Breathing New Life into Death Certificates: Extracting Cause of Death from Handwritten Records

Supplemental Appendix:

B. Digitization of images

High quality images are essential for accurate handwriting recognition. Our project worked with images that had been microfilmed and scanned by other organizations. However, future researchers looking to apply handwriting recognition may be working with collections that have not yet been digitized. In this appendix, we discuss important issues that may be encountered between encountering handwriting on paper and having images suitable for handwriting recognition. The aim is to equip researchers with the ability to approach digitization of collections aware of factors important for successful application of the methods in this paper, whether they are responsible for creating the images, or working with other organizations. We use the terms “imaging” and “digitization” interchangeably to refer to the process of creating a digital image of a handwritten record. Imaging of large collections should not be undertaken casually. However, with well organized manuscript material and the appropriate physical set-up for photography or scanning, it is realistic for researchers to digitize up to 2,000 – 10,000 page images in a week. At this volume of material, handwriting recognition of the text may be efficient, even after allowing for the fixed costs described above.

We discuss five issues that researchers should consider when approaching digitization.

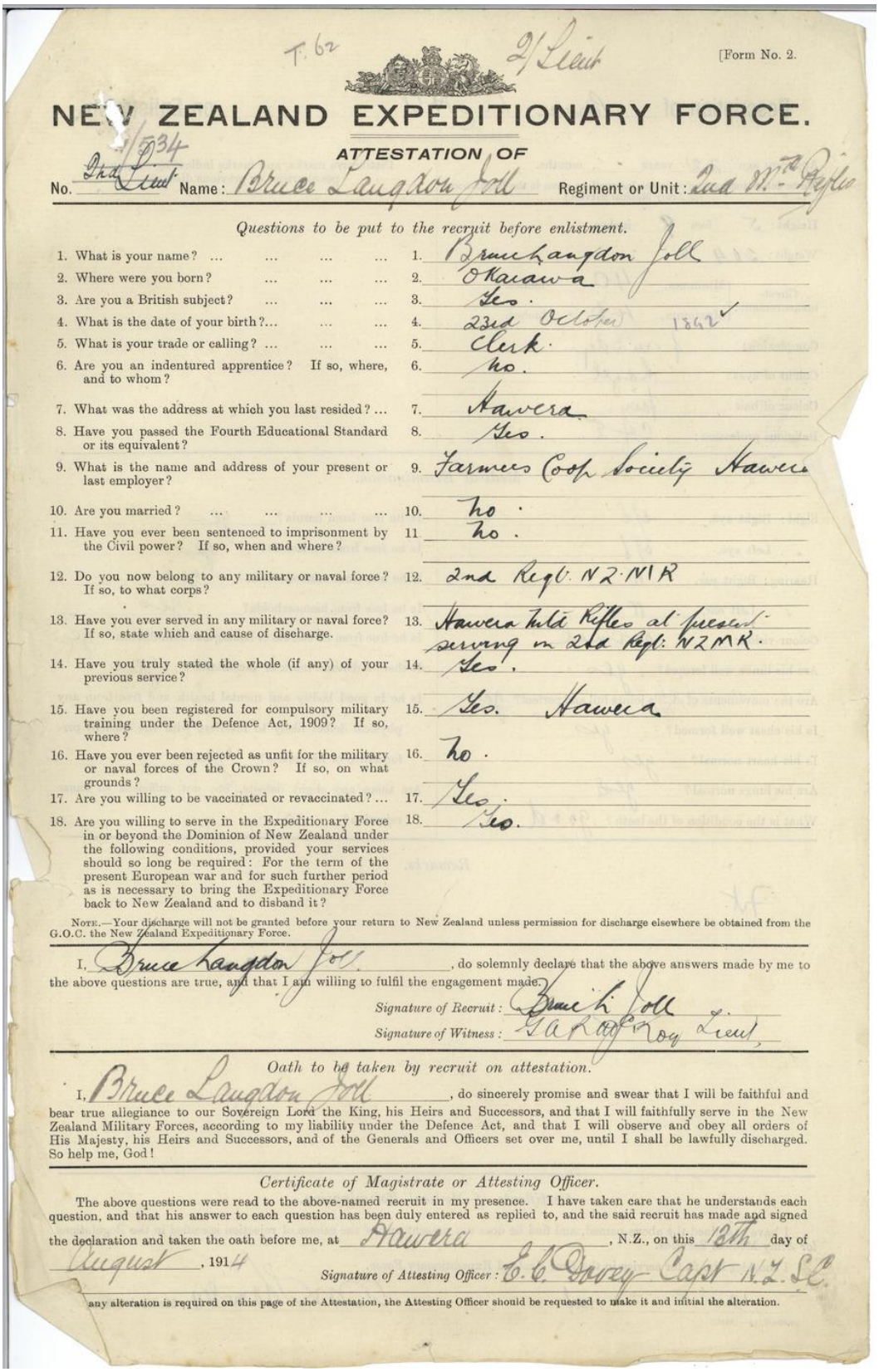
1. Resolution
2. Sharpness
3. Contrast
4. Coordinate locations of handwriting
5. Organization of image collections

Some of the examples illustrating these issues are drawn from an Archives New Zealand collection of digitized World War I enlistment records.[1] The digital images in this collection were created in a variety of different processes that usefully illustrate different quality images of similar source material. Some material was scanned directly from original paper records between 2010 and 2014. However, other material in these collections was microfilmed in the 1960s by microfilm operators with varying training. Some of the original paper forms were then destroyed. Material from the microfilms was printed in the past two decades, and then re-scanned. Thus, the collection contains substantial variation in image quality, illustrating some of the challenges that may be encountered.

To fix ideas about appropriate standards for image quality, we begin with an example of an image scanned directly from the original paper records. Note that the original paper forms are beige, but the handwriting is largely in sharp black ink, while there are also handwritten elements that have faded or were originally penned in lighter ink. As an aside, the layout of these forms is more challenging for handwriting recognition than the death certificates.

[1] https://www.archives.govt.nz/research-guidance/research-guides/war-records/world-war-one-1914-1918

Figure B.1: Image scanned at acceptable resolution, sharpness and contrast



T. 62 2/Lieut [Form No. 2.

NEW ZEALAND EXPEDITIONARY FORCE.

ATTESTATION OF

No. 1/534 2nd Lieut Name: Bruce Langdon Joll Regiment or Unit: 2nd N.Z. Rifles

Questions to be put to the recruit before enlistment.

1.	What is your name?	1. Bruce Langdon Joll
2.	Where were you born?	2. Okaiawa
3.	Are you a British subject?	3. Yes.
4.	What is the date of your birth?	4. 23rd October 1892 ✓
5.	What is your trade or calling?	5. Clerk.
6.	Are you an indentured apprentice? If so, where, and to whom?	6. No.
7.	What was the address at which you last resided?	7. Hawera
8.	Have you passed the Fourth Educational Standard or its equivalent?	8. Yes.
9.	What is the name and address of your present or last employer?	9. Farmers Coop Society Hawera
10.	Are you married?	10. No.
11.	Have you ever been sentenced to imprisonment by the Civil power? If so, when and where?	11. No.
12.	Do you now belong to any military or naval force? If so, to what corps?	12. 2nd Regt. N.Z.M.R
13.	Have you ever served in any military or naval force? If so, state which and cause of discharge.	13. Hawera Mtd Rifles at present serving in 2nd Regt: N.Z.M.R.
14.	Have you truly stated the whole (if any) of your previous service?	14. Yes.
15.	Have you been registered for compulsory military training under the Defence Act, 1909? If so, where?	15. Yes. Hawera
16.	Have you ever been rejected as unfit for the military or naval forces of the Crown? If so, on what grounds?	16. No.
17.	Are you willing to be vaccinated or revaccinated?	17. Yes.
18.	Are you willing to serve in the Expeditionary Force in or beyond the Dominion of New Zealand under the following conditions, provided your services should so long be required: For the term of the present European war and for such further period as is necessary to bring the Expeditionary Force back to New Zealand and to disband it?	18. Yes.

NOTE.—Your discharge will not be granted before your return to New Zealand unless permission for discharge elsewhere be obtained from the G.O.C. the New Zealand Expeditionary Force.

I, Bruce Langdon Joll, do solemnly declare that the above answers made by me to the above questions are true, and that I am willing to fulfil the engagement made.

Signature of Recruit: Bruce L Joll

Signature of Witness: [illegible] Lieut.

Oath to be taken by recruit on attestation.

I, Bruce Langdon Joll, do sincerely promise and swear that I will be faithful and bear true allegiance to our Sovereign Lord the King, his Heirs and Successors, and that I will faithfully serve in the New Zealand Military Forces, according to my liability under the Defence Act, and that I will observe and obey all orders of His Majesty, his Heirs and Successors, and of the Generals and Officers set over me, until I shall be lawfully discharged. So help me, God!

Certificate of Magistrate or Attesting Officer.

The above questions were read to the above-named recruit in my presence. I have taken care that he understands each question, and that his answer to each question has been duly entered as replied to, and the said recruit has made and signed the declaration and taken the oath before me, at Hawera, N.Z., on this 13th day of August, 1914

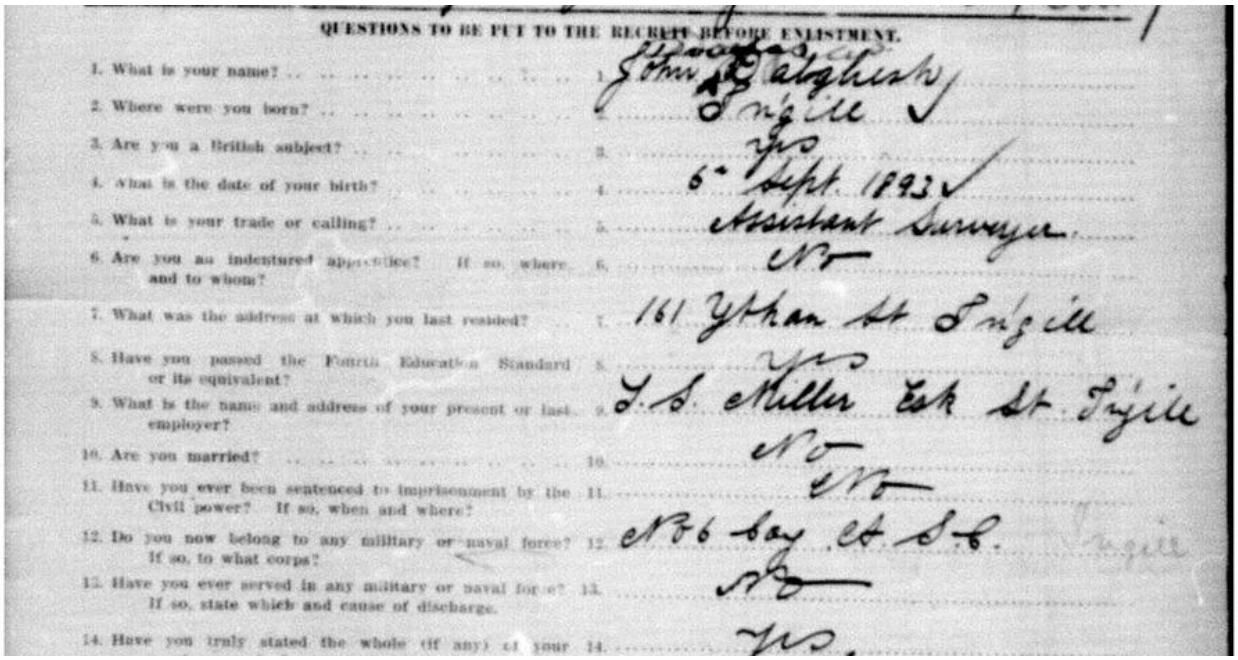
Signature of Attesting Officer: E. C. Davey Capt N.Z.S.C.

any alteration is required on this page of the Attestation, the Attesting Officer should be requested to make it and initial the alteration.

Resolution: Researchers should begin with the highest resolution possible and experiment with how much image compression can be obtained without compromising readability (for humans or computers). In most settings, a minimum acceptable resolution for printed text is 150 dots per inch (dpi). For both handwriting and printed text it is preferable to work with images photographed or scanned at a resolution of at least 300-400 dpi (within the capabilities of today's digital cameras). We recommend that researchers begin imaging around these resolutions, and test their workflow through handwriting recognition before deciding to capture images at lower resolution. While higher resolution images require more storage space, this is unlikely to be a constraint until researchers are working with collections of 100,000 images or more. At this volume of records it is likely researchers will be collaborating with archivists who can advise on appropriate imaging techniques.

Sharpness: Information is lost in successive reproduction and copying of material between formats. Thus, material originally converted to microfilm and then scanned is likely to be less sharp than the same material captured with modern digital imaging equipment. However, conversion from microfilm is a relatively straightforward process, and much quicker than native capture of images from manuscript material. Thus, researchers should examine the quality of converted microfilm images and consider whether new images are necessary. While much material microfilmed in the twentieth century and subsequently converted to digital images is of acceptable quality for handwriting recognition, this is not uniformly the case. Figure B.2. illustrates how the smaller printed text on the page has become harder to read. The image in Figure B.2 was first microfilmed, reprinted, and then scanned. At each stage, sharpness was lost. Because the handwriting is larger, it is less affected by these issues. Were readers to encounter handwriting in a smaller script, the lack of sharpness would likely prove problematic for handwriting recognition.

Figure B.2. Image with fuzzy elements through multiple reproductions and copying



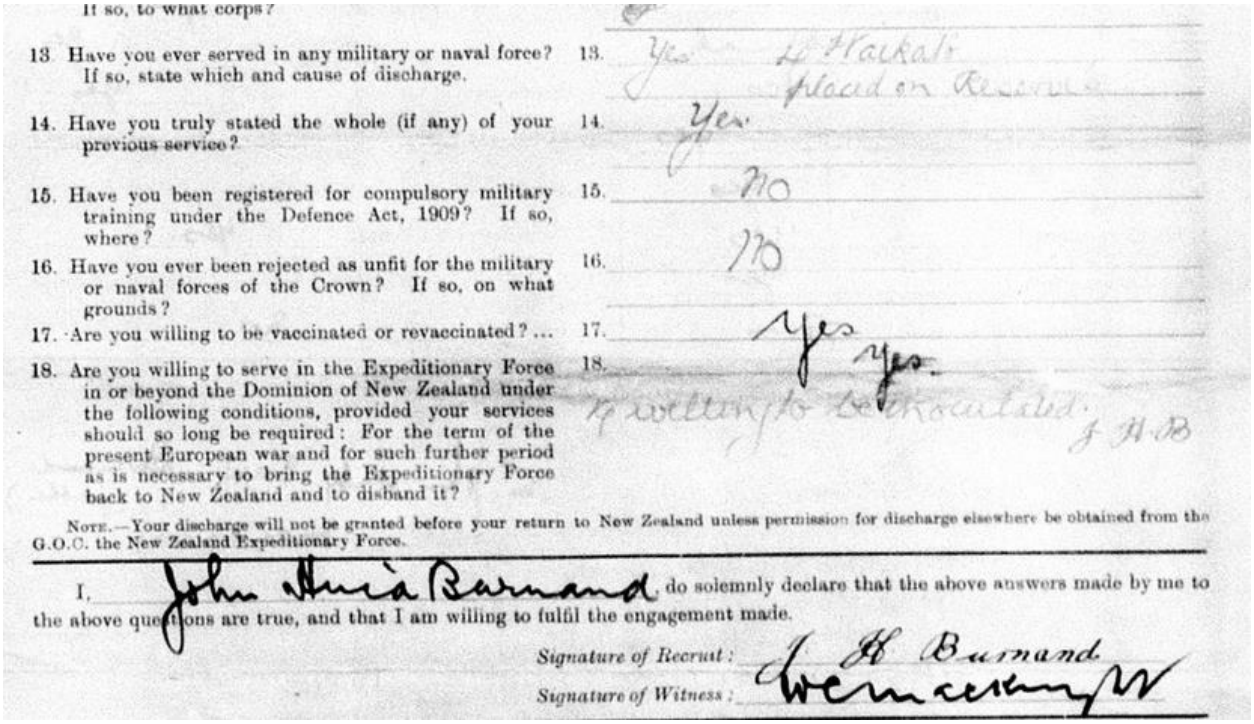
QUESTIONS TO BE PUT TO THE RECRUIT BEFORE ENLISTMENT.

1. What is your name? .. .. 1. John [illegible] Laughlin
2. Where were you born? .. .. 2. Orillia
3. Are you a British subject? .. .. 3. Yes
4. What is the date of your birth? .. .. 4. 6th Sept 1893
5. What is your trade or calling? .. .. 5. Assistant Surveyor
6. Are you an indentured apprentice? If so, where and to whom? 6. No
7. What was the address at which you last resided? .. 7. 161 Ythan St Orillia
8. Have you passed the Fourth Education Standard or its equivalent? 8. Yes
9. What is the name and address of your present or last employer? 9. T. S. Miller Esk St Orillia
10. Are you married? .. .. 10. No
11. Have you ever been sentenced to imprisonment by the Civil power? If so, when and where? 11. No
12. Do you now belong to any military or naval force? If so, to what corps? 12. No 36 Bay Ct S.C.
13. Have you ever served in any military or naval force? If so, state which and cause of discharge. 13. No
14. Have you truly stated the whole (if any) of your 14. Yes

Contrast. Contrast between the handwritten text and the page background is important for accurate machine recognition. Non-white paper, or paper that has yellowed over time due to age may produce images where the handwriting cannot always be clearly distinguished. In Figure B.3. the two “Yes” answers in response to questions 17 and 18 are likely to be of sufficient contrast that handwriting recognition software will identify them. However, questions 13-16 have low contrast, and handwriting recognition software may struggle to read the full response. Note in Question 13 that the contrast of the text varies, and the final word of the second line is unlikely to be successfully picked up by handwriting recognition tools (the final line of question 13 says “Placed on Reserve”). Low contrast images may be inherent in the physical materials that are being examined if the paper or ink (or both) have degraded over time. Careful imaging of this material is necessary. Post-processing of images to enhance contrast can improve the contrast of some material that is hard to read on the original paper documents. Particular care should be taken with handwriting that is not in black or blue ink. Ideally, researchers will make color images to retain details that would be lost with a conversion to black and white. It is not uncommon for color to be used as a variable in data of the form that economic historians will be working with. For example, some death registers record different modes

of death in different colors; while red and black have substantive meaning in accounting data and can be observed in use in primary source material.

Figure B.3. Image illustrating handwriting with low contrast



If so, to what corps?

13. Have you ever served in any military or naval force? If so, state which and cause of discharge. 13. Yes 2 Waikato Placed on Reserve

14. Have you truly stated the whole (if any) of your previous service? 14. Yes.

15. Have you been registered for compulsory military training under the Defence Act, 1909? If so, where? 15. No

16. Have you ever been rejected as unfit for the military or naval forces of the Crown? If so, on what grounds? 16. No

17. Are you willing to be vaccinated or revaccinated? ... 17. Yes

18. Are you willing to serve in the Expeditionary Force in or beyond the Dominion of New Zealand under the following conditions, provided your services should so long be required: For the term of the present European war and for such further period as is necessary to bring the Expeditionary Force back to New Zealand and to disband it? 18. Yes. Prepared to be vaccinated. J H B

NOTE.—Your discharge will not be granted before your return to New Zealand unless permission for discharge elsewhere be obtained from the G.O.C. the New Zealand Expeditionary Force.

I, John Henry Barnard, do solemnly declare that the above answers made by me to the above questions are true, and that I am willing to fulfil the engagement made.

Signature of Recruit: J H Barnard

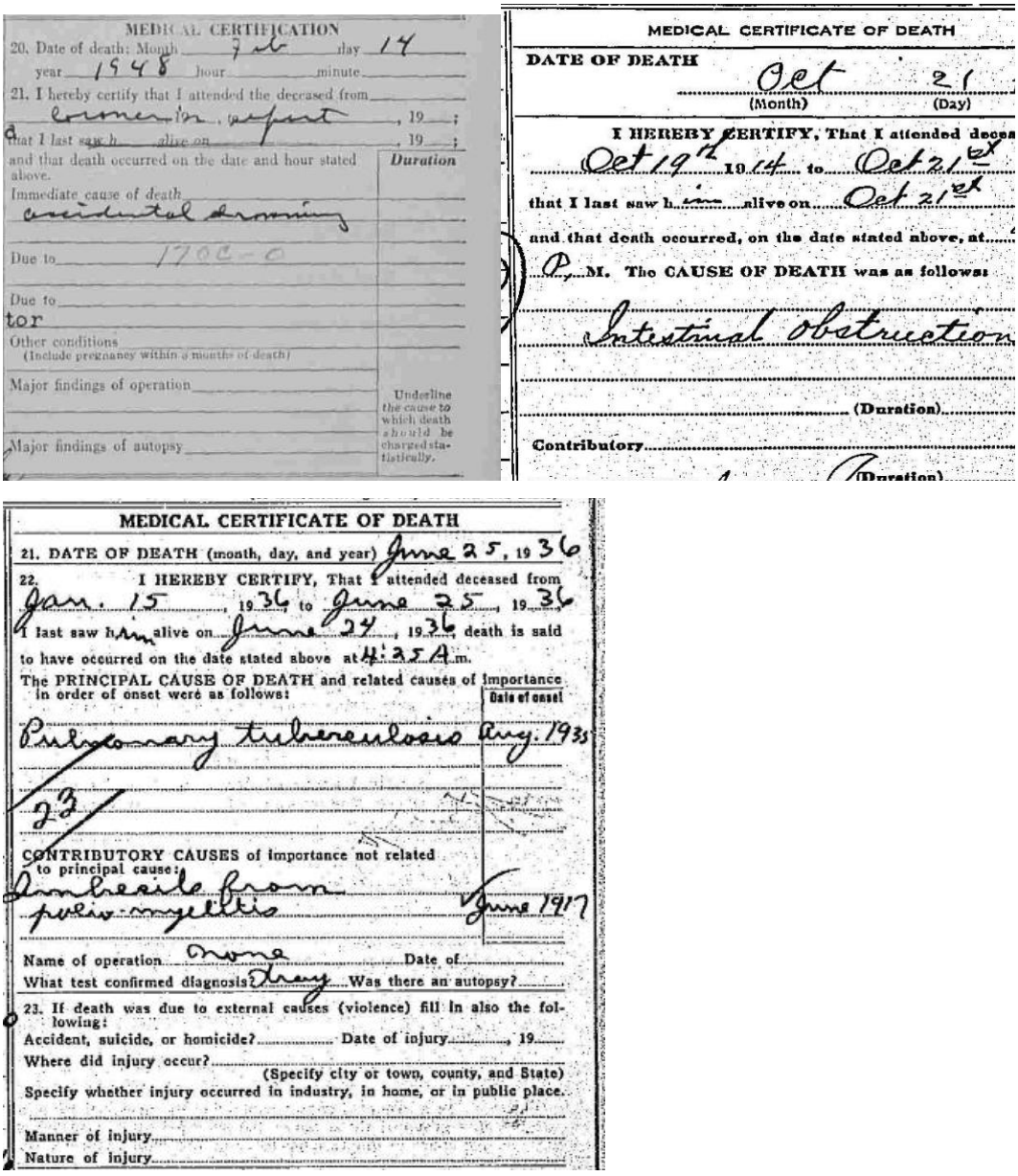
Signature of Witness: [illegible]

Coordinate location of variables: Economic historians will often be interested in identifying and recognizing multiple variables on the same page. An unlabeled string of words is unlikely to be useful without significant post-processing and parsing of the data. Moreover, the same words used in handwritten responses may refer to different concepts. Abstracting from any issues with the handwriting itself, researchers want to know whether “Columbus”, for example, refers to a birthplace, place of residence, or place of death. “Potter” may be an occupation or a surname. Thus, as well as recognizing where handwriting is located on the page it is important to segment the page into distinct variables.

Often the location of handwriting of interest can be identified through the presence of printed text. When digitizing collections that span a long period of time, the text identifying handwriting of interest may change. In Figure B.4, both the text identifying the primary cause of death changed over time, being described as the “cause of death,”

"immediate cause of death," and "principal cause of death". Similarly, the language describing contributory causes varied from one word ("contributory") to a fuller explanation of what the concept meant. As discussed in the body of the article, OCR, Mask-RCNN, or citizen-science "human computation" methods can be used to identify where the handwriting of interest occurs.

Figure B.4: Printed text identifying variable of interest changes over time



Researchers can improve the accuracy of locating variables on the page by capturing images at a consistent size, and with a consistent layout. For loose leaf manuscript material, this may be achieved by scanning or photographing using a copy stand or scanner with edge guides to align paper in a fixed location. This is particularly important for paper that varies in size. Edge guides can also reduce the incidence of skewed images. Skewness of even 3° from the grid can significantly degrade the accuracy of both OCR and handwriting recognition.

Figure B.5. Skewed images can degrade the accuracy of handwriting recognition

DEPARTMENT OF HEALTH
DIVISION OF VITAL STATISTICS
CERTIFICATE OF DEATH

1 PLACE OF DEATH
County Columbiana Registration District No. 23 File No. 59996
Township Perry Primary Registration District No. 4343 Registered No. 190
or Village ..... No. ..... St., ... Ward
or City of Robert J. Adams (If death occured in a hospital or institution, give its NAME instead of street and number)
2 FULL NAME ..... Did Deceased Serve in U. S. Navy or Army .....
(a) Residence. No. Perry Twp St., ... Ward. (If nonresident give city or town and State)
(Usual place of abode) How long in U.S., if of foreign birth? yrs. mos. ds.
Length of residence in city or town where death occurred yrs. mos. ds.

PERSONAL AND STATISTICAL PARTICULARS

3 SEX: Male | 4 COLOR OR RACE: White | 5 Single, Married, Widowed or Divorced (write the word):
5a If married, widowed or divorced HUSBAND of (or) WIFE of
6 DATE OF BIRTH (month, day, and year) 9/10 1926
7 AGE: Years | Months: 1 | Days: 3 | If LESS than 1 day ..... hrs. or ..... min.
8 OCCUPATION OF DECEASED
(a) Trade, profession, or particular kind of work
(b) General nature of Industry, business, or establishment in which employed (or employer)
(c) Name of employer
9 BIRTHPLACE (city or town) Perry Twp. Col Co.
(State or country) Ohio
PARENTS
10 NAME OF FATHER Wells Adams
11 BIRTHPLACE OF FATHER (city or town) ..... (State or country) Ohio
12 MAIDEN NAME OF MOTHER Sabra Grace
13 BIRTHPLACE OF MOTHER (city or town) ..... (State or country) Ohio
14 Informant Mrs. W. Adams (Address) Salem R. D. 3
15 Filed 10-14-1926 J. J. Church REGISTRAR

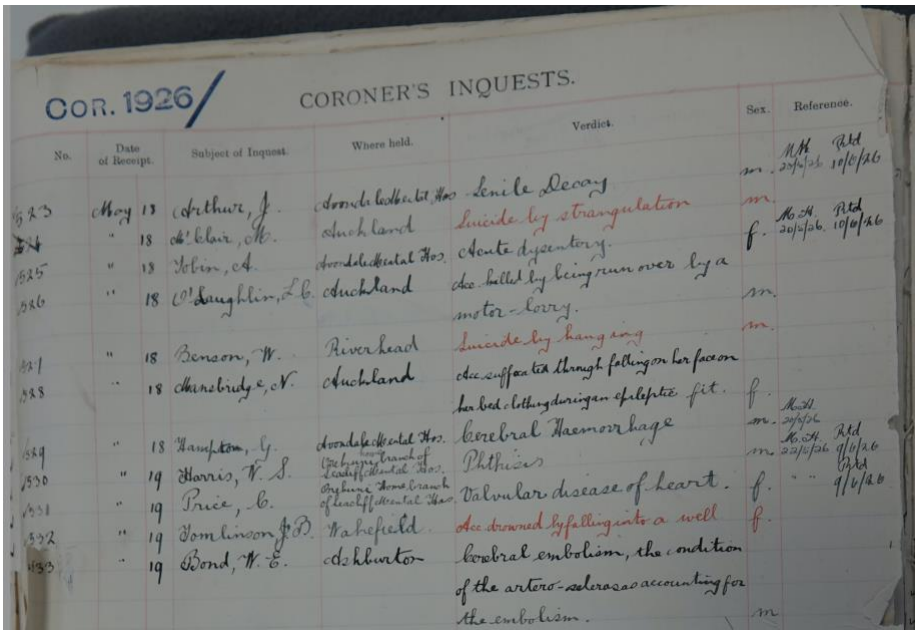
MEDICAL CERTIFICATE OF DEATH

16 DATE OF DEATH (month, day and year) 10/13 1926
17 I HEREBY CERTIFY, That I attended deceased from 10-8- 1926, to 10-13- 1926
that I last saw h... alive on 10-13- 1926
and that death occurred on the date stated above, at 7 a.m.
The CAUSE OF DEATH* was as follows:
Acute Enteritis
(duration) ..... yrs. ..... mos. 5 ds.
CONTRIBUTORY (SECONDARY) Pneumonia (Terminal)
(duration) ..... yrs. 1 mos. 3 ds.
18 Where was disease contracted if not at place of death? At place of death
Did an operation precede death? No Date of —
Was there an autopsy? No
What test confirmed diagnosis? Clinical Findings
(Signed) Chas A. Rood M. D.
10-14- 1926 (Address) Salem Ohio
*State the DISEASE CAUSING DEATH, or in deaths from VIOLENT CAUSES, state (1) MEANS AND NATURE OF INJURY, and (2) whether ACCIDENTAL, SUICIDAL, or HOMICIDAL. (See reverse side for additional space.)
19 PLACE OF BURIAL, CREMATION, OR REMOVAL Grand View | DATE OF BURIAL 10/14 1926
20 UNDERTAKER, License No. 64 ADDRESS
W. D. Arbaugh Salem

PHYSICIANS: "AGE" should be carefully supplied. EXACTLY. Exact statement of information should be stated. Exact statement should state CAUSE OF DEATH in plain terms, so that it may be properly classified. of OCCUPATION is very important. See instructions on back of certificate.

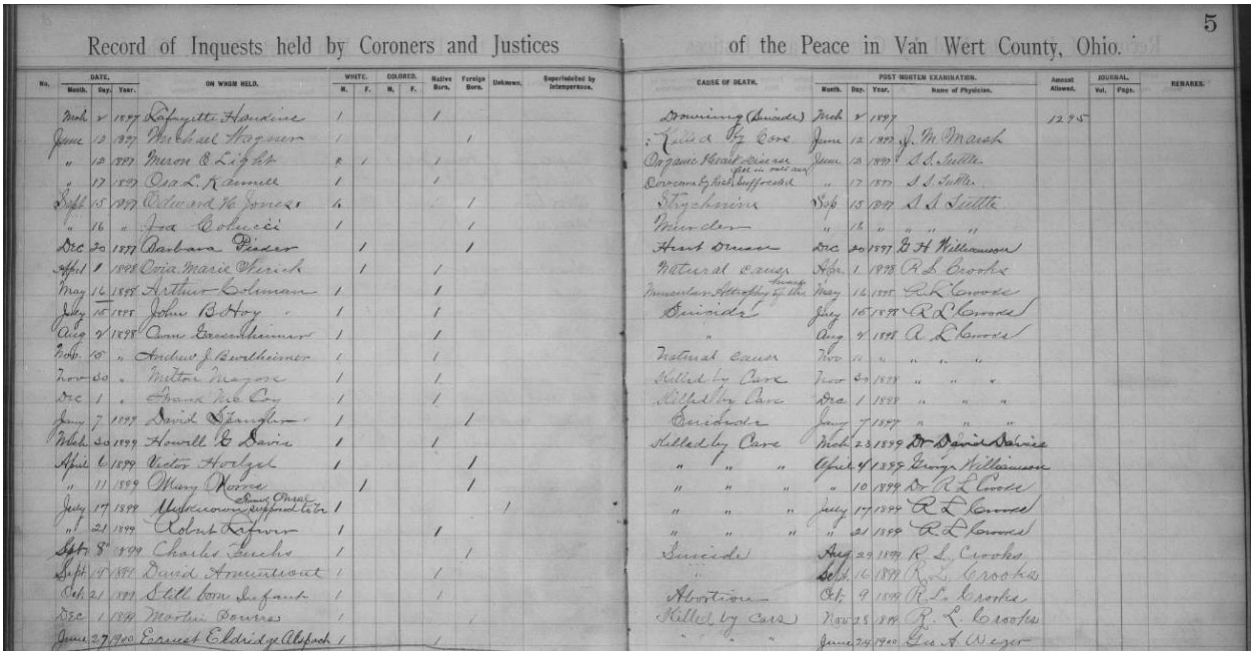
In general, loose-leaf material is more straightforward to image with consistent alignment and size. Bound volumes are commonly encountered in economic and social history research, as they were a convenient format for recording data systematically collected over time. Bound volumes, such as ledgers illustrated in the example below (Figure B.6), may need to be photographed on book cushions, or specially angled stands. Curvature of the image is common, particularly near page gutters. Imaging such documents to avoid skewness can be more difficult if the text is aligned to a grid. The example in Figure B.6 comes from a ledger of around 400 sheets of paper in a bound volume recording the conclusion of coroner's inquests. Causes of death are recorded in the column labeled "Verdict". Several features of this ledger would make data entry or handwriting recognition straightforward, including clear vertical lines denoting columns, adequate spacing between lines, and clear horizontal lines that prompted clerical staff to keep handwriting mostly aligned to the grid. Descenders (such as the tail of a g or y) can cause problems in this format, but this document would support the application of handwriting recognition software.

Figure B.6. Bound volumes with a spreadsheet style layout: New Zealand Coroner's Inquest Index



A slightly different format for a ledger containing cause of death information is illustrated in Figure B.7. In Figure B.6 the challenge is that individual data may span a varying number of lines, but all information related to one event is one side of the ledger. In Figure B.7 the information for each record spans the page gutter.

Figure B.7. Bound volume containing Causes of Death from Van Wert County, Ohio



Creating images of bound volumes that are only minimally skewed can be performed in capital or labor intensive ways. The capital intensive option is using “V-cradle scanners,” which cost around $100,000 at the time of writing. Thus, this option is only feasible in collaboration with an archive or library who already own the technology. Major universities are increasingly likely to have this technology for digitization of library books. These scanners can digitize an entire bound volume at high resolution with minimal skew in a couple of hours.

The labor intensive option is digital photography, but this may still be quite cost effective. High resolution and minimal skew can be achieved by manually adjusting the camera angle on each page and composing the picture. Because the volume is being photographed at an angle this technique can be physically tiring for the photographer. A tabletop tripod with a ball head mount, or an articulating arm mounted on a tripod, reduces

the physical labor involved for fixed capital costs closer to $1,000, and the variable costs of the research team's time. Use of a tripod or other stabilizer such as a copy stand allows photography with low "film speed" and remote shutter release, while obtaining adequate lighting, sharpness, and contrast to attempt machine handwriting recognition.

Readers previously unacquainted with photography beyond use of a phone camera may be deterred by the concise description of this work. They should not be. There is a competitive labor market for amateur or part-time professional photographers who can be engaged to provide advice on the specific requirements for photographing a collection of interest. A competitive market also exists for camera and lens rental, and universities may have loan equipment. Even using a hand held camera, it is possible for one person to photograph 1,000 – 2,000 pages per day from bound volumes. Ledgers, like the example illustrated above, may contain information on 20-50 individual events or individuals per page. Therefore, even in the worst case scenario of entirely hand-held photography it is possible to create images for tens of thousands of individual records in a week, for a cost still an order of magnitude less than the specialized scanner.

Bound volumes are more likely to contain information in a spreadsheet style format, with column headings identifying variables at the head of the page for multiple individuals. This format, illustrated in Figure B.6 or B.7 brings challenges slightly distinct from those we encountered with death certificates which were a single page per record. The example on the left side of Figure 1 shows some of the issues involved, even with well organized material. Each new record is identified by an index number in the left-most column. Typically one entry spans one line only, but several span multiple lines. Parsing this page into variables based on the columns delimited in red will efficiently split lines of text into variables. Subsequent lines of text that do not begin with an index number in the left column will need to be attached to the preceding record within the same variable. Researchers should examine their material to see whether additional lines within the same entity span columns, as this practice can be seen in some contexts. Finally, researchers will note the use of ditto marks to indicate the continuation of the same month. The general issue here is that "tidy data" that is stored efficiently and comprehensible to a computer is not entirely congruent with how humans worked in an analog world

(Campbell-Kelly et al. 2003). All of these issues can be resolved, but emphasize the need for researchers to examine their material for its idiosyncrasies and variations.

Organization of image collections: Researchers undertaking imaging themselves, or in collaboration with archives, will also need to organize a large volume of electronic images. Creating a structure congruent with the original archival organization of the material is a recommended starting point, so images can be easily connected with original manuscript sources. Archives are likely to have existing series, volume and file identifiers, which uniquely identify a source at the level of a folder or a single bound volume. The process of imaging creates a large number of new, unique individual files which exist as "child" records to the "parent" folder or bound volume. Scanners and digital cameras are likely to give file names that have no relation to the content of the image. Thus a critical step is renaming individual page images to both uniquely identify the image, and connect it to the parent records. As well as naming photographs of individual manuscript pages to identify them uniquely within the original archival arrangement, researchers should consider embedding text identifying the manuscript page on the bottom of the image. If files are ever renamed accidentally, this information helps preserve the information necessary to identify its provenance and logical location. File renaming and image processing can largely be done automatically, with batch operations on folders of images. Researchers can expect to spend 1-2 hours of active time organizing and processing images for every 40 hours of time photographing. These time costs can be minimized by organizing the photography or scanning process efficiently. For example, using separate, clearly identified memory cards for each bound volume will allow researchers to transfer the images for each volume into separate folders, and then process filenames and images accordingly. Scanner operators or photographers can also create spreadsheets of image names and their corresponding archival references during the imaging process. These spreadsheets can then be used to quickly generate renaming statements, and insert visible file identifiers into the image.