

Building a Responsible Historical Dataset

COLLEGE AND ADULT EDUCATION · HISTORY, DATA LITERACY, INFORMATION SCIENCE, JOURNALISM · GLOUCESTER, ROCKPORT, ESSEX, MANCHESTER-BY-THE-SEA

When old records name the same vessel or person more than once, or name different ones the same way, how do we decide what belongs together, and what do we owe the people in the records?

One 90-minute session (or two 75-minute sessions; see setup for the longer plan)

Cape Ann has kept lists of its people lost at sea for well over a century. A town history printed in 1895 lists Manchester's losses back to 1717, and the City of Gloucester keeps a register of thousands of names. Turning lists like these into a searchable dataset is useful for families, researchers and memorials, but it is easy to get wrong. Two schooners can share a name. Two men named Joseph Allen can be lost decades apart. A date can be a guess. In this session you will make the decisions a careful data editor makes, and write rules that respect the people in the records.

Primary source

R1. From Manchester's 1895 town history

Transcribed from a scan and checked against the page images; spelling as printed. Superscript note numbers are shown as [1] and [2].

Page 110 (chapter text):

"In 1764, Benjamin Andrews, Charles Leach and Daniel Foster were lost on the return voyage from the West Indies. In 1772, Capt. Daniel Edwards, Samuel Edwards, Benjamin Hill, Samuel Perry and Frank Silva were lost coming from the West Indies. In 1766, no less than ten were lost at sea. In September, 1843, the schooner "Vesper," of about sixty tons, owned by Jacob Cheever and his two sons, was lost with the whole crew, Capt. John Cheever, Rufus Cheever, Hilliard Morse, David Hall, Nathaniel Morgan and Merritt Lennon, all but the captain married men and fathers.[1] Nothing was ever known of their fate. When last spoken they had a full fare and were bound for home."

Note 1 on page 110: "Of those who were thus left widows, two, Mrs. Morse and Mrs. Lennon survive."

Pages 358 to 359 (introduction to the list):

"In order to secure as full a list as possible of those who have been lost at sea, a request for information was

published in Salem, Gloucester and Boston papers as well as in the Manchester Cricket. As the result of this inquiry the following list is appended, which is probably as complete as can now be made."

Page 359 (four lines from the list):

"1758. Joseph Allen, John Day, Simeon Wilson, John Driver, Richard Lee, Jun., John Dennis, Josiah Lee, John Sears, Samuel Morgan, Jun., Joseph Lee.[1]"

"1767. Benjamin Allen, Thomas Ayer, Benjamin Allen, Ezra Allen, John Leach, Richard Kitchin, Timothy Donnaway, Joseph Easty, John Marshall, Josiah Hassam."

"1768. Benjamin Andrews, Charles Leach, Daniel Foster, coming from West Indies."

"1772. Daniel Edwards, Samuel Edwards, Samuel Perry, Frank Silva, Benjamin Hilton, coming from the West Indies."

Note 1 on page 359: "There is some confusion in names this year."

Source: D. F. Lamson, History of the Town of Manchester, Essex County, Massachusetts, 1645-1895 (Town of Manchester, 1895), page 110 and Appendix L, pages 358 to 359. <https://archive.org/details/historyoftownofm00lams>. Rights: Public domain (published 1895)

Data table

R2. Linkage cards: 20 records from two sources

Each card is one row in one source. Dates are copied as the source records them, including brackets, slashes and question marks. Details are summarized in our words. Family details and remarks about other people that appear in some register entries are left out on purpose; Activity A3 asks why.

Card	Source and page	Name as recorded	Date as recorded	Vessel as recorded	Other details (our summary)
L01	Gloucester register, PDF p. 139	Costello, Martin	1870 [Sep]	Dauntless, sch.	Whole crew of 12 lost with the vessel on passage to the Bay of St. Lawrence, mackerel fishery
L02	Gloucester register, PDF p. 141	Craig, James/Jas. G.	1870 [Sep]	Dauntless, sch.	Master; same circumstances as L01
L03	Gloucester register, PDF p. 20	April, Frank	1900 Dec 10	Dauntless	Age 20, single, born in St. Malo, France; lost overboard on passage to the Bay of Islands, Newfoundland
L04	Gloucester register, PDF p. 291	Harding, Edward	1908 Dec 12	Dauntless	Born at the Bay of Islands, Newfoundland; lost in a dory at the Bay of Islands with two shipmates

Card	Source and page	Name as recorded	Date as recorded	Vessel as recorded	Other details (our summary)
L05	Gloucester register, PDF p. 20	Andrews, Theodore	1849 Oct 6	Levi Woodbury, sch	Age 21, married, born and lived in Essex; whole crew of 10 lost with the vessel near Boon Island, shore mackerel fishery
L06	Gloucester register, PDF p. 691	Simmons, James/Samuel	1891 Feb 15/19	Charles Levi Woodbury	Single, born in Barrington, Nova Scotia; washed overboard on passage from Newfoundland
L07	Gloucester register, PDF p. 612	Pinkham, Franklin D.	1876 Dec/Nov10	D. E. Woodbury, sch.	Master, single, born in Boothbay, Maine; whole crew of 10 lost; left Gloucester in late October and last seen in mid-December, as recorded
L08	Gloucester register, PDF p. 821	Woodbury, Lewis	1876 Oct 16	Joseph F. Allen, sch.	Single, born in Nova Scotia; whole crew of 10 lost with the vessel. The Fishermen's Own Book (1882) prints this crew member as 'Woodbury Lewis.'
L09	Gloucester register, PDF p. 551	Nickerson, Woodbury	1874 Mar	White Fawn	Single, born in Argyle, Nova Scotia; whole crew of 12 lost with the vessel on the Grand Banks
L10	Gloucester register, PDF p. 7	Allen, Joseph & crew	177-?	--	Master of a trading voyage home from Malaga; lost off Sable Island
L11	Lamson (1895), p. 359, MBTS-0011	Joseph Allen	1752	None given	Manchester list; no other details
L12	Lamson (1895), p. 359, MBTS-0027	Joseph Allen	1758	None given	Manchester list; the book's note says there is some confusion in names this year
L13	Gloucester register, PDF p. 391	Leach, Edward	1894 Apr	Martha and Susan, sch.	Married; born in Manchester, Mass.; lived in Manchester and Gloucester; whole crew of 13 lost with the vessel on Georges
L14	Lamson (1895), p. 361, MBTS-0130	Edward W. Leach	1894	None given	Manchester list; no other details
L15	Lamson (1895), p. 359, MBTS-0047	Benjamin Allen	1767	None given	First of two 'Benjamin Allen' names printed in the 1767 line
L16	Lamson (1895), p. 359, MBTS-0049	Benjamin Allen	1767	None given	Second 'Benjamin Allen' name in the same 1767 line
L17	Lamson (1895), p. 359, MBTS-0003	Benjamin Allen	1748	None given	Manchester list; no other details

Card	Source and page	Name as recorded	Date as recorded	Vessel as recorded	Other details (our summary)
L18	Gloucester register, PDF p. 339	Jacobson, John	1894 Mar/Apr	Ambrose H. Knight, sch.	Single, born in Norway; whole crew of 16 lost with the vessel on passage to Iceland
L19	Gloucester register, PDF p. 339	Jacobson, John	1896 Aug 26	William H. Rider	Age 60, single, birthplace recorded as Russia-Finland; died at sea off Nova Scotia, and the register records the illness that caused his death
L20	Lamson (1895), p. 360 (appendix) and p. 110 (chapter), MBTS-0120	John Cheever	1843 (appendix); September, 1843 (chapter)	Vesper (chapter only)	Captain; schooner lost with all six aboard; never heard from

Source: Facts from the City of Gloucester Archives Committee, Lost at Sea register (2010), based on the work of Roberta Sheedy, cited by PDF page; and from D. F. Lamson, History of the Town of Manchester (1895), Appendix L and page 110, cited by printed page and Cape Ann Current record ID. The Fishermen's Own Book (Procter Brothers, 1882) is cited for one spelling. <https://gloucester-ma.gov/74/Lost-at-Sea>. Rights: Facts from the Gloucester register (facts only) and Lamson (public domain), restated in our own words; table compiled by Cape Ann Current

Summary by Cape Ann Current

R3. How Cape Ann Current's maritime dataset handles these problems

One row per source entry. Every entry keeps its own record ID (MBTS-0001 to MBTS-0131 for Lamson's list; register rows are numbered in document order). Rows from different sources are linked, never overwritten.

Nothing is merged automatically. When the same name appears twice, as with the two Benjamin Allens of 1767 or the Edward Lee listed in both 1773 and 1774, both rows stay, with a note that identity review is required.

No invented names. Lamson's list includes a group of five unnamed men in 1773 and ten unnamed men in 1776. Each group is one row with a count of the people it represents, not a set of made-up names.

Two date fields. "Date as recorded" copies the source exactly. "Date precision" says what kind of date it is: year only, month only, a year range, month only with a year conflict, or an uncertain year. Of the 131 Lamson records, 120 have a year only.

Conflicts are written down, not settled quietly. Lamson's chapter and appendix give different years (1764 and 1768) for the same three men; the record keeps the appendix year and notes the conflict.

Verification status. Each Lamson row says it is a historical source whose individual entry has not been independently corroborated.

Facts only from the Gloucester register. The city's register is used as a research copy. Facts from it may appear in tables written in our own words, with a citation. Its pages are not reproduced and it is not mirrored column for column. Entries the register marks as not maritime, entries about children and entries after 1950 are left out of anything published.

Page links. Each Lamson row links to the page image so anyone can check the transcription.

Source: Written by Cape Ann Current from its own research files: the Lamson 1895 transcription (131 records, checked against the page images), the private research copy of the Gloucester register and the newsroom's rights review of maritime sources (checked Oct. 5, 2026). <https://archive.org/details/historyoftownofm00lams>. Rights: Text written by Cape Ann Current

Summary by Cape Ann Current

R4. Rights at a glance

Three rules underlie the table. Facts and names are not protected by copyright, though the way they are expressed can be (U.S. Copyright Office). Works of the U.S. government are not protected by copyright (17 U.S.C. 105); this does not apply to state or town works. As of 2026, U.S. works published in 1930 or earlier are in the public domain.

Source	Classification	May publish	Must not publish
City of Gloucester Lost at Sea register (2010)	Facts only	Per-person facts in our own words, cited	Copies of the PDF, a column-for-column mirror, the compilers' notes word for word
Lamson, History of Manchester (1895)	Public domain	Everything in the 1895 book	Nothing restricted; do not imply endorsement
Procter Brothers memorial books (1873, 1882)	Public domain	Full lists, transcriptions, page images	Additions in modern reprints
Library of Congress, Chronicling America newspapers (before 1931)	Public domain	Page images, text and facts	Later syndicated material without a check
U.S. Coast Guard history pages and NOAA	Public domain (federal)	Federal text, data and photos with credit	Items credited 'courtesy of' a third party; emblems; implied endorsement
Mass. Board of Underwater Archaeological Resources pages	Facts only	Facts in our own words; short quotations	Page text wholesale; its photos
Town of Rockport, Charles S. Haight page	Facts only	Facts in our own words, cited	Page text; photos

Source	Classification	May publish	Must not publish
Down to Sea website	Facts only	Individual facts, cited and cross-checked	Its text, images, lists or their arrangement
Manchester-by-the-Sea newspaper archive	Facts only	Facts from any issue; quotations from issues before 1931	Bulk scans or text; later text beyond fair use
Essex Shipbuilding Museum vessel list	Restricted	Individual facts cited in a story, with a link	Any dataset or table copied or adapted from the list
Cape Ann Museum digital collections	Restricted	Facts from item records, with links	Any image or bulk records without paid permission

Source: Summary of Cape Ann Current's rights review of maritime history sources, checked Oct. 5, 2026. A research summary, not legal advice. <https://www.copyright.gov/help/faq/faq-protect.html>. Rights: Text and table written by Cape Ann Current

A1. Linkage decision log

Work in pairs with R1 and R2. For each cluster, choose one decision: same (confirmed), probable, possible, different or cannot tell. Record your evidence and name the next source that could settle it. Enter your answers in the data sheet.

1. C1, cards L01 to L04: How many vessels named Dauntless do these records show, at minimum? Which cards, if any, describe the same vessel?
2. C2, cards L05 to L09: A search for the word Woodbury returns all five. Which are vessel names and which are people's names? Are any of the vessels the same?
3. C3, cards L08 (its vessel), L10, L11 and L12: Which records are about the person Joseph Allen and which about something else? Should any of the people be linked?
4. C4, cards L13 and L14: Are Edward Leach and Edward W. Leach the same man? How confident are you, and what would raise or lower your confidence?
5. C5, cards L15, L16 and L17: How should a dataset handle three Benjamin Allen entries, two of them in the same line?
6. C6, cards L18 and L19, plus the carpenter of the revenue cutter Gallatin, recorded as J. Jacobson, who died off Manchester on Jan. 6, 1892: How many different men are these?
7. C7, card L20 and R1: The Vesper's month comes from the chapter, not the list, and one crew member's first name is printed 'Hilliard' in the chapter and 'Hillard' in the list. How should the record show both?
8. Compare the chapter and the list in R1 for 1764 and 1768, and for 1772. Name every difference you find.

A2. Encoding uncertain dates

For each date as recorded, fill in the data sheet: the earliest possible date, the latest possible date, the precision (exact, month only, year only, range, decade, conflicting) and a short note. Never change the date as recorded.

1. 1849 Oct 6 (card L05)
2. 1870 [Sep] (cards L01 and L02)
3. 1891 Feb 15/19 (card L06)
4. 1876 Dec/ Nov10 (card L07)
5. 177-? (card L10)
6. 1749 or '50 (Lamson's list, page 359)
7. 1843 in the list, September, 1843 in the chapter (card L20)
8. 1764 in the chapter, 1768 in the list, for the same three men (R1)
9. An 'on this day' anniversary feature needs a month and day. Which of these records could honestly appear on a specific day, and how would you label the others?

A3. Rights, privacy and dignity: scenario cards

Use R3 and R4. For each card your table receives, decide what you would publish, what you would leave out and why. Name the rule you are applying.

1. Card 1. A register entry from 1849 names a man's parents and says his father and brother were also lost at sea. A genealogist asks you to include all family details in the public dataset.
2. Card 2. A compiler's note in one register entry describes members of an earlier crew with an unflattering word. Do you publish it?
3. Card 3. A register entry records the illness that caused a sailor's death aboard ship in 1896.
4. Card 4. A reader asks you to add her uncle, lost from a fishing boat in 1987, to the archive, with his photo.
5. Card 5. A volunteer offers to post the full Gloucester register PDF as a download on your site.
6. Card 6. You are profiling an Essex-built schooner, and the only source for its builder is the Essex Shipbuilding Museum's online vessel list.
7. Card 7. A website gives a sailing date that no other source you have found confirms.
8. Card 8. A colleague suggests using an AI tool to fill in the missing first names in Lamson's list.
9. Card 9. The 1895 town history names two widows who were still living. A descendant asks why their names appear in your dataset.

Talk about it

- Why is a false merge worse than a missed link in a memorial dataset? Is it always worse?
- The Gloucester register says it is 'unavoidably incomplete.' How should a dataset built from it describe what is missing?

- Who should be able to correct a historical dataset: families, historians, anyone? How would you check a correction?
- Does a person who died in 1767 have privacy? Does their family? What changes as the date gets closer to today?
- When two parts of the same book disagree, which part should win, and who decides?

Show what you learned

Prepare a short data handling memo for a community archive that wants to publish a Cape Ann lost at sea dataset.

Part 1 (10 points): Submit your completed linkage log for clusters C1 to C7, with a decision, a confidence level, the evidence and a next source for each. Part 2 (5 points): Propose a date scheme, with field names, that keeps the date as recorded and handles exact, month only, year only, ranged and conflicting dates; show three examples from the session. Part 3 (10 points): In 400 to 600 words, write publishing rules covering rights (with at least three sources classified), privacy and dignity (including living families and recent losses) and how corrections are received and shown.

Points: 25