A Unique Arrangement: Organizing Collections for Digital Libraries, Archives, and Repositories

Jeff Crow, Luis Francisco-Revilla, April Norris, Shilpa Shukla, and Ciaran B. Trace

School of Information, The University of Texas at Austin,
Austin TX, USA
{jcrow,revilla,anorris,shilpa,cbtrace}@ischool.utexas.edu

Abstract. Digital libraries increasingly host collections that are archival in nature, and contain digitized and born-digital materials. In order to preserve the evidentiary value of these materials, the collection organization must capture the general context and preserve the relationships among objects. *Archival processing* is a well-established method for organizing collections this way. However, the current archival workflow leads to artificial boundaries between materials and delays in getting digitized content online because physical and born-digital materials are processed independently, and digitized materials not at all. In response, this work explores the approach of processing materials in a digitized form using a large multi-touch table. This alternative workflow provides the first step towards integrating the archival processing of digital and physical materials, and can expedite the process of making the materials available online. However, this approach demands high quality digitization and requires that archivists perform additional tasks like matching multi-sided, multi-paged documents.

Keywords: Multi-touch, archival processing, digitized materials.

1 Introduction

The core mission of digital libraries is to facilitate the use of the information that they host. This mission is challenging because people create and use information in increasingly different ways. Furthermore, as technology evolves, digital objects increase in complexity and their file formats change. This creates dependencies on legacy hardware and software [1]. Consequently, researchers have been investigating how to provide long-term access [1], storage, and preservation [2, 3] of digital objects. Many of these solutions address the issue at object-level (e.g., creating smart digital objects that can automatically copy themselves [4]). However, focusing on the long-term use of these objects requires rethinking how collections are organized and managed as a whole.

Digital libraries and archives host innumerable digital objects of significant scientific, legal, economic, cultural, and historic value [1]. Currently, many digital libraries are built on the premise that these objects can be organized and made accessible as independent units. This makes sense when the items (e.g., books and journals) each

P. Zaphiris et al. (Eds.): TPDL 2012, LNCS 7489, pp. 335–344, 2012.

have a distinct internal cohesion and contain all the necessary information within them to be read, analyzed, and understood. However, digital libraries are increasingly incorporating digital objects (e.g., scientific data, and personal, organizational and government records) that are archival in nature. These archival records are distinguished by the fact that they are created as a by-product or instrument of some practical activity, and are set aside by their creator for future action or reference [5]. As such, these records constitute a "primary and privileged source of evidence about the activities and the actors involved in them" [6]. These archival records, though they can be read as individual units, lose much of their meaning as evidence when managed and accessed independently.

The unique nature of archival records has major implications for collection organization and system design. Records serve as evidence of the actions of the creating entity, and derive much of their meaning from the context in which they are created and filed. A key part of this context is the *archival bond* – the notion that a relationship exists between all records created as part of the same activity. Rather than treating records as standalone objects, archival thinking requires that the archival bond be maintained and preserved in order for records to retain their meaning and evidentiary nature [8]. While some digital library platforms support a certain level of grouping (e.g., volumes for journal issues), this is insufficient from an archival perspective.

Archival science addresses the requirement of preserving the evidential value of records through a well-established method for organizing and describing collections: *archival processing*. The activity of archival processing requires completing two steps: arrangement and description. *Archival arrangement* is the method for organizing the collection and involves establishing or re-establishing the original intellectual and physical order of records in a collection. In this iterative process the archivist is looking for clues of organization and order within the records and aggregations of records in order to restore or recreate the original filing system. *Archival description* is the "creation of an accurate representation of a unit of archival material by the process of capturing, collating, analyzing, and organizing information that serves to identify archival material and explain the context and records system(s) that produced it" [9]. Finding aids are similar to library catalogues in that they allow physical and intellectual management of the collection and facilitate user access to the collections. Supporting the activity of archival processing is crucial for digital libraries that aim to support activities such as scientific discovery and historical research.

In recent years, the need to reduce large backlogs of unprocessed collections has prompted a call for new ways of thinking about all aspects of this core activity [7]. Currently, digitized materials are not part of the processing workflow. Physical materials are processed before they are digitized. Born-digital materials are processed separately following a different methodology. In hybrid collections, this workflow can create an artificial boundary, potentially disrupting the archival bond.

This paper looks at the role of technology in supporting the activity of archival processing among practitioners as a key step in organizing groups of records before they become part of a digital library system. It focuses on recasting the workflow of archival practitioners by moving from a model where paper-based collections are processed first and digitized second, to one in which collections could be digitized

first and then processed second such that all materials (physical and born-digital) can be considered together. Specifically, this paper introduces the Augmented Processing Table (APT) project. APT pioneers the use of surface computing devices for processing collections of digitized archival material. Just like a hybrid collection combines physical and digital materials, APT creates a space that allows for the processing of digitized materials in combination with born-digital material, integrating both modalities (paper and digital) in one workspace.

2 Interactive Surfaces

In addition to archival science the APT project is informed by previous work on interactive surfaces and tangible user interfaces (TUI). Interactive surfaces that support multi-touch interactions play an important role in a variety of settings where people are engaged in information intensive activities such as office work [10], disaster control management [11], and leisure activities [12, 13]. Recently, researchers have been investigating the use of interactive surfaces in complex information applications such as document review in legal cases [14] and collaborative search among co-located group members [15].

In the design of multi-touch interfaces, designers often draw from the physical world, whether it is through the use of metaphors to describe interaction or behavior, or using embodying aspects of the physical world that are thought to be relevant to human interaction [16]. Tangible User Interfaces (TUIs) go beyond the use of metaphors, utilizing physical objects to represent, display, and/or act as a physical control of the digital representations on the multi-touch surfaces. For example, PaperView uses pieces of plain paper that act as personal, location-aware, interactive screens [17]. Designing TUI interfaces can be difficult [13] because it is necessary to decide when to provide physical or digital interface elements, and to what degree digital elements should emulate real-world interactions. However, evaluations of TUI systems show that interfaces that rely on familiar objects (e.g., paper) provide predictable and straightforward interactions [17]. Furthermore, the approach of integrating digital material into established paper-centric processes such as literary criticism has been proven beneficial [18].

APT explores the use of interactive surfaces for archival processing and studies if archival processing is amenable to be conducted using digitized documents. Like Terrenghi et al. the APT project is interested in "understanding not only people's expectations and mental models about digital versus physical media, but also an understanding of the associated affordances for interaction in these different situations" [16].

3 Design

Previous research about the affordances of paper and digital media [16, 19, 20] indicate archival processing of digitized materials is viable, and interactive surfaces can ease this transition from paper to digital. Similarly to Family Archive [21], APT aims

to let users interact with both born-digital and digitized material, and thus facilitates the study of how interactions with these objects differ. Although not the primary focus of this paper, this understanding will be crucial for developing a fully-hybrid environment for archival processing.

APT was built following a collaborative design process. The five-person research group consisted of three digital library/HCI researchers and two archival researchers (one of who has professional experience in archival processing). The team met weekly throughout the duration of the project, setting up tasks and tracking their progress. The archival researchers served as domain experts, providing crucial knowledge about the problem space. In processing collections, archivists need a quiet atmosphere and a large flat work surface (e.g., table). Typically, processing a collection requires several sessions to reconsider and fine tune the arrangement. In terms of time, the archivists expressed that for a disorganized collection of 40 items needing item level arrangement (this is a typical assignment in a graduate course on Archival Enterprise) it takes 2-3 hours, and is normally done in one or two sessions. In this scenario, archivists manipulate objects individually, and create and refine groups that reveal the relationships between objects (archival bond). This work is highly visual. Archivists often do not fully read the documents, but pay attention to the documentary form, general appearance, and certain internal metadata of the objects. While groups are expressed implicitly (e.g., piles) or explicitly (e.g., areas), archival arrangement has strict rules about group hierarchy, limiting the types of groups (*sub-group*, *series*, *files*) and the order in and among groups.

The archivists' domain knowledge facilitated identifying the following design implications:

- maximize the surface area on a dedicated workspace
- allow for the creation of ephemeral and permanent groups
- support user manipulations at the object and group level
- allow revisiting and/or reverting back to any previous states
- allow for note taking while processing
- allow metadata manipulation of objects and groups

In order to meet this requirements APT's core functionality was designed as a spatial hypermedia application [22]. Spatial hypermedia allows users to interact with objects and metadata, and can automatically infer groups (both implicit and explicit) based on visual structures such as piles and lists. Since a key goal behind the design of APT is creating a platform that allows for the study of archival processing, APT tracks the history of the workspace in a manner similar to spatial hypermedia systems such as VKB [22].

4 System

APT consists of a custom-made, large surface (5'x5' total, 47"x28" interactive), interactive tabletop computer that runs a specialized spatial hypermedia application for digital archival processing (see Figure 1).

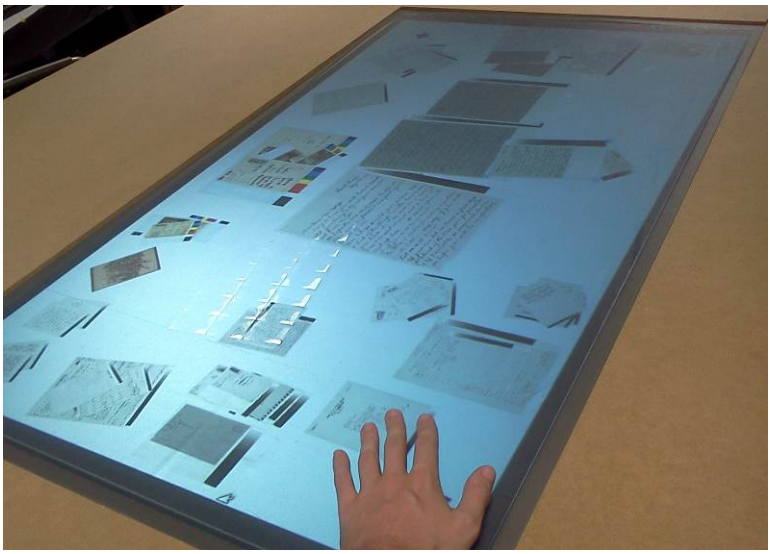

Fig. 1. The Augmented Processing Table

When document images are first imported into APT, they are tiled across the workspace to give an overall impression of the size of a collection. In the initial state all items are located at the root level of the workspace hierarchy. Users can move, scale and rotate items freely in 2D. Users can create groups, and add items to them, which in turn can be added to higher level groups, like series and subgroups. These groups correspond to specific levels of the archival hierarchy (see Figure 2).

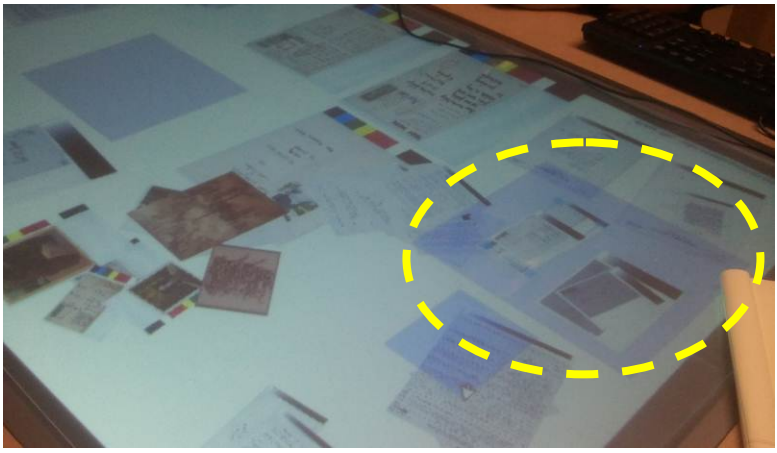

Fig. 2. Grouping and sub-groupings

Archivists may add metadata (e.g., title, description) to items and groups. APT saves the state of the workspace whenever any change occurs, requiring no input on the part of the user to save their work.

5 Pluralistic Walkthrough Evaluation

A key aspect of the project is to study if archival processing can be conducted using digitized materials on an interactive surface. In order to determine this, APT was evaluated using a pluralistic walkthrough (where a team with varied expertise walks through a scenario of use to uncover possible interaction and usability issues). One archival researcher created a test-collection using a subset of an existing collection. This researcher processed the test-collection physically, creating a baseline for the experiment. The other archival researcher served as the participant, processing the collection as part of the walkthrough.

The walkthrough studied the tasks and activities of the participant as she processed the test-collection using a think-aloud protocol to externalize her thoughts and motivations. The rest of the research team observed and took notes unobtrusively. The walkthrough was conducted in a single session that lasted four hours (including a break in the middle). After the task, an unstructured interview was conducted where the participant answered questions from the panel. Finally, after a week of individual reflection, the group met, compared notes and discussed the tasks and activities of processing using the interactive surface.

In general, the participant processed the digitized materials as if they were physical materials. The observation and analysis of the participant's activities and comments revealed some aspects about the system functionality and the differences between processing digitized collections and processing collections physically that are worth mentioning:

Quality of Digitization. Understanding the actual physical characteristics of the objects is extremely important for the task of archival arrangement. Improper digitization (e.g., deformed images) and lack of information about their physical characteristics (e.g., actual size of a document) can hinder the task. For example, when the participant was looking at an image, and particularly when she resized the image to read the content, the lack of information about the original dimensions of the document meant that she could not always decide, for example, if it was a postcard or a large painting.

Rotating and Resizing Controls. In APT the ability to scale is combined with the ability to rotate, and can be done at any corner of a document. During the experiment, the participant performed these functions repeatedly, and having a single control for both, resulted in unintentional rotation or resizing actions. The participant thought that the controls should be separated.

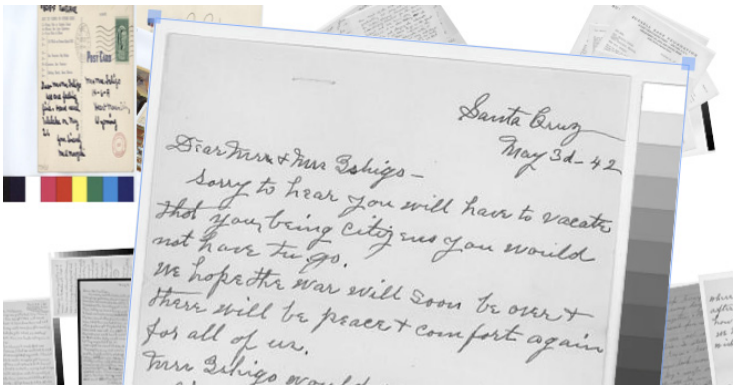

Fig. 3. A zoomed item

Lack of Digital Representation for Physical Characteristics. While the participant often resized documents in order to look at particular details, such resizing and zooming in groups made it hard to determine the size in relation to other documents in the workspace that potentially exist at many different scales (see Figure 3). This was aggravated in APT because the zoom factor was not shown and many of the digitized images lacked any information about scale. To address this issue, the next iteration of APT needs to show the zoom factor or include the physical size as metadata.

Matching Documents. Materials such as double-sided documents created a new *matching task* because each side of the document was captured in a different image. This led to issues as the participant had to determine if items were distinct documents or the front-and-back of a single document. The participant commented that a 'staple' function could solve the issue of multi-sided and multi-paged documents.

Creating and Utilizing Metadata. A common practice in archival processing is to take notes in order to facilitate the arrangement process. The participant specifically described the need for a digital equivalent, saying that she would use these temporary notes to capture titles, descriptions and group types (these were included in the system but not displayed outside of the editing menu), and key metadata from the documents themselves (such as enclosure annotations and reference initials). While note taking is common in traditional archival processing, APT can allow for notes to be directly appended to the images, something that is not done with physical items due to preservation concerns.

Searching and Fetching. Once the participant had done the initial sorting and added metadata to some items, her workspace had become "crowded" and she had difficulty finding specific items in the workspace. Even though APT can zoom in/out, she noted that it would be extremely beneficial to be able to find items using their metadata rather than having to move documents around trying to find a specific item.

Managing Groups. APT represents groups as distinct regions on the work space. Tasks like adding items to a group, manipulating the item order within a group, and creating a hierarchy of groups require additional support, especially when the

participant starts focusing on creating a presentation of the arrangement. For example, the participant expressed a desire to have documents adopt a 'snap to grid' behavior once put inside a group, and to be able to hide items behind a single representation of a group and only display the contents when needed.

Presentation Mode. After the participant had created a number of groups, she asked how to acknowledge in APT that she was finished arranging, expecting a separate mode for displaying and exploring a processed collection where the workspace would be "locked" and no changes could be made.

6 Discussion

Overall, the walkthrough evaluation revealed that archival arrangement includes several stages:

1. Triage – quick sort of materials into temporary groups (e.g., piles).
2. Group refinement – revision of all groups (one by one), validating the inclusion of every item (or moving them if necessary). At this stage archivists also start working on presentation aspects.
3. Object matching – formalization of relationships between objects (e.g., matching front/back of postcards and “staple” them together). This stage requires a lot of searching and parsing to find documents.
4. Metadata and archival hierarchy – adjustment of groups according to the formal archival structure, and entering the permanent metadata.
5. Overall workspace organization – ordering ‘messy’ parts of the workspace and making it suitable for presentation.

For the archival researchers, this basic articulation of the stages of processing was significant, because this process has never been systematically studied. Instead the traditional focus in the archival literature is on articulating processing principles or on determining costs.

While there are changes that are needed in order to accommodate archival practices, the evaluation showed that archival processing can be conducted digitally using digitized materials. This supports the approach of digitizing first and processing second.

The evaluation highlighted the need to pay attention to the digitization process, as some information can be lost or distorted. However, some of these issues can be solved by providing additional functionality to APT. For example a *stapling* function could help users match double-sided documents, and combine multiple pages that make a single item (e.g., multiple pages in a letter) into a single representation.

In terms of functionality, the evaluation of the initial iteration of APT called for functions similar to those provided by spatial hypermedia systems [22] including searching and fetching, and visual presentation operations. The evaluation also revealed the need for functionality specific to archival processing, including visualizations for metadata, physical characteristics, and relative state of the items and the workspace.

7 Conclusions and Future Work

Digital libraries are increasingly hosting collections that contain digitized and born-digital archival materials. For collections of an archival nature it is critical to arrange the materials in a way that captures the context of the overall collection and the relationships between the individual objects in it. Archival science shows that archival processing produces a proper arrangement and collection description that protects the evidentiary nature of materials in the collections.

Archival processing has traditionally followed a workflow of "process first, digitize second". This workflow has some drawbacks that impact "principled practice" and work productivity. This workflow may also lead to delays in getting digitized content online, because digitization needs to wait for processing to take place. Further, in traditional archival processing, physical and born-digital materials are processed separately and differently. Arguably, this creates an artificial boundary between the objects.

The APT project shows that archival processing is amenable to be conducted digitally using interactive surfaces such as multi-touch tabletops. This is highly significant as it represents the first step to integrate the archival processing of digital and physical materials, and allow a workflow of "digitize first, process second". While processing is still a labor-intensive task, this approach has the advantage that it can potentially augment the availability of items in digital archives.

The "digitize first, process second" approach demands a high quality digitization phase, and requires that the processing archivist performs additional tasks such as matching multi-sided, multi-paged documents.

APT provides a platform that allows researchers to investigate future directions of digital archives. The APT project is interested in evaluating the effectiveness of APT as a tool for studying, documenting and teaching archival processing, as well as for exploring new ways to do archival processing including remotely and collaboratively.

The results of the pluralistic evaluation are guiding the design of a second APT prototype, which better represents the objects' physical characteristics, and provides advanced functionality for creating and presenting the archival arrangement. This second prototype will have a formal evaluation with a larger sample of archivists.

Acknowledgements. This work is partially funded by a University of Texas, School of Information Temple fellowship. An alphabetical author sequence is used to acknowledge the equal contribution made by each group member.

References

1. Woods, K.A.: Preserving Long-Term Access to United States Government Documents in Legacy Digital Formats. Ph.D. Dissertation, Indiana University (2010)
2. Galloway, P.: Preservation of Digital Objects. ARIST 38, 549–590 (2010)
3. Woods, K., Lee, C.A., Garfinkel, S.: Extending Digital Repository Architectures to Support Disk Image Preservation and Access. In: 11th Annual International ACM/IEEE Joint Conference on Digital Libraries (JCDL 2011), pp. 57–66. ACM, New York (2011)

4. Cartledge, C.L., Nelson, M.L.: Unsupervised Creation of Small World Networks for The Preservation of Digital Objects. In: 9th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2009), pp. 349–352. ACM, New York (2009)
5. Duranti, L.: The Long-Term Preservation of Authentic Electronic Records: Findings of the InterPARES Project (2001), http://www.interpares.org/book
6. Thibodeau, K.: Building the Archives of the Future. D-lib. Magazine 7, 2 (2001)
7. Greene, M.A., Meissner, D.: More Product, Less Process: Revamping Traditional Archival Processing. The American Archivist 68(2), 208–263 (2005)
8. Duranti, L.: The Archival Bond. Archives and Museum Informatics 11(3-4), 213–218
9. Society of American Archivists. Describing Archives: A Content Standard (DACS). Society of American Archivists, Chicago (2004)
10. Wigdor, D., Perm, G., Ryall, K., Esenther, A., Shen, C.: Living with a Tabletop: Analysis and Observations of Long Term Office Use of a Multi-Touch Table. In: Tabletop 2007, pp. 60–67 (2007)
11. Nebe, K., Klompmaker, F., Jung, H., Fischer, H.: Exploiting New Interaction Techniques for Disaster Control Management Using Multitouch-, Tangible- and Pen-Based-Interaction. In: Jacko, J.A. (ed.) HCII 2011, Part II. LNCS, vol. 6762, pp. 100–109. Springer, Heidelberg (2011)
12. Shen, C., Lesh, N., Vernier, F., Forlines, C., Frost, J.: Building and Sharing Digital Group Histories. In: Proceedings of CSCW 2002 Conference on Computer-Supported Cooperative Work, pp. 324–333. ACM, New York (2002)
13. Kirk, D., Sellen, A., Taylor, S., Villar, N., Izadi, S.: Putting the Physical into the Digital: Issues. In: Designing Hybrid Interactive Surfaces. People and Computers, pp. 35–44. British Computer Society (2009)
14. O'Neill, J., Privault, C., Renders, J.-M., Ciriza, V., Bauduin, G.: DISCO: Intelligent Help for Document Review. In: Global E-Discovery/E-Disclosure Workshop – A Pre-Conference Workshop at the 12th International Conference on Artificial Intelligence and Law, Barcelona (2009)
15. Morris, M.R., Wigdor, D., Lombardo, J.: WeSearch: Supporting Collaborative Search and Sensemaking on a Tabletop Display. In: 2010 ACM Conference on Computer Supported Cooperative Work (CSCW 2010), pp. 401–410. ACM, New York (2010)
16. Terrenghi, L., Kirk, D., Sellen, A., Izadi, S.: Affordances for Manipulation of Physical versus Digital Media on Interactive Surfaces. In: SIGCHI Conference on Human Factors Computing Systems (CHI 2007), pp. 1157–1166. ACM, New York (2007)
17. Grammenos, D., Michel, D., Zabulis, X., Argyros, A.A.: PaperView: Augmenting Physical Surfaces with Location-Aware Digital Information. In: 5th International Conference on Tangible, Embedded, and Embodied Interaction (TEI 2011), pp. 57–60. ACM, New York (2011)
18. Deininghaus, S., Möllers, M., Wittenhagen, M., Borchers, J.: Hybrid Documents Ease Text Corpus Analysis for Literary Scholars. In: ACM International Conference on Interactive Tabletops and Surfaces (ITS 2010), pp. 177–186. ACM, New York (2010)
19. Piper, A.M., Hollan, J.D.: Tabletop Displays for Small Group Study: Affordances of Paper and Digital Materials. In: CHI 2009, pp. 1227–1236. ACM, New York (2009)
20. Sellen, A., Harper, R.: The Myth of the Paperless Office. MIT Press, Cambridge (2002)
21. Kirk, D.S., Izadi, S., Sellen, A., Taylor, S., Banks, R., Hilliges, O.: Opening up the family archive. In: 2010 ACM Conference on Computer Supported Cooperative Work (CSCW 2010), pp. 261–270. ACM, New York (2010)
22. Shipman, F.M., Hsieh, H., Maloor, P., Moore, J.M.: The Visual Knowledge Builder: a Second Generation Spatial Hypertext. In: 12th ACM Conference on Hypertext and Hypermedia, pp. 113–122. ACM, New York (2001)