Envisioning a Sustainable Future for Archives: A Role for Visual Analytics?

Victoria L. Lemieux, University of British Columbia, Vancouver, Canada
vlemieux@mail.ubc.ca

Abstract: This paper surveys recent research on the application of visualization and visual analytic technologies to address archival and information challenges of sustainability related to access, preservation, and security of archival documents, drawing, in particular on research being conducted at the School of Library, Archival and Information Studies at the University of British Columbia. The paper concludes that visualization and visual analytics are new technologies that hold much promise for a sustainable future of archives, but that much more research is needed to transfer these technologies from the lab to production systems working in archives.

Introduction: The Need for Sustainability in the Era of Big Data

Like many other organizations, archival institutions are facing a data tsunami. Data are growing massively and rapidly: About 2.5 exabytes (2.5 million terabytes) of data are created every day.[1] Ninety percent of the data in the world today has been created in the last two years.[2] Not surprisingly, it is a challenge to store and process data that is growing so rapidly. Data are also changing quickly to become more complex and less structured: According to IBM, an estimated "80% to 90% of any organization's data is what is referred to as unstructured...[and] that data is growing at 40% to 60% per year."[3] International Data Corporation (IDC) projects that unstructured data will grow at 60% plus compounded annually, far higher than transactional (structured) growth rates of 20% plus.[4] There are endless new types and sources of data: IBM cites "the rising use of interactive web technologies, such as blogs and social media platforms" as one of the big drivers in the growth of complex data. Traditional approaches to managing archival documents seem unlikely to remain tenable in the face of the oncoming data tsunami.

Visual Analytics as a new solution?

Visual Analytics is being applied in many domains to assist analysts where there is a need to process masses of complex data; answer an array of often ambiguous questions; keep humans in the loop and at the centre of analysis; blend computational analysis with interactive visualization of the results of that analysis; provide quick answers with on demand improvement of analytical results; incorporate presentation linked with analysis; and export easy to understand representations of results in real time.[5] What is Visual Analysis (VA)? VA is defined as "the science of analytical reasoning facilitated by interactive visual interfaces."[6] As the name implies, VA has to do with the visualization of information. Visualization is nothing new: pictures were used as a form of communication in early civilizations and cartographers have been making maps for centuries. Business information has been visualized in tables, outlines, pie charts, line graphs and bar charts for a long time.[7] The history of computer-aided visualization is, of course, much shorter. Arising from the computer graphics community, the study of the use of graphics for visualizing data traces its origin to the 1987 report of the Workshop on Visualization in Scientific Computing held in Washington, DC.[8] Visualization of data that lack spatial coordinates is known as Information Visualization

---

[1] IBM, "IBM PowerLinux Big Data Analytics Solutions," July 2012, accessed on 30 July, 2012 at *public.dhe.ibm.com/common/ssi/ecm/en/pos03099usen/POS03099USEN.PDF*

[2] J. Manyika et al., "Big Data: The next frontier for innovation, competition and productivity," May 2011, accessed on 30 July, 2012 at *www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation*

[3] IBM, "TAKMI: Bringing Order to Unstructured Data,"n.d., accessed on 30 July, 2012 at *www-03.ibm.com/ibm/history/ibm100/us/en/icons/takmi/*.

[4] Guest Blogger, "Huge data growth has led to new standards and powerful analytic and search tools to mine for business insights," 15 April, 2011, accessed on 30 July, 202 at *www.businessofgovernment.org/blog/strategies-font-color-redcut-costsfont-and-improve-performance/data-management*.

[5] VisMaster *Mastering the Information Age*, 2010, accessed on 30 July, 2012 at http://www.youtube.com/watch?v=5i3xbitEVfs>.

[6] J.J. Thomas and K.A. Cook, eds., *Illuminating the Path: The Research and Development Agenda for VA*, IEEE Computer Society Press.2005.

[7] D.P. Tegarden, 'Business InfoVis'. Communications of the Association for Information Systems, (1999), 1(4): 2-38.

[8] B.H. McCormick, T.A. DeFanti and M.D. Brown, eds., Visualization in scientific computing'. ACM SIGGRAPH Computer Graphics (1987) 21(6).

(InfoVis). The first IEEE InfoVis workshop was held in Atlanta Georgia in 1995.[9] It was the challenge of building an interdisciplinary approach to visualization in real-world applications that led to the creation of VA. Perhaps the most evocative call for this came from a short paper by a pioneer in computer graphics, Bill Lorensen. His 2004 paper "On the Death of Visualization" proposed that visualization researchers first "embrace our customers: find out the important problems they face." Second, they should "form alliances with other fields", and finally, they should "define some grand challenges."[10] In response to issues such as these, a panel of researchers was asked to define a research agenda for a new interdisciplinary effort to design and evaluate technologies for strategic and operational decision-making. Their efforts led to the publication of "Illuminating the Path: A National Research Agenda in Visual Analytics."[11] VA is seen as complementary to visualization in its focus on analytic task performance per se, and in its choice of a scientific methodology as the mechanism for doing so. In keeping with Lorensen's recommendations, VA places a good deal of emphasis on development of technologies and analytical methods focused on the needs of a particular set of "customers", their data, problems, working methods, and organizational structures.

This calls for a "translational research" approach.[12] The VA translational research cycle includes working with decision-makers in the context of their organizations to characterize data and solutions in the situations in which analysis takes place. This work defines research questions for laboratory investigation, the results of which guide the design of new interactive visualization technologies and analytical methods that are evaluated in partnership with the decision-makers and their organizations. This approach should enable organizations and individuals to more effectively utilize interactive visualization to solve problems that are ill-constructed, where the characteristics of a successful solution are not well-defined, and where data are massive, uncertain and changing over time. Thus, VA may be an approach that lends itself well to archival challenges in the era of big data.

Whether we are speaking of InfoVis or VA, visualization harnesses the power of human visual perception and cognition, lending truth to the old adage that "a picture is worth a thousand words." Visualization has advantages over other modes of communication because humans have evolved visual and spatial skills that include the ability to detect edges and discontinuities, things that stand out, variations in color and shape, and motion; to recognize patterns; and to retrieve information using visual cues.[13] Each of these visual and spatial attributes can be transformed into a graphical image to provide a rich visual description of data. As these features can be observed with "pre-attentive processing"; that is, they are perceived prior to conscious attention, they are understandable at a glance and much more rapidly than words.[14] Through encoding of data items and attributes into graphical images, visualizations can act as a repository of data which allows people to offload cognition to the perceptual system, using visuals as a form of virtual memory.[15] This can enlarge problem-solving capabilities by enabling the processing of more data without overloading the decision maker.[16] Visual images also have the advantage of being able to transcend the barriers of human culture since a map or a graph may be interpreted even by people who speak different languages.[17] Finally, because visual cues stand out to human perception more than words, a picture sometimes "forces us to notice what we never expected to see."[18]

Information Visualization and Visual Analytics in the Archives

---

[9] W. Wright, 'Information animation applications in the capital markets'. In Proceedings of the 1st IEEE Symposium on InfoVis 1995 (INFOVIS '95): October 30-31, 1995: Atlanta, Georgia, S. Eick and N. Gershon (eds), IEEE Computer Society Press.

[10] W. Lorensen,'On the death of visualization: can it survive without customers?'. In Position Papers of the NIH/NSF Proceedings of the Fall 2004 Workshop on Visualization Challenges: September 22-23, 2004: Bethesda, Maryland, National Library of Medicine.

[11] Thomas and Cook, eds., 2005.

[12] B. Fisher, B, T.M. Green, and R. Arias-Hernández, R. 'VA as a translational cognitive science'. Topics in Cognitive Science, (2011): 3(3): 609–625.

[13] S.M. Kosslyn, *Image and Mind*. Harvard University Press, 2005. ; N.H. Lurie, and C. Mason,, 'Visual representation: implications for decision making'. Journal of Marketing, (2007)71: 160-177.

[14] M. Ward, G. Grinstein, G, and D. Keim, *Interactive Data Visualization: Foundations, Techniqu,es, and Applications.* A.K. Peters Ltd, 2010.

[15] T. Munzner,'Visualization'. In *Fundamentals of Graphics*, 3rd ed. P. Shirley, M. Ashikhminand and S. Marshner, eds, A.K. Peters, 2009.

[16] Tegarden, 1999.

[17] Ward, Grinstein and Keim, 2010.

[18] J.W. Tukey, *Exploratory Data Analysis*. Addison-Wesley. 1977.

No systematic study of the use of visualization in Archives has been done, though this would be quite a useful contribution to the professional archival literature. Thus, the overview provided in this paper draws on illustrative examples and noteworthy recent developments in relation to the application of visualization to archival endeavours.[19]

Certainly there are many examples of archives having used visualizations to communicate information about their holdings. The most obvious example of this is to be found in the frequent use of photographs from archival collections to illustrate or "tell a story" about an archival *fonds* or to add visual interest to a web site. The use of interactive infographics is of more recent vintage. The JP Morgan Chase Archive, for example, has used an interactive infographic to convey the company's corporate history.[20] As in the case of this infographic, many of the visualizations used in archives represent and communicate archivists' analysis of an organization's history. This approach is in contrast to allowing researchers to explore archival material using an interactive visual interface in order to facilitate their own analysis, as would be the case with visual analysis.

There are increasing numbers of visualizations that have also been used to assist researchers to navigate archival *fonds*. Traditional means of browsing or displaying search results, such as lists and directories, restrict users' ability to see records in context.[21] Providing cross-references to subject keywords, functional descriptions, person, place and corporate names can only go so far in addressing this problem. For example, points at which these cross-references intersect cannot easily be displayed and do not support users' desire to navigate to another search, to repeat searches or to navigate up and down the hierarchy. This problem increases exponentially where related material is held in different series, collections or repositories. In these circumstances trying to follow a particular person, function or responsibility is extremely difficult if not impossible. In following one path, users lose sight of others, where they cross and what their relationship is. An early and common visualization aimed at overcoming such limitations is a hierarchical node link graph, which is also typically used in representing organizational charts. These graphs have been used to represent the relationship between *fonds*, *sous-fonds*, series, and items in an archival *fonds*.[22] Such representations provide a researcher with a sense of the contextual locus of particular archival record(s), but have little flexibility to support the researcher in visually exploring or re-representing archival records to generate analytic insights: the researcher is presented with the archivist's representation of the hierarchical arrangement of the *fonds*. Researchers are restricted to clicking on nodes to view the full hierarchy or to hiding hierarchical levels.[23]

The year 2005 brought new developments in the application of interactive visualizations to archives. The introduction of XML encoded finding aids, particularly EAD, and wide spread implementation of descriptive standards such as ISAD(G)2, created opportunities for the introduction of new visualization tools that leverage archival descriptive metadata. As more archival finding aids, of increasing complexity, were becoming available online the difficulty of seeing the 'wood from the trees' was increasing, particularly when these are implemented in EAD (Encoded Archival Description). To address this issue, Andersen (HATII, University of Glasgow) began exploring the use of visualizations to represent multiple "dimensions" (e.g. relationships among persons, committees, etc.) in online finding aids.[24] His project sought to visualize archival information by applying Ted

[19] Note: Michael Whitelaw's work should also be mentioned in this survey of recent developments. However, it is not included in this paper as in order to avoid duplication as this paper is presented as part of a session at the ICA's 2012 Congress in Brisbane in which Michael Whitelaw will present on his own work. For more information about Whitelaw's work with information visualization, see, for example: http://www.youtube.com/watch?v=i8JO0KkYvow.

[20] See http://www.jpmorgan.com/pages/jpmorgan/about/history.

[21] I. Andersen and S. North, "Multidimensional Visualization of Archival Finding Aids," 13 January, 2009, accessed on 30 July, 2012 at http://www.youtube.com/watch?v=i8JO0KkYvow.

[22] The following early example derives from Library and Archives Canada: accessed on 30 July, 2012 at http://collectionscanada.gc.ca/pam_archives/index.php?fuseaction=genitem.displayHierarchy&lang=eng&rec_nbr=135001&back_url=()&&back_url=().

[23] Allen (R.B. Allen, "Using Information Visualization to Support Access to Archival Records, *Journal of Archival Organization* (2005) 3,1: 37-49) suggests extending the analytic capabilities of the static hierarchy through the use of an interactive hierarchical browser. He notes that several hierarchy viewers have been developed and are now familiar in tools such as the Windows file manager. These, suggests Allen, can be particularly helpful but it can have even greater power in conjunction with search. For instance, a search on "Alaska" could give pointers to all agencies dealing with Alaska.

[24] Andersen and North, 2009.

Nelson's ZigZag™ structure to two existing EAD finding aids.[25] As Andersen explains: "ZigZag structures finding aid content as a series of cells, these are then linked together to form dimensions (for example, function), which are essentially relations between entities or objects. As each cell can belong to more than one dimension this allows a visualisation that combines selected dimensions (for example files by functional activity), but also displays other available dimensions without cluttering users current view."[26] Andersen's experiments with the use of the ZigZag visualization supports a much wider range of interactivity and analytic exploration than do hierarchical node link graphs. By using the ZigZag representation of multidimensional information, such as repository, collection, date and function, Andersen has aimed to provide researchers with the ability to view multiple connections that can exist in archival finding aids. Andersen explains, ". . . a user viewing the person name dimension (or line) would see each individual represented in a finding aid as a cell. This person may appear in different parts of a collection, separate collections at the same repository and at other repositories, quite possibility related to different organisations, functions or roles."[27]

The approach outlined by Andersen is dependent on pre-visualization archival analysis. As Andersen notes: "The number and extent of dimensions it is possible to represent does, of course, depend upon the quality and extent of the underlying data."[28] For his project, he used two finding aids provided by the University of Glasgow Archive Services. Both finding aids provide the project with the opportunity to test the concept against EAD, ISAD(G) and ISAAR(CFP) with one of the finding aids providing a further test as it included function and activity 'cross walks' within it. Both finding aids cross multiple collections and repositories.[29] While Andersen's approach may (it is untested) help researchers better understand and conduct research into archival holdings, the need to conduct deep archival analysis prior to creating the ZigZag representation presents a constraint on the extent to which it can help to achieve sustainability for archives in the era of big data. What we need are tools that support the archivist to more rapidly undertake the type of analysis that uncovers the relations that form the basis of Andersen's dimensions.

Allen has explored the possibility of visually expressing hierarchies, networks, processes and timelines in US government archival holdings using Encoded Archival Description (EAD) to extract data and structure from source documents.[30] Motivated by a desire to provide effective interfaces to a growing number of online archival finding aids, Allen has described five approaches from the hypertext and visualization research communities which can be used to improve such access: 1) navigating hierarchical structures, 2) illustrating networks of relationships, 3) viewing processes, 4) using time and space as organizing structures, and 5) spatializing arguments and discussions. Similar in purpose to Andersen's work, Allen proposes an interactive approach to visualizing links between archival materials and has created a prototype interface that uses a "mass-spring model" to spatialize the relationship among the concepts.[31] In his prototype three major nodes are apparent: in the upper left is the Coast Guard Group (RG#026); on the right are agencies relating to oceanography and water resources; in the lower left is the Department of the Interior (RG#048). The edges or links between nodes represent the relationships between the record groups. In contrast to Andersen's work, Allen's paper is silent as to the specifics of the data used to define the relationships between nodes and on the technical details of his prototype. Nor is much said about the elements of interactivity available in the prototype that could assist the researcher to further explore and generate new insights from their visual interactions with archival material. However, by comparing the Andersen's ZigZag structure and Allen's space-filling node-link graph, it is possible to understand that different visualizations and elements of interactivity can be used to support the same analytic objective: in this case, the aim of revealing relationships between archival holdings by extracting and visually re-representing descriptive metadata elements from EAD-encoded archival finding aids.

---

[25] For more information about ZigZag ™, see http://users.ecs.soton.ac.uk/lac/zigzag/

[26] Andersen and North, 2009.

[27] I. Andersen and S. North, "Multidimensional Visualization of Archival Finding Aids," 13 January, 2009, accessed on 30 July, 2012 at http://www.hatii.arts.gla.ac.uk/research/visual/background.htm

[28] Ibid.

[29] [29] I. Andersen and S. North, "Multidimensional Visualization of Archival Finding Aids," 13 January, 2009, accessed on 30 July, 2012 at http://www.hatii.arts.gla.ac.uk/research/visual/sysarch.jpg

[30] R.B. Allen, "Using Information Visualization to Support Access to Archival Records, *Journal of Archival Organization* (2005) 3,1: 37-49.

[31] For more information about mass spring models, see X. Provot, "Deformation constraints in a mass spring model to describe rigid cloth behaviour, *Graphics Interface* (1995): 147-154.

ArchivesZ, developed by Jeanne Kramer-Smyth, is a prototype of an information visualization tool that, in the same vein as Andersen and Allen's work, leverages the structured data available in EAD encoded finding aids.[32] By representing the distribution of subjects and time periods using the metric of total aggregate linear feet, Kramer-Smyth argues that ArchivesZ enables tool users to view total available research materials more quickly than they would by viewing a standard search result list. Kramer-Smyth's claims about the utility of the tool remain untested, however, as she only has the feedback of a single domain expert as evidence that the tool generates the intended results.

Kramer-Smyth states that ArchivesZ was created to support three major audiences: archivists, researchers and students. She claims that archivists might use the tool to compare their holdings with those of other archives and that metadata associated with their collections match their understanding of their holdings.[33] Researchers, claims Kramer-Smyth, might use ArchivesZ to permit easy identification of institutions with archival collections fitting the criteria of their research. For students in the university setting who are not aware of what primary sources are available, a tool like ArchivesZ might encourage browsing and open ended exploration of the available collections, according to Kramer-Smyth. Though Kramer-Smyth claims ArchivesZ can support these use cases, in developing the tool she worked only with a single archivist as domain expert. The interaction with the domain expert was not to closely study the archivists' cognitive tasks in order to determine tool requirements, but rather to request feedback from the archivist as to the possible uses to which the tool might be put. While this is not a totally invalid approach, it lacks the rigour of cognitive systems engineering design that one might expect in a tool aimed at achieving enhanced analytic capabilities.[34]

ArchivesZ employs a dual-sided histogram, inspired by PB Browser, to support exploration of the multiple subjects assigned to each collection.[35] As subject terms are selected, the dual-sided histogram chart is generated to display related subjects. The tool combines the dual-sided histogram with a more traditional histogram displaying year data to permit tightly coupled, multi-dimensional browsing of subject and time period metadata. Kramer-Smyth's work does not discuss alternate designs and visualizations that may have been considered as a means to achieve the search objectives; however, as seen in the work of Andersen and Allen, it is possible to achieve the same or a similar objective with very different visual mappings and representations.

The previous examples of the application of visualization and visual analysis to archives have all sought to leverage archival descriptive metadata derived from online archival finding aids. Such an approach, however, still requires that archivists spend time analyzing archival documents in advance of producing visual representations of them. As archival data grows in volume it may be more sustainable to produce interactive visual tools that assist archivists and researchers to directly analyze massive amounts of archival data. With this goal in mind, early work at UBC on the application of visual analytics in the archival domain used a pre-existing visual analytics tool (In-Spire ™) to explore scanned and XML-encoded images of British Cabinet Papers, dating from 1915-1977, held at the UK National Archives. Thus, rather than working with archival finding aids, the research work at UBC has sought to explore the application of visualization and visual analytics directly to archival documents.[36] The analysis was

---

[32] http://www2.archivists.org/sites/all/files/KramerSmyth-AbstractBio.pdf. Coded in the Ruby scripting language. Kramer-Smyth notes that Standards have been entering the archival lexicon at a fast pace to ensure data reliability, enable data aggregation, and manage data over the long term. However, we have not yet examined the use of these standards across the archival community. As we move into the next phase of standards-creation, a broad look at current implementations will help to inform the next generation of these standards. Kathy Wisser (Simmons College) and Jackie Dean (UNC Chapel Hill) have conducted research on EAD tag usage in the encoding community. The goal of research was to identify encoding behavior to note the presence and absence of elements and attributes and the way that elements are used within the context of an EAD instance. This is an important precursor to visual encoding of archival metadata, as automation of the process requires that the data elements be used consistently.

[33] J. Kramer-Smyth, M. Nishigaki, T. Anglade, "ArchivesZ: Visualizing Archival Collections," accessed on 30 July, 2012 at https://wiki.cs.umd.edu/cmsc734/images/0/08/ArchivesZ.pdf

[34] For more information about cognitive systems engineering, see G. Lintern, "Cognitive Systems Engineering," n.d., accessed on 30 July, 2012 at http://www.cognitivesystemsdesign.net/Workshops/Cognitive%20Systems%20Engineering%20Brief.pdf

[35] M. Derthick and J. Zimmerman, "The perspectives browser: Exploratory data analysis for everyone," Submitted to the 2005 IEEE Symposium on Information Visualization, Carnegie-Mellon University and Human-Computer Interaction Institute, accessed on 30 July, 2012 at http://www.cs.cmu.edu/ sage/PDF/DerthickZimmerman.pdf, 2005.

[36] This research was undertaken as part of a larger research collaboration between the UK National Archives, University College London Department of Information Studies, the University of British Columbia School for Library, Archival and Information Studies together with the Multimedia and Graphics Interdisciplinary Group, and Simon Fraser University Centre for Interactive Technologies.

performed by: uploading the XML text into visual analysis software in the lab.[37] Among existing VA tools, IN-SPIRE was chosen because it supports the rapid perception of key information characteristics in a collection, navigation through the information space, foraging for critical evidence and patterns, and organizing evidence for reasoning. Users can run queries, create groups of documents that are perceived to be of interest, and see correlations among such groups. IN-SPIRE also offers the facility to record the sources of documents and to arrange them in clusters, functionalities that we found of value in the conduct of this research experiment. In short, the Project Team viewed IN-SPIRE as a suitable tool with a good overall fit with the research objectives while providing for flexibility in the application of VA to the problem area.

Once the data were loaded into the visual analysis software, the analyst produced what InSPIRE calls a "theme view" that illustrates the relative frequency of term occurrence and co-occurrence. In this particular application, co-occurring key terms that occur more frequently have a higher peak (or 'spire') and the use of color coding creates a heat-map style visualisation from a bird's eye view. The analyst next produced a "galaxy view", wherein each dot represents a document derived from the top 200 key terms in the documents. The documents were clustered using a common clustering algorithm. In this case, the higher is the spire, the more common the occurrence of the keywords in the cluster. The analyst used this approach successively on clusters (represented in this case by spires) of interest by selecting an interesting cluster and dropping everything else out of the calculations (e.g., making the remaining documents temporarily into outliers) and then evaluating and corroborating (or modifying by labelling) the key topics of the cluster, but also finding peculiar documents for the given cluster. Using this technology with an iterative analytic technique, users are able to hone in quickly on the key topics in a large data set, rather than having to guess what might be covered and come up with key words that may or may not reveal relevant documents, as would be the case if access were provided through a conventional search engine. A major advantage of this approach is that the user does not necessarily need to know what he/she is searching for in advance. In this case, the VA tool overcomes this cognitive barrier by applying computer intelligence to analyse large datasets to "see the unseen." Combining text mining capabilities with VA's capability of showing clustering of documents in an archival *fonds* by topic prevents a problem that often occurs with keyword searching when documents can fall through the cracks of a search because users do not know what keywords to search on. By way of explanation of how VA overcomes this barrier, an analogy might be trying to search for the proverbial needle in a haystack without knowing that it is a needle that you are searching for or what a needle is called; given this scenario, actually finding the needle becomes very difficult. In contrast, VA technology applied to a proverbial "haystack" of data is able to tell the user there is a needle in the haystack; the user would then be able to use this information to drill down into the visualization to locate the needle. In this way, VA can present all documents and cluster them together, so no document is missed out accidentally.

Other researchers have applied visualization and visual analytic approaches to unprocessed documents or archival material. Though researching visual analysis of emails as opposed to historical archives, researchers at the University of Texas (UT) have worked on visual tools that use analysis of the textual content of archival records to establish contextual relations among them. Maria Esteva, Weijia Xu, Jaya Sreevelsan-Nair, Ashwini Athalye and Merwan Hadethe point out in a recent research paper that finding a trail of documents relating to a specific business activity as opposed to an email chain – an important part of the archival research process – remains a research challenge.[38] The authors have attempted to automate the identification of business relationships between documents (relationships which they refer to as the 'archival bond', using L. Duranti's terminology), by using information retrieval techniques to compute the frequency of the occurrence of particular terms within documents and document paragraphs, and then using this computation to quantify similarities among documents and among paragraphs to construct constructing visualizations of the results. While there are limitations in the UT approach, the combination of information retrieval and text mining with VA extends and complements extraction and use of existing metadata which need not be limited to e-mail but can be applied to a wide range of archival materials.

---

[37] Pacific Northwest National Laboratory, "In-Spire [computer software], 2011, accessed on 30 July, 2012 at http://in-spire.pnnl.gov/.

[38] Esteva, Maria et al. "Finding narratives of activities through archival bond in electronically stored information (ESI)" (Paper delivered at the 2009 Society of American Archivists' Research Forum, Austin, Texas, August 2009) and M. Esteva, W. Xu, J. Sreevelsan-Nair, A. Athalye, and M. Hade, "Finding narratives of activities through archival bond in electronically stored information (ESI). DESI III Global E-Discovery/EDisclosure Workshop: A Pre-Conference Workshop at the 12th International Conference on Artificial Intelligence and Law '2000, accessed on 30 July, 2012 at http://www.law.pitt.edu/DESI3_Workshop/Papers/DESI_III.Esteva-Xu-Nair.pdf adresinden erişildi.

More recently, researchers at the UT have developed a prototype VA system to aid archivists in identifying files requiring digital preservation and as a possible VA tool in support of archival research.[39] In order to make decisions about the long-term preservation and access of large digital collections, the authors observe that archivists gather discrete pieces of information such as the collection's organizational structure and its technical composition. This process from data gathering to exploratory analysis, to insights is generally aided by pen and paper, note these researchers. As digital collections become larger in size and more complex in structure, conducting analysis with manual methods limits the possibility of understanding these collections, and accurate notions are buried in details or get lost in generalities. To help archivists synthesize and obtain integrative insights from large and complex digital collections, these researchers investigate the use of interactive visualization to aid the analysis through visual exploration. They developed a visualization based on a space-filling treemap to present digital file-related metadata extracted from the collection at different levels of aggregation and abstraction. [40] A treemap visualization is a logical choice, as this visual form maps to the hierarchical semantic structure of the data the team chose to associate with elements in the visual display (i.e. file labels and other file metadata). Nevertheless, space-filling treemaps do have their limitations: the larger the holdings, the more detail can be lost as each rectangle in the treemap takes up a proportionally smaller amount of screen real-estate. As a result, smaller and rarer groupings of file formats, the very formats that can cause the greatest preservation challenge, can become lost in the noise of larger groupings of more common file formats. The researchers do not discuss these limitations, nor alternate designs and considerations. In addition, they remain silent on the quality of the file format metadata and its implications for preservation decision-making using their interactive visual interface. Yet, file metadata can be notoriously unreliable as migration between systems often overrides historical metadata with new metadata; indeed, this has been a common problem in accurately ascribing dates to files in digital forensics.

The UT researchers provide two largely hypothetical use cases to show how their application allows learning about a collection, by providing an integrative analysis experience that facilitates preservation decision-making by archivists. The other use case relates to support for researcher exploration of archival holdings (i.e. to determine trends in the collection and to investigate the contents and correlations). Similar to the Kramer-Smyth visual interface, the UT research team remain relatively silent on the methods used to understand archivists' analytic requirements, though they do mention that they worked with archivists at NARA to identify requirements for preservation decision-making. This might explain why the archival use case is the more compelling of the two use-cases for the tool. It is unclear how the UT research team determined the capabilities that would be required of the tool to support the analytic tasks of researchers. Both use cases remain untested.

More recent work at UBC is based on the premise that to build visual analytic applications for archival requires a solid understanding of how archivists and researchers think: that is, it is necessary to understand the analytic tasks that visual analytic systems are being designed to support. Efforts at applying visualization in the domain of archives have, so far, not placed great emphasis on the perceptual, cognitive and analytic aspects. Most efforts have made assumptions about the analytic requirements, rather than actually rigorously researching those requirements. For example, research into the application of visualization and visual analytics to archives has not made much of a distinction between groups of researchers. ArchiveZ functions the same way whether one is a genealogist or a historical researcher. Whitelaw's visualization of archival series, likewise, functions the same way, regardless of analytic requirements of different researchers.[41] This is not to suggest that such visual applications are not useful, but that an approach that places greater emphasis on combining cognitive systems engineering with visualization, is likely to better support real visual analysis. To this end, UBC has been conducting research into the archival reasoning process in order to better understand how the analysis that archivists undertake as they arrange and describe archival *fonds* can be supported using interactive visual interfaces.

Surprisingly, a review of the archival literature on arrangement and description has revealed that explanations of the archival reasoning process, by and large, are absent from the archival literature. Most of the literature on arrangement and description can be grouped in two categories. The first discusses the main concepts involved in

[39] M. Esteva, W. Xu, S. Dutt-Jain, J.L. Lee, W.K. Martin, "Assessing the Preservation Condition of Large and Heterogeneous Electronic Records Collections with Visualization," *International Journal of Digital Curation* (2011) 6, 1:45-57, accessed on 30 July, 2012 at http://www.ijdc.net/index.php/ijdc/article/view/162.

[40] For more information about space-filling treemaps, see B. Johnson and B. Shneiderman, "Treemaps: A spacefilling approach to the visualization of hierarchical information structures," IEEE Information Visualization, 1991: 275-282.

[41] See Whitelaw op. cit.

these two archival functions, whereas the second category deals with the practicalities archivists have to deal with when actually performing these functions (i.e. by providing procedural steps). Neither category of literature on archival arrangement and description is very clear about the thought processes involved. Since the 2000's, however, a third category of studies has been slowly emerging. By explicitly or implicitly acknowledging the role of archivists in shaping the meaning of records and archives, these works offer an important step forward in understanding how archivists think as they conduct arrangement and description. Yakel sees arrangement and description as a fluid, evolving and socially constructed practice, which she names "archival representation".[42] Heather MacNeil calls attention to the selectivity involved in archival description, as well as to its incomplete character.[43] Jennifer Meehan sees arrangement as "identifying and creating the contextual relationships of a body of records", and points to the historical standpoint of the archivist, the use of evidence and the role of inference as central factors in the analytical process in arrangement and description.[44] Though the works of Yakel, MacNeil and Meehan provide a solid starting point for research into how archivists think, these works also demonstrate the pertinence and necessity of a thorough investigation of the cognitive and perceptual tasks involved in archival arrangement and description.

In order to fill this gap in knowledge, the UBC visual analytics research team has been exploring archivists' cognitive processes. This fits with the study of analytic provenance in the field of VA. Analytic provenance is an area of research that focuses on examining a user's interactions and reasoning process when working with visualizations. It examines not only the final result garnered by working with a visualization or VA tool, but the user's insight on how they reached their conclusions. [45]

Experiments at UBC have entailed observing current archivists arranging and describing archival data. Archival records used in the experiment have not been previously processed, so as to obtain a clear understanding of the entire process rather than repeating arrangement and description on a known set.[46] This approach has also allowed uncertainties to arise as normal during the process, since with previously processed material such problems may have been expected or already dealt with. Participants have been asked to conduct their standard archival procedure on these records, while being observed using the think aloud cognitive task analysis method.[47] The think aloud protocol is a method of gathering data where the researchers ask the participant to describe what they are doing as they perform a task. This might include descriptions of what they are thinking, looking at, feeling, any unexpected problems they might be having, and so on. This allows any internal information that may be apparent to the subject but not the researchers to be made externally evident, and insight can be gained into the user's experience. The method can help reveal things like how an analyst resolves uncertainties determining the implicit heuristics or qualitative mental models that a user might be employing, or identifying what the common conceptual simulations that arise during an analysis might be.[48] Two archivists were used in total, who worked in a pair. This modified approach to the use of the think aloud protocol, inspired by pair analytics, overcomes one of the weaknesses of the think aloud method, which is that as individual participants become more cognitively engaged in their work, they articulate their thoughts less frequently.[49] The research team found that working in pairs encouraged the archivists to continue to articulate and communicate their thoughts to each other.

---

[42] E. Yakel, "Archival Representation," *Archival Science* (2003) 3,1: 1-25.

[43] H. MacNeil, "'Picking Our Text': Archival Description, Authenticity and the Archivist as Editor" *American Archivist* (Fall/Winter 2005) 68,2: 264-278.

[44] J. Meehan, "Making the Leap from Parts to Whole: Evidence and Inference in Archival Arrangement and Description," *American Archivist* 72,1: 72-90//

[45] W. Dou, D. H. Jeong, F. Stukes, W. Ribarsky, H. R. Lipford, and R Chang, "Recovering reasoning process from user interactions," IEEE Computer Graphics and Applications (2009). 29(3): 52–61.

[46] Archival documents encompassed mostly analog images and files, but also contained some floppy disks and a film reel. Sensitive information within the fonds was removed prior to the arrangement.

[47] S.B. Trickett, J.G. Trafton, L.D. Saner and C.D. Schunn,"'I don't know what's going on there': The Use of Spatial Transformations to Deal With and Resolve Uncertainty in Complex Visualizations," In: M.C. Lovett and P. Shah (eds.), Thinking with Data, New York, NY: Lawrence Erlbaum Associates, 2007: 65-86, Sessions lasted between 80 and 100 minutes depending on the participants' pace, and involved a full archival arrangement and description. Sessions were conducted in the participants' normal archival location, so as to achieve a more naturalistic recording of the methods involved in archiving. Participants were audio-recorded with their consent, and text transcriptions were made of each session. A brief interview followed each session. Observational notes on the environment or process were taken by the observers.

[48] Trickett et al., 2007

[49] R. Arias-Hernández, L. T. Kaastra, T.M. Green, and B. Fisher, 'Pair analytics: Capturing reasoning processes in collaborative VA'. In Proceedings of the 44th Annual Hawaii International Conference on System Sciences: January 4-7, 2011, Koloa, Kauai, Hawaii, R. Sprague (ed), IEEE Computer Society Press, 2011.

The UBC research team observed that archival processing followed a three-stage sense-making process: in phase one, from first contact with the archival records, the archivists were concerned with gaining an overview of the structure of the archival *fonds* and creating a draft arrangement. In the second phase, the archivists confirmed and refined the arrangement structure and re-ordered archival documents, as necessary, and in the third and final phase, they described the final arrangement, documenting their description in an archival finding aid. One significant observation is that the initial process of analyzing the archival *fonds* and developing a mental model of its contents creates a significant cognitive load. Participants reported being much more fatigued in the first phase than they did in subsequent phases: there was less joking during the process, and there appeared to be greater uncertainties. Part of this may be accounted for by the fact that the participants were still adjusting to being observed and performing the think aloud protocol, but observers remained silent throughout, and both subjects worked in their standard location and using their own methods. More time was devoted to figuring out the timeline and context of the *fonds*; after this was determined, particularly in the second and third sessions, both participants were able to proceed more quickly. Further contextual information was easy to integrate once they had determined an overall mental model of the total arrangement, and fewer uncertainties were shown when new information became evident. The archivists were even able to infer the date and purpose of various files once this general mental framework was established. A more solid understanding of the *fonds* appears to have been established in the second session, when the participants began deciding on the series and their arrangement, despite not having completed arrangement on all of the files, and having focused mostly on rehousing. Even with these unknowns, participants' final decisions on series closely mirror the ones they considered at this point. Similarly, initial predictions on the content of the individual folders or containers were more accurate once they had established more information about the creator of the files and had an initial mental model of the creator's life. Any points of confusion that arose in the later sessions occurred because of inconsistencies in what they had established in their interpretation of the timeline and the context of the files they were working with. While neither archivist explicitly mentioned the development of a mental model, results of the experiment suggest that creating a mental model is a major part of the process of arrangement and description.

Another of the major findings is that the actual form of the archival documents is very important in conducting archival analysis. While this is not in itself surprising giving the well-known archival adage of "form following function", the importance of form is of interest. The archival *fonds* with which the subjects worked was relatively small, comprising only three major divisions, but it still contained a large amount of information. It would have been impractical to have attempted to read every single document, or to explore every film reel or floppy disk. The large amount of content in even a small *fonds* means that trying to create series using only content will run into difficulties, as this would require examining in depth every single file. However, both archivists were able to quickly examine and categorize hundreds of files into their various series by paying attention to form. For example, grant applications were very common throughout the files that the subjects received. Different grants will have different application forms, but they tended to follow a general structure, and at the very least were simple to quickly identify and recognize as being grants as opposed to correspondence. After a fairly in-depth examination of a few of these grants, the archivists were able to quickly identify files that fit the visual description of a grant application, and were then quickly able to pick out relevant content, such as the size of the grant, the year, or the organization providing the grant. This then allowed the archivist to place the grant in a timeline of the author's life, and to draw inferences as to what the records creator may have been doing with regard to their professional career, or even to ascertain the creator's geographical location at the time..

While the observations were done only on physical archives as opposed to digital ones, relevant information can still be drawn and applied to the design of a VA tool for working with digital archives. Identifying similarities in form is itself a sort of pattern matching. By quickly searching through various digital files and identifying similarities in form, an archivist could quickly cluster files, and then rearrange them into suitable series, and outliers can quickly be brought to attention and categorized into an appropriate cluster, or form their own series. This potentially also allows for easier detection of duplicates, as well as cutting down the cognitive load for the archivists. Text mining for dates may assist in building a timeline, providing context for the archivist when trying to gain a sense of the fond as a whole, as well as being of possible use to future researchers using that archive. The subjects also used a number of other programs or sources to help their process. Dictionaries and an online encyclopedia were used to define or elaborate on terms they were not familiar with, while physical implements like rulers were used to help mark their current location in a box as they arranged it, and using a text editor, they made note of the original organization of the archival data. In some cases, there were no available aids, such as when comparing two photographs to

determine if they were duplicates, or trying to find whether an article had been published in a journal. Many of these aids could be integrated into a tool to assist the archivist's analysis.

Key conclusions

Visualization and visual analytics hold much promise as tools to support a sustainable future of for archives. Yet in spite of the great potential that visualization and visual analytics hold for archives, much more research is needed in order to understand how these technologies can be applied effectively to support archival functions and in order to transfer novel tools from the lab to production systems working in archives. Work on visualization and VA approaches to archival endeavours would benefit from focussing more on developing technologies to help archivists perform archival analysis on unprocessed archival documents or to assist researchers to explore archival documents directly, whether processed or not, than on creating tools that visually re-represent the results of completed archival analysis. As useful as tools that support visual interaction with archival descriptive metadata are, they still require much archival pre-processing before they can be used unless the metadata are created upstream before transfer to the archives and as a natural part of other processes. Such archival pre-processing is likely become increasingly untenable in the era of big data without new tools that cognitively aid archivists in the task of archival analysis or which aid researchers in the analytic tasks associated with their research. As such, information visualization and VA as applied to the domain of archives would also benefit from the study of different archival functions and their associated high and low-level cognitive and perceptual tasks, together with the range of visual metaphors or mappings and representation design alternatives. With such an approach, there is the potential for a better understanding of how archivists actually think or undertake archival analysis. Such an understanding is a critical precursor to developing effective tools to support visual analysis. It would also help to better understand the ways in which archival analysis and the analysis undertaken by researchers differ at a cognitive and perceptual level, and to distinguish more effectively between different types of researchers so as to build tools that reflect cognitive and perceptual differences opposed to designing 'one size fits all' tools. Current research on the application of visualization and VA to archives would also benefit from rigorous testing of the efficacy of design patterns against task, data type and other pertinent factors. Many claims have been made by tool developers about how the tools support archivists or researchers, but so far such claims are unverified. To move the research forward, both long-term user studies and experimental laboratory testing would be useful. Allen urges moving beyond identifying isolated visualization tools to consider how they may be combined into a system.[50] He suggests that by putting several of visualization tools together along with other data management tools we could create an archival-researcher's or a record-manager's workbench for supporting complex searches and combining materials from many sources. This is likely to be necessary to see visualization and VA really support a more sustainable future for archives. Though a number of researchers have experimented with combining information retrieval, text mining and natural language processing approaches with interactive information graphics, effective application of visualization and VA to archival functions is likely to require more extensive experimentation with combining these other technologies with visual technologies in support of archival tasks.

The research needed to advance the application of visualization and VA in archives, and the understanding of visual technologies needed to enable archivists to use interactive visual interfaces to assist archival analysis or research into archival holdings will be dependent on incorporating knowledge about visual literacy and visual technologies into the archival curriculum. To this end, the School of Library, Archival and Information Studies at UBC now offers a course on Information Visualization and Visual Analytics as part of its Masters of Archival Studies Programme.

In short, there is a number of ways information visualization and VA can sustain archives in the future, but there is still a long way to go before these technologies can have a major impact on daily archival work. Nevertheless, an archival future in which interactive visual tools help archivists perform archival analysis and assist researchers to explore archival documents is both achievable and likely to be at least one answer to achieving a more sustainable archival future.

[50] Allen, 2005.