,149. The project automatically generates custom designed script charts. Scholars control variables such as manuscripts, letters, the number of examples, and image size. They can also adjust charts “on the fly” by adding or deleting letter rows or manuscript columns. Scholars can also drag rows and columns to change their order. © The British Library Board.

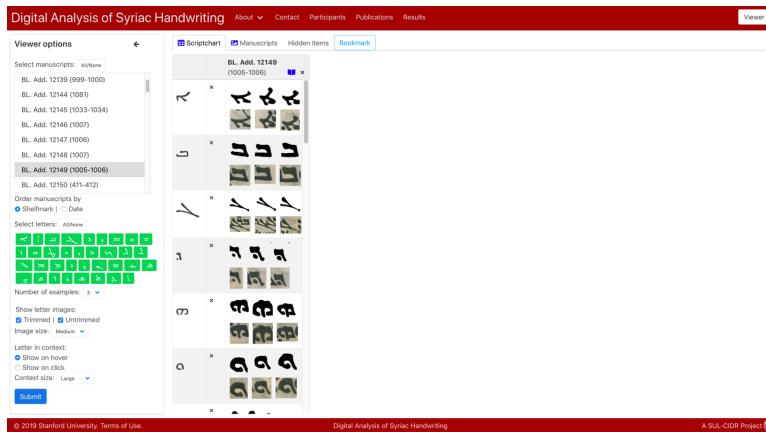


Figure 4. Script Chart of British Library Additional 12,149 Including Untrimmed Images. In addition to displaying binarized images, the interface can also show the letter images in the form they were when first identified by a research assistant. Although not as clean-cut as the binarized images, these untrimmed images are generally of higher resolution. They can also help the scholar verify finer details in a specific letter image and identify any artifacts created by the binarization process. © The British Library Board.

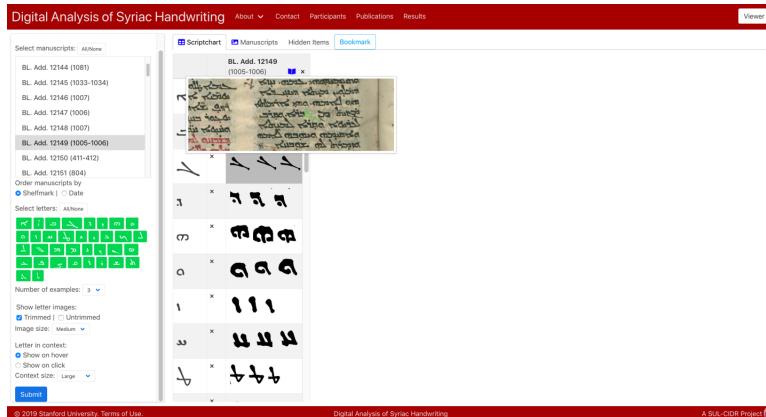


Figure 5. Script Chart of British Library Additional 12,149 Showing a Letter in Context. Although a collection of single letters creates a very clear script chart, it also eliminates many of the details that are most important for paleographers. Hovering the cursor over any letter in the

script chart shows that specific letter example (which the interface highlights in green) in a larger context. The scholar can help specify how large an area surrounding the letter is displayed. This allows researchers to investigate traits, such as how letters connect to each other, which would be hidden in a single-letter script chart.

It turns out that a year before Yeshua wrote *British Library Additional 12,149* he copied *British Library Additional 12,148*. This provides a rare opportunity to examine how similar a given scribe's handwriting is across manuscripts. *Figure 6* shows what happens when one adds *British Library Additional 12,148* to the display options. Now Yeshua bar Andrew's two works appear in the manuscript viewer next to each other. If one returns to the script chart, it now shows letter examples from both manuscripts, allowing one to see how similar the individual letter forms actually are. But it turns out that in the early eleventh century Joshua was actually quite busy and he copied two additional manuscripts. By selecting all four shelf-marks one can see, as in *Figure 7*, all four of Yeshua's manuscripts at one time. A return to the script chart (*Figure 8*) shows how similar all of these letters are to each other.

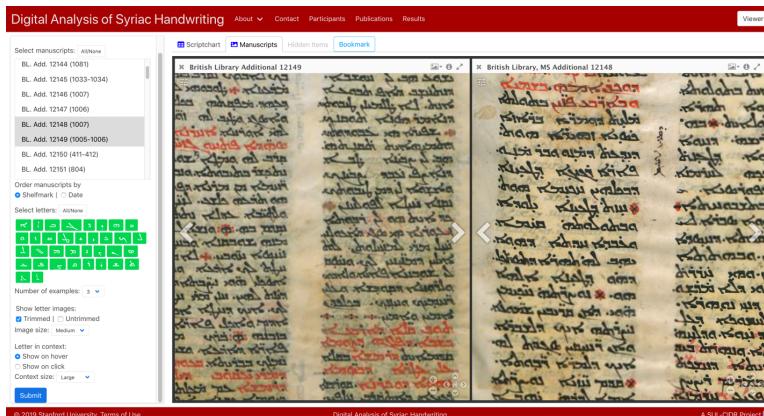


Figure 6. British Library Additional 12,148 and 12,149 in the MIRADOR Manuscript Viewer. Scholars can simultaneously view multiple manuscripts. *British Library Additional 12,148* and *12,149* were both written by an early eleventh-century scribe named Yeshua, son of Andrew. Displaying these two manuscripts next to each other illustrates how little Yeshua's handwriting varied across documents. © The British Library Board.



Figure 7. British Library Additional 12,146, 12,147, 12,148, and 12,149 in the MIRADOR Manuscript Viewer. The front-end interface allows scholars to view up to four manuscripts simultaneously. In this case, all four were written by the same scribe. © The British Library Board.

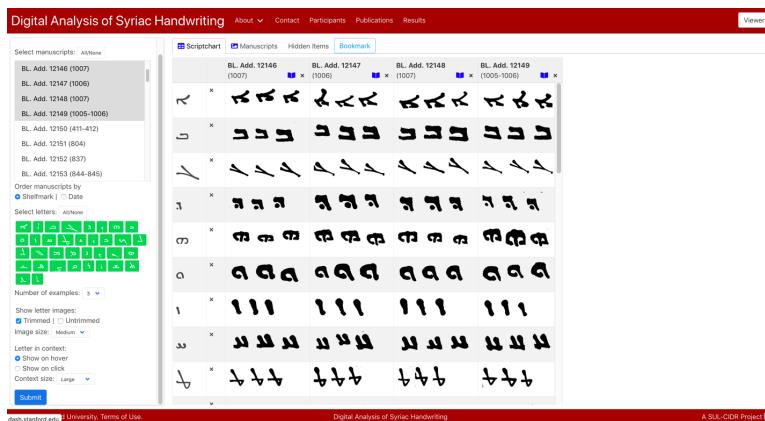


Figure 8. Script Chart of British Library Additional 12,146, 12,147, 12,148, and 12,149. Charts of letters from multiple manuscripts allow for easy comparison. In this case, all four manuscripts come from Yeshua son of Andrew and show how little the scribe varied his handwriting between documents. Alternatively, one can choose a given letter or series of letters and display all securely dated examples of that form chronologically, essentially creating a timeline for the development of Syriac script. © The British Library Board.

DASH also helps trace the chronological development of Syriac script. Instead of specifying a single or small group of manuscripts, one can instead choose a single letter and view how that letter form changes across all securely dated manuscripts. Alternatively, one can view the development of multiple letter forms in relationship to each other. As with other script charts, chronological charts such as these retain the option of viewing any given letter example in context or toggling to a full manuscript page.

If one wants to share such discoveries, a quick click pastes a URL onto the clipboard that can be sent to a colleague by e-mail or even by a text message. With a single click they will see the exact same visualization. Alternatively, one can paste the URL into an article footnote (or use a Tiny URL to shrink it). If in an on-line article or a PDF, the article's reader can click on the note

and it will immediately open exactly the same visualization on the DASH site or a print reader can type the URL into a browser to access the visualization.

...

In terms of Syriac Studies, DASH has had two overarching goals. First, the project provides access to a much greater number of securely dated manuscript pages than previously available. In this sense, it is an expansion of Hatch's *Album of Dated Syriac Manuscripts*. The second, field-specific goal can also be seen as an extension of Hatch's book. Like Hatch's work almost seventy-five years ago, DASH is designed to support other scholars as they develop and implement their own studies of specific manuscripts or of Syriac script more generally.

The project, however, also aspires to be of help outside of the field of Syriac Studies. It uses Syriac manuscripts as an extended case study for exploring how a wide range of visualization techniques might assist manuscript studies as a whole. DASH thus illustrates a conceptual model for thinking beyond the tradition of a printed manuscript album as the primary paleographic resource. As an open source project, it also provides code that those in other subfields can adopt and adapt for their own digital endeavors.

Scholars of Syriac are lucky to have a relatively large corpus of very early, securely dated manuscripts. This not only facilitates the creation of paleographic tools, it also makes their development that much more imperative. We certainly would like the Digital Analysis of Syriac Handwriting project to be of use to individual scholars and their research. Our greatest hope, however, is that both for those who focus on Syriac manuscripts and for those who work in other linguistic traditions, it might inspire additional projects in computer assisted paleography.