Knowledge creation through recommender systems

Taylor Arnold
Department of Math & Computer Science, University of Richmond, USA

Peter Leonard
Digital Humanities Lab, Yale University, USA

Lauren Tilton
Department of Rhetoric & Communication Studies, University of Richmond, USA

Abstract

The way materials are archived and organized shapes knowledge production (Derrida, J. *Archive Fever: A Freudian Impression*. Vancouver: University of Chicago Press, 1996; Foucault, M. L'archéologie du savoir. Paris, France: Éditions Gallimard, 1969; Kramer, M. Going meta on metadata. Journal of Digital Humanities, 3(2), 2014; Hart, T. How do you archive the sky? *Archive Journal*, 5, 2015; Taylor, D. Save As. e-misférica, 9, 2012). We argue that recommender systems offer an opportunity to discover new humanistic interpretative possibilities. We can do so by building new metadata from text and images for recommender systems to reorganize and reshape the archive. In the process, we can remix and reframe the archive allowing users to mine the archive in multiple ways while making visible the organizing logics that shape interpretation. To show how recommender systems can shape the digital humanities, we will look closely at how they are used in digital media and then applied to the digital humanities by focusing on the Photogrammar project, a Web platform showcasing US government photography from 1935 to 1945.

Correspondence:
Lauren Tilton, University of Richmond, Department of Rhetoric & Communication Studies, 402-C Weinstein Hall, 28 Westhampton Way, VA 23173, USA.
E-mail:
ltilton@richmond.edu

1 Introduction

Recommender systems are procedures built into digital media that offer suggestions of content that may be of interest to a given user (Ricci *et al.*, 2011). These systems are now a common aspect of many popular Web sites. E-commerce sites such as Amazon and eBay suggest products related to recent searches or purchases (Sarwar *et al.*, 2001). The front pages of media sites like Youtube, Netflix, and Imgur suggest records that have recently spiked in popularity. Likewise, pages for particular content elements display thumbnails of similar videos and images in the sidebar (Davidson *et al.*, 2010). Social media sites suggest other users one might know based on matching self-reported metadata about education and occupation or using overlapping friend networks (Gupta *et al.*, 2013). At their best, these systems increase user discovery and engagement by exposing forgotten or unknown elements of a collection to a captive audience, while simultaneously making it easier to find specific content of interest.

To build recommender systems, Web-based companies utilize both the historical behavior of

doi:10.1093/llc/fqx035 Advance Access published on 11 August 2017

other users (collaborative filtering), as well as explicit user preferences and metadata associated with items in their collection (content-based filtering). Both of these approaches are made possible due to the large amount of user data that is frequently available to commercial Web sites (Melville *et al.*, 2002). Social networking sites require users to login before creating a profile or interacting with other users, and media sites require logging in before posting new content or commenting on current content. Some media sites, such as Flickr, offer only a very limited experience to anonymous users. Digital journalism Web sites often have explicit paywalls blocking unregistered users from accessing most content. E-commerce sites can typically be viewed anonymously but require registering to make purchases and incentivize being logging in while browsing. In nearly all of these examples, Web sites can track nearly every action a user takes on their Web site and tag this behavior to personally identifiable information such as credit card numbers, billing addresses, full legal names, and self-reported preferences.

The ultimate goal of the recommender systems employed by commercial Web sites is to make a profit for their associated companies. The exact time horizons and details may differ, but in most cases recommender systems strengthen eventual profits by improving user engagement and, where applicable, increasing the total volume of sales made on the Web site (Herlocker *et al.*, 2004). The presence of quantifiable metrics along with a large amount of user data makes the process of building recommender systems in this space amenable to predictive analytics (Breese *et al.*, 1998). One very well-known application was the Netflix Prize competition that ran from 2006 to 2009, where teams publicly competed to build predictive recommender systems in the hopes of winning a 1 million dollar prize (Bennett and Lanning, 2007). A technique combining a large collection of models into one model proved superior, and is still used today as a technique for training recommender systems (Zhou *et al.*, 2008).

Not limited to the commercial domain, recommender systems also offer the digital humanities a method for exploring and remixing humanities data. They build off of and extend work in archive studies and information science to harness metadata for discovery (Bates, 1989; Derrida, 1996; Foucault, 1969; Hart, 2015; Ingwersen, 1996; Kramer, 2014; Taylor, 2012). They allow users to reframe the archive, in turn allowing for new organizing logics that open up new questions and knowledge production. Yet, implementing recommender systems in the digital humanities involves a series of challenges.

First, digital archives and other online cultural heritage projects generally have limited, if any, explicit data about individual users. These sites want to reach as wide an audience as possible and do not want to burden users with creating accounts and having to login before accessing content. Billing information is rarely needed, as their content is freely available, and they do not engage in any e-commerce activities with rare exceptions. Online privacy regulations in both the USA and European Union also impose substantial extra technical costs for storing personally identifiable information (Rosen, 2012; Simitis, 1995). Second, sites hesitate to use information contained in user agent strings (i.e. Web browser and operating systems) or coarse IP-based location information, as this may conflict with their commitments to user privacy (Ramakrishnan *et al.*, 2001; Yen *et al.*, 2012). Even in cases where digital, public projects desire to provide usernames and login information, structural barriers such as expertise and funding can preclude this from going forward. Servers maintaining the backend of Web sites that host user-generated content require significantly more resources and demand constant maintenance, neither of which many projects have the necessary funding to allocate to these systems (Baeza-Yates and Ramakrishnan, 2008). Third, projects hosting user data may also be subject to additional legal issues such as responding to digital search warrants and content takedown requests (George and Scerri, 2007).

Even when there are data about users, challenges persist. While granular user-based data are rarely available to such projects, aggregated statistics such as page views, bounce rates, and average time on the site may often still be collected (Borgman, 2009). Yet, predictive analytics are still a challenge because it is often unclear how to train recommender systems in digital, public projects. A method trained to increase user engagement may come at the expense of the project's educational or research aims. Additionally, many of these projects, particularly digital archives, have a large ratio of items in their collection to the number of users that frequent the site over even a long time period. Social networking sites by design have one user per item, and e-commerce sites typically have many users per item due to the high carrying cost associated with stocking any individual item. So, even if digital public projects focus solely on a single quantifiable metric and use only aggregate user-based data, the sparseness of the data set will often lead to unsatisfactory results.

There are solutions. Recommender systems for public digital projects can use static algorithms based solely on item-based metadata. The long, dedicated work of archivists and librarians in contrast has provided many digital archives with extensive metadata about each and every item within an archival collection. The resulting metadata can be used to produce excellent item-based recommender systems (Aktas *et al.*, 2004). To show these methods in action, we turn to Photogrammar.

Photogrammar is a Web-based platform for organizing, searching, and visualizing the 170,000 photographs from 1935 to 1945 in the US Farm Security Administration and Office of War Information (FSA-OWI) collection. To build support for and to justify government programs during the Great Depression and the World War II, the FSA-OWI set out to document America and the successful administration of new government services. They produced some of the most iconic images of the era and employed prominent documentary photographers such as Arthur Rothstein, Dorothea Lange, Gordon Parks, and Walker Evans, all of whom shaped the visual culture of the era both in its moment and in American memory. Unit photographers were sent across the country. Over 170,000 negatives were sent to Washington, D.C. for processing. The Photogrammar project uses digital humanities methods including spatial, image, and text analysis to increase discoverability of a photography archive while also changing humanities scholarship in media studies, visual culture studies, and 20th-century US cultural history. See Fig. 1 for an example of the user interface displaying photography over a historical map.

Exploration and scholarship on the FSA-OWI collection have focused on the most prominent photographers such as Dorothea Lange and Walker Evans. To show thematically similar pictures to users, Photogrammar uses term frequency–inverse documentary frequency (TF-IDF) to suggest images whose captions bear a mathematical similarity to a given photograph on the screen. Approximately 80,000 images in the collection have captions, and TF-IDF uses this semantic signal as the base for a series of quantitative transformations to turn English words into points in an abstract space. The distance between these vectors can then be measured and used as a proxy for semantic distance in the captions.

The first transformation involved in our use of TF/IDF weights less common terms more heavily in its calculations, so that extremely common words (such as 'the' or 'a') are less influential than more specific words such as 'church' or 'store'. In our implementation of TF/IDF, we also perform a second transformation of the textual signal to optimize results. Although the photographs in the FSA/OWI collection provide an explicit location in their metadata (such as 'Chicago, Cook County, Illinois'), they often repeat this geographic information in their captions. A caption such as 'A man waits in line in Chicago, Illinois for bread.' thus contains a kind of in-band metadata that can skew TF/IDF results. The relative rarity of Chicago and Illinois in the entire corpus ensures that the captions with these terms will appear just as close to the original caption as those with 'bread' and 'line'. To avoid this locational bias, we subtract a list of all city, county, and state names from the caption before analysis. This approach is a kind of 'geospatial stoplist' that results in broad thematic similarity and reduces banal results. For example, the relative rarity of 'German' in the broader corpus ensures that photographs of German Americans, German social clubs, and similar content will appear highly ranked as results for a caption with that word. The result, as seen in Fig. 2, is a set of photographs that may otherwise never have been placed in conversation with each other. This allows users to explore

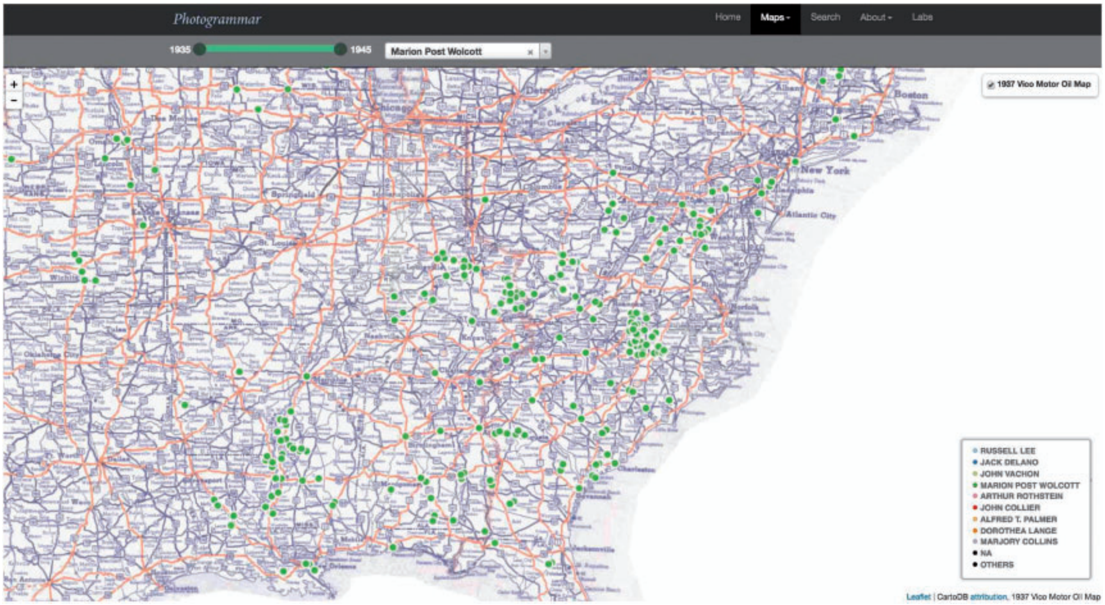


Fig. 1 Locations of photographs taken by Marion Post Wolcott for the FSA-OWI archive in the southeastern USA. Points are overlaid on the 1937 Vico Motor Oil map

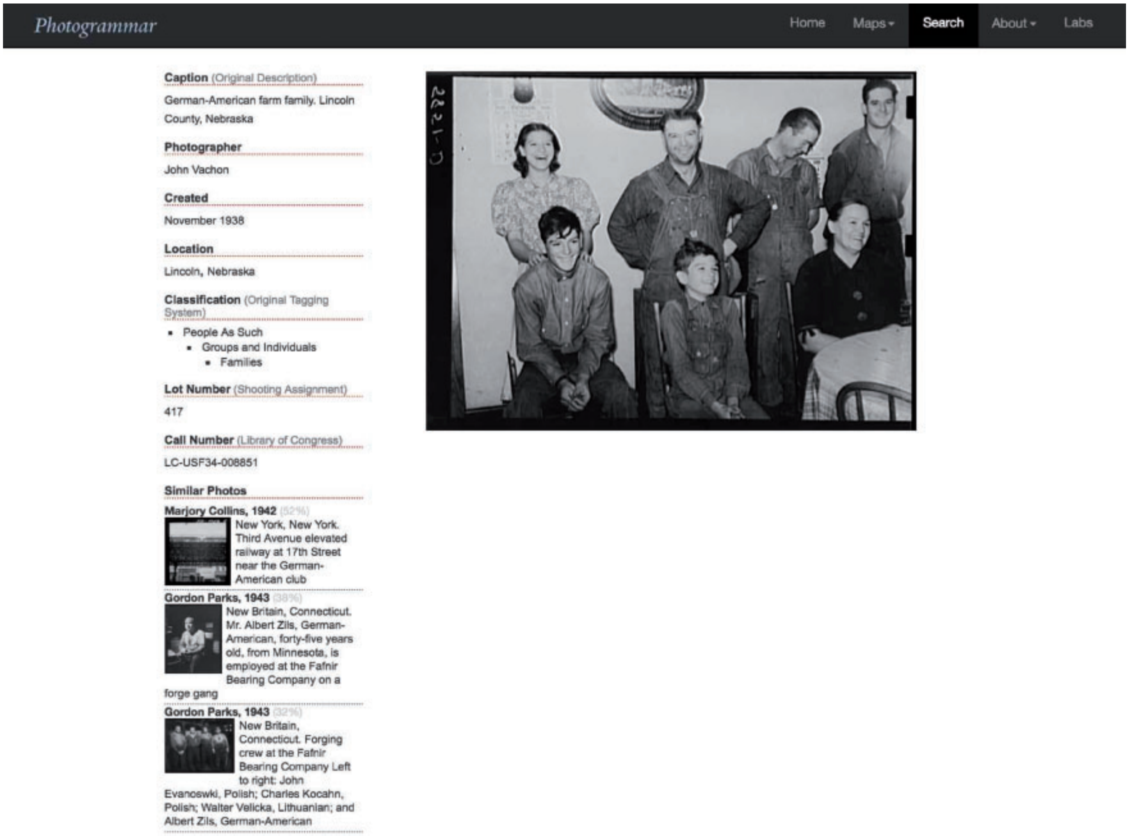
the archive in new ways shaped by the recommender system in turn decentering the continued focus on the most prominent photographers and places.

Another approach is to measure visual, as opposed to textual, similarity. Although the English language tokenizes trivially and yields an easily actionable data set for approaches such as TF/IDF, the pixel data of raster images are much more difficult to analyze and process. One emerging approach, most often identified with the Software Studies Initiative (Manovich *et al.*, 2012) and its software project ImagePlot, is to measure the mean color values of each image along a set of visual features such as hue, saturation, luminosity, and RGB values, for example. This approach has yielded aggregate patterns in such data sets as thousands of European Impressionist paintings (Manovich, 2015) and the covers of the fashion magazine *Vogue*. We have applied color analysis—brightness, frequency, hue, and saturation—to the approximately 1,600 color photographs. The color space laboratory, Fig. 3, is currently available on the Photogrammar Web site.

The averaged colormetric dimensions of the color images in the Photogrammar collection is used to build an exploratory visual recommender system. A user who is studying agricultural images of orange harvesting in Santa Fe, New Mexico might be shown images of oranges for sale in California that share the same RGB and saturation values as the original image, even if they were taken by a different photographer in a different year.

At the same time, the abstractions inherent in such a visual analysis involve trade-offs. Since the measurements are based on averages (specifically, the mean values) of all pixel data, there is a risk for artifactual or nonsensical results. For example, a woman wearing a blue dress standing in a yellow field would present the same average RGB values as a green forest. To engage more fully with the visual complexity of each image, we are turning to Convolutional Neural Networks (CNNs), a promising new field for this kind of applied 'machine vision'.

Benoît Seguin has shown the applicability of intermediate layers of CNNs (those between the initial



Caption (Original Description)

German-American farm family. Lincoln County, Nebraska

Photographer

John Vachon

Created

November 1938

Location

Lincoln, Nebraska

Classification (Original Tagging System)

- People As Such
  - Groups and Individuals
    - Families

Lot Number (Shooting Assignment)

417

Call Number (Library of Congress)

LC-USF34-008851

Similar Photos

Marjory Collins, 1942 (52%)
New York, New York. Third Avenue elevated railway at 17th Street near the German-American club

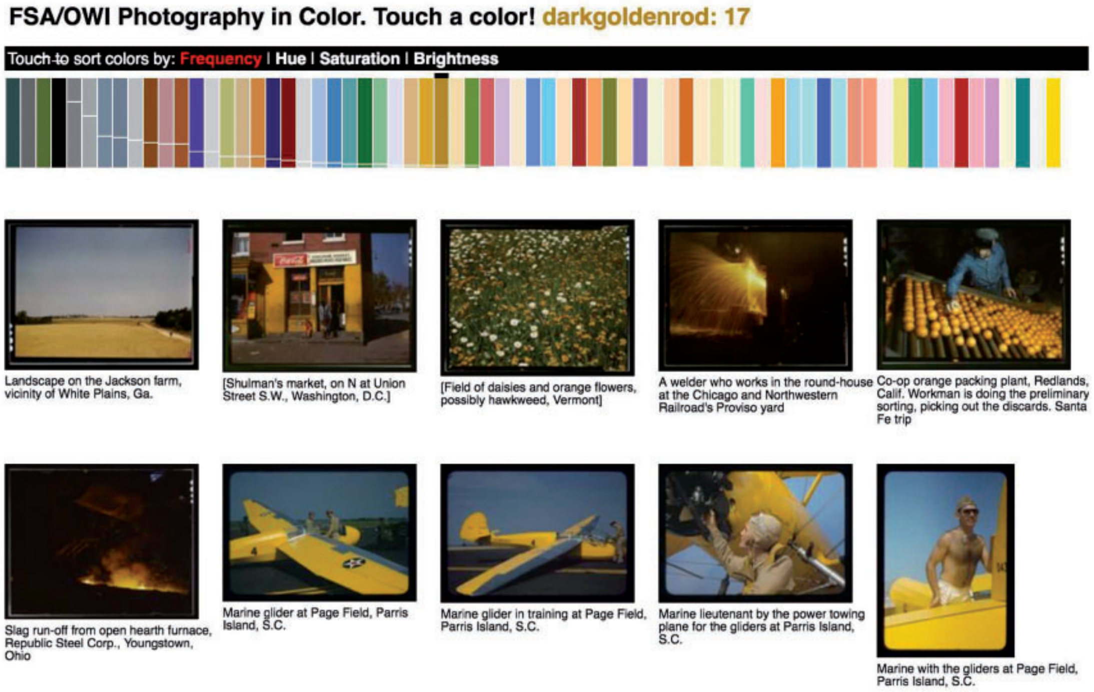
Gordon Parks, 1943 (38%)
New Britain, Connecticut. Mr. Albert Zils, German-American, forty-five years old, from Minnesota, is employed at the Fafnir Bearing Company on a forge gang

Gordon Parks, 1943 (32%)
New Britain, Connecticut. Forging crew at the Fafnir Bearing Company Left to right: John Evanoswki, Polish; Charles Kocahn, Polish; Walter Velicka, Lithuanian; and Albert Zils, German-American

Fig. 2 Item page from the Photogrammar Web site showing a photograph taken by John Vachon in 1938. Similar photos based on TF-IDF are displayed in the lower left-hand side of the page

primitives and the final image classification layer) for use in image similarity problems using cultural heritage material. Although the penultimate layers of these neural networks are difficult to describe intuitively, they can be 'thought to represent high-level characteristics directly usable for the classification tasks' (di Lenardo *et al.*, 2016). They capture a high-dimensional (usually in the thousands) representation of the image content, and can be used in much the same way as TF/IDF textual vectors to recommend similar content. We plan to implement such a neural network-based recommendation system on top of the Photogrammar collection in the near future.

Current and future recommender systems in Photogrammar offer an opportunity to construct alternative organizational structures within a digital, public archive. The linked metadata approach of the aforementioned projects offers great starting points, but only serves to reinforce the archival structure already exposed by extant search features and other user interfaces. In these cases, recommender systems do more than simply provide helpful suggestions. They produce a structural argument and provide a narrative that differs from other organization systems within a digital archive. Latent links between diverse parts of an archive are made to guide users to relatively unexplored parts of a large collection of items. Framing recommender systems as organizational structures highlights their importance as a means of knowledge production. The ways we construct knowledge shape the kinds of questions that can be answered, and recommender systems offer an opportunity to reshape and remix in the digital humanities.



FSA/OWI Photography in Color. Touch a color! darkgoldenrod: 17

Touch ~~to~~ sort colors by: Frequency | Hue | Saturation | Brightness

Landscape on the Jackson farm, vicinity of White Plains, Ga.

[Shulman's market, on N at Union Street S.W., Washington, D.C.]

[Field of daisies and orange flowers, possibly hawkweed, Vermont]

A welder who works in the round-house at the Chicago and Northwestern Railroad's Proviso yard

Co-op orange packing plant, Redlands, Calif. Workman is doing the preliminary sorting, picking out the discards. Santa Fe trip

Slag run-off from open hearth furnace, Republic Steel Corp., Youngstown, Ohio

Marine glider at Page Field, Parris Island, S.C.

Marine glider in training at Page Field, Parris Island, S.C.

Marine lieutenant by the power towing plane for the gliders at Parris Island, S.C.

Marine with the gliders at Page Field, Parris Island, S.C.

Fig. 3 Photogrammar color laboratory, showing the nearly 1,600 color photographs from the FSA-OWI archive. Here all photographs with a concentration of Dark Golden Rod are shown, including images of planes, oranges, and flowers

References

Aktas, M. S., Pierce, M., Fox, G. C. and Leake, D., (2004, November). A web based conversational case-based recommender system for ontology aided metadata discovery. In *Proceedings of the 5th IEEE/ACM International Workshop on Grid* Computing. IEEE Computer Society, Washington, DC USA, pp. 69–75.

Baeza-Yates, R. and Ramakrishnan, R. (2008, March). Data challenges at Yahoo! In *Proceedings of the 11th International Conference on Extending Database Technology: Advances in Database Technology*. New York, NY, USA: ACM, pp. 652–655.

Bates, M. J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Review*, **13**(5): 407–424.

Bennett, J. and Lanning, S. (2007, August). The Netflix prize. In Proceedings of KDD *Cup* and *Workshop*, Vol. 2007, p. 35, New York, NY USA: ACM.

Borgman, C. L. (2009). The digital future is now: a call to action for the humanities. *Digital humanities quarterly*, **3**(4).

Breese, J. S., Heckerman, D. and Kadie, C. (1998, July). Empirical analysis of predictive algorithms for collaborative filtering. *In Proceedings of the Fourteenth Conference* on Uncertainty in *Artificial Intelligence*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA pp. 43–52.

Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., Gupta, S., He, Y., Lambert, M., Livingston, B. and Sampath, D. (2010, September). The YouTube video recommendation system. In Proceedings of the fourth ACM *Conference* on Recommender *Systems*. New York, NY, USA: ACM, pp. 293–296.

Derrida, J. (1996). *Archive Fever: A Freudian Impression*. Vancouver: University of Chicago Press.

di Lenardo, I., Seguin, B., Kaplan, F. (2016). Visual patterns discovery in large databases of paintings. In Digital Humanities 2016: Conference Abstracts. Kraków: Jagiellonian University & Pedagogical University, pp. 169–172.

Foucault, M. (1969). *L'archéologie du savoir*. Paris, France: Éditions Gallimard.

George, C. E. and Scerri, J. (2007). Web 2.0 and user-generated content: legal challenges in the new frontier. *Journal of Information, Law and Technology*, **2**: 1–22.

Gupta, P., Goel, A., Lin, J., Sharma, A., Wang, D. and Zadeh, R. (2013, May). Wtf: The who to follow service at twitter. In *Proceedings of the 22nd International Conference* on World Wide *Web*. New York, NY, USA: ACM, pp. 505–514.

Hart, T. (2015). How do you archive the sky? *Archive Journal*, **5**. http://www.archivejournal.net/issue/5/archives-remixed/how-do-you-archive-the-sky/

Herlocker, J. L., Konstan, J. A., Terveen, L. G. and Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, **22**(1): 5–53.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory. *Journal of Documentation*, **52**(1): 3–50.

Kramer, M. (2014). Going meta on metadata. *Journal of Digital Humanities*, **3**(2). http://journalofdigitalhumanities.org/3-2/going-meta-on-metadata/

Manovich, L., Douglass, J., and Zepel, T. (2012). How to compare one million images? In *Understanding Digital Humanities*. UK: Palgrave Macmillan, pp. 249–278.

Manovich, L. (2015). Data science and digital art history. *International Journal for Digital Art History*, **1**: 13–35

Melville, P., Mooney, R. J. and Nagarajan, R. (2002, July). Content-boosted collaborative filtering for improved recommendations. In *Proceeding Eighteenth national conference on Artificial Intelligence*, pp. 187–192. Edmonton, Alberta, Canada, AAAI Press.

Ramakrishnan, N., Keller, B. J., Mirza, B. J., Grama, A. Y. and Karypis, G. (2001). Privacy risks in recommender systems. *IEEE Internet Computing*, **5**(6): 54.

Ricci, F., Rokach, L. and Shapira, B. (2011). *Introduction to Recommender Systems Handbook*. United States: Springer.

Rosen, J. (2012). The right to be forgotten. *Stanford Law Review Online*, **64**: 88.

Sarwar, B., Karypis, G., Konstan, J. and Riedl, J. (2001, April). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*. New York, NY, USA: ACM, pp. 285–295.

Simitis, S. (1995). From the market to the polis: the EU directive on the protection of personal data. *Iowa Law Review*, **80**(3): 445–470.

Taylor, D. (2012). Save As. *e-misférica*, **9**(1). http://hemisphericinstitute.org/hemi/en/e-misferica-91/taylor

Yen, T. F., Xie, Y., Yu, F., Yu, R. P. and Abadi, M. (2012, February). Host fingerprinting and tracking on the web: privacy and security implications. In *The 19th Annual Network and Distributed System Security Symposium*. Reston, Virginia, USA, Internet Society (ISOC).

Zhou, Y., Wilkinson, D., Schreiber, R. and Pan, R. (2008, June). Large-scale parallel collaborative filtering for the netflix prize. In *International Conference on Algorithmic Applications in Management*. Vancouver: Springer Berlin Heidelberg, pp. 337–348.