Early American Cookbooks: Creating and Analyzing a Digital Collection Using the HathiTrust Research Center Portal

Gioia Stevens
The Graduate Center, City University of New York

Recommended Citation
Stevens, Gioia, "Early American Cookbooks: Creating and Analyzing a Digital Collection Using the HathiTrust Research Center Portal" (2017). CUNY Academic Works.
https://academicworks.cuny.edu/gc_etds/1840

How does access to this work benefit you? Let us know!
Follow this and additional works at: https://academicworks.cuny.edu/gc_etds
Part of the American Studies Commons, Cataloging and Metadata Commons, and the Digital Humanities Commons

This Capstone Project is brought to you by CUNY Academic Works. It has been accepted for inclusion in All Dissertations, Theses, and Capstone Projects by an authorized administrator of CUNY Academic Works. For more information, please contact deposit@gc.cuny.edu.
EARLY AMERICAN COOKBOOKS:
CREATING AND ANALYZING A DIGITAL COLLECTION USING THE
HATHitrust RESEARCH CENTER PORTAL

by
GIOIA STEVENS

A capstone project submitted to the Graduate Faculty in Liberal Studies in partial
fulfillment of the requirements for the degree of Master of Arts,
The City University of New York

2017
Early American Cookbooks:
Creating and Analyzing a Digital Collection Using the HathiTrust Research Center Portal

by

Gioia Stevens

This manuscript has been read and accepted for the Graduate Faculty in Liberal Studies in satisfaction of the capstone requirement for the degree of Master of Arts.

Date

Matthew K. Gold

Capstone Advisor

Date

Elizabeth Macaulay-Lewis

Acting Executive Officer

THE CITY UNIVERSITY OF NEW YORK
ABSTRACT

Early American Cookbooks:
Creating and Analyzing a Digital Collection Using the HathiTrust Research Center Portal

by

Gioia Stevens

Advisor: Matthew K. Gold

The Early American Cookbooks project is a carefully curated online collection of 1450 cookbooks published in the United States between 1800 and 1920. The purposes of the project are to create a freely available, searchable online collection of early American cookbooks, to offer an overview of the scope and contents of the collection, and to use digital humanities tools to explore trends and patterns in the metadata and the full text of the collection. The project has two basic components: a collection of 1450 full-text titles on HathiTrust and a website site to present a guide to the collection and the results of the digital humanities analysis.

Early American Cookbooks collection
URL: https://babel.hathitrust.org/cgi/mb?a=listis&c=1934413200

Early American Cookbooks website
URL: https://wp.nyu.edu/early_american_cookbooks/
# TABLE OF CONTENTS

Abstract
Introduction
Environmental Scan
Digital Humanities Tools and New Approaches to Library Metadata
Project Development
Work Plan
Evaluation
Challenges and Solutions
Best Practices
Relationship to Track and Previous Course of Study
Continuation of Project
Appendix A
Appendix B
Appendix C
Appendix D
Appendix E
Appendix F
Bibliography
Introduction

The history of cookbooks in the United States can tell us not only about food preparation and consumption, but also about the history of American culture. Reading the recipes people used every day shows us how they lived their lives. Most cookbooks were written by and for women, so they are especially important for exploring women’s history. Cookbooks focus on women’s skills in the kitchen, but they also reveal how they saw themselves, their families and their communities. A large collection of cookbooks spanning a wide historical period can illustrate changing attitudes toward diet and nutrition, new ingredients and styles of cooking, homemaking, and the commercialization of food production.

The idea for this project came to me after cataloging hundreds of early print cookbooks for the Marion Nestle Food Studies Collection at the Fales Library & Special Collections at New York University Libraries. Most early cookbooks in printed form are accessible only in special collections libraries or private collections. These books are an incredibly important resource for food historians or anyone with a passion for old cookbooks, but there are very few full text resources available online. Digitization of library collections has made high-resolution, searchable scans available to the public. These scans are available as individual books on Google Books, the Internet Archive, HathiTrust and a variety of websites, but there has been no central online resource to unite them as a searchable, curated collection.
The Early American Cookbooks project was designed to meet this need. The purposes of the project are to create a freely available, searchable online collection of early American cookbooks, to offer an overview of the scope and contents of the collection, and to use digital humanities tools to explore trends and patterns in the metadata and the full text of the collection. The project has two basic components: a collection of 1450 full-text titles on HathiTrust and a website site to present a guide to the collection and the results of the digital humanities analysis.

**Early American Cookbooks Collection**

URL: [https://babel.hathitrust.org/cgi/mb?a=listis&c=1934413200](https://babel.hathitrust.org/cgi/mb?a=listis&c=1934413200)

The Early American Cookbooks collection is a separate, permanent, public collection within the HathiTrust Digital Library that was created specifically for this project. The HathiTrust Digital Library is a partnership of academic & research institutions, offering a collection of millions of titles digitized from libraries around the world. Within the Early American Cookbooks collection, users can browse the titles, read them online, and search the contents of each book. The collection as a whole can be searched independently of the rest of the digital library. Keyword searching across all 1450 titles allows the user to find particular recipes, ingredient names, or anything else and go immediately to the results in the full text. This type of search would be very valuable for a food historian tracing something such as the history of the hamburger or when Americans first started eating spaghetti.

**Early American Cookbooks website**
This WordPress site, also entitled Early American Cookbooks, serves as a gateway and guide to the HathiTrust collection. The site gives a general introduction to the history of Early American Cookbooks, an overview of the scope and contents of the collection, a discussion of certain interesting books or groups of titles within the collection, and links to other online resources and library collections specializing in early American cookbooks. The site also explores the metadata and the full text of the collection by presenting and interpreting the results of analysis using digital humanities tools. Text analysis and data visualization show trends and patterns in the collection that may add to our understanding of early cookbooks and the history of American food.

Environmental Scan

There are several online resources for early cookbooks, but there is no central online resource to unite them as a searchable, curated collection. An overview of the available resources includes the following:

The Feeding America Digital Archive (Michigan State University Libraries)

http://digital.lib.msu.edu/projects/cookbooks/

This is an important resource for early cookbooks and it contains outstanding interpretive articles. However, the digital archive contains only seventy-six full-text titles. The site is now over ten years old and the search interface is limited and cumbersome.
American Foodways: Jewish Contribution (University of Michigan Libraries)
https://www.lib.umich.edu/online-exhibits/exhibits/show/jewishfoodways

This digital exhibit presents the history Jewish cookbooks with content drawn from the Janice Bluestein Longone Culinary Archive at the University of Michigan. It does not offer full text of the titles and it has no search features.

British Library Books for Cooks
http://www.bl.uk/learning/langlit/texts/cook/cookery2.html

This “Texts in Context” section of the British Library website includes extracts from seventeen texts, ranging from Medieval recipe manuscripts to 1940s cookbooks. It does not include any full text titles and has no search features.

Other sites are essentially blogs with selected links to individual full-text titles in Google Books or the Internet Archive, but no ability to search over multiple titles. These include:

Savoring the Past

Vintage Cookbooks
http://vintagecookbooks.healthyeatingandlifestyle.org/

Culinary Lore

Historic Cookbooks Online
Digital Humanities Tools and New Approaches to Library Metadata

The Early American Cookbooks project is a demonstration of how digital humanities tools and new approaches to library metadata can improve user access to special collections materials. The presence of individual scans from multiple libraries in the HathiTrust Digital Library is a wonderful resource for users who are unable to visit special collections libraries in person. However, the material in HathiTrust is not collected or curated by subject area and a researcher must sift through millions of records to find pertinent items. Search results frequently contain duplicate records from different libraries (each with its own idiosyncratic catalog record) and many records offer multiple scans from different libraries. It can be quite complex to compare many similar records and determine which are duplicates and which represent different editions or printings of a book.

Creating the Early American Cookbooks collection on HathiTrust involved considerable labor in searching, evaluating and de-duplicating the record set (see Work Plan section below). Eleanor Lowenstein’s Bibliography of American Cookery Books, 1742-1860 and the interpretive essays in The Feeding America Digital Archive were especially helpful in evaluating the results. Other useful bibliographies include Gastronomic Bibliography by K.G. Bitting, Culinary Americana by Eleanor Parker Brown and Bob Brown, American Books on Food and Drink by William R. Cagle, Lisa Killion Stafford,
and Cecily Brownstone, and American Cookbooks and Wine Books, 1797-1950 by Janice Bluestein Longone et al. Evaluating the initial search results and distinguishing and de-duping multiple printings, editions, and scans reduced the collection from over 2000 records to a final set of 1450. The resulting curated and cleaned-up Early American Cookbooks collection is an example of how librarians with subject area knowledge and metadata expertise can make a broader public contribution beyond the walls of their own individual libraries.

Creating a collection on HathiTrust also offers opportunities to explore trends and patterns in the collection using the tools available in the HathiTrust Research Center (HTRC). The HTRC provides research access to the public domain corpus of the HathiTrust Digital Library. The HTRC is a collaborative research center launched jointly by Indiana University and the University of Illinois, along with the HathiTrust Digital Library. The HTRC provides an infrastructure to search, collect, analyze, and visualize the full text of nearly 3 million public domain works and is intended for nonprofit and educational researchers (“HTRC Portal - Home” 2016).

The HTRC Portal https://analytics.hathitrust.org/ offers an efficient way to analyze the metadata to explore trends and patterns in a collection. The Marc Downloader tool (see Work Plan section below) provides catalog records for all of the titles in the collection in XML format. This data allows a researcher to assemble the records as a set and look for new ways to maximize the value of the information. For many 19th century books, such as the ones in the Early American Cookbooks collection, these records are idiosyncratic,
legacy data, much of it drawn from old catalog cards. These records contain information such as “Brooklyn, opposite New York” as the place of publication as well as many different forms of abbreviation, terminology, and cataloging conventions. When this legacy data is cleaned up and presented as a dataset rather than as individual records, it can yield new information about a collection in aggregate. Metadata about authors, publishers, places of publication, dates, and subject content can offer insights into a collection and its significance within a broader historical scope. This information can be visualized and presented to users as a new way to gain understanding of the depth and breadth of a collection.

The HTRC also offers an array of text analysis tools to explore trends and patterns in the full text of a collection. These algorithms include a range of functions (Naïve Bayes classification, OpenNLP dates, OpenNLP entities, topic modeling, Dunning Log-likelihood comparison, spellchecker to detect OCR errors, tag cloud, and word count) that are built into the HTRC portal as ready-made tools. The user can build a set of HathiTrust records as a workset within the portal and then run these algorithms on that workset. The portal offers an interface for both assembling and analyzing a corpus of digital texts that is especially valuable for users who are new to computational methods. The HTRC algorithms that proved most fruitful for the Early American Cookbooks project were the Meandre Topic Modeling algorithm and the Meandre Dunning LogLikelihood to Tagcloud algorithm.
The Meandre Topic Modeling algorithm was created by Loretta Auvil. It serves to identify “topics” in a workset based on words that have a high probability of occurring close together in the text. Topics are models trained on co-occurring text using Latent Dirichlet Allocation (LDA), where each topic is treated as a generative model and volumes are assigned a probability of how likely each topic is to have generated that text. The most likely words for a topic are displayed as a word cloud. The algorithm loads each page of each volume from HTRC, removes the first and last line of each page, joins hyphenated words that occur at the end of the line, removes all tokens that do not consist of alphanumerical characters, filters stop words, replaces "not " with "not_" to deal with negations, creates a topic model using Mallet and displays the top 200 tokens in a tag cloud as well as exporting the topics in an XML file (“Description of the HTRC Portal Algorithms” 2016). This tool has been very useful in analyzing the subject content of the entire corpus of the Early American Cookbooks collection as also a subset of government publications within the collection.

As Megan R. Brett explains in *Topic Modeling: A Basic Introduction*, topic modeling is a form of text mining, a way of identifying patterns in a corpus. “You take your corpus and run it through a tool which groups words across the corpus into ‘topics’” (Brett 2013). Miriam Posner has described topic modeling as “a method for finding and tracing clusters of words (called “topics” in shorthand) in large bodies of text” (Posner 2012).

Topic modeling is an automated text mining technique that offers a “suite of algorithms to discover hidden thematic structure in large collections of texts” (Blei 2013, 7). Topic modeling is a methodology developed in computer science, machine learning, and natural
language processing that has recently become very popular in the digital humanities (Meeks 2013). New digital tools such as MALLET (McCallum 2002) generate comprehensive lists of subjects through statistical analysis of word occurrences in a corpus. The content of the documents, not a human indexer, determines the topics (Jockers 2013, 124). Unlike traditional classification systems with a pre-existing taxonomy of terms, topic modeling creates topics by clustering words that frequently occur together in a text. The resulting topical clusters can be readily interpreted as subject facets by human readers, allowing them to browse the topics of a collection quickly and find relevant material using topically expanded keyword searches (Mimno and McCallum 2007). Scholars such as Jeffrey Binder have recently critiqued topic modeling as a method for the study of literary and cultural texts because it is “overly dependent on the “literal” definitions of words and, at worst, complicit in the marginalization of nonstandard…modes of expression” (Binder 2016). These critiques raise some important concerns, but topic modeling does provide useful results for a corpus of cookbooks that contain many standard terms and very little figurative language.

The Dunning Log-likelihood statistic was developed by Ted Dunning at the University of New Mexico. It employs statistical a measure based on likelihood ratios that can be applied to the analysis of text (Dunning 1993). The statistic has been employed by digital humanities researchers as a way to compare corpuses of text and discover “subtle differences between closely related sets” (Schmidt 2011). The Meandre Dunning LogLikelihood to Tagcloud algorithm, also created by Loretta Auivil, compares and contrasts two worksets by identifying the words that are more and less common in one
workset than in another workset. The algorithm calculates Dunning Log-likelihood based on two worksets provided as inputs: an “analysis workset” and a “reference workset.” It loads each page of each workset, removes the first and last line of each page, joins hyphenated words that occur at the end of the line, performs part of speech tagging, lowercases the tokens remaining, counts the tokens remaining for all volumes for each collection, and compares counts from each collection using the Dunning Log-likelihood statistic. The “overused” tokens and the “underused tokens” (200 by default) in the analysis collection (relative to the reference collection made available via a csv file (“Description of the HTRC Portal Algorithms” 2016). This tool has been very useful in analyzing how different subsets of the Early American Cookbooks collection (government publications, vegetarian cookbooks, Fannie Farmer’s cookbooks, and the different census regions of the United States) differ from the collection as a whole.

**Project Development**

The first stage of developing the Early American Cookbooks project involved basic planning and decision-making. Research on available digital resources for early cookbooks and the potential uses for tools available in the HTRC Portal was essential for determining scope and utility of the project. Answering basic questions about who the users would be and what purpose the site would serve helped to narrow the focus and make decisions about basic content areas. The next steps were to research and evaluate different platforms such as WordPress, Omeka, and Scalar. I decided to create a WordPress site on the New York University Web Publishing platform.
The basic steps in my work for building the project were:

- Build the initial collection on HathiTrust
- Upload sample sets to HTRC to test out metadata analysis and HTRC algorithms
- Evaluate accuracy and significance of these tests
- Finalize the HathiTrust collection (evaluate and de-duplicate records)
- Design WordPress site (select a theme and decide on basic architecture)
- Find images (appropriate usage rights, resizing and storing image files)
- Build the WordPress site (draft versions of all pages and posts)
- Do metadata work and text analysis work (see detailed Work Plan section below)
- Create visualizations and upload them to WordPress
- Write analysis and interpretation for the visualizations
- Re-write and expand content for all pages of the site
- Finalize the design of the home page and overall quality check

Work Plan

This section describes the workflow steps needed to build the collection, analyze the metadata, and analyze the full-text.

Building the Collection:

- Create public permanent collection in HathiTrust from full-text search results
- Evaluate contents of collection, narrow time span to pre-1920 public domain titles
- Check and de-dupe individual records (multiple printings, editions, and scans)
• Evaluating and de-duping reduced the collection from over 2000 records to a final set of 1450

• Download basic collection metadata and upload volume ID numbers to create workset in HTRC

Analyzing Collection Metadata:

• Download catalog records in MARCXML format using Marc Downloader tool

• Convert MARCXML to MARC21 and join records into one file using MarcEdit

• Export selected fields as csv file using MarcEdit

• Clean up metadata using OpenRefine (for example “Brooklyn, opposite New York” was modernized to be “Brooklyn, NY” and “[1893].” was changed to be “1893” by removing library catalog punctuation)

• Create topical subsets of records using OpenRefine

• Upload csv files to Tableau to explore data

• Produce visualizations in Tableau (maps, timelines, author chart, books by year)

Analyzing Full Text:

• Search selected terms (“vegetarian” “frugal”) as keywords in collection and download metadata for the results

• Upload new worksets in HTRC for keyword search sets

• Upload new worksets in HTRC for topical subgroups of titles created using OpenRefine (see metadata section above)
- Run Meandre Topic Modeling algorithm on selected sets
- Run Meandre Dunning LogLikelihood to Tagcloud algorithm to compare and contrast two sets
- Evaluate results, look for errors in data, and re-run as needed
- Export results from HTRC

**Evaluation**

The Early American Cookbooks project has successfully captured the objectives outlined in the original plan for the project as described in the June 2016 MALS capstone abstract form. The project has succeeded in the three main goals:

- create a freely available, searchable online collection of early American cookbooks
- offer an overview of the scope and contents of the collection
- use digital humanities tools to explore trends and patterns in the metadata and the full text of the collection.

The project has also succeeded in completing the two main deliverables:

- collection of 1450 full-text titles on HathiTrust
- website to present a guide to the collection and the results of the analysis.

The aspects of final product that differ from the original capstone proposal include the following. The capstone estimated that the collection would contain approximately 2500 cookbooks, but evaluating and de-duping the collection reduced the number to 1450. The
capstone also anticipated that the site would contain embedded full text for a selection of key titles plus a categorized, linked list of all the titles in the collection. This proved to be impractical because embedding the full text onto the WordPress site was not permitted by NYU Web Publishing’s limitations on the use of iFrames. In addition, it became clear that the WordPress site should be a guide and gateway to the full text on HathiTrust and not a platform to host a duplicate version of the content.

**Challenges and Solutions**

This project encountered several challenges in the course of its development and finding solutions was a very valuable learning experience. The HTRC Portal algorithms sometimes failed to work or created poor results. Tracing the cause of these problems meant questioning the validity of upload data, the suitability of the tool, and/or the composition of the dataset. Sometimes the answer was simple, such as when I found that an algorithm repeatedly failed to process a workset because a library call number had been substituted for a HathiTrust volume ID number in the upload metadata. Other problems seemed to stem from a mismatch between the chosen algorithm and my dataset. For example, the OpenNLP Entities algorithm is designed to generate lists of people and place names, but my results included terms such as “Brown Sauce” and “Butter Taffy” as if these were names of individuals.

More complex problems involved the composition of the dataset and how choices I made in subdividing my data had influenced the results. For example, the word “feces” was a
valuable clue in interpreting and correcting a problem with one of my datasets. The word first appeared in a Meandre Dunning LogLikelihood tag cloud of over-represented terms for books published in the Southern census region of the United States. It seemed hard to believe that cookbooks on Southern cuisine featured feces, so a re-examination of the dataset was in order. Washington, D.C. is part of the Southern census region, but it is also the place of publication for large numbers of government documents. Separating out the books published by government agencies from the larger Southern set proved to be the answer to the problem. Running the Meandre Dunning LogLikelihood algorithm and the Meandre Topic Modeling algorithm on the government documents alone showed that these publications were concerned with scientific approaches to human nutrition. Feces along with other terms describing digestion were prominent in this dataset. Rerunning the algorithms on the set of Southern books without the government publications created results that did not contain the word feces and that were more consistent with the subject matter of the titles in the set.

Creating a map visualization in the Tableau data visualization platform presented some challenges. The map shows place of publication data from the library catalog records for the full set of books in the Early American Cookbooks collection. The catalog records included needed a great deal of clean up work in OpenRefine to standardize old state abbreviations such “Calif.” or “Mass.” plus many different types of punctuation and cataloging terminology. The next step was to load the cleaned up data into Tableau, which has many features and excellent mapping tools. After tinkering with the format of
the state names, I was able to get Tableau to auto-generate the latitude and longitude for all the place of publication data in my set. The Tableau map was very easy to create once I had the latitude and longitude data, but setting up an appropriate color range for visualizing the number of records required additional experimentation.

WordPress also presented some challenges, mostly resulting from the limitations of widgets, plugins, and design elements. NYU Web Publishing would not permit me to embed a HathiTrust search widget because of their restrictions on iFrames in WordPress. I needed to come up with an alternate solution so I learned to make minor customizations using HTML and created my own link button by using the HathiTrust logo in a free text widget. The limitations on iFrames also prevented me from embedding HathiTrust books on my site. The existing gallery plugins on the NYU Web Publishing instance did not allow me to create a satisfactory gallery of images with external links, so I used HTML to create a table displaying book covers with links to HathiTrust. I later discovered that this table did not display properly on smaller computer screens, so I deleted it and used a slider plugin instead. WordPress has a wide array of off the shelf tools, but adapting them to suit the needs of a particular project can be challenging.

**Best Practices**

The practices that worked best to realize the project goals centered on planning, project management, and iterative approaches to testing and building. My initial plan was to do all of the metadata and text analysis first work first, create the data visualizations, and then think about my results and design the WordPress site with all of content as the last
step. I revised this plan when I decided to do a proposal to present the project at a library conference. The proposal recommended including a URL so I changed my plans and built a working version of the site with draft content.

Making unfinished work public is not my natural inclination, but I learned in my coursework (see section below) that sharing work in an early stage is a good way to solicit valuable feedback and find collaborators. I was pleased that my conference proposal was accepted, but the decision to build the site first had other valuable benefits as well. By building a working prototype of the project, I was able to focus more precisely on what the finished version should contain. My work on the metadata and text analysis was more targeted and efficient as a result. I did the analysis work in planned batches and uploaded the results to the WordPress site as it was completed. As a result, I did not spend much time and effort creating work that would not be used in the final version. I focused on planning and creating the minimum viable product and I was able to meet all of the project goals and deliverables by the deadline.

**Relationship to Track and Previous Course of Study**

The Early American Cookbooks project is directly related to my coursework in the Digital Humanities track of the Master of Liberal Studies program. The Introduction to Digital Humanities with Matthew Gold and Stephen Brier was my first encounter with topic modeling and other computational methods and I began to see how these methods are related to library metadata work. In the Digital Humanities Methods and Practice
course with Matthew Gold, I gained valuable hands on experience in project planning and agile project management and learned about the value of making unfinished work public. The Interactive Technology and Pedagogy courses with Stephen Brier, Luke Waltzer, and Michael Mandiberg taught me more about designing a project to meet a clear user need, the importance of focusing on the minimum viable product and the need to show a proof of concept at the outset of a project. The Analyzing Cultural Data course with Lev Manovich taught me about using data visualization to tell a story and how to present complex information in a simple way. Elizabeth Macaulay-Lewis’s courses, Islamic Art and Architecture and Great Digs, were not directly related to the Digital Humanities track, but they taught me a great deal about using WordPress, copyright and permissions issues, using images, and how to write for an academic audience online rather than in print.

Continuation of Project

The Early American Cookbooks project has succeeded in its goals, so I have no immediate plans to expand it beyond its current scope. Possible future plans may include writing about the project as a case study for librarians interested in using digital humanities tools to enhance access to special collections material. The sustainability of the HathiTrust collection will depend on HathiTrust’s long term commitment to maintaining individual collections as a part of their interface. The WordPress site will be hosted by NYU Web Publishing as long as I continue to be a member of the NYU community. If I leave NYU, the files for the site will be available to export to another WordPress instance.
Appendices

Several screenshots from the Early American Cookbooks website are included as appendices: Collection home page (Appendix A), Books by year: 1800 to 1920 (Appendix B), Regional cookbooks (Appendix C), Vegetarian timeline (Appendix D), Topic modeling (Appendix E), and Fannie Farmer (Appendix F).
Early American Cookbooks
Read and search 1450 full text titles

Collection

Early American Cookbooks is a carefully curated collection of 1450 cookbooks published in the United States between 1800 and 1920. All of the titles in the collection are in the public domain and are available in full-text on the HathiTrust Digital Library. Browse the titles, read them online, and search through the contents of each book. Search across all 1450 titles to find a particular ingredient, recipe name, or anything else and go directly to the results in the full text.

Cookbook History

The history of cookbooks in the United States can tell us not only about food preparation and consumption, but also about the history of American culture. Reading the recipes people used every day shows us how they lived their lives. Most cookbooks were written by and for women, so they are especially important for exploring women's history. Cookbooks focus on women's skills in the kitchen, but they also reveal how they saw themselves, their families and their communities. A large collection of
Cookbooks spanning a wide historical period can illustrate changing attitudes toward diet and nutrition, new ingredients and styles of cooking, homemaking, and the commercialization of food production. Please see the Resources page to learn more about cookbook history.

Full text searching and data visualization

Most early cookbooks in printed form are accessible only in special collections libraries or private collections. Digitization of library collections has made high-resolution, searchable scans available to the public. Early American Cookbooks is not simply a collection of digital cookbooks which can be viewed one at a time. Keyword searching across all 1450 titles allows exploration of the content of the full collection. In addition, this site offers analysis of the collection using digital humanities tools. Text analysis and data visualization show trends...
and patterns in the collection which may add to our understanding of early cookbooks and the history of American food. Please see the About page for more information.
Early American Cookbooks
Read and search 1450 full text titles

Books by year

The number of cookbooks published per year grew steadily during the 19th century and bounded upwards in the early 20th century. In the early 19th century, most families used collections of handwritten recipes, often handed down through generations and shared with neighbors and friends. The publishing industry in the United States expanded rapidly during the late 19th century and commercially produced cookbooks became widely available. The market for cookbooks continued to expand with...
Early American Cookbooks
Read and search 1450 full text titles

Regional cookbooks

This map shows the number of books published per state for the full collection of 1450 cookbooks. New York has the greatest number of books published, followed by Massachusetts, Illinois, Pennsylvania, and California. These numbers align with the growth of the book publishing industry in the United States. New York City, traditionally the publishing center of the United States,
published the greatest number over time, followed by other publishing centers in Boston, Philadelphia, Chicago, San Francisco, and Los Angeles. The trend in the numbers also shows the history of westward expansion from 1800 to 1920, with the greatest total numbers in the East and much lower numbers in the West.

Dividing the collection into United States census regions shows that 61.1% of the titles were published in the Northeastern region, 24% were published in the Midwest, 8.8% in the West, and 6.1% in the Southeast.

The map was created by downloading the MARCXML catalog records for the collection from the HathiTrust Research Center. The records were converted using MarcEdit and then sorted and cleaned using OpenRefine. The records were then loaded into Tableau to create latitude and longitude data from state names and to make a filled map displaying the number of records per state.

Tagged on: map, MarcEdit, Midwestern, Northeastern, OpenRefine, publishing, Southern, Tableau, Western

← Authors chart

Topic modeling for early American cookbooks →

APPENDIX C

Government publications
Marginalia and clippings
Regional cookbooks
Special equipment
Topic modeling
Unusual recipes
Vegetarian cookbooks
Wartime shortages

Data visualizations

Authors chart
Books by year
Fannie Farmer
Frugal cookbooks over and under-represented terms
“Frugal” timeline
Government publications over-represented terms
Regional cookbooks
Northeastern
Southern
Midwestern
Western
Topic modeling for early American cookbooks
Topic modeling for government publications
Vegetarian over-represented terms
“Vegetarian” over time
Early American Cookbooks
Read and search 1450 full text titles

“Vegetarian” timeline

The number of books in the Early American Cookbooks collection which contain the word “vegetarian” in the text increases slowly in the late 19th century and then grows substantially in the years from 1900 to 1920. The vegetarian movement in the United States grew over the same timespan and publishers began producing cookbooks devoted to a purely vegetarian diet. The timeline also reflects the increased number of references to a vegetarian diet not only in books such as How to Cook Vegetables (1891) by the bestselling author Sarah Tyson Rorer, but also in general cookbooks such as Fannie Farmer’s A New Book of Cookery (1917).

Full Text Search
Search the collection on HathiTrust

Categories
Cookbooks for men
Data visualizations
Famous cookbooks
Fannie Farmer
Frugality
### Data visualizations

**Authors chart**

**Books by year**

**Fannie Farmer**

**Frugal cookbooks over and under-represented terms**

**“Frugal” timeline**

**Government publications over-represented terms**

**Regional cookbooks**
- Northeastern
- Southern
- Midwestern
- Western

**Topic modeling for early American cookbooks**

**Topic modeling for government publications**

**Vegetarian over-represented terms**

**“Vegetarian” over time**
Early American Cookbooks

Read and search 1450 full text titles

Topic modeling for early American cookbooks

Topic modeling shows some interesting trends and patterns in the text for the 1450 books in the collection. The ten word clouds in the chart below show different topics or clusters of words that recur across all of the texts. The names of the topics were not generated by the algorithm but rather added as a way to label and interpret the clusters. While it is impossible to draw definitive analytical conclusions, the topics do provide a interesting snapshot of the subject matter.

Early American cookbooks had many common themes, largely because the diet and cookery techniques in the 1800 to 1920 period were far more homogenous than they are today. Nearly every cook used salt, paper, and butter as the primary methods of seasoning (topic 1), boiled kettles over a fire for long periods of time (topic 2), prepared meat, most frequently with gravy or sauce (topic 3), made bread (topic 6), cake (topic 9), and various fruit based items such as jelly, lemonade, or ice cream (topic 10). Some topics are more sparse and hard to interpret. Topic 5 possibly represents boiling vegetables and topic 7 seems to be about pickling or similar processes. Topic 4 includes words such as place, time, made, long, heat, air, and dry. The significance is unclear, but the topic may possibly refer to storage of food in cupboards, drying fruit or other related processes. Topic 8 reaches beyond the ingredients and instructions into the how and why of cooking and homemaking. Words such as food, time, good, made, great, people, work, body, give, family, years are commonly present in the forewords and introductions to cookbooks which sought to provide inspiration for readers.
many famous titles produced in multiple printings and editions. Community cookbooks (published by a church or charitable organization to raise funds) also increased in number after the Civil War. The growth of the packaged food industry led to the publication of cookbooks tied to commercial products (such as Crisco or Jell-O) in early 20th century.

Tagged on: community cookbooks, handwritten, publishing

Data visualizations

- Authors chart
- Books by year
- Fannie Farmer
- Frugal cookbooks over and under-represented terms
- “Frugal” timeline
- Government publications over-represented terms
- Regional cookbooks
  - Northeastern
  - Southern
  - Midwestern
  - Western
- Topic modeling for early American cookbooks
- Topic modeling for government publications
- Vegetarian over-represented terms
- “Vegetarian” over time
Topic 5: Pickling?

Topic 6: Bread

Topic 7: Boiling vegetables?

Topic 8: Family
APPENDIX E

Topic 9: Cake

Topic 10: Jelly/lemonade/ice cream

Tagged on: algorithm  LDA  topic modeling

← Regional cookbooks  Books by year →

Fannie Merritt Farmer (1857-1915) was a major figure in American cooking in the late 19th and early 20th centuries. Her most successful cookbook, The Boston Cooking-School Cook Book, was first published in 1896 and sold millions of copies in many subsequent printings and editions. The 100th anniversary edition was published in 1996 and the book is still in print today.

The book was the first to introduce precise measurement and Farmer later became “the mother of level measurements.” Her discussion of food composition, caloric calculations and the
body’s need for nutrients, formed a systematic view of cooking that influenced cooking instruction for decades to come (Feeding America).

Text analysis of Farmer’s books clearly illustrates her emphasis on precise measurements. When her books are compared to the full set of titles in Early American Cookbooks, the over and under-represented terms show that measurement is the key difference. In the tag clouds below, the over-represented terms are tablespoons, teaspoons, and cup. The under-represented terms are teaspoonful, tablespoonful and cupful which were frequently used in cookbooks of the era. Farmer insisted upon the difference between a vague “teaspoonful” and an exact “teaspoon.”

Fannie Farmer over-represented terms (Meandre Dunning Log Likelihood to Tagcloud Algorithm)
Fannie Farmer under-represented terms (Meandre Dunning Log-Likelihood to Tagcloud Algorithm)

This visualization was created by comparing two sets of texts, Fannie Farmer cookbooks and the full Early American Cookbooks set, using the Meandre Dunning Log-likelihood to Tagcloud algorithm in the HathiTrust Research Center Portal.

Tagged on: Boston Cooking-School Cook Book Fannie Farmer measurement

Search this site

Bibliography


