Universiteit Leiden

Digitization Initiatives For Preservation & Access :

The Transkribus Project As a Stepping Stone For a New Archival Era

Archival Studies (MA) Faculty of History

A Master Thesis by Filotas Liakos

June 2019

Contact Details:

liakosfilotas@gmail.com

Supervised By: Prof. Paul Brood

Second Reader: Dr. Bart van der Steen

Table of Contents

Introduction

Context

In March of 1997, Gordon Bell and Jim Gray of Microsoft's research department, with their article titled "The Revolution Yet to Happen", predicted that in half a century – that is, the year 2047 – almost all information about physical objects, people, buildings, processes, and organizations would be digitally available online (Bell & Gray, 1997). Cyberspace, the global, non-physical space of interconnected digital communication networks (Dodge & Kitchin, 2001) would set a new perception in people's minds about informational technology, science, entertainment, and education.

Twenty-one years after Bell & Gray's prophetic article, technology gradually became involved in every aspect of everyday life. Nowadays, everything around us seems to be automated and technologically sophisticated in ways we could not even imagine before. Science and technology have enabled every-day objects like cell phones, watches, and even vehicles to advance their competencies; we now see basic functions like phone calls or text messaging to have advanced to sophisticated procedures involving cutting-edge technology such as machine learning and virtual assistants, making our daily tasks easier than ever (Wortmann & Flüchter, 2015). A very interesting antecedent of this progress is the vast production of digital data, as well as the transition towards digital repositories of material such as online collections. What is more, the majority of humanity's technological memory has been in digital form since the early 2000s, which was 94% digital in 2007 (Hilbert & Lopez, 2011).

General Problem

Notwithstanding, this innovative spirit does not charm all sides of modern life yet. Informational technology, from the moment it was introduced to archival science, caused an almost chaotic situation for the archival community (Runardotter, 2007). Even before explicitly entering the archival field, technology offered the capability of producing vast amounts of recorded information, more than any previous technologically advanced decades of human activity. Humans create 2.5 quintillion bytes of data each day at our current pace,

a pace that is only hastening with the ever-growing Internet of Things (Marr, 2018). All these records, sooner or later, will likely need to be managed. The paradox of this technological miracle is that it made data less available than ever, due to the fact that digital information is less reliable, retrievable, and accessible than the good old preserved analog documents (Duranti, 2001). In other words, the public is now able to generate an endless amount of information, but can no longer guarantee its long-term access or preservation (Duranti, 2001; Runardotter, 2007). Therefore, the case of managing vast amounts of digital data has become an immense challenge for the archival community, one that has not yet been fathomed.

Specific Problem

Preservation and access are critical values in the archival field, values that were hugely affected by new technological affordances (Hedstrom, 1997; Moss, Thomas & Gollins, 2018). Digital preservation and access have always been appealing topics for modern archival research (Mnjama, 2008; Kemp, 2015; Cunningham & Phillips 2005; Runardotter, 2011). Both these concepts are central in this research, which takes them as the starting point of its inquiry. Starting from the modern issues that make preservation and access complicated, this research is gradually expanding towards digitization initiatives that help tackle such challenges through the solution of digital preservation and access.

Archival preservation today is a fundamental issue for archivists all over the globe. At the beginning of the first quarter of the 21st century, keepers of records are still confronted with the possibility of damaging or losing essential parts of humanity's scientific and cultural heritage which consists of records and collections of nations' archives, libraries, and other repositories of immense value. A recent example is the loss of records and artifacts at the fire that destroyed Brazil's National Museum in September 2018 (Gorman, 2018), libraries and archives destroyed by natural disasters (UNESCO, 1996), by wars (UNESCO, 1996), or more recently, by digital obsolescence, also known as the danger of losing the ability to ‘read' specific types of digital files (Anderson, 2015). A future where all archival material is safe and accessible online may sound truly attractive and sustainable, but as we will discuss later on this paper, archival preservation, and especially digital preservation is a complex and challenging operation (Adu, Dube & Adjei, 2016). Problems like physical and digital storage, budget and ethics codes that are able to create an

unfavorable environment for archival preservation, as well as possible solutions and examples, will be discussed thoroughly throughout this paper.

Except for preservation, another focal point of this study is the modern challenge of archival materials' **access**. Access is an integral part of archival science, as it provides the affordance of constructing, refining, correcting or reassuring memory by material that is preserved whenever it is needed (Menne-Haritz, 2001). Online access has become an imperative requirement, especially for new generations, who seem to follow every technological innovation with remarkable ease. Online access carries some profound advantages, as well as some disadvantages that implicate both professionals, as well as the public. Some of the challenges that archivists have to face nowadays include securing archival material (Sewdass, 2014), the future of reference for archivists (Trace & Ovalle, 2012) and archives as institutions and their visibility which may be erased from the public memory. These issues also relate to various moral principles and ethics in the archival field (Danielson, 2010), especially when online access and digital material distribution is at stake. In that case, we pass through a whole new level of challenges for archivists (Holzmann, Goel & Anand, 2016). Online archival access requires serious technological infrastructure and further technological advancements, and at the same time, it can bring additional responsibilities for archivists. At this point, online access has to deal with multiple issues such as budget, copyrights, deterioration, and security of digital material; all these aspects will be dealt with in detail in the following paragraphs.

a. Proposed Solution

Digitization may be an excellent opportunity for institutes to evolve and adjust their core principles closer to modern age standards, even though as we already stated, may pose new challenges and struggles. According to the Universal Digital Library: "for the first time in history, all the significant literary, artistic, and scientific works of mankind can be digitally preserved and made freely available, in every corner of the world." But at what cost? It is indeed true that digitization has alleviated some of the tension between the desire to provide access and the need to preserve originals (Matusiak & Johnston, 2014). Digital technology is the main idea for unlimited new possibilities in the information science world. As technology moves forward, more and more pathways are cleared for scientists to explore history and heritage. A prime example is the development of the Transkribus project, a

supervised machine learning software that can be trained to automatically transcribe handwritten documents and that various institutes around the world have already adopted.

Some of the questions that this paper attempts to answer are:

1. What does the current digitization landscape look like?
2. What are some noteworthy initiatives regarding digitization that occurred until today? Are digital preservation and access the optimate solutions for archival security?
3. What is the opinion of professional archivists about digital infrastructures in the archival scene today?

b. Outline of paper

This research aims to investigate current digitization initiatives for preservation and access and shed some light on the possible future pathways of archival science. First, an extensive literature review regarding prior history, preservation, access as well as digitization will be analyzed in order to present the current state of the art, as well as the problems that these concepts have in store. Secondly, an extensive analysis of the Transkribus project will be presented, so that the pros and cons can be detected. In that part, an overview of two selected noteworthy digitization initiatives will be presented, in order to explore innovative ways that were adopted by institutes globally in order to move forward and evolve. Last but not least, the results of a global survey conducted during the summer months of 2018 will be showcased, in order to grasp the opinions of archival professionals and detect future directions.

Literature Review

Introduction

Archives are collections of documents, or records, which have been selected for permanent storing and preservation due to their increased value as evidence or as sources for historical or other research (definition given by the British National Archives). The activities of individuals and organizations create records; these records not only serve a purpose while in use, but also some of them are later selected and preserved as part of an archival collection (British National Archives). The need to methodically assess the methods and handling of archival material brought the birth of archival science into place. Duranti and MacNeil (1996) give the following definition about archival science: "*Archival science, which emerged out of diplomatics in the nineteenth century, is a body of concepts and methods directed toward the study of records in terms of their documentary and functional relationships and the ways in which they are controlled and communicated* (p.47)". It can be argued that archival practice, as an organized science with standardized concepts, was initiated by the publication of the "*Manual of an Archival Arrangement and Description*" ("Vor Handleiding van het Ordenen in Bescheijven Archieven") , written by Muller, Feith, and Fruin back in 1898, a work which assembled a series of assumptions, or rules, and created a general consensus on the area of archival management (Horsman, Ketelaar & Thomassen, 2003). The Dutch were the first to articulate and standardize these principles, concerning the nature, arrangement, and management of the archives in a proper manner as a manual to help professionals do their job in a concise and proper manner (Cook, 1996).

Menne-Haritz (2001) strongly highlights that the primary function of archives is amnesia prevention, i.e., to construct, refine, correct or reassure collective memory.

From a more practical standpoint, the archival profession mainly consists of collecting, managing but also providing access to archival collections and records for the long term (Duranti & Franks, 2015). Archival professionals must do that, while at the same time, must ensure several original characteristics of the material, such as the survival of the provenance, which translates to maintaining all available information about the originator of the archives, in order to preserve the context and secure the survival of significant content within the archive (Sweeney, 2008). Another significant activity is the primary process of being able to keep the original order, which refers to the responsibility of keeping the records in the arrangement which the creator put them, so as to maintain relationships between

records and thus provide evidence about the way that the creator carried out their activities (Niu, 2014). All these can be easier said than done, as in many cases creator has long since disappeared and archives could have been moved around or profoundly used. What is more, these procedures become harder as we transit towards the digital age (Niu, 2014). In the following paragraphs, a brief theoretical background of each concept of interest will be presented, in order to grasp the topic as fully as possible.

Preservation

Introduction

"*The selection of records of enduring value is the archivist's first responsibility. All other archival activities hinge on the ability to select wisely*"; the above phrases are indicative of the main principles and values that the Society of American Archivists reported as principal back in 1986. Cambridge Dictionaries define preservation as "*the act of keeping something the same or of preventing it from being damaged"* (Cambridge Dictionary Online). Archival preservation has very much to do with protecting archival material from potential harms of any kind, from physical disasters like floods or fires to the eventual natural corrosion that happens over time. It involves controlling for things such as the environment of the collection, storage and handling, reformatting, emergency preparedness, and disaster planning, as well as conservation.

It is an undisputed fact that preservation decision making is a heavy burden that follows archival professionals, as the amount of material, as well as the time and financial cost of preserving everything, do not allow hasty judgments (Walters, 1996). Preservation decision making comes after appraisal, i.e., the process when the value of a collection is estimated by institutional representatives (Rockembach, 2018). Walters (1996), based on previous observations by Conway (1990) and Cox (1992), supports the view that preservation is the extension of the appraisal processes, and indeed it would be paradoxical not to prioritize preservation processes based upon the appraisal verifications. The main idea of Walters (1996) is the application of appraisal strategies to the preservation decision making, as well as the synchronization of these two processes into a seamless operation.

Standards related to preservation practices

Unquestionably, preservation is a vital principle of archival science, and there are several standards (technical standards, conventions, and guidelines; from very restrictive and specific to relatively permissive and general in the application) to ensure preservation all over the world (Irons-Walch, 1994). For example, one of the preservation standards that the British National Archives implement is the PD 5454 British Standard for storing material, which includes instructions regarding, among others, optimal lighting, temperature, and mechanically controlled conditions. Irons-Walch (1990) summed up all international and American standards that had been identified by the Society of American Archivists and relate to preservation practices. This study, although quite practical as it was executed to help organize and present the pros, cons, and costs of each standard that was available back

then, is of particular significance, as it not only showcases the multiplicity of available standards available for traditional preservation back then, but it also stresses the importance of archival participation in the development of such standards (Iron-Welch, 1990).

Preservation and Conservation

At this point, a critical distinction must be made. The term conservation is in many instances used interchangeably with the term preservation. However, it is essential to make clear that these two terms do not denote the same idea. Put simply; preservation can be seen as a form of protection, that includes the elements we talked about above. Conservation, on the other hand, should be seen more as a fix towards something that has undergone any type of damage (Ball, 2005). For example, it is an archivist's job to preserve a historical manuscript from humidity, but in case humidity reached and destroyed the manuscript in any way, it then should be turned towards a specialized conservator to take care of it and return it as close as possible to its initial state (Ball, 2005).

Digital Preservation

Countless manuals provide detailed directions of how to digitally preserve a collection, whichever form it may have (Leake, 1960; Ritzenthaler, 2010; Maynard & Foster, 2012; Forde & Rhys-Lewis 2013; Boyda, 2013). According to Forde & Rhys-Lewis (2013), preservation is "(...) the means by which the survival of selected material is ensured for enduring access". That may involve activities such as collections' care (Powell, 2015; Allen, 2016), security (Sewdass, 2014), conservation and restoration (Kathpalia, 1973), as well as disaster planning (Fleischer & Heppner, 2009).

The new ways that technology brought about made digital preservation an option; digital-born material, such as the residue of e-communications (digital-born videos, photographs, e-mail correspondence, and every kind of potential archival material that was born digitally and was never in an analog form), as well as analog data that became digital through digitization, i.e. exactly processes that transform analog material into digital information, that must then itself be preserved (Dictionary of Science and Technology, 2013), as it cannot assure their trustworthiness or longevity, for example through migration or refreshing (Lynch, 2000). The purposes of digitally preserving an archive include the protection of original sources, the representation of original sources for the sake of research

or other purposes, and the transcending of originals, i.e. generating a digital product with added affordances that are impossible to achieve with original sources, for example magnetic resonance imaging (MRI) or technology that incorporates searchable marked-up or raw full text, like the Optical Character Recognition (OCR) or Handwritten Text Recognition (HTR) systems (Besser et al., 2000).

Many scholars have stated the benefits of this method. Forde and Lewis (2013) support that preservation and access are two sides of the same coin; the material is being preserved with the goal of it being widely accessible. Therefore, the benefit of access expansion can be reached through digital preservation. Secondly, digital preservation must be either way existent for digitally-born data (Weston, Garbe & Baldini, 2017). Another valuable benefit of preserving digitally is that archival personnel can have access to all of the collected material and its metadata; that way, knowledge sharing is enhanced, time-consuming activities such as duplications are diminished, and a better communication between internal stakeholders (or even whole networks of institutions) is achieved, making desired goals more easily accessible (Wang & Zhu, 2011). Conway (2010), on the other hand, listed some possible drawbacks of digital preservation, namely the financial strain that a digitization process might bring, the need of full integration of technology and information management processes, and last but not least, the deep-rooted commitment as well as resilient leadership that can orchestrate these long and demanding procedures.

What is more, the role of archival mediation in the digital archive, when stakeholders choose not to interact with a physical repository or archivist, also needs to be taken into account. To conclude, one should also not forget the need for the digitized information to be preserved themselves as well, since digital obsolescence is an evident hazard (Edwards, 2015). At this point, it must be mentioned that a vital requirement of proper preservation of digital objects is the successful preservation of their authenticity (Factor et al., 2009).

Authenticity & Integrity

The last part that should be examined regarding digital preservation is authenticity, meaning the trustworthiness, the reliability that an object carries (Adam, 2010). Factor et al. (2009) support that:

"authenticity is not only a factor of successful preservation; it is a requirement, a necessary condition without which a failure of the preservation system is implied."

Hence, excessive consideration must not only be given to the authentic elements of each preserved object; when assessing authenticity, archival professionals are investigating

whether the object is what it indicates to be (Duranti, 1998). For analog objects that turn digital (for any purpose), their digital copies are precisely that; a copy, a reproduction of the original (Adam, 2010). However, special attention must be given to digital files, as its copies are generally understood as "identical and indistinguishable from the original digital object" (Lynch, 1994). Adam (2010) argues that even though in theory every digital copy is capable of maintaining its authenticity, alterations are still taking place, giving as an example the process of migration, where the format of digital files is altered (hence an essential part of their authentic self) in order to fight the dangers of obsolescence, putting the object's integrity (its relationship with the original) into question.

Adam (2010) proposes the steps that an archivist must take to ensure and preserve digital authenticity. More specifically, a professional should first ascertain where the digital object came from and what it implies to be. Secondly, they must decide which of the qualities of the object, its context, and its content must be preserved in order to remain authentic and, in so doing, set criteria for its authenticity. These criteria are established by researching of its context, its creators, metadata or other technical information that came with it, and on top of all that, ensure their validity in order to form these criteria. Lastly, they must verify that the original digital object, as well as any copies made for any purpose, continues to meet these criteria over time. Then and only then, authenticity can be reached. Unlike traditional, analog objects, digital data and their dynamic nature are in need of continuous authenticity reviews to maintain their legitimacy (Adam, 2010).

Following the same logic, Factor et al. (2009) made a similar point about digital objects' integrity; they supported than in digital files, what matters is not so much their physical matter (their original bit stream written in binary code), but it is their essential components and the content structure that have to remain the same. Indeed, technological advancements compels us to such changes in the physical format of digital objects; what matters is that the information they carry, as well as their facet, stays the same.

Access

Digital Access

Back in 2001, Angelika Menne-Haritz recognized the paradigm shift that took place in archival science, from storage to access. This shift influenced all fundamental values and methods of archival practice considerably (Menne-Haritz, 2001). The International Council on Archives, in its publication of the ‘Principles of Access to Archives' describes access as a metaphor; "(...) access is the link between preserved archives and the public" (ICA, 2011). Access can be understood as form (meaning that everyone who is in need of information sources may have the tools and infrastructure to have access to them), as well as an attitude (meaning that archivists should respect users' competencies and do not provide support for the understanding of the records), from a theoretical viewpoint that puts access and the use of archives in the center of archives' reasons of existence (Menne-Haritz, 2001).

New technological advancements brought online access into the picture. The report of Australia's Council for the Arts showcases some of the crucial steps that must be taken in the process of providing digital access, namely;

1. talking to experts,
2. planning the archives
3. selecting the content
4. preparing and preserving the content
5. managing the archive and
6. delivering the content (Australia Council for the Arts, 2011, p.19).

Online access of digital material can have multiple benefits, including the control of spatial and time challenges, making everything instantly accessible from any part of the world (Hansen & Sundqvist, 2012), minimizing human intervention and therefore median interpretation biases (Conway, 2000), new research possibilities through full-text search and sophisticated, cross-collection indexing (Conway, 2000), as well as reaching wider audiences of any kind, twenty-four hours of the day, seven days a week. Brynjolfsson et al. (2003) quantify the benefit of access to the full list of books at Amazon in contrast to, say, the 100,000 books locally available to a consumer. Digitally accessible archival collections are indeed a desperate need in the Information Age when plenty of users search for

information online. At the same time, digital access may impose great challenges to archival material (copywriting, physical deterioration, security), archival institutions (financial cost, unusability of physical infrastructure and staff), as well as archivists themselves (need of new training), and these challenges should be taken into account by the very pros of the science in order to be met (Senturk, 2014).

It is already mentioned that scholars support the fact that preservation and access should be working in sync to reach their desired goals (Forde and Lewis, 2013). Archival institutions that only provide preservation and storage for the future, but not for access in the present, do not fulfill their function as memory service providers (Menne-Haritz, 2001). As archives are the vehicle of (re)creating a memory, one of their primary services is access to all material that can be used to carve memory (Menne-Haritz, 2001).

Therefore, if the media on which records are stored or the software and hardware used in the rendering process becomes technically obsolete, it can threaten the accessibility of digital files.

Digitization

Introduction

It can be argued that the origin of digitization can be traced back in the 17th century, when Gottried Leibniz (1646 - 1716) developed the binary arithmetic system; this arithmetic system has the function of transforming complex information into dyads represented by ones and zeros, and lead to the development of following technological innovations such as the Mors alphabet, the telegraph, and of course computing and programming (Vogelsang, 2010). In the archival world, there are two main types of archival material when we talk about digitization and digital management. First of all, there is material that was born-digital, meaning in a digital form from the beginning; for example, a picture that was taken from a smartphone and that was never printed on paper, but circulated via digital devices like a laptop, or a tablet. These files are intrinsically digital, and some researchers have looked into them, mainly focusing on lists of digital-born material or manuals that refer to proper ways to handle such material (Nelson, 2012; Dekker, 2010; Dekker, 2013; Peet, 2017).

The second type of material is the one that is analog per se and needs to be firstly digitized in order to reach a digital format, i.e., transform it into ones and zeros for a modern computing device to understand them. For example, if we have a printed photograph that we need to send to someone we know via online media, (e.g., Facebook's Messenger), we will first need to digitize it by scanning it with a digital scanner, or by simply take a digital picture of it and then share the reproduced item. In theory, any machine capable of presenting two differentiated states (the ones and the zeros) can be used to store and communicate digitized signals. Once more, it is important to declare that digital preservation does not refer to the same notion as digitalization, even though they are sometimes used interchangeably; digitization refers to the transformation of processes (for example, the banking sector is being digitized through banking apps and websites), while digitalization refers to objects and material.

These examples of digitization spread wide beyond pictures or paper documents; nowadays, all kinds of textual as well as signal material can be digitized, from music, films, theatrical shows, manuscripts of all kinds, even 3-dimensional objects. While some observe how digitization inevitably strips communication of its interesting imperfections, others dispute that digitization, by decreasing communication to its essential components, produces a lingua franca, able to facilitating universal communication (van Dijk, 2006). Being stripped of errors and repetitions permits digitized information to be easily deposited and transferred,

allowing the "easy manipulation and display of these data" (Verhulst, 2002: 433). Information that has been digitized also affords *data compression* (Negronponte, 1995, 15), that permits for controlled storage in large volumes (Verulst, 2002: 433). In other words, being easily manipulated allows digital data to provide users with additional control over information (Owen, 1997: 94; Beniger, 1986). This additional control allows users to shape their own experiences of it (Feldman, 1997: 4). In more simple terms, digitization permits an extensive degree of interactivity between user and information. This is, probably, most compulsorily stated in legal scholar Lessig's comprehensive idea about digital technologies (2008), through his support of a liberating model of "remix culture", implying that digitized information has the ability to be controlled but also the capacity to be easily and meticulously transferred among points. As digital bits have only two possible forms, 1 or 0, receiving nodes will possibly create fewer errors in transferring and decoding data process than usually occurs in analog systems. Many scholars argue that this process possibly results in a "*lossless*" transmission, causing "*less faults and replication of mistakes and more opportunities for exact processing and calculation*" (van Dijk, 2005: 44).

However, this underscores that shifting digital information does not include any actual transfer of physical materials. Alternately, there is only the transference of information about the configuration of transistors, meaning there is only copying. Some scholars see this as decaying the distinction between the original material from a copy (Groys, 2008: 91), an approach that holds remarkable relevance for legal matters of intellectual property (see Benkler, 2006). As Lessig (2008, 98-99) remarks: "*The law regulates 'reproductions' or 'copies.' However, every time someone uses a creative work in a digital context, the technology is making a copy. When someone reads an ebook, the machine actually is copying books original text from your hard drive, or possibly from a hard drive on a network, straight to the memory of your computer. That 'copy "triggers" the copyright law. When you choose to play a CD on your personal computer, the recording automatically gets copied into memory on its way to your earphones or speakers. Independently of what you do, your actions trigger the law of copyright. In this way, every action must be justified as either licensed or 'fair use*".

The protection of copyright is not the only legal concern that implicates in digitization. Lately, many have examined the associations between digitization and surveillance. More specifically, Nicholas Negroponte acknowledged two decades ago that the digitization process produces "metadata," or "a bit that provides you information about the other bits" (1995: 8). In other words, metadata are resulting by the radical simplification or conversion of the information in digital form. The system has the ability to produce information about digital

streams by extracting signals down to their most basic form. Metadata authorize computer systems to index, search, and finally store digitized information. Digital metadata are frequently produced by users themselves in ways that classify and at the same time index the information (Mathes, 2004). Metadata have been exceptionally significant features of digital media in contexts that varying from knowledge production, social and most important scientific research to government supervision. It has contributed to the rise of ‘big data' social science efforts, from exposing the networked construction of blogs and patterns of social relations on Facebook to the patterns of social media use for political news sites, but also patterns of health messages distribution. Metadata have also proved exceptionally valuable for state agencies seeking to monitor people. The legal context of state supervision using metadata have been the case of the open-ended deliberation about the National Security Agency's use of digital media in order to supervise citizens around the globe, the extent to which was unveiled by the former Central Intelligence Agency Edward Snowden. In the context of this deliberation, Jason Healy (2013) who after the events identified as the first historian of cyber conflict, demonstrated the power of metadata, using organizational affiliations to "uncover" Paul Revere and his revolutionaries fellow without needing to examine the content of their conversations.

Across disciplines, many scholars have concentrated on warning citizens about the fundamental uniqueness of digitization and digitized information. Several researchers have supported that digitizing information endows it with important and essential qualities. Scholars perceive these as the components of digital information and the inevitable outgrowths of digitization. Although, it is an undisputed fact that digitization radically transforms the entire landscape of media and gradually has become universal. Nowadays, the vast amount of media technologies that we routinely interact with are digital. Subsequently, there are no analog equals to be posed against the power of digital technologies.

Benefits & Challenges

The past twenty years in archival science have been observed a revolutionary change in the ways in which scholars and the wider public, access and use manuscripts. Nowadays, digitization has become a quite popular approach among GLAM institutes around the globe. The digital involvement in archival science offers unlimited tools regarding fundamental values in archival science such as accessibility and preservation. Although,

despite the beneficial nature of digitization, remains an approach that entails potential benefits and challenges which archival experts and institutes must be aware of.

Experts define digitization as *"the process of converting, creating, and maintaining books, artworks, historical documents, photos, journals, etc. in electronic representations so they can be viewed via computer and other devices"* (Institute of Museum and Library Services, 2002). Digitization practices have changed and are still changing the archival scene in multiple directions, although some of its most significant advantages remain a controversial issue among archivists and institutes. Digitization as a practice has the ability to enhance archival accessibility and concludes to the mass availability of archival material. Users have the opportunity to interact with countless archival collections and documents of various GLAM institutes from the ease of their home. This mass availability of the archival material has created new standards for the spread of archival information among users, due to the gradually growing digital reproduction of material. Users can achieve better analyses based on existing online material, as they can use a large spectrum of tools in order to conduct their research. Magnifiers, high-resolution zoom functions and data exportation with the press of a button are only some of the tools that users can use, as these functionalities offer to users new ways to find, collect and comment on content. These online tools can be truly valuable for discovering new layers of content.

Furthermore, except from the benefits of instant and massive archival accessibility which digitization offers to both archivists and users, the reduction of physical storage space is another important advantage of digitization. Archival production is a process in progress and institutes should have the capacity of physical space in order to safely deposit and protect analog archives. With the introduction of digitization in archival science, traditional procedures, for example the way that records are being stored, were significantly altered, as institutes have to deal also with digital born archives.

Digital archives can be stored in an incredible range of digital devices creating this way infrastructures which are allowing further accessibility of the archival content. However, not all records are in digital form and most of them are not easily accessible. Numerous physical collections that are protected by archival institutes still require preservation and storage in suitably configured spaces where temperature and humidity are constantly controlled. Otherwise, physical archives could be damaged. There are still countless repositories that aim to protect analog archives from damage, as digital conversion is a work in progress. However, with the passage of time and as long as the archival material continues to turn digital, primary archives make their appearance to the audience less and

less. This is happening because digitized or digital born archives require much less space, effort, and care for their conservation than analog ones.

Moreover, digitization can be a cost-efficient operation as the cost of printing and paperwork can be exorbitant. The management of analog material usually involves various sub-costs like equipment management, maintenance, and cost of space. All these factors seem to be critical components for the financial state and future strategies of archival institutes as digitization might be a costly operation however comparing to additional expenses which analog records creates for its preservation, digitization appears to be the most cost-efficient solution.

Security is another field where digitization contributes, as creates series of potential restrictions which contribute to the long term security and confidentiality of the archival content. Additionally, digitization offers to archives a disaster recovery potential as there is always a risk of disaster, whether it is natural or manmade. Nevertheless, digital archives are not excluded from the risk of being destroyed by the same causes, though their rescue and recovery resemble far more feasible due to the extensive digital reproduction which is taking place at GLAM institutes.

Another significant benefit of digitization is that the process itself is environmentally friendly. Document imaging and overall document digitizing appear to be an environment-friendly initiative which most of the institutes appear to take into account. Digitization removes the needs of creating multiple backup copies and unnecessary printing, increasing this way the eco-friendly quotient of institutes while at the same time reduces the commonplace necessity for funds.

On the other hand, digitization is also creating new challenges for the archival community. The majority of libraries, archives, and museums nowadays, appear to battle with getting up to speed and remaining current in matters concerning both digitization and digital preservation. Digitization appears to be a time-consuming process, mainly depending on the state of the potential holdings before being digitized. Some elements are so delicate that experiencing the digitization process could possibly damage them irreversibly; for example, laser from a scanner can cause damage to old photographs and documents. Although, despite the potential damage, a critical reason for digitizing archival material is because some of them are so heavily used that digitization appears to be the only solution in order to preserve the original document long past what its life would have been as a physical holding.

Additionally, the digitization process can be quite costly. Institutions require the best image quality for their digital copies so when records are digitally converted from one format to another, only a high-quality copy is maintained. Smaller institutions possibly cannot afford

such equipment. Human resources at most of the facilities also define the amount of material that is appropriate for digitizing. Furthermore, funding can limit digital preservation measures in many institutions. The cost of continually updating the hardware and the software can also be prohibitively expensive. Training is an additional issue, since many librarians and archivists do not possess a computer science background.

To conclude, the digitization of vast holdings of special collections has enabled an archival regeneration. This renewal not only has sparked a return to questions about materiality and the material text, but it also invites scholars to consider how libraries reorganize, re-access, and reproduce historical materials (Hardy, 2018). Hardy (2018) supports the perspective that viewing digitization as merely scanning or copying documents is easily translated as a task that can be done mainly by machines, and does not reflect the actual procedures that involve human labor.

Digital Materiality

Still, we need to keep in mind the oxymoron of digital materiality. Digitization seemingly replaces the materiality of media artifacts with a simultaneously ephemeral and ubiquitous immateriality, as digital material is increasingly accessed through multiple technological devices, seemingly everywhere and nowhere at once (Lischer-Katz, 2017). On the one hand, it can be argued that digital media do not comprise by anything more than electrical signals (which essentially translate to the ones and zeros) and code, (which essentially translates into the direction that is given to the computational device to form these ones and zeros). As Paul Leonardi says in his 2010 research:

"[D]ata and electricity are not objects. They are 'stuff' without a tangible character. You cannot touch the data. You can interact with the paper (an object) upon which data is written; you can interact with the screen (an object) upon which data is displayed, but you cannot touch the data itself."

However, many scholars have recently challenged that opinion, focusing on the importance of digital materiality. Lischer-Katz (2017) brought the issue of micro- and macro-materiality of digital objects; micro-materiality involves material issues such as code, electric protocols and all things small involving the realization of digital material, whereas macro-materiality involves the mega-infrastructures such as social networks and infrastructures that are needed for all digital life to exist, making an interesting point about the neglect these material go through when arguing about digital immateriality. Mardon and

Belk (2018) investigate why collectors pay for and go after digital objects when there is a general assumption that such objects are abundant and ubiquitous. The authors make an interesting point by stressing the fact that digital media, be it cultural artifacts such as music, videos or photographs or (more prominently) digital profiles or avatars, are now of great value as ‘personal possessions', which gain their materiality by appearing on a screen or a speaker, i.e., a tangible object through which we access digital material (Mardon & Belk, 2018). Some researchers have even argued that data can be solemnly be seen as a material in itself, without a need for understanding or appreciation (Salter & Murray, 2014). Hence, digital data is still, in the final instance, restricted in the forms of such objects. So again, through this lens, digitization flows between the material and the immaterial, making it a one of a kind process that needs thoughtful examination and strategic planning. The answer behind whether digital data's materiality or immateriality may not have a definitive answer, and it may be a matter of perspective. Nevertheless, from an archivists' perspective, the theoretical division about materiality and the demarcation or union between the medium and the content must be taken into consideration, as both these notions play an important role when deciding about the future of valuable historical sources and information.

Another exciting aspect of materiality is the human labor behind the digital files; for example, an algorithm may reproduce and replicate itself through machine learning and artificial intelligence (Louridas & Ebert, 2016). However, the initial creation, or the breath of life, if we can call it that, comes by human programmers, and it is based upon engineers that make sure to care for the tangible objects through which digital data are accessed. Finally, at the core of digitization processes themselves, it is human agents who delegated about particular decisions regarding the algorithms that digitization processes are being performed upon.

Migration

Digital progress has brought all these new opportunities for archival management. Nevertheless, as it continuously moves forward, it also brings about the dangers of digital obsolescence and the need for continually upgrading and migrating formats. Digital migration refers to the occasional alteration of digital material from one hardware or software configuration to another, or from one generation of computer technology to the next generation, in order to resist the threat of technical obsolescence, something that Garett and Waters (1996) believed to be an essential function of digital archives. In the 14 years since the report was published, migration has become an increasingly common practice in the

field. Before taking severe decisions regarding the digital configuration of archival material, professionals must take into account the constant need for migration and its antecedents, like financial costs and human resources with adequate knowledge on the matter. Regardless of its benefits or limitations, the adoption of migration as a tool for preservation brings with it new questions regarding the previously stated issues of authenticity and integrity of digital copies.

Digitization as a mediator for informational discoverability

To begin with, duplication or multiplication of archival material through digitization eliminates the risk of content being lost whether that's due to natural disasters, such as flood or fire, or even merely a potential loss of the original copy if it is removed from a room and never brought back. Physical media can be very delicate from a risk perspective, and this is why archival institutes, as well as libraries or museums, rightly prefer to keep the majority of valuable physical documents safe in repositories away from the public audience. Multiplying the archive in physical forms such as paper (through books or publications) could be a possible solution. However, the paper archive is limited to data that can be explicitly stored on paper. By using a digital solution, all content associated with the collection can be stored in one place. This could be 3-dimensional projections of objects, or scanned images in high definition. Digitization offers easy access to such sensitive material, extending access to fragile resources. Thousands of terabytes of archival content can be provided to historians and the public audience with information that was never accessible before, and could probably not be accessed in person in the future due to their sensitivity. Thus, digitization can be a tool that combines preservation and access.

Nowadays, access to content should not be limited in only one location or a connection to an internal server. By digitizing collections, giving access to the material to a broader audience becomes an option, and creates a better end-user scenario where they can consume the content they need and receive possible updates in real-time. All of that is more convenient when the organization is able to store all of its data in the same place, something possible with digitized collections. Digitization makes it easier for historians to collaborate online on specific documents or artifacts that have been digitized, and create collective knowledge which in the past would only be possible if both professionals and the archival material was physically present in a room. For example, an Australian historian can analyze a medieval transcript that is physically stored in the Hague but is made available online for professionals to examine through digitization. From the ease of her home, the

historian can collaborate on a document with another professional residing in a different place of the world, and together they can combine their knowledge and shed light to new information or patterns.

Another fascinating aspect of digitization is related to structured data, metadata, and search tags; with digital data, search capability to entire collections is possible, meaning that anyone could use it to find any piece of information on any part of the digital document and its metadata easily. With the new advances in technology and especially HTR transcription, the vast quantity of data that can become digitally available becomes clearer. One could imagine the possibilities that Big Data and its applications in archival science could bring. Even if archival institutions manage to excerpt all possible information of digital collections in a searchable form (and especially on the handwritten character level), it would be an overwhelming task for both professionals as well as the public to make sense of that data or combine related topics meaningfully.

Personalization: each end-user, practitioner or department can form their classification systems based on keywords or topics of interest. For example, an institution might find it meaningful to create classifiers based on the chronological order, while an end-user might classify the same collection based on keywords, or word-clouds, or even combinations of digitized handwritten words. This creates a more natural way for all possible users to utilize collections on their best interest easier and more efficiently. We can understand a lot about the content of each document by studying at fingerprints, such as who uploaded it, who is the creator and the document title. From these elements, we often have the ability to infer a classification. Moreover, we can also study and explore unknown parts of the content of the document using linguistic analysis techniques in order to classify it appropriately.

To begin with, digitization can help utilize the benefits of Machine Learning, which is an application of artificial intelligence (AI) that provides systems the ability to learn and improve from experience without being explicitly programmed automatically. At its simplest level, machine learning might offer additional improvements to existing technologies, such as improving the accuracy of search and discovery by better classification of content, or (semi)-automated appraisal and classification of born-digital archives and records, which can be too labor-intensive due to the vast volume of such data. What is more, machine learning might assist in ‘cleaning up’ unstructured stores of records by automating classification and sentencing. Another possible application can come through the applications of Unsupervised Machine Learning, which investigates the way that systems can infer a function in order to describe a hidden structure from unlabeled data. Meaningful patterns hidden in texts can be

revealed, exposing additional information (such as context, or clusters) that would either not be visible or would be too hard to spot. Last but not least, digitization offers the ability to analyze documents with new innovative techniques, such as historical text analytics, or quantitative historical linguistics.

The trustworthiness of digital archives

Since the early days of the internet, civilization has been wrestling with the concept of how much information we should share and how we limit sharing. The movement to “share” information has resulted in many different and inspiring trends, not the least being social media applications, including Facebook and Twitter. More recently, networking specialists have begun to admit that the Internet was really designed for performance, not security. (Yeo, 2013).

In an extended survey that took place in the summer of 2018, archivists from all over the world were questioned about their opinions on digital information and trust. Via a survey questionnaire that was distributed in more than 17 countries, archival experts were explicitly asked whether they believe that the Internet is a threat to archival integrity, and if they believe that digital archival collections lack security, compared to analog forms of archival collections. In the first question, only 37.0% of the candidates responded that online technology and the Internet can be a threat to archival integrity while the majority of candidates with a percentage of 45.5% supported that, compared to analog, digital collections lack security. These figures bring many new questions to the surface as archivists today seem to be particularly suspicious about the digital material that exists on cyberspace, concerning the digital renewal that archival science is currently experiencing.

Nowadays, the trust of digital records has been interpreted in many different ways. The people who do not own the experience to evaluate the authenticity of a record, commonly rely on the credentials of the experts who authenticate it. The digital environment that archives are being displayed nowadays possesses specific difficulties in establishing trustworthiness in electronic documents (Raab & Szekely, 2017). In archival science, records are considered trustworthy only if they are reliable, accurate, and authentic. Reliability is identified as the trustworthiness of a file as a statement of fact, based on the proficiency of its author, its completeness, and the controls on its creation. Although, as technological evolution has become a new reality for over half a century now, things regarding trustworthiness in digital documents tend to be more and more controversial. Our

insatiable appetite for technological innovation has raised a host of challenges to privacy, security, and trust. In the same context, as the cyberspace grows without universal standards of operation, secure organizations are struggling to preserve security. However, users seem willing to take the fact that that archival information is accessible online for granted, not considering such deeper issues. In reality, archives can hardly withstand the temptation to discharge the burden of processing materials on the community of users, while most users are not particularly bothered by the exact sources of the hits for their searches (Szekely, 2017).

In a modern environment, the value of trustworthiness may experience a decline because communication is depersonalized and the identity of users remains hidden behind websites or other digital proxies (Yeo, 2013). Informational technologies nowadays are becoming unreliable in trust issues and of course, further difficulties arise from increasing skepticism regarding the assumption that pieces of information can be definitively classified as true or false, as deserving or undeserving of belief. All these concerns are unified when we consider records and archives in cyberspace. Many archival institutes chose to store and give access to files in online environments, often by using cloud computing infrastructures established by commercial suppliers losing this way the absolute control of the accessible material. While archivists usually welcome the convenience these environments offer, when they encounter records or objects that purport to be recorded in the digital realm, they may be unsure how far they can trust what is being purveyed to them (Yeo, 2013).

In the past, trust in archival records was said to be reinforced by faith in archivists and in the institutions where archives were kept. Archival institutions were relied on to preserve records in a way that inspired trust, and individuals who lacked the knowledge needed to evaluate a record could call on professionals to perform the necessary authentication (Szekely, 2017). Archivists and archival institutions were seen as neutral and objective third parties that could be trusted to protect records and not tamper with them. However, not only is trust in professional experts and institutions now subsiding (Duranti & Rogers, 2012) but archivists also face issues of disintermediation: online users cannot interact with archivists or sense the physical institution in the way that traditional users could. Perhaps archival experts need to reinvent old ways to make them fit for the digital age (Duranti & Rogers, 2012).

In the same way, current digital repositories seem to create a new attitude for the informational world while at the same time play an imperative role in the long term preservation of documents. Although besides the beneficial aspect of digital access and preservation, electronic records are experiencing a trustworthiness crisis that can cause

multiple issues to the archival communities and public audience. In order to prevent this disturbing phenomenon which is taking place in an extended level in multiple digital repositories and online libraries, archivists are trying to establish a global system of archival standards to keep digital archives unspoiled and secure from cyber threats. This vision of a distributed system of trusted repositories of digital collections was and still is a noble goal, but achieving it would require considerable understanding about the components and attributes of such a repository. Simply put, a high degree of standardization was needed. One of the first standards that archivists chose to establish was the Open Archival Information System Reference Model (OAIS) which became an International Organization of Standardization (ISO) standard (ISO 14721) in 2002. Having a model of a digital repository was an essential first step, but the model did not address the matter of trust. Five years later and since the digital preservation of the archives was a global rather than a national issue, people in the UK and the Netherlands were working on DRAMBORA (Digital Repository Audit Method Based on Risk Assessment) which shared a number of commonalities while at the same time represented different approaches, and further work was required to achieve a single international standard (Dryden, 2011).

In the same way, a working group of ISO in 2007, was formed under the auspices of the Consultative Committee for Space Data Systems (CCSDS) to produce an international standard to serve as the basis for a full inspection and certification program for digital repositories. The criteria are organized into three areas: organizational infrastructure, digital object management, and infrastructure and security risk management. According to Jean Dryden's article (2011), the last stage of standards development, was taking place in June 2011 as series of test audits were conducted at multiples digital repositories in the United States and Europe, using ISO 16363. Today within eight years from last audits, CCSDS has already managed to develop data standards and information system structures covering a diversity of areas including data creation, transmission, administration, and preservation as well as the systems supporting that kind of data (Dryden, 2011). This worldwide innovation changed the way that archivists and the public audience experience and understand online digital records, but the journey has not ended yet. This attempt might be the most successful so far from an international organization, although considering the fast pace that digital archives are distributed through the Internet today, global standards must continuously remain up to date and in a position to certify digital repositories as trustworthy.

Example: The National Archives of the Netherlands

All the above principles and assumptions that have been discussed earlier in this chapter, find good ground in the most important archival institute of the Netherlands, the National Archives. Nowadays, in the Netherlands, archival science is experiencing its digital renaissance and influential archival institutes like the National Archives seem to be vital coefficients of this intellectual and technological progress. For the archival field, technology has always been a value that allows people to interact with the past more naturally, and for scholars, a mean to conduct their research more accurately. According to former actions and strategies, the National Archives seems to be an institute that embraces every kind of motivation about archival science, and that is why there are many initiatives concerning access, preservation, and security that are currently taking place in the institute.

The *Transkribus* platform, an ambitious initiative for HTR recognition, which will be analyzed in the following chapters, is part of this renaissance that the Nationaal Archief represents, as the technology this platform carries is capable of broadening the horizons both for archivists as for the public audience, regarding archival research and information extraction.

Accessibility and preservation regarding digitization were always values that push archival institutes to move forward and continuously create new approaches for better administration of the archival material. Every year, millions of documents experience the digitization process. At the moment, the National Archives are planning to digitize approximately 10% of the National Art Collection in the coming 15 years. That is a total of more than 20 kilometers of paper (Nationaal Archief, 2018). An impressive initiative regarding archival accessibility in combination with digital innovation is the DUTO project which brought together the Dutch National Police forces and the National Archives. In the Netherlands since 2013, all police forces, the National police services force and an IT department have been absorbed by the National Police. DUTO is a program of requirements which is developed by the National Archives and aiming to provide sustainable accessibility in the information systems of government organizations. This program requirements have the status of a standard and are able to provide information professionals with tools in order to make information sustainably accessible. Since its establishment, DUTO has become an important instrument for the National Police and for the management of its archives. This initiative also had as one of its objectives to enhance the accessibility of the National Police

archive and for this reason, the conversion of the archive material into a digital format was essential (Nationaal Archief, 2018).

Furthermore, besides accessibility, Nationaal Archief is aware of the fact that privacy plays a vital role in the relationship between the citizen and the government and is therefore high on the legislative agenda. The National Archive has a responsibility concerning personal data and data exchange in all areas in which the organization is active. Furthermore, the institute is also obliged to handle the collection, storage, and management of personal data of data subjects with care, proportionally and confidentially. This fact applies to tasks in the field of acquiring archival material, archive management and making archive records available (Nationaal Archief, 2018). That is why the National Archives support that is essential to be transparent about how the organization handles personal data and guarantees privacy. Additionally since the September of 2018 and through the project "Making the iceberg visible" with the support of the Transkribus platform as the only tool to achieve this goal, the Nationaal Archief battling in order to be able to convert and provide analog information of complex manuscripts, like the VOC collection, into fully searchable digital texts until 2020. In short, the National Archives of the Netherlands is a remarkable example of institutional action regarding the digital renewal of archival material that contributes to the informational regeneration that archival material currently experiencing.

Conclusion

The question remains, are new technologies creating better frameworks for achieving best archival practice? Calahan & Hujda (2018) in their recent study of new archival tools at the Archives and Special Collections Department at the University of Minnesota, came to the conclusion that introducing new technological tools brought internal practices in closer alignment with archival standards, and in doing so drastically changed the way professionals intellectually and physically manage their repositories. Last but not least, a number of researchers focused on the impact that digital archives have on historical research. Sinn (2012) in his empirical study using quantitative citation analysis, concluded that the use of secondary materials was prominent, as well as the fact that usage of digital sources considerably increased in terms of intensity (amount of materials in each type of source) as well as extensity (how widely a material is being used). It also has been argued that the archival profession should rebrand itself and see technology as a vehicle that can move the profession forward, rather than a means of professional destruction (Garaba, 2015).

Archives are not needed as historical institutions, especially when historical research at the universities is reaching a level of professionalization with which they cannot compete. However, they are needed as providers of access to the past so that everybody can investigate it for his own questions. Archival institutes that provide service on a high professional level get the image of useful social institutions that can be trusted. Access in the following is understood as the key that allows archives to acquire a profile as service-oriented competent professionally managed institutions (Menne-Haritz, 2001).

Transkribus

Introduction

In the last years, a massive number of historical artifacts from libraries, museums, and archives have started to make their presence online appreciable. Documents that were never before accessible to the public eye are now counting thousands of terabytes of digitized images waiting to be transcribed by scholars and history enthusiasts. Ancient manuscripts and medieval documents that were not easily readable by the vast majority of historians are now transcribed in their entirety. Transcription is the process in which historical artifacts are turned into editable text, and in this case, into digitally editable text. Thanks to the technological evolution nowadays, HTR (Handwritten Text Recognition) technology offers the ability to explore the past like never before. Only a few years ago, today's computational power belonged to the sphere of the imaginary. Computer systems evolved tremendously, and are now able to not only "read" historical scripts, but also automatically transcribe manuscripts and archival documents created in the previous centuries. Automated recognition of historical artifacts is a challenging task and demands a transdisciplinary approach. Handwritten documents are as unique and individual as their writers. In the last decade, the scenery of HTR technology has significantly changed so that today we can identify the most promising factors which will make the reformation of access to historical handwritten documents achievable. Technologies like pattern recognition, computer vision, and document image analysis are only some of the related fields that have accomplished remarkable progress during the last decade. Additionally, powerful machine learning algorithms involved in the vast development of new extraction methods and document layout analysis algorithms which recently have successfully applied to the HTR field (Kahle, Colutto, Hackl, Muhlberger, 2018).

Another essential factor that involves in HTR consolidation is the availability of digitized archival documents. Nowadays, more and more institutes perceive digitization as a natural component of their mission and invest significant resources into large scale digitization initiatives. Subsequently, each year thousands of volunteers collaborate with institutes and genuinely contribute to the improvement of the accessibility of digitized collections. Fortunately, all these notions found common ground in one platform. Transkribus is considered as one of the most critical initiatives for the introduction of HTR technology to the public. This software is a revolutionary tool based on the JAVA programming language together with a graphical widget toolkit. This platform was created as a part of the University of Innsbruck's contribution to the TranScriptorium e-Research Consortium (2013 - 2015), a

project that was funded by the European Union and can be considered the alpha version of the software. Professor Günter Mühlberger, the head of the Digitization and Electronic Archiving group at the University of Innsbruck, along with his team, are leading the development of this service platform, which is aimed explicitly towards archival institutes and history specialists. Their team received financial support from the European Union, which was initiated with the TranScriptorium project, and continued through a new project, named "Recognition and Enrichment of Archival Documents" (READ, 2016 - 2019) (Kahle, Colutto, Hackl, Muhlberger, 2018). This project combines groundbreaking research, humanities scholarship, digitization initiatives, and a crowdsourcing marketing strategy. Last but not least, with this project, they are aiming to implement a virtual research environment where archivists, volunteers, scholars, and computer scientists will be able to innovate for the enrichment of handwritten archival material using cutting-edge technology and achieve never-before-seen access to archival material with the support of HTR technology.

Teaching a computer how to read

The most basic type of digitizing an analog document, for example, a manuscript, is image scanning, with a regular or professional scanner. What the scanner does, is take a digital picture of the manuscript and send it to the computer. From there, users can store it, distribute it or analyze it. However, the contents of this image file cannot be edited, as the scanner does not recognize each individual element of the image, like words of pictures.

On a second level, Optical Character Recognition (OCR) has been used in converting scanned or printed text images into editable and searchable text for further processing (Patel, Patel & Patel, 2012). The typical process of an OCR software contains some standard steps; first, the software analyzes the document layout, and it identifies where the text is. Afterward, it identifies the rows, then the words, and finally individual characters. After computing each character's features, it classifies them, and it cleans them up using a standard dictionary or a language model.

In its core, this involves processing a scanned image of a printed page and using software to distinguish pixel patterns within the image. In a final stage, these patterns are translated these into alphanumeric characters which the computer recognizes. With the ability that OCR technology has, i.e., to scan each character individually, the final result comes in files that have a text format, rather than an image format. It is this technology which enables users to search not just the title or date which until then was manually added

by the institute, but actually the words written inside the document, paving the path to more natural search, categorization and accessibility of each document, as it is cannot be disputed that being able to explore all the contents of each document digitally provides far more information that may be proven useful for researchers.

Since the early innovations in OCR technology, both the software and methodologies have improved and OCR has been applied successfully to a wide range of material using a variety of software, commercial and open-source (Blanke, Bryant, & Hedges, 2012). Blanke, Bryant, and Hedges (2012) support the view that open-source technologies have many advantages for historical collections that are difficult to OCR, without rejecting the significant edge in performance that commercial products have. More specifically, after a thorough analysis, the authors conclude to the realization that open source tools, apart from being effectively free to use, tend to be more flexible and have lower administrative expenses compared to comparable commercial offerings while at the same time can deliver decent results on many types of source material (Blanke, Bryant, & Hedges, 2012). There are several open and closed-source OCR tools available, and the accuracy rate of any OCR tool varies from 71% to 98% (Patel, Patel & Patel, 2012).

OCR & HTR

Nowadays, HTR is an important research topic for the scientific community and should not be confused with the Optical Character Recognition (OCR) technology. OCR is the recognition of printed or written text characters by a computer. OCR systems are capable to offer reliable text recognition but they are not able to provide answers in the level that HTR technology can. OCR systems tend to examine the surface of a document, more or less in the fashion that regular scanners do, and recognize the whole text as an individual unit. This technology was originally designed to recognize printed text and it is based on digital text exportation. In some cases can convert the characters into editable text directly from the image and also used to convert a hard copy of a document into an electronic version.

In recent years, Handwritten Text Recognition (HTR) has been a challenging and interesting research area in the field of pattern recognition and image processing. It is a challenging issue to develop a practical cursive, handwritten text recognition system which can maintain high recognition accuracy and is independent of the quality of the input documents. Character Recognition is the mechanical or electronic translation of images of handwritten, typewritten or printed text into machine-editable text. Handwritten Character

Recognition (HCR) is the process to classify characters from the input handwritten texts, as per the predefined character classes. Whereas, HTR is the process to segment line from the text and finally classify characters from the line.

On the other hand, an HTR system is able to provide indexing consulting by analyzing the index on the character level. Each character of the index is treated as an individual unit and allows the system to recognize the meaning and the region of the character in the sentence. This technology can be a productive alternative to the implementation of artificial intelligence, pattern recognition, machine learning, and natural language processing technologies. These kind of technological advancements are the reasons why archivists and historians are now in a beneficial position to provide credible transcriptions for ancient and medieval manuscripts. The demand of the scientific world for further analysis of historical documents began to inspire many scholars and the scientific field of humanities to collaborate and set the initial base for digitization initiatives not only in government but also in the institutional level.

Even though in its newest technological stages, handwriting can be recognized by OCR software, there are many more obstacles, for example, different types of languages and their scripts, obsolete dialects and words, and different types of handwriting that existed over the centuries, such as *fraktur* or *antiqua.* The most basic type of digitizing a manuscript is the actual scanning of it, with a regular or professional scanner. What the scanner actually does, is take a digital picture of the manuscript and send it to the computer. However, this content of this image file cannot be edited, as the scanner does not recognize each individual character. On a second level, OCR technology has the ability to scan each character individually, resulting in files that have a text format, rather than an image format.

What is more, there is a big gap in the recognition of handwritten texts in special documents, such as mathematical documents, maps, musical sheets or blueprints. For those who work on archival material, the OCR technology and its revolutionary possibilities came with a very important limitation; character segmentation is not possible in unconstrained handwritten text images like those of most historical documents of interest (Sánchez et al., 2013). Rather than focusing on individual characters, Handwritten Text Recognition (HTR) engines process the entire image of a word or line, scanning it in various directions and then putting this data into a sequence, without the need of the characters or words to be isolated beforehand (Seaward & Kallio, 2017; Sánchez et al., 2013). HTR can be defined as the problem of finding the most likely word sequence for a given handwritten sequence image (Romero, Serrano, Toselli, Sánchez & Vidal, 2014). In handwritten documents,

character context is important, as the same character can be written in many different forms depending on its context and its turn in one word.

Furthermore, there is a big amount of challenges that historical texts pose when we talk about digitization. First and foremost, the age and quality of the paper can create a big amount of noise among characters, distorting the initial font. Secondly, the same type of distortion can happen from the binding. The inking may be poor, the users' annotations add more information that may not be relevant to the original text, and last but not least, there is a lack of dictionaries that can help with obsolete words or bad spelling. The answer to all these problems can be found in *image analysis,* which refers to several analyses to the scanned text image before it is entered into the software so that unnecessary noise can be eliminated as much as possible.

Even though handwriting can be recognized by OCR technology in its newest technological stages, there are several obstacles, for example, different types of languages and their scripts, obsolete dialects and words, and different types of handwriting that existed over the centuries, such as fraktur or antiqua. (Scaling up algorithms to Google size?). What is more, there is a big gap in the recognition of handwritten texts in special documents, such as scientific documents, maps, lyric sheets or blueprints.

For those who work on archival material, the OCR technology and its revolutionary possibilities came with a fundamental limitation; character segmentation is not possible in unconstrained handwritten text images like those of most historical documents of interest (Sánchez et al., 2013). Rather than focusing on individual characters, Handwritten Text Recognition (HTR) engines process the entire image of a word or line, scanning it in various directions and then putting this data into a sequence, without the need of the characters or words to be isolated beforehand (Seaward & Kallio, 2017; Sánchez et al., 2013). HTR can be defined as the problem of finding the most likely word sequence for a given handwritten sequence image (Romero, Serrano, Toselli, Sánchez & Vidal, 2014). In handwritten documents, character context is important, as the same character can be written in many different forms depending on its context and its turn in one word.

In recent years, Handwritten Text Recognition (HTR) has been a challenging and exciting research area in the field of pattern recognition and image processing. It is a challenging issue to develop a practical cursive handwritten text recognition system which can maintain high recognition accuracy and is independent of the quality of the input documents. This technology can be a productive alternative to the implementation of artificial intelligence, pattern recognition, machine learning, and natural language processing

technologies. These kind of technological advancements are the reasons why archivists and historians are now in an advantageous position to provide credible transcriptions for ancient and medieval manuscripts. The demand of the scientific world for further analysis of historical documents began to inspire many scholars and the scientific field of humanities to collaborate and set the initial base for digitization initiatives not only in government but also in the institutional level.

Interface & Design

The Transkribus project, which according to its webpage aspires to be a personal learning network, but also a natural component of successful crowdsourcing citizen science ecosystem, seems to be one far-reaching tool for the digital transformation of historical research. Transkribus' interface is a platform-autonomous JAVA tool, with which users can reach the services offered by the platform. Users can download Transkribus free of charge, from a comprehensive Wiki webpage which is additionally available as a user guide. During the last six months, I had the opportunity to discover many aspects of this revolutionary software during an internship at the National Archives of the Netherlands. Transkribus was the reason why I decided to explore the new digital perspectives in the archival field. This platform offers a chance to explore the past and the space to deepen into the meaning of historical documents from multiple angles. For the last six months, I have been using this platform nearly every day as part of my research, and in this chapter, I will analyze Transkribus' potential design issues.

When I started to work with this platform as a user, I immediately realized that this software was not necessarily user-friendly. At first glance, the platform's digital interface was looking kind of peculiar and hard to understand. There were many strange-looking buttons and functions that users have probably never used before, creating an intimidating first impression. In my opinion, the absence of any built-in introduction, or pilot, in order to help users understand how to use the platform sufficiently, classifies Transkribus as a specific category of software, academic software.

The software is indeed well designed, but not addressed to the average user. A promising platform like Transkribus, which tries to introduce HTR technology to the academic world, not being designed based on a user-friendly conception seems rather odd, as it cannot be assumed that academics have advanced experience on these types of platforms. Academics for archival or historical experts profoundly design this crowdsourcing ecosystem.

Average users must spend several hours or maybe days until they are able to use the software and work efficiently with it on their projects. I believe that the Transkribus team is aware of this issue and this is the reason why they have created a series of downloadable manuals in PDF form, which aim to introduce the users to the software features and functions. Each manual analyzes a different aspect of the system and explains to users how they can operate the system efficiently. These manuals are undisputedly helpful, but compared to modern standards, this kind of approach can be perceived as an old-fashioned method from some of the users. Most of the digital applications today tend to include introductory instructions and hints in order to appeal to the users' interest and make them feel capable enough to continue to operate the application. Confidence should be one of the first feelings that users experience while using the software, and with the Transkribus platform, confidence must be gained gradually through constant studying and experimenting.

However, it would be unfair to claim that the Transkribus' manual approach was not helpful nor educational, as eventually, users will be able to understand the core idea behind this platform and also train themselves on how they can operate it sufficiently. The only disadvantage from a functional perspective, according to my opinion, is that users can educate themselves and achieve a certain level of knowledge about the platform's operation, but in order to do so, they will have to spend a significant amount of time on studying and manual reading. This is not necessarily a bad condition, but according to modern standards, time-consuming processes usually are a negative characteristic for computer applications (Mancinelli, 2016).

Besides the interface challenges, Transkribus offers some unique and compelling tools for its users. The platform is well designed, but the interface remains, in my personal opinion, a severe issue for the users. Transkribus' major selling point seems to be the use of HTR technology, but as a crowdsourcing platform, the Transkribus team should consider that the platform is not used only by experts, but also by scholars, students, and individual researchers. Hence, the designing of a more user-friendly interface seems essential for the software's success.

Project Bentham

Despite computer software, the Transkribus team appears to promote its crowdsourcing strategy even further, with another project called "Transcribe Bentham." Transcribe Bentham is an award-winning initiative which launched back in 2010 and which is

based in the Bentham Project, whose aim was to introduce to the public online transcriptions of the original and unstudied manuscripts of philosopher and jurist Jeremy Bentham (1748-1832) (Moyle, 2011). Users do not need to download or install a specific transcription tool because they can find projects "transcription desk" straight into their web browser. The project was an initiative of UCL London's global university, and It is now funded as part of the READ project, which is working on the Automated Text Recognition of historical manuscripts. Besides "Transcribe Bentham" Transkribus recently created a brand new e-learning website, which is created by the University of Innsbruck, also as a part of the READ project. This web-page promises to help users get familiar with all kinds of historical handwriting and gain experience with HTR technology. In both cases, users have the chance to work with the available documents, for example, to modify or transcribe them. The online interface for editing those documents is a remarkable web space but does not contain the majority of the functions that are included on Transkribus' computer interface.

To conclude, those initiatives give the opportunity to the public to experience the advantages of the central platform and also educate themselves about the transcribing method. In my opinion, both websites can be viewed as provisional applications for the Transkribus platform. Although, from a marketing perspective, it is fair to assume that the creation of these applications is a brilliant idea because both projects as crowdsourcing initiatives are especially engaging for humanities scholars, archival institutes, volunteers and computers scientists. Those four public categories appear to be the primary target groups of Transkribus' marketing policy and at the same time potential clients and coworkers of the platform.

Additionally, crowdsourcing strategy does not stop in online transcribing tools as Transkribus team has developed its equipment and mobile scanning application. DocScan is a mobile app which allows users to take high-quality pictures within archives and to upload them into Transkribus platform directly. This application offer users the chance to become more independent from the digitization activities in archives while they are in an advantageous position to contribute to the digitization of documents. DocScan includes advanced algorithm and a "series mode" system that allows users to capture photos with the turn of a page. According to Transkribus, this mobile application can take more than five hundred images within an hour. In addition, Transkribus team decided to accessorize the mobile application with a specially constructed tent. Scan Tent is portable scanning equipment that supports users mobile devices in place above the document. This tent is made from a nylon silk fabric and contains built-in led lightning with USB power supply.

Moreover, the base of this tent is made from a black felt fabric in order to provide a more standardized background for the mobile application. Thus it is clear that Transkribus marketing strategy is undeniably impressive. According to my opinion, both of those tools have the potential to revolutionize historical research while they are also useful for converting any old material into a machine-readable format and distributing it online for open use.

In conclusion, the marketing strategy of the platform is undeniably brilliant. The team of professor Mühlberger has successfully developed:

1. A pioneering computer platform for automated HTR transcription
2. An intelligent web-based initiative called "Transcribe Bentham."
3. An e-learning website in order to introduce the audience into HTR technology
4. A sophisticated mobile application (DocScan) for document scanning which is co-working with Transkribus computer application
5. A piece of portable scanning equipment in order to make the scanning experience of DocScan users even more reliable.

All these tools have the ability to work seamlessly together allowing users to experience the power of HTR technology like never before. Transkribus is in a position to offer the users a complete ecosystem for document transcribing and enrichment, and this is a genuinely credible selling point of this platform.

Transkribus' selling point

The Transkribus project was established in 2010 under funding from the Arts and Humanities Research Council, and until today, the idea behind this project remains as pioneering and innovative as nine years ago. Transkribus' target groups can be divided into four categories:

1. Humanities scholars who they are high-level experts and they can provide an accurate transcription of a document. Also, they desire to manage scholarly digital editions of manuscripts.

2. Archival institutes that want to analyze and restore information from a vast amount of digitized records, and at the same time are actively involved in crowdsourcing operations in order to enrich the produced data.

3. Volunteers who know to operate the platform efficiently and can take part in significant transcription projects like READ's "Transcribe Bentham."

4. Computer scientists who aim to develop new algorithms and methods for information extraction and they can contribute with their methods to the technological progress of the platform.

Furthermore, the system's developers are already planning to make Transkribus commercially available to users around the world. Until today, major archival and historical institutes are in touch with Transkribus, expressing their interest. In 2018, the National Archives of the Netherlands used Transkribus as their central ecosystem to base notable projects on cutting edge HTR technology. According to the READ project, which is funding Transkribus until 2019, "the main objective is the advance access to historical, handwritten documents from all over the world, regardless of their alphabet, language or the date of their creation." Transkribus, as part of the READ project, follows the same promoting strategy and promises to its users the ability to transcribe historical documents in a highly standardized, flexible and reliable way. Mainly for the archival field, Transkribus offers a path for new opportunities to access, enrich and explore archival material like never before.

It can be argued that Transkribus fulfills the standards and the expectations of its users. Among the unique features of this platform are the keyword spotting tool, the automated transcription, the advanced layout analysis and the custom manufacturing of HTR models. Each feature has its advantages and disadvantages, but according to Transkribus, what counts most is the transcriptions' safety. The Transkribus server ensures that users will never lose their transcriptions or the documents that they have uploaded on the system. However, besides safety, this platform also ensures long term accessibility of the historical and archival artifacts while also contributes to the preservation of the handwriting material through its server. Thus, accessibility in combination with high-end HTR technology that provided through a well organized technological ecosystem are elements that make this platform an almost irresistible combination for Transkribus' users.

Transkribus has the ability to simplify tasks that would often take years of work, helping scholars with complex handwriting and unusual layouts. Nevertheless, high-end technology combined with the unique features of the platform is also the dominant

characteristics of this system. The servers at the University of Innsbruck use machine learning algorithms in order to teach new writing styles to the system. The system can transcribe the text in any language and handwriting type. After a user transcribes part of the text manually, the software engine learns to identify the characters and then finishes the task automatically with impressive accuracy. Thus, the idea behind the platform seems exceptionally simple and pioneering. All the user needs to do give an image to the software and a part of the corresponding text and based upon this text; the software can learn the handwritten script and similar fonts. However, in order to do this properly, users must create certain circumstances, under which their documents will finally be automatically transcribed.

Moreover, Transkribus' ecosystem is undeniably a big plus that involves in the commercial success of the platform. Users have the opportunity to be part of a very pioneering cutting edge cluster from technologically perspective that allows them to grow their capabilities and knowledge about handwritten documents and application of HTR technology. All those services are provided by the READ project combined with Transkribus' team expertise free of charge. The standard requirement for the use of this ecosystem is average mobile devices and personal computers. The Scan Tent might be a brilliant approach that possible capture the interest of the potential users but even without this piece of equipment, users are still able to produce sufficient and credible scan and text information for their work in the platform.

On the other hand, automatic transcription might be one of the highest selling points of Transkribus, but the success of this operation is dependent on each document's needs. Each text has its unique characteristics and requires special personalized treatment. Users must be able to provide to the system accurate human-made transcription of the document that desires to transcribe, and after that, they must build a model that is intelligent enough in order to decode the handwriting types that the document includes. In short, we are concluding that this platform does not have one but several selling points. The technology that Transkribus platform is handling can provide it is entirely advance HTR technology for academic and research needs, while the increased accessibility of the transcribed documents through the servers at the University of Innsbruck ensures that this project will keep bringing together the scientific and the technological world.

Although besides platforms advanced features, competition in marketing approaches today seems to be an essential issue for every service platform. So far there is currently no service worldwide which offers services similar to those of Transkribus. In the same context, nowadays there are many high-level companies out there that can possibly outcast Transkribus. E.g. In October of 2014, Google announced its plan to digitally convert fifteen

million historical books and distribute them for free online until the end of 2015. In this scenario, if such a tech giant decides to enter the archival field and offer free digitization and provision of digital surrogates like Transkribus, profoundly the commercial success of the platform would change dramatically. Although if the involvement of Google to archival field seems utopian, today more and smaller companies seem to compete Transkribus and trying to provide the same services on HTR field. Companies like the Dutch Picturae and the project MONK of the University of Groningen, gradually adapting the same marketing strategy as Transkribus project and appeal many scholars, volunteers, and tech experts. Nevertheless, Transkribus team feels confident that in the future will be able to find attractive strategies to answer those challenges adequately (Sánchez et al., 2013).

The technology behind Transkribus

Transkribus is a comprehensive platform for the computer-aided transcription, recognition, and retrieval of digitized historical documents (Seaward & Kallio, 2017; Sánchez et al., 2013). This platform has the ability to participate and benefit from the most recent researches in Handwritten Recognition. The primary user interface of this platform is provided through an open-source desktop application (Kahle, Colutto, Hackl, Muhlberger, 2018). Through this desktop application, users have the chance to use several tools for document image analysis and also collaborate with other users on the transcription field and share results. Furthermore, Transkribus is a client-server system where users can carry out scan-based operations through a desktop client in order to produce the ground-truth versions of the documents that they are working with. Archivists use the term ground-truth in order to state that a document is in its absolute truth condition and that it is not an object for further analysis, meaning that the digital transcription is a perfect and reliable reproduction of the information on the digital image of the original manuscript.

Moreover, Transkribus, as a client-server system, uses a protected network cluster based on multiple computers and can distinguish one or more clients on the server in order to execute the commands of the users in real time without problems (Kahle, Colutto, Hackl, Muhlberger, 2018). This means that the software itself has not the capacity to manage uploaded files, and the whole process is performed on Transkribus' central server; the user sees a projection of that on their screen. When users are working on the platform's interface on their personal computers, they are sending requests and expecting to receive responses

from the system's server simultaneously with other users of the platform. Transkribus ensures that users' data are all stored on a central repository and the main advantage of this is that training data can be shared anonymously without sharing the actual content of the documents to the server. Thus, the desktop application of this platform is associated with a server application, allowing users to co-operate on their work and achieve progress on each inventory number with a version history option available in the system.

Transkribus' programming code is written in Java 8 language and is based on the graphical Standard Widget Toolkit (SWT) (Kahle, Colutto, Hackl, Muhlberger, 2018). The Java language can be used from software developers for the creation of complete applications that can be allowed to run on a single computer or be distributed among servers and clients in a network (Lewandowski, 2015). Transkribus' desktop application operates completely offline, so users need only internet access in order to upload image files of the original manuscripts to the server. After this Transkribus can sufficiently store users files into its file system and loading image data offline. The platform supports various types of image files like PNG, JPEG, and TIFF. After the initial loading of the files to the server, document types transformed according to Page Analysis and Ground-Truth Elements (PAGE) format. This document format has an XML based page image representation framework and can recognize multiple image characteristics like the image borders and to layout structure the content of the page (Kahle, Colutto, Hackl, Muhlberger, 2018).

Similarly, the platform uses this format in order to store the respective layout and the transcription data of the documents and also classify them into the system by alphabetical order according to their filenames. From the moment that the document is successfully uploaded into Transkribus' server, users can start to segment the image into regions, lines, and baselines. The platforms layout analysis tool provide to users an effortless and intelligent way of automated text file segmentation. After layout analysis users can add the text content (human-made transcription) to the segmented elements of text manually. Both image and text areas are displayed in Transkribus' graphical user interface, and then the system has the aptitude for synchronizing the segmented location automatically to the cursor position in the textfile area and vice versa.

Furthermore, the Transkribus software contains a unique source code collection which is the fundamental component for every computer application and is usually created by program developers. Transkribus' source code is publicly accessible, and it was released under General Public Licence V3 (GPLV3)(Kahle, Colutto, Hackl, Muhlberger, 2018). This public license is used as a free software license and ensures that the users have the right to share and modify the software via online platforms like Github. Although, as a condition of

granting those permissions, the licensee requires contractually to agree on the policy terms of the license. If users desire to make the software or derivative works of the platform available to others, then they must provide them along with that license, and thus the new products or the new derivatives have to be subject to that license, including the availability of the source code of the software. This fact has been a controversial piece of the GPL agreement. Subsequently, many software companies, do not wish to make the source code of their products available to licensees or much less to the public at large because doing so will eliminate trade secret protection for their products. In short, software creators use the GPL as an underline foundation for their commercial products that they plan to distribute, but users or licensees will never have genuine access to the source code of the platform.

Equally important with the source code is the backend system of the Transkribus platform. Software engineers describe the term backend as "the part of a computer system or application that is not directly accessed by the user, typically responsible for storing and manipulating data." In the same way, the backend of Transkribus consists of several Java application running on an open-source Java Servlet Container backed by a database management system. That means that the users are restricted to use various features of the system in this operation mode. Once the desktop interface is connected to the backend mode, starts using users credentials and then the system gives access to user collection. Similarly, the documents that are uploaded into the Transkribus system can be absorbed by various HyperText Transfer Protocol Secure (HTTPS) and File Transfer Protocol (FTP) channels in order to organized into collections. The HTTPS protocol regulates how data are sent between user browsers and the website that users are connected to, in our case Transkribus interface (Kahle, Colutto, Hackl, Muhlberger, 2018). The FTP protocol, on the other hand, is a way to transfer files online and one of the most convenient ways to move files online. Although the collections in which these protocols creates automatically into the system are private by default and the only person that has access to them is the owner. Additionally, Transkribus allows the creator of the collection also to invite other users into the collection and work together on a specific file. Last but not least a single document can be linked to an erratic number of collections and be accessible from the total number of users that have access to the collection.

Moreover, after the ingestion process by various channels, the Transkribus platform uses an independent file - image storage utilization which assigns to a unique file key on each image and XML document of the initial file (Kahle, Colutto, Hackl, Muhlberger, 2018). This unique file key secures documents integrity and can also be used by the users in order to retrieve the original document via a series of channel request to a specific Uniform

Resource Locator (URL). Such URL requests, including the unique key of the document as a query parameter. The query parameters for collection resources are usually used in order to tailor and filter the response output of channels request. Although in the case of images, these parameters can be used from the software in order to adjust the quality of the image and to trigger fast image operations whose output is easily restored. Another essential point is the fact that the unique file keys combined with document metadata and pagination information are securely saved in the database of the system. In the same way, each time that users want to add and save a new transcription into the system, this mechanism creates and uploads an enriched version of the erstwhile XML document and generates a new file key, securing this way that the file version of each document will be preserved into the system as users modify them. Subsequently, the most recent XML PAGE file version of each page is indexed into an Apache Solr. The Apache Solr defined by experts as an open source search platform which is built based on a Java library. This open source platform provides full-text search functionality to Transkribus users. Last but not least, Transkribus' job management tool provides to its users, a cluster of worker modules which are able to process scheduled jobs and heavy tasks with the support of the backend suppliers (Kahle, Colutto, Hackl, Muhlberger, 2018).

To elaborate, the server of Transkribus exposes all its functionalities via a Representational State Transfer (RESTful API) ecosystem. The API functionality is essentially a way for the applications to borrow characteristics and data from each other. It is designed to take advantage of existing protocols and basically is a messenger which communicate with the system and receives requests, while has the ability to transfer them to the system and then returns the response back to users. Moreover, the API works as a third-party application that can ingest files to the Transkribus server, execute the given job orders such as HTR and exports the results computationally. Thus, all the server clients of Transkribus are available for working with API. The java client on the other hand and more specifically the client which is included in Transkribus' desktop application, continuously check the state of the server and is available to developers. The source code availability contains several critical components so as the REST API of the server and are accessible from users via Github, a web-based hosting service which is used for computer code.

To conclude, the Traskribus platform is a significant technological advancement especially for archival science because it is one of the most valuable medians that archivists need in order to explore the past successfully. This platform includes HTR engines which are based on supervised machine learning algorithms. By this term software, specialists want to describe an HTR engine that uses instances with known labels or "correct examples" of

documents in order to make the HTR engines more intelligent and proficient to successfully transcribe a handwritten manuscript. Particularly in collections that contain a vast variety of different writers or writing styles, HTR engines in order to provide accurate results to users, need as much training data as possible from users and by this, we mean human-made transcriptions in ground truth version. Thus, Transkribus appears to be a great combination of low and high technology at the same time. This kind of technology has the potential to reform the way researchers contract with archival collections. Archival science is in desperate need for new technological tools, and Transkribus platform appears as a significant piece of technology in this analog base science.

Interface tool analysis

Transkribus platform comprises multiple functionalities in order to assist users needs. The platform architecture includes built-in tools that are reducing user's effort for searching or using third-party applications or even seeking specialized HTR tools. In this chapter, will be addressed some of the most potent interface tools of this software, like the layout analysis, text recognition, keyword spotting, HTR+, and data exportation. Each of these tools allows users to edit and enrich the uploaded manuscripts through the platform's interface.

Layout analysis: Transkribus users in order to achieve the ultimate level of HTR automated transcription through the platform uses the layout analysis tool to mark up the segmentation of uploaded images and prepare the system for the insertion of human-made transcription. Historical handwritten documents do not follow specific layout rules, and subsequently, a layout analysis method must be invariant to layout inconsistencies and distortions of the text. This tool can comprise the production of a text block and layout region component in the layout. Segmentation can be done manually or executed automatically by Transkribus, but after this procedure, they must carry out the insertion of human-made transcription manually and also adjust with high precision to the uploaded images in order to create the ground truth version of the manuscript. Furthermore, Transkribus system can integrate mechanisms to detect text blocks and lines on an individual level but also both at the same processing stage. Especially handwriting recognition is a very time-consuming task for the system as the segmentation elements are represented by complex schema structures such rectangles or polygons. This approach analyzes the context of the uploaded file and identifies the Text regions (TR), the Lines (L) and Baselines (B) which is the only segmentation component which consists of just a line with several points as a polyline (Kahle, Colutto, Hackl, Muhlberger, 2018). Last but not least, these segmented regions are

known as elements and follow a hierarchical order as a baseline needs to be part of a line region, and a line region needs to be part of a text region. Thus, Layout analysis seems to be an integral tool for users in order to processed handwritten manuscripts in the Transkribus platform.

Text recognition: There are three types of text recognition technology that unified in Transkribus platform. As we mentioned, the platform previously includes two off-line HTR engines which are specialized in the manuscript analysis. The University of Valencia has contributed to the intelligence of these engines offering pattern recognition and human language technology(Kahle, Colutto, Hackl, Muhlberger, 2018). This technology contains Optical Models (OM) and Language Models (LM) which are essential for the automated text recognition from the Transkribus platform. Furthermore, statistical models like Hidden Markov Models ensures the proper functioning of the Transkribus system and contain information regarding the appearance and characters aspects in the manuscript (Kahle, Colutto, Hackl, Muhlberger, 2018). On the other hand, a supplementary Language Model offers to the system a possible allocation over sentences of words. Basically, it is a distribution function considering the next word in a sentence and is used from Transkribus in order to compute the probability of a potential text in the language but also to predict the next word in a sentence. During the training operation in the platform, optical and language models collaborate in order to create a custom vocabulary for each file. The recognition output clusters that arise from this cooperation, representing the most credible transcription sets and their probabilities. Every text line is represented by the system as a word-graph whose edges are linked with words. The word-graphs actually allow the usage of more advanced internal application, such as the Keyword Spotting which we will analyze further in this chapter.

Moreover, Transkribus engines including another spectrum of computational technology regarding HTR automated transcription. The second engine of the platform is developed by Combinatorial Interaction Testing (CITlab) and Planet intelligent systems GmbH and is based on Recurrent Neural Networks (RNN) a function which works independently of optical or language models (Kahle, Colutto, Hackl, Muhlberger, 2018). This network allows the Transkribus system to recognize handwritten digits with great accuracy. Neural networks are inspired by human brain function and actually is a cluster of artificial networks which produce human-like intelligent and have the ability to make software smart and capable to understand human logic. Nowadays such technology is used in every virtual assistant but also in every computer platform like Windows and IOS. In Transkribus platform

those networks have the endowment to create a confidence matrix on every text line and character of the uploaded file, trying this way to create links between the trained systems alphabet and between the inspected position on each line in order to produce accurate automated transcription. Transkribus platform is brilliantly designed to use every information source that users provide in order to train its HTR system sufficiently. Each page, each image including segmentation settings, human-made transcriptions, and baseline levels can be processed for software advantage.

Nevertheless, the goal of the training process is to create an entirely new HTR system which will be techno scientifically unique and based on the given script. Although HTR training can also be used as an enhancement tool for existing transcriptions but with new training data. This tool allows users to establish a training set which will be used as the training base of the model and an optional test set which used from the system as experimental space in order to measure the aptitude level of the model. Furthermore, users must take into consideration some basic parameters before they proceed into the training function. Every document has its own unique demands and users must be aware of this because they should manually define some of the basic parameters for this method such as the number of epochs but also the train and the learning rate of the system. Once the system finishes with the training model production, the model is ready to be implemented on the uploaded file scan. The system has draw information from the transcription that users have manually provided and is instantly applicable for recognition processes on the segmented pages. Additionally, Transkribus provide its users with the ability to measure the accuracy of the produced model via built-in tools such as the computing accuracy and the textfile comparison between the human-made and the machine made a transcription. The system is aiming to produce custom models with a character error rate (CER) as much as closer to 0% on both sets.

To the date of this writing, Transkribus' HTR training parameters remain a complex calculation system. Transkribus platform remains an expert domain, but as HTR training is an upscale procedure, computational power must evolve in order to provide the majority of users with more accurate HTR results and new innovative ideas for the HTR files interpretation.

Keyword spotting: Keyword spotting (KWS) has drawn the attention of the research community as an alternative means to solve difficult cases of handwriting text recognition. This tool is a standard full-text search option that is implemented in the Transkribus platform.

Keyword spotting gives users the ability to search easily distinct words in their documents and subsequently, has a broad scope of application and is divided into an offline and an online stage. In the offline phase, the system extracts information from the word images, and text lines from the pages and they represented by feature vectors into the system. In the online stage, users have the ability to formulate a query either by choosing a tangible example from the collection or by typing a keyword into the systems search bar. This is an intensive computing task, and the final results of this operation are stored in the Transkribus server in order to be accessible from the users any time. As in the text recognition method, users once again must be aware of the appropriate feature selection for this operation; otherwise, a faulty element selection can have a significant impact on the performance of the keyword spotting system. In a more straightforward interpretation, this system is an intelligent extension feature of Transkribus platform that allows users to spot words or regular expression easily into a digital scan without the necessity of human or machine made a transcription. The results of this procedure are displayed into Transkribus' interface projection as a list that gives users information about the word region in the text, the number of the page that the word had been found and the automated machine made transcription in which the word is embedded. Although this tool includes another uncommon mechanism which is called the confidence level. The confidence level signifies the frequency of potential confidence intervals that include the actual value of the unknown population parameter. In other words, it is a mechanism that gives the user the ability to adjust the systems intelligence and particularly in the Transkribus platform a confidence threshold is a figure between 0 and 1. When the confidence threshold is above 0.5, the system can be considered as very confident in spotting keywords which are matching with the search query. On the other hand, if the confidence level is below 0.1, this means that the software is able to expose more potential matches for the keyword, but it will be less sure about those matches.

HTR+: We have already mentioned the advanced capabilities of the HTR system which are highly involved in the automated text recognition process. The HTR+ in the Transkribus platform is a new approach of HTR technology inspired by the CITlab team of the University of Rostock and actually is a powerful update of the previous HTR system. In order to evolve this technology, the Transkribus team used Google's advanced machine learning library making this way its algorithms even smarter. The TensorFlow system has been created by the Google Brain team and is an open source library for numerical

computation and large-scale machine learning. This platform has the ability to perform a complex aggregate of machine learning, deep learning models and algorithms.

Furthermore, it uses Python in order to provide a suitable front-end Application Programming Interface (API) for building applications while performing those applications in the high-performance C++ programme language. The factor that makes Tensorflow such a major contribution to Transkribus platform is the fact that is designed to execute train operations and run deep neural networks for handwritten digit analysis, image recognition, word embeddings, recurrent neural networks, machine translation, and natural language processing. Although, there are significant technical differences between HTR and HTR+. Except for Tensorflow usage, there are two other significant changes, first on hardware which changed from Central Processing Unit (CPU) training to Graphics Processing Unit (GPU) and second in Neural Network function which now operates in a more profound and larger computational scale. According to Transkribus, this tool can accelerate the training speed of the models by a factor 10 to 100 and perform CER reduction by 50% to 75%. To summarize, this upgrade allows users to produce better HTR models faster and more accurate than before. The technological improvement of the Transkribus algorithm is an ambitious movement that makes platforms potentiality even more attractive for scholars and archivists.

Data exportation: Document exportation tends to be one of the most essential features and a major selling point for many software out there. Transkribus supports a wide variety of export formats and allows users to generate various output formats like Page XML, Metadata Encoding and Transmission Standard (METS), Portable Document Format (PDF), Text Encoding Initiative (TEI), Office Open XML (DOCX) and Excel Sheet format. Users have the option to export the files using the server or client exportation. Server exportation is a relatively easy and quick task as the entire operation takes place into Transkribus server. After the end of the process, the user will receive a mail with the server location link, in order to download the exported files.

On the other hand, the client exportation is the process in which the files are exported directly into users personal computer. There are significant differences between those approaches, but it is crucial to focus on the exportation functionality itself as an individual part of Transkribus application. The fact that the users have the ability to export their own files in every possible format make this platform a truly pioneering system. Exportation is an important issue especially for archivists because it contributes to archival preservation and access. This option is already providing many individuals and archival institutes with the capability of displaying unique transcribed documents for the first time, e.g.: the national

archives of the Netherlands have already contacted research about the ways that transcribed documents can be presented online. Last but not least, exportation functionality is crucial for this project as the archivists must decide in which format the exported data will remain safe and presentable to the public audience.

Thus, It has been shown that maintaining and scaling Transkribus platform is a consistent commitment as for Transkribus team so as for the University of Innsbruck. This platform is expected to reach crucial breakthroughs in the near future and lead the way for a new digital era in the computational handwritten text understanding. The interface tools that Transkribus holds is a major selling point for the platform itself so as for the third party applications which are taking part in the system. Hence, this project creates a sustainable virtuous ecosystem which describes as a transcription and recognition platform.

Conclusion

From all the above it is evident that this chapter deals with a unique initiative for the future of Handwritten Text Recognition. Transkribus platform within a short period of nine years has managed to reform the way that current archivists use HTR technology.

This piece of machinery will give the opportunity to the scientific audience to develop further research methods and establish brand new ways of operating and extracting information from handwritten archival manuscripts. What counts more is the fact that Transkribus is a powerful medium that users have free access and by this, they can reach the majority of modern technological discoveries on HTR field as part of the Transkribus platform. Transkribus offers a variety of advanced tools to its users, including:

1. Archiving of text collections and associated scans or transcriptions
2. Enrichment with metadata
3. Automatic and manual segmentation of the text
4. Tag setting, commenting and annotation
5. Automatic transcription
6. Use of automatic HTR functions

7. Training your own HTR model for a specific typeface
8. Error rate measurement of HTR and OCR

Every tool offers users a different editing perspective, and they are all equally compelling. Transkribus, in order to provide such advanced text recognition and analysis, incorporates advanced Machine Learning algorithms and Natural Language Processing. Furthermore, the system contains a unique piece of source code and engines that exploit the power of Neural Networks. This complex computational structure in combination with the vision that project READ delivers are the primary reasons why this system is such an ambitious initiative for the archival society.

Hence, we are concluding that Transkribus is an exceptional piece of technology because it is constructed under technologically high- end specifications as its own ecosystem can accommodate almost every piece of machinery available today. Transkribus from a technical perspective might not be perfect yet but the growth that the platform has shown into the last decade guarantees the future prosperity of the software. Last but not least, it is fair to acknowledge that Transkribus project is indeed a powerful stepping stone that will inspire a new archival era as has already transformed the ways that most of the European Archival institutions contacted research on handwritten material.

Digitization Initiatives In Digital Humanities

Introduction

Nowadays, digitization has become such a buzzword concerning galleries, libraries, archives, and museum (GLAM) institutes. Digitization regards to constructing a digital representation of physical objects. In other words, digitization is about converting something non-digital or analog into a digital reproduction or artifact (Maurya, 2011). Additionally, as the technology that digitization method requires has become quite affordable, multiple workshops and symposiums are becoming more and more popular. Many universities, archives and national libraries around the world have their own digitization programs for smaller scale projects while conduct research in order to explore the technological spectrum

around the digitization process (Swanepoel, 2008). Although a significant factor that contributes to the popularity of this technique is that digitization usually perceived from archivists and scholars as the ultimate solution for the challenges that arise from preservation and access (Maurya, 2011). In a technologically advanced era, digitization initiatives have the privilege to offer for the first time in history important scientific and literacy material freely via the Internet (Swanepoel, 2008). Nevertheless, a big question that emerges is: Why we should digitize the analog material available today?

In the past twenty years, we have seen a cataclysmic change in the ways that scholars and the broader public have access and use archival manuscripts. Among librarians, scholars, and other users, the consensus is that the increased availability of digital surrogates is a positive thing, and digital technology is frequently portrayed as a means of democratizing the archive (Tanner, 2015). Although there are two reasons why the available archival material is digitized. Firstly, preservation seems to be a very important concern for archival science today. As society progressively moves forward into new aspects of archival treatment, preservation has become an essential principle for more and more institutes and universities. Digitization initiatives contribute to the security of digitized material and at the same time provide long term preservation. Secondly, the enhancement of archival access is another essential benefit of digitization as billions of people have internet access and the majority of digitized material is available online. Subsequently, countless digital libraries have arisen since the beginning of digitization initiatives. This phenomenon is the product of digitization initiatives around the world and is a new approach regarding what could be possibly retrieved by browsing and exploring the internet (Swanepoel, 2008).

Thus, in order to observe applicable digitization methods, this chapter deals with the analysis of two of the most important digitization initiatives that currently take place in the Netherlands, project Republic and project Triado. Each of those two projects investigates a different aspect of digitization practices and makes possible for archivists to conduct large corpora of digitized data multidisciplinary and thus creating new research perspectives.

Project Triado

TR.I.A.D.O (Tribunaalarchieven als digitale onderzoeksfaciliteit) is a comprehensive project that investigates the way in which the most wanted Second World War archives of the Netherlands will become digitally available and searchable online. This project aims to

develop an efficient and productive method that can convert large quantities of unstructured, analog data from archival collections into usable digital research data (Triado, 2018). In other words, the institutes try to improve the digital access of the documents, enrich the available sources and making them useful for purposes. TRIADO project is funded by the Royal Dutch Academy of Sciences (KNAW) since 2016 and in order to reach its goals, corporates with four other partners including the Institute for War, Holocaust and Genocide Studies (NIOD), the Network of War Sources (NOB), the Institute of History and Culture research Huygens ING and the National Archive of the Netherlands. This project started in January 2017 and it will run until July 2019. Within these two and a half years, institutes will try to make the files of WW2 findable and available from a digital perspective but also bridge the gap between advancements in the field of Digital Humanities and collection-based institutions (TRIADO, 2018).

In the Netherlands, there are more than four hundred organizations that hold documents from WW2 in their collections. Although only a small percentage of the original source material is digitized and an even smaller fraction is searchable online. Through digitization, researches aiming to investigate on a much larger scale specific details that historical files and audiovisual material contains and finally creating new links between historical facts that took place in the war period and explore the deepest insights of the archives. Furthermore, digitization also makes the research more effective regarding patterns of violence, collaboration, and repression in this study case. Although two important questions must be answered through this inquiry are:

1. Which digital techniques are most suitable in terms of quality for the digitization aim of the project?
2. On the basis of the chosen access system, is it possible for researchers to answer in statistical scientific research questions?

Researchers want to provide sufficient answers for both questions but the main collection that TRIADO project focuses on is the “Centraal Archief Bijzondere Rechtspleging” (CABR) collection. The CABR files contain information that falls under the General Data Protection Regulation in the Netherlands. This collection, managed by the National Archive and is the most consulted WW2 archive in the Netherlands and is interesting because it reflects a unique form of justice, created after a disruptive period of large scale violence. It comprises 4.5 kilometers of procedural documents and appendices of approximately 300,000 persons who, after the Second World War were suspected of collaboration with the

enemy. That is why this archive is limited in public. Access can be obtained under certain conditions (Klijn, 2016).

In 2016, the Network of War Sources (NOB), together with the National Archives collaborates with the Centre for Language Studies department of Radboud University manage to conduct exploratory research on the quality of Optical Character Recognition (OCR) and named entity software. The main object that was under investigation was the CARB collection. The investigation showed that if researchers use a standard software approach for the majority of archival examination, text can be properly machine readable, and at the same time searchable. Subsequently, this approach allows institutes to develop their own research strategies and finally give answers to more complex statistical research questions. Through the OCR analysis archivists are able to discover important elements and accelerate their own research. The use of such technology can reveal important historical aspects and actually lead the investigation into new scientific findings (TRIADO-voorstel, 2016).

Triado's project workflow is divided into four simple steps:

1. The first step is the digitization of the available material. Thanks to high-end scanners researchers are able to scan meters of the archive material and produce metadata in accordance with the standard quality standards of the National Archive (TRIADO-voorstel, 2016).

2. The second step is auto sortering. Through this software-based procedure archivist filtering the information that the archival material contains and sort the more targeted use of transcription software for specific types of documents (TRIADO-voorstel, 2016).

3. The third step is the transcription. In this stage, archivists have to turn approximately 100,000 scans into machine-readable text. For the needs of this step, experts use ABBYY FineReader 11 an operating system which will allow researchers to convert the available scans into ALTO format. As for the handwritten material, the institutes will experiment with a different workflow via the Transkribus tool (TRIADO-voorstel, 2016).

4. The fourth step deals with the enrichment of the archival material. Using external data like existing lists of names of persons and organizations a post-correction is made on top of the existing machine-readable text. The enrichment process applied to the files with a standardized approach and this way at the end of the workflow

there is an enriched set of digital research data remains that can be used for further research. Last but not least, the findings that resulting from this method are included in a research report, which is intended to investigate the process from analog archive to usable digital archive (TRIADO-voorstel, 2016).

Through this well-organized workflow, institutes desire to create a credible system for the extraction of valid data sets. In those steps, archivists seem to take advantage of every existing method and technology about the file processing and according to the official statement of TRIADO project, the results will be publicly available in order to inspire further digitization initiatives in the archival field.

Project Republic

Republic is a digitization initiative that has as a center of its interest the Republic of the Seven United Netherlands. According to history, the Dutch Republic was one of the world's surpassing powers in the long 17th century which after the so-called Resolutions between the period 1576 - 1796, ruled by the representatives of the provinces the States General. These rulers governed both society and the colonial empire for two hundred and twenty years from 1576 until 1796. The republican form of the Dutch government itself was something exceptional at the time. The country was led not by a monarch, but by delegates of the Provinces, the States General. During the period 1576 - 1796 the resolutions and the unbalanced political condition of the Dutch government were reasons which played an influential role in the political status of Europe (Republic (Samenvatting), 2018).

Although, even today the resolutions of the Dutch nation are used for restricted research on specific topics, locations, and people. The so-called "resolution" of the Dutch nation and the accessibility difficulties that historians facing in order to find informative material about this topic even online, is the reason why this initiative takes place now in the Netherlands. Republic research plan is a project of historical interest and an initiative of Huygens ING humanities center which is currently funded by the Netherlands Organisation for Scientific Research NWO with 2.5 million euros. This project aims to perform an extended series of research in order to conduct a comprehensive investigation about controversial historical facts so as about historical, political and economical vaguenesses in Dutch history (Huygens ING, 2017).

Nowadays, the Dutch republic is a topic of tremendous interest to the scholarly and archival community globally. Through this project, the experts of both institutes will try to provide credible answers to fundamental questions regarding the Dutch nation. Questions which are relating to early modern institutional innovation, political reconstruction, “regime change”, networking, politics, language, and representation. Republic initiative is planning to provide access to digital images of approximately half a million pages of handwritten and printed political and historical material of the Dutch Republic. In order to secure the access of the digitized content REPUBLIC will perform automated transcriptions and structured text that will be developed by both machine learning and human correction. This will allow historians to provide answers to numerous critical questions about the Dutch golden age. The investigating teams of both institutes planning to develop a powerful algorithm that will automatically identify languages, dates, named entities, and certain text elements, while that information will be automatically available in an annotation layer (Republic (Samenvatting), 2018).

Last but not least, a framework for structured data will provide to users long term accessibility to the content of the resolution files by topics. These revolutionary approaches and tools will secure adequate access to the complete collection of resolutions passed by the States-General, 1576-1796. Thus, REPUBLIC will present an online, Open Access Research Environment that will empower and motivate the research by historians, political scientists and all those interested in the practice and evolution of politics, both in the Netherlands and abroad (Huygens ING, 2017).

Conclusion

In conclusion, digitization has revolutionized the way that archivists and historians perceived how information can be stored and preserved. Digitization initiatives and initiate collaborative programs such as TRIADO and REPUBLIC will revolutionize the future access of the archival material. On the other hand, through such initiatives, there is a major contribution to the preservation of analog and digitally born material for the sake of informative posterity. Nowadays, thanks to digitization and modern technology, collaborative programs extend far beyond the traditional modes of preservation and access. Thus, as long as the primary aim of the digitization initiatives will be the long-term preservation of the information in a form that will guarantee sustained access and dissemination in the future, archival science has the chance to broaden its horizons to a more superior and risk-free future regarding digital and analog information.

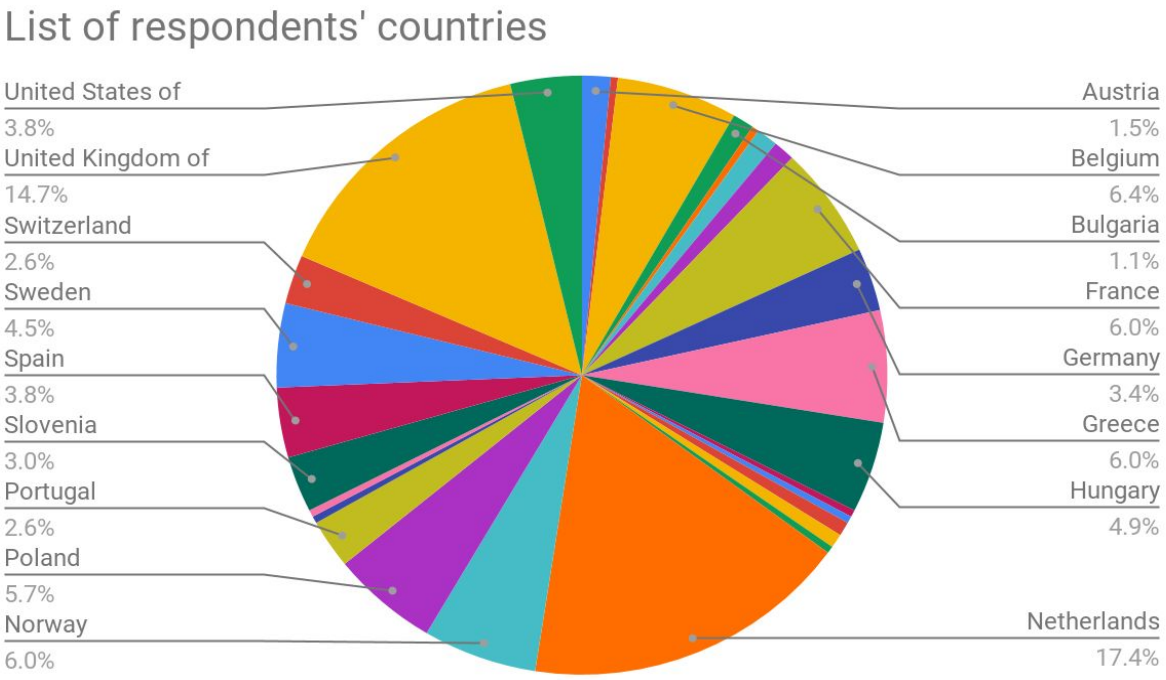
Survey Results

In order to form a general idea of how professionals of the field utilize technology, an online survey was designed and distributed. The target population was professionals with an association to Galleries, Libraries, Archives & Museum (GLAM) institutions. The questionnaire consisted of 27 questions about demographics, personal use of technological tools and the Internet, the use of technology in their workplace, their opinion about HTR software, as well as the needs and future of the profession. Last but not least a significant part of this questionnaire deals with the relation between archivists and digital archives as with the level of trust which they have regarding online and digital records.

Methodology

After the questionnaire was formed, it was distributed online in an online questionnaire form. Due to the irregularity and scarcity of the target population, a non-probability sampling procedure was chosen, and specifically, convenience sampling. The survey was distributed to major GLAM institutes, and specifically archives, located mainly in Europe and North America. The majority was selected from formal lists published by the Archival Portal Europe Foundation, as well as from acquaintances of respondents who were asked to share the questionnaire to peers in the field. The responses were anonymous.

The sample consisted of 308 respondents, of which 268 were associated with GLAM institutes (233 with archives, 58 with museums, 55 with libraries and 7 with galleries - some of the respondents were associated with more than one institution). The most prevalent age groups of respondents was 35-44 (26,9%) and 45-54 (27.3%), followed by 25-34 (18.8%) and 55-64 (16.6%). 48,9% of the respondents obtained a Master's degree, and 22.0% a doctorate degree, with the rest 25.4% having a 4-year degree or lower. The percentages of each of the most prevalent percentages can be found in Figure 1. Some of the institutions involved, among others, the Dutch National Archives, the Swedish National Archives, National Archives of Iceland, and the University of British Columbia.



Results and discussion

As far as personal use of technology is concerned, the answers were quite resolute. 96,7% of the respondents, as expected, use at least one personal digital device (smartphone, personal computer, laptop, tablet) and an astonishing 99.2% stated that they use the Internet daily. On the other hand, when asked about the credibility of information retrieved from the web, 75.2% of respondents were ambivalent in their response, saying that it "might or might not" be credible, when a 17.8% said that it is probably credible. From the above, we can conclude that even though the vast majority of professionals use the Internet daily, there is still uncertainty regarding its credibility of information. Another interesting item regarded their preference on saving their own private documents, with the majority claiming that they prefer to keep their documents saved offline on their computer, rather than on a cloud service online or in an analog form (see Table 1).

"I prefer my document collection (photographs, documents, books, music) to be stored…"

	Frequency	**Valid Percent**
Online, on some kind of cloud service	79	32.64%

Offline, on my computer (or hard disk)	114	47.11%
Offline, on a physical/analog form	49	20.25%
Total	242	100.00%

Table 1. Preference of document storage.

Half of the respondents (52.8%) stated that they can access archival material online somewhat easy, with another 25.1% stating ambivalence on the matter (neither easy nor difficult). Furthermore, the vast majority (92.8%) supports that the Internet and new technologies have made access easier. However, another 37.0% believes that online technology and the Internet can be a threat to archival integrity. Same results came when respondents were asked about digital security, with the majority (45.5%) agreeing that digital collections lack security, compared to analog forms of collections. Their attitude on digitization for preservation and access can be seen in Table 2. It can be argued that even though there are concerns regarding the integrity and security of collections, the majority of professionals believe that digitization is the best solution for preserving analog collections. I believe that digitization [the process of transforming analog material into binary electronic (digital) form] is an optimal solution for the preservation of analog collections.

"I prefer my document collection (photographs, documents, books, music) to be stored…"

	Frequency	**Valid Percent**
Online, on some kind of cloud service	79	32.64%
Offline, on my computer (or hard disk)	114	47.11%
Offline, on a physical/analog form	49	20.25%
Total	242	100.00%

Table 1. Preference of document storage.

Although half of the respondents (51.6%) have used OCR technology before, only 9.4% have used HTR technologies, while 22.4% have never used any of these technologies. 18.1% were familiar with the Transkribus software, from which 25.0% thought that it was user-friendly (Table 3). A vast 83.7% believes that HTR technology will help future research.

I believe that digitization [the process of transforming analog material into binary electronic (digital) form] is an optimal solution for the preservation of analog collections.	
	Valid Percent
Strongly agree	24.79%
Somewhat agree	35.47%
Neither agree nor disagree	16.67%
Somewhat disagree	14.53%
Strongly disagree	8.55%
Total	100.00%

Table 2. Attitudes on digitization

What is more, 74.6% of the respondents disagree with the fact that the digitization and initiatives pose a threat towards the archival profession, opening a very interesting discussion towards the future of the profession and the new pathways towards professionals should put their focus on, in order to keep up with the needs and new possibilities in the field.

Last but not least, the respondents were asked their opinion about the difficulties and future of the archival profession. An impressive 68.7% believes that their respected archival institutes do not receive adequate (governmental or private) funds to support their preservation and access efforts, when at the same time 83.6% support the view that financial resources have a significant role in the advancement of archival science.

An impressive 85.5% believes that the archival profession is misunderstood by the many. This brings up another very important theme that seems to come up repeatedly in all kinds of archival research, namely the re-branding of the profession in order to bestow all of its aspects and attract more people towards it, as innovation is believed to also play a significant role in the development of archival science (68.9% of the respondents find innovative ideas to play a significant role in the advancement of the field).

Conclusion

This master thesis is an extended research which investigated some crucial aspects of archival renewal practices towards archival institutes and aimed in giving answers in critical questions regarding digitization practices and innovative tools that are being used at the moment.

Firstly, this research was concerned with digitization and more explicitly with the advantages and disadvantages of digitization practices for preservation and access both in an individual and institutional level. Digitization helps archival institutes to create infrastructures for further accessibility regarding the new technological standards. These practices allow both archivists and the public to interact in a digital manner with the past via digitized material. Digitization in archives turns records into an instantly accessible source which can be accessed through multiple technological devices regardless of place and time. This material is easily deposited as it does not require physical space for storage and transferable across devices allowing this way users to manipulate and control the information more efficiently. Furthermore among the advantages of digitization is included the mass availability of digitized information across the cyberspace, the fact that digitization strengthens the spread of knowledge as a universal language that technological devices can understand and display, and more importantly, offers a great range of tools that users can work with, such as the ability to enlarge scanned manuscripts or save them to their personal devices.

On the other hand, digitization is not always the optimum solution regarding archival preservation and access. Digitization is an expensive and time-consuming operation which requires high-end machines and many institutes around the globe cannot respond to this challenge due to funding issues. Moreover, digital media do not comprise by anything more than electronic signals of ones and zeros. Materiality in archival science is a controversial issue among archivists as digital media are copies that creating a series of metadata which trigger copyright restrictions. Last but not least, trustworthiness and security of digitized records is a matter that creates many opposing views both for archivists and users as the frequency of digital crimes and frauds across the cyberspace is becoming more and more intense.

Secondly, another aim of this research was to analyze new tools that do not focus merely on preservation and access but are also able to provide additional information levels for further analysis. This research specifically investigated new tools that the Transkribus

software brings to the archival science. Transkribus, a software that generated from the European project R.E.A.D, brought a new spectrum of new tools (prominently through handwritten text recognition) that enable researchers and practitioners to examine multiple levels of information that was in some way hidden in archival manuscripts. The introduction of HTR recognition via this software triggers archivists' motivation for further analysis of records. This platform, powered by advanced machine learning, can bring ingenious functions applicable to manuscripts, like layout analysis, text recognition, keyword spotting, and data exportation; all of which are complex tasks to be performed manually by archivists and institutes. Another unique feature of the software, the HTR model manufacture, allows archivists and users to train their own data in order to create an intelligent model capable of recognizing different languages, characters, and handwritten styles*. Apart from its disadvantages, we can conclude that, through a new way of digitization, the Transkribus platform can not only provide solutions for further accessibility and preservation but also offer new perspectives and features that were not analyzable with such ease before. In short, these innovative tools have already laid the foundation for a richer digital future for archival material, through which new types of data and metadata that can be available for archivists and end-users.

Thirdly, another objective of this research was to shed light regarding which of the modern tools or practices that were mentioned previously, are being used at the moment in numerous digitization initiatives in the Netherlands. For this purpose, an analysis was carried out regarding two important digitization initiatives that are currently taking place in the country, project TR.I.A.D.O and project REPUBLIC. Both projects incorporate remarkable technological practices which are used for archivists in order to deepen into archival artifacts. More specifically in TR.I.A.D.O project, the usage of OCR and HTR technology so as the extended practice of Transkribus platform and the enrichment of archival material utilizing external data, were essential methods which took advantage of modern technological standards and applied them in collections, aiming to extract valid data sets. For the needs of this project, archivists from various institutes turned nearly 100.000 scans into machine-readable text converting them into ALTO format, which is one of the handiest and accessible formats for both archivists and software developers. This technique in combination with the advanced machine learning abilities that incorporates Transkribus in its tools, provide archivists with a remarkable range of high-end utilities, which are intelligent enough to examine and analyze every single aspect of the documents, while at the same time provide valid outputs.

In the same way, project REPUBLIC incorporates sophisticated technological tools in its agenda, which aim to develop a powerful algorithm that will automatically identify text elements and essential components of the records. In this project, both archivists and computer developers aspire to perform automated transcriptions for the available material so as to produce structured text which was developed by advanced machine learning algorithms under human supervision. This research approach shows many similarities with the previous initiative, although, in this project, Transkribus was not the software of choice. To summarize, both projects incorporate sophisticated technological tools such as unsupervised machine learning and automated transcription, and prove that advanced technological evolvement in archival science is not just an assumption, but an undisputed fact of current digitization initiatives.

Finally, this research in order to be in a position to provide valid answers regarding archival experts perception and opinion of what the present and the future of archives looks like, includes a 27 question survey that consisted of 308 respondents in more than 17 countries. The results projected that the majority of respondents:

A. Prefer to store their documents in an offline environment.

B. Consider that the internet has enabled easier access to archival material.

C. At the same time, they feel unsure about the informational credibility of cyberspace, and

D. Think that the internet is a threat to archival integrity.

E. Believe that digital archival collections lack security comparing to analog ones.

F. Affirm that digitization is the best solution regarding archival preservation.

G. Sense that the archival profession is misunderstood by the many.

These results, reflect concerns and beliefs that modern archivists possess. The aim of this survey was to clarify if modern archivists believe that technology is a valuable asset to the archival community and, eventually, if they feel ready to become part of this technological change in the archival landscape. Current archival professionals do believe that digital innovation is the optimum way that leads archival components like preservation

and access into a new sophisticated level however, they seem particularly cautious about digital archives' credibility and trustworthiness. Most of the candidates prefer to store their data in a secure environment as they perceive that cyberspace is an obscure and hostile environment for archival information. In short, according to survey outputs, current archival professionals, on the one hand, understand the need of the digital advancements in the archival field, although, on the other hand, they believe that the Internet and modern informational practices are still not suitable for the management of the archival material.

Discussion

In this part follows a discussion regarding the contribution of individual archivists to archival science. The modern archival scene has significantly changed the last decades causing multiple disagreements between beliefs and possible future plans regarding the progress of archival science. What this research tried to transmigrate in the modern archival community is: The archival profession is a demanding challenge between the present and the past. Archivists must be able to educate themselves as the future of archival practices is going to be digital as well as analog. Digitized material and digital born archives creating an unprecedented scene where archivists must be in a position to face every possible challenge. For this reason, they should keep up with the new tools and techniques that are introduced every year in the archival community either via institutes or digitization initiatives.

Furthermore, they should be critical and protective regarding the fundamental values of our science. For this reason, archivists should deal more directly with modern archival professionals except for the academic aspect. Another essential aim of this research was to contribute to the discussion regarding GLAM institutes. Archival science like any other scientific field contains unpredictable and unconventional theoretical and technical aspects, so institutes should investigate future plans regarding digitization as it will become soon an imperative and inevitable reality for the current archival field. Additionally, this research displays the significance of the human factor. The institutes should be committed to the technological growth of our science and for this reason, they must provide the necessary human resources in order to succeed. This way, institutes can ensure the prosperity of an objective and critical mentality which will help them to soberly evaluate new tools, approaches or funds regarding archival digitization. Moreover, should be clear that each institute is a unique organism with different needs and ambitions. There is no general rule of how GLAM institutes should approach or evolve every new advancement in archives, however, a possible collaboration between institutes and archival experts for the establishment of a universal system of standards and practices maybe is the key in order to move archival science even further. In short, GLAM institutes should possess a critical mindset in order to use tools and practices cautiously as every software or algorithm is invented by the human factor and after all technology without human proficiency is just another insufficient component among humans abilities to solve problems.

Bibliography

Adam, S. (2010). Preserving authenticity in the digital age. Library Hi Tech, 28(4), 595–604. https://doi.org/10.1108/07378831011096259

Adu, K. K., Dube, L., & Adjei, E. (2016). Digital preservation: The conduit through which open data, electronic government and the right to information are implemented. Library Hi Tech, 34(4), 733–747. https://doi.org/10.1108/LHT-07-2016-0078

Allen, S. (2016). Collections Care and Stewardship: Innovative Approaches for Museums (Decker, ed.). Museum Anthropology Review,10(2), 122.

Anderson, D. (2015). The digital dark age. Communications of the ACM, 58(12), 20–23. https://doi.org/10.1145/2835856

Australia Council for the Arts. (2011). S. Langley (Prepared), J. Bailey (Ed.). Archives in the digital era: Scoping study report. Sydney.

Ball. (2005). Preservation and Conservation for Libraries and Archives. Adolescence, 40(159), 673. https://doi.org/10.1103/PhysRevLett.101.245302

Bell, G., & Gray, J. N. (1997). The Revolution Yet to Happen. In Beyond Calculation (pp. 5–32). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4612-0685-9_1

Besser, H., Chapman, S., Conway, P., Fenton, E. G., Frey, F., Gertz, J., … Vogt-O'Connor, D. (2000). Handbook for Digital Projects: A Management Tool for Preservation and Access. (M. Sitt, Ed.), Northeast Document Conservation Center (Vol. 2, pp. 1–182). https://doi.org/10.1353/pla.2002.0021

Blanke, T., Bryant, M., & Hedges, M. (2012, February). Ocropodium: Open source OCR for small-scale historical archives. Journal of Information Science. https://doi.org/10.1007/978-3-642-10625-5_41

Boyda, J. (2013). Preserving the immaterial: Digital decay and the archive. Spectator, Xxxiii(2), 1-63.

Conway, P. (2000). Overview: rationale for digitization and preservation. Handbook for Digital Projects: A Management Tool for Preservation and Access, 5–20. https://doi.org/10.1353/pla.2002.0021

Cook, T. (1996). What is past is prologue: A history of archival ideas since 1898, and the future paradigm shift. Archivaria, 43, 17–63. https://doi.org/10.1088/0508-3443/7/5/302

Cunningham, A., & Phillips, M. (2005). Accountability and accessibility: Ensuring the evidence of e-governance in Australia. Aslib Proceedings, 57(4), 301–317. https://doi.org/10.1108/00012530510612059

Dekker, A., (2010). Archive 2020: Sustainable archiving of born-digital cultural content. Amsterdam: Virtueel Platform.

Dekker, A. (2012). Born-Digital kunstwerken in Nederland. Virtueel Platform. Retrieved January 7, 2019 from https://dare.uva.nl/search?identifier=795f4ea1-846a-4f62-8392-e1797ec5d97a

Digital Preservation. (2013). Dictionary of Information Science and Technology.

Dodge, M., & Kitchin, R. (2001). Atlas of Cyberspace. New York (Vol. 93, pp. 946–948). Addison-Wesley Professional. https://doi.org/10.1111/j.1467-8306.2003.09304014_7.x

Dryden, J. (2011). Measuring Trust: Standards for Trusted Digital Repositories. Journal of Archival Organization, 9(2), 127–130. https://doi.org/10.1080/15332748.2011.590744

Duranti, L. (1991). Diplomatics: New Uses for an Old Science (PART I). Archivaria Vol. 28, 7-2. Retrieved January 20, 2019, from https://archivaria.ca/index.php/archivaria/issue/view/387/showToc

Duranti, L. (2001). Concepts, principles, and methods for the management of electronic records. Information Society, 17(4), 271–279. https://doi.org/10.1080/019722401753330869

Duranti, L., & Rogers, C. (2012). Trust in digital records: An increasingly cloudy legal area. In Computer Law and Security Review (Vol. 28, pp. 522–531). Elsevier Ltd. https://doi.org/10.1016/j.clsr.2012.07.009

Edwards, M. (2015). Some Darker Sides of Digitization; or, Disappearing Data, Doubtful Descriptions, and Other Deformations of Print. Style, 49(3), 321-333.

Garrett, J. & Waters, D. (1996). Preserving digital information: Report of the task force on archiving of digital information: Commissioned by the Commission on Preservation and Access and the Research Libraries Group. Washington, DC: Commission on Preservation and Access, 1996. 59 pp. $50.00 softcover. ISBN 1887334505. (1997). Library Acquisitions: Practice and Theory, 21(3), 413-414.

Hansen, L. E., & Sundqvist, A. (2012). To Make Archives Available Online: Transcending Boundaries or Building Walls? Journal of Archival Organization, 10(3–4), 207–230. https://doi.org/10.1080/15332748.2013.795784Z

Hayles, N. K. (2003). Translating Media: Why We Should Rethink Textuality. The Yale Journal of Criticism, 16(2), 263–290. https://doi.org/10.1353/yale.2003.0018

Hilbert, M., & López, P. (2011). The world's technological capacity to store, communicate and compute information. Science, 332(6025), 60–65. https://doi.org/10.1126/science.1200970

Hughes, Lorna M. (2004). Digitizing Collections: Strategic Issues for the Information Manager. London: Facet Publishing. ISBN 1-85604-466-1. Chapter 1, "Why digitize? The costs and benefits of digitization", p. 3-30; here, especially p. 9-17.

Huygens ING receives NWO-large of 2.5 million euro for REPUBLIC. (2017, September 28). Retrieved March 12, 2019, from https://www.huygens.knaw.nl/huygens-ing-ontvangt-een-nwo-groot-van-25-miljoen-euro-voor-republic/?lang=en

Irons Walch, V. (1990). Preservation Standards: Checklist of Standards Applicable to the Preservation of Archives and Manuscripts. The American Archivist, 53(2), 324–338. https://doi.org/10.17723/aarc.53.2.k2p3m81751465218

Irons Walch, V. (1994). Standards for Archival Description: A Handbook. Society of Amer Archivists. ISBN-10: 0931828961

Factor, M., Henis, E., Naor, D., Rabinovici-Cohen, S., Reshef, P., & Ronen, S. (2009). Authenticity and Provenance in Long Term Digital Preservation: Modeling and Implementation in Preservation Aware Storage. In TAPP'09 First Workshop on Theory and Practice of Provenance. San Francisco, CA. https://doi.org/http://portal.acm.org/citation.cfm?id=1525938

Fleischer, S., & Heppner, M. (2009). Disaster Planning for Libraries and Archives: What You Need to Know and How to Do It. Library & Archival Security, 22(2), 125-140.

Forde, H. (2007). Preserving archives (Principles and practice in records management and archives). London: Facet Publishing.

Forde, H., & Rhys-Lewis, J. (2013). Preserving archives / (Second ed., Principles and practice in records management and archives). London: Facet Publishing.

Garaba, F. (2015). Dodos in the archives: rebranding the archival profession to meet the challenges of the twenty-first century within ESARBICA. Archives and Records, 36(2), 216–225. https://doi.org/10.1080/23257962.2015.1030609

Gorman, J. (2018, September 04). The Brazil Museum Fire: What Was Lost. The New York Times. Retrieved January 2, 2019, from https://www.nytimes.com/2018/09/04/science/brazil-museum-fire.html

Hardy, M. (2018). Digitization. Early American Studies: An Interdisciplinary Journal, 16(4), 637-642.

Hedstrom, M. (1997). Digital Preservation: A Time Bomb for Digital Libraries. Computers and the Humanities, 31(3), 189–202. https://doi.org/10.1023/A:1000676723815

Holzmann, H., Goel, V., & Anand, A. (2016). ArchiveSpark: Efficient Web Archive Access, Extraction, and Derivation. Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries, 83–92. https://doi.org/10.1145/2910896.2910902

Horsman, P., Ketelaar, E., & Thomassen, T. (2003). New Respect for the Old Order: The Context of the Dutch Manual. The American Archivist,66(2), 249-270.

Huvila, I. (2014). Archives, Libraries, and Museums in the Contemporary Society: Perspectives of the Professionals. Conference 2014 Proceedings, 66(6), 45-64.

Institute of Museum Library Services. (2002). Status of technology and digitization in the nation's museums and libraries: 2002 report. Institute of Museum and Library Services.

Kahle, P., Colutto, S., Hackl, G., & Muhlberger, G. (2018). Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. In Proceedings of the International Conference on Document Analysis and Recognition, ICDAR (Vol. 4, pp. 19–24). IEEE Computer Society. https://doi.org/10.1109/ICDAR.2017.307

Kemp, J. (2015). How digitization integrates in the world of archives preservation. Journal of the Institute of Conservation, 39(1), 1-9.

Klijn, E. (2016, December 06). KNAW-onderzoeksfonds toegekend aan TRIADO-project. Retrieved March 12, 2019, from https://www.oorlogsbronnen.nl/nieuws/tribunaalarchieven-als-digitale-onderzoeksfaciliteit

Leake, C. (1960). Preserving Our Science Archives. Science, 132(3420), 158-160.

Leonardi, P. M. (2010). Digital materiality? How artifacts without matter, matter.First Monday, Volume 15, Number 6. Retrieved January 20, 2019, from https://firstmonday.org/ojs/index.php/fm/article/view/3036/2567

Lewandowski, C. M., Co-investigator, N., & Lewandowski, C. M. (2015). Java 8 in Action. The effects of brief mindfulness intervention on acute pain experience: An examination of individual difference (Vol. 1, pp. 1689–1699). https://doi.org/10.1017/CBO9781107415324.004

Lischer-Katz, Z. (2017). Studying the materiality of media archives in the age of digitization: Forensics, infrastructures, and ecologies. First Monday, 22(1). https://doi.org/10.5210/fm.v22i1.7263

Louridas, P., & Ebert, C. (2016). Machine Learning. Software, IEEE, 33(5), 110-115.

Lynch, C. (2000). Authenticity and integrity in the digital environment: an exploratory analysis of the central role of trust. Authenticity in a Digital Environment, 32–50. https://doi.org/10.1007/s11837-007-0139-8

Mancinelli, T. (2016). Early printed edition and OCR techniques: what is the state-of-art ? Strategies to be developed from the working-progress Mambrino project work Tiziana Mancinelli. Historias Fingidas, 4(2016), 255–260. https://doi.org/10.13136/2284-2667/65

Mardon, R., & Belk, R. (2018, December 1). Materializing digital collecting: An extended view of digital materiality. Marketing Theory. SAGE Publications Ltd. https://doi.org/10.1177/1470593118767725

Marr, B. (2018, July 09). How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read. Retrieved January 2, 2019, from https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#6fdefdd960ba

Matusiak, K., & Johnston, T. (2014). Digitization for Preservation and Access: Restoring the Usefulness of the Nitrate Negative Collections at the American Geographical Society Library. The American Archivist, 77(1), 241-269.

Maurya, R. N. (2011). Digital Library and Digitization. International Journal of Information Dissemination and Technology, 1(4), 228–231.

Maynard, J., & Foster, A. (2012). Preserving the Audio Arts Archive. Journal of Conservation and Museum Studies, 10(1), 59-63.

Menne-Haritz, A. (2001). Access - The reformulation of an archival paradigm. Archival Science, 1(1), 57–82. https://doi.org/10.1007/BF02435639

Mnjama, N. (2008). The Orentlicher Principles on the Preservation and Access to Archives Bearing Witness to Human Rights Violations. Information Development, 24(3), 213-225.

Moss, M., Thomas, D. & Gollins, T. (2018). Artificial Fibers—The Implications of the Digital for Archival Access. Frontiers in Digital Humanities, 5, Frontiers in Digital Humanities, 01 August 2018, Vol.5

Moyle, M. (2011). Manuscript transcription by crowdsourcing: Transcribe Bentham. LIBER Quarterly, 20(3–4), 347–356. https://doi.org/10.18352/lq.7999

Muller, S., Feith, J.A, & Fruin, R. (1921). Handleiding voor het ordenen en beschrijven van archieven : Ontworpen in opdracht van de Vereeniging van Archivarissen in Nederland (2e dr ed.). Groningen: Erven B. van der Kamp.

Nationaal, A. (2018, November 16). Politie werkt samen aan duurzame toegankelijkheid. Retrieved April 16, 2019, from https://www.nationaalarchief.nl/archiveren/nieuws/politie-werkt-samen-aan-duurzame-toegankelijkheid

Nationaal, A. (2018, May 07). Privacyreglement Nationaal Archief. Retrieved April 16, 2019, from https://www.nationaalarchief.nl/privacyreglement-nationaal-archief#collapse-6719

Nationaal Archief, Digitaliseren. (n.d.). Retrieved April 16, 2019, from https://www.nationaalarchief.nl/archiveren/digitaliseren

Nelson, N., & Association of Research Libraries. (2012). Managing born-digital special collections and archival materials / (SPEC kit ; 329). Washington, D.C.: Association of Research Libraries.

Niu, J. (2014). Original order in the digital world. Archives and Manuscripts, 43(1), 1-12.

Patel, C., Patel, A., & Patel, D. (2012). Optical Character Recognition by Open source OCR Tool Tesseract: A Case Study. International Journal of Computer Applications, 55(10), 50–56. https://doi.org/10.5120/8794-2784

Peet, Lisa. (2017). LC's new born-digital archives. Library Journal, 142(15), 14.

Powell, B. (2015). Collection Care: An Illustrated Handbook for the Care and Handling of Cultural Objects. Rowman & Littlefield.

Raab, C., & Szekely, I. (2017). Data protection authorities and information technology. Computer Law and Security Review, 33(4), 421–433. https://doi.org/10.1016/j.clsr.2017.05.002

Republic (Samenvatting). (2018, December 01). Retrieved March 12, 2019, from https://www.nwo.nl/onderzoek-en-resultaten/onderzoeksprojecten/i/05/32205.html

Ritzenthaler, M., & Society of American Archivists. (2010). Preserving archives & manuscripts (2nd ed., Archival fundamentals series II). Chicago: Society of American Archivists.

Rockembach, M. (2018). Archival appraisal: An analysis based on a systematic literature review. Encontros Bibli, 90-98.

Roberts, J. (1987). Archival Theory: Much Ado about Shelving. The American Archivist, 50(1), 66-74.

Romero V., Serrano N., Toselli A.H., Sánchez J.A. & Vidal, E. (2011). Handwritten Text Recognition for historical documents. Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage, Hissar, Bulgaria, pp. 90-96

Runardotter, M. (2007). Information Technology, Archives, and Archivists — An Interacting Trinity for Long-term Digital Preservation. Social Informatics. Retrieved January 2, 2019, from http://epubl.ltu.se/1402-1757/2007/08/LTU-LIC-0708-SE.pdf

Runardotter, M. (2011). Organizational cooperation for cultural heritage: A viable systems approach. Systems Research And Behavioral Science,28(1), 77-90.

Salter, A., & Murray, J. (2014). New Media Art. Flash: Building the Interactive Web. MIT Press. https://doi.org/10.1111/j.1467-8357.2005.00520.x

Sánchez, J. A., Mühlberger, G., Gatos, B., Schofield, P., Depuydt, K., Davis, R. M., … de Does, J. (2013). tranScriptorium: a European project on handwritten text recognition. In Symposium on Document Engineering (pp. 227–228). ACM. https://doi.org/10.1145/2494266.2494294

Seaward, L., & Kallio, M. (2017). Transkribus: Handwritten Text Recognition technology for historical documents. Digital Humanities 2017. Retrieved 27.01.2019 from https://dh2017.adho.org/abstracts/649/649.pdf

Sewdass, P. (2014). Management of library and archival security: From the outside looking in O'Neill, Robert K. ed. South African Journal of Libraries and Information Science, 66(4), South African Journal of Libraries and Information Science, 01/26/2014, Vol.66(4).

Şentürk, B. (2014). Arşivler, Arşivciler ve Arşiv Malzemesi İçin Bir Tehdit Unsuru Olarak Online Erişim: Bir Grup Arşivci Gözüyle Değerlendirme. Türk Kütüphaneciliği, 28(4), 496-509.

Shapley, M. (2015). Preserving Archives. Archives and Manuscripts, 43(2), 1-2.

Sinn, D. (2012). Impact of digital archival collections on historical research. Journal of the American Society for Information Science and Technology, 63(8), 1521-1537.

Swanepoel, M. (2008). Digitization Initiatives: a Reconnaissance of the Global Landscape. In 29th IATUL Conference. Aucland: The International Association of Technological University Libraries (IATUL). Retrieved from http://docs.lib.purdue.edu/iatul/2008/papers/5/

Sweeney, Shelley. (2008). The ambiguous origins of the archival principle of "provenance." Libraries and the Cultural Record, 43(2), 193-213.

Szekely, Ivan (2017) "Do Archives Have a Future in the Digital Age?," Journal of Contemporary Archival Studies: Vol. 4 , Article 1.
http://elischolar.library.yale.edu/jcas/vol4/iss2/1

Tanner, S. (2015, February 05). Democratisation of Collections through Digitisation. Retrieved March 12, 2019, from https://www.libfocus.com/2015/02/democratisation-of-collections-through.html

Trant, J. (2009). Emerging convergence? Thoughts on museums, archives, libraries, and professional training. Museum Management and Curatorship, 24(4), 369-387.

TRIADO. (2018, March 16). Retrieved March 12, 2019, from https://www.nationaalarchief.nl/archiveren/nieuws/triado

TRIADO-voorstel. (2016, December 13). Retrieved March 12, 2019, from https://www.oorlogsbronnen.nl/triadovoorstel

UNESCO (1996), Memory of the World; Lost memory - Libraries and Archives Destroyed in the 20th century / prepared for UNESCO on behalf of IFLA by Hans van der Hoeven and on behalf of ICA by Joan van Albada. Paris, 1996. Retrieved January 2, 2019, from https://unesdoc.unesco.org/ark:/48223/pf0000105557

Van Dijk, J. A. G. M. (2006). The network society: Social aspects of new media. (S. Edition, Ed.), Book (Vol. 17, p. 292). Sage Publications. https://doi.org/10.1586/14737175.7.7.887

Vogelsang, M. (2010). Digitalization in Open Economies (Contributions to Economics). Heidelberg: Physica-Verlag HD.

Walters, T. O. (1996). Archival appraisal methods and preservation decision-making. The American Archivist, 59(3), 322–338. https://doi.org/10.17723/aarc.59.3.w4th5pp861802870

Wang, Z. & Zhu, Y. (2011). Research on Knowledge Sharing Technology of Digital Library Based on Web 2.0. Energy Procedia, 13, 8588-8593.

Weidling, T. (2013). Den äldsta arkivläran: Jacob von Rammingens båda läroböcker i registratur- och arkivskötsel från 1571, samt en monografi om arkiv från 1632 av Baldassare Bonifacio [The oldest archival science: Jacob von Rammingen's two manuals of registry and archival management from 1571, and a monography on archives from 1632 by Baldassare Bonifacio]. Scandinavian Journal of History, 38(2), 270-271.

Weng, C. (2016). Knowledge discovery of digital library subscription by RFC itemsets. The Electronic Library, 34(5), 772-788.

Weston, P., Carbé, E., & Baldini, P. (2017). If bits are not enough: Preservation practices of the original contest for born-digital literary archives. Bibliothecae.it, 6(1), 154-177.

Wortmann, F., & Flüchter, K. (2015). Internet of Things. Business & Information Systems Engineering, 57(3), 221-224. https://doi.org/10.1007/s12599-015-0383-3

Yeo, G. (2013). Trust and context in cyberspace. Archives and Records, 34(2), 214–234. https://doi.org/10.1080/23257962.2013.825207

Zhang, Jane. (2012). Archival Representation in the Digital Age. Journal of Archival Organization, 10(1), 45-68.