Visualisation of Hard Drive Content to Support Archival Processes for Personal Digital Archives.

Zoe Bartliff[1], Yunhyong Kim[1], Guy Baxter[2]

Zoe.Bartliff@glasgow.ac.uk, Yunhyong.Kim@glasgow.ac.uk, g.l.baxter@reading.ac.uk

[1] Information Studies, School of Humanities, University of Glasgow, Glasgow, UK

[2] Special Collections Service, University of Reading, Reading, UK

Abstract

This research explores visualisation of data for working with personal digital archives (PDA). Large scale PDAs, comprising content from several personal hard disk drive images, are not receptive to 'open the box and take a look' approaches to appraisal traditionally adopted by analogue archives. By employing a *Sunburst visualisation* to represent file directory structures, this paper demonstrates that it is possible to gain an 'at a glance' comprehension of content organisation and date distribution, whilst concurrently allowing for the dynamic and interactive exploration of information such as usage patterns, content metadata and *original order* relevant to archival appraisal processes.

1 Introduction

Within an archive, the next step after material acquisition is usually appraisal, whereby materials are reviewed for sensitivity, meeting legal mandates, and determining archival value. Standards and workflows defining this selection and appraisal process are less than consistent for digital archives (Smith, Gooding and Mann, 2019). One of the most challenging, yet important, aspects of appraisal and cataloguing of PDAs is the vast quantities of material they contain. With the potential to contain millions of diverse individual records, this archival challenge only becomes more prominent as cheaply available large-scale storage increases. Oftentimes, manual workflows are applied in the digital context (Chassanoff and Altman, 2019), which is labour-intensive for the archivist and prone to errors. They are therefore largely agnostic of archivists' appraisal and cataloguing needs, for example, to reflect the recent trend of adopting minimal processing approaches.[1]

Visualisations are becoming increasingly popular for navigating large digital collections (cf. Windhager et al., 2018). Some have been used for specific types of content (Hangal, Lam, and Heer, 2012), summaries of documents (Collins, Carpendale and Penn, 2009), or appraisal (Xu et al 2010). Each of these methods, however, remove the content from its *original order* and/or context of creation, making it difficult to 'capture, collate, analyze, and organize information about material that serves to identify it and to explain its context and the systems that produced it' (Bureau of Canadian Archivists, 2008). The research presented here distinguishes itself by

[1] cf. https://blogs.loc.gov/thesignal/2012/08/more-product-less-process-for-born-digital-collections-reflections-on-curatecamp-processing/

exploring a visualisation method that captures the file directory structure as a bridge between the interests of keeping original order and the movement towards prioritising selection and appraisal (Cook, 1997).

2 Supplementing the workflow with Sunburst Visualisations

The Sunburst visualisations have been employed in various sectors focused on hierarchical data (Collins et al., 2009, Stasko, 2000). The primary reason for their popularity is the immediacy of both analysis and data access for the user with minimal training, something highlighted by the pioneering scholars of the technique (Stasko, 2000).

Further to this, they allow for the embedding of great quantities of additional metadata information, whilst preserving the simple, easy-to-engage-with presentation. The hover-over metadata function and the dynamic animations offered by the Plotly Sunburst library[2] make the visualisation appealing and engaging. It ensures a high level of detail comparable to Xu et al's (2010) comprehensive treemaps whilst also being intuitive to use. The colour contrasts can also be employed to differentiate further categories within the data relevant to appraisal. The Sunburst visualisation is an ideal interface for keeping the file directory order as it was created while accentuating the multifaceted categories of interest to archival appraisal processes.

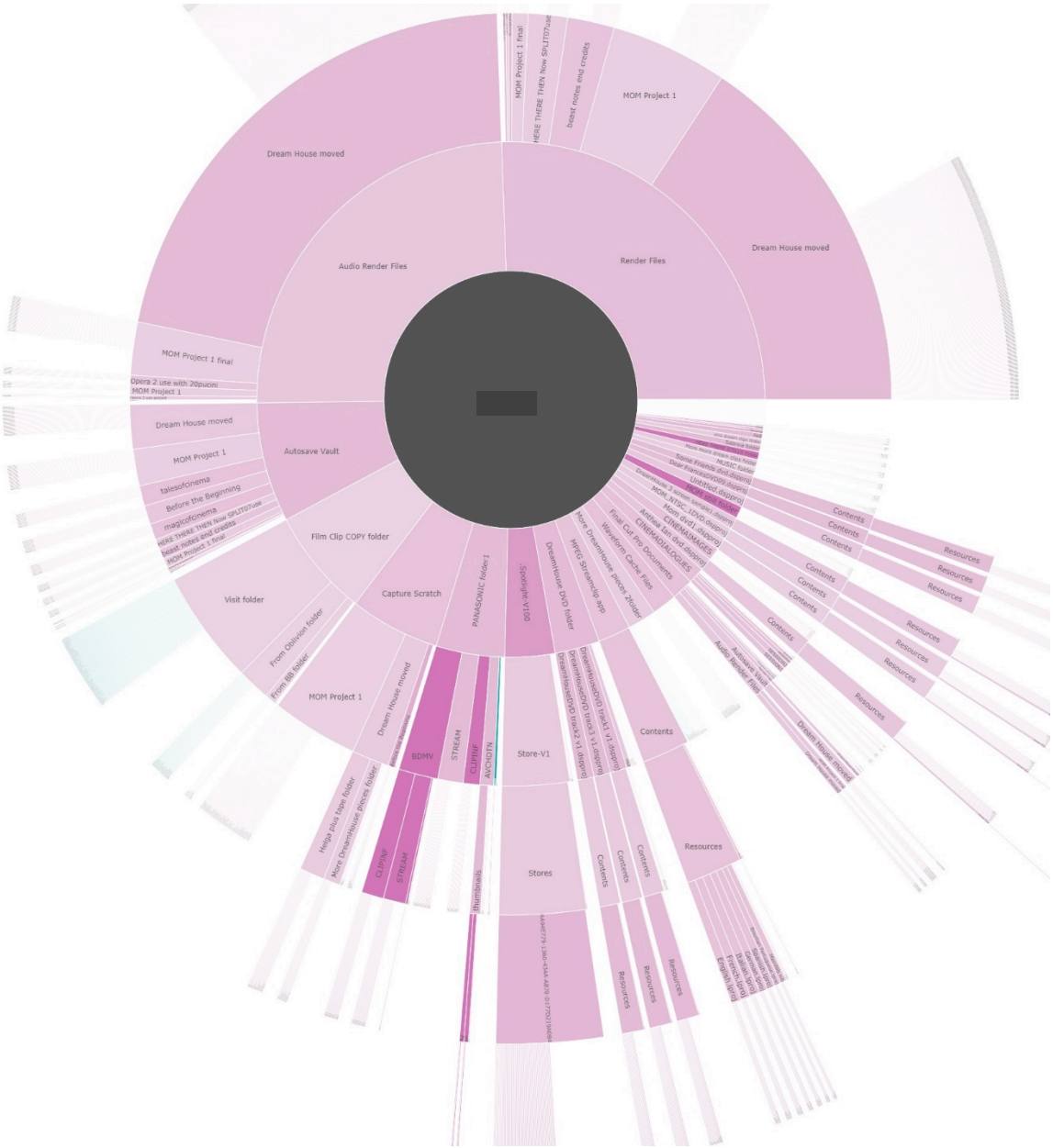
The Sunburst format is attuned to the archival need to gain an 'at a glance' idea of the structure and distribution of the data. To demonstrate, we have visualised two hard drives from a filmmaker's collection,[3] to display the file directory structures (Figure 1 & 2). These demarcate, by colour, the dates at which the creator last modified the file where dark teal represents dates in 2000 and dark pink those in 2012. From this it is immediately evident that Figure 1 largely represents older data than Figure 2, but also that both drives include activity throughout the date range. Figure 1 suggests later evidence of modification on this drive to have been quite focused, yet earlier data to have been maintained. Figure 2, conversely, suggests, that data on this drive was modified during a concentrated period with a limited number of earlier files mostly buried in the deeper levels of the directory structures. This suggests distinctly different patterns of usage. Careful examination of the directory structures also reveals repeated structures related to selected project software. For the archivist, such demarcation can aid in identifying potential relationships between concurrent data and relevant software, something that could be particularly useful across a multi-drive collection. Knowing the period of modification for a project and seeking those files within a similar categorisation could help to identify, without disturbing the original order of the drive, related material which could otherwise have been missed.

In addition to this timeline focus, other information related to the working environment and/or content could be easily incorporated (cf. Xu et al 2010). This would allow us to avoid additional data processing and the creation and analysis of frequency-based visualisations (e.g. line graphs or bar charts) therefore lowering the overall workload.

[2] https://plotly.com/python/sunburst-charts/

[3] University of Reading Special Collections MS5502


Maxtor int HD 2 bkup folder
lost+found
Final Cut Pro Documents
Audio Render Files
Irene 'C' Project_02 JUN 05**
Oblivion Project_fm JUN 05**
Render Files
Autosave Vault
Contents
Resources
Adobe InDesign CS
InDesign CS-app
Help
Plug-Ins
Microsoft Office 2004 french
Office
Mode`les
Clipart
Themes
Aide
Fonts
Contents
Resources
Images
Version_Cue
Audio Files
more RAM
less RAM
Demo Songs
SharedSupport
Presentations
Conceptions
Required
MacOS
Online
Proximity
AUMLibrary.cfm
Dictionaries


Dream House moved
MOM Project 1 final
HERE THERE THEN Now SPLIT07.sae
beast notes end credits
MOM Project 1
Dream House moved
Audio Render Files
Render Files
MOM Project 1 final
Opera 2 use with 20pucini
MOM Project 1
Dream House moved
MOM Project 1
talesofcinema
Before the Beginning
magicofcinema
HERE THERE THEN Now SPLIT07.sae
beast notes and credits
MOM Project 1 final
Autosave Vault
Film Clip COPY folder
Capture Scratch
Visit folder
From Oblivion folder
From BB folder
MOM Project 1
Dream House moved
PANASONIC folder 1
SpotOn-V100
DreamHouse DVD folder
MPEG Streamclip app
Home Dreamhouse pieces folder
Final Cut Pro Documents
Waveform Cache Files
Audio Render Files
Autosave Vault
Contents
Resources
Store-V1
Stores
Contents
Resources
BDMV
STREAM
CLIPINF
AVCHDTN
Helga plus cage folder
Mom DreamHouse pieces folder
CLIPINF
STREAM
thumbnails

3 Conclusion and future work

We have demonstrated that Sunburst visualisations have exceptional potential to allow the archivist to take full advantage of the flexibility of the digital format, whilst also honouring the original order of the material (Cook, 1997).

The next steps forward for this work would be to evaluate the visualisations in collaboration with archivists (cf. Lemieux 2015) and to integrate them with common archival workflows, whether directly or through a simple lightweight proxy. This would allow archives to supplement their catalogue/database and employ the visualisation as part of archival management and access strategies. For example, the possibility of integrating hyperlinks into the segments which connect to file content and appropriate software, could further ease the burden upon the archivist's cognitive processes during appraisal permitting both `overview' and `details on demand' (Lemieux 2015). Such a visualisation would also align with observations made elsewhere to aid researchers seeking information within the archive (cf. Windhager et al., 2018).

To summarise, the visualisation serves as a foundation for 'a generous interface' which 'would also enrich interpretation by revealing relationships and structures within a collection' (Smith et al., 2019) and facilitate actions based on these relationships.

Acknowledgement

This work was supported by the Arts and Humanities Research Council [AH/R007012/1].

References

[1] Bureau of Canadian Archivists. (2008) *Rules for Archival Description*. url:http://www.cdncouncilarchives.ca/RAD/RAD_Principles_July2008. pdf.

[2] Chassanoff and Altman. (2019) "Curation as "Interoperability With the Future": Preserving Scholarly Re-search Software in Academic Libraries". In: *Journal of the Association for Information Science and Technology* doi:10.1002/asi.24244.

[3] Collins, Carpendale, and Penn. (2009) "DocuBurst: Visualizing Document Content using Language Structure". In: *Computer Graphics Forum* 28.3, pp.1039–1046. doi:10.1111/j.1467-8659.2009.01439.x.

[4] Cook. (1997). "What is Past is Prologue: A History of Archival Ideas Since 1898, and the Future Paradigm Shift". In: *Archivaria* 43

[5] Smith, Gooding and Mann. (2019) "The forensic imagination: interdisciplinary approaches to tracing creativity in writers' born-digital archives". In: *Archives and Manuscripts* 47.3, pp.374–390. doi:10.1080/01576895.2019. 1608837.

[6] Hangal, Lam, and Heer. (2012) "Processing Email Archives in Special Collections". url:http://xenon.stanford.edu/~hangal/dh2012.pdf.

[7] Lemieux (2015). Visual analytics, cognition and archival arrangement and description: studying archivists' cognitive tasks to leverage visual thinking for a sustainable archival future in: *Archival Science*, 15(1), pp.25-49.

[8] Stasko. (2000) *Information Interfaces: Sunburst.* url:https://www.cc.gatech.edu/gvu/ii/Sunburst/.

[9] Windhager et al. (2018) "Visualization of cultural heritage collection data: State of the art and future challenges". In: *IEEE transactions on visualization and computer graphics* 25.6, pp.2311–2330.

[10] Xu, Esteva, & Dott. (2010). "Visualization for archival appraisal of large digital collections". In: *Archiving Conference*. *Society for Imaging Science and Technology*. pp.157-162