

Building a Responsible Historical Dataset: teacher guide

Teacher guide. Keep separate from student materials.

Grade band	College and adult education
Subjects	History, Data Literacy, Information Science, Journalism
Towns	Gloucester, Rockport, Essex, Manchester-by-the-Sea
Duration	One 90-minute session (or two 75-minute sessions; see setup for the longer plan) (sequence totals 90 minutes)
Prerequisites	Comfort reading a table and sorting rows in a spreadsheet.; No prior training in history, statistics or programming is needed.
Essential question	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?
Version	1.0, 2026-10-05
Review status	Draft: awaiting classroom teacher review
Lesson ID	CAC - LES - 011

Objectives

Students will be able to:

- Participants will be able to explain record linkage and the difference between a false merge and a missed link, using examples from Cape Ann records.
- Participants will be able to classify at least seven related record clusters as same, probable, possible, different or cannot tell, and justify each decision with evidence and a named next source.
- Participants will be able to encode historical dates that are exact, month only, year only, ranged or conflicting without losing the date as recorded.
- Participants will be able to sort sources into public domain, facts only and restricted, and state what each allows a publisher to do.
- Participants will be able to write publishing rules that protect the dignity of people in the records and the privacy of their living families.

Massachusetts curriculum frameworks

No Massachusetts curriculum framework governs college courses. Massachusetts adult education programs align to the College and Career Readiness Standards for Adult Education, but specific codes are not cited here because the official document could not be checked. As a reference point only, the skills here extend the high school Digital

Literacy and Computer Science standard on analyzing a complex data set (9-12.CT.c.4) and the high school News/Media Literacy standards on verification; they are not college or adult standards.

Standards checked against the official Massachusetts DESE framework documents on Oct. 5, 2026.

Materials

- Participant handout with readings R1 to R4 and activities A1 to A3
- Data sheet (spreadsheet) with the linkage decision log and the date encoding table
- Laptops or printed copies of the data sheet; one per pair
- Optional: projector to show the Lamson page images on archive.org and the City of Gloucester Lost at Sea page

Setup

For the 90-minute plan, run the sequence as written. For two 75-minute sessions, use Session 1 for the opening, the briefing and an extended linkage exercise (add 20 minutes for pairs to enter their decisions in the spreadsheet and 15 minutes for a whole-group reconciliation of disagreements, plus 10 minutes to compare the Lamson chapter and appendix in R1). Use Session 2 for dates, rights and privacy, then add 25 minutes for teams to draft a one-page data dictionary and 10 minutes of peer review. Before class, read R3 and R4 closely and skim the City of Gloucester Lost at Sea page (<https://gloucester-ma.gov/74/Lost-at-Sea>), which says its list is incomplete. Do not print or project pages of the Gloucester register itself; the lesson uses only facts from it, in our own words.

Teaching sequence

10 minutes Opening: How many ships named Dauntless?

Teacher: Show records L01 to L04 from R2. Ask: "How many vessels are these? How many events? What would you need to know to be sure?" Take answers without confirming any. Introduce the session's question.

Students: Discuss in pairs and write a first guess with one reason.

15 minutes Briefing: linkage, merges and uncertainty

Teacher: Explain record linkage: deciding which records refer to the same real person, vessel or event. Contrast a false merge (two people become one, and one disappears from the record) with a missed link (one person appears twice). Explain why a memorial dataset should prefer missed links to false merges. Walk through R3, which describes how Cape Ann Current's own maritime dataset keeps records separate, records date precision and limits what it publishes.

Students: Read R3 and note two rules you would keep and one you would question.

30 minutes Hands-on linkage exercise

Teacher: Pairs work through Activity A1 using R1 and R2. Circulate and push for evidence: "What in the record supports that? What would change your mind?" At 25 minutes, ask two pairs that disagree on cluster C4 to explain

their reasoning.

Students: Complete the linkage decision log on the data sheet.

15 minutes Dates that are not dates

Teacher: Introduce Activity A2. Model the first row. Ask: "If a website shows an 'on this day' anniversary, which of these records could honestly appear on a specific day?"

Students: Encode each date in the table and answer the anniversary question.

12 minutes Rights, privacy and dignity

Teacher: Assign two scenario cards from Activity A3 to each table. Use R4 for rights. Keep the focus on the families of the people in the records.

Students: Decide what to publish, what to leave out and why; report one card to the room.

8 minutes Wrap up and assessment launch

Teacher: Collect one sentence from each table on the rule they think matters most. Explain the assessment.

Students: Write the rule and begin the assessment plan.

Background for the teacher

Record linkage is the task of deciding which records describe the same real entity. In memorial data the two errors are not equal: a false merge erases a person, while a missed link only shows one person twice. Cape Ann Current's maritime dataset therefore keeps one row per source entry, links rows rather than combining them, keeps the date as recorded beside a precision field and writes conflicts into a caveat field. The examples in R2 are real records. Facts from the City of Gloucester register are restated in our words and cited by PDF page, which the newsroom's rights review classifies as facts only; Lamson's 1895 history is public domain. In preparing this lesson, a check of Lamson's 1767 line against the page image found that the printed name is Timothy Donaway, while the newsroom's working transcription read Dounaway. That is a useful example of why every transcription links to its page image and why checks never end. The 1764 and 1768 conflict, Hill and Hilton, Ayres and Ayers (chapter and list for 1756), and Hilliard and Hillard all come from within Lamson's own book. The Gallatin's carpenter, J. Jacobson, appears in the 1892 Treasury letter and muster roll printed by Congress in 1900; he is not on Lamson's list and not in the Gloucester register.

Handling sensitive topics

Every record in this session is a person who died at sea, and some participants may have family in these records or have lost someone at sea recently. Say so at the start. Use the names respectfully, avoid jokes about errors in the records and do not speculate about how anyone died. Use 'lost at sea' or 'died' rather than 'victim' unless a source documents something more. If a participant recognizes a family name, invite them to share only if they wish, and point them to the City of Gloucester Archives Committee, through the city's Lost at Sea page, for corrections to the city's list.

Discussion prompts

- 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?

Formative checks

- After the opening, each pair states one reason the Dauntless records might not describe one vessel.
- During A1, check that every decision has evidence and a named next source.
- Exit sentence: the one publishing rule your table thinks matters most.

Assessment

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

Answer guidance

A1

- C1: At least two vessels. L01 and L02 describe one schooner lost with its whole crew in 1870, so the Dauntless of 1900 (L03) and 1908 (L04) cannot be that hull. L01 and L02 are the same event: same (confirmed) as a vessel and event link. Whether L03 and L04 are the same vessel cannot be told from the register; the vessel field differs only in format. Next source: the annual List of Merchant Vessels of the United States (official numbers and home ports) or Gloucester newspapers.
- C2: Vessels are Levi Woodbury (L05), Charles Levi Woodbury (L06) and D. E. Woodbury (L07). People are Lewis Woodbury or Woodbury Lewis (L08) and Woodbury Nickerson (L09), whose first name is Woodbury. The Levi Woodbury was lost with its whole crew in 1849, so the 1891 Charles Levi Woodbury

is a different vessel with a different name; D. E. Woodbury is different again. Lesson: a text search matches strings, not entities, and similar names are not evidence of identity.

- C3: L08's vessel, the schooner Joseph F. Allen, is not a person. L10, L11 and L12 are people named Joseph Allen. L11 (1752) and L12 (1758) are both on the Manchester list six years apart; one man cannot be lost twice, so either they are two men or one entry is an error, and Lamson warns the 1758 names are confused. Keep both, flag them. L10 is a Gloucester register entry for a trading master lost in the 1770s; nothing ties him to Manchester and the decade does not fit L11 or L12. Decision: different, or cannot tell, never same.
- C4: Probable. Same name apart from a middle initial, same year and the register records a Manchester birth and residence, which fits a Manchester list. Not confirmed: Lamson gives no vessel or month, and the middle initial is not in the register. Raise confidence with a Manchester death record, an 1894 Manchester Cricket notice or a crew list for the Martha and Susan. Link the rows; do not combine them.
- C5: Keep all three rows. L15 and L16 may be two men of the same name or a duplicate in printing; the dataset cannot tell and marks identity review required. L17 (1748) is a different loss year, so not the same man as either 1767 entry unless one date is wrong.
- C6: Three different men. L18 died with the whole crew of the Ambrose H. Knight in 1894; L19 died at sea from the William H. Rider in 1896 at 60; the Gallatin's carpenter died on Jan. 6, 1892, and appears in neither list. Common names are the classic false merge risk.
- C7: Keep the list's year in one field and the chapter's month and vessel as a sourced addition, with both spellings recorded (Hilliard in the chapter, Hillard in the list) and the chosen display spelling explained. Cape Ann Current's record does this and notes the source of the month.
- Chapter versus list: 1764 versus 1768 for Andrews, Leach and Foster; Benjamin Hill versus Benjamin Hilton in 1772; the chapter calls Daniel Edwards Capt.; the chapter says ten were lost in 1766 while the list has ten names under 1767.

A2

- 1849 Oct 6: earliest and latest Oct. 6, 1849; precision exact (as recorded).
- 1870 [Sep]: September 1870; month only, and the brackets show the month was supplied by the compiler, so mark it inferred.
- 1891 Feb 15/19: Feb. 15 to 19, 1891; range or conflicting days; no single anniversary day.
- 1876 Dec/ Nov10: Nov. 10 to December 1876; conflicting; the register itself notes the vessel left Gloucester in late October and was last seen in mid-December.
- 177-?: 1770 to 1779; decade, uncertain.
- 1749 or '50: 1749 to 1750; year range.
- Vesper: year 1843 from the list, month September from the chapter; month only, with the source of each part noted.
- 1764 versus 1768: conflicting years within one source; keep both, display the conflict.

- Only exact day dates such as Oct. 6, 1849 can honestly appear on a specific day. Others can appear in a 'this month' or 'this year' list with their precision shown, or not at all.

A3

- Card 1: Family relationships from 1849 are historical public record and matter to genealogy, but publish only what serves the remembrance, in plain words, cited. Many archives publish names, dates and vessel and leave detailed family notes to the source.
- Card 2: Leave it out. It is a judgment about other, unnamed people, not a fact about the man listed, and the compilers' notes are not to be copied word for word.
- Card 3: A historical cause of death is usually acceptable, but ask if it adds anything; 'died aboard' is often enough.
- Card 4: The 1987 loss falls outside the dataset's 1950 limit and involves living family. Thank her, explain the rule and route the request to an editor; any story would need family consent and a cleared photo.
- Card 5: No. The register is facts only for reuse; link to the city's own copy instead.
- Card 6: The museum list is restricted. An individual fact may be cited in a story with a link, but nothing from the list goes into the dataset, and a second source should be sought.
- Card 7: Facts only; cite it, label it single source and keep looking for corroboration.
- Card 8: Never. A dataset of the dead must not invent names. Missing names stay missing, with a count of people represented.
- Card 9: The names are printed in a public domain 1895 book, but they add nothing to the record of the men lost. A good answer weighs this and may choose to keep them out of the dataset while leaving them in the cited source.

Rubric

Criterion	4	3	2	1
Linkage decisions	All seven clusters decided with accurate evidence, appropriate confidence and a specific next source; no false merges.	Six or seven clusters decided with evidence; at most one weak next source; no false merges.	Several decisions lack evidence, or one false merge.	Decisions are mostly unsupported or include more than one false merge.
Date scheme	Keeps the date as recorded and handles all five kinds of date with clear field names and correct examples.	Handles at least four kinds of date with correct examples.	Handles some kinds of date but loses the date as recorded or misclassifies one example.	No workable scheme.

Criterion	4	3	2	1
Rights	Classifies three or more sources correctly and states exactly what each allows and forbids.	Classifies three sources correctly with minor gaps.	One or more misclassifications.	Rights are not addressed.
Privacy, dignity and corrections	Clear, specific rules for living families, recent losses, sensitive details and a visible correction process.	Rules cover most of these areas with some specifics.	Rules are general or miss living families or corrections.	No rules or rules that would harm families.

Differentiation

Multilingual learners

- Provide a glossary: record linkage, merge, entity, precision, public domain, facts only, corroborate.
- Let pairs discuss in any shared language and write decisions in English using the five fixed labels.
- Explain the abbreviations in R2 (sch. for schooner, b. for born) before the exercise.

Students with disabilities

- Offer the data sheet in an accessible spreadsheet with clear column headers and no merged cells.
- Provide R2 in large print or as a screen-reader-friendly table; never rely on color to mark clusters.
- Allow the memo to be submitted as a recorded explanation with the linkage log.

Advanced learners

- Write matching rules in plain language or pseudocode for C1 to C6 and list the cases your rules would still get wrong.
- Compare Procter Brothers' Fishermen's Own Book (1882) with the Gloucester register for the Joseph F. Allen crew and log every difference.
- Design a corrections form for the public and the review steps behind it.

Print, low-bandwidth and extension options

Print: Print the handout and the data sheet. The whole session runs on paper; pairs can mark decisions on the printed log.

Low bandwidth: No internet is needed. Download the Lamson page images in advance if you want to show them.

Extensions:

- Use the full Lamson list of 131 records to count how many dates are year only, month only and ranged, and explain what that means for any chart made from the list.

- Find the Joseph F. Allen crew in the 1882 Fishermen's Own Book on archive.org and decide how to record 'Woodbury Lewis' versus 'Woodbury, Lewis.'
- Draft a one-page data dictionary for a community memorial dataset.

Google Classroom: ready to paste

Assignment title: Building a Responsible Historical Dataset: Data Handling Memo

Student instructions:

Open your copy of the handout and the data sheet. During the session, complete the linkage log and the date table with your partner and discuss the scenario cards with your table. Then write the data handling memo described in the assessment section and attach your completed data sheet.

Suggested points: 25

Attach:

- **student-handout.docx:** Make a copy for each student
- **slides.pptx:** Students can view file
- **data-sheet.xlsx:** Make a copy for each student

Never attach the teacher guide to student work. See [how to use this lesson in Google Classroom](#).

Citations and rights

- D. F. Lamson, History of the Town of Manchester, Essex County, Massachusetts, 1645-1895 (1895), pages 110 and 358 to 361. <https://archive.org/details/historyoftownofm00lams>. Public domain (published 1895)
- City of Gloucester Archives Committee, Lost at Sea register (2010), based on the work of Roberta Sheedy. <https://gloucester-ma.gov/74/Lost-at-Sea>. Facts only; facts restated in our words and cited by page; no pages reproduced
- Procter Brothers, The Fishermen's Own Book (Gloucester, 1882). <https://archive.org/details/fishermensownboo00procrich>. Public domain (published 1882)
- U.S. House of Representatives, Report No. 1830, 56th Congress (1900), muster roll and Treasury letter on the revenue cutter Gallatin. https://www.govinfo.gov/app/details/SERIALSET-04027_00_00-075-1830-0000. Public domain (U.S. federal government work)
- U.S. Copyright Office, What Does Copyright Protect? (FAQ). <https://www.copyright.gov/help/faq/faq-protect.html>. U.S. federal government work

Lesson text by Cape Ann Current, free to use and adapt under our [Free Access and Attribution Policy](#). Credit: "Source: Cape Ann Current." with a link to <https://capeanncurrent.com/classroom/building-a-responsible-historical-dataset/>. Public domain excerpts remain public domain; other sources keep their own terms, listed above. Version 1.0, 2026-10-05. Draft: awaiting classroom teacher review.