DIGITIZING CULTURE

A Research Paper submitted to the Department of Engineering and Society
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By

Jessica Xu

March 27, 2020

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISOR
Catherine D. Baritaud, Department of Engineering and Society

The world is currently experiencing a massive digital revolution with the rapid growth of internet users and internet searches. According to a survey by Internet World Stats, over four thousand million people, which accounts for roughly sixty percent of the world's population, is accessing the internet as of January 2020 (Internet World Stats, 2020). In order to meet the growing online demand, more and more institutions and businesses are moving information online. Digital archives are becoming more common as primary sources and historical records are being converted to digital formats. However, several problems and biases arise when attempting to digitize resources with cultural heritage. The STS research will focus on complications in attempting to digitize resources with a cultural heritage background, and result in a research paper that answers the questions of if those complications can be mitigated by investigating the situation through the lens of the Social Construction of Technology theory (Pinch & Bijker, 1984). The STS paper will answer the question of if an integrated approach to digitizing resources with cultural backgrounds can improve results.

The STS project is loosely coupled with the technical project, which will focus on developing a tool to benefit Social Networks and Archival Context (SNAC), an archival organization. SNAC's current workflow for adding new data to its archives requires deep and extensive knowledge about the data model of SNAC itself. technical project will seek to create a custom plugin that will provide a more intuitive way to refine and update data according to SNAC's data model. This will include basic functionality such as uploading simple data to the plugin and processing it. The technical team will also receive feedback from the client and make changes accordingly. A finalized product with all extra constraints implemented as well as a technical report will be produced by the end of the spring semester. The product will effectively allow users to upload,

clean and publish data more quickly and accurately than before. Given that SNAC is an archival institution, it was very important that the information being uploaded is correct.

Both the technical and STS projects are heavily involved with digital archives. Social Networks and Archival Context (SNAC) itself is an archival organization which focuses on representing relationships between primary resources and people or organizations (Social Networks and Archival Context, n.d.). The STS topic investigates better methods to accurately represent cultural heritage digitally, while the technical topic seeks to create a tool that will increase the efficiency of a specific archival institution. Both projects are concerned with improving the accuracy of digitizing resources.

THE SHIFT TO THE DIGITAL AGE

The pace of digital transformation is accelerating worldwide. Private businesses continue to invest in cutting-edge technologies in order to best their competition, and they modify their business models to meet changing and increasing customer expectations. The shift to the digital age is blurring the boundary between the physical and digital world. Increased availability of resources and more accessible means of communication brought on by digitization can promote unprecedented levels of collaboration and information spread (Royakkers, Timmer, Kool, van Est, 2018). These new possibilities allow the quick spread of information across borders that can cause "processes of hybridization of cultural forms and (local) cultural change" (Primorac & Jurlin, 2008, p. 71). New forms of intercultural communication are being developed and new cultural identities are being created and redefined through the digital domain.

The public, researchers, and historians, and scientists now have more access to online materials than ever. Digital resources are more easily manipulated and changed than their physical counterparts (Royakkers et al., 2018). Thus, ensuring that digital information is both accurate and correctly represented digitally is more important than ever. Any inaccurate information stored in a digital archive can be propagated much faster than its physical counterpart, and can be passed down for generations to come.

CONCEPTS IN DIGITIZATION

Digitization is "the creation of digital objects from physical, analogue originals by means of a scanner, camera or other electronic device" (Manžuch, 2017, p. 2). Apart from just scanning documents, digitization of resources must also go through several steps: selection of materials, assessment of needs, metadata collection and creation, digitization and creation, and submission to repositories (Manžuch, 2017).

One immensely important aspect of digitization is the creation of metadata for the resource. Metadata is a set of data that gives more information about other data. All archives use some form of metadata for description, administration, and preservation of an archived resource (Hodge, 2000). A metadata schema consisting of multiple metadata elements can represent the relationships between its elements digitally (Giannoulakis, Tsapatsoulis, Grammalidis, 2018). Usually, the main purpose of metadata is to assist users in locating information and discovering resources. Metadata is also crucial in the organization and preservation of digital resources. According to the Library of Congress, there are three types of metadata: descriptive, structural, and administrative. Descriptive metadata helps users to search for and find objects and

information such as the author or title of a piece. Structural metadata describes the relationship between objects, and administrative metadata helps with maintenance of a resource; examples include the upload date of a resource, file type, and other technical information (Giannoulakis et al., 2018). When implemented correctly, metadata can improve accessibility and operability of information. A plethora of different standards and schemas for the digitization of a resource and the creation of its metadata exist, and thus makes true standardization difficult (Hodge, 2000).

CULTURE IN ARCHIVES

Digital archives are used as a store for information, and can also be used to conserve "cultural memory" (Evens & Hauttekeete, 2011 p. 164). Indigenous cultures have long been using digitization to "regain control over traditional cultural knowledge that is excessively commercialized and used without the consent of the communities where it originated" (Manžuch, 2017, p. 3). In fact, the World Intellectual Property Organization is encouraging indigenous communities to digitize their intangible heritage such as dance, song, and rituals as conventional intellectual property systems have no protections for these (Manžuch, 2017). The past of these indigenous and marginalized communities has been represented in archives and libraries, but through the lens of "dominant, powerful social groups or societies" (Manžuch, 2017, p. 3). Now, these marginalized groups can use digital collections as a platform for regaining control over their heritage. However, digitizing cultural content has several downfalls and biases, especially if the culture is from a marginalized community.

In the selection and interpretation of cultural heritage for digitization, bias is apparent. Pickover, a member of Historical Papers Research Archive (Research Gate, n.d.), notes that there

is a predominately western approach to the selection process of resources to represent the past (Manžuch, 2017). The standardized metadata schemas used are also predominantly developed with a western perspective. The metadata schemas used often try to fit "indigenous knowledge and spirituality into a western worldview" (Manžuch, 2017, p. 5), which has a negative impact on the culture and on the integrity of the information. Providing digital access to objects that are considered limited use or sacred objects is a form of bias as well. Providing open access online to these objects is a form of neglecting community needs that reinforces a "discriminatory approach to the community" (Manžuch, 2017, p. 5). Perhaps the biggest source of inaccuracies is the lack of local information surrounding a resource. When this information is not present, it is difficult to interpret cultural and historical nuances the lie within the resource itself. Some cultural elements are considered rudimentary in their place of origin, but require a great deal of refinement to make them "meaningful to the global audience" (Akinde, 2007, p. 44). Some related concepts are literacy and language barriers. An estimated eighty seven percent of documents on the internet are written in English, yet only approximately twenty percent of the world can speak English (Akinde, 2007). This great disparity means that the volume of content available in digital archives is disproportionately western, crowding out many local content and local views.

Digitization has the potential to allow marginalized communities become creators of indigenous information and knowledge rather than "passive consumers of imported information" (Akinde, 2007, p. 46). However, all of the aforementioned biases conflict with "obligations of memory institutions to present multiple perspectives on the subject and to promote cultural diversity and dialogue" (Manžuch, 2017, p. 5). Many of these biases stem from the fact that a majority of archival institutions are based in western societies, and thus many of the archivists

that work on these indigenous projects have a western perspective as well. This problem is detailed in a diagram in Figure 1 below. The linear handoff model is used to represent the flow of the information before it reaches the user base (Pinch & Bijker, 1984). Each time the information is handed off, it is interpreted and changed by each of the actors. First, the indigenous community creates the resource and the information with social and cultural nuances

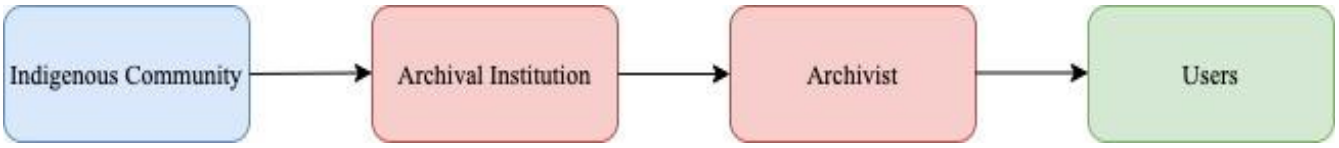

Figure 1: Digitization Linear Handoff Model: A representation of the flow of information throughout the digitization process (Adapted by Jessica Xu (2020) from W.B. Carlson, 2009).

embedded within. The resource is then filtered with many other similar resources by an archival institution, and selected to be added to the archive. An archivist is then assigned the project, wherein he/she digitizes the information, translating as needed and creating metadata tags as seen fit. Finally, the user uses the metadata tags created previously to discover the resource and reads the archivist's interpretation of the resource and its surrounding context. In this model, there are many points of entry where biases can be introduced, and thus many potential inaccuracies in the data exist.

DIGITIZING CULTURE: AN INTEGRATED APPROACH

The biases that exist in the digitization process can be disastrous for the integrity of information represented in a digital archive. Yet, having poorly formed data that does not comply with standards can also be disastrous to an archive. Only by having both a strong technical understanding and a strong cultural understanding of a resource, can an accurate and efficient

digitization take place. A model that could help mitigate some of the biases mentioned before is detailed in Figure 2. The archivist, the archival institution, the indigenous community, and a local expert if needed all work together to complete the digitization process. The archivist and archival institution will provide the technical expertise needed to create well-formed data and metadata. They can also use their expertise to select works that would be valuable to the archive. The local expert and indigenous community can work together to create a more detailed cultural understanding of the information. Altogether, this creates an archive entry that is more understandable, yet still well formed and technically correct.

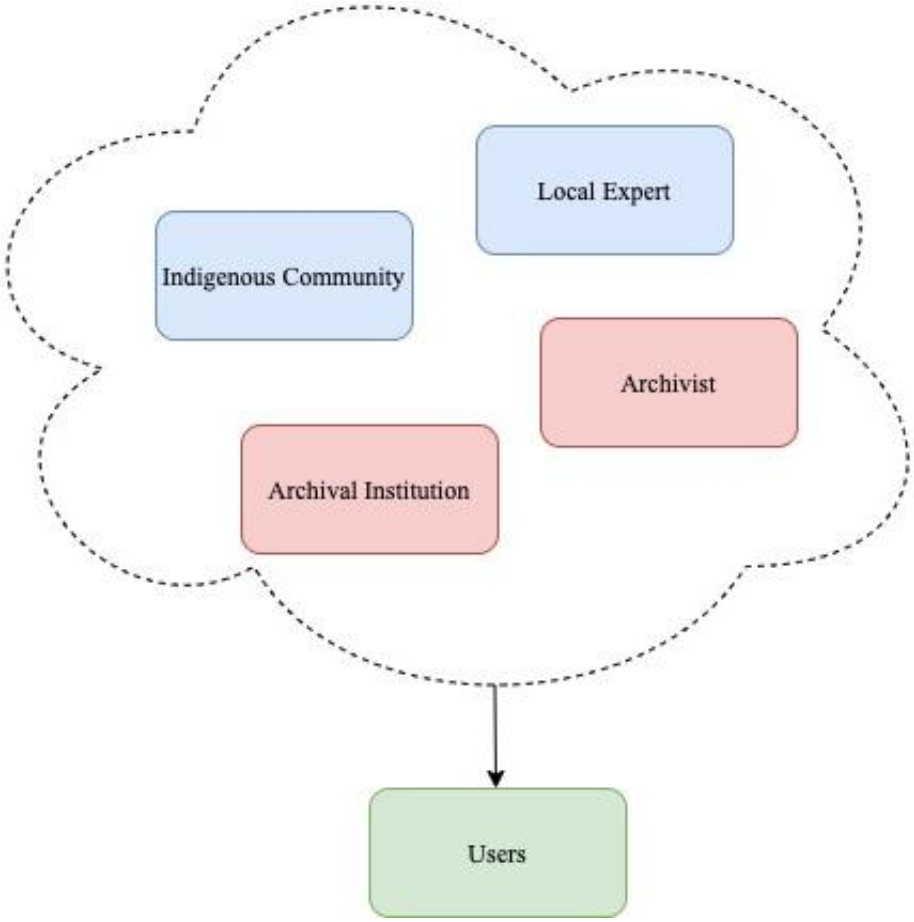

Figure 2: Modified Digitization Model: A representation of the flow of information throughout an improved design of the digitization process (Adapted by Jessica Xu (2020) from W.B. Carlson, 2009).

Many biases exist in the commonly implemented digitization techniques today, includisng selection, language, cultural understanding, and access. These biases and lend themselves to inaccuracies to the archival entries used to represent a cultural resource. A slight

modification to the linear handoff model that represents the current workflow can help mitigate many of these biases. By having local experts and archival experts work closely together in a team, a more accurate representation of the cultural resource can be created and made accessible to the public. Future work on this research paper could include investigating and analyzing several existing digital cultural archives and their digitization processes.

WORKS CITED

Akinde, T.A. (2007). Digitizing African local content: the way forward. *Continental Journal of Information Technology*, 1, 44-50. Retrieved from http://eprints.rclis.org/15579/1/44-50.pdf

Evens, T., Hauttekeete, L. (2011). Challenges of digital preservation for cultural heritage institutions. *Journal of Librarianship and Information Science*, 43(3), 157-165. doi:10.1177/0961000611410585.

Feige, D. M. (2019). The culture of digitalization and the digitalization of culture. In B. Münzberg, S. Konietzko, J. Goudis, J. Mester (Eds.), *Contemporary research topics in arts education*. (pp. 52-59). Hamburg, Germany: Rat für Kulturelle Bildung.

Giannoulakis S., Tsapatsoulis N., Grammalidis N. (2018). Metadata for intangible cultural Heritage – the case of folk dances. *VISIGRAPP*, 634-635. doi:10.5220/0006760906340645

Hodge, G.M. (2000). Best practices for digital archiving. *D-Lib Magazine*, 6(1). Retrieved from http://www.dlib.org/dlib/january00/01hodge.html

Internet World Stats. (2020, March). *Internet growth statistics*. Retrieved from https://www.internetworldstats.com/emarketing.html

Manžuch, Z. (2017). Ethical issues in digitization of cultural heritage. *Journal of Contemporary Archival Studies* 4(4), 1-17. Retrieved from https://elischolar.library.yale.edu/jcas/vol4/iss2/4/

Pinch, T., & Bijker, W. (1984). The social construction of facts and artefacts: or how the sociology of science and the sociology of technology might benefit each other. *Social Studies of Science, 14*(3), 399-441. Retrieved from www.jstor.org/stable/285355

Primorac, J., Jurlin, K (2008). Access, piracy, and culture: the implications of digitalization in southeastern Europe. In A. Uzelac & B. Cvjecticanin, *Digital culture: the changing dynamics*. (pp. 71-90). Zagreb, Croatia: CultureLink.

Research Gate (n.d.). *Michele Pickover Profile*. Retrieved from https://www.researchgate.net/profile/Michele_Pickover

Royakkers, L., Timmer, J., Kool, L. van Est, R. (2018). Societal and ethical issues of digitization. *Ethics Inf Technol*, 20, 127–142. doi:10.1007/s10676-018-9452-x

Social Networks and Archival Context. (n.d.). *About SNAC*. Retrieved from https://portal.snaccooperative.org/about

Xu, Jessica (2020). Figure 1: Digitization Linear Handoff Model.

Xu, Jessica (2020). Figure 2: Modified Digitization Model.

BIBLIOGRAPHY

Akinde, T.A. (2007). Digitizing African local content: the way forward. *Continental Journal of Information Technology*, 1, 44-50. Retrieved from http://eprints.rclis.org/15579/1/44-50.pdf

Ark, T. V. (2018, November). Why social studies is becoming AI studies. *Forbes*. Retrieved from http://forbes.com/

Calmon, F.P., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K.R. (2018). Data pre-processing for discrimination prevention: Informatic-theoretic optimization and analysis. *IEEE Journal of Selected Topics in Signal Processing*, 12(5), 1106-1119. Doi:/ 10.1109/JSTSP.2018.2865887

Columbus, L. (2019, September). State of AI and machine learning in 2019. *Forbes*. Retrieved from http://forbes.com/

Evens, T., Hauttekeete, L. (2011). Challenges of digital preservation for cultural heritage institutions. *Journal of Librarianship and Information Science*, 43(3), 157-165. doi:10.1177/0961000611410585.

Feige, D. M. (2019). The culture of digitalization and the digitalization of culture. In B. Münzberg, S. Konietzko, J. Goudis, J. Mester (Eds.), *Contemporary research topics in arts education*. (pp. 52-59). Hamburg, Germany: Rat für Kulturelle Bildung.

Giannoulakis S., Tsapatsoulis N., Grammalidis N. (2018). Metadata for intangible cultural Heritage – the case of folk dances. *VISIGRAPP*, 634-635. doi:10.5220/0006760906340645

Hodge, G.M. (2000). Best practices for digital archiving. *D-Lib Magazine*, 6(1). Retrieved from http://www.dlib.org/dlib/january00/01hodge.html

Ham, K. (2013). Free, open-source tool for cleaning and transforming data. *Journal of the Medical Library Association*. Retrieved from https://www.ncbi.nlm.nih.gov/

Hill, K. M. (2016). In search of useful collection metadata: Using OpenRefine to create accurate, complete, and clean title-level collection information. *Serials Review*, 42(3), 222-228. Retrieved from https://www.sciencedirect.com/journal/serials-review

Internet World Stats. (2020, March). *Internet growth statistics*. Retrieved from https://www.internetworldstats.com/emarketing.html

Langley, P. (2011). The changing science of machine learning. Machine Learning, 82(3), 275-279. doi:10.1007/s10994-011-5242-y

Law, J. & Callon, M. (1988). engineering and sociology in a military aircraft project: A network

analysis of technological change. Oxford University Press, 35(3), 284-297. doi:10.2307/800623

Manžuch, Z. (2017). Ethical issues in digitization of cultural heritage. *Journal of Contemporary Archival Studies* 4(4), 1-17. Retrieved from https://elischolar.library.yale.edu/jcas/vol4/iss2/4

OpenRefine. (n.d.). *Introduction to OpenRefine*. Retrieved from http://openrefine.org/

Park, J-R. (2008). Metadata quality in repositories: A survey of the current state of the art. *Cataloging & Classification Quarterly*, 47(3), 213-228. doi:10.1080/01639370902737240

Pinch, T., & Bijker, W. (1984). The social construction of facts and artefacts: or how the sociology of science and the sociology of technology might benefit each other. *Social Studies of Science, 14*(3), 399-441. Retrieved from www.jstor.org/stable/285355

Primorac, J., Jurlin, K (2008). Access, piracy, and culture: the implications of digitalization in southeastern Europe. In A. Uzelac & B. Cvjecticanin, *Digital culture: the changing dynamics*. (pp. 71-90). Zagreb, Croatia: CultureLink.

Research Gate (n.d.). *Michele Pickover Profile*. Retrieved from https://www.researchgate.net/profile/Michele_Pickover

Royakkers, L., Timmer, J., Kool, L. van Est, R. (2018). Societal and ethical issues of digitization. *Ethics Inf Technol*, 20, 127–142. doi:10.1007/s10676-018-9452-x

Sammy, A. (2019). Bias in the machine. *Internal Auditor*, 76(3), 42-46. Retrieved from https://na.theiia.org/Pages/IIAHome.aspx

Shein, E. (2018). The dangers of automating social programs: Is it possible to keep bias out of a social program driven by one or more algorithms? *Communications of the ACM*, 61(10), 17-19

Silberg, J., Manyika, J. (2019, June). Tackling bias in artificial intelligence (and in humans). *McKinsey Global Institute*. Retrieved from http://mckinsey.com

Social Networks and Archival Context. (n.d.). *About SNAC*. Retrieved from https://portal.snaccooperative.org/about

Xu, Jessica (2019). Figure 1: GANTT chart.

Xu, Jessica (2019). Figure 2: SNAC plugin model.

Xu, Jessica (2019). Figure 3: Simple machine learning workflow.

Xu, Jessica (2019). Figure 4: Machine learning bias actor network theory.