Manuscript Fragments, Reused in Bindings, as Untapped Genealogical Sources

Opportunities for Family History and Manuscript Studies to Easily Provide Information for One Another, All with the Help of Machine Learning

David Black Keith Prisbrey

FamilySearch Library Volunteers.
However, this research is personal and not promoted by nor tied to FamilySearch

(The program for the Workshop/Conference is below the paper, with my part highlighted.)

1. Introduction

We present a novel use of computer vision with the objective of searching for untapped information written on bindings made from previously-used manuscripts. The main data for which we are searching are instances of a trash piece of parchment or paper having a manuscript – such as a contract, deed, arrest record, roll call of monks, or part of a literary or theological work – as part of its binding.

It is important to note that many advances in computer vision currently come from the use of foundational models[1][2][3] that incorporate the latest advances in neural networks for image classification. Our main procedure consists of training a Residual Neural Network to identify cases of reused manuscript fragments from collections of document images. The collections may contain images from one codex, i.e. one book, or from many codices.

Specifically, as to our usage of a ResNet, we use the *ResNet-50* model[4]. Using the pre-trained weights from the huge ImageNet[5] dataset as a starting point, we trained on over 600 images, about 1 in 10 of which contained examples of Reused Manuscript Fragments.

With a random 95:5 split between training and test sets, we obtained a 96.77% accuracy on the training set after 30 epochs of training.

To further test the applicability of such an approach in finding other interesting objects in manuscripts, we used a set of over 200 images, about half of which showed pages containing stitching used for parchment repair, to train and test. Again using a 95:5 split between training and data sets, we obtained a training accuracy of 100%.

We conclude that such transfer learning from ResNet-50, using initial weights from the ImageNet dataset, is capable of finding information that is new and useful to both genealogists and experts in manuscript studies. A tool to find such information has a very cheap price in training.

It is our hope that large, archival and family-history databases such as those held by FamilySearch and similar corporate organizations will be searched for genealogically relevant fragments as well as for fragments of literary, theological, and other works that are interesting to the manuscript studies community. As far as we know, no searches for Manuscript Reuse Fragments have been performed on the genealogical and archival databases. Along with this search, it is hoped that we can search the university and state libraries used by those in manuscript studies; we would find new material for them and would also have new sources of names, dates, and relationships to forward our work in family history.

2. The Manuscript Fragments: Definitions and Examples

We first show example document images where part of an old and unwanted manuscript has been detached from its original codex and used to help bind and protect another document. (We will continue to refer to such detached parts as Reused Manuscript Fragments, also sometimes using the initialism, RMF, and sometimes using phrases such as "examples of reuse".) These images, their sources, and their usage information are

available (and are bigger) at the GitHub of David Black, specifically at https://github.com/bballdave025/manuscript-waste-reuse-finder. They follow in Figure 1.

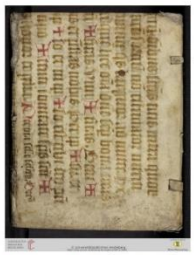
Fig. 1a (reuse)


Fig. 1b (no reuse)


Fig. 1c (no reuse)


Fig. 1d (no reuse)

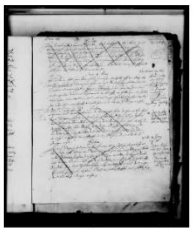
Fig. 1e (no reuse)


Fig 1f (reuse)


Fig. 1g (reuse)

1a-b : Heidelberg, Cod Sal VII,73
1c-e : FamilySearch, DGS 007996631 (item 1)
1f : Bibliothèque Nationale de France
1g : FamilySearch, DGS 00

We will continue to refer to such detached parts as Reused Manuscript Fragments, also sometimes using the initialism, RMF, and sometimes using phrases such as "examples of reuse". A whole field, known as fragmentology, deals with such materials. An open-access and peer-reviewed journal, appropriately named Fragmentology[6], has been publishing a yearly issue since 2018. Many consider the work of Ker to be foundational, especially his *Fragments of Medieval Manuscripts Used as Pastedowns in Oxford Bindings with a Survey of Oxford Binding c. 1515-1620*[7]. It has been a known concern, if not a defined area of manuscript studies and codicology since, at latest, the dissolution of the monasteries in England (1536-1540), as studied by A. Reynolds[8].

There are several international efforts, whose names and websites[9-14] are given in the References, to catalog RMFs and to reconstruct as much of the originals from which they came as possible, inspired to use then-emergent technological tools. However, to our knowledge, the searches are done manually, with some reference to the metadata kept by libraries and collections. The only reference to AI we found in the field of manuscript studies was the mention of a grant to help use handwriting recognition and similar technologies to catalog Hebrew Manuscript Fragments, supported, among others, by the European Research Council and the National Library of Israel[15].Also, to our knowledge, no attempts to find, catalog, and reunite RMFs use such collections as those at FamilySearch and Ancestry®. We have the opportunity to add powerful tools and a huge dataset to such efforts, as well as to receive information about fragments with genealogical value.

This leads me to a favorite example of the new and important information that can be gleaned from RMFs. Librarian C. de Hamel bought some dirty old fragments in 2001[16]. He later sent them to two colleagues he thought could help unravel their origin. A few weeks later they had solved the mystery. National Geographic reported on it, as did the UK's online newspaper, *The Independent*. From the latter:

> Dr Simon Corcoran and Dr Benet Salway of the history department at University College London have found fragments of an important Roman law code that previously had been thought lost forever[17][18].

The dirty old binding fragments were part of the known but (until then) lost *Codex Gregorianus,* part of the Justinian Law Code that underpins much of today's Western laws.

References:

[1] Rick Merritt. "What Are Foundational Models?." nVidia Blog. Posted online March 13, 2023. Accessed January 30, 2024. URL : https://blogs.nvidia.com/blog/what-are-foundation-models

[2] BasicAI Marketing Team. "The Foundational Model: Key Facts and Insights." BasicAI Blog. Posted online January 3, 2024. Accessed January 30, 2024. URL : https://www.basic.ai/post/what-is-the-foundation-model.

[3] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein et al. "On the opportunities and risks of foundation models." arXiv preprint arXiv:2108.07258 (2021).

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition." in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. Accessed at arXiv:1512.03385 on January 30, 2024.

[5] J. Deng, W. Dong, R. Socher, L. -J. Li, Kai Li and Li Fei-Fei. "ImageNet: A large-scale hierarchical image database." in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Miami, FL, USA. (2009). pp. 248-255. DOI: 10.1109/CVPR.2009.5206848.

[6] William Duba, Chstoph Flüeler, Veronika Drescher, editors. "A Journal for the Study of Medieval Manuscript Fragments." *Fragmentology* [Online], Volume VI (2023), published on December 31, 2023, accessed January 27, 2024.
URL: https://www.fragmentology.ms ;
DOI: https://doi.org/10.24446/vb1n.

[7] Neil Ripley Ker. *Fragments of Medieval Manuscripts Used as Pastedowns in Oxford Bindings, with a Survey of Oxford Binding c.1515–1620*. Oxford: Oxford Bibliographical Society, A.T. Broome, 1954.

[8] Anna Reynolds. "'Such dispersive scattredness[sic]': early modern encounters with binding waste." *Journal of the Northern Renaissance (8)*. (2017).

[9] Fragmentarium - International Digital Research Lab for Medieval Manuscript Fragments: http://fragmentarium.ms/

[10] Broken Books Project - Center for Digital Humanities, College of Arts and Sciences at Saint Louis University : https://brokenbooks.omeka.net/

[11] Digital Analysis of Chant Transmission (DACT) : http://dact-chant.ca/

[12] Medieval Manuscript Fragment Project - University College, London : http://dh2015.org/abstracts/xml/TERRAS_Melissa_Collaborative_Digitisation_UCL_s_/TERRAS_Melissa_Collaborative_Digitisation_UCL_s_Mediev.html

[13] Lost Manuscripts - Centre for Bibliographical History, University of Essex : http://www.lostmss.org.uk/project

[14] Books within books: Hebrew Fragments in European Libraries : http://hebrewmanuscript.com/

[15] National Library of Israel Announcements. "New AI Project to Enable Full-Text Searches of Medieval Manuscripts." in National Library of Israel Announcements. (2022). Accessed January 30, 2024.
URL: https://www.nli.org.il/en/at-your-service/announcements/european-research-council-awards

[16] Christopher de Hamel, "Chapter 10: Theodor Mommsen", in *The Manuscripts Club: The People Behind a Thousand Years of Medieval Manuscripts*, United States: Penguin Publishing Group, 2023. 363–364.

[17] Simon Corcoran, "The Gregorianus and Hermogenianus assembled and shattered." *Mélanges de l'École française de Rome - Antiquité* [En ligne] 125-2 (2013), posted on December 19, 2013, accessed January 30, 2024. URL: http://journals.openedition.org/mefra/1772. DOI: https://doi.org/10.4000/mefra.1772.

[18] Malcolm Jack. "Cracking the codex: Long lost Roman legal document discovered." *The Independent* [online], posted on January 28, 2010, accessed January 30, 2024. URL : https://www.independent.co.uk/life-style/history/cracking-the-codex-long-lost-roman-legal-document-discovered-1881769.html

BRIGHAM YOUNG UNIVERSITY

Family History TECHNOLOGY WORKSHOP

2024 Program

February 27, 2024

Events take place in the Upstairs Hall of the Hinckley Center

9:00 a.m. Welcome

9:05 a.m. Keynote

Anna Scius-Bertrand (University of Fribourg and HES-SO)

9:45 a.m. Developer Talks /Research Talks (15 minutes each)

1. Storied Biography Extraction Technology - Laryn Brown
2. Connected Cities - Mike Ostler, Debbie Ostler
3. LivingHistory.AI - Jonathan Gibson
4. EmulateMe - Ariel Mathov, Maximiliano Ejberowicz

10:45 a.m. Break

11:00 a.m. Lightning Talks (3 minutes each)

1. Our Beginnings - Jonah Austin
2. Powerlinker - Sam Carlsen and Zarin Loosli
3. Yellow temple project - Tiberius Baker
4. Segmenting US census records - Jackson Roubidoux
5. US Surname Lexicon - Spencer Timmerman
6. Using family history tasks to create meaningful employment - Meg Wright
7. Sparse-data linking with land records - Zach Flynn
8. Who, Me? Gather Israel? - William Mickelson
9. User-friendly family history textbook - Kristilee J. Manuel
10. Using Google Gemini to Translate 133 Languages - James Tanner
11. Family Tree Validator - Bob Scott
12. Kindex - Kimball Clark
13. Family Migration Research Using FamilySearch Data: Updates and Potentials - Sam Otterstrom
14. Multiplying handwriting training data in lower-resource settings - Seth Stewart

12:00 p.m. Lunch

1:00pm. Developer Talks/Research Talks (15 Minutes Each)

1. Weakly Supervised Information Extraction from Semi-Structured Document Images - Fabian Wolf, Oliver Tuselmann, Christoph Rass, Gernot A. Fink
2. Extracting Handwritten Historical Data from Registry Forms - Jade Martinez, Lyla Wortham, Nishatul Majid, Elisa H. Barney Smith
3. Manuscript Fragments, Reused in Bindings, as Untapped Genealogical Sources - **David Black**, Keith Prisbrey
4. Universal Foundational "Large" Language Models for Learning Genealogical Processes from Inception - Patrick Schone

2:00 p.m. Afternoon Break

2:15 p.m. Developer Talks / Research Talks (15 Minutes)

1. The Impact of Family History Technology on Social and Emotional Well-being - Emma Ausman, Isabella Stephens, Christian Hall
2. A Prototype for Splitting Munged Persons in Family Tree - Randy Wilson

2:45-3:15 p.m. Closing Keynote

Goldie May - Richard Miller

3:15-3:30 p.m. Concluding Remarks