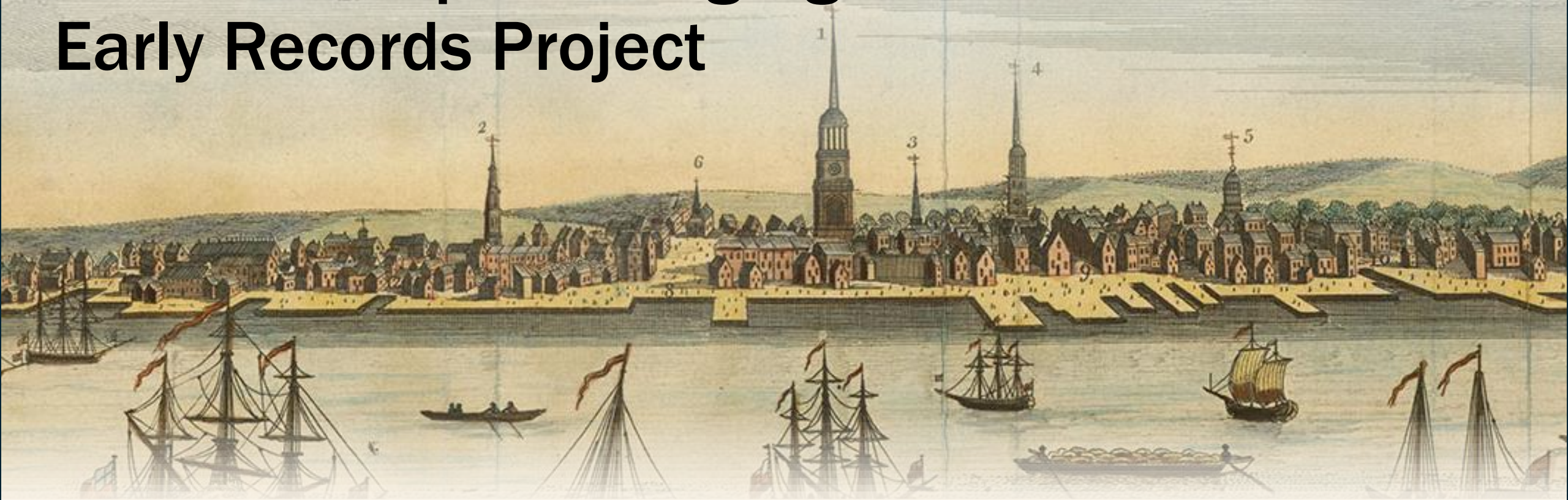


AI in the Archives:

The Philadelphia Congregations

Early Records Project



Walt Rice, Jr., Technology Consultant, R&R Computer Solutions, walt@r2infosys.com

Carol W. Smith, Independent Curator and Archivist, carolwsmith@msn.com

Philadelphia Congregations Project

- Gather the early records of Philadelphia's earliest congregations
- Directly digitized records from 11 congregations and the Episcopal Diocese (~46,000 pages) through the Athenaeum of Philadelphia
- Received contributions of scans from the Friends Historical Library, the Archdiocese of Philadelphia, and the Moravian Archives (~96,000 pages)



Uses of AI in the Archives



Transcription



Metadata enhancement / creation



Research

1. Transcription

- As part our initial project, we built an interface for volunteers to register and transcribe the records, page-by-page
- Over 6 years, we have about 19,000 human-transcribed pages
- COVID gave us a head-start!

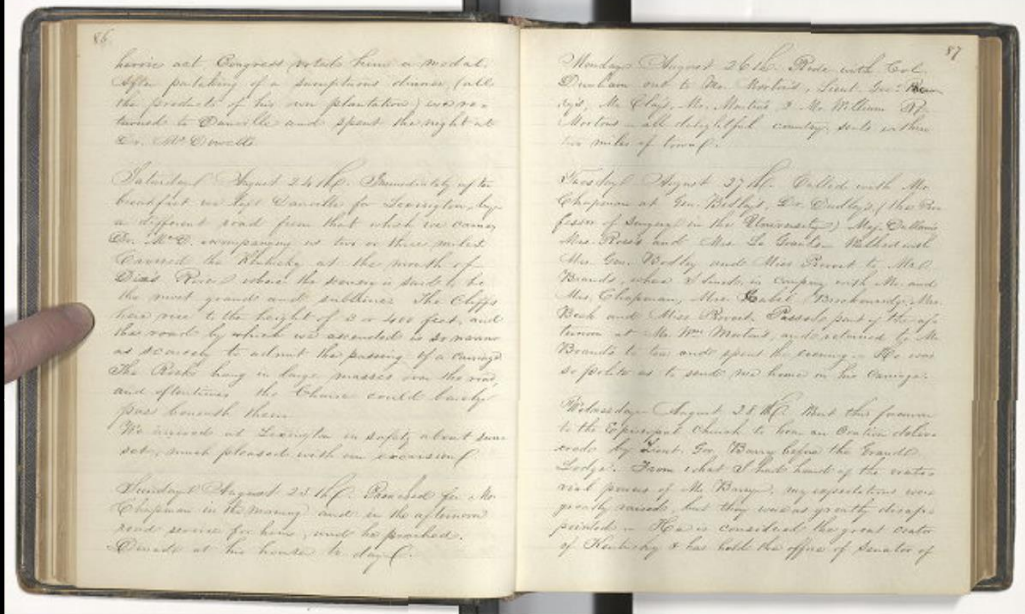
Go to page: Find Unfinished Transcription Page | Transcription Guidelines | Admin: Logout

Transcription Dashboard | Edit Work

Image 46 of 273 Go

CONTENTS

A Journal of Tours through the United States From August 26, 1819 to...



[Page 84]
when large mounds stood in the centre of what is now a hollow, and what is very remarkable, that they have been eaten away by the Buffalo, Elk, Deer, & other wild animals, which resorted here in immense numbers to eat the earth, for the salt with which it is highly impregnated. An old inhabitant of the place stated to the Dr. that he had seen 10,000 Buffalo at these Licks at one time. For a considerable distance around, the ground was trodden by them as hard as a turnpike road, and even now the roots of the trees are left bare by their continual tramping. Large cavities would sometimes be eaten into the sides of these mounds, and then the earth would fall in and bury numerous animals beneath it. I tasted the water which issues from a ledge of rocks, and found it to be

AI Transcription - Then

- ChatGPT 4 was the first useful model (Dec 2023)
 - Behind the scenes, Tesseract for OCR – horrible output
 - The language model was amazingly good at taking very poor OCR (from handwriting) and deciphering what the flow of the text must be
 - Not great at names and numbers
- Discovered that multiple steps produces much better output
 - Send images to Microsoft Azure Cognitive Vision first
 - Ask ChatGPT for an independent transcription, and then a comparison to produce the best text
 - Second round through ChatGPT for hyphenation

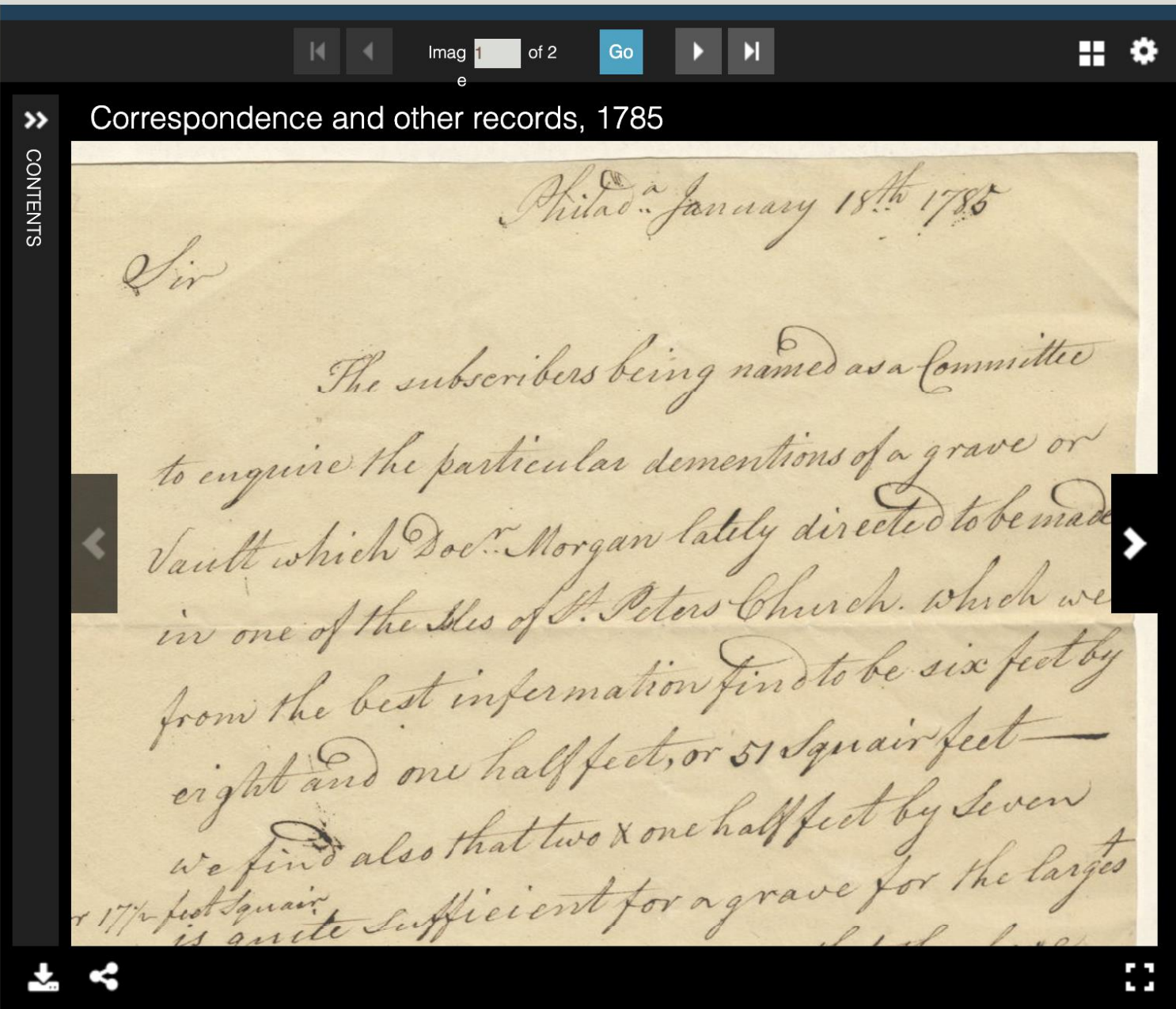
AI Transcription - Now

- Gemini 3 Pro “solves” transcribing handwritten text (Nov 2025)
 - Mark Humpries (Wilfrid Laurier University) quantified CER and WER as equivalent to the best human transcribers ([source](#))
 - We have NOT experienced that level of accuracy, but it is useful out-of-the-box for just about every purpose
 - Names and numbers now largely accurate
 - AI now also follows transcription rules: strikeouts, insertions, hyphens
 - Not perfect, but much more consistent
 - Cost about 7-10¢ per page
- Pre-reading with Azure Cognitive Vision STILL produces better output!

Risks!

- Hallucination is STILL a huge risk
- You can instruct the AI not to guess, but in reality, that's exactly what AI is always doing
- Names are most challenging, because language probability doesn't work the same
- But sometimes whole paragraphs or letters can be created (seemingly) out of nowhere
- As the AI gets better, the hallucinations get more convincing!

Correspondence and other records, 1785



← **Text**

[Image 1]

This page was automatically transcribed and has not been edited.

102
The Revd
Dr Wm White -

Sir

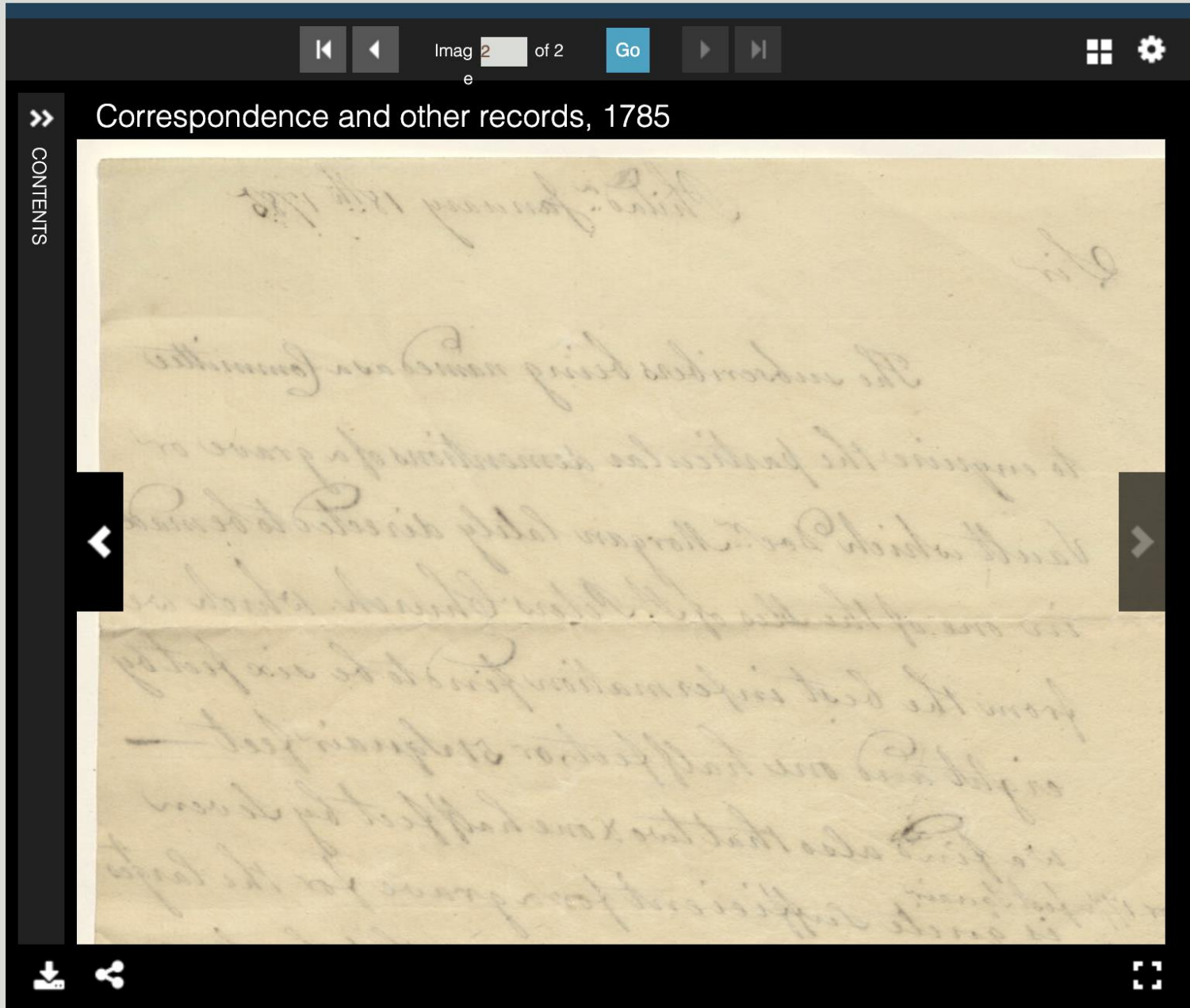
The subscribers being named as a Committee to enquire the particular dementions of a grave or Vault which Doctr Morgan lately directed to be made in one of the Iles of St Peters Church. which we from the best information find to be six feet by eight and one half feet, or 51 Squair feet - we find also that two & one half feet by Seven [* or 17 1/2 feet Squair] is quite sufficient for a gra person, from which it appears that the above grave or Vault occupies three times the space of a Common grave, and we are of opinion ought to be paid for in the proportion of three to one. we are very respectfully

Sir Your Humble Ser
Robt Towers
Frans Gurney

Doctr William White

[Help transcribe this resource](#)

Correspondence and other records, 1785



← **Text**

[Image 2]

This page was automatically transcribed and has not been edited.

Philada Jany 15 1828

The subscribers being named a Committee
to enquire the particular circumstances under
which Mr Chas J Henry holds the pew
No 50 in the Church – Report

That from the best information they can obtain
it does appear that Mr Henry
holds said pew in right of his Wife
who was the daughter of the late Mr
John Stille who was the owner of the
same at the time of his death &
under whose right the same is now held

Whereupon your Committee offer the
following Resolution

Resolved that Mr Chas J Henry be
confirmed in the possession of Pew No 50
in right of his Wife

Joseph [illegible]
Thos [illegible]

[Help transcribe this resource](#)

ENTIRELY FABRICATED!

Transcription Prompt, part 1

Your task is to accurately transcribe handwritten historical documents, minimizing the CER and WER. Work character by character, word by word, line by line, transcribing the text exactly as it appears on the page. To maintain the authenticity of the historical text, retain spelling errors, grammar, syntax, capitalization, and punctuation as well as line breaks (with the exception of the hyphenation rules).

Transcribe all the text on the page including headers, footers, marginalia, insertions, page numbers, etc. If insertions or marginalia are present, insert them where indicated by the author (as applicable).

Transcription Prompt, part 2

Also follow the Philadelphia Congregations Early Records transcription guidelines (source URL for humans: <https://philadelphiacongregations.org/home/transcription-guidelines/>).

Key rules excerpt:

- Preserve original line endings.
- Maintain spacing between paragraphs.
- Hyphenated line breaks within a page: remove the hyphen and join the word on the previous line; do not move whole lines.

Transcription Prompt, part 3

- Use tags when applicable:
 - Uncertainty (complete): [?]
 - Uncertainty (partial): [???omplete] (use ? for unknown letters)
 - Deletions/strikethrough: [DE: vanilla]
 - Marginal notations: [* ...]
 - Editor note: [ED: ...]
- Do not mark up superscripts or subscripts. Include the characters inline with the text. Do not include ^ or brackets for superscripts or subscripts.
- If you can provide a modernized search term for a clearly-identified proper noun (name/place) without altering the original text, you may add it as a tag after the original form: Ace Blinkin [Abraham Lincoln].

Transcription Prompt, part 4

BLANK / BLEED-THROUGH RULE (critical):

- Some scans are blank on the scanned side (recto) but show faint ink bleed-through / ghost text from the reverse side.
- If the scanned side has no intentional, readable text (only bleed-through/ghosting, stains, show-through, paper texture), output EXACTLY `[[BLANK_PAGE]]` and nothing else.
- Do NOT guess or reconstruct text from bleed-through. If the Azure OCR contains text that appears to be bleed-through, ignore it.

If a word or phrase is illegible, indicate it with `'[illegible]'`.

Transcription Prompt, part 5

Preserve line breaks that occur in the original document. Put spaces before indented lines to preserve indents in the original. The image may be either one page or a spread of pages. If it is two pages, transcribe each page separately. If you can't transcribe the image for some reason, return an error explaining why.

If the transcription is in a language other than English, please transcribe it in the original language, and after the transcription, provide a translation of the transcription in English. Do not include any other comments, just the transcription and the translation.

Transcribe the document from the image alone. You must maintain the line breaks of the original document. Output only the final best transcription, and do not include any other text in your response.

2. Metadata Enhancement or Creation

- Once we have reasonably good transcriptions, next step is to ask AI to summarize and describe the content
- All models particularly good at structured output, i.e., Dublin Core or other metadata schemas
- Not so good at authority lists for names and subjects
- Models improving on (1) following directions and (2) citing sources
- Responses are often *good*, but lack *consistency*
- Cost about 15-50¢ per request, depending on options

Two Methods





Direct Enhancement

- AI directly rewrites the metadata fields
- Abstract with citations, but other fields without
- Difficult to distinguish what has been (subtly) changed, and much harder to validate
- Feels fast, magic, and SCARY

Suggested Improvements

- AI provides bulleted suggestions for improvements
- All suggestions with cited pages in original
- Integration is the responsibility of the archivist
- Easier to validate each “fact” and incorporate

Direct Enhancement

Field	Current Metadata		Enhanced Metadata (Gemini 3 Pro)	Enhanced Metadata (GPT-5.2)	Enhanced Metadata (Claude Opus 4.5)
Name (Title)	<div>A Journal of Tours through the United States From August 26, 1819 to May 8, 1857 By Rev. Benjamin Dorr A. M.</div>		A Journal of Tours through the United States From August 26, 1819 to May 8, 1837 By Revd. Benjamin Dorr A.M.	Journal of tours through the United States and Canada, 1819-1864	A Journal of Tours through the United States From August 26, 1819 to May 8, 1857 By Rev. Benjamin Dorr A. M.
Creator	<div>Dorr, Benjamin, 1796-1869</div>				
Date	<div>1819/1864</div>				
Description	<div><p>Reverend Benjamin Dorr served as rector for Christ Church from 1839 to 1868, and undertook the difficult job of holding the congregation together during the Civil War.</p><p>Dorr traveled extensively throughout his life. This handwritten manuscript contains Dorr's notes about his many travels through the United States from 1819 to 1864. He often traveled through New England, but also visited the mid-Atlantic states, and even ventured into the deep South and as far West as the Green Bay area, stopping in Detroit and Chicago. He traveled by steamboat, stagecoach and horse; stayed with family and friends; dined with persons prominent in the military, church, politics and</p></div>		<p>Reverend Benjamin Dorr (1796-1869) served as rector for Christ Church, Philadelphia, from 1839 to 1869. This handwritten manuscript contains Dorr's detailed journals of his travels throughout the United States between 1819 and 1864.</p> <p>Dorr's travels took him through New England, the Mid-Atlantic, the South, and the Midwest. He records observations on landscapes, cities, and institutions, often preaching and collecting funds for domestic missions. Notable journeys include a 1822 tour to Ohio and Kentucky where he met Henry Clay; a 1834 visit to Washington D.C. meeting President Andrew Jackson and Martin Van Buren; a 1836 tour of the South including New Orleans; and a 1836 trip to</p>	<p>Reverend Benjamin Dorr served as rector for Christ Church from 1839 to 1868, and undertook the difficult job of holding the congregation together during the Civil War.</p> <p>Dorr traveled extensively throughout his life. This handwritten manuscript contains Dorr's notes about his many travels through the United States from 1819 to 1864, including a tour to Montreal in 1821. He often traveled through New England, but also visited the mid-Atlantic states, and even ventured into the deep South and as far West as the Green Bay area, stopping in Detroit and Chicago. He traveled by steamboat, stagecoach and horse; stayed with family and friends; dined with persons prominent in the</p>	<p>Reverend Benjamin Dorr served as rector for Christ Church from 1839 to 1868, and undertook the difficult job of holding the congregation together during the Civil War.</p> <p>Dorr traveled extensively throughout his life. This handwritten manuscript contains Dorr's notes about his many travels through the United States from 1819 to 1864. He often traveled through New England, but also visited the mid-Atlantic states, and even ventured into the deep South and as far West as the Green Bay area, stopping in Detroit and Chicago. He traveled by steamboat, stagecoach, railroad, and horse; stayed with family and friends; dined with persons prominent in the military, church, politics and</p>

Suggested Improvements

Field	Current Metadata	Enhanced Metadata (Gemini 3 Pro)	Enhanced Metadata (GPT-5.2)	Enhanced Metadata (Claude Opus 4.5)
Name (Title)	A Journal of Tours through the United States From August 26, 1819 to May 8, 1857 By Rev. Benjamin Dorr A. M.		<ul style="list-style-type: none">Revise to reflect the full chronological span and scope shown in the journal/index (tours extend beyond 1857 and include 1864; also includes Montreal/Canada) [p.1; p.528; p.i]Consider removing honorific/degree punctuation that is not consistently used in the manuscript title page (e.g., “A.M.” appears; “Rev. Dr.” appears elsewhere) and align with the title as written on the journal title page [p.1]	
Creator	Dorr, Benjamin, 1796-1869		<ul style="list-style-type: none">Consider adding an authorized form/qualifier for the creator’s clerical title as used in the journal (e.g., “Rev.” appears on title page) while keeping the same name string structure [p.1]	
Date	1819/1864		<ul style="list-style-type: none">Use EDTF interval for the journal’s coverage as evidenced by entries and index (e.g., 1819/1864) [p.3; p.528]If describing the added index pages separately, consider noting their creation as undated editorial additions rather than altering the main date span [p.1]	
Description	Reverend Benjamin Dorr served as rector for Christ Church from 1839 to 1868, and undertook the difficult job of holding the congregation together during the Civil War. Dorr traveled extensively throughout his life. This handwritten manuscript contains Dorr's notes about his many travels through the United	<ul style="list-style-type: none">Dorr records meetings with significant figures such as Henry Clay in Kentucky [p.73], Bishop Philander Chase in Ohio [p.59], and Henry Schoolcraft in Detroit and Mackinac [p.346].The journal includes detailed	<ul style="list-style-type: none">Add that the volume includes an editorial index page noted as not part of the original journal [p.1]Add that the journal includes a copied letter from George Washington (1779) transcribed by Dorr during the 1822 tour [p.21; p.22]	

Suggestion Prompt, part 1

You are a metadata assistant.

We attempt to use DACS for the title field. See: (link)

Other fields are generally consistent with Dublin Core / DCMI Terms. See: (link)

Provide suggestions for a human archivist on what to add, remove, or revise in the metadata, keeping the same field names.

Do not rewrite the full field values. For each field you suggest changes for, return a multi-line string that is only a bullet list, like:

- item one [p.3]
- item two [p.4]

Suggestion Prompt, part 2

If a field does not need a change, return an empty string. For empty fields, provide suggested improvements instead of filling the field.

The Subject field must use OCLC FAST subject terms; for Subject suggestions, list only the subject terms as bullets (no explanations), still with citations.

The Format (Original) field must use FAST genre/form terms, most likely one or more of: Manuscripts; Contracts; Account books; Registers (Lists); Correspondence; Records (Documents); Deeds.

The Date field should be formatted in EDTF format, see: [\(link\)](#)

Suggestion Prompt, part 3

If it makes sense, you may also suggest edits to the Description field, but do not remove specific details the human cataloger included; keeping or adding detail is preferred.

The Abstract field is meant to provide a longer summary of the work, and relevant context. The Abstract field should contain specific page citations from the transcript in square brackets (for example, [p.3]). If the transcript lacks page markers, use the image number (for example, [image 1]) for citations.




Return only a JSON object where keys match the input fields. Do not add new keys or commentary. Every bullet must include page citations from the transcript in square brackets (for example, [p.3]). If the transcript lacks page markers, use [page unknown] for citations.

3. Researching with AI

- With (good enough) transcriptions and decent metadata, the next step is to have AI help readers discover and understand the archival materials
- Familiar interface is a chat
- Two possible approaches

← Philadelphia Congregations Early Records

Ask the Archives



New Chat

The Christ Church bells were removed from Philadelphia during the Rev War? What information do we have about that?

Yes. In the Revolutionary War crisis of **September 1777**, there was an official push to **take down Philadelphia's "publick" bells** so the British could not seize and melt them for military use—and **Christ Church's bells were caught up in that effort**.

What the Christ Church records say (primary sources)

1) Congress and Pennsylvania ordered bells removed (Sept. 14, 1777)

Christ Church's vestry minutes preserve copies of the relevant orders:

- **Continental Congress (Sept. 14, 1777):** a resolve directing that steps be taken to "remove all publick Bells in Philadelphia to a place of security" as the enemy approached. ^[1]
- **Pennsylvania Supreme Executive Council (also Sept. 14, 1777):** ordered Col. Benjamin Flower to hire workmen to "take down the Bells of all the publick buildings in the City, and convey them to a place of Safety." ^[1]

2) Christ Church's clergy/vestry tried to stop or delay it (Sept. 16, 1777)

Ask a question...

Send

Full Content / Context

- For smaller sets of text, for instance a single minute book, AI can keep the whole text in the conversation
- Provides deeper understanding of the one work, but unable to make connections to others
- Practical limits are about 400 pages
 - Depends on the model and the content
 - Higher the number of pages, worse the output from the AI
- You can do this yourself by adding a text file to ChatGPT and asking away

Hybrid Search (RAG)

- RAG is **Retrieval-Augmented Generation**
- AI decides, based on the user's question, what to search for in a special database of chunked text (and can also search the web)
- Retrieved text is used by the AI as context to answer the user
- Great for *broad* coverage of material, but not necessarily for *deep* or *thorough* coverage
- Best compromise so far for end users
- Requires programming to implement
- Rough estimate for cost is about 10¢ per user question

Ask the Archives / Ask a Historian

- Live demo time!

Cost Summaries

- Because of the constant experimentation on our side, and the constant changes from providers, costs are hard to estimate
- Ballparks so far:
 - Transcription: 10¢ per page
 - Metadata: 15-50¢ per item
 - Research Questions: 10¢ per question
- For something like a book of 400 page book of minutes, that looks like \$40 to process.

Environmental Costs

- Google probably has the best ability to understand actual costs per query because of their scale and control of their own data centers
 - Elsworth, et al., “Measuring the Environmental Impact of Delivering AI at Google Scale.” <https://arxiv.org/abs/2508.15734>
- Using Google’s numbers, processing that 400 page minute book would use/produce about:
 - Energy: 0.36 kWh (4 hours of TV)
 - Water: 390 mL (0.39 L) (small water bottle)
 - CO₂e: 45 gCO₂e (1/10 mile in a car, or 45 minutes of you breathing)

Final Thoughts

- AI in transcription
 - Volunteers love it, speeds up work considerably
 - Does not replace humans for accurate transcriptions
- AI in metadata creation
 - Shines with abstracts, descriptions, and search terms
 - In an era of less processing rather than more, an incredible aid
 - Requires diligent checking
- AI in research
 - Reading the originals is critical
 - Not always easy to know how to phrase questions to elicit a full response
 - Lacking understanding of unspoken/unwritten stresses and context

Q&A