JDH

Journal of Digital Humanities

Vol. 3, No. 2 Summer 2014

ISSN 2165-6673

ROY ROSENZWEIG Center FOR History AND New Media

A PRESSFORWARD PUBLICATION

Table of Contents

Features

Focus

Review

Digital Contexts

The Editors

The digital contexts of our scholarly practice impact not only the kind of work that we may do as humanists, but also how we represent changes in theory and methods over time. Whether we are preserving, analyzing, or representing cultural heritage collections, interpreting digital media, or communicating through open repositories or social media, our activities are doubly informed by digital modes of production and digital professional practice. Every time we participate in a conference panel that others tweet or blog about, deposit our pre-print article in an institutional repository, or even offer an online version of our course syllabus, the technical situation of our work as teachers, researchers, or students responds, knowingly or not, to a digital condition.

The articles in this ninth issue of the *Journal of Digital Humanities* consider ways in which digital contexts challenge scholarly practice, from the creation or engagement with digital source materials to new methods for sharing results, interpretations, or ideas. In particular, the authors grapple with reconciling the theories and values of one's discipline with today's shifting digital landscape.

What opportunities, as well as risks, do remediations of informal conference conversations through shared Google Docs or Twitter present? Bethany Nowviskie's "On the Origin of 'Hack' and 'Yack'" recuperates an anecdotal history of the phrase "hack vs. yack," which is often used as a shorthand for opposing approaches to professional practice. Nowviskie resituates the phrase "more hack; less yack" within the first professional context in which it was deployed: to eschew the staid roles of active speakers and passive audiences and to foster active, non-hierarchical engagement among participants at THATCamp Prime in 2009. Prompting us to consider how recirculation of a phrase like "hack vs. yack" over social media divorces it from its necessary context, Nowviskie concludes, "… to pretend or believe that 'more hack; less yack" represents *a fundamental opposition in thinking* between humanities theorists and deliberately anti-theortical digital humanities 'builders' is to ignore the specific history and different resonances of the phrase, and to fall into precisely the sort of zero-sum logic it seems to imply."

This issue's focus section features papers by Katharina Hering, Michael J. Kramer, Kate Theimer, and Joshua Sternfeld originally presented at the 2014 American Historical Association annual conference. In "Digital Historiography and the Archives," the authors explain that as prompts for lively engagement with fellow panel participants and the audience, the papers are a representation of, but not a substitute for, the roundtable itself. Likewise, they that argue professional theory and practice are changing due to the current digital contexts of historical and archival work. Katharina Hering's "Provenance Meets Source Criticism" considers an increasingly important ethic of attaching source criticism and provenance to a digital object's record, while Kate Theimer in "A Distinction Worth Exploring: 'Archives' and 'Digital Historical Representations'" discusses the salient distinctions between the digital practices of archivists and historians. Joshua Sternfeld contributes to the conversation in "Historical Understanding in the Quantum Age" by insisting that big data methods hold the potential to resituate digital collections within a much broader empirical context. Finally, Michael Kramer's "Going Meta on Metadata" responds to his co-authors by contemplating the ways in which the increasingly digital landscape occasions a possible flip between historical roles and professional practices.

In "Using Computer Vision to Increase the Research Potential of Photo Archives," John Resig presents a case study that combines TinEye's Match Engine with custom tools to perform image analysis of Italian anonymous art in the Frick Photoarchive. In the context of a digital collection, Resig responds to the challenges multiple, duplicate copies present to archivists and collectors by presenting a case study and custom toolkit that allows researchers to compare large collections of images and identify new relationships among items in the collection. These items include images before and after conservation, copies of the same artwork, detail shots of the same images, and cataloging errors. A key study for digital collections management, Resig’s contribution exemplifies in practice one way in which digital humanists are responding to questions raised by Hering, Kramer, Theimer, and Sternfeld about scale, provenance, and practice.

Finally, Alex Christie explores the challenges of theorizing entwined scholarly fields in his review of *The Johns Hopkins Guide to Digital Media* by editors Marie-Laure Ryan, Lori Emerson, and Benjamin J. Robertson. He argues that the book offers a touchstone for both experienced and new audiences interested in digital media studies now and in the future.

What responsibility do we have to situate our representations of informal and formal scholarly conversation within its original context? How does our/the digital situation inform how we engage in scholarly debate? As the works in this issue suggest, scholars have an ethical, as well as a scholarly, imperative to place our object of study–whether it’s the creation of an archive or a conference presentation or a Twitter conversation–within a context that represents the condition of its creation. The goal of the *Journal of Digital Humanities* and its sister publication *Digital Humanities Now* has been to retain a sense of each entry’s original context as we recirculate and preserve informal scholarly discourse. By redirecting readers of *Digital Humanities Now* back to each entry’s original site of publication, and by citing the origin of each entry in the *Journal*, we strive to contextualize current, informal scholarship that can inform future discourse and research.

Joan Fragaszy Troyano and Lisa M. Rhody, Editors

Using Computer Vision to Increase the Research Potential of Photo Archives

John Resig

In art history research, photographs of art are the lifeblood of study. Since it’s usually impossible for a scholar to travel the globe and visit an artwork as need arises, there is substantial demand for archives of photographs of artworks for reference and study. There are photo archives around the world with millions of photographs in them, including the prestigious Frick Photoarchive. These archives aggregate photographs from many institutions and private collections. It is their job to make sure the photos are maintained and the works of art they document cataloged, changes in attribution or ownership updated, and that they have properly identified and merged duplicate photographs and entries relating to a single artwork.

This process of finding duplicate artworks can be breathtakingly time-consuming. Many professional researchers spend years correcting and merging entries in even a moderately-sized archive. For an archive with over a million photographs, that process becomes impossible. This says nothing of the difficulty of sharing images between institutions where cataloging standards or metadata may differ drastically.

Image similarity analysis is an exciting computer vision technique for matching photos whose image content is substantially or completely similar. Through image similarity analysis, it is highly likely images depicting the same object will be found and matched.

The application of computer vision to art photo archives has largely been unexplored up to this point. Lev Manovich has explored ways of analyzing images of artworks while looking for trends in an artists oeuvre or entire artistic movements. However, most institutions have used large scale image analysis primarily for cases of copyright enforcement, face detection, or color/composition analysis.

To explore what image similarity analysis was capable of, I completed an analysis of the digital images of Italian anonymous art at the Frick Photoarchive. The image similarity analysis, using TinEye’s MatchEngine service, was automated using newly-developed tools. I further processed and dissected the data using custom tools. The analysis was able to confirm some of the existing relationships between photographs that were manually generated by researchers. The analysis was also able to discover a number of completely new relationships, including: works of art before and after conservation, copies of the same artwork, cropped detail shots of the same artwork, and cataloging errors.

The custom toolkit developed to analyze the Italian anonymous archive will be publicly released as a generic image similarity analysis tool. Comparable results could be easily achieved by other institutions for a minimal cost using these tools.

The results of the image similarity analysis of a photo archive are extremely exciting and could completely change how the process of cataloging images is completed. It could also make some impossible tasks, such as merging multimillion image archives, a reality.

The Frick Photoarchive

Started in 1920, The Frick Photoarchive has continually expanded over nearly a century and now contains over 1.2 million photographs of works of art. In addition to sponsoring original photography of art around

the world, the Frick has benefited from photograph donations from both institutions and scholars. To this day the library still actively purchases photographs.

The Frick Art Reference Library recently contacted me when they saw image analysis work that I was doing with my Ukiyo-e.org: Japanese Woodblock Print Search and Database project (which deserves a separate essay). They were curious if image analysis could work for photographs of paintings, three-dimensional artworks (instead of prints), and their collections in particular. Additionally, they were interested in where image analysis could aide in the process of merging multiple photo archives

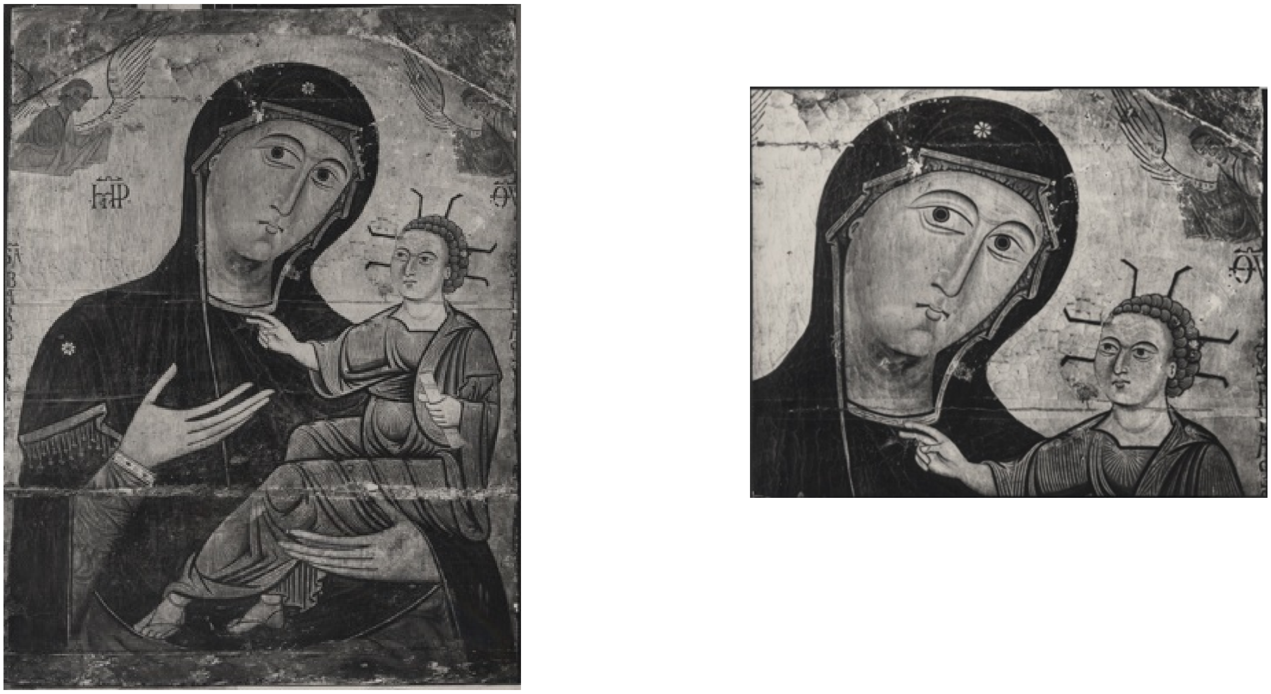
The Frick Library is a member of the newly-formed International Digital Photo Archive Initiative, a consortium of fourteen photo archives from Europe and the United States with an aggregate 31 million photos of art. Nearly all of these institutions are in the process of digitizing their photo archives. They see the tremendous power of sharing photos and photo metadata amongst institutions: the aggregated information can yield a better understanding of the artworks (works before and after conservation, works that have been stolen or are missing can be revealed, and provenance and general scholarship can be accelerated).

The Frick Library is still early on in the digitization of their collection. Thus far, they've digitized about 70,000 photographs. Their in-house digitization lab has just recently been set up and will allow for a far greater volume of photos and increases in metadata quality. They've also received grants to digitize their collection of 57,000 original negatives of artworks, most of which is already available online in the Frick Digital Image Archive.

Frick Italian Anonymous Digital Archive

The first digitization project undertaken by the Frick Photoarchive, sponsored by the Pernigotti S.p.A., Averna Group in Milan, was to digitize 18,548 photographic reproductions of 14,284 works of anonymous Italian art and turn it in to a digital photo archive. This photo archive is made available to researchers through the Frick Digital Image Archive. The digitization was undertaken by an outside lab long before the Frick Photoarchive had its in-house digitization lab set up.

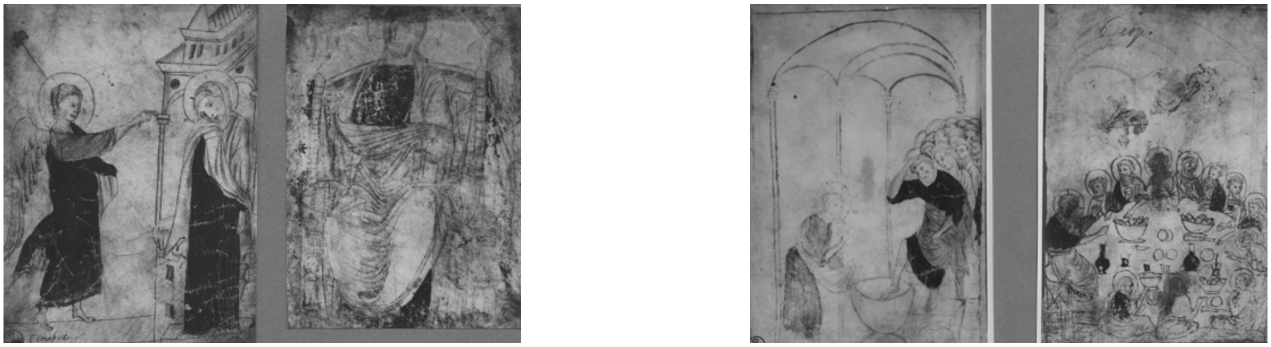
The artworks represented in the Italian anonymous archive are largely from around the time of the Renaissance and are either unattributed or considered to be anonymous. The archive is not limited to just two-dimensional paintings, but also includes frescos, drawings, prints, and sculpture. A representative example of the artworks and photos in the archive is shown below:



Madonna and Child, 13th century, La Chiesa di S. Eufrasia, Pisa.

In the case of this artwork, there are two separate photos representing the same piece: the full panel and a close-up detail shot. Note that the photos are in black-and-white: this is the case for nearly all the photos in this particular archive.

In the Italian anonymous digital archive, the photos are generally organized into groups with all photos from the same work of art clustered together under a single number (for example `10383a.jpg`, `10383b.jpg`, `10383c.jpg`, etc.). This clustering was done manually by the original digitization team using metadata associated with the photos in the archive. However just because the photos are of the same work of art does not guarantee that they'll be depict an image that is identifiably the same work of art. For example the following two photos depict different portions of the same work of art with no overlapping imagery:



Florentine, 13th century, Uffizi Museum in Florence

The Italian anonymous archive poses particular challenges to researchers at the Frick Photoarchive. Most of the photos in the photo archive are organized by attributed artist, making it easy to find duplicate, or alternate, photos of the same work of art. The fact that none of the works in this particular archive are attributed makes it extremely hard to guarantee that every alternate photo of an artwork will be grouped together.

Correcting Merged Photo Archives with Metadata

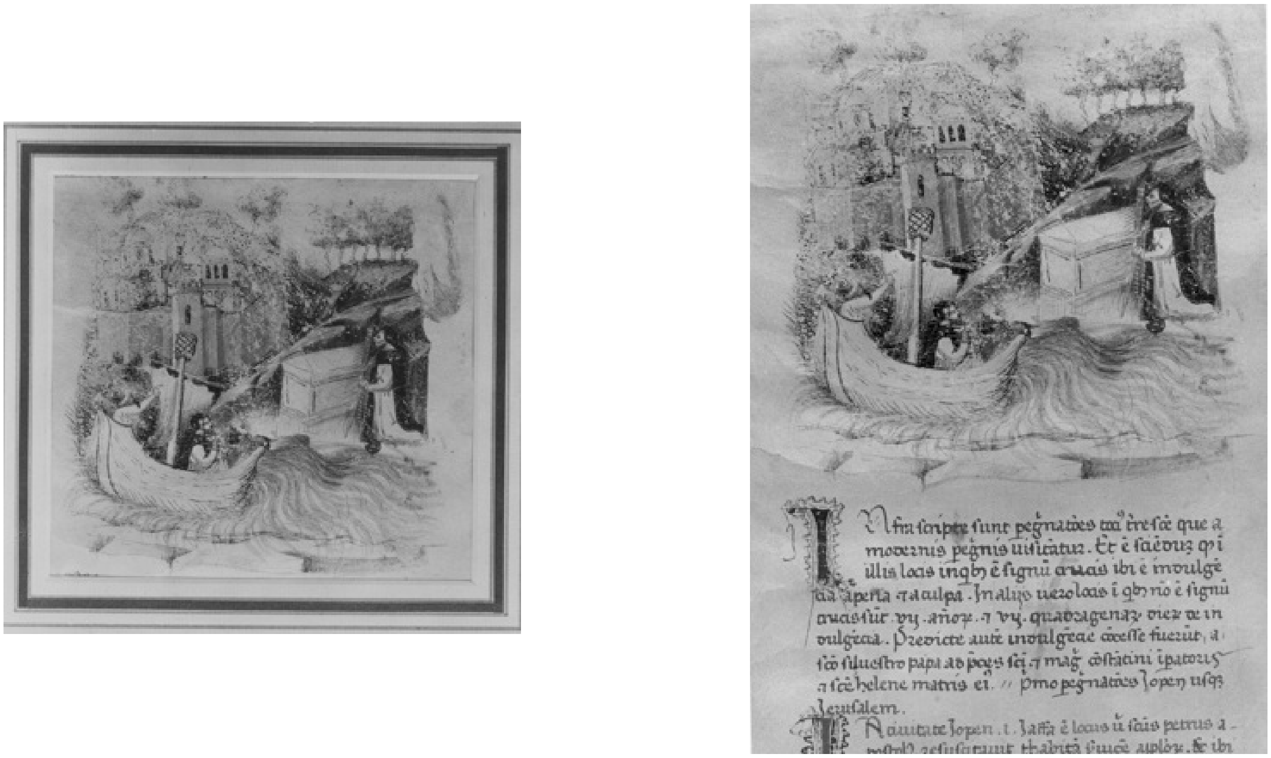
Interestingly, the problem of grouping related art photographs is actually quite similar to the problem of grouping images across multiple major (and sometimes international) photo archives. If one were given two sets of images, each with thousands (or millions!) of images in them, it would be physically impractical for humans to go through all of the entries for a particular artist and cluster every identical work of art. When faced with a problem of this magnitude the smart thing to do would be to turn to the metadata associated with the images to support the merging.

To appropriate a famous quote from the programmer Jamie Zawinski:

> Some people, when confronted with a problem, think
> “I know, I’ll use metadata.” Now they have two problems.

In theory good metadata attached to records should be able to solve most problems that come with merging or correcting problems in a collection (or between collections). However, in practice, it’s very likely that institutions will have varying interpretations of quality, make mistakes in cataloging, and make mistakes in data entry. When merging multiple collections whose metadata is written in different languages or between collections that are missing critically important metadata (as is the case with the Italian anonymous archive’s missing artist names), the challenge becomes even more difficult.

This is where the effectiveness of computer vision and using image analysis to correct archives becomes crucial. Accurately matching two images that have identical visual characteristics in two different collections can reveal missing or mistaken data. As a representative example two images found to be similar through the analysis are shown below: one is a photo from a Christie’s auction catalog dating to 1936 and the other is a photo from the Harvard Art Museum in Cambridge.



Tuscan, 15th century, Harvard Art Museum.

Naturally the artist is unknown in both of these cases, but it’s very possible to have found a match after the fact if the metadata was good enough. Unfortunately, for these two images that was not the case. For whatever reason, the Harvard Art Museum fails to mention that this piece came from an auction at Christie’s (or that the owner who donated it had purchased it at Christie’s). Given that there is no identifiable artist, title, or date of this piece, it thus makes it incredibly unlikely that a human would’ve been able to discover that these two images were of the same work of art.

Put simply: there is frequently not enough information for humans to intervene and make a connection between images in a scalable manner. Individual researchers can certainly hunt through photos that have been organized (hopefully correctly) by artist or national school and century and attempt to make the associations manually. However, this process is painstaking at best and does not work well across hundreds, or thousands, of artists and potentially millions of images.

If all metadata associated with an image is ignored, and only the contents of the image were analyzed, it becomes possible to find interesting image matches that were likely undiscoverable using raw human power. A computer vision image analysis algorithm that's capable of finding matches between images that have a set of identical content would be the perfect tool for performing the analysis. With such a tool any matches that occur would likely indicate that they are different images of the same artwork.

It's possible that some researchers may become skittish at the prospect of ignoring all the painstakingly-generated metadata that's been associated with their images (for the purpose of finding similar images, at least). However, it's important to note that images rarely lie. When they do, there's likely something interesting happening that would be a good area for further research (such as copies of the same work of art).

Image Similarity Analysis Implementations

Computer Science research into computer vision and image comparison techniques has been going on for decades. Research and implementation is finally at the point where image analysis can be performed against millions of images simultaneously (as can be seen in the services provided by Google, Yahoo, and Bing Image Search). The general availability of this technology however, has been mixed. There are some freely available, open source, tools such as imgSeek and libpuzzlea>, which bring rudimentary image comparison technology to a larger audience. There are also commercially-available tools that provide fast image analysis with a greater level of clarity, such as TinEye's MatchEngine.

Finding the right tool that would work for the print images that were collected from the various institutions was especially tricky. The features needed for an effective print image search are:

- The process of adding in a new image, and performing a search with an image, must be fast. (If searches and comparison are too slow it'll be too hard to use effectively.)
- The engine should be capable of scaling up to hundreds of thousands, if not millions, of images.
- The engine should be able to find exact matches (cases where an artwork is definitively contained within an image). Inexact matches tend to confuse the results and make the matches hard to discern.
- The engine must be able to ignore differences in color, even differences between a color photograph and a black-and-white photograph. (Many institutions provide images only in black-and-white. Comparing those images with color matches at other institutions would be very useful.)
- It must be possible for an image of an artwork detail (part of a larger artwork) to match an image of the complete artwork.
- Images that have watermarks or other invasive imagery should still be matched (and not only match other images that also have watermarks).

Initially, imgSeek was explored because it did direct image comparison, worked quickly, and was open source. However, there were many difficulties in its practical use. imgSeek only analyzes pieces of an image (the colors and where those colors are located in the image), which causes similarly-composed images to appear as matches, even though they may be entirely different. For example, an image of blue sky with green grass would match all images that were blue at the top and green at the bottom, rather than just images of sky and grass. Additionally, it's unable to effectively find images that are in black-and-white or match details to a complete image of an artwork.

The MatchEngine tool, while a commercial service, is much better suited for finding images that are exact matches of one another or even details embedded inside a larger image. In all of the testing, MatchEngine outperformed the imgSeek service in quality. MatchEngine was much better at finding exact matches, ignoring differences in color, and finding details inside images.[1]

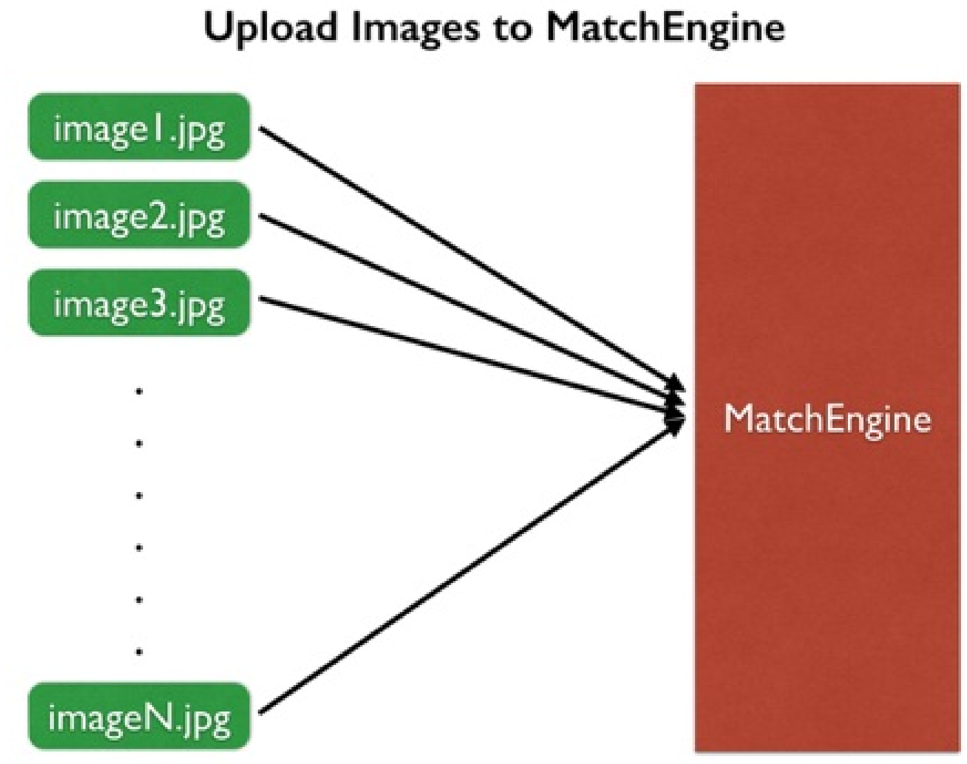
Implementation

With an image analysis utility in place, it is now possible to create a tool for automatically finding interesting new matches, correcting cataloging mistakes, and validating some of our existing matches.

The Frick Photoarchive provided an export of the 18,548 images in the Italian anonymous archive. MatchEngine will automatically scale down any image that is over 300 pixels tall or wide. Thus, to simplify the transfer, the Frick Photoarchive reduced the size of all the images before passing them along. In total, the size of these images was about 2 Gigabytes. Additionally, the Frick Photoarchive provided a CSV dump of all of the metadata associated with the images.

A number of tools were developed to perform the image analysis, collect the data, and analyze the results of the analysis.

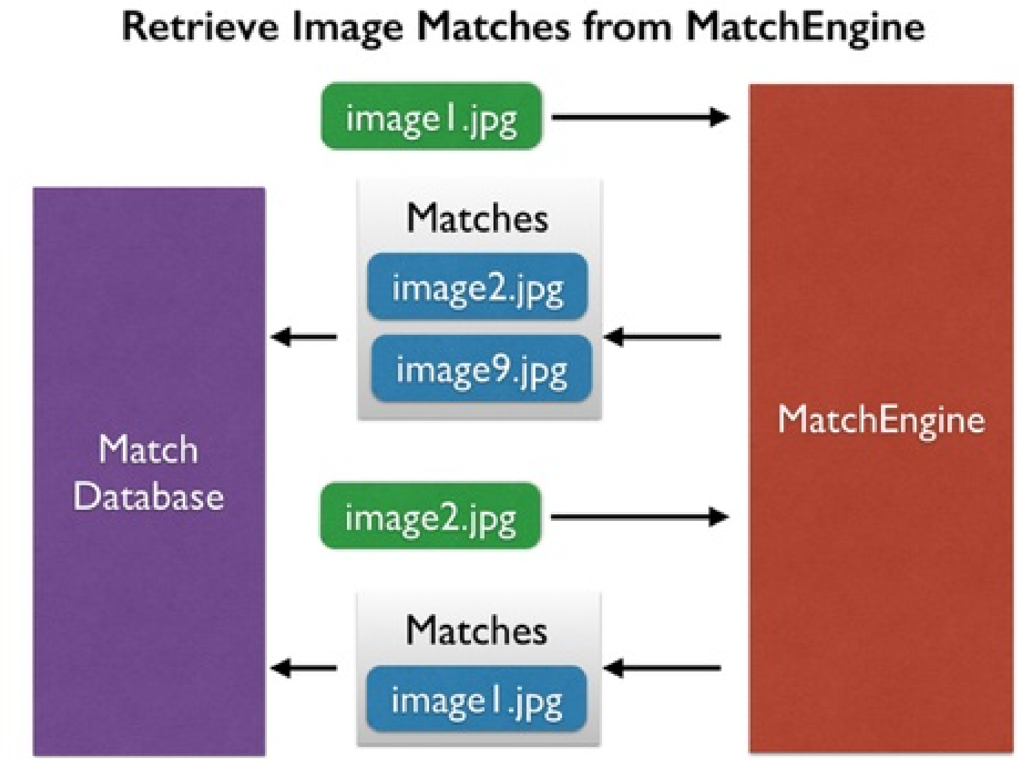
The first tool was a utility for uploading all of the images to the MatchEngine service through their private REST API.



The MatchEngine API supports uploading up to 1,000 images simultaneously. While the uploading is occurring, no other operations can be performed with the API. For the 18,548 Italian anonymous images, it took about 3 hours to complete over a standard home cable Internet connection.

Conventionally the MatchEngine service is used for two purposes: 1) providing a list of similar images for every uploaded image and 2) allowing a user to search images by uploading a photograph. Normally most users of MatchEngine keep images in the service over a long period of time to handle user search queries. For the analysis performed on the Italian anonymous archive, there was no need to keep the images in the MatchEngine service for any significant duration: only a bulk list of the similarities between the uploaded images was needed.

Another tool was then built to query MatchEngine for every previously-uploaded image to determine if any similar images had been found. MatchEngine’s indexing of the images was performed immediately upon upload and was made available for querying. Thus every single uploaded image could be queried and a full relationship graph could be downloaded.


Retrieve Image Matches from MatchEngine

The MatchEngine results for an image may look something like this:

```
&quot;frick-anon-italian/13291.jpg&quot;: [
    {
        &quot;score&quot;: &quot;27.80&quot;,
        &quot;target_overlap_percent&quot;: &quot;100.00&quot;,
        &quot;overlay&quot;: &quot;...&quot;,
        &quot;query_overlap_percent&quot;: &quot;47.18&quot;,
        &quot;filepath&quot;: &quot;frick-anon-italian/13291b.jpg&quot;
    },
    {
        &quot;score&quot;: &quot;12.50&quot;,
        &quot;target_overlap_percent&quot;: &quot;100.00&quot;,
        &quot;overlay&quot;: &quot;...&quot;,
        &quot;query_overlap_percent&quot;: &quot;20.93&quot;,
        &quot;filepath&quot;: &quot;frick-anon-italian/13291a.jpg&quot;
    }
]
```

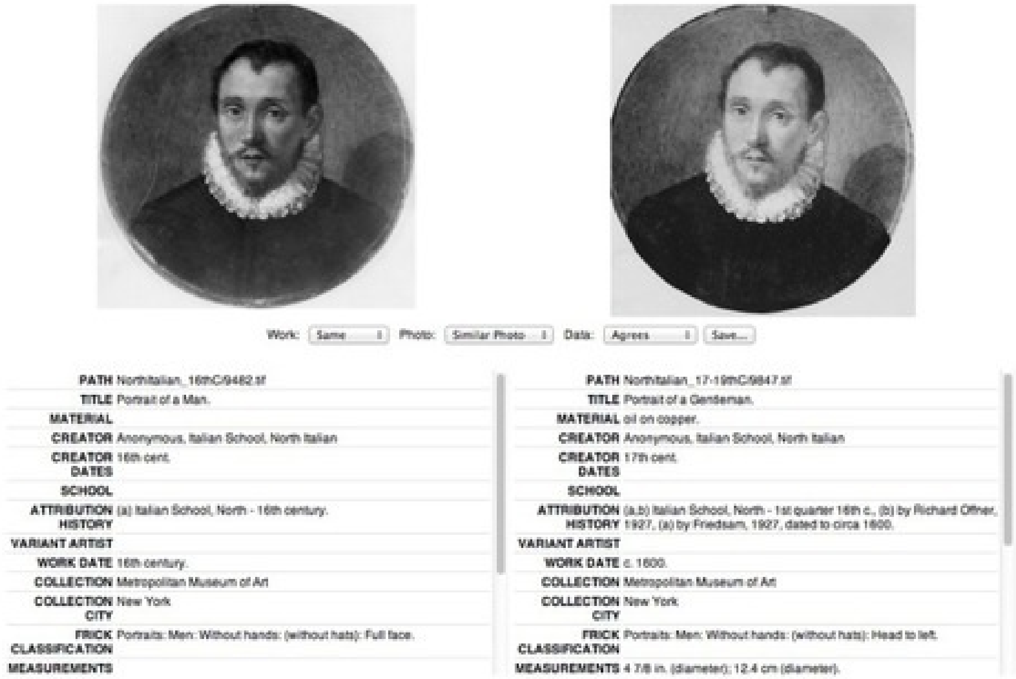
In this case, a query with the image `13291.jpg` received matches for the images `13291b.jpg` and `13291a.jpg` (I anticipated this result: all of these images were previously cataloged as being the same work of art depicted in alternate photographs or detail shots). The results show the “score” of the result, as specified by MatchEngine. The score represents how closely two images are deemed to be related. In practice, even very low-scoring images still appear to be the same work of art. MatchEngine also provides data regarding how much of the images were overlapping and provides some details on how to line up the images with one another; however, none of that is needed for this particular analysis.

The MatchEngine similarity data can be downloaded in parallel (using up to four simultaneous API connections). On a home cable Internet connection it took about an hour to retrieve all of the image similarity data for the entire Italian anonymous archive. All of the similarity data was then cached in a local JSON file for later retrieval. At this point the MatchEngine service was no longer needed or used. All of the images could then be deleted, using the API, from the MatchEngine servers.

Once all the image similarity matches have been downloaded to a local data store, the next step is to review all of the results and categorize the newly-matched results (this step is only performed for any previously unknown matches). The categorization of the matches isn't completely necessary: the matches could be passed off directly to researchers and catalogers instead. However, performing a basic organization of the results could help optimize researcher effort and focus attention on particular results or problem areas.

With a result categorization tool I was able to easily categorize all of the image matches. This could easily be achieved by other non-experts, or at least by people who have a basic familiarity with the subject matter being depicted in the images.

The categorization tool provides the user with a view of the two images that were matched by MatchEngine paired together with the raw data provided in the CSV data dump.


Work: Same Photo: Similar Photo Data: Agrees

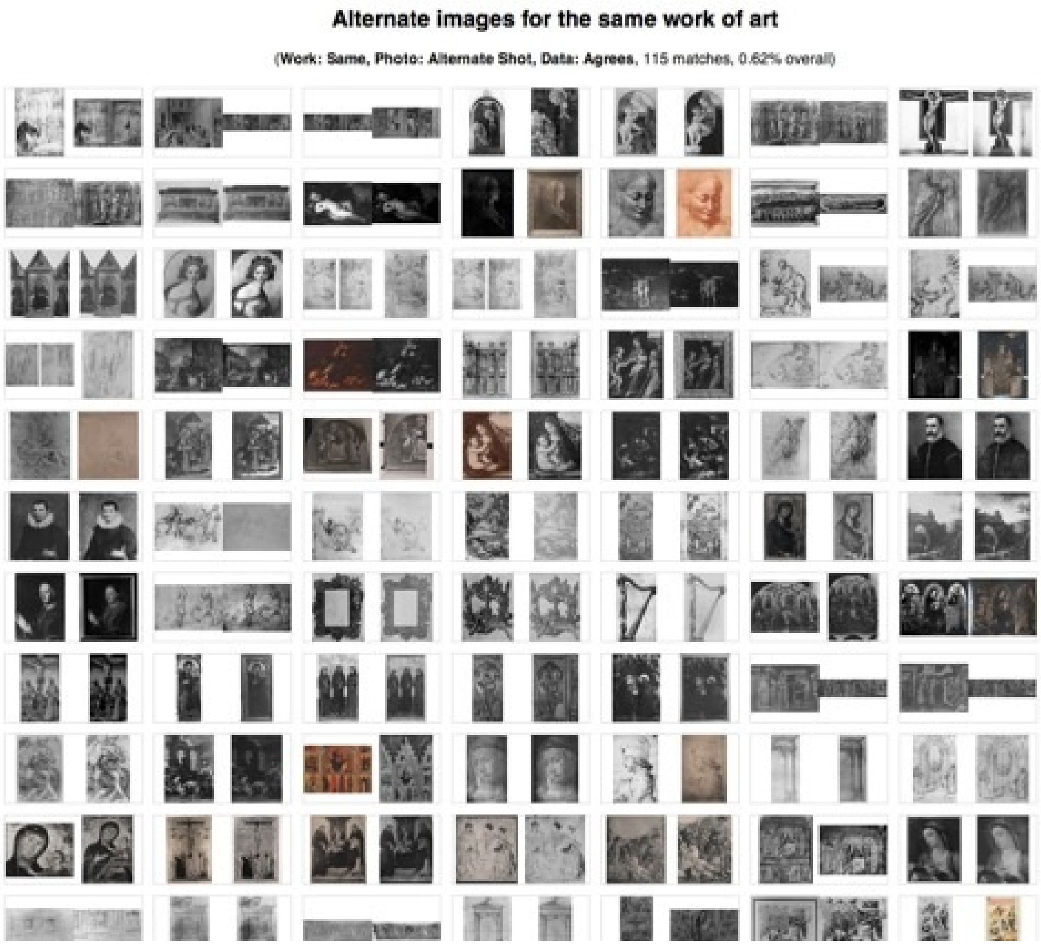
PATH	NorthItalian_16thC/9482.tif
TITLE	Portrait of a Man.
MATERIAL	
CREATOR	Anonymous, Italian School, North Italian
CREATOR DATES	16th cent.
SCHOOL	
ATTRIBUTION HISTORY	(a) Italian School, North - 16th century.
VARIANT ARTIST	
WORK DATE	16th century.
COLLECTION	Metropolitan Museum of Art
COLLECTION CITY	New York
FRICK CLASSIFICATION	Portraits: Men: Without hands: (without hats): Full face.
MEASUREMENTS	

PATH	NorthItalian_17-19thC/9847.tif
TITLE	Portrait of a Gentleman.
MATERIAL	oil on copper.
CREATOR	Anonymous, Italian School, North Italian
CREATOR DATES	17th cent.
SCHOOL	
ATTRIBUTION HISTORY	(a,b) Italian School, North - 1st quarter 16th c., (b) by Richard Offner, 1927, (a) by Friedsam, 1927, dated to circa 1600.
VARIANT ARTIST	
WORK DATE	c. 1600.
COLLECTION	Metropolitan Museum of Art
COLLECTION CITY	New York
FRICK CLASSIFICATION	Portraits: Men: Without hands: (without hats): Head to left.
MEASUREMENTS	4 7/8 in. (diameter); 12.4 cm (diameter).

This view gives a user, theoretically, everything that they need in order to determine what this newly-discovered match is and how these two images are related. The match was categorized on three axes:

1. **Work:** whether the artwork being depicted was the same work, a different work, or the same work but modified some how (e.g., before and after restoration).
2. **Photo:** whether the photograph was the same photo (100% identical), a similar photo (similar framing and composition with slight differences), or an alternate shot (such as a detail shot).
3. **Data:** whether the corresponding metadata of the two images agreed, disagreed, or was ambiguous. (When looking at the data it was only marked as 'agreed' if the data was obviously referring to the same artwork, typically held at the same institution.)

After I manually completed the categorization of all 446 matches between 815 images, the results were sorted into appropriate “bins” that denoted interesting trends.



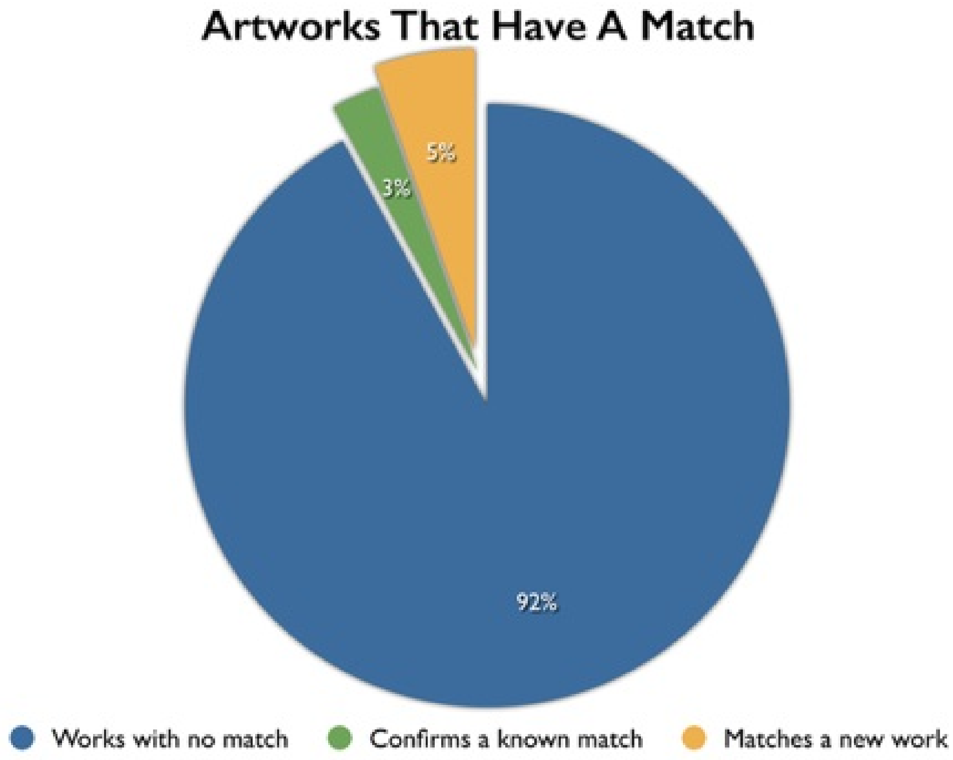
Alternate images for the same work of art

(**Work:** Same, **Photo:** Alternate Shot, **Data:** Agrees, 115 matches, 0.62% overall)

All of these binned matches were then passed on to researchers at the Frick Photoarchive for further analysis and record correction.

Results

The Italian anonymous photo archive was represented by 14,284 artworks. The image analysis found a match in 1,135 artworks (8%), including both newly-discovered matches and confirmations of existing relationships. Of those matched, 770 artworks (5%) had at least one new match with another distinct artwork, producing a total of 385 previously unknown inter-artwork relationships.



Out of the total 18,548 images, 1,187 images matched a known work of art and 446 new image pair matches were discovered. (An artwork can be represented by many individual images. In fact, one artwork alone had 152 photos associated with it.)

A complete examination of the image similarity analysis performed upon the Italian anonymous photo archive requires an understanding of three areas of results:

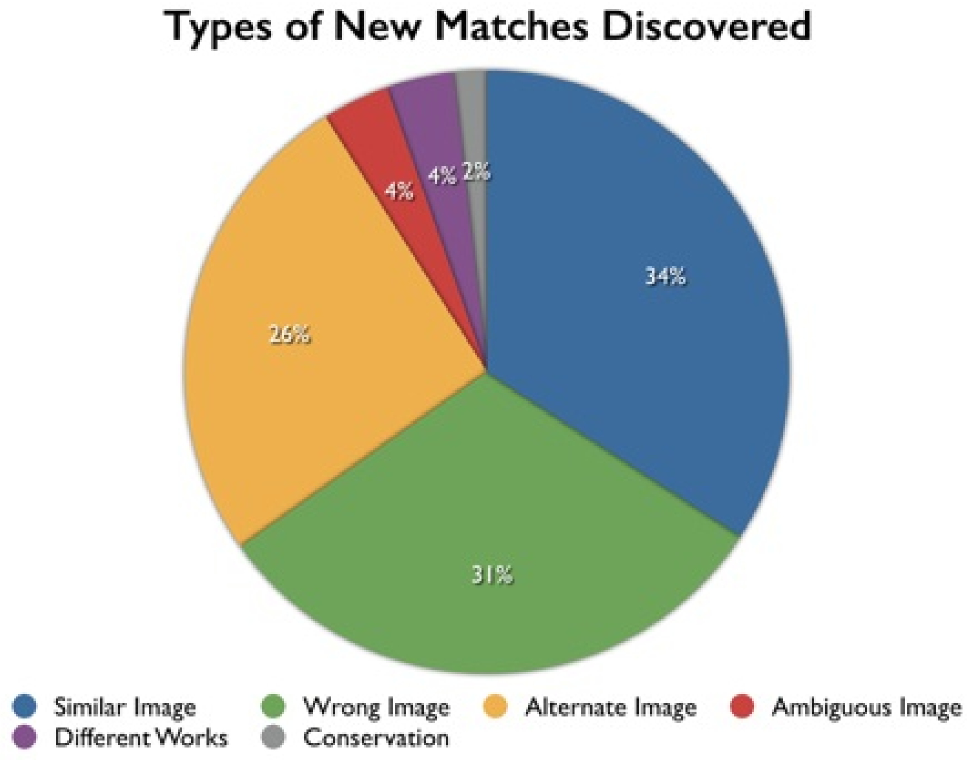
1. **New Matches:** completely new, previously un-cataloged, relationships between images discovered using the image similarity analysis.
2. **Confirmation of Known Matches:** confirming previously-cataloged relationships between images using the image similarity analysis.
3. **Unconfirmed Known Matches:** previously-cataloged relationships between images that the image similarity analysis failed to identify.

These studies were performed in order to look at all aspects of the image similarity analysis and determine what the analysis was capable of and what its limitations were. Learning that it was capable of confirming existing matches created by a researcher, as well as learning what matches it was unable to confirm, can help to set some expectations about how image similarity analysis can work for other photo archives.

New Matches

The new matches discovered by the image similarity analysis were certainly the most exciting for the researchers at the Frick Photoarchive. The analysis was able to accelerate their understanding and correction of the metadata associated with the digitized images.

The types of new matches broke down into a number of different areas:



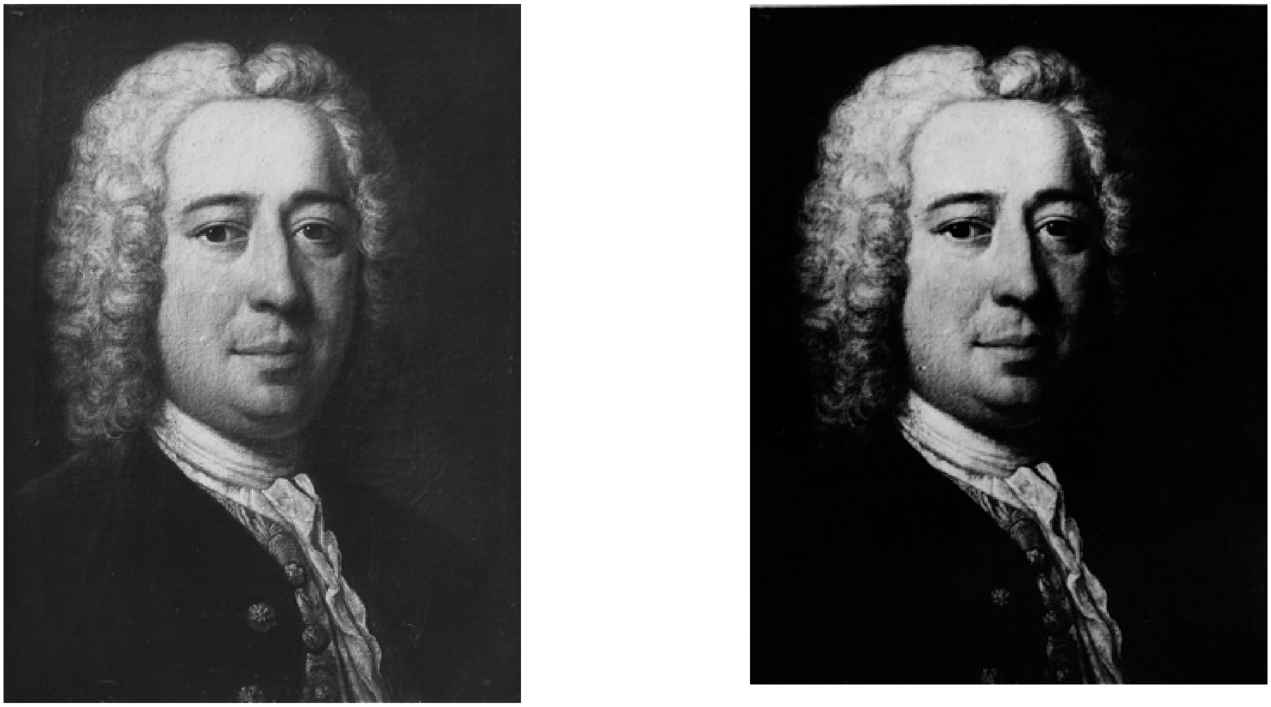
1. **Similar Images:** photographs that are highly similar (with the only differentiating factors being the difference in scan or lighting).
2. **Alternate Images:** matches where one photograph is an indirect, alternate, view of the same artwork (such as close-up of a detail or the same artwork viewed from an alternate angle).
3. **Conservation:** photographs of the same artwork most likely taken before and after conservation or during the process of conservation.
4. **Different Works:** photographs of two different artworks that are highly similar.li>
5. **Wrong Images:** the same, or similar, photograph but with the metadata in strong disagreement (likely resulting from a cataloging error).
6. **Ambiguous Images:** the same, or similar, photograph, but with ambiguous metadata (could be the same artwork but it's unclear).

The majority of the new matches (65%) were legitimate new discoveries previously missed by researchers. The remaining 35% of the matches were potential cataloging errors (most of which likely happened during the digitization process of the images).

Similar Images

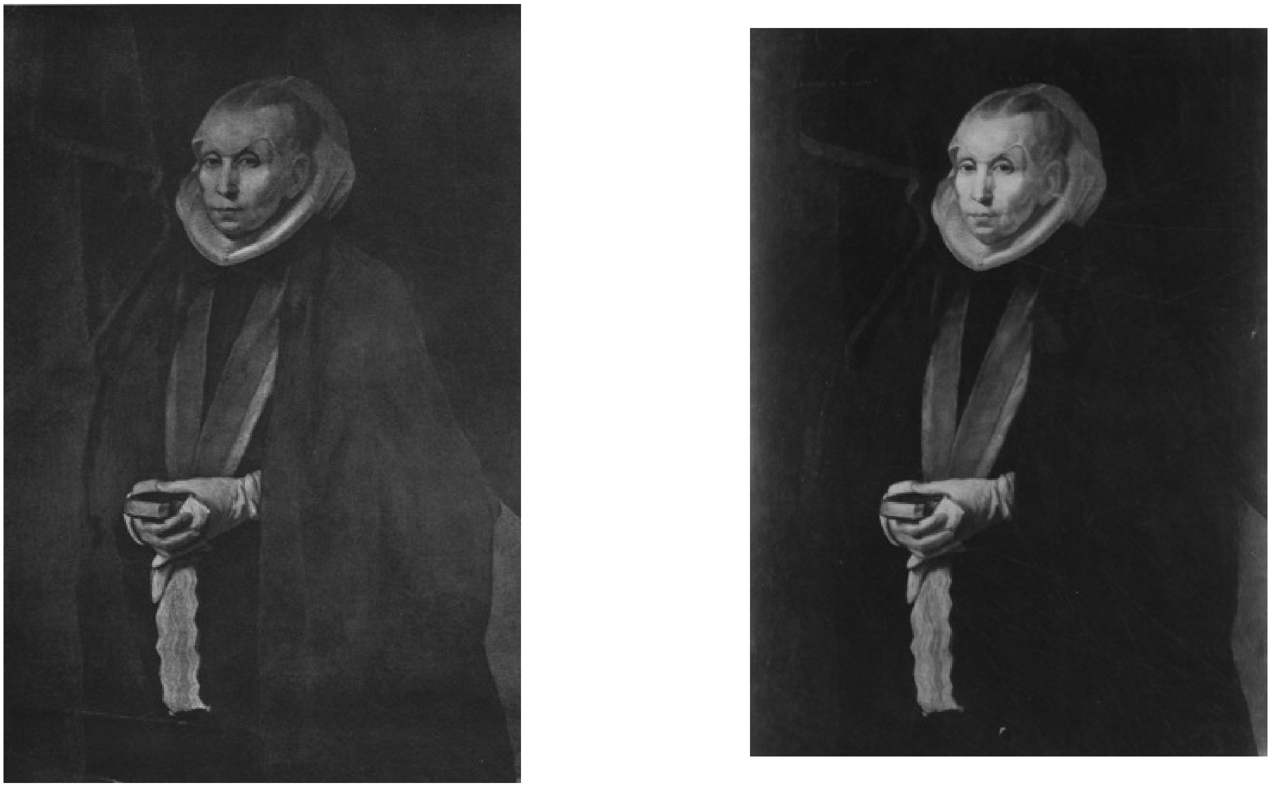
These are the same works that had a highly-similar photograph (of which there were 152 matches). This is the most obvious level of similarity: everything agrees (both the image and the data) in a very obvious way. Often times, these photographs would have similar cataloging details but were organized into different time periods or regions of Italy (thus making it more difficult for researchers to spot the discrepancy and correct it).

The first image shows the same work of art simply presented in two different, but similar, photographs. The only major difference is the lighting (obscuring a large portion of the painting). This was, by far, the most common type of similar image discovered through the analysis.



***New Match:** different lighting, same work of art.*

Another similar pair of images was discovered in which virtually everything agreed except for a critical piece of cataloging: one was cataloged as a full-length portrait of a man, the other as a portrait of a lady.

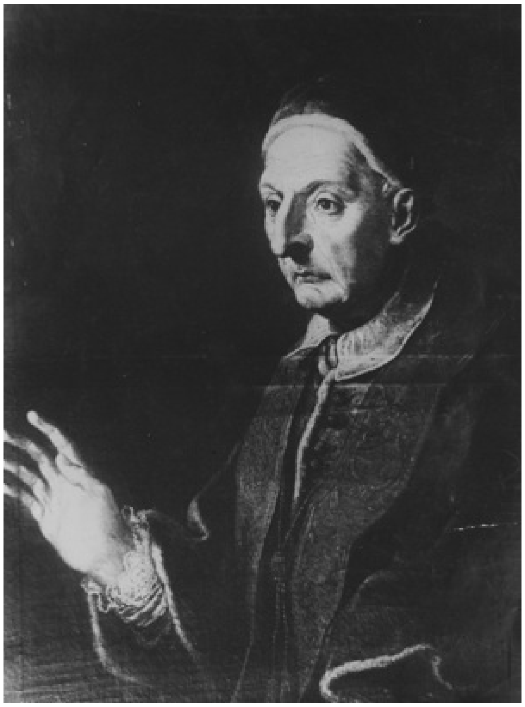
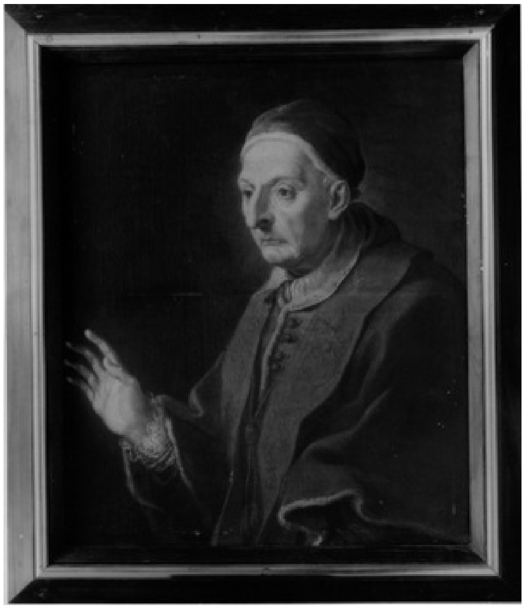


***New Match:** different lighting, same work of art.*
(One categorized as a full-length portrait of a man, the other as a portrait of a lady.)

Alternate Images

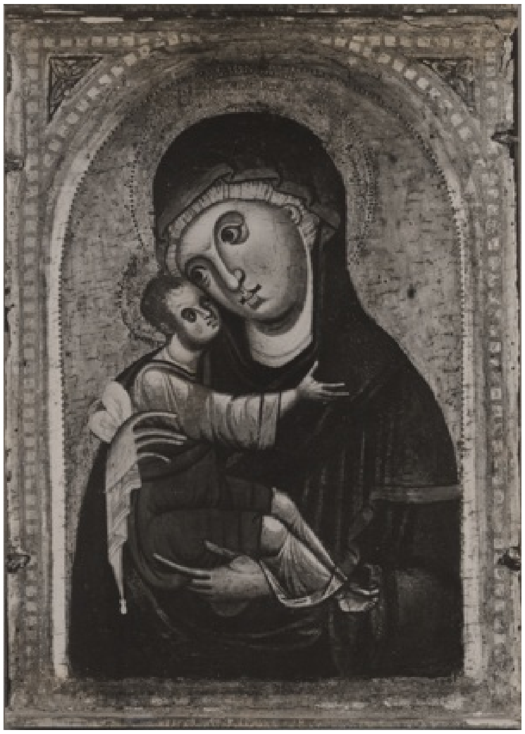
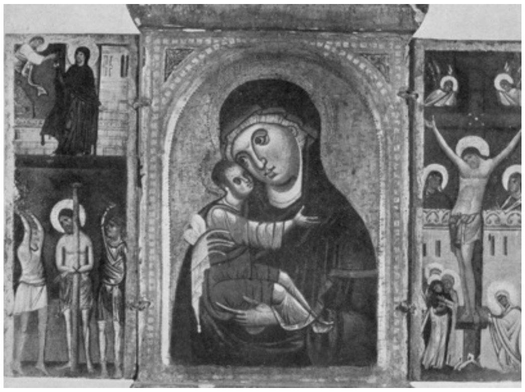
These matches were photos that both depicted the same work of art but showed alternate views (for a total of 115 matches). Frequently this was some sort of detail shot of the work. In all of these cases both the images and the data agreed. These matches were particularly interesting as finding a portion of an image inside another one can be quite technically challenging. Seeing the results provided by MatchEngine were quite heartening and suggested the possibility of finding many detail shots of a work of art.

The first work of art shows a dramatic difference in lighting as well as cropping. The photo on the right includes the frame of the work whereas on the left the image is cropped dramatically (into the painting itself).

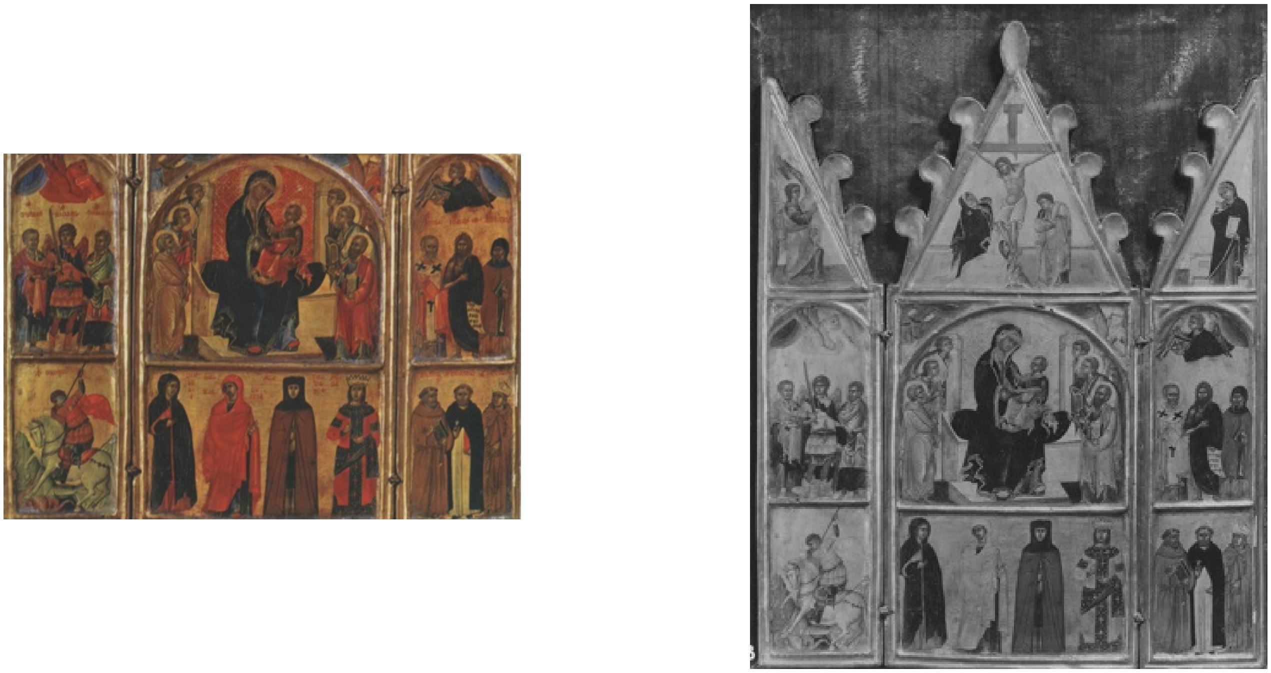
New Match: *different cropping and lighting, same work of art.*

The next work shows a close-up of the center portion of the work. Both photos are also in black-and-white.

New Match: *detail of the same work of art.*

This final representative on an alternate, match is both a close-up detail shot and in color, compared with the black-and-white full shot.

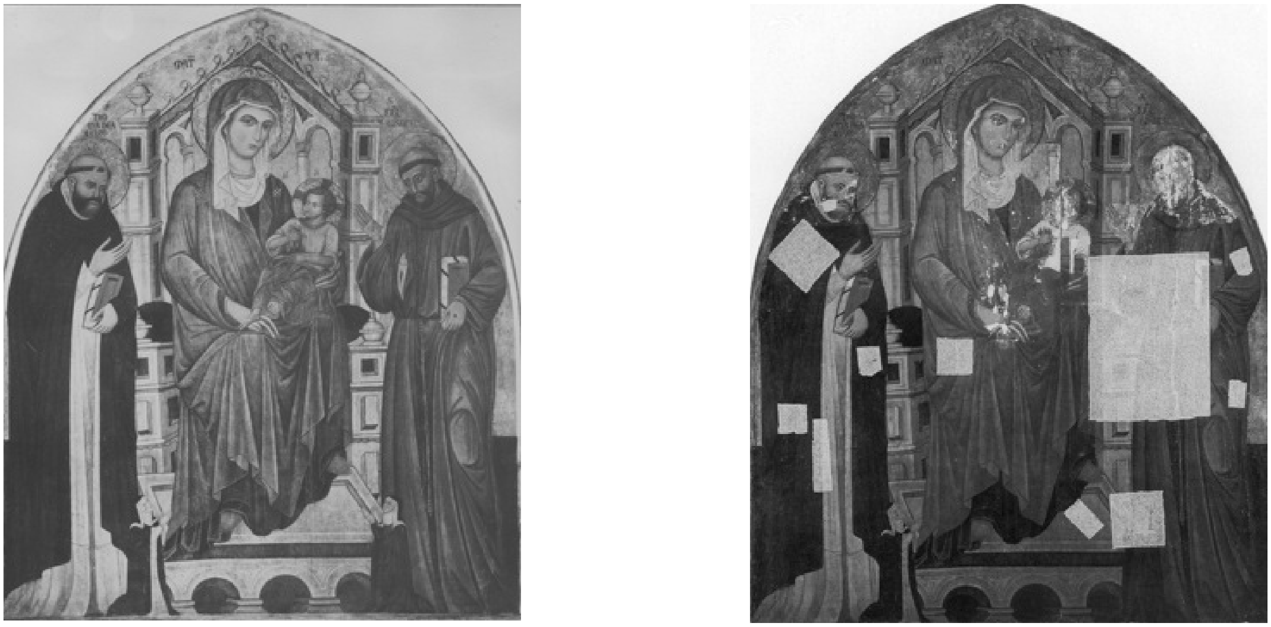


***New Match:** detail shot, color vs. black-and-white, same work of art.*

Modified/Conservation Works

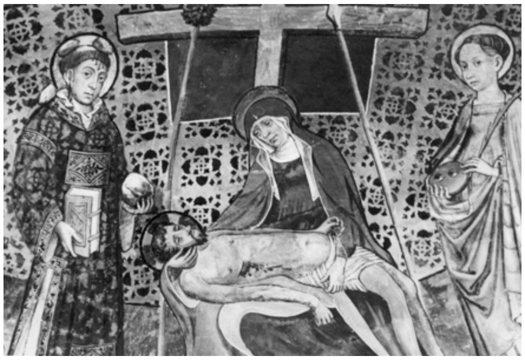
The same artwork before and after the conservation process were discovered during the image similarity analysis. Since the photos in the Frick's collection span many years, there are many instances where there are early photos of an artwork (from the early 1900s) together with photos from later in the century. Occasionally, an artwork will be in the process of restoration or will have undergone restoration at some point in the interim. Eight works were discovered in which possible restoration had been undertaken.

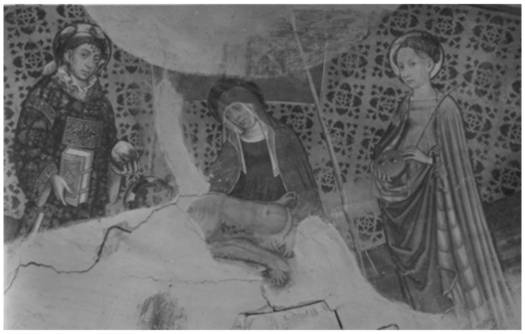
In the first work, restoration is in progress (seemingly an x-ray photography of the work):



***New Match:** same work of art, seemingly an x-ray or an in-progress restoration.*

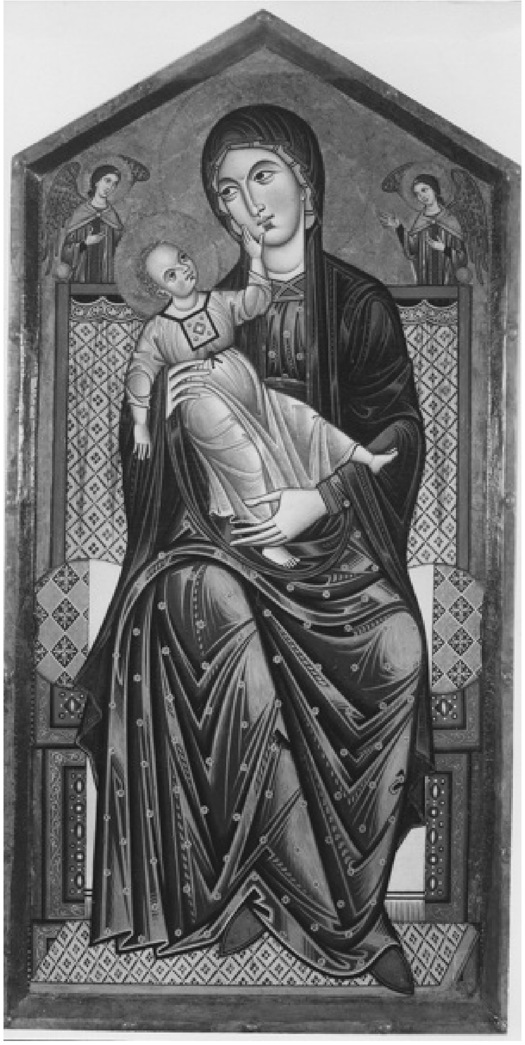
In another match, extensive restoration has been completed. Large portions of the fresco have been rebuilt and re-painted.

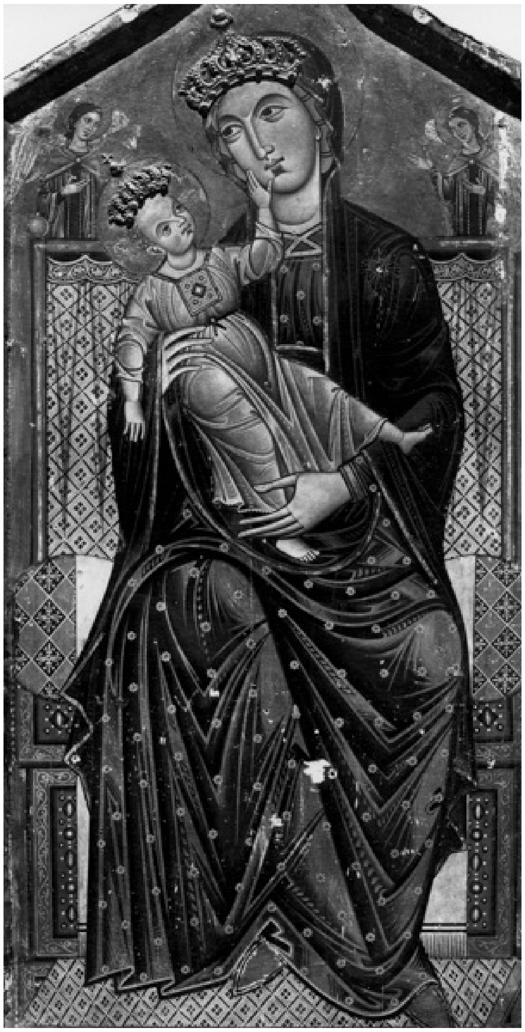




New Match: *same work of art, before and after restoration.*

Finally, a more subtle example: chipped paint has been repaired, the frame has been repaired, and seemingly extraneous crowns have been removed.



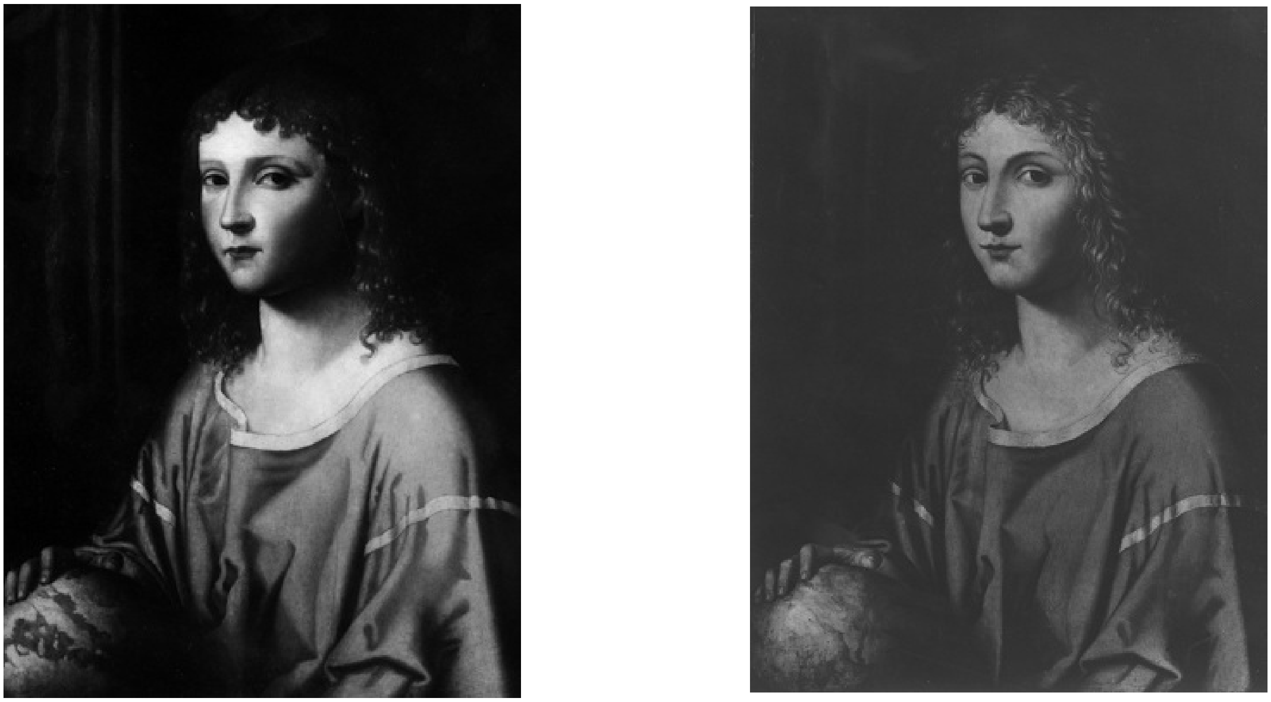


New Match: *same work of art, before and after restoration.*

Copies

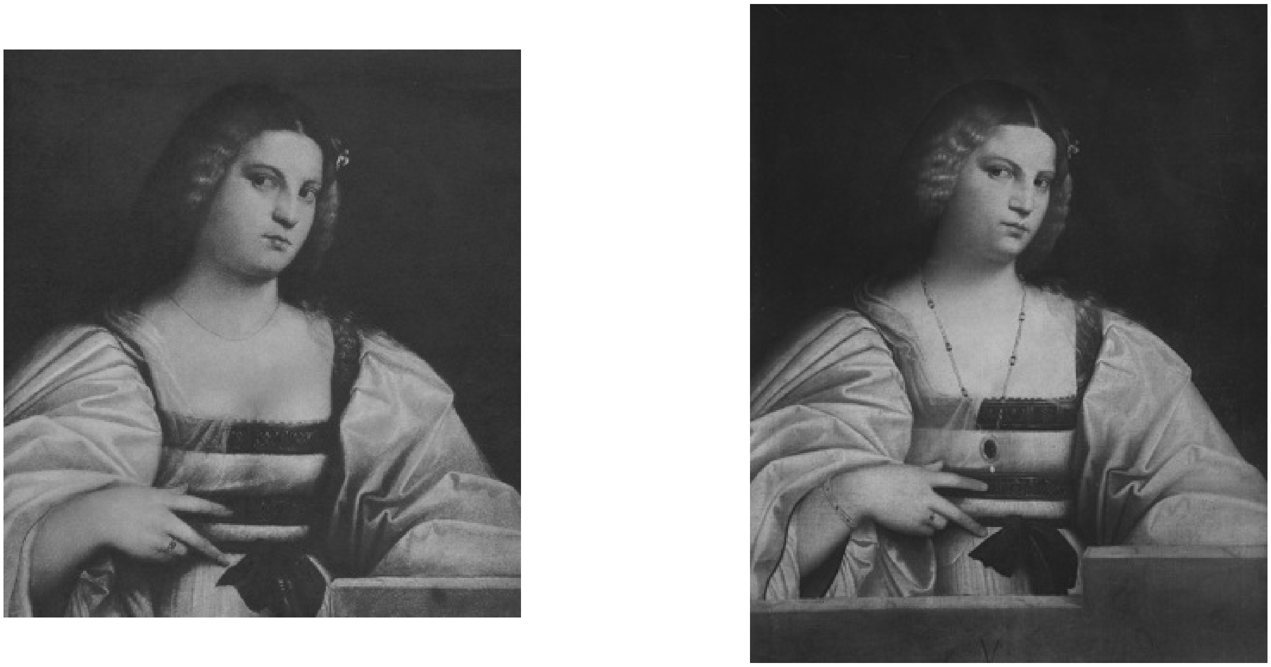
16 pairs of similar, but slightly different, artworks were discovered: the artworks were both copies of each other or of a third artwork. This discovery was especially interesting as it showed how potentially powerful MatchEngine's algorithm is. Even though the photographs aren't of the same work, it's still able to find the strong similarities between the works and expose them as a strong match.

The first two works are both later copies of the same work by Leonardo Da Vinci. Note the differences in the faces and in the globe.



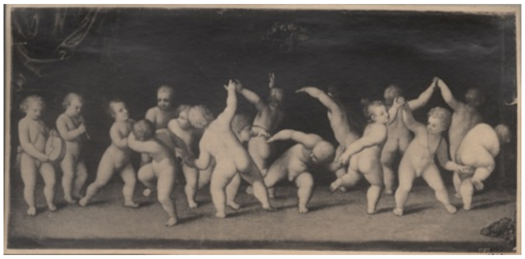
***New Match:** different work of art. Note the different face and globe.*

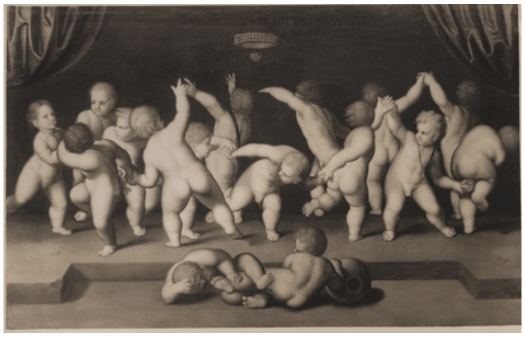
In another case, both works of art are copied from a third work (with slightly different faces and different necklaces).



***New Match:** different work of art. Note the different face and necklace.*

In this final case, both works are seemingly quite similar, with changes to the positioning of the children, the addition (or removal) of some children at the bottom of the work, and a change in the chandelier.



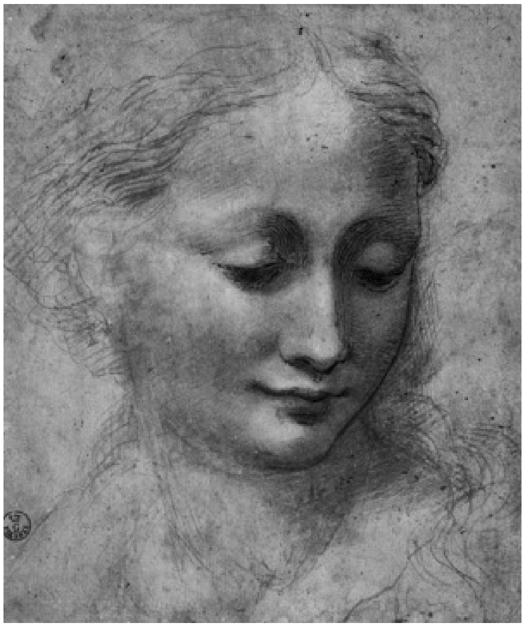


***New Match:** different work of art. Some children missing, added, changed.*

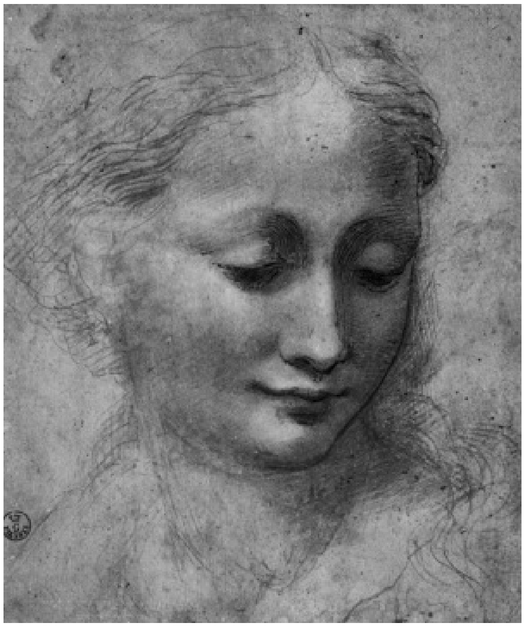
Digitization Errors

The image similarity analysis was also able to uncover 138 unexpected matches: cases of identical artworks with metadata in strong disagreement. These seemed to be the result of either the wrong image being uploaded for a work or the wrong metadata being used. Either way, it appears as if most of these problems occurred during the digitization process by the outside vendor because the Frick's internal physical records are still correct. Such discoveries are especially useful: the Frick Photoarchive has been able to correct the erroneous data and provide a better digital archive as a result.

The following works exemplify the kind of cataloging errors that were exposed. The images appear to be virtually identical yet have very different metadata. It's likely that the wrong image was paired with a metadata record, in this case:



Arms with Folded Hands
Castello sforzesco, Milan.



Female Head
Gabinetto disegni e stampe degli Uffizi, Florence.

First work doesn't match description, wrong cataloging.

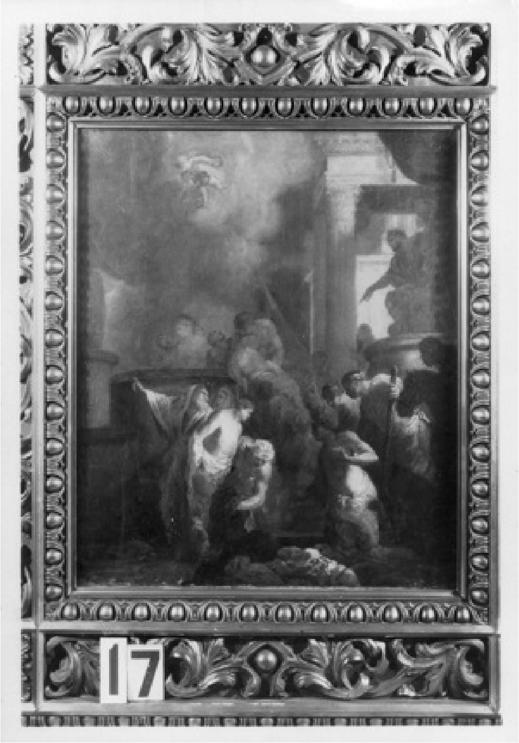
Additionally, these photos are in color and black-and-white but disagree on the metadata. In this case, it's likely that the correct image was uploaded but the wrong metadata was used.

Still Life with a Bottle, a Plate, a Mortar and Pestle, a Bowl, a Pot, Game and a Cat on a Stone Ledge.

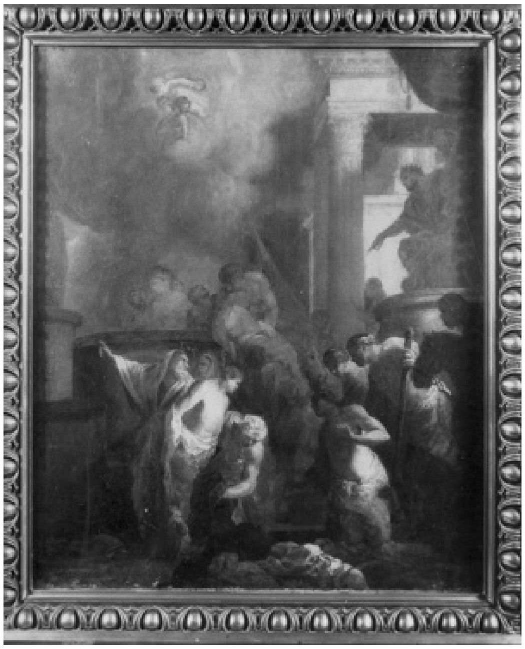
Virgin Entrhoned Nursing Christ, Between Two Saints.

Second work doesn't match description, wrong cataloging.

There were an additional 16 matches which may have been a cataloging mistake, or may actually be correct and require additional exploration by a researcher. The following match is such an example:



A Martyrdom
Accademia di San Luca, Rome.small>



The Corporal Works of Mercy
The Faringdon Collection Trust, Buscot Park.

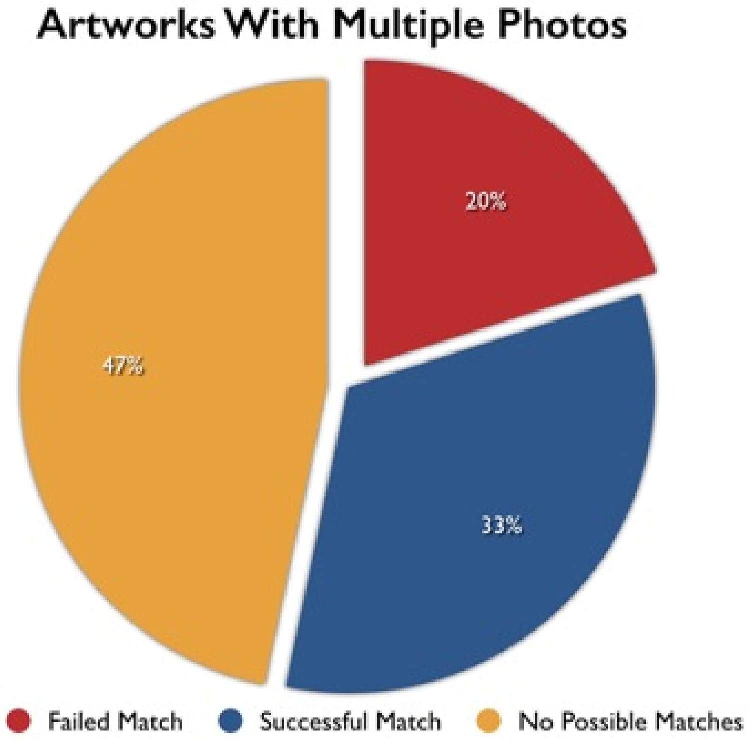
Same work, perhaps changed collections?

Measuring Image Analysis Efficacy

Even with all of these interesting new matches being discovered using image analysis it's important to attempt to understand how effective the MatchEngine algorithm is at finding matches. The best way to quantify this is by looking at images where a match should have occurred but did not.

Within the Italian Anonymous Art archive there are 1357 works of art associated with more than one photo. These photographs were manually grouped together by researchers at the Frick Photoarchive. The photos associated with a single artwork aren't always alternate views of the same work. Frequently, they are multiple photos of an artwork from different angle, the front and back of a work, or pictures of a three-dimensional artwork. Sometimes they are photos of different aspects of an artwork (for example three photos, each of a different panel in a triptych).

To better understand the types of photographs that were available for the artworks, a full survey was done of all 906 artworks that have multiple photographs but were not explicitly matched by the MatchEngine algorithm. The artworks were broken down into two categories: artworks for which there was no obvious visual relationship between the presented photographs and artworks for which there was some strong visual similarity between two or more of the photographs.

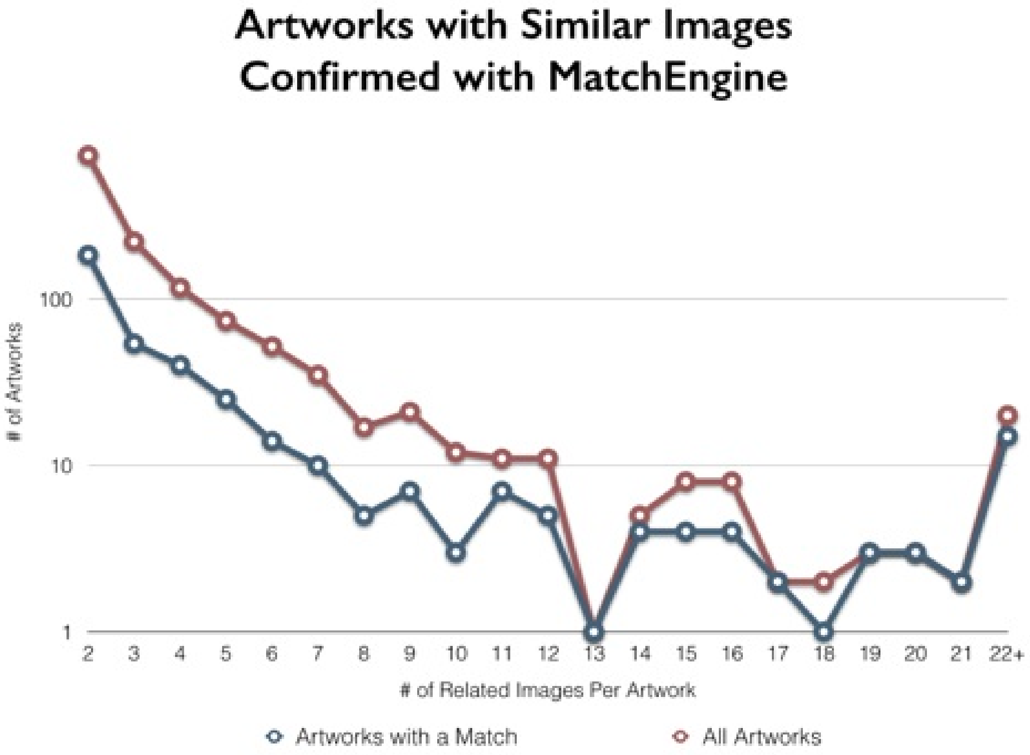


47% of all artworks with multiple photos had no two photos that were visually similar to each other. In those cases, the MatchEngine algorithm was incapable of finding any relationship: MatchEngine is only able to examine what is presented in the image itself. For example, the following artwork depicts two separate panels in the same piece:

em>Two different panels from the same artwork, no overlapping details.

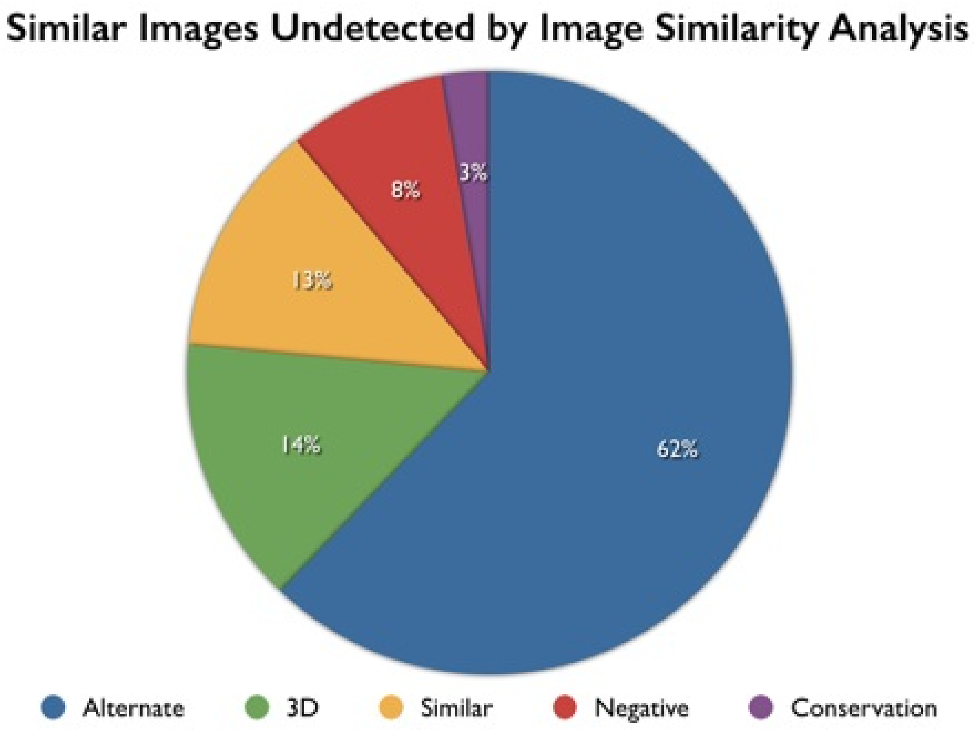
Of the remaining 53% of the images that did have a visual relationship between two or more of the images 33% were successfully matched and 20% were not.

Initially it was assumed that there might be a correlation between the number of photographs made available for an artwork and the likelihood of there being a confirmed match. An analysis was completed looking at artworks broken down by the number of photographs associated with the artwork:



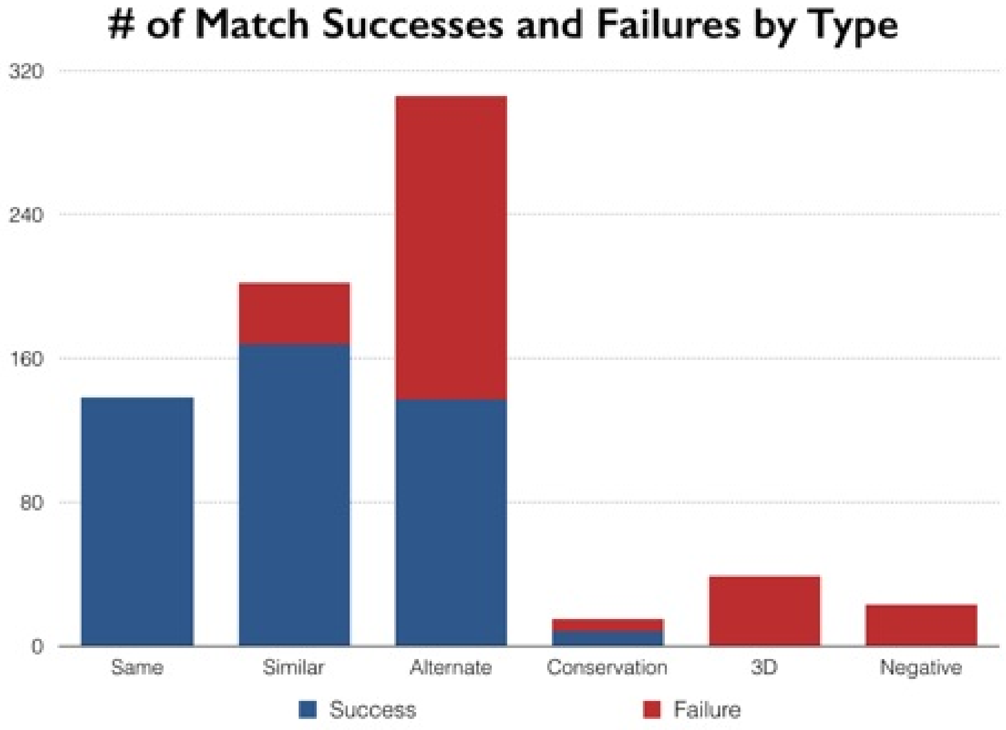
Looking at these numbers there does not appear to be a strong correlation between the number of photos associated with an artwork and the likelihood of there being a match. Only at the upper-end of the spectrum (for artworks associated with 19, or more, photographs) is there a strong correlation with a successful match occurring.

In order to understand where these matches come from and where the failings are, the images that were not matched need to be examined. This process will lead to a better understanding of the limitations of the MatchEngine technology and can help to set researcher expectations appropriately. A full breakdown of the types of images that weren’t matched included:



To arrive at this breakdown, I performed a full survey of all the 272 artworks that have at least two photographs with a strong visual relationship. Where there were multiple potential matches between photographs, the best possible photograph pair was chosen to be representative for that artwork.

At first glance, the types of matches appear to be similar to the types of matches that MatchEngine successfully discovered. However, a final breakdown of the images that failed to match was compared to the images that successfully matched using MatchEngine:



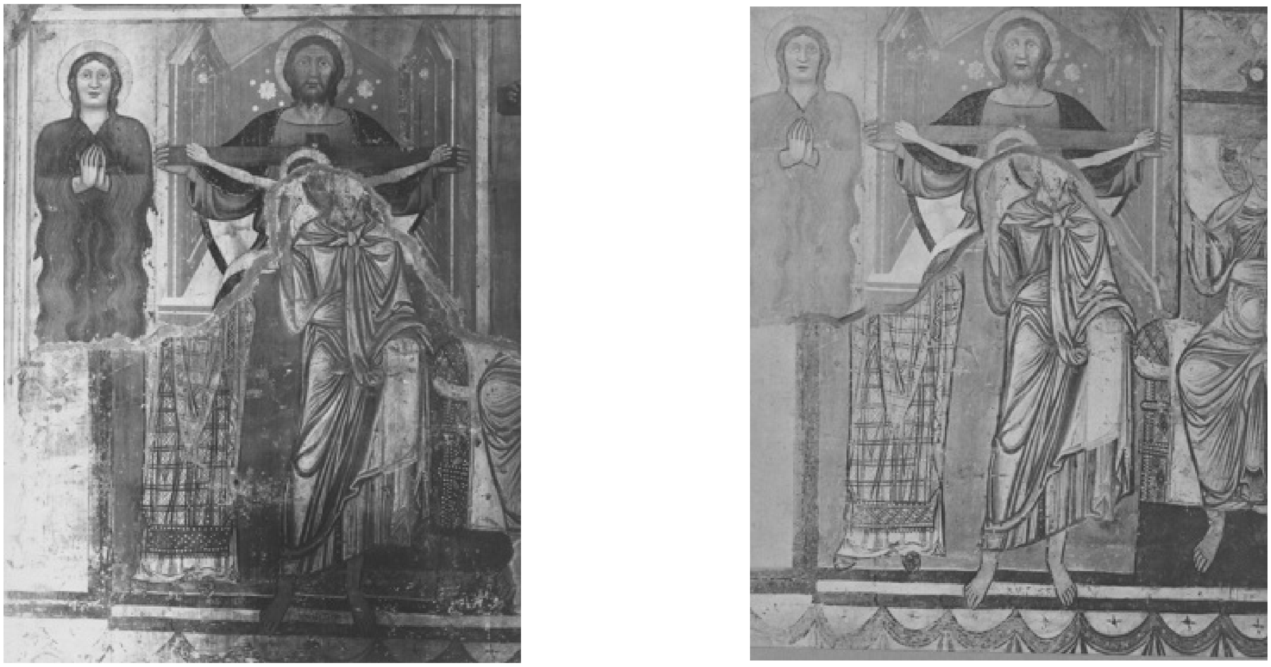
This is where the shortcomings of MatchEngine became apparent: every single three-dimensional and negative match failed in MatchEngine, as did the majority of alternate shots. Also note that there were comparatively very few failures where the photos were similar and no failures when the photos were near (or nearly) identical.

An analysis of all the individual types of match failures will help us to better understand what, specifically, MatchEngine struggled with in the identification of these images.

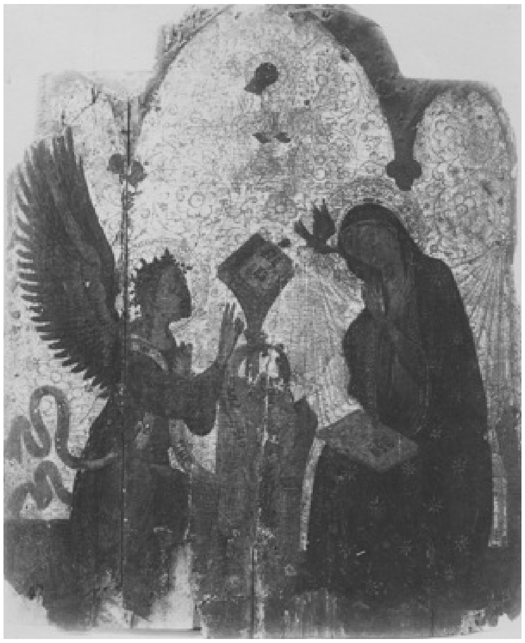
Similar Images

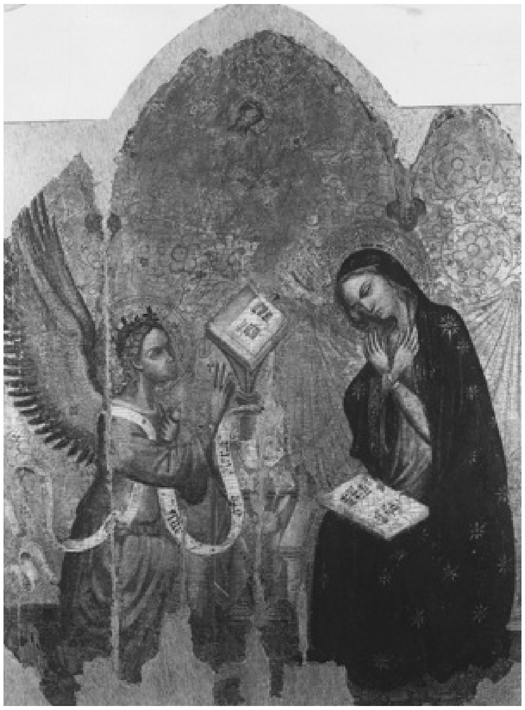
The following images are cases where the framing of the photographs are similar but the lighting between the shots is different. MatchEngine was able to successfully discover a number of these cases so it’s a bit surprising that that it struggled with these results (there were 168 successful similar image matches and 34 unsuccessful matches – or about 17% of all similar image matches failed).

It’s likely that the MatchEngine algorithm is looking at edges within the image, so with sufficiently different lighting some cases are no longer easy to pair. Below are some example of similar images that were not matched by MatchEngine:



Seemingly, the difference is between direct and raking lighting.



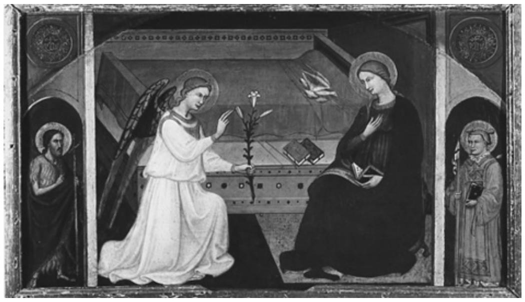


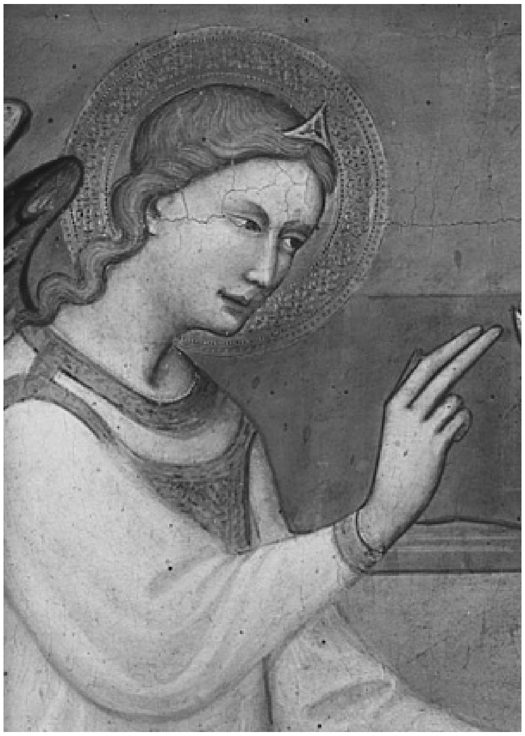
Very different lighting and exposure.

Alternate Images

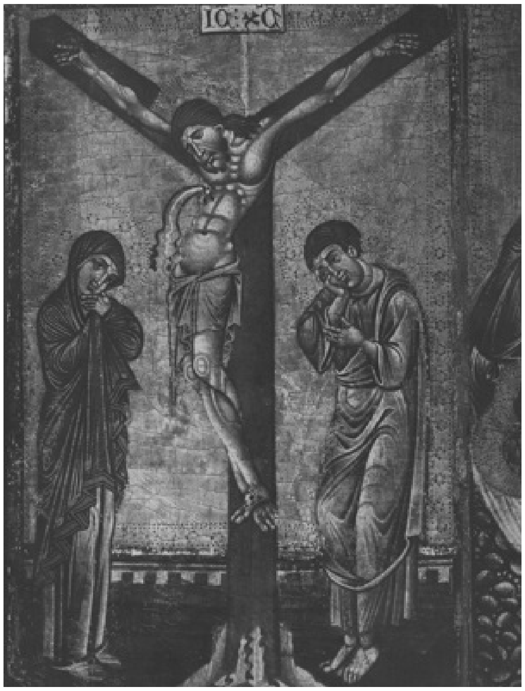
Alternate images of the same artwork produced the largest number of failed matches. In all of these cases a portion of an image contained within another image was not successfully matched. 137 alternate image pairs were successfully matched, whereas 169 alternate image pairs failed to match for success rate of only 45%.

Below are some examples of image pairs that failed to match, all of which were detail shots of small portions of the overall image:

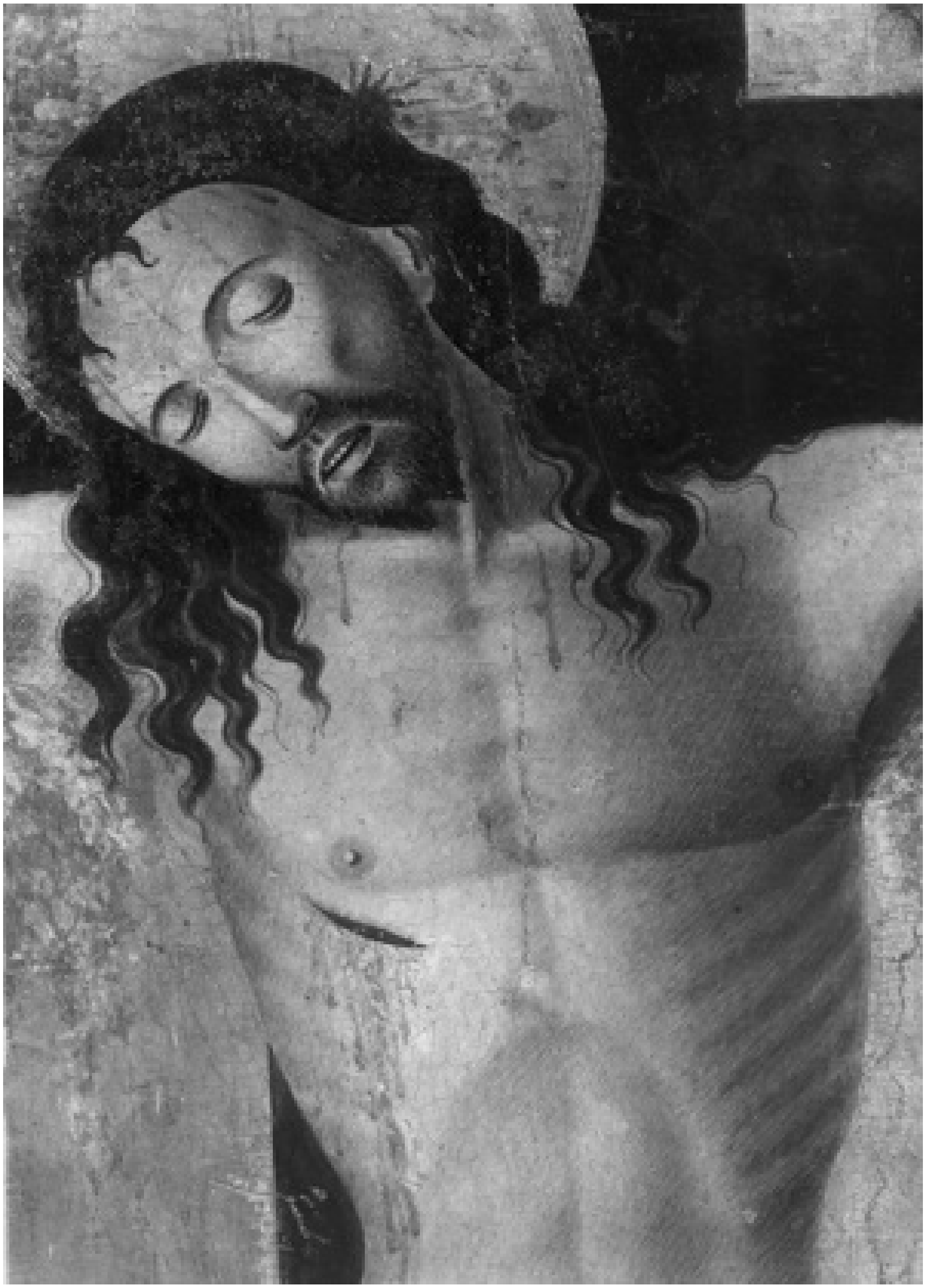




A small detail of the angel's head and arm.



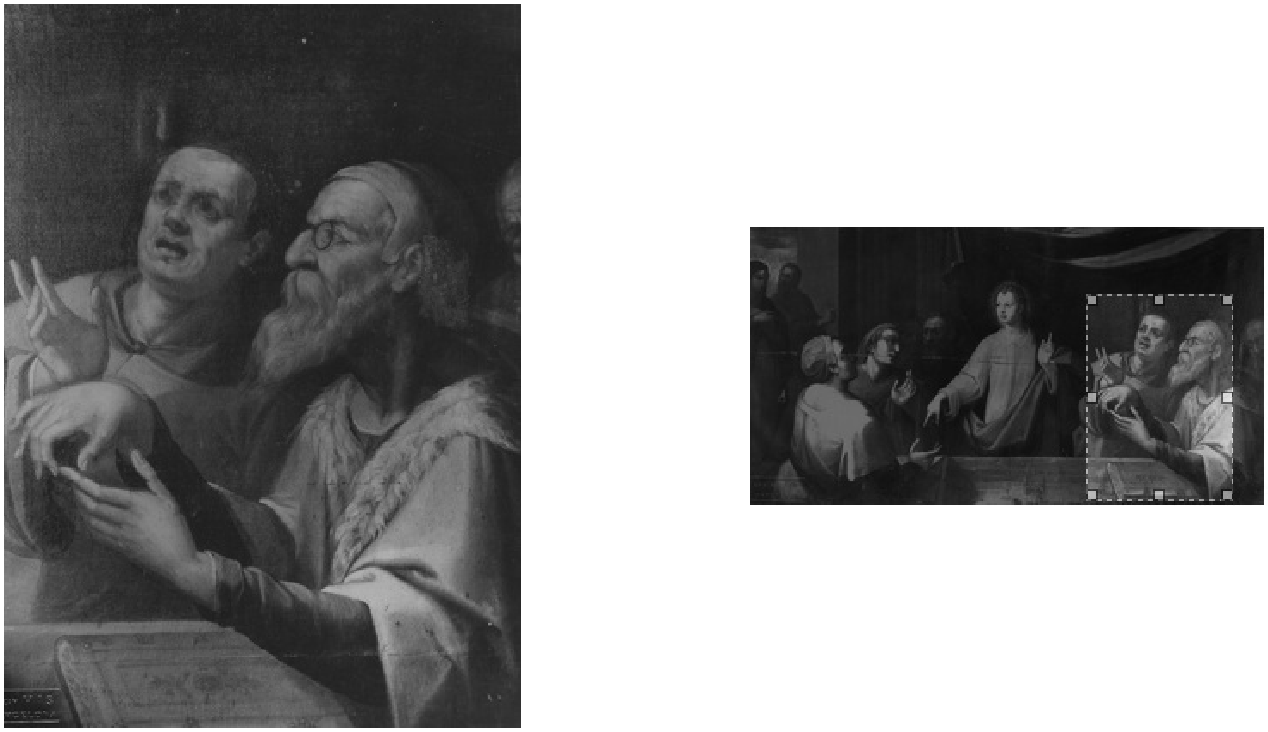
A tiny panel from the middle-right-hand side of the altar piece.



An extreme detail shot of the head of Jesus.

The poor results seemed particularly contradictory, as there was a large number of successful alternate image matches. However, one critical detail of the MatchEngine implementation is important to understand (this is also the case for most computer vision techniques): the image must be reduced in size before it can be successfully analyzed. In the case of MatchEngine, all images are reduced to 300 pixels in the smallest dimension before being processed. Taking this into account, and looking at the above example failures, an assumption can be made that there is a significant loss of detail during the processing of these images making a match difficult.

I also hypothesized that there is a correlation between the percentage of the image overlap between two images and the likelihood of there being a match between them (the larger the percentage the greater the likelihood of a match). To test this hypothesis all of the failed alternate image matches were analyzed. The overlapping portion of the image was manually selected to determine what percentage of the image was matching. A resulting selection would look something like this:

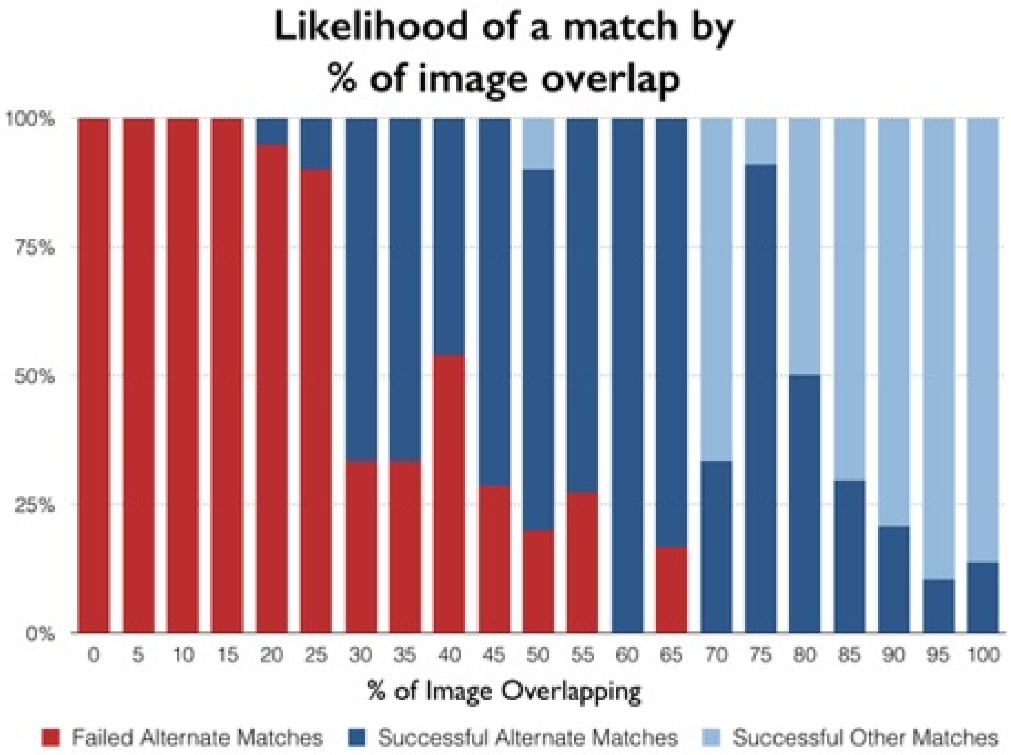


Manually selecting the portion of an artwork that overlaps with the corresponding alternate image.

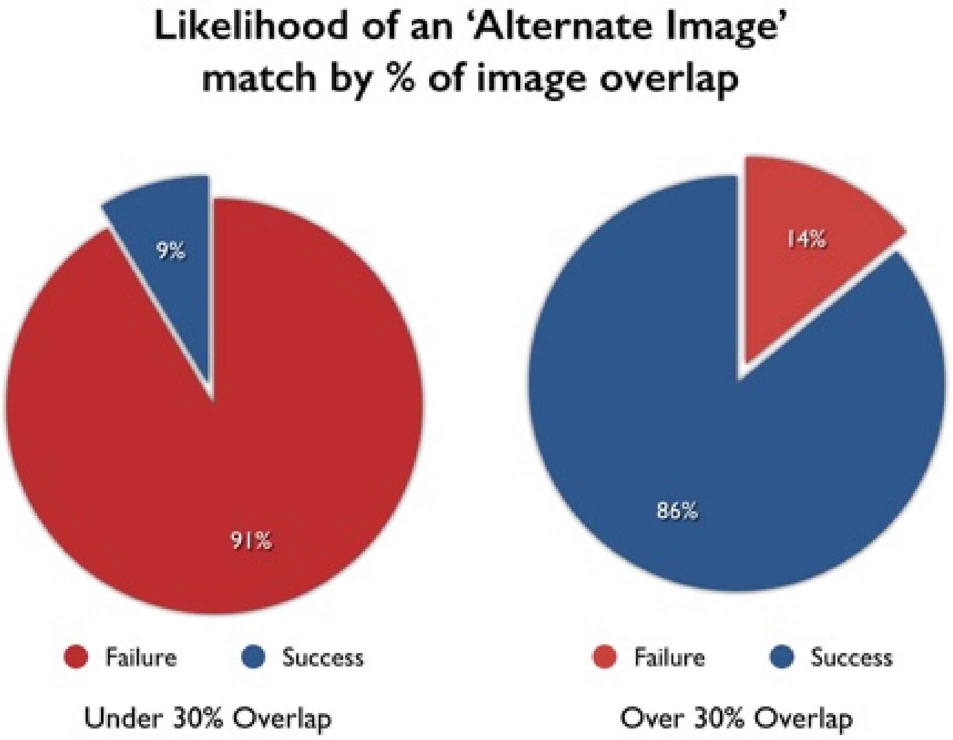
Thankfully, MatchEngine already provides the overlapping percentage for successful matches via their query API:

```
&quot;frick-anon-italian/13291.jpg&quot;: [
    {
        &quot;score&quot;: &quot;27.80&quot;,
        &quot;target_overlap_percent&quot;: &quot;100.00&quot;,
        &quot;overlay&quot;: &quot;...&quot;,
        &quot;query_overlap_percent&quot;: &quot;47.18&quot;,
        &quot;filepath&quot;: &quot;frick-anon-italian/13291b.jpg&quot;
    }
]
```

All that was left was to plot out the alternate image match failures, the alternate image match successes, and the other successful matches.



Looking at these results, it becomes immediately apparent that there is a strong correlation between the percentage of the image overlapping and the likelihood of there being a successful match. Below 30% of the image overlap, there are almost no successful matches between images. If the results are broken down to show the matches with less than 30% overlap and the matches with over 30% overlap these striking results are generated:

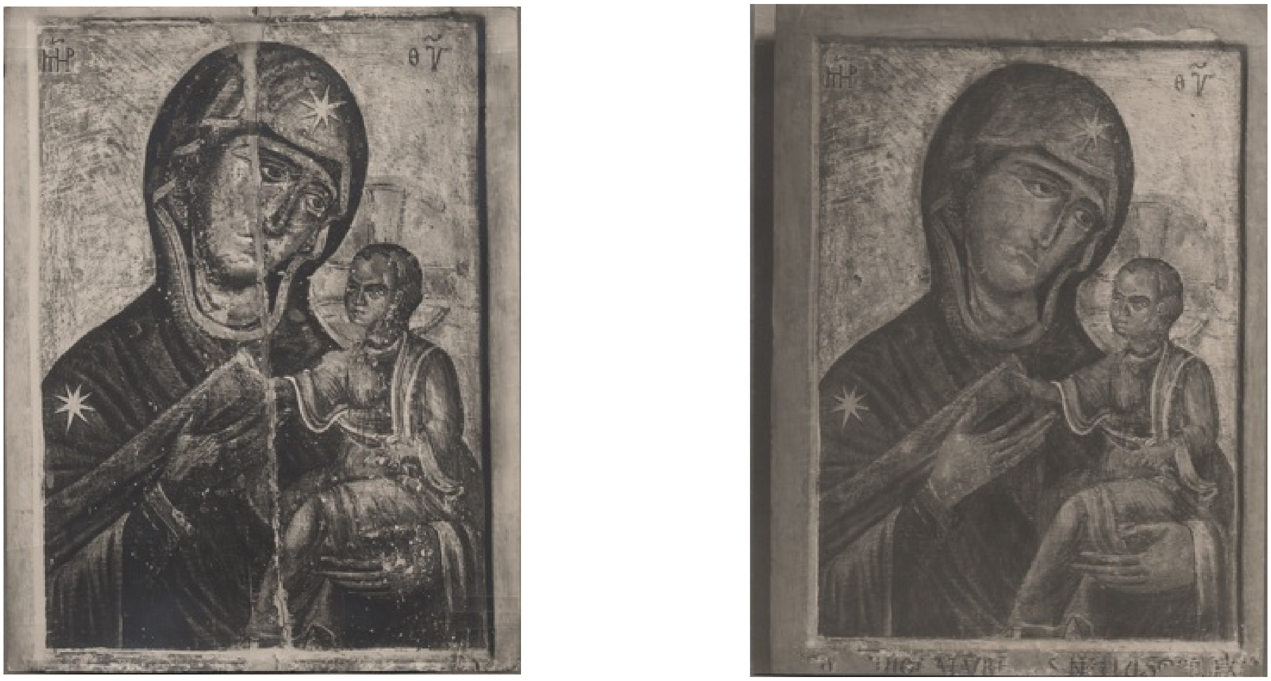


The results indicate that MatchEngine is not designed to adequately handle cases where there is less than 30% of the image overlapping. This is important to understand, as it can help catalogers better understand the limitations of computer vision systems such as MatchEngine. In many cases, when such a small fragment of the images overlap it is almost exactly like searching for a needle in an image haystack.

Conservation

As was the case with the successful matches there were a few cases where there were images of an artwork before and after the process of conservation. It was rather surprising that any images were able to match after conservation so it was unsurprising that nearly half of the conservation cases resulted in a failure to match (8 successful matches, 7 unsuccessful matches).

An example of a work, after conservation, that failed to match:



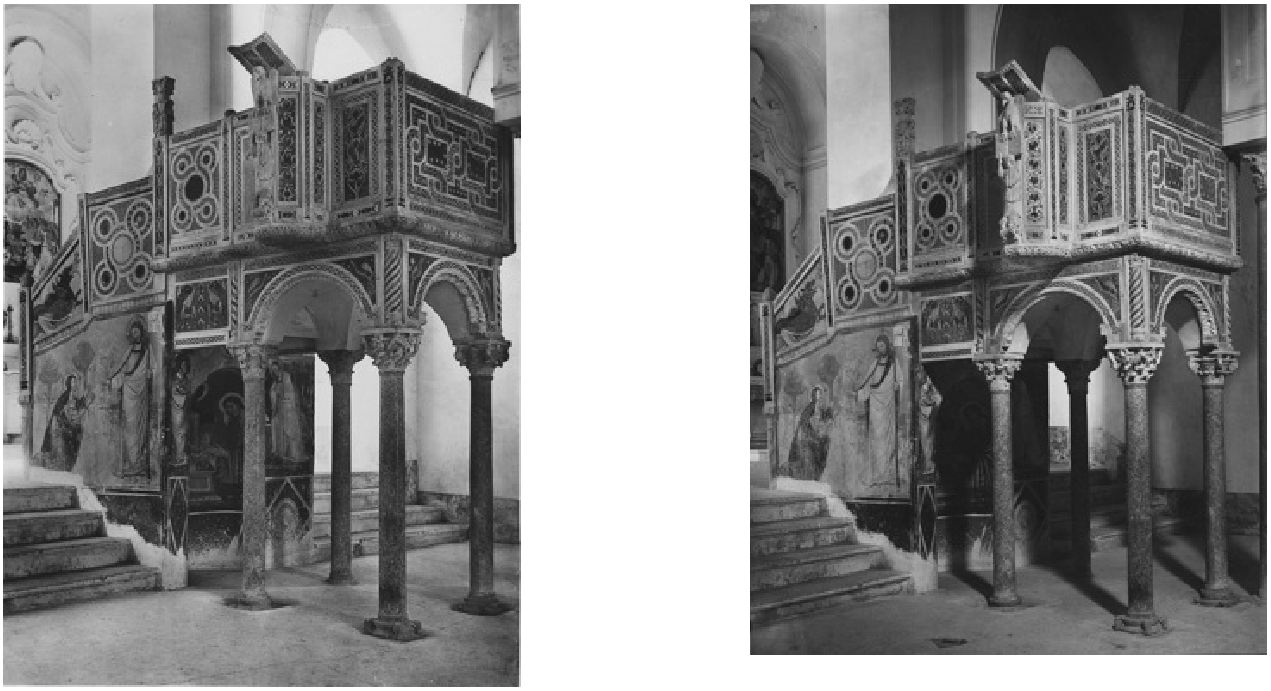
Work after conservation with different lighting.

Three-dimensional Works

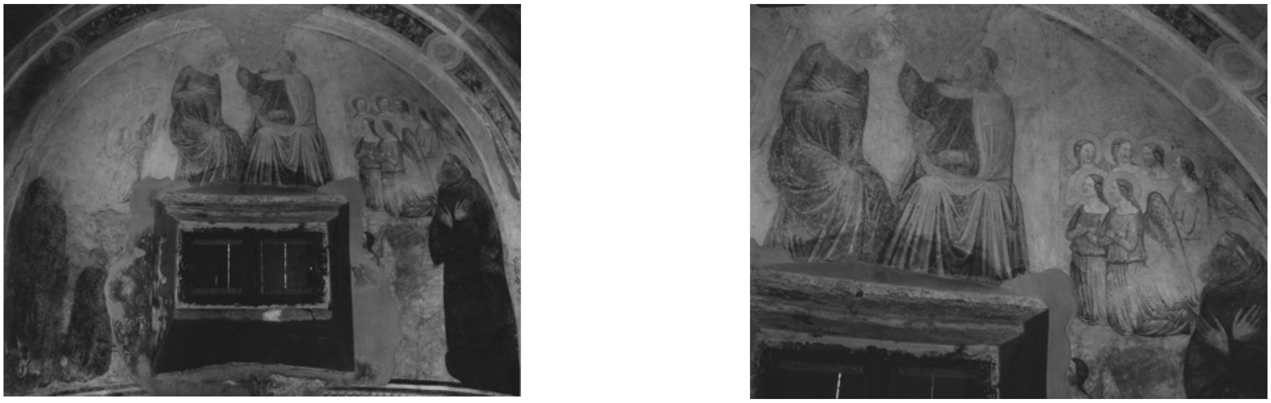
According to the MatchEngine web site, MatchEngine “cannot be used for identifying 3D objects.” Analyzing the failures tends to come to the same point of agreement: none of the 39 three-dimensional artwork images successfully matched each other.

Presumably, a different service would need to be used to find three-dimensional matches of this nature. Unfortunately, I am not aware of any services that provide this technology in a way that is able to gracefully scale to thousands of images in the way that MatchEngine can.

The results included the following incomplete matches:



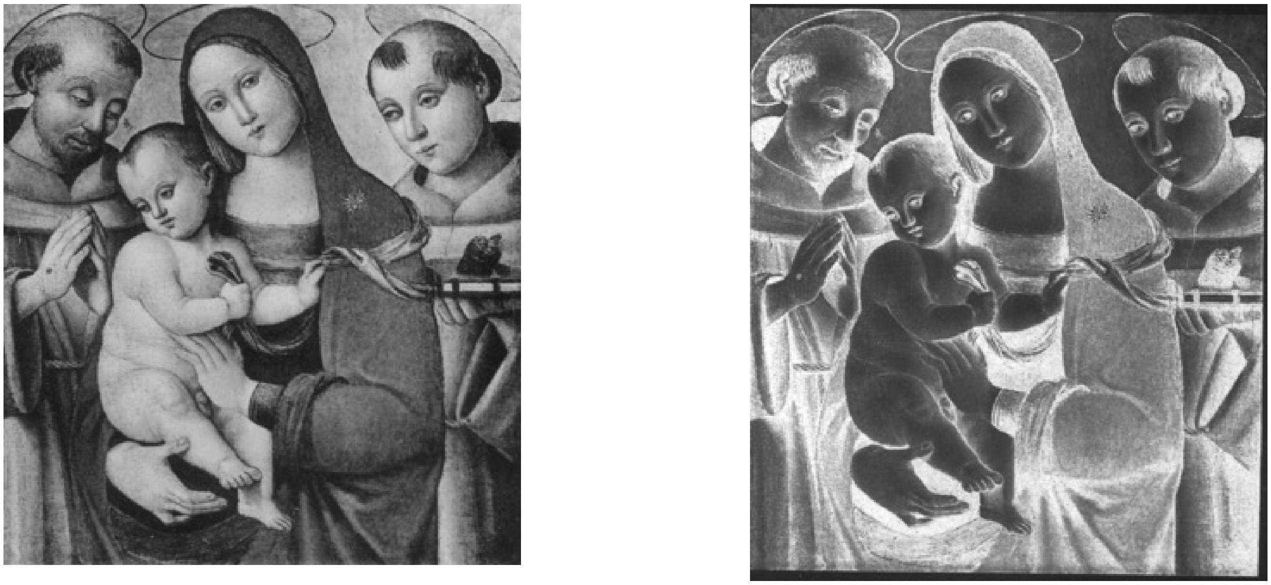
Same object with different lighting.



Same object at a different angle (even though it is a fresco, it’s observed from different angles, causing a failure).

Negative Images

The anonymous Italian art archive contains 23 artworks whose only alternate image is a negative. In all 23 cases, MatchEngine failed to find a match between the primary image and the negative. Considering that MatchEngine never claimed to match these types of images, it is safe to assume that this is not a use case that MatchEngine was designed to handle.



The same artwork in normal and negative views.

A proper match between the positive and negative forms of the image would be possible with MatchEngine if all negative images were first converted to their normal, positive, form. There would be extra work involved in making the match happen and since so few of the images in this particular archive fall under this criteria it was not deemed worthwhile to make this conversion.

Conclusion

This initial foray into using computer vision techniques to enhance the research potential of photo archives was exceedingly successful. A number of unknown relationships were discovered between images, digitization mistakes were detected, and corrections were offered. Additionally, the MatchEngine service used for performing the image analysis worked better than either the Frick Photoarchive or I could have hoped.

The potential of the MatchEngine service for the image set was fully explored: it works exceptionally well for images that are very similar, or for photographs that have moderate lighting changes, or for detail shots of the same artwork. However, MatchEngine is not a good tool for analyzing three-dimensional objects, detail shots with small amounts of overlap, and photographs with drastically different lighting.

Taking all of this into account, the overall quality of matches that MatchEngine provided within the anonymous Italian art archive was around 88%:

While there are limitations to computer vision techniques on the whole, these results are very promising. This high rate of match implies that there could be relatively few undiscovered new matches. Moreover, even after looking through all of the matches, MatchEngine never presented a single mistaken match. Every match had a high level of similarity between the two images and made sense to the catalogers.

It’s important to note that this particular archive is likely one of the most challenging use cases for using computer vision techniques in general (other archives are likely to have a much higher rates of match). The fact that most of the images in this archive were black-and-white (lacking additional information about the colors of the work) was a major hindrance to improved matching. The less data that the analysis engine has to work with, the harder it is to make a successful match. Additionally, many of the photographs in the set had drastically different lighting between shots, making it very hard to do comparisons. Presumably, another archive that had consistent lighting would fare much better.

With this new, powerful image analysis, the real fun begins: looking for other ways in which this analysis can benefit archives. There are three areas in which this image analysis would have immediate impact:

1. **Analysis and Error Correction:** the case demonstrated in this paper. Analyzing an established archive and using image analysis to look for undiscovered connections and to correct potential cataloging mistakes.
2. **Digitization:** performing image analysis during the digitization process. This analysis would provide the digitizer with contextual information about the work they're processing and help them to spot possible duplication or errors before they update the catalog.
3. **Merging:** given two archives of photographs, detect similar images and automatically merge the metadata records for a photograph. At the moment, the only solution to merging two archives is to attempt to rectify all of the metadata (which can be especially challenging if the archives are in different languages). If image analysis was used then all of the troublesome metadata could be ignored and relationships would be discovered purely based upon the images themselves.

The potential for computer vision and image analysis to change how photographs and images are managed in archives, libraries, and museums is absolutely staggering. Tasks that previously were insurmountable (such as merging two million-photograph archives) are now in the realm of possibility. The implications of this technology are still being explored and are likely going to completely change photo archives as they currently exists.

Originally published by John Resig on February 10, 2014. Revised for *Journal of Digital Humanities* July 2014.

Thanks

I would like to thank the Frick Art Reference Library for their interest and collaboration in exploring the potential of image analysis for photo archives. I received tremendous encouragement from them to explore this research and I'm very excited about collaborating with them more.

The Tineye team have been a pleasure to work with. I've been extremely pleased with the quality and reliability of their MatchEngine API. A few years ago, I explained to them some of the projects that I wanted to work on and they were excited to support me in their development by providing me with free access to their MatchEngine service. They've asked for nothing in return but I feel duty-bound to point out how good the service is and why you should use them if you have similar image matching needs.

I would also like to thank the Kress Foundation for providing a grant to fund future collaboration with the Frick Art Reference Library in developing Open Source tools for art photo archives to perform image analysis on their collections.

[1] At the moment the pricing for MatchEngine only works on a monthly payment cycle and doesn't exactly match the use case outlined here. Presumably, this exact analysis could've been achieved by signing up for a "Basic" plan, which has a $500 one-time setup fee and a monthly cost of $500. It supports an image collection size up to 20,000 images and supports 30,000 searches – both of which would've been enough to perform the analysis outlined here. It's almost certain that the TinEye team will have better ideas on how to perform this analysis in the most cost-efficient manner possible. ↩

About John Resig

John Resig is the creator of the Ukiyo-e.org Japanese woodblock print database and search engine. He develops tools to aid in the research of Ukiyo-e and other art history subjects. A Visiting Researcher at Ritsumeikan University, he recently presented at the 2013 Japanese Association for Digital Humanities conference in Kyoto, the Japanese Art Society of America, and the Digital Humanities 2014 conference. Mr. Resig is the Head of Computer Science at Khan Academy and is a renowned computer programmer, having created the jQuery JavaScript library used by over two-thirds of all web sites. He has also published two books on JavaScript programming: *Pro JavaScript Techniques* and *Secrets of the JavaScript Ninja*.

On the Origin of “Hack” and “Yack”

by Bethany Nowviskie

One of the least helpful constructs of our “digital humanities” moment has been a supposed active opposition, drawn out over the course of years in publications, presentations, and social media conversation, between two inane-sounding concepts: “hack” and “yack.” The heralding of digital humanities as the academy’s “next big thing” has been (depending on whom you ask) over-due or overblown, unexpected or contrived, refreshing or retrograde—but one thing is clear: everyone has a rhetorical use for it. The uses of “hack vs. yack,” on the other hand, rapidly became so one-sided that I find it odd the phrase retains any currency for critique.

After waffling through the winter, I’m finally publishing a brief note on the history of “more hack; less yack.” I do this not to reignite debates nor to comment on recent uses, but to provide a concise, easy-to-find, easy-to-cite account of its origin. I suspect the absence of such a thing a tricks us into repeating the phrase un-critically. This is ironic, because it now most often appears as short-hand for a supposedly un-critical, anti-theoretical, presentist, cheerleading, neoliberal digital humanities culture, standing in active opposition to… whatever the speaker or writer understands as salutary humanities *yack*. However, to contextualize “more hack; less yack” is not to defend it. It went viral at a moment when the last thing the digital humanities needed was an anti-intellectual-sounding slogan. It was perhaps objectionably pat, a little tone-deaf, and too easy to align with the “brogrammer” stereotype shortly to emerge from hacker culture. You might also rightly fire on it for its meme-like occlusion of implications beyond its immediate context, and for being chirped at you a few times too many, ca. 2009-2011.

It strikes me as more useful to offer an account of the early days of “more hack; less yack,” than to catalogue its later appearances in articles and blog posts. I can do this, because I attended the first several THATCamp meetings, and remember well how “more hack; less yack” evolved. It began as a goofball joke.

In 2008, a small group of graduate students, technology staff, and contingent and junior faculty at George Mason University founded THATCamp as a humanities-and-technology un-conference, meant to transplant into academic conference culture some aspects of the user-generated, self-assembling bar-camp format often encountered at tech gatherings. THATCamps do not feature peer-reviewed papers or invited talks. With only a few recent exceptions (*keynotes?* really?), no formal or pre-determined presentations are made at them, at all. Instead, “un-conference” participants are invited to propose ideas for informal sessions. These can range from open discussion and hands-on collaboration to demos and workshops—and a mashed-up schedule is built on the fly, by rough consensus and with opportunity for input from all attendees, in an open meeting on the morning of the event. THATCamps have rapidly become a relaxed and often exceptionally fruitful complement to formal, peer-reviewed digital humanities conferences like the one sponsored annually by the Alliance of Digital Humanities Organizations. And many see them as a refreshing, affordable, interdisciplinary supplement to disciplinary or thematic symposia and large humanities conferences of long standing. THATCamp, not DH-writ-large, was the context in which “more hack; less yack” first appeared; THATCamp is the context in which it spread—until it seemed to be taken, largely by colleagues newer to digital scholarship, as something of a capsule summary of an interdisciplinary and inter-professional community of practice with roots in fact stretching back some sixty years.

Two of our hosts at George Mason’s Center for History and New Media grew up listening to working-class radio stations in 1980s New England—the kind where a hyper-masculine disk jockey promised you, “Less talk, more rock!” We laughed when Dan Cohen, a pre-tenure History prof in shorts and sandals, combined this memory of his mis-spent youth with a science fiction classic to promise us a rockingly Martian good time: if it could foster learning and deeply-felt, immediate exchange in the absence of performative conference papers, THATCamp might offer everyone "less talk, more *grok*." But the *Stranger in a Strange Land* metaphor didn't hold up, and we all knew it—because in fact the un-conference model was meant to promote *more talking*, not less, and among a broader group of people. In Cohen's words:

> the core of THATCamp is its antagonism toward the deadening lectures and panels of normal academic conferences and its attempt to maximize knowledge transfer with nonhierarchical, highly participatory, hands-on work. THATCamp is exhausting and exhilarating because everyone is engaged and has something to bring to the table. *Thoughts on One Week, One Tool*

If anything was meant to be curtailed by THATCamp's challenge to 20-minute papers, 3-paper panels, and a few beats reserved for "this-is-more-a-comment-than-a-question"—it was not the talking. It was the overwhelming amount of time spent in passive listening. THATCamp offered an alternative to some established conference practices that seemed out of line with new opportunities for scholarly communication and in-person exchange. However, "fewer instances of paper-reading, grand-standing, and reinforcement of disciplinary divisions and the academic caste system; more grok" is not exactly catchy.

So, when Dave Lester, a software developer working at CHNM, quipped "More hack; less yack!" it made a silly kind of sense. Specifically, it made sense *as a comment on the dominant structure of academic conferences*, not as a condemnation of the character and value of discourse-based humanities scholarship. And it particularly resonated with the largely alt-ac crowd of humanities practitioners in the room that day—some fifty of us, by my estimate. And it seems to have resonated in particular with many of the librarians, programmers, and instructional technology staff who would find subsequent THATCamps such a delightful and *too-rare* opportunity to participate on near-equal terms with faculty attendees. This leads me to some editorializing on perhaps the least appreciated social aspect of "more hack; less yack."

If you are a scholar of (say) history or literature, yacking—by some definition of the term—*is your work.* It's how you think through your ideas, it's how you test and put them into circulation among your peers, it's how you teach: and may the best yacker (that is to say, the most informed theorist, clever and effective writer, erudite presenter, and thoughtful, decisive, fluent interlocutor) win. It's easy to see why so many humanities scholars who encountered Lester's phrase, often out of context, were inclined to understand "yack" as "deeply theorized, verbal and written exchange," and were therefore surprised and insulted to see it apparently denigrated. If, on the other hand, you are a staff member in a digital center, or an academic service professional like a librarian, instructional technologist, or digital archivist, a significant portion of your work progresses and is rewarded differently. You just might read something else in the juxtaposition of "hack" with "yack." Yacking is a part of everything people in these employment categories do, of course (because that's one way we all learn, think, and share)—but we are also asked to produce work, in service to humanities scholarship, of a different kind. The endemic, hour-by-hour "meeting culture" of an increasingly bureaucratic, often ill-managed, and top-heavy university means that, for many, time spent yacking is *the number-one thing* preventing us from doing our jobs.

In other words, "less yack; more hack" has a different valence for people whose productivity and performance is rarely judged on *les mots justes*. For humanities faculty, the academic workplace is predominately a site of expert verbal interchange. For staff asked to produce or maintain technical systems, run intellectual and social programs, or develop spaces and collections for scholarship, "yacking" may connote "wasting time." For better or worse, too much yack and not enough hack in the working day makes us come in early and stay late, just to keep our heads above water. (And I think we can acknowledge this common difference in expectations and accountability for time, while giving our staff and alt-ac colleagues credit for understanding what can be gained and lost in conversation, for striving to strike the right balance, and for their awareness of the deeper, structural problems in the systems within which they labor. Complicity is a complicated thing.)

I have an inkling that—just as its initial spread in the THATCamp community was predicated on a lack of appreciation for how the phrase might read to humanities scholars new to digital collaboration and the unconference format—the long, grumpy afterlife of “more hack; less yack” has depended on some elision of the daily challenges facing digital humanities service personnel.

Besides, isn’t “more hack; less yack” really just a strawman? I only find it being used in earnest rarely and beyond the academic digital humanities community. When pressed, even critics who continue to conflate it with DH practice and offer it up for ridicule are becoming more quick to modulate, clarify, and step away. Maybe it’s satire, now. In my view, to pretend or believe that “more hack; less yack” represents *a fundamental opposition in thinking* between humanities theorists and deliberately anti-theoretical digital humanities “builders” is to ignore the specific history and different resonances of the phrase, and to fall into precisely the sort of zero-sum logic it seems to imply. Humanities disciplines and methods themselves are not either/or affairs. The humanities is both/and. We require fewer slogans – and more talk and grok, hack *and* yack.

Originally published by Bethany Nowviskie on January 8, 2014.

About Bethany Nowviskie

Bethany Nowviskie is Director of Digital Research & Scholarship and Associate Director of the Scholarly Communication Institute at the University of Virginia Library, and President of the Association for Computers and the Humanities (ACH). She holds a doctorate in English from the University of Virginia and, in addition to her work on a number of notable digital archives and tools, has taught courses in writing, poetry, bibliography, and new media aesthetics and design. Nowviskie has been a practicing digital humanist since the mid-1990s.

Digital Historiography and the Archives

The following pieces by Joshua Sternfeld, Katharina Hering, Kate Theimer, and Michael Kramer are based on our session at the American Historical Association (AHA) meeting in 2014, “Digital Historiography and the Archives,” and the series of blog posts based on our presentations that we posted on Michael Kramer’s blog, Issues in Digital History, and cross-posted on AHA Today.[1] We are thrilled the JDH invited us to submit our blog posts for publication. We had conceptualized the session as an interdisciplinary roundtable discussion with short presentations by each panelist, followed by a rich and multifaceted discussion with the audience.

AHA 2014
Roundtable
Discussion:
Digital
Historiography
and the
Archives
CONTEXTCONTEXTCONTEXTCONTEXTCONTEXTCONTEXTCONTEXTCONTEXTCONTEXTCON

Expanding upon this format and our engaged in-person discussion, we had hoped that the series of linked blog posts would serve as a virtual, extended roundtable. Because it exists in the liminal, intermediary space between traditional conference session and formal publication, we imagined it as an experiment in new forms of “open source” scholarly communication. In sharing our work in provisional form, we hoped to encourage readers to continue this discussion on the blog. While we are publishing the discussion papers in the JDH in a more traditional format, the original idea behind the roundtable remains unchanged, and the blog continues to offer a space for posting criticisms, comments, ideas, and reflections.

Introduction

The preservation, analysis, and representation of digital information in digital collections, archives, and other media poses complex, challenging, and often confusing, issues for historical researchers, archivists, digital humanists, and librarians alike. Whether we even call these digital materials “archive” is at stake.

We hoped to address some of these issues in the session and subsequent blog posts, while also discussing some of the elements of a framework or vocabulary that can support a critical appraisal of digital information. In his 2011 article in the American Archivist, panelist Joshua Sternfeld introduced such a framework called *digital historiography*, which he defined as the “critical, interdisciplinary study of the interaction of digital technology with historical practice.” The AHA panel was originally organized, in part, as a response to that work.

All participants in the panel emphasized how archival theory and practice need to be an integral element of such a critical framework, along with evolving historiographical and professional practices. The digital medium has challenged historians to expand their knowledge about archives, and understand their function in generating scholarship and knowledge. But what might be the key theoretical and methodological questions surrounding the intersection of digital archives or digital collections and historical practice? What materials do archives collect and preserve, and why? Which materials are selected, and which are excluded? What are the driving forces and principles guiding the contextual information about collections provided by archives? Which political, social, economic, and cultural power relationships now structure the archives? How do we cope with the sudden, and at times unexplained, disappearance of collections in digital archives (portions of American Memory being a prominent example, as one audience member suggested)? How important is contextualization of collections in the digital environment? How can archival metadata be better situated in place and in time? These are questions that archivists and historians might come together to confront in critical and productive ways.

Fortunately, not only has the digital medium unleashed a heightened awareness of established archival principles and historical practice, it has also introduced new lines of theoretical inquiry. Historians and archivists are beginning to work with sources of varying scope, format, and provenance, thereby challenging both fields to reconsider the limits of historical inquiry, the contextualizing properties of metadata, the design of access systems, and the engagement of new audiences. In short, trends in digital scholarship and practices have contested our collective conception of “the archive” as well as the role of the twenty-first-century historian.

What became evident from the session was that historians must collaborate with information professionals, including archivists, to create critical contextual information for sources, reference resources, and repositories as well as new kinds of scholarly work that harnesses the power and registers the challenges of the digital archive, while serving a diverse community of users composed of researchers, educators, information professionals, students, artists, policymakers, and members of the public as a whole. The question is how? What areas of research should be explored and what methodologies, theories and practical models are already under development?

We hope that sharing materials from the roundtable on the blog and now in the JDH, even in provisional form, will continue to provide a catalyst for sustained discussion.

Originally published by Katharina Hering, Michael J. Kramer, Joshua Sternfeld, and Kate Theimer on January 21, 2014. Revised for *Journal of Digital Humanities* in August 2014.

[1] The authors wish to thank Vanessa Varin from the AHA for cross-posting our pieces on *AHA Today*. ↩

Historical Understanding in the Quantum Age

Joshua Sternfeld

The following remarks were delivered at the AHA Roundtable Session #83 Digital Historiography and Archives*. They have been slightly modified and annotated for the* Journal of Digital Humanities*. Please note that the concepts presented here are a work-in-progress of a much larger project about digital historiography; I welcome additional comments or feedback. Finally, the statements and ideas expressed in this presentation do not necessarily reflect those of the National Endowment for the Humanities, or any federal agency.*

What do the contours of conducting history in the twenty-first century look like and how are they changing? As we face a whirlwind of activity in reshaping historical methods, theory, and pedagogy, one thing is certain: The twenty-first-century historian has access to more varied and immense amounts of evidence, with the ability to draw freely from multimedia sources such as film and audio recordings, digitized corpora from antiquity to the present, as well as born digital sources such as websites, artwork, and computational data.

To get a sense of this abundance of evidence, let's consider for a moment a very contemporary historical record, one that is undergoing perpetual creation and preservation. In April 2010, the Library of Congress signed an agreement with Twitter to preserve all 170 billion of the company's tweets created between 2006 and 2010, and to continue to preserve all public tweets created thereafter. On a daily basis, that translates to roughly half a billion tweets sent globally.[1]

The case of preserving Twitter is a task that should excite both archivists and historians. For archivists, the primary technical and conceptual challenge is to provide timely access to such a massive corpus. As of last year, a single computational search by a lone researcher would have taken an unreasonable twenty-four hours to complete, which is why the Twitter archive remains dark for now.[2]

For historians, the challenge to analyze the Twitter corpus is equally if not more complex. Our traditional methods of close reading and deep contextualization fail to penetrate the social media network. To illustrate my point, let's conduct a "close" reading of a single tweet. At 8:16 PM, on November 6, 2012 – election night – President Barack Obama tweeted "Four more years" followed by an image of him embracing Michelle Obama.

Figure 1: "Four More Years" Tweet Sent by President Barack Obama

Granted, we could analyze the visual composition of the image that was selected for the tweet that accompanied Obama's re-election announcement, but the textual statement itself, "Four more years," leaves little to the analytic imagination.[3]

But when you consider that this single tweet, containing a three-word sentence and an image, was retweeted 784,170 times and favorited nearly 300,000 times, you begin to get a sense of the vast network of people that lies just beneath the surface. As historians, we ought to be curious about the transmission of this tweet, including the demographics of who sent it, how quickly it circulated, and whether additional information was delivered as it raced through the Twittersphere. When you consider that at least thirty-one million additional tweets were sent on Election Day, you begin to realize that as historians, we will need to reorient our approach to studying the past so that it does not involve reading every one of those thirty-one million lines of text.

I have selected the Twitter corpus purely for illustrative purposes as just one of several possible examples of how digital media is transforming our relationship with historical materials. And of course, digital humanists are comfortable with conducting distant reading for a variety of corpora. The point I would like to make is that historians now face a decision that would have seemed inconceivable just a few years ago: to work with a limited set of sources from a circumscribed set of archives or special collections, or with materials that would be impossible to digest in a lifetime? In other words, historians must learn to maneuver in not just the era of the million books, but the million financial transaction records, web pages, census records, and a wealth of other data points.

In today's talk, I would like to outline two concepts — scale and appraisal — that are critical for orienting how we work with an abundance of historical evidence made accessible by digital archives, libraries, and collections. To help us grasp the difference between digital and traditional modes of history, I propose we think about history in a *quantum* framework.

Just to be clear, I am not suggesting that a person could ever be in two places at once, or that historians will one day be able to "quantum leap" back in time with a companion named Al to rewrite the past! Rather, by placing history on a spectrum based on the scale of historical information, much like how our own physical environment can be placed on a scale from the subatomic to the astrophysical, we can reorient our understanding of human behavior, movement, networking, and activity.

Appraisal, to borrow a concept from archival theory, provides the framework to interpret the results of analysis conducted along the quantum spectrum. Whether we realize it or not, historians have always conducted appraisal of historical data. We have always assessed what information has value as evidence. And let's face it; we have been notoriously poor at explicating our methods of working in the archives. We have had a tendency to brush aside a detailed explanation for how we search for and discover archival materials, organize those findings, and then present them in a cogent argument.

In the past, our physical limitations of what we could read and collect forced us to adapt our modes of argumentation and analysis by drawing upon inference, annotation, instinct and most of all experience as guides toward appraising evidence. The sheer size and scope of today's digital sources demand a level of methodological rigor that we are not yet accustomed to applying.

My discussion about scale and appraisal of historical evidence, and digital historiography in general, is therefore grounded in a pursuit to define historical understanding in the digital age. Historical understanding explains how we learn about the past, authenticate evidence, and build arguments. In short, historical understanding answers the questions basic to all humanistic endeavors: how and why? Why did a phenomenon occur and why ought we to consider it significant? How do we explain a pattern of human behavior? These basic questions should be the guiding force behind all activities in digital history, from the building of new tools to interpretative projects. All too often, however, historical understanding has been drowned out in the noise of historical big data and the glitz of complex software, leaving skeptics to wonder whether digital history can fulfill its promise of revolutionizing the discipline. Applying methods of digital historical appraisal based on the scale of evidence sets us on a path towards validating conclusions and therefore contributing to historical understanding for a new era.

The Issue of Scale

Let's begin with the concept of scale, where I will borrow liberally from the sciences.[4] There is a theory in physics that the Newtonian laws of our physical environment may not necessarily apply at the level of the extremely small or extremely large.

The properties of gravity, mass, acceleration, and so forth begin to break down when you consider subatomic particles too small to detect using conventional instruments. Furthermore, mystifying substances such as dark matter seem to disrupt these same properties at an astrophysical level. While there are scientists who are working towards a unifying theory, for the time being these levels, and the laws that govern them, remain distinct.

I would argue that history operates in a similar fashion, with different levels of historical information, or data, with which one can choose to work. Up until very recently, the vast majority of historical work operated at what I call the Newtonian level, that is, the level at which a single historian can synthesize data into a coherent narrative argument. A basic definition of history, the study of change over time, reflects a Newtonian mindset:

> History: a continuous, systematic narrative of past events as relating to a particular people, country, period, person, etc., usually written as a chronological account; chronicle[5]

Much like Newtonian physics, Newtonian history has been incredibly successful in building an understanding of the past, particularly when accounting for the laws of causality and the interactions of individuals and societies.

Just as the principles behind inertia inform us that an object in motion tends to stay in motion or an object at rest stays at rest, we have developed over time principles of rhetoric and logic that allow us to work within a degree of epistemological certainty. History in the modern era, in other words, works best when investigating the likelihood that Event/Person/Society A may or may not have contributed to Outcome B, however abstract that outcome may be. My point is that the type of linear narrative history that has developed over the course of centuries has depended in part upon the degree of access to historical data, our ability to synthesize that data into a cogent argument, and our modes of representation.

What happens to the fabric of historical work if we extend our access to information along a quantum spectrum?

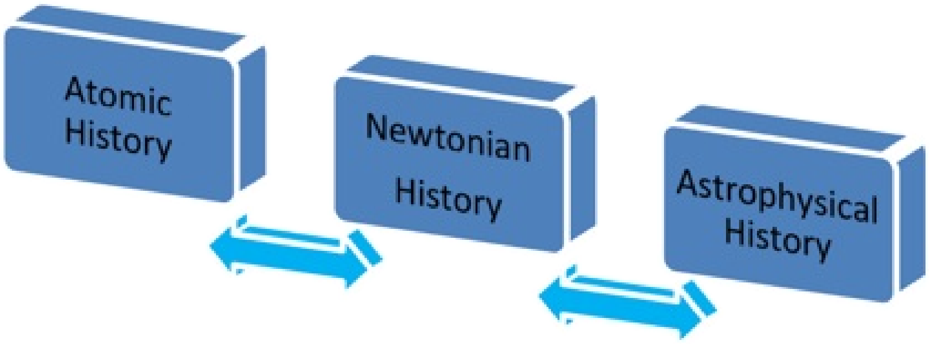


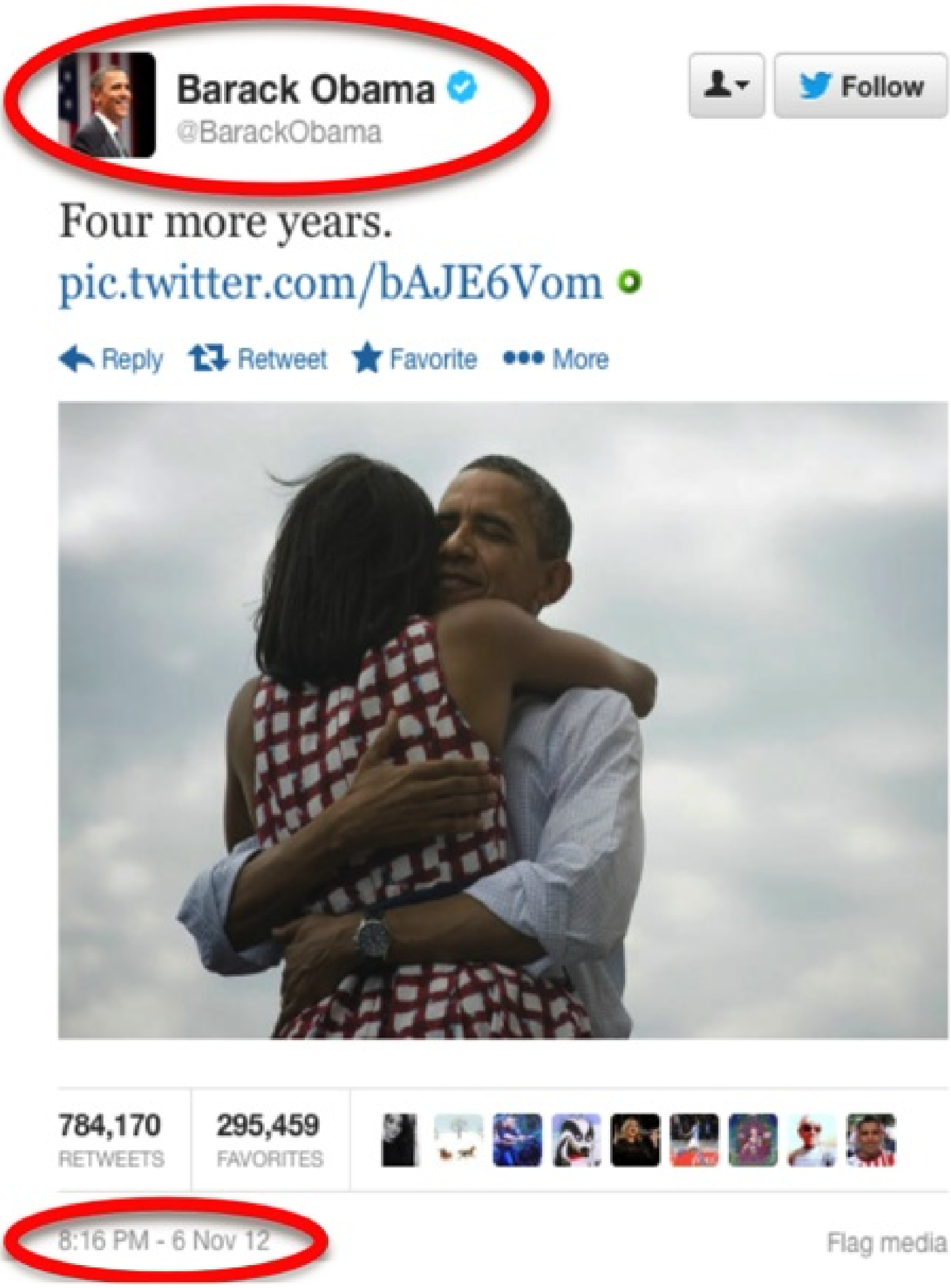
Figure 2: Quantum Spectrum of History

We begin to open new levels at which we can do history. On one end of the spectrum, there is the study of a micro piece of evidence, much as we have always done but at a level of precision that may not have been previously possible. We can study the digitized manifest of a slave ship, or a criminal trial in London, each with a rich history worthy of unraveling.[6]

Digital technologies have also allowed historians to probe the materiality of artifacts in ways that have unearthed new findings. Methods of spectral analysis, for example, have detected pigments underneath parchment, such as was the case with the discovery of the Archimedes Codex[7], or the revelation that Thomas Jefferson originally intended for the word “citizens” to be “subjects” in the drafting of the Declaration of Independence.[8]

Returning to our example of Twitter, we might consider the single tweet. While earlier our close reading of

Obama's tweet seemed one-dimensional, there is in fact an extraordinary amount of visible and hidden information surrounding the text message that one can extract from a published tweet that conveys a history unto itself. According to the Library of Congress, each tweet can contain 50 unique pieces of information, or metadata, such as the time it was published, the location from where the tweet was sent, and information about the person who sent it.[9]



Barack Obama
@BarackObama

Four more years. pic.twitter.com/bAJE6Vom

784,170 RETWEETS
295,459 FAVORITES

8:16 PM - 6 Nov 12

Figure 3: Provenancial Metadata of a Tweet

Besides information contained within the tweet, we may wonder about external circumstances surrounding its creation. For example, exactly where was Obama when he sent the announcement? Did he even compose it himself? Metadata can get us part of the way toward reconstructing the context behind the creation of a tweet, but we likely will need other sources of information outside Twitter, such as journalistic accounts of election night, to assemble a richer portrait of a moment in time.

On the other end of the spectrum, as suggested earlier, are the millions of tweets that may be connected to a single event, person, or community.

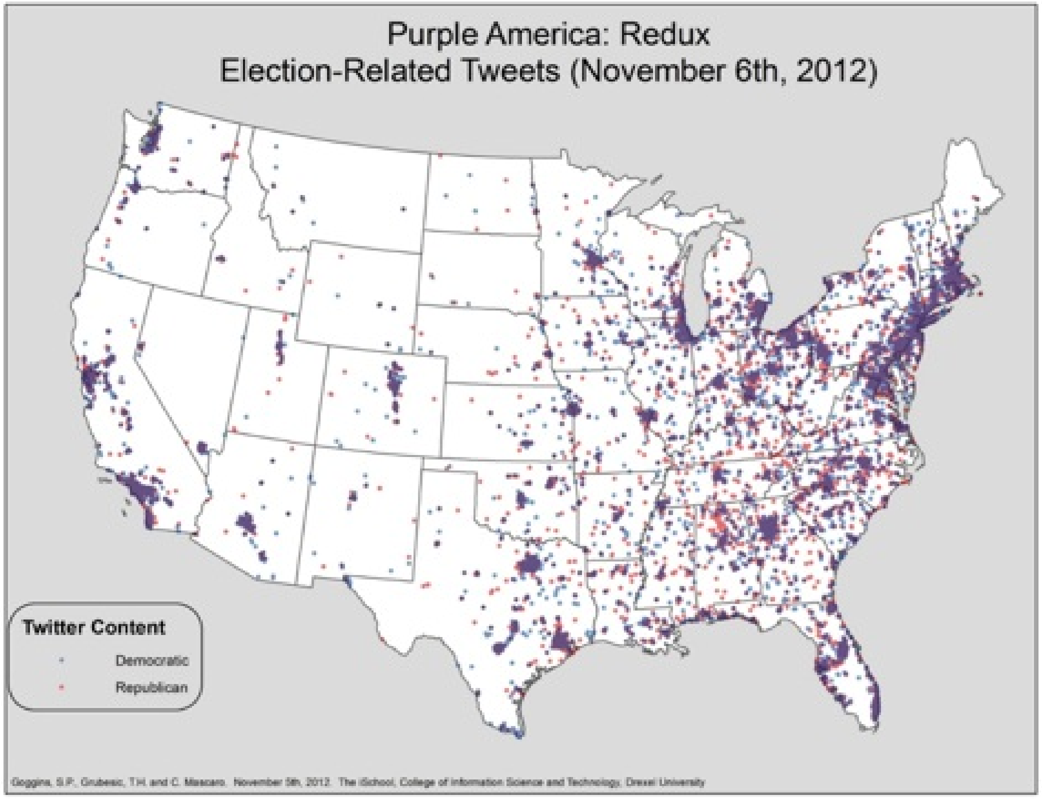

Purple America: Redux
Election-Related Tweets (November 6th, 2012)
Twitter Content
Democratic
Republican
Goggins, S.P., Grubesic, T.H. and C. Mascaro. November 5th, 2012. The iSchool, College of Information Science and Technology, Drexel University

Figure 4: S.P. Goggins, T.H. Grubesic, and C. Mascaro, “Election 2012 – Election Day Partisan Tweets across the USA and in Ohio,” (Drexel University, 2012), http://www.groupinformatics.org/election2012b. Accessed July 19, 2014. For more information on how these Twitter visualizations were generated, see: Alan Black et al., “Twitter Zombie: Architecture for Capturing, Socially Transforming and Analyzing the Twittersphere,” in Proceedings of the 17th ACM international conference on Supporting group work (2012).

Besides tweets, this extreme end of the spectrum also includes massive databases, storehouses of information of all varieties from complex texts, to databases filled with statistics, to non-textual media. One ship’s manifest is combined with nearly 35,000 others to form the *Trans-Atlantic Slave Trade Database*[10] and one London trial is collated with over 197,000 others conducted between 1674 and 1913 to form *The Proceedings of the Old Bailey*.[11] For the remainder of my talk, I want to focus on this end of the historical data spectrum.

As you may imagine, when you move into the realm of the vast, the traditional laws governing historical understanding begin to breakdown. Take the graphs produced by the scientists behind the Google Ngram Viewer, an analytic tool that counts the frequency of words and phrases across a portion of the millions of books scanned by Google. The scientists have demonstrated, for example, that by tracing the appearance of selected artists’ names we can generate graphs that reveal periods of state censorship, revealing that between 1933 and 1945 National Socialists in Germany suppressed discussion of artists deemed Jewish or degenerate.[12] On the one hand, these graphs are compelling for suggesting the effectiveness of the censorship campaign. On the other hand, they don’t necessarily reveal anything historians don’t already know, and they certainly don’t explain the causes behind the statistical trend. We have no idea from a graph alone who enforced the artist ban or the rationale behind its enforcement. In other words, big data visualizations such as those produced by the Ngram Viewer wipe away any remnant of historical causality. If you feel a sense of unease at this prospect, you are not alone.[13]

Our unease should suggest that we need to retake the reigns from the computer scientists and build a new framework for working with big historical data, one that can harness computational power while exploring deeper, less obvious connections. But where do we begin?

Returning to the physics analogy, the massive particle accelerator known as the Large Hadron Collider built in Switzerland was designed in part to test for the existence of the Higgs Boson, or “God particle,” by smashing together atoms and computationally sifting through the enormous amounts of resultant data. Of course, before the Large Hadron Collider could be trusted to produce reliable data, it needed to be calibrated and recalibrated according to stringent standards established by the scientific community. I would argue that digital history is lodged in a perpetual state of experimentation, smashing data together to see what is produced, but not yet at a stage of calibration to produce a substantial body of worthwhile evidence. We are simply not yet accustomed to working at different scales of data, and a great deal of calibration is still required.

Why is this the case? Why don’t the laws of Newtonian history apply at different scales? The reason is that we have yet to calibrate the digital tools and methods according to historiographical questions both old and new that are suitable for investigation. When we transition into a new scale of data, we naturally introduce variables that have the potential for invalidating our findings. Skeptics declare – often rightly so — that historical data is incomplete, fuzzy, and ill-suited for methods that operate on a presumption of scientific certainty. They also argue that the representational properties of digital media can supply a false sense of objectivity by denying data proper contextualization.

There is a lot of validity to these criticisms. More often than not, historical data sets *are* messy, and vulnerable to an assortment of critical attacks. Only by applying our skills of critical analysis will we feel comfortable participating in the creation and use of digital tools such as the Ngram Viewer. My point is that we must use our natural critical tendencies – in short, our skepticism — to calibrate our data sets and the methods used to interrogate them. Only then will we begin to fulfill what Alan Liu calls a “new interpretive paradigm” in the digital humanities, and digital history in particular.[14] This brings me to *digital historical appraisal*, which, in a nutshell, is the process by which we profile a dataset.

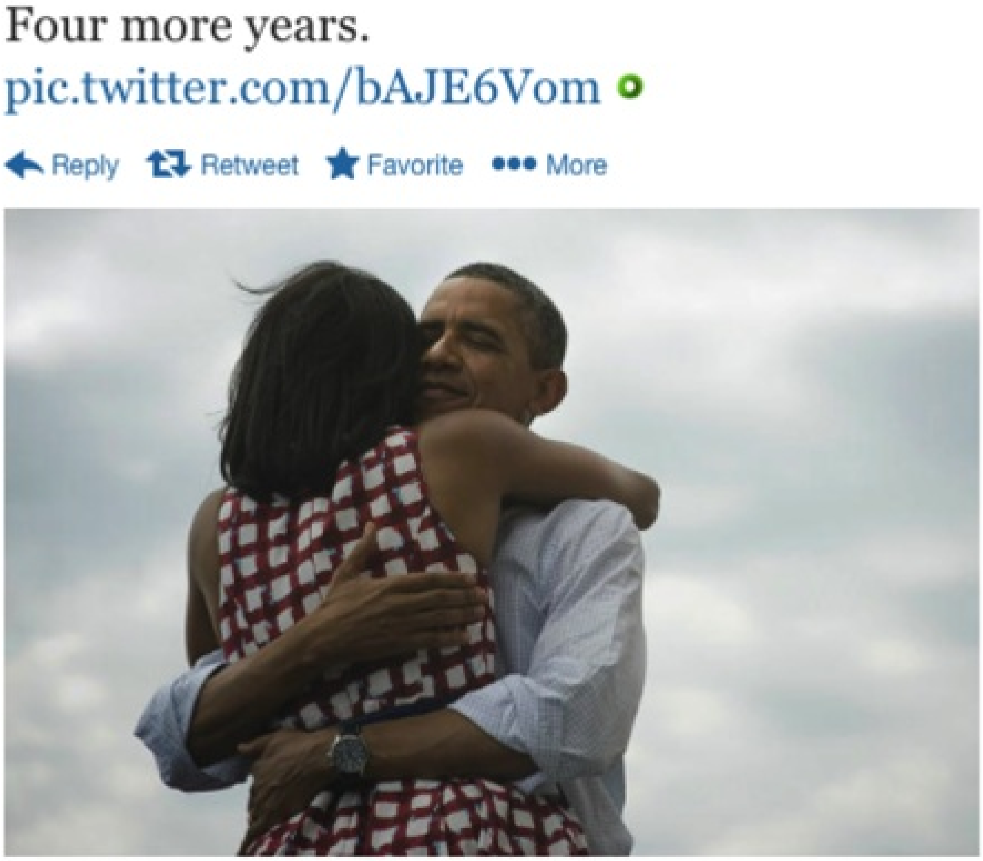
Digital Historical Appraisal

As I suggested earlier, we rarely stop to appreciate the complex analytic assessment that occurs during our appraisal of archival materials. Who has the time to acknowledge systematically the reams of materials that we reject before we stumble across the items we deem as possessing evidentiary value? Archivists would explain that evidentiary value derives from basic archival principles such as *respects du fonds*, which assures us that the original order of the materials has been preserved. As my fellow presenters discuss further, the endeavor of organizing archival materials is an intellectual, subjective exercise. The decision-making process to arrange and describe a collection contributes to the collection’s contextualization and influences historians’ access to materials.

Contextualization also applies in the digital environment, although the risk of separating a digital record or artifact from its provenance raises the stakes considerably. Whereas historians feel comfortable assessing a collection defined in linear feet, they have yet to find reliable methods for assessing collections defined in terabytes. The absence of visual cues thanks to cold, monolithic servers and hard drives, however, need not deter historians from gaining greater intellectual control over a digital collection. We simply need new methods to apply historical appraisal.

There are two elements vital to conducting an appraisal: scope and provenance. As with analog materials, we want to consider what digital materials have been selected for an archive or collection. By determining which items were kept and which ignored or discarded, we can begin to construct the contextual boundaries, or scope, of a collection. Besides the selection of materials, we also must account for whether one can trace data back to their point of origin, what archivists call provenance. For materials that were originally in an analog format, we would want to know their original archival location, whereas for born digital information we may want to know under what conditions data are generated. Without such information, or metadata, digital records pose a greater risk of becoming de-contextualized, that is, they have the potential to lose the value they may possess according to their relationship to other records. Just as we would be wary to trust a lone document that has been divorced from its archival folder or box, a digital item without metadata places additional strain on validating its authenticity and trustworthiness.

Think of how we might treat Obama's tweet if we didn't have the unique markers that signal that he was the author, or the time and location from which the tweet was sent.

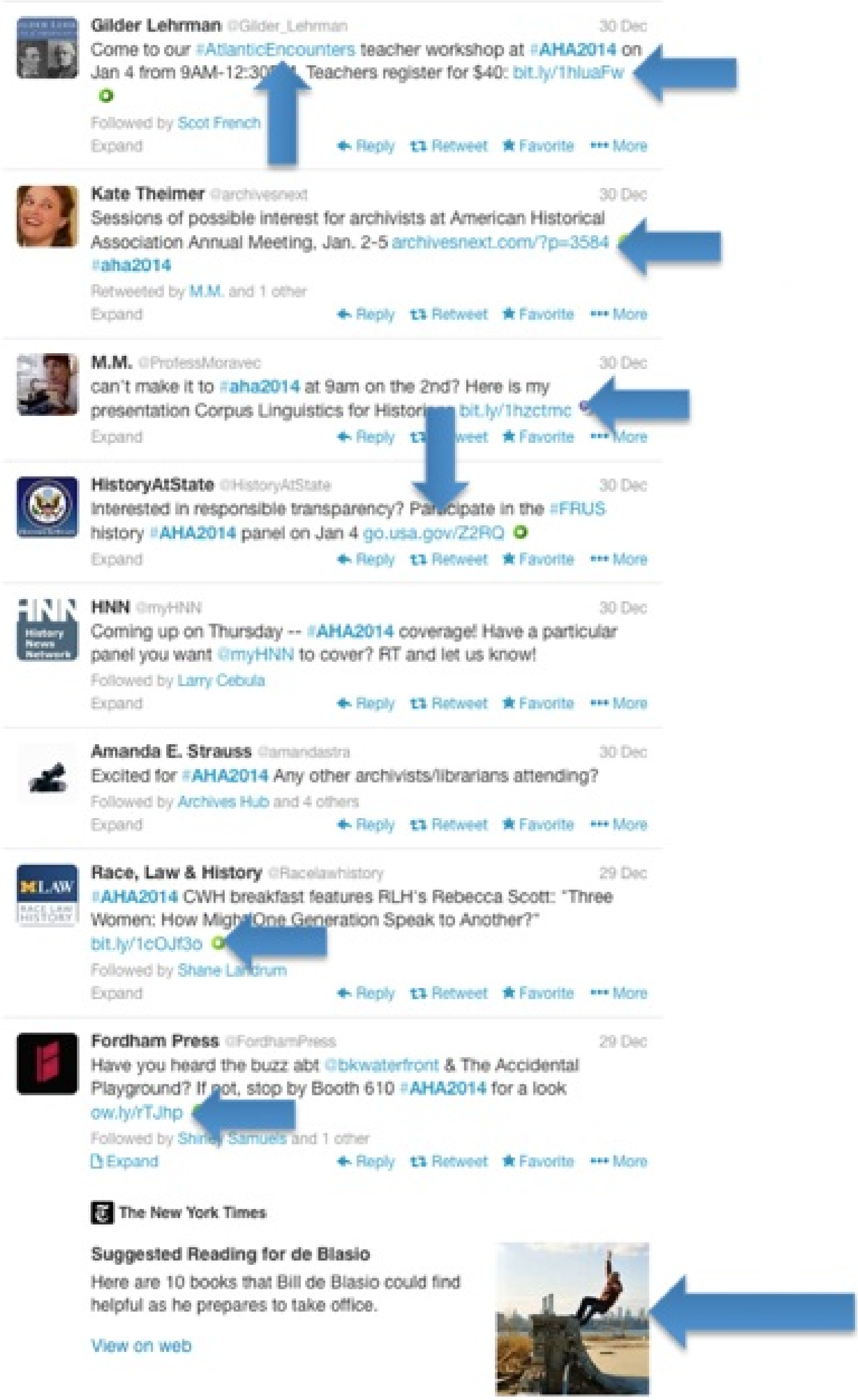


Four more years.
pic.twitter.com/bAJE6Vom

Figure 5: Obama's Tweet Minus Provenancial Metadata

Would we still trust it as reliable evidence? Consider also the provenance of a tweet. Although we may have a record of a tweet's original time posting and possibly geographic point of origin, tracking its distribution in subsequent retweets, quotes, and conversations introduces a host of challenges that requires a combination of precision, deftness, and creativity in critical thought.

We can also expand our discussion of provenance beyond the tweet in question. Consider that the historical value of tweets often comes from referencing events, conversations, and websites that have a provenancial record *outside* the Twitter communication stream.



Gilder Lehrman @Gilder_Lehrman 30 Dec
Come to our #AtlanticEncounters teacher workshop at #AHA2014 on Jan 4 from 9AM-12:30PM. Teachers register for $40: bit.ly/1hluaFw
Followed by Scot French

Kate Theimer @archivesnext 30 Dec
Sessions of possible interest for archivists at American Historical Association Annual Meeting, Jan. 2-5 archivesnext.com/?p=3584 #aha2014
Retweeted by M.M. and 1 other

M.M. @ProfessMoravec 30 Dec
can't make it to #aha2014 at 9am on the 2nd? Here is my presentation Corpus Linguistics for Historians bit.ly/1hzctmc

HistoryAtState @HistoryAtState 30 Dec
Interested in responsible transparency? Participate in the #FRUS history #AHA2014 panel on Jan 4 go.usa.gov/Z2RQ

HNN @myHNN 30 Dec
Coming up on Thursday -- #AHA2014 coverage! Have a particular panel you want @myHNN to cover? RT and let us know!
Followed by Larry Cebula

Amanda E. Strauss @amandastra 30 Dec
Excited for #AHA2014 Any other archivists/librarians attending?
Followed by Archives Hub and 4 others

Race, Law & History @Racelawhistory 29 Dec
#AHA2014 CWH breakfast features RLH's Rebecca Scott: "Three Women: How Might One Generation Speak to Another?" bit.ly/1cOJf3o
Followed by Shane Landrum

Fordham Press @FordhamPress 29 Dec
Have you heard the buzz abt @bkwaterfront & The Accidental Playground? If not, stop by Booth 610 #AHA2014 for a look ow.ly/rTJhp
Followed by Shirley Samuels and 1 other

The New York Times

Suggested Reading for de Blasio
Here are 10 books that Bill de Blasio could find helpful as he prepares to take office.

View on web

Figure 6: Sample Twitter Stream from #AHA2014. Arrows Point to External Links and Media

In other words, historians may be interested not just in the provenance of the tweet, but the vast corpus of materials such as websites, reports, slides, and other digital materials associated with those tweets.

There are researchers in other disciplines who are conducting experiments to answer these very questions in information science fields such as socioinformatics and alt-metrics. My point is that historians must play a role in this research, as we realize that the digital media of today becomes the artifacts of tomorrow. Participation begins with what historians do best, applying a critical framework to appraise historical data.

Thankfully, we are beginning to see examples of what digital historical appraisal may look like. Pragmatically, the methods of appraisal, and even the modes for explaining these methods will depend on the size of the collection and the nature of the project. At times, appraisal may require algorithmic computations measuring particular elements of the data. One such example worth noting is the *Trans-Atlantic Slave Trade Database*. The lead historian on the project, David Eltis, wrote an essay entitled "Coverage of the Slave Trade" that outlines the quantitative methods employed to estimate the data set's completeness. Drawing upon accepted evidence from the field, Eltis contends that the nearly 35,000 Trans-Atlantic voyages documented in the database can help scholars "infer the total number of voyages carrying slaves from Africa," by concluding that the database represents "some trace of 81 percent of the vessels that embarked captives."[15] (Note that the author elected to communicate his appraisal of a complex database using an "analog" format –the essay– that ought to remind us that visualizations may require just as much written explanatory text as a scholarly article or monograph.) In short, relying upon a combination of statistical analysis and historiography, the developers generated a claim about their database's representativeness. The interplay between historiography, appraisal, and digital modes of representations will require much more consideration as we continue to shape digital historiography.

Digital Historiography

At its core, digital history has enhanced the capacity of historians to investigate the past at different levels of inquiry. By considering the concepts of scale and appraisal in tandem, my hope is that the field will move toward a pragmatic approach to conducting digital history. The size and scale of information will determine, in the end, the mode of inquiry and the results that are possible.

In previous work, I labeled this pragmatic approach to conducting history *digital historiography*, a term that has had some bearing on the title of the AHA session. I defined digital historiography as "the interdisciplinary study of the interaction of digital technology with historical practice."[16]

This definition provided an opening to consider digital history at every stage of production by encouraging practitioners to consider how digital historical understanding should determine which theories and methods to adapt for a given pursuit.

Digital historiography recognizes that historical understanding possesses fundamentally different qualities in analog versus digital environments. Instead of adapting the historiographical questions of yesterday to the tools and methods of today, we ought to recognize that digital history will yield new understanding, new modes of inquiry that can complement our Newtonian tendencies to want to explain causes and effects.

In short, we can characterize digital historiography as mediating among the different levels of quantum history. By refining our ability to appraise historical data, we will become adept at moving back and forth among the levels. In Twitter terms, this would be the negotiation of going from a single tweet, to a circumscribed network of tweets such as a professional society or state, to the national and even global expanse of tweets, and back again.

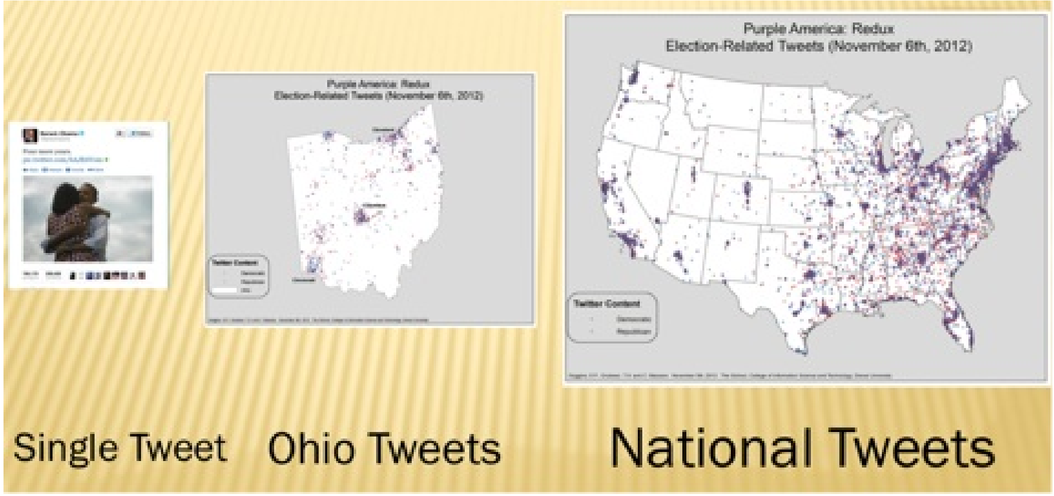


Figure 7: Spectrum of 2012 Election Day Tweets

How does each level inform the other and where are the connections that permit us to trace activity within the network?

In terms of the Ngram Viewer and similar tools, a query produced at a big data level may compel us to return to the underlying sources for additional close reading and analysis. In their macro-analysis of Victorian literature, Fred Gibbs and Dan Cohen suggest developing a method that can accommodate both the macro and micro sets of data:

> Any robust digital research methodology must allow the scholar to move easily between distant and close reading, between the bird's eye view and the ground level of the texts themselves…. The hybrid approach we have briefly described here can help scholars determine exactly on which books, chapters, or pages to focus, without relying solely on sophisticated algorithms that might filter out too much. Flexibility is crucial, as there is no monolithic digital methodology that can be applied to all research questions.[17]

Gibbs and Cohen are correct in asserting that any viable methodology will require "flexibility" in moving among the different levels of data. What their preliminary findings do not explicate is *how* the digital historian achieves such flexibility. What sort of intellectual rigor, whether represented by shared standards, practices, or theories, must be in place to provide the integrity necessary to sustain an argument from one level to the next?

The answer, as I hope my talk begins to outline, can be found by returning to the relationship between historians and archivists. Only by understanding one another's domain can we begin to bridge the disciplinary divide that will enable us to pinpoint the bits of data or texts that warrant our attention. In other words, we need to develop creative methods for reducing a million books down to a more manageable set of materials.

Historians and archivists, therefore, ought to concern themselves with the areas of transition, the connections *between* macro datasets and those that can be consumed at a human level.

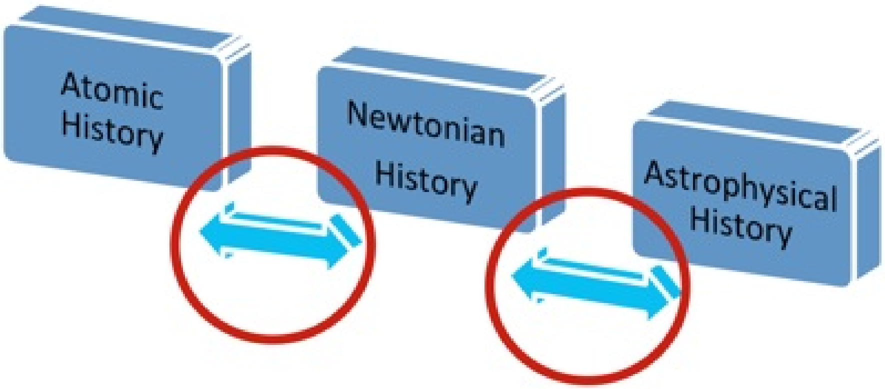


Figure 8

Liu writes: “[T]he interpretive or analytical methods at the two ends of the scale, macro and micro, are anything *but* seamless in their relationship…. It may be predicted that one of the next frontiers for the digital humanities will be to discover technically and theoretically how to negotiate between distant and close reading.”[18] This challenge requires knowledge of how digital information is managed, organized, and made accessible as well as deep mastery of relevant historiographical matters. It requires the ability to contextualize datasets at scales that historians may not be accustomed to analyzing. In short, it requires coordination among historians, archivists, and other information professionals. By appraising more precisely the size, scope, and completeness of a given collection of historical data, we will begin to construct a pathway towards posing and answering some of our most complex and enduring questions about the human condition.

Originally published by Joshua Sternfeld on January 20, 2014. Revised for the *Journal of Digital Humanities* in August 2014.

References

Aiden, Aviva Presser, and Jean-Baptiste Michel. *Uncharted: Big Data as a Lens on Human Culture.* Riverhead Hardcover, 2013.

Black, Alan, Christopher Mascaro, Michael Gallagher, and Sean P. Goggins. “Twitter Zombie: Architecture for Capturing, Socially Transforming and Analyzing the Twittersphere.” In *Proceedings of the 17th ACM international conference on Supporting group work*, 229-38, 2012.

Eltis, David, and Martin Halbert. “The Trans-Atlantic Slave Trade Database.” Emory University, 2009. http://www.slavevoyages.org/tast/index.faces.

Gibbs, Frederick W., and Daniel J. Cohen. “A Conversation with Data: Prospecting Victorian Words and Ideas.” *Victorian Studies* 54, no. 1 (2011): 69-77.

Goggins, S.P., T.H. Grubesic, and C. Mascaro. “Election 2012 – Election Day Partisan Tweets across the USA and in Ohio.” Drexel University, 2012. http://www.groupinformatics.org/election2012b.

Hitchcock, Tim, Robert Shoemaker, and Clive Emsley. “The Proceedings of the Old Bailey: London’s Central Criminal Court, 1674 to 1913.” University of Hertfordshire and University of Sheffield, 2013. http://www.oldbaileyonline.org.

Library of Congress. “Subject to Change.” In *Wise Guide*: Library of Congress, 2010. http://www.loc.gov/wiseguide/aug10/subject.html.

———. “Update on Twitter Archive at the Library of Congress.” 2013. http://www.loc.gov/today/pr/2013/files/twitter_report_2013jan.pdf.

Liu, Alan. “The State of the Digital Humanities: A Report and a Critique.” *Arts & Humanities in Higher Education* II, no. 1-2 (2011): 8-41.

Michel*, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, William Brockman, The Google Books Team*, et al.* “Quantitative Analysis of Culture Using Millions of Digitized Books.” *Science* (2010).

Netz, Reviel, and William Noel. *The Archimedes Codex: How a Medieval Prayer Book Is Revealing the True Genius of Antiquity’s Greatest Scientist*. Da Capo Press, 2009.

Nunberg, Geoffrey. “Counting on Google Books.” *The Chronicle Review* (2010). Published electronically December 16, 2010. http://chronicle.com/article/Counting-on-Google-Books/125735.

Osterberg, Gayle. “Update on the Twitter Archive at the Library of Congress.” In *Library of Congress Blog*, edited by Erin Allen. Washington, D.C.: Library of Congress, 2013. http://blogs.loc.gov/loc/2013/01/update-on-the-twitter-archive-at-the-library-of-congress/.

Randall, Lisa. *Knocking on Heaven’s Door: How Physics and Scientific Thinking Illuminate the Universe and the Modern World*. HarperCollins, 2012.

Sternfeld, Joshua. “Archival Theory and Digital Historiography: Selection, Search, and Metadata as Archival Processes for Assessing Historical Contextualization.” *The American Archivist* 74, no. Fall/Winter (2011): 544-75.

[1] Gayle Osterberg. “Update on the Twitter Archive at the Library of Congress.” Library of Congress Blog, 2013, http://blogs.loc.gov/loc/2013/01/update-on-the-twitter-archive-at-the-library-of-congress/. Accessed July 19, 2014. ↩

[2] Library of Congress, “Update on Twitter Archive at the Library of Congress,” (2013), http://www.loc.gov/today/pr/2013/files/twitter_report_2013jan.pdf. Accessed July 19, 2014. ↩

[3] Citing a single “original” tweet is an exercise in online sleuthing. In theory, the tweet may be accessed at @BarackObama’s profile — https://twitter.com/BarackObama — although the user may have to wade through over 12,000 tweets to locate it. Obama’s Organizing for Action references the tweet here: https://twitter.com/BarackObama/status/266031293945503744. The image used for this presentation was accessed January 2014, but the tweet seems to continue to have life. According to Twitter’s “Golden Tweets” it has been retweeted 810,000+ times and favorited 300,000+ times: https://2012.twitter.com/en/golden-tweets.html. ↩

[4] I owe much of my (admittedly coarse) understanding of quantum mechanics to Lisa Randall, *Knocking on Heaven’s Door: How Physics and Scientific Thinking Illuminate the Universe and the Modern World* (HarperCollins, 2012). ↩

[5] http://dictionary.reference.com/browse/HISTORY?s=t. Accessed July 19, 2014. ↩

[6] For typical examples of both, see the Register of Africans from the schooner “Virginie” found in the Trans-Atlantic Slave Trade Database,

http://www.slavevoyages.org/tast/resources/images.faces;jsessionid=2E7D3E3977482C7211A842 and a murder trial held on July 4, 1730: http://www.oldbaileyonline.org/images.jsp?doc=173007040011. Accessed July 19, 2014. ↩

[7] For the complete story of this fascinating discovery, see Reviel Netz and William Noel, *The Archimedes Codex: How a Medieval Prayer Book Is Revealing the True Genius of Antiquity's Greatest Scientist* (Da Capo Press, 2009). ↩

[8] An explanation of which can be found here: Library of Congress. "Subject to Change." Wise Guide, 2010, http://www.loc.gov/wiseguide/aug10/subject.html. Accessed July 19, 2014. ↩

[9] "Update on Twitter Archive at the Library of Congress." ↩

[10] David Eltis and Martin Halbert, "The Trans-Atlantic Slave Trade Database," (Emory University, 2009), http://www.slavevoyages.org/tast/index.faces. Accessed July 19, 2014. ↩

[11] Tim Hitchcock, Robert Shoemaker, and Clive Emsley, "The Proceedings of the Old Bailey: London's Central Criminal Court, 1674 to 1913," (University of Hertfordshire and University of Sheffield, 2013), http://www.oldbaileyonline.org. Accessed July 19, 2014. ↩

[12] This argument first appeared in the article that introduced the concept of "Culturomics": Jean-Baptiste Michel* et al., "Quantitative Analysis of Culture Using Millions of Digitized Books," *Science* (2010). A more comprehensive discussion of the National Socialist censorship campaign can be found in the authors' follow-up book: Aviva Presser Aiden and Jean-Baptiste Michel, *Uncharted: Big Data as a Lens on Human Culture*, (Riverhead Hardcover, 2013). Chapter 5. ↩

[13] See, for example, Geoffrey Nunberg, "Counting on Google Books," *The Chronicle Review* (2010), http://chronicle.com/article/Counting-on-Google-Books/125735. Accessed July 19, 2014. ↩

[14] Alan Liu, "The State of the Digital Humanities: A Report and a Critique," *Arts & Humanities in Higher Education* II, no. 1-2 (2011). 21. ↩

[15] David Eltis. "Construction of the Trans-Atlantic Slave Trade Database: Sources and Methods." http://www.slavevoyages.org/tast/database/methodology-02.faces. Accessed July 19, 2014. ↩

[16] Joshua Sternfeld, "Archival Theory and Digital Historiography: Selection, Search, and Metadata as Archival Processes for Assessing Historical Contextualization," *The American Archivist* 74, no. Fall/Winter (2011). ↩

[17] Frederick W. Gibbs and Daniel J. Cohen, "A Conversation with Data: Prospecting Victorian Words and Ideas," *Victorian Studies* 54, no. 1 (2011). p.76. ↩

[18] Liu, "The State of the Digital Humanities: A Report and a Critique." 27. ↩

About Joshua Sternfeld

Joshua Sternfeld has served since 2009 as a Senior Program Officer at the National Endowment for the Humanities Division of Preservation and Access in Washington, D.C. Prior to his arrival at the Endowment, Josh was Assistant Director and Postdoctoral Scholar for the UCLA Center for Information as Evidence and Information Studies Department from 2007-2009. He holds a B.A. in History from Princeton University and a Ph.D. in History from UCLA (2007), specializing in modern European cultural history and jazz studies. Josh has presented, taught, and published on the theoretical and pedagogical attributes of digital historiography. His publications have appeared in The American Archivist and Digital Humanities Pedagogy: Practices, Principles, and Politics.

A Distinction Worth Exploring: “Archives” and “Digital Historical Representations”

by Kate Theimer

In the original presentation of these papers at the AHA session, I was the final speaker on the panel, and so my talk was framed as a response to and expansion of the points made by the previous speakers.

In preparing for the panel “Digital Historiography and Archives” at the 2014 meeting of the American Historical Association, I had my usual trepidations about how the other speakers and the audience would frame their conception of “archives.” In writing my talk I read an article Josh had written for an archival journal in 2011[1] and was pleased to see his careful usage of the phrase “digital historical representations” as an umbrella term covering some of the resources presented by archives, as well as a range of products from other sources.

In discussing archives with historians and other humanities scholars, I often feel somewhat pedantic in my continual emphasis on the meaning of words.[2] But after all, words represent concepts and perceptions of reality, and if those words aren’t clearly communicating what we intend, then it’s hard to achieve meaningful progress. The approach I chose for my remarks at the digital historiography session was to illustrate the points the other speakers had made about the importance of questioning, understanding, and articulating the context of creation of digital historical representations by discussing the differences between different types of digital information sources created and used by historians—many if not most of which are often all referred to as “archives.” In all of these cases the context of the creation of the information sources is critical to understanding the problems that may be inherent in that source and which the researcher should take into consideration. I am not a historian, but I would think that understanding why and how an information resource was created—that is to say, its context—is more valid than ever in digital historiography.

Most readers will be familiar with what for lack of a better term I’ll call “traditional” archives—that is, primarily paper-based (or non-digital) largely unique materials, brought together in repositories in aggregations either created by the originating organization or person, or by a third party, such as a scholar, manuscript dealer, or the repository itself (as in special collections). Appraisal and selection of such materials is a multi-dimensional process with many factors involved, often including political influence, censorship on the part of the creator or collector, resource limitations on the part of the repository, random chance, and acts of God. How and why the materials on our shelves end up there is not always a straightforward story and one that is usually not captured in detail in the public description of the materials. How the materials were aggregated and for what purpose is usually described at some level in the finding aid, but documentation in this area can be sporadic. I would guess most archivists believe—rightly or wrongly—that metadata fields like “Custodial History,” “Appraisal, Destruction and Scheduling Information,” and “Administrative/Biographical History” are not valued by most users. Even among historians I’m not sure how often they are of interest, or at least how often historians ask the archivist for more information if the finding aid is skimpy in this regard.

Again, that’s “traditional” physical archival materials, represented digitally by descriptions in online finding aids, catalog records, etc. For these materials, I think what has changed for historians in the modern digital age is the increased expectation—and reality—that more descriptive information about materials will be made available online, and also the ability to easily create their own digital copies with digital cameras and smart phones.

Next we have collections of digitized analog historical materials—sometimes called “digital archives.”

These may be topically based—assembled from holdings of many repositories, like the William Blake Archive or the Wilson Center Digital Archive. Or they may be all from one repository—as in the recently launched FRANKLIN site, which provides online access to digitized collections from the Franklin D. Roosevelt Presidential Library and Museum. These collections may be created by archivists, librarians, historians, passionate amateurs, nonprofit organizations or for-profit companies. Because these digital historical representations are created by such a wide range of sources, it's critical to know about the context of these collections—including who assembled them, what their purpose was, and what criteria they used.

Often when historians are talking about archives, when I probe to see what they mean, it is these kinds of collections they are referring to. In her paper Katja observed that it's important to know where the individual original materials are located and where they fit in their archival context and that is certainly true. But it's also important to understand where materials fit in the context of the new digital collection. On what basis were items added to this collection? Why were some items excluded? To what extent is what's being presented a subset of what's available? Where does the metadata come from? How was it created and reviewed? As with online finding aids for physical collections, what is being accessed in this kind of digital collection is a surrogate—a description of that object or aggregate created by a person to represent it. Even a scanned image of a document is a surrogate, although hopefully an accurate one. Descriptions and metadata can be subjective and also subject to errors.

It seems to me as if these kinds of collections—or "digital archives" as they're commonly called, would raise a host of questions in terms of digital historiography—some similar to those presented by online information for "traditional" archives, but many others that are different.

Yet a different kind of aggregate, also sometimes called "digital archives" are groups of born-digital materials as opposed to digital surrogates of analog originals. These types of aggregates, kept together because they come from a single source or creator, reside primarily within archives and special collections repositories, and consist of records created or received by an organization in the course of business, maintained by them and transferred to their associated archival repository. The electronic records created by the Census Bureau and transferred to the National Archives are an example of this kind of aggregate. Another example can be found in the equivalent of the "papers" of a person or family, such as Salman Rushdie collection at Emory, which contains the contents of his personal computers. For these kinds of aggregates archives have most of the same kinds of issues with selection, appraisal, and custodial history as they do with non-digital materials, but with additional issues raised by their digital format related to reliability and authenticity as well as how to provide access.

And last but not least, you can have assembled collections of born-digital materials—yet another category of what are termed "digital archives." The September 11 Digital Archive, created by the Roy Rosenzweig Center for History and New Media, is a good example of this type of collection. In this case—and also with the Internet Archive—the collection serves a critical function: acquiring born-digital materials that might not otherwise survive. Many born-digital materials are more fragile than their analog counterparts for various reasons, and so some of these collections are similar in function to special collections libraries, which pull together valuable individual items for preservation. It's also worth noting that in digital collections, copies of materials can reside in more than one collection. For example, in the September 11 collection there are copies of documents created by the New York City Fire Department (Incident Action Plans). Presumably there are also copies of these born-digital records being transferred to the official repository for the municipal records of New York City. These kinds of "digital archives" combine the issues related to assembled collections—that is, the necessity of exploring who is creating them, for what purpose and using what methods— and those concerns related to born-digital materials as far as preservation and authenticity.

Coming back to the term "digital historical representations," I'm happy to see this broader term being used in discussions about "archives" and digital historiography. Many products that could fall into this category

—such as databases and sources like Google Books—would be removed one step (or more than one step) too far to be categorized as “archives.” I would consider these as separate intellectual products created from archival sources. And, indeed, in a way, so are any of the collections in which copies of archival materials are removed from their original context and “re-mixed” to be part of a new creation—a new “digital archives” like Valley of the Shadow, to use a classic example. In fact, in a pre-digital era analogous versions of the scholarly products mentioned here (other than databases) would still have existed, I think, and been called something other than “archives”—they would have taken the form of exhibits, edited volumes of letters or printed collections of documents, assembled and edited by historians or other sources. The question of why the word “archives” has been adopted to refer to collections of materials is one for a different discussion, but I do think it’s worth noting that this co-opting of the word does seem to be a rather recent development.

I hope the efforts discussed in this session encourage more rigorous assessment of digital historical representations and will result in a greater understanding and appreciation of what makes archives distinct from these other kinds of products. I often fear that this appreciation and understanding is being lost as fewer historians work with “old-fashioned” physical archival collections, and do most of their work online, where it is easy to think that all digital collections are the same. The value of the collections of materials preserved in archives often lies in the relationship of the records to each other—what’s called the archival bond—which means that the whole is greater than the sum of the parts. As a whole, the materials provide evidence about the activities of their creator or the person or organization who brought them together.

Discussions of digital historiography and the archives should be a two way street. It was heartening to see archival concepts such as appraisal and provenance being discussed at an AHA session and so seeing information flow from the archival literature to that audience. It is unclear what kind of awareness most historians have of archival theory or practice. Anecdotal evidence provided by many archivist colleagues suggests that such knowledge is, at best, uneven.[3] In return it is certainly also the case that digital historiography, that is the study of the interaction of digital technology with historical practice, can inform the work of the archival profession.

The papers from this session discussed how technology has changed the way historians do their work, and certainly it has also effected the way archivists do our work as well. Among the most significant of those ways is in the increased workload placed on archivists to create descriptions and digital copies to share online, to find ways to collect and preserve digital materials, and of course, to actively connect with the public via the ever widening world of digital tools and social media. Digital technology has also increased the user base for archival resources, meaning that the connection between our historian users and archivists is more diluted than it was in the past. In prioritizing our work and establishing our practices, archivists are trying to meet the needs of the broadest range of users. In so doing, it’s possible that the more specialized needs of historians—if indeed they are different from other users—are not being met. We need to keep an ongoing dialog between our two professions to ensure that we’re all working together as effectively as possible to support the historical enterprise.

Originally published by Kate Theimer on January 20, 2014. Revised for *Journal of Digital Humanities* August 2014.

[1] Joshua Sternfeld. “Archival Theory and Digital Historiography: Selection, Search, and Metadata as Archival Processes for Assessing Historical Contextualization.” *American Archivist* Fall/Winter 2011, 544-575. ↩

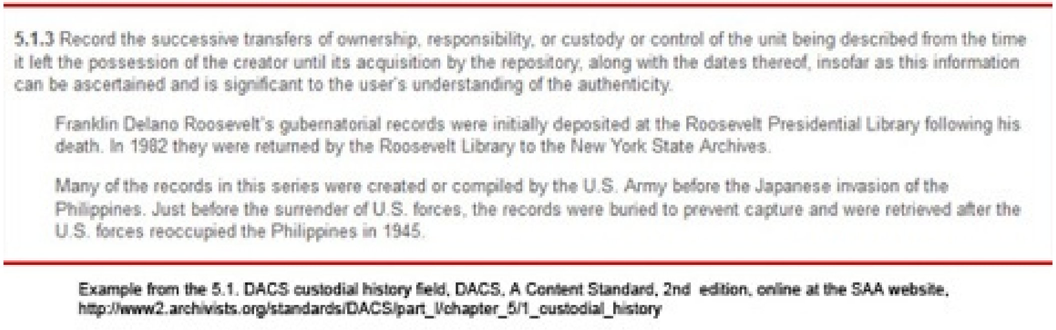
[2] See, for example, Kate Theimer. “Archives in Context and as Context.” *Journal of Digital Humanities* Vol. 1, No. 2 Spring 2012, http://journalofdigitalhumanities.org/1-2/archives-in-context-and-as-context-by-kate-theimer/. ↩

[3] The need for greater communication between historians and archivists was discussed in the concluding chapter of Francis Blouin and William Rosenberg, *Processing the Past* Oxford University Press, 2012. ↩

Provenance Meets Source Criticism

Katharina Hering

(This is a slightly revised version of the paper from which I spoke at the AHA. I added links, references, a few images, and an introductory paragraph. I also changed the title.) Archival and historical theory and methodology emerged in the late 19th century in a related historical and disciplinary context. Yet, to the degree that archives and history have evolved as separate disciplines and professional fields,[1] they tend to be treated as separate traditions. As Joshua Sternfeld and the other panelists have emphasized, however, critical digital historiography brings together elements of both archival and historical theory and methodology. The archival principle of provenance and the historical tradition of source criticism, in particular, complement another in underscoring the importance of providing context for documents, records, collections, archives, and digital historical representations. Traditionally, the archival concept of provenance refers “to the individual, family, or organization that created or received the items in a collection,” and the principle of provenance suggests that records originating from the same source should be kept together, and should not be interfiled with records from other sources to preserve their context.[2] The Canadian archivist Laura Millar emphasizes the importance of recognizing provenance as the key element of archival arrangement and description.[3] At the same time, Millar argues for an expanded understanding of provenance as a combination of *creator history, records history, and custodial history*. Her broadened concept of archival provenance is based on a comparative analysis of the concepts of provenance in museology, archaeology, and in archives, all of which have slightly different traditions and meanings. Millar further argues that a respect de provenance offers a more useful and realistic basis for a broader contextualization of records than the principle of the respect des fonds,[4] which has traditionally shaped descriptive practice, especially in government archives.



5.1.3 Record the successive transfers of ownership, responsibility, or custody or control of the unit being described from the time it left the possession of the creator until its acquisition by the repository, along with the dates thereof, insofar as this information can be ascertained and is significant to the user's understanding of the authenticity.

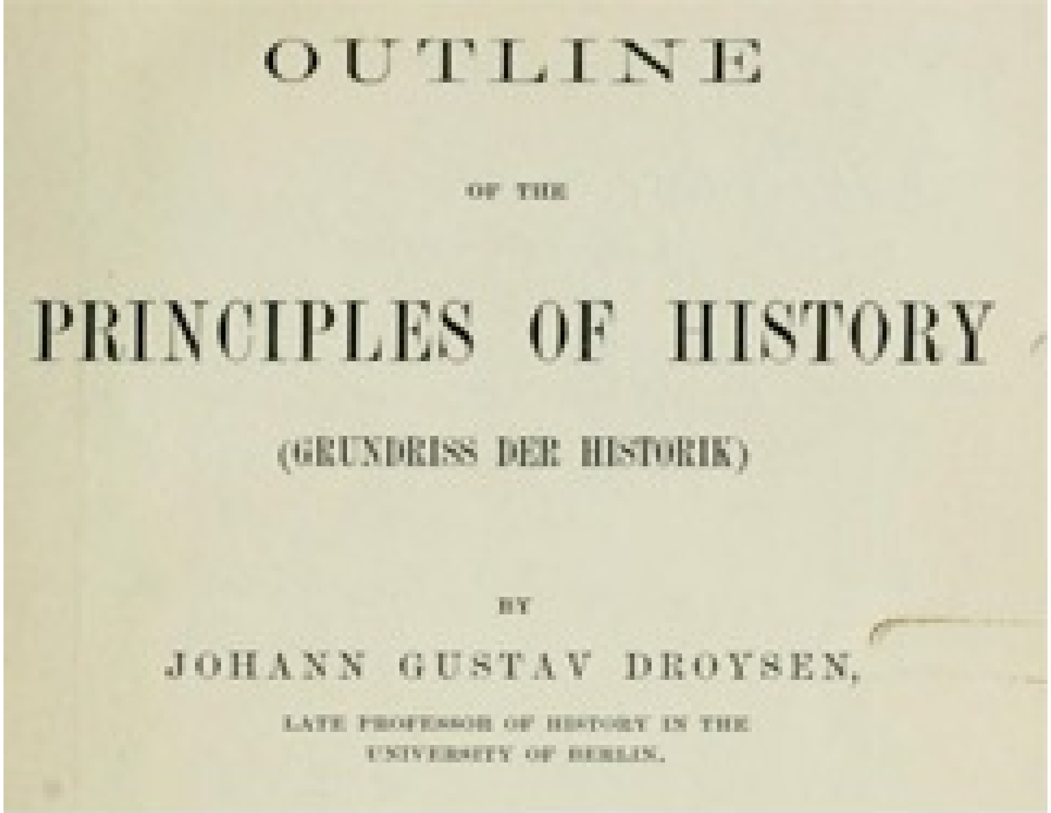
Franklin Delano Roosevelt's gubernatorial records were initially deposited at the Roosevelt Presidential Library following his death. In 1982 they were returned by the Roosevelt Library to the New York State Archives.

Many of the records in this series were created or compiled by the U.S. Army before the Japanese invasion of the Philippines. Just before the surrender of U.S. forces, the records were buried to prevent capture and were retrieved after the U.S. forces reoccupied the Philippines in 1945.

Example from the 5.1. DACS custodial history field, DACS, A Content Standard, 2nd edition, online at the SAA website, http://www2.archivists.org/standards/DACS/part_I/chapter_5/1_custodial_history

Example from the 5.1 DACS custodial history field, DACS, A Content Standard, 2nd edition, online at the SAA website.

While emphasizing the importance of provenance as a key archival principle, Millar—among others—acknowledges that there is a discrepancy between aspiration and practice: While archivists generally agree that providing information about the provenance of records and collections is critical, this understanding is not always reflected by archival practices. While many archives and archivists make an effort to provide detailed information about the provenance of specific records or collections, the fields in archival finding aids, or in catalog records, where information about provenance is supplied, often remain sparsely populated, or empty.[5] In absence of a systematic study of existing attitudes and practices, the reasons for this remain speculative. It may be because information about provenance is not considered relevant enough, or because this contextual information is not available, or because it was not transferred from institutional documentation to finding aids. Whether or not users care about this information is a another, related question, which sparked an interesting discussion on *Archives Next* in 2012 – the consensus seem to have been: it depends.[6] The lack of information about the provenance of collections, or individual items, is exacerbated in digital archives and collections, or collections of digital historical representations. As Joshua Sternfeld has highlighted, items that become part of digital collections can easily get detached from their original collection context, and in that process, existing information about the original provenance of the item frequently gets lost. This can also happen with digital collections that are removed from their original creation context. Just as in many physical archives, the contextual information about the provenance of digital collections, or digital objects that are part of digital collections, may not have been collected in the first place. Supplying information about provenance in digital archives is also more complicated due to the massive scale of many collections, and due to the fact that one has to distinguish between the provenance of the original record, item, or collection (if it was a physical object that has been digitized), and the provenance of the digital historical representation, or collection of digital historical representations. Thus, digital collections often require additional layers of information about provenance. The reasons for the lack of adequate contextual information about provenance of digital historical representations are complex, and there are many challenges to providing this information or metadata – technical, conceptual, institutional, and economical. These challenges, however, do not diminish the importance and the ethical obligation for providing adequate contextual information about items, collections, or digital historical representations. “Is it ethical for archivists to detach digital items from their archival context in order to make them more ‘digital friendly’ and more accessible to meet needs of some users?” Jane Zhang asks.[7] I believe that the tradition of source criticism in historical theory and methodology complements the archival principle of provenance, and that it adds an important historical perspective to the archival obligation to provide information about the provenance of digital historical representations. Source criticism has been an integral part of historical theory and method since the concept was introduced by Prussian historian and philosopher Johann Gustav Droysen, and then further developed by historical theorist and philosopher Ernst Bernheim.



OUTLINE

OF THE

PRINCIPLES OF HISTORY

(GRUNDRISS DER HISTORIK)

BY

JOHANN GUSTAV DROYSEN,

LATE PROFESSOR OF HISTORY IN THE
UNIVERSITY OF BERLIN.

Outline of the Principles of History, 1893, title, cropped from the Internet Archive copy.

Outline of the Principles of History, 1893, title, cropped from the Internet Archive copy.

Droysen in his *Outline of the Principles of History* (1st German edition in 1867, first English ed. in 1893) defined the task of criticism as to "determine what relation the materials still before us bears to the acts of will whereof it testifies."[8] Droysen distinguished several elements of criticism: criticism determines the genuineness and authenticity of a source, it addresses the development from earlier to later forms of the materials, it questions the validity of the information and the source, and the correctness of the information. Source criticism also includes the critical analysis of the information in the source itself: which events and developments does it reflect, how was the description influenced by its contemporary context, who was the author, how did it relate to other sources of the time?[9] Droysen emphasized that the outcome of the critical analysis was not the "exact historical fact," but rather the ability to place "the materials in such a condition as renders possible a relatively safe and correct judgment."[10] Historian and philosopher Ernst Bernheim later expanded and specified Droysen's theory. [11]

Lehrbuch

der

Historischen Methode

und der

Geschichtsphilosophie.

Mit Nachweis der wichtigsten Quellen und Hilfsmittel
zum Studium der Geschichte.

Von

Ernst Bernheim,

Professor der Geschichte an der Universität zu Greifswald.

Bernheim, Lehrbuch, 1908, cropped
from the Internet Archive copy.

Bernheim, Lehrbuch, 1908, cropped from the Internet Archive copy

Especially relevant in this context is Bernheim's distinction between internal and external source criticism – the critical analysis of the content of the source versus the analysis of the creation context or provenance of the source. Based on Laura Millar's definition of provenance, one could also understand external source criticism – when applied to individual records as well as collections — as the investigation of creator history, records history, and custodial history. Certainly, the tradition of source criticism has to be situated in its time and historical context in the late 19th century, and the notion of a "source" itself can be problematic, as Michael Kramer highlights in his perceptive criticism of the essentialist connotation of a "source" as something that can be exploited by the historian (see Michael Kramer, "Going Meta on

Metadata”). Still, the tradition of source criticism and the archival principle of provenance complement another in highlighting the importance of collecting and providing contextual information for sources or collections. Combined with a broadened understanding of provenance, the tradition of source criticism can support archivists, historians, librarians, digital humanists, and others with developing a set of questions and a vocabulary that can aid the analysis and description of digital collections, or digital historical representations alike. Digital source and resource criticism—in a revised, modern version, which was developed and theorized by the late Swiss historian Peter Haber[12]—as well as provenance are important elements of critical digital historiography. But how can such an ambitious goal, framed by archival and historical theory, be implemented? What are the challenges at specific institutions? What are possible practical approaches for archivists, historians, librarians, and others to collaborate to collect and provide adequate, critical, contextual information about digital historical representations? How can contextual information that historians gather in the course of their research make their way into archival finding aids or catalog records? How can the contextual knowledge about collections that archivists have gathered help historians with developing source critical analyses? What can researchers and archivists do if they find that digital historical representations lack adequate contextual information? How can source criticism lead to resource and database criticism? How can information professionals, including archivists, and researchers, including historians, voice their concerns when faced with a lack of contextual information provided by big commercial databases, such as JSTOR, Ancestry.com, EBSCO, and, of course, Google, over which they have no control? How can collaborative teams of history liaison librarians, archivists, and historians develop portals that help with supplying contextual information about the provenance of specific collections made available by commercial databases that allow a critical, informed use of these resources?[13] When confronting these complex questions, the principle of provenance and the tradition of source criticism can provide a familiar basis for historians and archivists, while serving as guides for developing new, collaborative models of providing archival and historical context for digital surrogates. Originally published by Katherina Hering on January 20, 2014. Revised for *Journal of Digital Humanities* August 2014.

[1] Robert Townsend, *History’s Babel: Scholarship, Professionalization, and the Historical Enterprise in the United States, 1880-1940.* Chicago; London: University of Chicago Press, 2013. ↩

[2] See the definition of provenance in SAA’s online glossary of archival and records terminology. Last visited: July 15, 2014. ↩

[3] Laura Millar, “The Death of the Fonds and the Resurrection of Provenance: Archival Context in Space and Time,” *Archivaria* 53 (Spring 2002): 1-15. ↩

[4] See the definition of fonds in SAA’s online glossary of archival and records terminology. Last visited: July 15, 2014. ↩

[5] Depending on the content and cataloging standards used, the relevant fields differ. In APPM, which was replaced by DACS, existed a field for information about provenance. In DACS, the relevant fields to include information about provenance are administrative/biographical history; custodial history, and immediate source of acquisition, in ISAD (G), the field for custodial history is called archival history. Related and relevant MARC fields are 541, 545 and 561. Provenance is not one of the 15 elements of the simple Dublin Core metadata element set, but part of qualified Dublin Core. Thanks to Kate Theimer for suggesting the clarification regarding the relevant fields. ↩

[6] See: Kate Theimer, “Debate: The majority of users don’t care about provenance. They just want access to information,” *Archives Next*, May 2012. Last visited: July 20, 2014. ↩

[7] See Jane Zhang's reference to her 2012 paper on *Archival Context, Digital Content, and the Ethics of Digital Archival Representation* in the 2012 discussion about provenance in *Archives Next*. ↩

[8] Johann Gustav Droysen, <em>Outline of the Principles of History (Grundriss der Historik)</em>. Translated by E. Benjamin Andrews. New York: Howard Fertig Inc. edition, 1967. Originally published in English in 1893. German original published in 1867, p. 20. 1893 edition available from the Internet Archive. Last visited July 30, 2014. ↩

[9] Droysen, *Outline*, ibid., paragraph 26-36, pp. 21-26. ↩

[10] Droysen, ibid. ↩

[11] Ernst Bernheim, *Lehrbuch der historischen Methode und der Geschichtsphilosophie*, 1st German ed. 1889. Bernheim's Lehrbuch was never translated into English. The 5th and 6th 1908 Leipzig edition is available from the Internet Archive. ↩

[12] See: Peter Haber, *Digital Past*, München: Oldenbourg 2011, among other publications. ↩

[13] Beyond Citation, a very interesting project exactly along these lines that "seeks to encourage critical thinking about academic databases and their impact on research and scholarship," has recently been been developed as part of the Digital Praxis seminar at CUNY. See Eileen Clancy's Feb. 10, 2014 comment on our blog posts. ↩

Going Meta on Metadata

Michael Kramer

I once joked to an archivist that all I really do as a historian is add *meta*-metadata to the archival database.

What I meant was that if we understand metadata—the information that accompanies artifacts—as not merely descriptive, but also already on its way to interpretation, then what is historical scholarship but a further extension of this elaboration of the evidentiary record? The joke was intended as humorous (at least for the geeky among us) because typically historical work gets separated from, and often privileged over, archival labor. The archive is there, historians mistakenly believe, solely to be mined by them alone for scholarly production. Archivists, of course, know better. The archive serves many other purposes than just fodder for historical inquiry. And as archives move into the digital domain, whether through digitization of artifacts or with born-digital materials, the archive's many uses expand even more. Might it be more accurate, then, my joke implied, to reverse the hierarchy of archive and historical research as we move into the digital realm? Perhaps the digital archive becomes the final product—something that now absorbs historical research and publishes into it—rather than just a storehouse for information hidden in the stacks; and perhaps in this new context, historical findings become nothing but a new field added to the interactive, accessible, crowdsourced, hyperlinked records management system!

To call what historians do meta-metadata is to go meta, as it were, on the intersections between what archivists and historians do, from how they conceptualize their practices to how they work with (and sometimes against) each other. In the new online spaces where the digital archive meets digital history, the relationship between these two professions takes on new and unexpected possibilities—and tensions. We will need to think carefully about how the digital returns us to buried institutional wounds that date back in the United States to the 1930s, when archivists and historians parted ways in their professional affiliations.[1]

The American Historical Association, which sponsored our panel, was itself at the center of the controversial split at that time. So there is a history with which to reckon here. But the digital also presents new dilemmas and opportunities that require attention.

The concept of historiography—the history of historical writing itself, from the record of what has been said about a topic to the articulation of debates over interpretations to the awareness of different methods that historians have used to analyze their sources—might play a key role in helping both archivists and historians to navigate changes wrought by the digitization of both of their fields. What our panelists call "digital historiography" offers a means of more critically connecting archival theory and professional archival practices to historical theory and professional historical practices. More of this term historiography and its significance in a moment. But first I wish to make a few brief comments about the short but sharp presentations by Josh Sternfeld, Katja Hering, and Kate Theimer.

It is my hope that these mini essays themselves eventually find their way into the archives, whether those be the storified live tweet archives of the session, the official AHA archives, and certainly, to speak more metaphorically, the memory archives of your minds. For these presentations are important contributions to what we might call the archive of how we are to understand digital technologies as we all—historians, archivists, students, professionals, citizens—increasingly dip our virtual toes in digital waters and sometimes find ourselves with the distinct feeling that we might soon be drowning.

Our roundtable today is particularly focused on three keywords: digital, historiography, and the archives. There are other keywords that crop up as well in Josh, Katja, and Kate's comments: surface, context, provenance, metadata, scale, appraisal, calibration, evidence, criticism. These are worthy of further elaborations too. But for now I want to hone in on the three words in our roundtable's title as the crucial ones.

The Digital

First, the digital. Josh Sternfeld's wonderful suggestion is that we imagine a "quantum history" that moves beyond the scale of a sort of Newtonian historical middle ground in which evidence and convincing argument have largely stable properties and interact through mostly predictable and agreed-upon relationships. Going micro and macro have already become part of the historical repertoire, but I think Josh is correct to suggest that the digital affords new opportunities to revisit those strategies of analysis and think about how we might toggle, if you will, among them. As he puts it, we might "calibrate" our narratives in new ways. One way to do this is for historians to think more critically about the provenance of our sources, a term to which Katja draws our attention. Rather than treat evidence as transparent access to the truth, we might consider the how's and why's of the origins of our "evidence" from their starting point right through to the generations of archival creators, maintainers, and interpreters. We should also remember that the archival objects are themselves often surrogates (a wonderful term that Kate invokes), or what media studies scholars call "remediations," of older documentary forms. The digitized book or photograph is not, at its material level, the original version, but rather a copy of it rendered in a new medium of bits and bytes, data and code. And of course, historians might pay more attention to the ways in which many of the so-called "original" documents—whether they be paper, audio recordings, film, or photographs—are but representations of the past, and usually partial or distorted ones at that. The archive is our record of the past, not the actual thing itself.

The digitization of this archival record might, at first, seem like a further retreat, yet another step removed, from history as it happened. But this "remediation" of archival materials into a new form is also a tremendous opportunity to consider the past with more sensitivity, to pay far greater attention to how we access and analyze history itself. Digitization, in this sense, asks us to slow down rather than the more common assertion that it enables us to speed up. The transition into the computational domain can accelerate certain kinds of availability and manipulation of archival materials, but it also provides a glimpse of the process by which we preserve archival holdings and use them to endow the messy chaos and vast diversity of the past with meaning, structure, continuity, order, and significance.

Historians and archivists alike have long been aware of the ways in which archives shape our very perceptions of the past. In this sense, the traditions of historical and archival thinking, the methodologies and methodological debates of these fields, have as much to bring to the new digital domains as digital technologies do to these respective professions. For instance, digital history's "quantum" turn, as Josh Sternfeld asks us to imagine it, offers an opportunity to revisit the notion of the *long durée* and the macro-historical Braudelian ideas of the *Annales* School.[2] Similarly, digital technologies might allow us to rethink the concepts of "microhistory." We might fix our attention on what Josh calls the "dark matter" of cultural minutae as they lurk in the vast world of data, networks, and digital infrastructures. Similarly, Katja's notion of "source criticism," drawn from her reading of the nineteenth-century work of Johann Gustav Droyseen, bespeaks the productive effort to recover past methodological and historiographical approaches in order to grapple with new digital challenges and opportunities. So too does Kate's insistence that we use the word "archives" with care and precision—and even perhaps not use it at all when its digital incarnation diverges fundamentally from archival purposes of preservation.

Archives

Now to that loaded word: archives. Why the popularity of this term? Why also the pressure on this word now? This pressure comes not merely from the new representational and methodological qualities of digital technologies; it also comes from a longer running inquiry into the power of representation, particularly of the state's uses of official recordkeeping to wield power, secure legitimacy, obscure facts, and govern its citizens (and so too those deemed non-citizens) by tracking them as individuals or transforming them into abstract demographic statistics. Though I grant the legitimacy of the position among certain archivists such as Kate that perhaps we should employ the term archive with care and precision, in a limited rather than expanded way, I think that the critical inquiry into the power of the archive, its ability to wield knowledge in service of hierarchy and control, also asks us to crack upon the term—and the archive itself.

The convergence of mediated forms within the digital domain—the collapse of archives themselves, the curation of their holdings, the research conducted within and across them, the conversations they inspire, and the publications inspired by and grounded in their artifacts and materials—asks us to open up what we call an archive rather than close it down. Particularly in an era when both corporations and the government are using large, "official" digital archives for data mining of human individuals, we need to assert that the archive should be understood as a kind of commons, not a tool of totalized mastery and secretive information. Transparency and privacy must be renegotiated in the new technological structures of the digital archive. And despite the pressures of standardization and homogenization that the digital demands in order to function, we need to insist that it adjust to and respect more quirky, messy, and unofficial modes of archivization as well.

A long-running digital dream, dating back to the 1940s has been to assemble the information of the world into one linked archive of sorts.[3] But perhaps the digital can also, in ways we do not quite understand yet, enable strange, heterogeneous, and different kinds of memory and history too. The balancing of universal standards against particular contexts will become key here. The underlying architecture of the digital is, after all, archival in nature. Whether in the "chunking," "shells," "kernels," and the modularity of particular software or in the use of modular databases or in the entire functional necessities of the Internet, the digital medium balances universal protocols of containerization, record-keeping, and networked interfacing against singular and distinctive uses and spaces of activity within the digital domain.[4] How we develop digital archives so that they are not a one-size-fits-all platform, how we fight the urge for standardization while still harnessing the power of interconnectivity in the digital arena—this becomes one of the great challenges for archivists and historians alike.

Historiography

Historiography, the last term from our roundtable's title, provides a good starting point for archivists and historians to try to broaden what the archive might be and do in the digital domain. As I have already suggested, there are some real differences between archivists and historians that we need to consider. After all, what archivists by training are taught to call objects, artifacts, documents, and items historians, by contrast, refer to (sometimes with far too much unquestioned essentialism and also with an air of exploitative plundering) as sources. For archivists, the goal is to preserve, describe, and provide access to archives for a broad range of users, from professional historians to private archive owners to the public at large. For historians, the goal is, most of the time, to "mine" archival material for interpretation. These two approaches—archivist's and historian's—can go together, of course, but they only do so through slightly different imaginings of the stuff itself in the archives and the uses to which it should be put.

The digital domain brings these somewhat different goals, archival preservation and access, on the one hand, and historical interpretation and analysis on the other, into the same space. As the presentations richly suggest, archival theory brings to digital historiography far more sophisticated modulations between varying levels of scale and appraisal, text and context, and source preservation and source criticism. To these contributions, historians might add an additional element that draws upon the traditional use of the term historiography to signify the history of historical inquiry itself, which is to say the history of historical interpretations of the past and the attention to varying methodologies that have produced historical findings.

For historians, historiography signals a shift from “primary” sources—often archival ones—to “secondary” sources—or the historical arguments, interpretations, and interventions that use the archives to mount claims about the past. Of course, this distinction is rather artificial: today’s “secondary” sources often become tomorrow’s “primary” ones; what seems in the archive to offer direct access to the past is itself fundamentally representational and interpretive in nature already; and of course the very placement of certain materials in archives and the exclusion of other materials speaks to the power of the archive itself to shape what counts as history and what is delegitimized. But nonetheless, the term historiography points to these very complexities. It reminds us to always remember that the past arrives to us through layers of interpretation. We might even say that the past *is* interpreted. It is not relative or invented or a fiction. Certain things did occur and others did not. But rather it is messy and chaotic enough, multivalent and multifaceted in the extreme, that being aware of historiography makes us understand what the past is, and how archives both provide and deny access to it. Historiographical debates by their nature force us to be far more sensitive to competing versions of the past, to varying means and methods of making sense of the archival record, and to the ways in which history is no static thing even when its constituent elements get lodged and preserved in the metaphorical amber of the archive.

And then there is this: as we well know the digital tends to be far less static than prior mediums of preservation. So what, then, does it mean within the digital domain to address historiography when it is understood to be the collection of secondary sources and ongoing debates about a historical topic? It would mean, perhaps, rethinking the relationship between primary and secondary sources in new ways, not just going to the supposedly pure sources, fetishized as they are in the field of history. It would also mean that we might reconceptualize the preservation of historiography itself, that in the digital medium, we might link so-called primary and secondary sources in new, more fluid and dynamic ways that speak to the richness of their interconnections.

In shifting from thinking about digital history to digital historiography, there is a new kind of provenance at work, much as there seems to be in paying attention to the term in archival work and theory. We need to develop modes of preserving a historiographical sequence not of object ownership, but rather of interpretive ownership. This is, in some respects, a far more contested kind of ownership of course. Who “owns” what interpretation cuts close to the bone of prestige and status in the historical profession. In being so, in paying attention to how historiography gets digitized, we are forced to ask questions such as: which historians have looked at these archival things before? What sources have been ignored and why? What did historians have to say about archival sources and why? How did they temporalize them, contextualize them, conjoin them, or distort them? What methods and preoccupations, interests and worldviews, shaped their interpretations?

These questions are meant to suggest that within a historiographic context, within thinking about the history of historical interpretation, we need to grapple with continuities between and among generations of historians and also, of course, debates. We also need to think carefully about voices left out of these conversations and the kinds of questions and themes that drive them as well as the well-known, established voices. We need to confront the whole assemblage of the history of history, which is grounded not only in readings of primary sources in archives, but also readings of secondary sources previously understood as, in some sense, fundamentally outside the archives.

We might, in summation, even think of various historiographies themselves as archives. They may not be encased in walls or stacked in boxes on shelves, but they are, sure enough, constellations of materials brought together with a provenance secured and documented in literature reviews, encyclopedia entries, the background sections of articles and books, and in footnotes and endnotes. If primary sources exist as one kind of archive requiring more careful attention to methods of access and analysis, secondary sources are also an archive of sorts, brought together through interpretive practices, characterizations, and interventions in the field of history itself.

What a digital archive might do is provide a space for bringing these two kinds of archives into play with each other. It can, in Stuart Hall's sense of the word as the bringing together of disparate elements, "articulate" them to one another.[5] A digital architecture for a new imagining of the archive might be able to provide more dynamic linkages and movements among, on the one hand, materials being used as primary sources—put to service to represent the past as best as it can be factually reconstructed—and, on the other, materials being used more primarily as secondary sources.

A new kind of useful fluidity might emerge among linked open source archives and scholarship using the materials in those archives. The digital archive, with an expanded notion of what it does, has the opportunity for enriching history by more dynamically linking primary sources and their subsequent interpretations, and in doing so, of raising the question of what a source is exactly, and how we "appraise," to use Josh's term, the relationship of evidence to argument, sources to interpretations and ongoing conversations. Within new kinds of digitized settings, historiography can flourish as a key part of archives themselves and the historical narratives of the past they inspire.

In this sense, maybe history is just *meta*-metadata. Maybe that's not such a bad thing.

Originally published by Michael J. Kramer on January 20, 2014. Revised for *Journal of Digital Humanities* August 2014.

[1] Peter Novick, *That Noble Dream.* Robert B. Townsend, *History's Babel: Scholarship, Professionalization, and the Historical Enterprise in the United States, 1880-1940* (Chicago: University of Chicago Press, 2013), 181-182. ↩

[2] See David Armitage and Joanna Goldi, "The Return of the Long Durée: An Anglo-American Perspective," *Annales. Histoire, Sciences sociales* 69 (2014), for a polemical call to historians to use digital technologies in order to return to the project of the Annales School. ↩

[3] Among many examples, see Vannevar Bush, "As We May Think," *Atlantic Monthly* 176 (July 1945), http://www.theatlantic.com/doc/194507/bush; J.C.R. Licklider, "Man-Computer Symbiosis", *IRE Transactions on Human Factors in Electronics HFE-1* (March 1960), 4-11; and the work of Douglas Engelbart, "Toward Augmenting the Human Intellect and Boosting Our Collective IQ," *Communications of the ACM* 38, 30-32 (1995). Howard Rheingold, Tools For Thought: The History and Future of Mind-Expanding Technology (1985; reprint, Cambridge, MA: MIT Press, 2000). See also, Thierry Bardini, *Bootstrapping: Douglas Engelbart, Coevolution, and the Origins of Personal Computing* (Stanford, CA: Stanford University Press, 2000). ↩

[4] Tara McPherson, "U.S. Operating Systems at Mid-Century: The Intertwining of Race and UNIX," in *Race after the Internet*, eds. Lisa Nakamura and Peter A. Chow-White (New York: Routledge, 2012), 21-37; Alexander R Galloway, *Protocol: How Control Exists After Decentralization* (Cambridge, MA: MIT Press, 2006; Wendy Hui Kyong Chun, *Programmed Visions: Software and Memory* (Cambridge, MA: MIT Press, 2011); Andrew Blum, *Tubes: A Journey to the Center of the Internet* (New York: Ecco, 2012). Lev Manovich, "Database as Symbolic Form," *Convergence* 5, 2 (June 1999), 80-99. ↩

[5] See, for instance, Stuart Hall, "Race, Articulation, and Societies Structured in Dominance," *Sociological Theories: Race and Colonialism*(Paris: UNESCO, 1980), 305-345. ↩

About Michael Kramer

Michael J. Kramer holds a visiting assistant professorship at Northwestern University, where he teaches history, American studies, digital humanities, and civic engagement, and he works as an editor in the Design, Publications, and New Media Department at the Museum of Contemporary Art in Chicago. His book, The Republic of Rock: Music and Citizenship in the Sixties Counterculture, was published by Oxford University Press in 2013. He is the co-founder of the Northwestern University Digital Humanities Laboratory and is currently developing a multimedia project about the Berkeley Folk Music Festival and the history of technology and culture in the US folk revival. Additionally, he serves as director of the Chicago Dance History Project, a large-scale oral history and archival digital documentation of dance in the Chicago region, and he is the dramaturg for The Seldoms Contemporary Dance Company. He has written for numerous publications and blogs about art, culture, and politics at Culture Rover.

Review of The Johns Hopkins Guide to Digital Media (2014)

Alex Christie

Ryan, Marie-Laure, Lori Emerson, and Benjamin J. Robertson, eds. *The Johns Hopkins Guide to Digital Media*. Baltimore: Johns Hopkins UP, 2014. Print.

Current scholarly activity in digital media reflects a convergence of cultural and political critique with technological investigation, engagement, and practice, and the challenge to creating any guide or introduction to digital media that it risks codifying — and thereby diminishing — the diversity of approaches, methodologies, and theoretical approaches to be found. The *Johns Hopkins Guide to Digital Media*, edited by Marie-Laure Ryan, Lori Emerson, and Benjamin J. Robertson, resists collapsing the many and varied potential trajectories of digital media studies into singular narratives by weaving intellectual diversity and vibrancy throughout its representation of the disciplinary fabric of digital media studies.

Poignantly, Jussi Parikka captures the need for such a multivariate approach in his entry "History of Computers" when he writes: "There is just too much for a *single* history of the computer. Any history of computing becomes suddenly a metaquestion of how to write a history of such complexity" (249). By extension, digital media studies is best served when it resists representing affiliated scholarly activities through essential or normative practices; therefore, the entries in *The Johns Hopkins Guide to Digital Media* do well to reflect the intersections of multiple, interrelated inquiries and approaches that cohere along various, productive constellations. Readers of the *Guide* will find that many entries intersect in purpose and topic, even as they diverge along historical, theoretical, and methodological lines of inquiry.

With entries on media types ranging from sound and video games to code and poetry, *The Johns Hopkins Guide to Digital Media* surveys a recognizably diverse and evolving field. Entries, however, extend beyond media type elucidating key properties, such as immersion, mediality, and avatars. Other entries explore cultural and theoretical issues, ranging from cyberfeminism and gender representation to ontology and cognition. The *Guide* includes a combination of short and long entries by scholars whose work represents historical, empirical, theoretical, computational, and archival approaches. Taken together, this plurality of intellectual engagements opportunes multiple points of entry into the field, offering readers what the editors describe as "a GPS and a map of the territory of digital media, so that they will be able to design their own journey through this vast field of discovery" (xiii). To this end, each entry comes embedded with in-line references to other related entries in the *Guide* to facilitate the type of exploration and discovery the editors describe. The result is a highly accessible resource for both newcomers to digital media and seasoned scholars in the field looking for a snapshot of its current state. Students and teachers in search of a comprehensive guide will also find great pedagogical value in this book.

In addition to plotting related instances of digital media scholarship and creation, the *Guide* unpacks digital media, itself as a territory in which coordinates are enmeshed in humanities and social sciences research. At once historically aware and methodologically reflexive, the entries collectively situate the Guide's two key terms as complex zones of intellectual discovery. For example, in his entry on music, Aden Evans complicates neat compartmentalizations of the digital, suggesting: "Entirely digital music is out of the question, but the intersections of music and the digital are numerous and telling" (344). A similar troubling of the newness of new media comes from Jessica Pressman, who argues: "the work of the new is precisely what inspires us to reconsider the old and to recognize the intersections and convergent histories of old and new" (365). As a result, throughout the *Guide*, an awareness of digital media as comparative and hybrid in nature evolves.

Usefully, entries often situate contemporary technologies in relation to the historical and material practices that inform their development. Kristyn Leuner, for instance, traces the scroll bar to Pliny's documentation of turning papyrus into scrolls in "Book to E-Text", while Mark Nunes's entry unpacks the early modern development of postal routes and the Victorian-era telegraph as early instances of networking. Comparative approaches in other entries offer accessible examples of digital media concepts, such as Jake Buckley's documentation of the move from measuring to calculating time to explain the shift from the analog to the digital in "Analog vs. Digital." Likewise, Bethany Nowviskie draws from the everyday problem of searching for lost keys to offer a particularly lucid explanation of the difference between a heuristic and an algorithm, exemplifying the range of accessible material available throughout the book.

Complementing the many entries that unpack their topic's historical contexts, diverse disciplinary approaches are also brought to bear on digital media. In her entry "Graph Theory," Ryan demonstrates a rich combination of theory and practice as she works through the concept of the rhizome in light of the contrast between tree and network structures. Other contributors situate digital media in relation to important cultural and social contexts that inform their emergence, which they, in turn, influence, as well. Charles Ess's entry "Ethics in Digital Media" documents the evolution of the Pirate Bay from a file-sharing website to a Nordic political party advocating copyright reform. Brian Croxall similarly offers examples of political events documented in real time via microblogging, as well as mass social actions organized through social media platforms.

Throughout, contributors demonstrate not only the extent to which digital media is enmeshed in diverse sets of material practices, but also illustrate the key role media play in reshaping those practices. As Anna Munster writes of new materialist scholarship, "digital theorists, writers, media producers, and artists… [engage] the digital as a mode or cluster of operations in consort with matter, a way of materially *doing things* in the world" (330). In perhaps the most representative entry of such an approach, Matthew Kirschenbaum outlines "strategies or approaches for preserving bits and their contexts" in the digital preservation community, by unpacking the preservation of born-digital material as both a technical and cultural field of investigation — a move that becomes emblematic of Ryan, Emerson, and Robertson's editorial strategy (405).

The *Johns Hopkins Guide to Digital Media* represents a valuable and lasting contribution to the field of media studies by revealing current attitudes toward media as digital and material, preserved in bits while moving through multiple communities of practice, and key in unraveling the multiple entwinements between culture and technology. Comprehensive and accessible, there are an impressive 154 entries in the *Guide* that in combination offer a glimpse of the current state of scholarly work in digital media that is both detailed and broad.

The disadvantage to the multitude of entries, however, is that emerging scholars may find it difficult to trace nuanced distinctions between topics that are closely allied, but that represent distinctions between multiple modes of inquiry. For instance, the *Guide's* numerous entries on interactive media document different media types, but cover shared theoretical and methodological topics across those types, repeating content. These overlapping entries stand in contrast to those with the "Cyber" prefix, which demonstrate clear points of disciplinary convergence and divergence in the field. Indeed, such contrast speaks to ongoing discussions about the status of interactivity in digital media and digital humanities scholarship, identifying a rich zone of intellectual activity reminiscent, at least in part, of previous disciplinary discussions in the realm of cyber scholarship. Yet, the *Guide's* strength is its ability to negotiate these closely allied but historically distinct scholarly lines of inquiry.

One additional challenge the *Guide* faces is the lack of a keyword index. For example, terms such as gamification, ambient intimacy, and identity tourism represent highly focused topical areas, while entries such as "Cognitive Implications of New Media" are much more opaque. Tracing key terms and their intersections across related entries would make the *Guide* even more accessible.

Nonetheless, *The Johns Hopkins Guide to Digital Media* serves as an important scholarly reference, offering multiple points of entry into a complex area of intellectual activity. More than a comprehensive look at the detailed threads of inquiry in this field, it will serves as a key resource to which future students and scholars alike can turn for its representation of the current state of digital media studies.

About Alex Christie

Alex Christie is a doctoral candidate in English at the University of Victoria. He conducts research on 3D geospatial expression and scholarly communication for the Modernist Versions Project (MVP) and Implementing New Knowledge Environments (INKE) in the Electronic Textual Cultures Lab (ETCL). He is developing an open source toolkit for digital humanities pedagogy with grant funding from the Association for Computers and the Humanities (ACH); his dissertation traces experiments in rule-based literary expression across modernist poems and manuscripts.