While serendipity is chance-based and cannot be controlled, perhaps it can be engineered. A few existing digital humanities and cultural heritage projects experiment with engineering serendipity. Serendip-o-matic, launched in August 2013, aims to re-incorporate chance into the scholarly research process. On the website, users input a text; the tool identifies key words in the text and responds with primary source images from several online collections. The goal of Serendip-o-matic is to yield happy accidents for a wide range of users, whether students in search of inspiration for a paper topic or scholars looking for materials to enliven a current project.[4] Another example is Magic Tate Ball, a mobile application designed by digital studio Thought Den to encourage a general audience to discover works of art in the Tate's collection. Using GPS location, time of day, weather, and analysis of ambient noise, the application returns an artwork, explaining why this work was selected and providing content that allows the user to learn more.[5] Magic Tate Ball enables users to engage with works they would not have sought out otherwise while infusing play in the discovery process.

At HyperStudio, we hope to incorporate a similar sense of serendipity in ArtX. Serendipity has the dual advantage of skirting traditional boundaries and adding a playful element to the user experience, which serves both browser and researcher. As we aim to make meaningful and creative connections between the art objects that comprise our past and the events of the present, we believe we can incorporate both audience groups without sacrificing archival rigor. To do so, we will need a holistic, audience-centered approach to digital curation and recommendation.

To achieve this goal, we plan to start small. Through specific partnerships with museums in Boston, we are building a closed and controlled system that can serve as a testing ground for new models of recommendation. Free from industry demands such as growth and scale, we can perfect our schemas and our assumptions before expanding to other institutions. We are also hopeful about creating a collaborative, open-source approach to art recommendation, particularly given the close secrecy with which proprietary recommendation algorithms are guarded. By encouraging open conversation around the ways we recommend art, we may find unique approaches and ways in which current recommendation systems are insufficient or misleading.

We have many questions and challenges ahead. It will be important to understand our audience: How much control over the discovery process do users want, and how can we best balance the sliding scale between browser and researcher? We expect our primary audience to be Boston-area residents and university communities—a casual but informed audience that bridges aspects of both. We hope to instill a scholar's depth of interest and rigor in the casual user and we hope scholars too can employ the tool as serendipitous inspiration for their own work. But how transparent can we be about the logic behind our recommendations? How can we scale such a strategy, connecting artworks to books, lectures, music, movements and ideas?

Perhaps most importantly, while we have explained "why serendipity," we must address the "how." Serendipity involves more than simply selecting objects at random, but what signals are important? How can we prime a user for the mindset of serendipitous discovery, rather than rote research? Moreover, is it truly serendipitous if we are closely engineering the suggestion? We look forward to addressing these questions, but with care to not create our own faulty algorithms. One of the challenges in this process is to avoid reducing cultural objects to the level of products, and museum audiences to consumers. Looking past the current limitations of discovery will be vital for generating new connections and ideas.

References

1. **A.A. Kardan and M. Ebrahimi**, *A novel approach to hybrid recommendation systems,* Information

2. *Interview: Matthew Israel on The Art Genome Project,* September 21, 2013, Museum Geek, museumgeek.wordpress.com/2012/09/21/interview-matthew-israel-on-the-art-genome-project.

3. Scholarship includes Allen Edward Foster and Nigel Ford, "Serendipity and Information Seeking: An Empirical Study," Journal of Documentation, 59 (2003): 3, pp. 321-340; Sebastian Chan, "Tagging and Searching – Serendipity and museum collection databases" (paper presented at the annual meeting for Museums and the Web, San Francisco, California, April 11-14, 2007); and Anabel Quan-Haase and Kim Martin, "Digital Humanities: The Continuing Role of Serendipity in Historical Research" (paper presented at the annual meeting for iConference, Toronto, Canada, February 7-10, 2012).

4. *One Week | One Tool Team Launches Serendip-o-matic,* Roy Rosenzweig Center for History and New Media, Friday, August 2, 2013, chnm.gmu.edu/news/one-week-one-tool-team-launches-serendip-o- matic.

5. **Ben Templeton** (2012), *Mobile Culture and the Magic Tate Ball,* The Guardian, July 16, www.theguardian.com/culture-professionals-network/culture-professionals- blog/2012/jul/16/mobile-culture-magic-tate-ball-app.

Supporting "Distant Reading" for Web Archives

Lin, Jimmy
University of Maryland, United States of America

Kraus, Kari
karimkraus@gmail.com
University of Maryland, United States of America

Punzalan, Ricardo L. Punzalan
University of Maryland, United States of America

In a recent essay on the stock footage libraries amassed by Hollywood studios in the first half of the 20th century, Rick Prelinger—moving image archivist at the Internet Archive—laments that "archives often seem like a first-aid kit or a rusty tool, resources that we find reassuring but rarely use" (Prelinger 2012). Although he doesn't single them out by name, web archives are particularly vulnerable to this charge. User studies, access statistics, page views, and other metrics have in recent years told a consistent story: web content that has been harvested and preserved by collecting institutions, universities, and other organizations often lies fallow, and like Prelinger's rusty tool may be notable more for its latent potential than for having served any real purpose (Hockx-Yu 2013; Kamps 2013; Huurdeman et al 2013). While the reasons for neglect are myriad, this paper focuses on one: the lack of tools to support a wide range of interactions with the content. We describe initiatives underway at the University of Maryland to partially redress the problem and highlight the need for qualitative user studies.

The Internet Archive's Wayback Machine is perhaps the best-known and most widely available tool to browse captured content. Both the Internet Archive's main public site and Archive-It, its subscription-based web archiving service, replicate the experience of viewing web pages on the live web, thus reifying a "close-reading" experience. First developed in the mid-1990s, the software came of age at the same time digital humanities scholars were building the first generation of web collections aimed at providing high-resolution digital facsimiles of literary and artistic works by Blake, Rossetti, Dickinson, Whitman, and others. The emphasis on accurate rendering and display is thus a hallmark of both the Wayback Machine and many early DH projects, the latter of which likewise self-identify as "archives," albeit archives on a dramatically smaller scale.

Although the capabilities offered by the Internet Archive and other commercial services are significant, we believe considerable technical advances are needed if web archives are to fulfill their promise as tools of analysis as well as preservation. Within the field of DH, the big data vistas offered by scholars such as Matt Jockers and Ted Underwood provide

both inspiration and models on which to base these efforts (Jockers 2013). Unlike the boutique digitization initiatives that characterize the early wave of DH archives of the 1990s and early 2000s, which were often devoted to the works of a single author, the new macroanalytic approaches are premised on mass-digitization of print heritage. The paradigm they embody, moreover, is not digitization in the service of verisimilitude—reproductions that show exact fidelity to their originals—but rather digitization that produces terabytes' worth of intermediary copies that can be cleaned, normalized, segmented, tokenized, mined, and visualized to yield new insights about the cultural record writ large. Such a paradigm disrupts the usual data-information-knowledge continuum by taking the unitary wholes of creative expression—the "cooked" novels or poems or historical documents in print—and temporarily degrading them to a "raw" data state so that they can be analyzed at scale to make higher-order knowledge claims.

We believe that the technical infrastructure to support macroanalytics or "distant reading" on web archives today is inadequate. Existing tools were built before the coming of age of "big data" technologies and provide wobbly foundations on which to build analytical tools that scale to petabytes of data. As an example, the open-source Wayback Machine is implemented as a monolithic stack primarily designed to scale "up" on more powerful servers and expensive network-attached storage. Its architecture captures the ethos of "state-of-the-art" software engineering practices of the late 1990s. Not surprisingly, the field has advanced by leaps and bounds in the last decade and a half. In the 2000s, Google published a series of seminal papers describing solutions to its data management woes, which involve analyzing, indexing, and searching untold billions of web pages. Instead of scaling "up" on more powerful individual servers, the strategy entailed scaling "out" on clusters of commodity machines (Barroso et al., 2013). Before long, open-source implementations of these Google technologies were created, bringing the same massive data analytic capabilities to "the rest of us." These systems form the foundation of what we know as "big data" today, and provide the backbone of data analytics infrastructure at Facebook, Twitter, LinkedIn, and many other organizations. Three key systems are:

- The Hadoop Distributed File System (HDFS), which is a horizontally-scalable file system designed to store data on clusters of commodity severs in a fault-tolerant manner (Ghemawat et al. 2003). The largest known HDFS instance (by Facebook) holds over 100 petabytes.
- Hadoop MapReduce, which is a simple yet expressive programming model for distributed computations that works in concert with data stored in HDFS (Dean and Ghemawat, 2004). MapReduce models analytical tasks in two distinct phases: a "map" phase where computations applied in parallel, followed by a "reduce" phase that aggregates partial results.
- HBase, which is a distributed store for semi-structured data built on top of HDFS that allows low-latency random access to billions of records. Google's Bigtable (Chang et al., 2006), from which HBase descended, powers Gmail, Google Maps, as well as the company's indexing pipeline.

Modern big data technologies provide a technical path forward and an accompanying research agenda that does for web archives what macroanalytics or so-called "distant reading" has begun to do for digitized corpora in DH. As a first step in this effort, we are developing Warcbase, an open-source platform for storing, managing, and analyzing web archives built on the three technologies discussed above. The platform provides a flexible data model for organizing web content as well as metadata and extracted knowledge. We have built a prototype application that provides functionality comparable to the Wayback Machine in allowing users to browse different versions of resources in a web archive (typically as WARC or ARC files). Since Warcbase takes advantage of proven open-source technologies, we are confident of the infrastructure's ability to scale in a seamless and cost-effective manner.

Yet Warcbase is only the beginning. We believe that our prototype—and, more generally, the technologies described above—will provide new capabilities that support innovative uses of web archives. Responsive full-text search on massive collections of web pages, one of the first items on a scholarly wishlist, is within reach: the tools exist in various open-source projects, awaiting integration. Longitudinal analyses of web pages such as tracking the frequency of person or place names become possible if we integrate off-the-shelf natural language processing tools. Yet another possibility is topic modeling on a massive scale; a separate project at the University of Maryland has built Mr.LDA, an open-source Hadoop toolkit for scalable topic modeling (Zhai et al., 2012). To provide a hint of what's possible, we have been working with Congressional archives from the Library of Congress to explore topic modeling and large-scale visualizations of archived content, the results of which we will share during the conference presentation.

Why are large web archives so underused? It is surely not due to a lack of culturally significant material. Valuable content, ripe for exploration, ranges across topics such as electronic literature, alternate reality games, digital tools for human rights awareness, the Arab Spring uprising, and Russian parliamentary elections, to name just a few. Restrictive access regimes are partially to blame, but that alone does not provide a sufficient explanation. We believe that the issue, to a large extent, is a technological form of circular reasoning: scholars do little because the right tools don't exist, and tool builders are hesitant to build for non-existent needs and users. Progress is necessary to understand the essential activities, methods, and questions of researchers. Interviews with current web archive users are a start, but breakthroughs will require deep collaborations between scholars and technologists. The end goal is a comprehensive set of tools for researchers in the digital humanities and beyond to analyze and explore our digital cultural heritage.

References

Archive-It: Web Archiving Services for Libraries and Archives. archive-it.org

Barroso, Luiz Andre, Jimmy Clidaras, and Urs Holzle. (2013). *The datacenter as a computer: an introduction to the design of warehouse-scale machines* (second edition). Morgan & Claypool Publishers.

Chang, Fay, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Michael Burrows, Tushar Chandra, Andrew Fikes, and Robert Gruber. (2006). *"Bigtable: A distributed storage system for structured data."* Proceedings of the 7th USENIX Symposium on Operating System Design and Implementation (OSDI).

Dean, Jeffrey and Sanjay Ghemawat. (2004). *"MapReduce: Simplified data processing on large clusters."* Proceedings of the 6th USENIX Symposium on Operating System Design and Implementation (OSDI).

Ghemawat, Sanjay, Howard Gobioff, and Shun-Tak Leung. (2003). *"The Google File System."* Proceedings of the 19th ACM Symposium on Operating Systems Principles (SOSP).

Hockx-Yu, Helen. (15 February 2013) *"Scholarly use of web archives."* files.dnb.de/nestor/veranstaltungen/2013-02-27-scholarly-use-of-web-archives_public.pdf

Huurdeman, Hugo, et al. (2013). *"Sprint methods for web archive research."* WebSci 2013 Proceedings of the 5th Annual ACM Web Science Conference:182-190.

Jockers, Matthew. (2013). M*acroanalysis: digital methods and literary history*. Urbana-Champaign: University of Illinois P.

Kamps, Jaap. (1 August 2013). *"When search becomes research and research becomes search."* SIGIR'13 Workshop on Exploration, Navigation and Retrieval of Information in Cultural Heritage (ENRICH). Dublin, Ireland. www.slideshare.net/jaap.kamps/sigir-workshop-enrich13

Moretti, Franco. (2013). *Distant reading.* London: Verso.

Moretti, Franco. (2007). *Graphs, maps, trees: Abstract models for literary history*. London; New York: Verso.

Prelinger, Rick. (2012). *"Driving through Bunker Hill."* In Kraus, K. and Levi, A. (Eds.). Rough Cuts: Media and Design in Process. MediaCommons: The New Everyday. mediacommons.futureofthebook.org/tne/pieces/driving-through-bunker-hill

Zhai, Ke, Jordan Boyd-Graber, Nima Asadi, and Mohamad Alkhouja. (2012). *"Mr. LDA: A flexible large scale topic modeling package using variational inference in MapReduce."* Proceedings of the 21th International World Wide Web Conference (WWW).

The term "distant reading" was coined by **Franco Moretti** in *Graphs, Maps, Trees* (2007) and has undergone further elaboration in his newest book (2013). See the "Works Cited" section for full bibliographic information.

Developing a Physical Interactive Space for Innovative Digital Humanities Exhibition

Liu, Jyi-Shane
Natioanl Chengchi University, Taiwan, Republic of China

Liao, Wen-Hung
Natioanl Chengchi University, Taiwan, Republic of China

1. Introduction

Digital humanities empower a creative transformation in both humanities and computing research by inspiring and fostering interdisciplinary interaction. Recently, digital visualization has been considered and established as a scholar methodology for digital humanities (Jessop, 2008). Projects, such as "Tooling Up for Digital Humanities" and "The Spatial History" (White, 2010) at Stanford University, have explored and experimented with various forms of graphic representation of data. Visualization is insightfully considered as part of a research process that may induce powerful arguments or raise new questions. It is also pointed out that visualization seems to give a sense of objective and scientific communication in the scholarly, yet sometime ambiguous, activities of digital humanities.

One of the less addressed issues in digital humanities visualization concerns the exhibition facilities. Even though some display equipment and technologies have been developed for some times, their innovative integration with a large-scale auditorium space to create an exhibition facilities for digital humanities has actually been little reported. We developed an innovative exhibition facility for digital humanities visualization with a conceptual framework of place-making that exploits digital technological mediation of people and humanities. Similar to the museum experiences with innovative engagement (Falk & Dierking, 2000) (McCarthy & Ciolfi, 2008), the exhibition facility induces locative experience for sense-making and potentially plays a pivotal role in facilitating further advance of digital humanities. Our work provides a field tested contribution to the research community by engaging wider audience for digital humanities, facilitating its social impact, and filling the vacancy of building a physical platform for presenting and showcasing research results for better recognition.

2. Physical Interactive Space as Digital Humanities Exhibition Facilities

Following the notion of place-making in urban development and heritage studies (Malpas, 2008), a physical space forms an existential ground where people's senses of digital humanities are shaped and defined. Therefore, an innovative exhibition facility can serve as a social and technical infrastructure of cross-disciplinary interaction and allow for new experiences with tangible and intangible forms of digital humanities. This opens up new ways of exploring and articulating digital humanities visualization with physical and social settings, and potentially widening appreciation and deepening recognition of digital humanities for general audience.

We developed the exhibition facility by transforming a large room used for library reference service and installing an array of display equipment for various forms of interactive visualization. With a floor space of 810 square meters, the room was re-conceptualized as a mixture of digital gallery and auditorium by novel interior design and technology embedment. Figure 1 shows the floor plan of the exhibition facility that comprises an inner conference room, a flanked outer corridor, and a lobby. The inner and outer space are separated by sliding doors in the front opening, auxiliary doors in the corners, and entrance doors from the lobby.

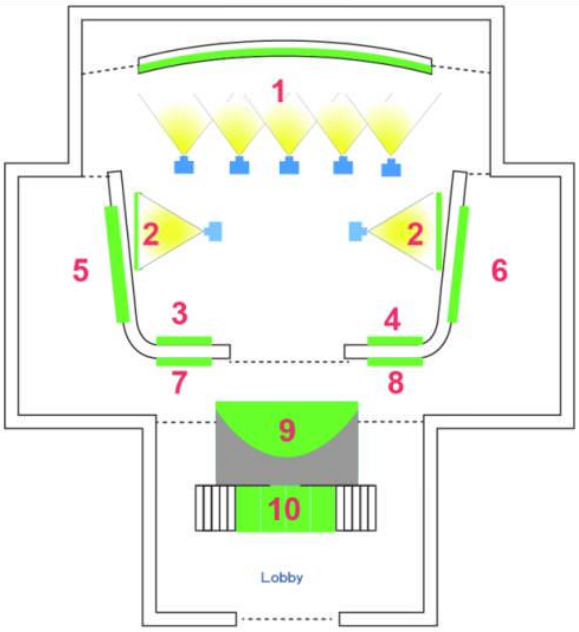

Fig. 1: Floor Plan of the Exhibition Facility for Digital Humanities

A number of ten display systems are either mounted or projected on walls in both parts of the facility, as listed below.

1. An arc wall in size of 12 meters by 2.5 meters (width and height) used as a touch wall display with projection blending of 5 projectors, rendering a surrounding effect of visualization.
2. Two 120-inch retractable projection screens, providing auxiliary displays.
3. Two 42-inch touch screens embedded in a wall book shelf, collaging digital and physical archival exhibition.
4. Two 12-inch monitors mounted on a photo collage wall, blending digital and physical image display.
5. A rear projection touch screen in size of 5 meters by 1.2 meters with projection blending of 3 projectors inside the partition wall, providing easy access and playful social interaction with digital images.
6. Two 60-inch 3D touch screens embedded in a partition wall, rendering 3D images of objects with 3D goggles.
7. A 55-inch touch screen embedded in a partition wall,
8. Two 42-inch transparent LCD boxes, exhibiting physical objects/materials inside the boxes while displaying digital information on the transparent screens.
9. A curvier arc wall in size of 8 meters by 2.5 meters used as a surrounded wall display with projection blending of 3 projectors, rendering immersive visualization.
10. A collage of wall-mounted four 46-inch screens in 4K2K resolution (ultra high definition), used as a digital signage board in the lobby.

Figure 2 through Figure 5 show actual images of the renovated results for an innovative exhibition facility.