Humanities 'Big Data'

Myths, challenges, and lessons

Amalia S. Levi
College of Information Studies
University of Maryland
College Park, MD 20742, USA
amaliasl@gmail.com

***Abstract*—This paper argues that there have always been 'big data' in the humanities, and challenges commonly held myths in this regard. It does so by discussing the case of transnational research on dispersed communities. Concluding, it examines the lessons humanities and sciences can learn from each other.**

Keywords—humanities; transnational; diasporas; archives; communities; web resources

I. THE THREE MYTHS OF HUMANITIES SCHOLARSHIP

'Big data' are hailed as a novel phenomenon, a wave of data so voluminous that we need new methods of inquiry. But, in humanities, there have always been 'big data'. Anyone who has ever stepped inside a museum, conducted research in an archives, pulled a book from a library shelf, or even wishfully looked at old family photos can attest to this fact. Piles upon piles of documents, objects, images, books, and the information they contain are eloquent sources of indomitable 'big data.' Navigating, using, and making sense out of these 'big data' is what humanities scholars have always done.

Humanities scholarship is then characterized by three myths. The first is that 'big data' is the purview of the sciences and that humanities must embrace this novelty in order to stay relevant in a world where the constitution of knowledge itself is being challenged [1]. The second is that, never mind 'big data,' humanities scholarship is a continuous struggle to overcome the scarcity of primary sources, especially if one's research is outside "mainstream" histories. The third is that everything humanities scholars need for their research exists in archives and museums.

These myths have been prevalent in the case of historical research on transnational populations such as diasporas, immigrants, ethnic minorities, or refugees. Research in transnational communities is characterized by dispersion of resources among institutions, and countries; inconsistency of formats, languages, and material; lack of a central "authority" (institution) collecting such resources, and, today, the dominant role of the Web in the communication flow among members of such communities.

Communication on the Web produces vast corpora of online data. In fact the Web is so integral in the lives of transnational populations that it has given rise to the term of e-Diasporas, i.e., collectives that are sustained and re-created as globally imagined communities [2]. At the same time, content generated by communities on the Web has only exacerbated the issue of dispersion of resources in yet more places and formats.

In their article, "Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon," boyd and Crawford question the assumptions embedded in Big Data, approaching the subject as social scientists and media studies scholars. In this article, I would like to contribute to this discussion as a historian and archives scholar. The points I raise are based on my research on and interaction with ethnic and diasporic communities and the scholars who study them. I believe that the particular—sometimes extreme—issues scholars face when conducting research on such communities can highlight concerns, considerations and challenges of humanities 'big data' and enrich the dialogue on this with other disciplines.

II. TRANSNATIONALISM AS A SOURCE OF 'BIG DATA'

Transnationalism, as a term, overcomes the limitations of the term migration that traditionally reflected a one-way process. Transnationalism denotes circular and dynamic cross-border mobility of people or groups between two or more locations over time [3].

'Big data' in the context of historical research on transnational populations can be understood in two ways:

1) Retrospective 'big data': Analog resources in libraries, archives, and museums (LAM) or private collections that lend themselves to a retrospective creation of big data through digitization. The issue with this kind of 'big data' is that not everything is digitized, and especially in the case of "marginal" histories exist out in the wild. In an era when users expect to find everything online, the disconnection between what is really there and what is visible—and readily available—is obvious.

2) 'Real' big data: Resources created in abundance today that already at their inception are digital, networked 'big data' (e.g., Twitter, Facebook etc). They are quantifiable and computationally malleable, but present challenges for humanities scholars, because they require new methods of analysis. The emphasis in this category of data is not on the "bigness"—they are without doubt huge—but in the way their value as data, their 'recordness,' is understood as such. 'Real' big data exist, but for the time being do not form part of historical scholarship, and the alarming speed with which they disappear, leaving considerable lacunae in our understanding

978-1-4799-1293-3/13/$31.00 ©2013 IEEE

about our very recent past, is currently the prerogative of information professionals, and not scholars.

In the case of historical scholarship, there is currently a disjunction between what kinds of primary sources are currently being produced, and what is preserved for the future—and in what ways. More specifically, in the case of transnational communities that span localities there is never a single, "authoritative" institution able to create representative archives out of all these data. It is only humanities scholars and social scientists conducting research in these communities who are able to create conceptual links between dispersed materials. Such links are externalized through published scholarship (journal articles, monographs, lectures etc., or in scholars' personal archives).

What is important in humanities 'big data' today is not the uniqueness of one record, but the volume of many. For community research, it's not even the records that matter but the network of relationships that constructs community, consolidates its identity and shapes its memories [4]. Furthermore, online 'big data' exemplify the ways that individual lives coalesce on social network platforms to form virtual personal and community archives [5]. At the same time, today's cult of "networked individualism" [6] has forever changed the notion of belonging to a community.

III. TRANSNATIONAL HUMANITIES RESEARCH

Today, it is mostly social scientists use web content as data corpora of "naturally occurring, textualized interactions" from which they harvest text and metadata and which they compile into web archives for systematic, fine grained analysis [7].

Historians who have different research questions might approach such archives with different methods, but for the time being the Web plays a limited role as a source material for them [8]. Increasingly though they will be expected to tackle dispersed and uneven web material and produce research that synthesizes inconsistent data into something that makes sense. What happens when humanities scholars that traditionally developed historical discourse by blending individual documents encounter 'big data'?

The greatest issue for humanities scholars conducting transnational research was up to now the limitations of the physical world. Before the wide availability and use of computers, constraints such as geographical distances, different institutional policies, and analog resources, humanities scholarship was doomed to be a monastic and laborious work. It was necessarily a close reading of our cultural heritage. With the advent of digitization and the Internet, humanities research has become data intensive and collaborative [9].

Cultural heritage institutions have tried to solve these limitations—and in the process, gain greater physical and intellectual control over records—with pragmatic approaches that at the same time are consistent with the aesthetic sensitivities, ideological background, and belief systems of people. In an institution's physical space, or online presence, what got to be exhibited or highlighted was not statistical representations of its holdings, but one out of many: an item that was unique in its quality to excite our senses, to capture our interest, to appeal to our ideological sensibilities, and to reinforce the institution's preeminence and superiority among others as the authoritative source of the story that it presented.

What we lacked were ways to computationally exert control over analog cultural heritage material. Metadata was one way that LAM employed in order to tame humanities 'big data': classifying data into categories made description and navigation relatively easy. At the same time it named things and tried to fit the diversity of our world into predetermined "buckets" that reflected more the biases of the classifying entity rather than our reality [10]. The quest for the unique and the limitations of the physical space led to the myth that everything humanities scholars need for their research exists in archives and museums, and in fact that there is a scarcity of primary sources if one's research is outside "mainstream" histories. Scarcity though boils down to the findability (and malleability) of material, and findability is a direct result of the human limitations of reaching dispersed material and making sense out of them.

IV. CHALLENGES OF 'BIG DATA' HUMANITIES RESEARCH

In the face of 'big data,' historians researching dispersed communities face a series of challenges:

A. A balancing act

In their research, humanities scholars deal with dichotomous, multidimensional and multidirectional 'big data.' Their scholarship must keep a fine balance that 1) spans borders and institutions, 2) amalgamates existing LAM cultural heritage collections with 'big data' produced in social media platforms, and 3) does not prioritize online data over "offline" ones, or vice versa. Data that span borders might refer to the same subject, but might not be homogeneous in terms of format, language, or unit. Their contextual background might exist in a previous, analog world. In fact, offline and online data exist interwoven, "mutually contextualizing," and scholars cannot rely only on one or the other—or if they do so, they have to justify their choice of one kind of data over the other [11]. Moreover, quantifiable 'big data' are usually seen as more authoritative or accurate. Scholars today have to keep in mind all of the above in order to give a balanced representation of the history of transnational communities.

B. The datafication of 'un-data'

Using the term 'data' in humanities is an oxymoron. Humanistic data are as un-data as they can get. For a start, they are never raw. Humanities data are not generated by instruments, but by people in the process of going about their everyday life. They are data because *we* tell so. They are important because people imbue them with significance, and as such they are always value-laden, representative of the time and place they are created in. They empower, and at the same time they create silences [12]. They acquire significance in increments, and produce reality in four crucial moments: "The

moment of fact creation (the making of *sources*); the moment of fact assembly (the making of *archives*); the moment of fact retrieval (the making of *narratives*); and the moment of retrospective significance (the making of *history* in the final instance)" [13].

If facts are factual because *we* tell so, then to what extent does the datafication process of our lives truly represent reality? In the online world, individuals can choose to interact with whomever and however they want through selective exposure to material consistent with their pre-existing attitudes and beliefs [14], avoiding the "mental and intellectual discomfort" that accompanies information that does not fit with what they think they know or believe in [15].

If the creation of 'big data' in humanities is in itself a decidedly human process, then we need humanistic methods for their study, now more than ever. Humanities 'big data' cannot be understood simply as data, independent from the process of their datafication, i.e., the process of turning human lives into facts and narratives. If we think that 'big data' will solve all our problems in humanities scholarship, we run into the danger of a "morality-free engagement with a positivist understanding of human history" [16].

C. 'Recordness'

What constitutes a 'record' in historical research today has dramatically changed in recent years. The dynamic and complex structures of communities and their cultural expressions warrant an "expanding" and "expandable" view of the record as including not only traditional documents, but also oral expressions, performances, monuments, commemorations, community festivals, parades, and even more [4]. But "minor narratives, the untold stories, the traces, the whispers and the expressions of marginalized identities" might not always be deemed valuable and archivists are urged to "embrace new ways of seeing and understanding records," to "recognize and accept this evidence into the archives" by "extending traditional boundaries of recordness" [4].

Previously historians relied on LAM holdings, particularly diaries, family letters, associational and congregational records, or newspapers, that today are increasingly replaced by born-digital material (e.g., diaries by blogs, and letters by e-mails). And while previously building cultural heritage collections was a systematic, but also to a great extent a serendipitous process, today the fragility of the digital medium exacerbates the serendipity of what will be remembered in the future.

What is a record today then? Are 'big data' corpora to be statistically mined for meaning or trends, or are they archives from where we can extract individual digital objects? New kinds of records require that scholars use new computational methods when dealing with them.

D. Violence

Retrospectively, it is not surprising that humanities 'big data' are seen as an oxymoron. To date, we have done everything possible in order *not* to have 'big data' in humanities. Humans have systematically and selectively destroyed what they did not like or agree with based on their ideological, political, or religious beliefs. Human history is a continuum of random acts of violence that shape our cultural heritage:

1) Catastrophic violence: When it was not natural disasters (such as fire, earthquakes, and floods) ravaging our cultural heritage, we did it ourselves either by the law or by the sword—in wars, ethnic cleansing, and civil strife. And while scientific 'big data' requiring computational manipulation in order to yield meaning remained relatively unperturbed, humanities 'big data' deemed "dangerous" or "corrupting" were obliterated.

2) 'Archival violence': We have also consistently appraised: through what Derrida has called "archival violence," we have selected what would be preserved in archives (thus, in perpetuity) and what would disappear. We have also selected what would be taken out of the storage and exhibited in museum showcases, so as to tell our story. These were mostly done based on biases and values, and not on the needs of future users. Today, many LAM appraise Web resources for inclusion into their holdings with the aim of enhancing existing collections [17], but this approach risks to perpetuate biases and skewed practices of the past.

3) Cyber-Infrastructural violence: As previously seen, human lives become 'big data' through an intricate process of datafication. Today, transnational mobility is being carried out on physical, as well as virtual platforms facilitated by technology. Such powerful and complex knowledge-based systems shape and constitute the ways transnational populations produce and consume 'big data' in networks around the world [18]. At the same time, Web resources are unstable and scholars in the future will have to make do with material that is incomplete [8], its intellectual coherence and meaning significantly damaged from having been migrated from one hardware or software platform to another [19].

V. SOME AFTERTHOUGHTS

The above challenges demonstrate that things that were always considered as given in humanities have profoundly changed: 'records' are not anymore what they have always been, and humanities are not really the serene and peaceful endeavor everyone thought it to be. Quite opposite, it is a charged and multidimensional process of interpretive dilemmas.

Recent scholarship critically examines 'big data' as a cultural, technological, and scholarly phenomenon [1], as information overload [20], as affecting our historical methodology [21], or as introducing ethical dilemmas in the way we interact with history [16]. Humanities and sciences have a lot to learn from each other, and such dialogue will enrich our understanding of the notion of "humanities 'big data'." Since—as we hope this paper has rendered obvious—humanities have always had 'big data,' humanities scholars' contribution in this dialogue can be enriching and constructive, even though in many Big Data debates they are usually sidelined by an existing "arrogant undercurrent" [1].

What sciences can learn from historians working on transnational communities is that sometimes data spread all over the globe are too big to be interpreted or tamed even with 'hard,' computational methods. When confronted with such infinite un-data, one needs a certain humility in order to explore them "effectively, efficiently, and as comprehensively as possible" [22]. Humility in front of data that are "too big" is one of the values that humanities scholars know well, because they have always been dealing with such data.

Conversely, humanities scholars have a lot to learn from scientists: Not in the ways they understand data—as we previously saw, humanities data are foundationally different from scientific data. They are complex, imbued with multiple meanings, carrying layered identities, and open to interpretation. In fact, humanities must not seek to imitate sciences in order to become more "scientific"—we have enough scientific data to deal with. Scholars working on transnational communities need to adopt science's exuberant approach to collaborating, sharing data and practices. They also need to learn to look more at the network level, rather than individual stories, and to interweave online data (data that already are 'big data') into their scholarship. By redefining the methods of humanities scholarship, they find new ways of creating links among cultural heritage material. Through this process we begin to contextualize online community data through existing cultural heritage material, and conversely, we include cultural heritage material into contemporary discourses. If 'big data' can sometimes be 'too much' information, then we need to involve scholars and communities of practice in a critical curation cycle [20] for the contextualization of such data.

Finally, humanities scholars need also to embrace and promote technologies that can help bring to light the 'big data' already inherent in humanities. Technologies that already information professionals are experimenting with, such as Linked Open Data and the Semantic Web, or research and applications stemming from Computer Sciences, such as machine learning and wikification, or new Internet architecture initiatives, such as the NDN project, may help to bring to surface entities and emphasize the links among them. The next frontier in humanities scholarship is then to enhance the humanities cyber-infrastructure through novel ways of accessing and presenting humanities 'big data' to the users.

REFERENCES

[1] d. boyd and K. Crawford, "Six provocations for Big Data," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 1926431, Sep. 2011.

[2] A. Alonso and P. Oiarzabal, Eds., Diasporas in the new media age: Identity, politics, and community. Reno, NV: University of Nevada Press, 2010.

[3] J. H. Cohen, Cultures of migration: The global nature of contemporary mobility. Austin: University of Texas Press, 2011.

[4] J. A. Bastian and B. Alexander, Eds., Community archives: the shaping of memory. London: Facet, 2009.

[5] S. McKemmish, "Evidence of me … in a digital world," in I, digital: Personal collections in the digital era, C. A. Lee, Ed. Society of American Archivists, 2011.

[6] L. Rainie and B. Wellman, Networked. Cambridge, MA: The MIT Press, 2012.

[7] S. Lomborg, "Researching communicative practice: Web archiving in qualitative social media research," Journal of Technology in Human Services, vol. 30, pp. 219–231, 2012.

[8] N. Brügger and N. O. Finnemann, "The Web and digital humanities: Theoretical and methodological concerns," Journal of Broadcasting & Electronic Media, vol. 57, pp. 66–80, 2013.

[9] C. L. Borgman, Scholarship in the digital age: Information, infrastructure, and the Internet. Cambridge, MA: The MIT Press, 2010.

[10] G. C. Bowker and S. L. Star, Sorting things out: Classification and its consequences. Cambridge, MA: The MIT Press, 2000.

[11] S. Orgad, "From online to offline and back: Moving from online to offline relationships with research informants," in Virtual methods: Issues in social research on the Internet, C. Hine, Ed. New York, NY: Berg, pp. 51–65, 2005.

[12] J. M. Schwartz and T. Cook, "Archives, records, and power: The making of modern memory," Archival Science no. 2, pp. 1-19, 2002.

[13] M.-R. Trouillot, Silencing the past: Power and the production of history. Boston: Beacon Press, 1997.

[14] S. J. Baran and D. K. Davis, Mass communication theory – Foundations, ferment, and future. New York: Cengage Learning, 2011.

[15] C. Baehr and R. C. Schaller Jr., Writing for the Internet: A guide to real communication in virtual space. New York: Greenwood, 2009.

[16] T. Hitchcock, "Academic history writing and the headache of 'Big Data'," (blog entry), http://historyonics.blogspot.com/2012/01/academic-history-writing-and-headache.html, January 30, 2012.

[17] National Digital Stewardship Alliance Content Working Group, *Web archiving survey report,* www.digitalpreservation.gov/ndsa/working_groups/documents/ndsa_web_archiving_survey_report_2012.pdf, June 2012.

[18] A. Elliott and J. Urry, Mobile lives. New York: Routledge, 2010.

[19] D. Anderson, J. Delve, and V. Powell, "The changing face of the history of computing: The Role of emulation in protecting our digital heritage," in Reflections on the history of computing, A. Tatnall, Ed. Springer Berlin Heidelberg, 2012, pp. 362–384.

[20] A. Burdick, J. Drucker, P. Lunenfeld, T. Presner, and J. Schnapp, Digital_Humanities. Cambridge, MA: The MIT Press, 2012.

[21] F. W. Gibbs and T. J. Owens, "Hermeneutics of data and historical writing," in Writing history in the digital age, J. Dougherty and K. Nawrotzki, Eds. 14-Mar-2012. [Online]. Available: http://writinghistory.trincoll.edu/data/gibbs-owens-2012-spring/.

[22] W. J. Turkel, K. Kee, and S. Roberts, "A method for navigating the infinite archive," in History in the digital age, T. Weller, Ed. New York: Routledge, 2012.