
# *Building and Maintaining the Worcester City Directory Digital Archive*

An Interactive Qualifying Project

by:

Stephen Blouin spblouin@wpi.edu

Alex Titus adtitus@wpi.edu

Submitted to:

Prof. Joseph Cullon, Project Advisor

Sponsoring Organization:

The Worcester Historical Museum, as represented by William Wallace

April 26, 2016

# Abstract

The Worcester Historical Museum (WHM) preserves a large collection of Worcester city directories. Many of the books are over a century old and are breaking from use. A way to preserve these materials is to digitize them. A previous project with the WHM produced a website, Digital Worcester. We studied websites from institutions we designated as peers of the WHM. We also started a plan to make Digital Worcester self-sustaining with a broad survey of funding options. We scanned 12 city directories, completing an effort to digitize every fifth directory from 1845 to 1945. The scanned batches were uploaded to Digital Worcester. For metadata we used a subset of the set created from peer institutions. Our final study compared a custom implementation of Digital Worcester to the hosted implementation.

# Acknowledgements

We would like to thank those who helped make this project possible.

Bill Wallace, Robyn Conroy, and the WHM for providing assistance in focusing our goals early in the project.

Joy Hennig of the Worcester Public Library for providing us with three directories that could be chopped to facilitate faster scanning.

Mike Kemezis, Jess Colati and the rest of the WPI Archives Staff for teaching and allowing us to use the Archive's scanners.

Professor Joe Cullon provided invaluable assistance in steering this multi-directional project. He also worked tirelessly to find destructible copies of each of the city directories that we scanned and chopped them himself to save us the time and hassle of scanning 1000+ pages books 2 pages at a time with the overhead scanner.

# Executive Summary

The Worcester Historical Museum (WHM) preserves a large collection of Worcester city directories. City directories are an important historical resource because they recorded the names, addresses, and professions of the population, the locations of businesses, and information about the city's government and organizations. Unfortunately, many of the books are over a century old and are breaking with use by a combination of historians, researchers, and genealogists. The main problems we identified with using city directories were:

- ease of search across and within directories
- accessibility of the physical book
- fragility of the physical book

A solution to lessen the severity of these problems is the digitization and online display of city directories. Implementing this solution requires the creation of a website to display the digitized directories, a standardized metadata value set, a plan to allow the site to be self-sustaining, a policy for dealing with the data produced by the digitization process, and the actual digitization process itself.

A previous project with the WHM produced a website, Digital Worcester, using the free, open source platform Omeka, a joint project of the Roy Rosenzweig Center for History and New Media and George Mason University. This website is hosted on the related site Omeka.net, which is dedicated to hosting Omeka sites. We built on this existing project to improve Digital Worcester by adding scanned directory content and developing metadata and file-naming standards.

We studied existing websites from institutions we designated as peers of the WHM. These institutions had similarities in scope, size, and financial resources. We used this study to produce a Dublin Core metadata value set for Digital Worcester, specifically focusing on the rights field.

Development of a plan to make Digital Worcester self-sustaining started with a broad survey of funding options. We studied grants, paywalls, advertisements, and donations. Grants were found to be more oriented toward short-term creation of projects and not coverage of operating expenses. Paywalls were found to be financially sound, but not appealing given the limitation they impose on accessibility. Advertisements require high traffic sites to generate meaningful revenue. Donations are difficult to estimate as a source of income, but a simple calculation gave the average monthly cost of hosting Digital Worcester on Omeka.net as

$83.33. Any donation or other option would have to supply that cost at a minimum to maintain the site.

We scanned 12 city directories, completing an effort to digitize every 5th directory from 1845 to 1945. Progress was greatly aided in this endeavor by the use of a feed scanner, which can scan loose pages very quickly. To get loose pages, the book must be unbound. Prof. Joe Cullon both located spare directories and unbound them for us, which was vital for achieving this goal.

The scanned batches were then split into smaller 30-page files, named according to their contents, and uploaded to Digital Worcester. For metadata we used a subset of the set created from the peer institutions, as we want to coordinate the metadata decisions with the WHM.

We created a data management policy to ensure that our data is appropriately backed up and shared in a manner that will ensure only relevant parties can access it. The data is currently located on a research server at Worcester Polytechnic Institute (WPI). WPI has its own backup solution, which suits our needs in this area. For sharing we determined that only Prof. Cullon, WHM staff, and ourselves would have access to the files, for security purposes.

Our final study compared a custom implementation of Digital Worcester to the hosted implementation. This was conducted because Omeka.net limits the types of plugins that its Omeka servers can use, barring some that we felt could be useful. The main drawback to the custom implementation was the technical expertise required to manage the server, while the main advantages were a lower operating cost and more flexibility with the website's themes and plugins. Unfortunately those plugins turned out to be less useful than we had hoped.

Important decisions determining the future of Digital Worcester center around the financial maintenance of the site and the server hosting location.

## Recommendations

We recommend that the WHM keep Digital Worcester as a hosted website on Omeka.net and that the site be maintained financially though a combination of grants and donations. The technical expertise required to successfully manage a server in this era far outweigh any potential benefits that an institution like the WHM could gain from owning and administrating the server themselves. For grants, we recommend the Massachusetts Cultural Council's Cultural Investment Portfolio to maintain the site and both the National Endowment for the Humanities' Humanities Collections & Reference Resources and the National Archives' Access to Historical Records to complete the digitization process. Additionally, we recommend

the implementation of a donation link on Digital Worcester that would allow users to support the site and ensure its continued operation.

# Table of Contents

# Chapter 1: Introduction

## 1.1 Goal

The overall goal of this project is, through digitization, to make the Worcester Historical Museum's city directory collection more accessible and easily searchable.

## 1.2 Objectives

To achieve our mission we need to:

- continue to digitize the city directory collection of the WHM.
- evaluate the suitability of Omeka for hosting digital collections.
- determine, evaluate, and implement best usability practices for the database and Omeka.
- identify, evaluate, and present options for the expansion and maintenance of the website.
- establish and unify best metadata practices with WHM.
- develop a data management policy for the digital directories.

By reaching these goals, the WHM's directory can be successfully digitized and the problems described below ameliorated.

## 1.3 Problem Outline

City directories are similar to today's phone books in that they contain information about the residents and businesses of a city. They offer more information than phone books, as they also provide records of the residents' addresses and occupations as well as advertisements from local and regional companies. Directories include other important details about the city, such as the population, members of the city government, fire department, police, and other organizations and even descriptions of school schedules. These historical directories were intended to help the users by providing information that people today would turn to the Internet for, such as the schedule for the school year or hours of operation for businesses or libraries.

The books are very useful to historians and researchers because of this information. They can be used for topics as varied as genealogy, statistical analysis, geographic visualization, and industrial history. The broad appeal of the information leads to a high demand

for them. Unfortunately, as many of the directories are over a century old, the directories are prone to breaking with the heavy use.

The people most interested in city directories are casual individuals and scholars. Casual individuals are often simply curious about their ancestors or prior residents of their homes. Scholars and researchers are also avid users of city directories and use them to identify people and trends in time. For example, Roy Rosenzweig used Worcester city directories ranging from 1871 to 1935 in his work *Eight Hours for What We Will* to trace the lives of individuals and map the locations of saloons around the city over time.

Searching for a specific person or object in a collection of directories is a cumbersome process at best. Each directory has its own order to the sections, requiring more time to even determine if a given name is present. This leads to a very time--consuming process when trying to follow a specific trend or name through time. The process becomes even more difficult the larger the project's scale. Rosenzweig, who manually mapped the locations of saloons and saloon owners over a fifty--year period for his book, noted that "all this tracing can become burdensome" (Rosenzweig, pg. 293). Additionally, searching directories by hand requires that the searcher be in physical contact with the book in an archive reading room. Some may not be willing or able to make the journey to the museum or archive where the directories are located or the directory may not be available because the archive may wish to protect its fragile few copies.

Thus, our primary problems that we identified are:

- Ease of search across and within physical directories
- Limited access with respect to hours of operation of the institution that stores the directories
- Limited access with respect to the physical location of the institution that stores the directories
- Fragility of the physical copies of directories

By digitizing city directories, these problems can be ameliorated, and additional preservation challenges can be overcome. Digitization of historical materials has many advantages. One advantage is the ability for more than a single individual to access records at the same time. With physical records, only a limited number of materials are available to an institution and subsequently, anyone who needs to access those materials. When such materials are digitized, an unlimited number of users could access them at once. Along with this, users can access these materials from anywhere at any time, rather than when an

institution is open. When records are digitized and placed online, there is no limit to where or when users can access those materials.

Another advantage of digitization is the ability to safely back up records. In disasters, institutional records can be lost. With physical materials, backing up is a labor--intensive process that involves a capture process to perform a physical reproduction of the materials. In addition, these backups take up physical space that an institution may not have. In many cases, backup copies of physical materials are not even made for these reasons. With digital collections, backing up the materials is as easy as downloading them or copying them to another location, without the need of physical reproduction.

Another advantage of digitization is that it reduces the need for physical storage space within an institution. Directories that have been digitized can be stored in a more specialized storage facility as most interaction with the directories will be with the digital copies.

Finally, digitization will protect the remaining physical copies by limiting the amount of handling the directories undergo. Research on the content of the directories, such as fonts and layouts as well as the actual information in the book, can be done with good quality digital images. Studies on the actual material composing the book would still require the physical copy.

# Chapter 2: Environmental and Literature Review

## 2.1 Omeka

Omeka is a free, open-source web-publishing and content-management platform designed for displaying collections from libraries, museums, and archives. Omeka was created as a joint project between the Roy Rosenzweig Center for History and New Media and George Mason University. Based on the recommendations in (Connolly, et. al., 2015), Digital Worcester adopted Omeka over competing content-management and web-publishing platforms CONTENTdm and SimpleDL.

Omeka's core functionality can be extended through the use of plugins, which add functions such as: the creation of text-only webpages, the replacement or augmentation of the existing Dublin Core Metadata Initiative (DCMI) basic metadata set and the embedding of PDF documents into the webpage.

Materials in Omeka are organized into collections, items, and files, each of which can have its own metadata. Collections allow multiple items to be grouped together, such as a city directory collection containing multiple city directory items, for example. Items are the basic unit of Omeka, and can have multiple files associated with them, for example a city directory item may contain files of scans of the directory's pages.

## 2.2 Omeka.net

Omeka.net is a website hosting platform that allows users to create and share their Omeka sites without having to buy, configure, and manage a web server. After making an account, users choose one of several plans, from a Basic plan offering 500 MB of storage, limited access to plugins, and a single website for free up to a Platinum plan offering 25 MB of storage, access to all Omeka.net plugins, and an unlimited number of websites for $999.00/year. Currently Digital Worcester is hosted on Omeka.net under the Platinum plan.

The version of Omeka offered on Omeka.net is limited compared to the version available for download from Omeka.org (the home site). A summary of major differences between an Omeka implementation from Omeka.org and one hosted on Omeka.net is shown in Table 2-1.

Table 2-1

| **Features, basic requirements** | **Omeka.org** | **Omeka.net** |
|---|---|---|
| Server | Server required (LAMP) | No server required |
| FTP Client | Required for file uploads and modifying Omeka | Not required |
| Web-based administrative interface | Yes | Yes |
| Storage Space | Determined by server admin | Determined by plan: 500MB, 1GB, 5GB, 10GB, or 25GB |
| File size limitations | Determined by server admin, with Dropbox plugin for files that exceed that limit | 64 MB maximum |
| Sites per Installation | One website per Omeka installation | Determined by plan: 1, 2, 5, 10, or unlimited |
| Custom Domain Redirects (base URLs) | Yes | No |
| Plugins and Themes | All plugins and themes listed on Add-Ons page | Determined by plan; only core plugins and themes available |
| Pricing | Free | Basic plan is free; other plans available with varying prices |
| Support | User forums | Detailed instructions at info.omeka.net; Troubleshooting help at info.omeka.net/contact |

Based on (Omeka.net)

## 2.3 Peer Institutions

We identified institutions that we felt were comparable to the WHM in scope, size, and financial resources. We then studied aspects of these peer institutions' websites to learn what the WHM would be capable of constructing and maintaining. The metadata used by these institutions and others were also studied to help produce a rights field for Digital Worcester's documents. The peer institutions identified were:

- Passport Holyoke's Creating Holyoke

- The Jones Library's Digital Amherst
- Lowcountry Digital History Initiative
- Wooster Digital History Project
- New Haven Free Public Library

We additionally examined the websites of the following:

- North Carolina Digital Heritage Center
- Pasadena Digital History Collaboration
- University of Pittsburgh Digital Research Library

We studied the peer institutions to determine, in conjunction with studying the DCMI's Metadata Terms, the appropriate values to use for each of the fields available to us. The results can be found in Appendix C.

Examples of rights field metadata used on the websites of both peer and non-peer institutions are displayed in Table 2-2.

Table 2-2

| **Institution** | **Rights Field Metadata Value** |
|---|---|
| Amherst (Jones Library) | "This digital file may be used for educational purposes, as long as it is not altered in any way. Prior written permission is required for any other use of the digital files from the Jones Library." |
| Holyoke (Wistariahurst Museum) | "This digital file may be used for educational uses, as long as it is not altered in any way. Prior written permission is required for any other use of the files from the Wistariahurst Museum." |
| Holyoke (Holyoke Public Library) | "Rights: This digital file may be used for educational uses, as long as it is not altered in any way. Prior written permission is required for any other use of the files from the Holyoke Public Library" |
| New Haven Free Public Library | "Digital Image © New Haven Free Public Library. All rights reserved. Image may be used for personal research or educational use without prior permission. For request to publish or exhibit, contact New Haven Free Public Library." |
| North Carolina Digital Heritage Center | "This item is presented courtesy of the North Carolina Collection, UNC-Chapel Hill, for research and educational purposes. Prior permission from the North Carolina Collection is required for any commercial use." |
| Pasadena Digital History | "The Pasadena Public Library makes no assertions as to |

| | |
|---|---|
| Collaboration | ownership of any original copyrights to digitized images. However, these images are intended for Personal or Research use only. Any other kind of use, including, but not limited to commercial or scholarly publication in any medium or format, public exhibition, or use online or in a web site, may be subject to additional restrictions including but not limited to the copyrights held by parties other than the Library. USERS ARE SOLELY RESPONSIBLE for determining the existence of such rights and for obtaining any permissions and/or paying associated fees necessary for the proposed use." |
| University of Pittsburgh Digital Research Library | "This work is believed to be in the public domain with no known restrictions on access and use." |

## 2.4 Financing and Maintenance of Site

There are several options to financing a website. The main methods used are subscriptions and advertisements. For the WHM, other means of financing exist because of the museum's status in the community and legally as a non-profit. These include federal, state, & local grants,  and donations & gifts.

Grants can be used to provide short-term funding for projects, usually for the development and implementation stages of a project. To receive a grant, an organization has to fill out an application detailing its activities and how the relevance of those activities to that particular grant. The National Endowment for the Humanities (NEH) has many grants for various projects within the humanities. Only one grant deals with activities relevant to Digital Worcester, the Humanities Collections & Reference Resources (HCRR) grant. The HCRR grant funds efforts to extend life of important collections and make their content widely accessible, often through digital means (NEH, 2016). Additionally the National Historical Publications and Records Commission (NHPRC) of the National Archives has a grant entitled Access to Historical Records, which is similar to the HCRR grant (National Archives, 2016). At the state level, the Massachusetts Cultural Council and Mass Humanities (an NEH affiliate) provide grants for the humanities. Mass Humanities won't fund operating expenses, but has funded websites dealing with presentation of documents of historical significance in the past. The Massachusetts Cultural Council's Cultural Investment Portfolio is one of the only grants that provides operating support to cultural institutions. The requirements are strict, requiring certain minimum operating expenses and legal statuses, including completing filings as a public charity

in Massachusetts and as a nonprofit organization. Local grants can be obtained through the Worcester Art Council.

Subscription-based services are a common option for websites, typically those that produce time-sensitive or live material such as newspaper-affiliated websites. Subscribers pay periodically, typically on a monthly or yearly basis, for access to the material. This is known as a "hard" paywall, in contrast to a "soft" paywall, where a limited amount (either in scope (pure soft) or in number (metered)) of materials are available for free before a paid subscription is required. Hard paywalls have fallen out of favor except where a particular expertise or reputation allows a site to keep its visitors, such as the Wall Street Journal (Filloux 2011).

Online advertisements are a common option for financing websites. Websites are paid by advertising companies to display ads on sections of their webpages. The pricing model for online advertising is broken down into two major categories: performance-based and impression-based, of which performance-based is the more popular, revenue-wise (PricewaterhouseCoopers, 2014). Performance-based models use a cost-per-click (also pay-per-click) metric to determine payment, which counts the number of clicks on ad banners. Impression-based models use a cost-per-mille (CPM) metric, which counts the number of times an ad loads on the webpage. Advertising thus works best as a funding mechanism for high-traffic sites. Discovering average payment rates for cost-per-click and CPM is difficult, as the rates depend upon both the advertising company and the website. The agreement between the two is confidential, so obtaining exact, specific numbers is nearly impossible, but consultants' blogs contain generalized numbers. These indicate that an average CPM would be at most $10 (Sonn), usually nearer to $1 (Zwilling), for a rate somewhere between $0.01 and $0.001 per view. For cost-per-click, it is somewhat higher, ranging between $0.20-0.60 per click (Statistic Brain), because of the lower probability of occurrence.

A final option would not attempt to have the site fund itself directly, instead relying on existing donations, gifts, and memberships for support. The WHM's current website is funded through a gift from the Fred Harris Daniels Foundation. Membership is an existing mechanism for private individuals and companies to support the museum. Either of these could be extended to encompass Digital Worcester.

# Chapter 3: Methods

## 3.1 Scanning

Gordon Library Archives acquired a new feed scanner, a Fujitsu fi-7160, which we found to be very quick and effective, greatly reducing the amount of time we spent scanning. The feed scanner requires that the directory be unbound -- each page removed from the binding -- which required that we find expendable copies of the directories to effectively destroy. Thanks to the diligent searching efforts of Prof. Cullon, Robyn Conroy, and Joy Hennig copies of all of the directories that we planned to scan were found.

During scanning, we used the following procedure:

1. Divide directory pages into batches of 150 pages
2. Load into scanner: page front onto feed tray; page bottom in the air
3. Open Acrobat
4. File>Create>PDF from Scanner>Custom Scan
5. Scan options:
   Sides: Both Sides
   Color Mode: Autodetect
   Resolution: 300/400 dpi depending on size of smallest font
   Page Size: Automatic
   New PDF Document
   Mark Optimize Scanned PDF: with 2/5 on slider
6. Scan
7. Wait
8. Save as batchX, where X is a number
9. Close Acrobat

Any damage to the directories pages (including missing pages) was noted during the scanning and uploading process.

## 3.2 Upload Procedure

After scanning, the files are in large batches with unhelpful names. The large file size makes it time-consuming for users to handle the files and for the PDF Embed plugin to display the files. The unhelpful names make it difficult to keep track of which files have been uploaded

and which remain to be uploaded. The names would also make managing files in the future difficult, such as finding a particular file to update with a better scan of pages.

To solve these problems, we decided to split the scan batches into roughly 30-page files and developed standardized file naming conventions for the smaller files. The details of this naming scheme, an example of the naming scheme, and a second justification for the splitting and naming can be found in Appendix A.

After the batches are split into upload-sized files, we ran Adobe Acrobat's Text Recognition program on them. This uses Optical Character Recognition (OCR) to create searchable text in the PDF image. The PDF Text plugin that we use on Digital Worcester can extract this text and make it searchable from the site. A walkthrough of the file splitting and OCR procedure can be found in Appendix B. The now searchable files are then renamed according to the naming conventions specified in Appendix A and uploaded to server.

## 3.3 Metadata Fields

An important aspect of this project was determining and inputting metadata for directories. After gathering data from the websites of peer institutions and studying the DCMI's description of metadata terms, we produced possible values for each of the terms in the full DCMI metadata element set. A comprehensive listing of these values can be found in Appendix C, but the most important fields currently being used are the title, subject, description, date, format, language, type, coverage, date created, and date copyrighted. Some terms were difficult to find unequivocal values for, and for these we chose a 2-3 of the most relevant choices based on what was observed in the peer institutions. Different vocabularies (a set style or selection of values for each term) also exist for filling in these terms, the most observed example was that of the Library of Congress (LoC).

We had hoped to finalize the vocabulary for Digital Worcester in conjunction with the WHM, however we were not able to due to timing and scheduling issues. As such, Digital Worcester currently uses a small set of metadata for each directory.

Searches across directories were enabled by the PDF Text plugin for Omeka. This plugin allows the site to automatically read text created by OCR and use it as a value for a special metadata field, which is added to each uploaded page batch. Omeka searches the metadata fields of each page batch, including the this special metadata field, enabling users to search for a particular name and receive the proper page batch.

## 3.4 Data Management Policy

The process of scanning directories and saving the files as backups for the main website generates a lot of data that must be kept secure. To this end we developed a data management policy which met our needs that the data be shared only with relevant parties and that the data could be recovered in the event of a disaster. This policy can be found in Appendix D.

# Chapter 4: Analysis

## 4.1 Custom Omeka Implementation and Omeka.net Hosting

By running a custom Omeka implementation alongside the Omeka.net site, we hoped to compare the strengths and weaknesses of both approaches.

The Omeka.net hosting service is a paid service with different tiers, each offering more features than the last. These tiers range from their Free service, which only gives 500MB of storage, 1 site, 14 plugins, and 5 themes, to their Platinum plan, which gives 25GB storage and unlimited sites, plugins and themes, and costs $999 per year. The custom implementation is running on a dedicated computer that was set up to serve Omeka. This computer was given a Linux installation, then had Apache, MySQL and PHP installed. Next, Omeka itself was installed and the system was brought up to spec with the Omeka.net site.

Moving to a custom implementation of Omeka offers a few important advantages. One pro of such a custom implementation is that it has a low maintenance cost, consuming about 100W. This translates to 100W * 1 year * 13.13¢ / kWh (Worcester average) = $115, the same cost as running a light bulb. Another benefit of running this custom system is that it gives unlimited storage space, as opposed to the limited storage space present on even the highest tier Omeka.net plan. This is important, since a complete digital collection could very well go over 25GB in size. Another positive of using the custom machine is access to unlimited plugins and themes, as well as access to low-level system resources to implement custom functionalities. Some such functionalities are custom search engine integration, such as Solr or Google searches, or custom plugins that could be made specifically for Digital Worcester.

The custom implementation does not come without pitfalls, however. Additional tech support resources would be required to install and maintain the Omeka system, which may negate the system's raw energy costs. Additionally, custom search features that are possible in the custom implementation are not all that much better than Omeka's default search system, and they would have the additional cost of development. The last difficulty is the fact that the physical machine which would be running the custom Omeka implementation has to be able to cope with relatively large user demand of the site, as well as have a decently fast internet connection.

In either case, the museum will also need a policy for storing and backing up these files when control of Digital Worcester passes to them. The initial storage server can be located in the museum itself, but the back ups should be stored off-site, to preserve the data should the

museum itself be damaged. This can be achieved either through a third party or by an off-site server managed by the museum itself.

## 4.2 Financing and Maintenance of Site

For financing and maintaining Digital Worcester we considered several options: grants, paywalls, advertisements, and donations & gifts. Each should be considered based on how effectively a mechanism fulfills the goal of making Digital Worcester self-sustaining and how applicable a mechanism is to a website like Digital Worcester. Analysis along these lines is hampered greatly by a lack of knowledge about our expected user base.

Grants provide funding in the short term–usually a year–and also usually stipulate that the funding not be used for operating expenses. This means that all of the grants mentioned in Chapter 2.4, except for the Cultural Investment Portfolio from the Massachusetts Cultural Council, cannot be used for the maintenance of the site. The grants can, however, be used to fund the completion of the digitization effort or finance additional user studies to determine best practices for Digital Worcester.

Paywalls can be erected in one of two manners. The first is the subscription method defined in Chapter 2.4, where users would pay a monthly or yearly fee for unlimited access to the site. The second would be a pay-per-view-type method, such as charging for each download. This could be compared to buying a section of the directory electronically. In terms of technical complexity for integrating each method into Omeka, the second is more difficult. To estimate the number of subscribers Digital Worcester would require to break even at different payment rates, the values in Figure 4-1 were determined using the annual cost of the hosted Omeka.net site ($1000/year).

Figure 4-1

| **Subscription cost per year** | **Break-even subscriber numbers** |
|---|---|
| $6.00 ($0.50 per month) | 167 |
| $12.00 ($1 per month) | 83 |
| $24.00 ($2 per month) | 42 |
| $60.00 ($5 per month) | 17 |
| $100.00 (yearly rate) | 10 |

| | |
|---|---|
| $120.00 ($10 per month) | 9 |

The number of people required to break even on the low monthly rates seems achievable. The break even calculations for the pay-per-view method were also determined using the annual cost of the Omeka.net site and are summarized in Figure 4-2.

Figure 4-2

| **Cost per download** | **Required number of downloads per year** |
|---|---|
| $0.25 | 4,000 |
| $0.50 | 2,000 |
| $1.00 | 1,000 |
| $2.00 | 500 |
| $5.00 | 200 |
| $10.00 | 100 |

When considering the price of any paywall, careful thought must be given to the trade-off between the accessibility of lower prices and the fewer required numbers of higher prices. A paywall would also have to be created, incurring additional costs as well as possible credit card fees from user payments. Also to be considered in a decision are the current cost of viewing a physical directory at the WHM and that the peer institutions offer their materials online for free.

Advertisements are largely traffic-based mechanisms, requiring large numbers of potential viewers to achieve meaningful revenues. Based on the values found in Chapter 2.4 and assuming that the hosted Omeka.net site is used, we arrive at the break-even values found in Figure 4-2.

Figure 4-2

| **Rate per ad** | **Annual Break-even Traffic** |
|---|---|
| $0.001 | 1,000,000 |
| $0.01 | 100,000 |
| $0.10 | 10,000 |
| $0.20 | 5,000* |
| $0.60 | 1,667* |

*Traffic number is number of clicks on advertisement

These values suggest that, except for the pay-per-click model, yearly traffic to Digital Worcester will have to be high for the website to be self-sustaining on ad revenue alone. It should also be considered that no peer institution has advertisements on their websites. This suggests that such an approach, while technically viable, may not be an applicable solution to helping Digital Worcester achieve a self-sustaining status.

Donations, gifts, and other options can also be considered. For donations, a simple calculation similar to that done for the subscriptions can be performed. Assuming the $1000 per year cost of the Omeka.net site (likely needed to have space to store all of the city directory collection), donations would have to supply $83.33 monthly to sustain Digital Worcester. Gifts and options involving the use of money from membership fees would follow the same calculation. Considerations for using donations, gifts, and/or membership fees include feasibility of donation drives and the impact of the monthly fee on the museum's budget.

# Chapter 5: Conclusions and Recommendations

## 5.1 Directory Scanning Achievements

A major part of this project was completing the scanning process of every fifth directory (those whose years end in a 0 or 5). We scanned 12 directories, completing the range from 1845 to 1945. This includes the previously unfilled, vital period of rapid growth Worcester experienced in the late 19th and early 20th centuries, where we added scans of the 1885, 1895, 1900, 1905, 1910, and 1915 directories.

## 5.2 Data Management

We recommend that the WHM look into the online backup service provided by their IT contractor, Centrend, for a solution to managing the city directory data. This service will store backups of the files in a separate facility, providing protection in the event that a disaster befalls the museum itself.

## 5.3 Metadata

Digital Worcester currently uses the metadata fields for title, subject, description, date, format, language, type, coverage, date created, and date copyrighted. However, this is only a small subset of the possible fields included in the DCMI Extended set. We recommend that the WHM use Appendix C as a guide to determine which fields to include and what to put in them.

## 5.4 Custom Implementation or Hosted Service

Based on the analysis in Chapter 4.1, we recommend going with Omeka.net hosting instead of using a custom implementation. This is because it comes with ease-of-use and offers a no-frustration implementation, whereas the custom machine will need to be set up manually. Additionally, tech support costs may end up outweighing the raw cost savings that the custom machine would offer.

## 5.5 Financing Digital Worcester

Based upon the analysis in Chapter 4.2, we recommend a two-fold approach to financing Digital Worcester. Firstly, we recommend applying for the Massachusetts Cultural Council's Cultural Investment Portfolio, as it is designed to aid operating expenses, and for the

federal HCRR and Access to Historical Records grants to finish the digitization of the city directory collection. Secondly, we recommend maintaining Digital Worcester through donations, likely with a donation link in the site itself. These options are the most applicable to a website like Digital Worcester and are capable of meeting its financial needs.

The website will likely be unable to generate the traffic it needs to fully finance itself through advertising alone. This, combined with the lack of ad banners on the websites of peer institutions, makes advertisements a poor option for Digital Worcester. Paywalls could finance the site, but would conflict with our goal of making the website accessible.

# References


About Us - Digital Amherst. (n.d.). Retrieved October 8, 2015, from http://www.digitalamherst.org/about

About Us - Wooster Digital History Project. (n.d.). Retrieved October 8, 2015, from http://woosterhistory.org/about

Centrend. (n.d.). Online Backup. Retrieved March 2, 2016, from http://centrend.com/index.php/managed-services/online-backup

Connolly, B. C., DiMaggio, P. N., Grande, D. (2015). *Digitizing Worcester's History* (Undergraduate Interactive Qualifying Project No. E-project-051415-130204). Retrieved from Worcester Polytechnic Institute Electronic Projects Collection: http://www.wpi.edu/Pubs/E-project/Available/E-project-051415-130204/

Creating Holyoke | About Us. (2009). Retrieved October 8, 2015, from http://www.creatingholyoke.org/about

Farris, P. W., Bendle, N. T., Pfeifer, P. E., & Reibstein, D. J. (2010). Marketing Metrics: The Definitive Guide to Measuring Marketing Performance (2nd ed.). Upper Saddle River, NJ, NJ: FT Press.

Filloux, F. (2011, June 5). Analyzing the metered model. *Monday Note*. Retrieved February 29, 2016, from http://www.mondaynote.com/2011/06/05/analyzing-the-metered-model/

Historic Pittsburgh City Directories. (2014). Retrieved October 8, 2015, from http://digital.library.pitt.edu/cgi-bin/t/text/text-idx?c=pitttextdir

Lowcountry Digital History Initiative | About. (n.d.). Retrieved October 8, 2015, from http://ldhi.library.cofc.edu/about

Mass Humanities. (n.d.). Grants. Retrieved February 29, 2016, from http://masshumanities.org/grants/

Mass. Cultural Council. (n.d.). Cultural Investment Portfolio. Retrieved February 29, 2016, from http://www.massculturalcouncil.org/applications/cip_portfolio_guidelines.asp

National Archives. (2016). Access to Historical Records. Retrieved March 2, 2016, from http://www.archives.gov/nhprc/announcement/access.html

National Endowment for the Humanities (NEH). NEH Grants. (n.d.). Retrieved February 24, 2016, from http://www.neh.gov/grants/

North Carolina City Directories. Retrieved October 8, 2015, from http://www.digitalnc.org/collections/city-directories/

Omeka.net. About. (n.d.). Retrieved March 1, 2016, from http://info.omeka.net/about/

Omeka.org. (n.d.). Project. Retrieved March 1, 2016, from http://omeka.org/about/

Pasadena Digital History Collaboration. (n.d.). Pasadena City Directories. Retrieved October 8, 2015, from http://collection.pasadenadigitalhistory.com/cdm/landingpage/collection/p16237

PricewaterhouseCoopers LLP. (2015, April). IAB internet advertising revenue report: 2014 full year results. Retrieved February 25, 2016, from http://www.iab.com/wp-content/uploads/2015/05/IAB_Internet_Advertising_Revenue_FY_2014.pdf

Rosenzweig, R. (1983). Eight Hours For What We Will: Workers and leisure in an industrial city, 1870-1920. Cambridge [Cambridgeshire];New York;: Cambridge University Press.

Sonn, D. (2011, November 18). Projecting Online Ad Revenue. [Web log]. Retrieved March 1, 2016, from http://blogs.imediaconnection.com/blog/2011/11/18/projecting-online-ad-revenue/

Statistic Brain Research Institute. (2015, March 3). Average Revenue from PPC Banner Ads. Retrieved March 1, 2016, from http://www.statisticbrain.com/average-publisher-revenue-from-ppc-banner-advertisements/

Sutton, P. (2012, June 8). Direct Me NYC 1786: A History of City Directories in the United States and New York City. Retrieved September 30, 2015, from http://www.nypl.org/blog/2012/06/08/direct-me-1786-history-city-directories-US-NYC

Worcester Arts Council. (n.d.). Retrieved February 29, 2016, from http://www.worcesterma.gov/development/cultural-development/worcester-arts-council

Zwilling, M. (2012, March 12). Website Ads are Not a Revenue Stream for Startups. [Web log]. Retrieved March 1, 2016, from http://blog.startupprofessionals.com/2012/03/website-ads-are-not-revenue-stream-for.html

# Appendix A: File Splitting and File Naming Conventions

Our current directory scanning process creates PDFs containing approximately 150 pages, which we combine into a single, large file containing the entire directory. We are also using Omeka's PDF Embed plug-in to allow users to automatically view the actual scans without having to download the files themselves. Embedding the file will also greatly decrease the chance a user will attempt to use the text created by OCR, which lacks the formatting of the physical page.

Omeka.net, who is hosting Digital Worcester, places a maximum limit of 64 MB on uploaded files. Our current practice creates batches of approximately a third that size (20-25MB). Thus, a single file containing the entire directory is infeasible for most directories. Files of the size of our batches (~150 pages) are also not preferential because of the time taken to load the file, especially on each directory's item page in Omeka, where PDF Embed will attempt to show all of the constituent files that make up an item (larger directories can be made of 6 or more large batches).

However, we assume that the majority of users will not browse to the item page and will instead use the search feature. Searches in Digital Worcester currently return the individual files as results. When viewing a file with PDF Embed, it will only show that file. In this case we should still avoid the large, 150-page files, as they still take a long time to load, especially on a slower Internet connection.

Thus we have decided to continue the practices of last year's group, which split the files based on the section of the directory they represented (examples: Names A-C, Advertisements). These sections tend to be approximately 30 pages in length, reducing load times greatly. We will now split the large scan batches into these smaller sections before uploading them. It also follows from this decision that files should be named according to what section they contain (of the form: year_pgxx-xx_descriptionX-Y.pdf). The part of each directory before the names is difficult to name in this way, as it contains advertisements, charts of city and town populations, directory notes, city government officials, and street locations, among other bits of information about Worcester and the surrounding towns. Last year's group chose not to name their files with a description besides the year and page numbers because of this. For this section, we can also leave it undescribed, or we can name it as "intro", "misc", or "cityinfo" to reflect its position in the book or its contents.

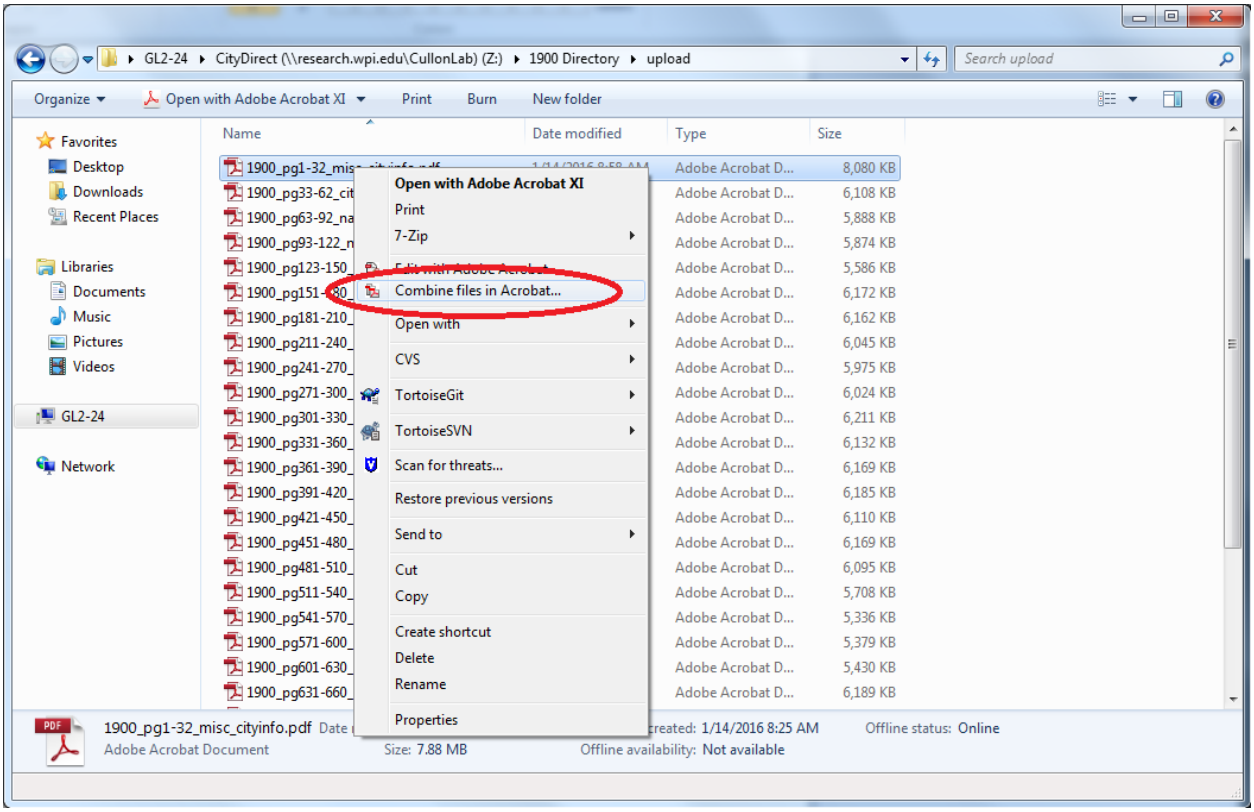
Sample directory file utilizing the above described naming scheme:

| | Notes on naming |
|---|---|
| 1900_pg1-30_cityinfo.pdf | |
| 1900_pg31-60_cityinfo_namesA.pdf | If page contains multiple categories |
| 1900_pg61-90_namesA-G.pdf | |
| 1900_pg91-120_namesG-P.pdf | |
| 1900_pg121-150_namesP-Z.pdf | |
| 1900_pg151-180_namesZ_busA-H.pdf | bus here means Business Directory |
| 1900_pg181-210_busI-S.pdf | |
| 1900_pg211-240_busS-Z_adverts1.pdf | Numbering the advertisement pages |
| 1900_pg241-270_adverts2.pdf | might be helpful |
| 1900_pg271-300_adverts3.pdf | |

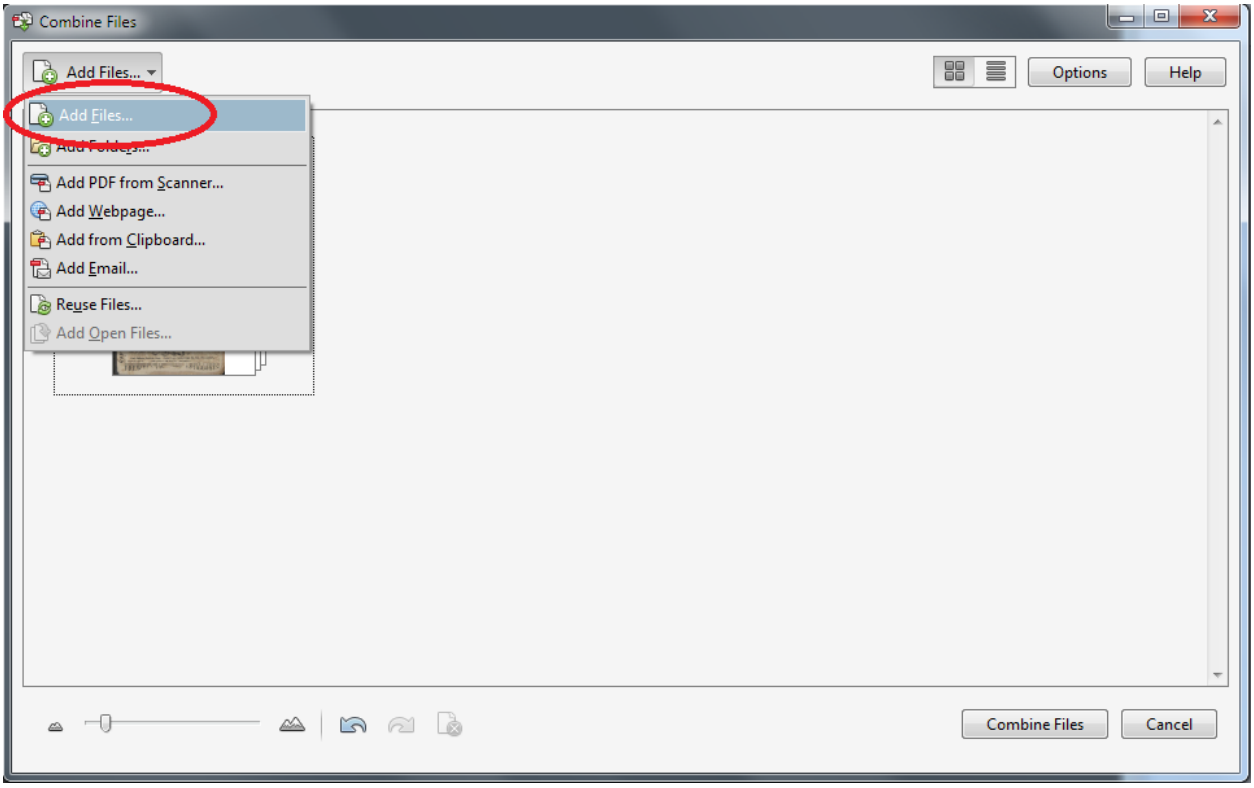
# Appendix B: File Splitting, OCR, and Upload Procedure

Combining files (to add separately scanned pages or covers to the batch-scanned files)

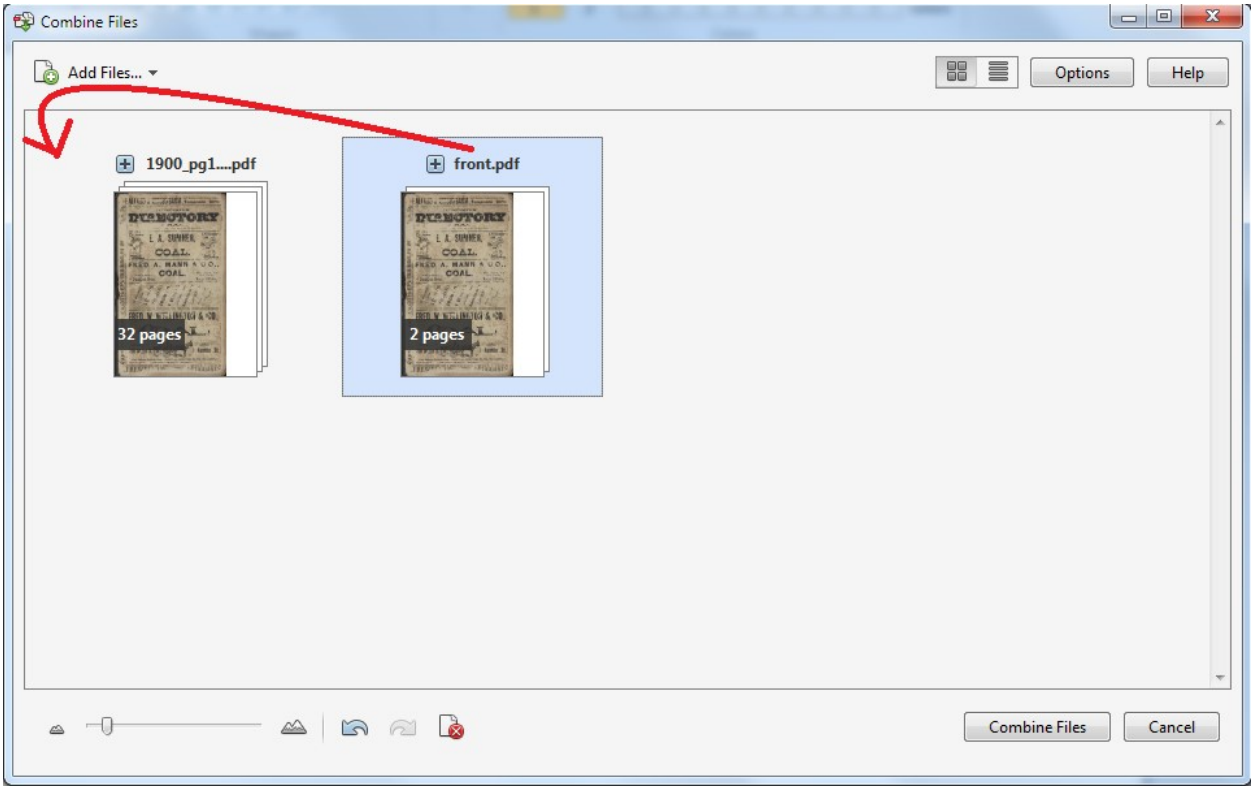
Select file to add to:



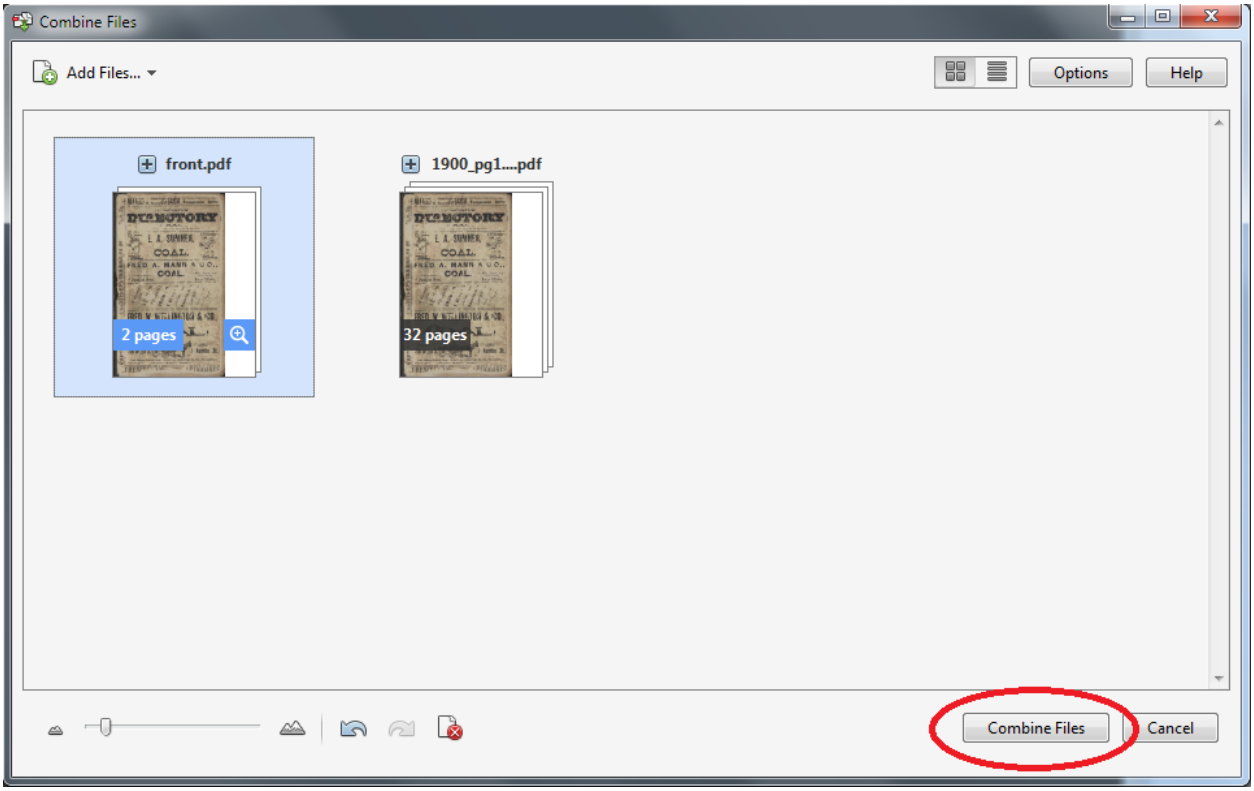
Select other file(s) to combine with this file:



Drag-and-drop the files into the proper order; files can be expanded by clicking on the + icon next to the file name to allow for the drag-and-dropping of individual pages into the correct order, if need be:

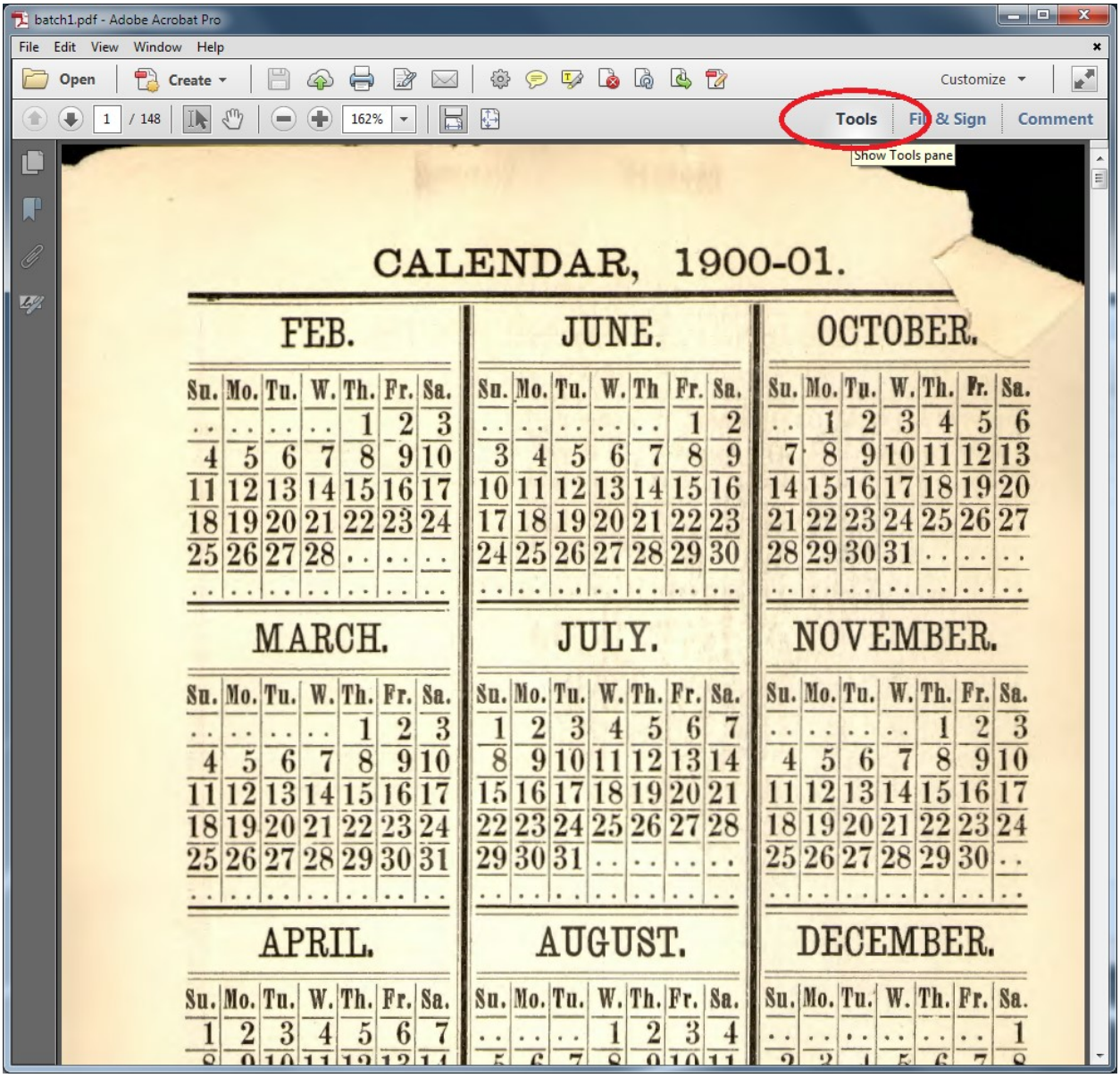


Select “Combine Files” to combine the files in the displayed order, which will create a new file that can be saved where need be:

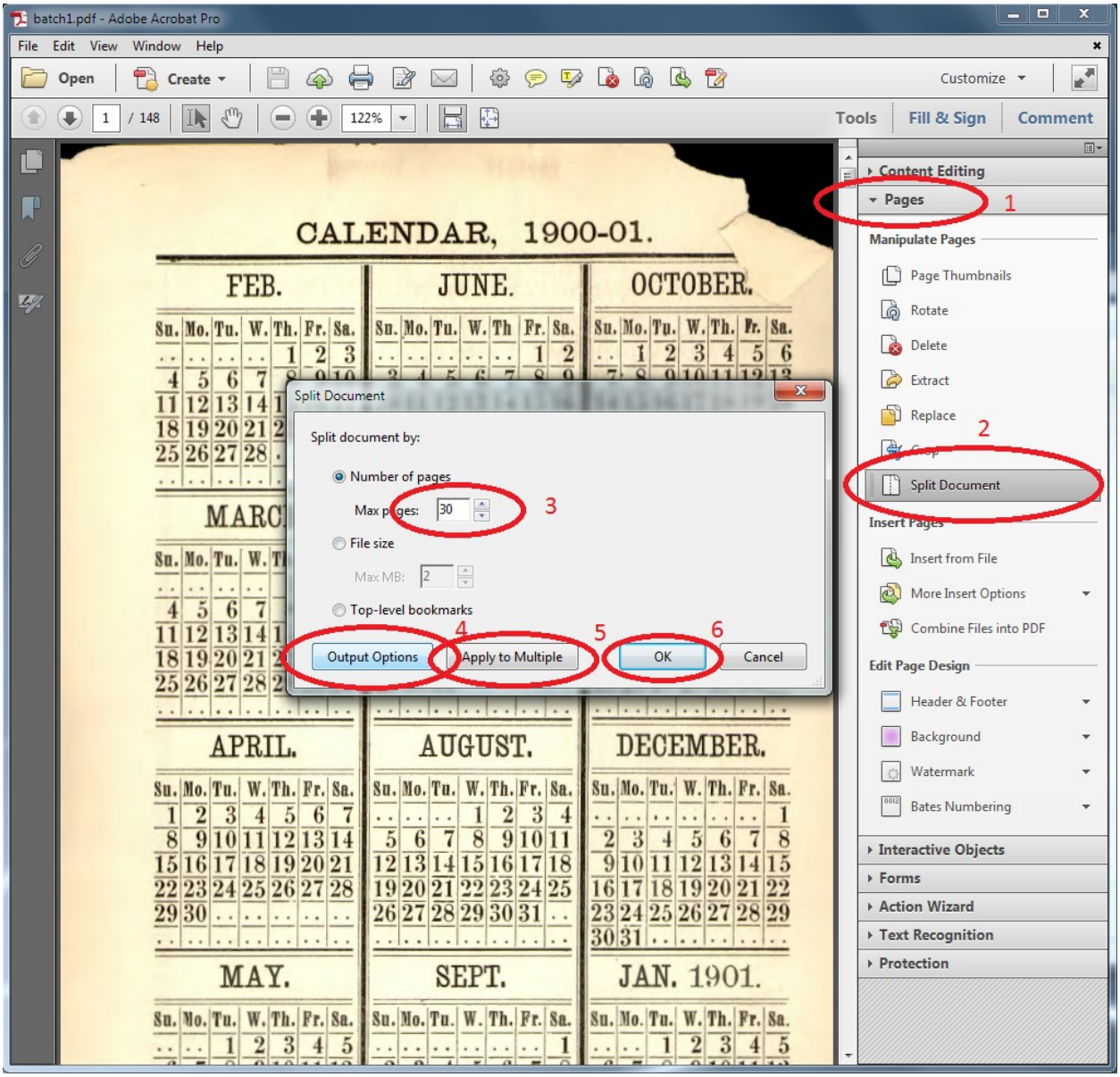


## Splitting documents into upload sized files

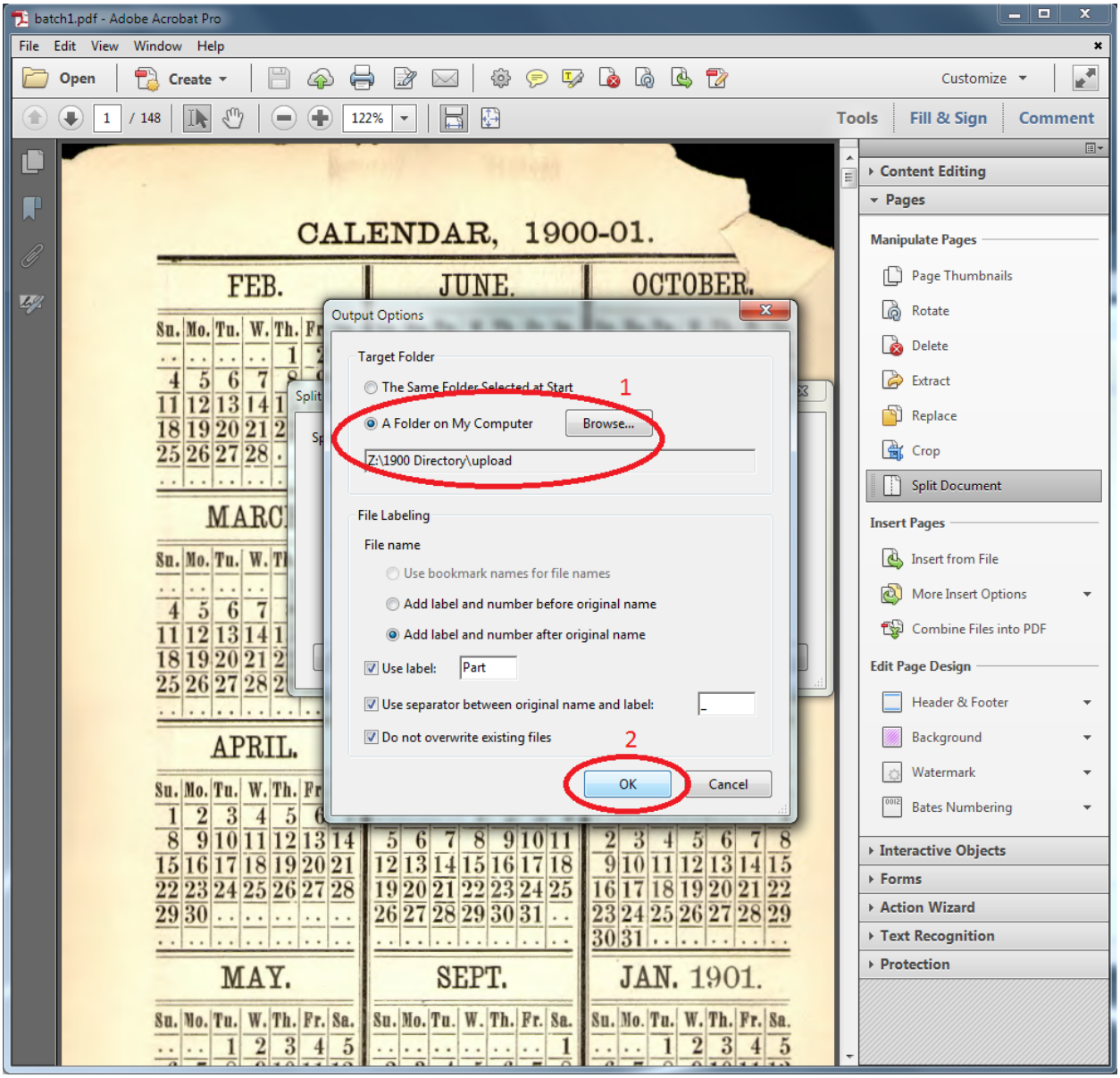
Open the “Tools” menu:



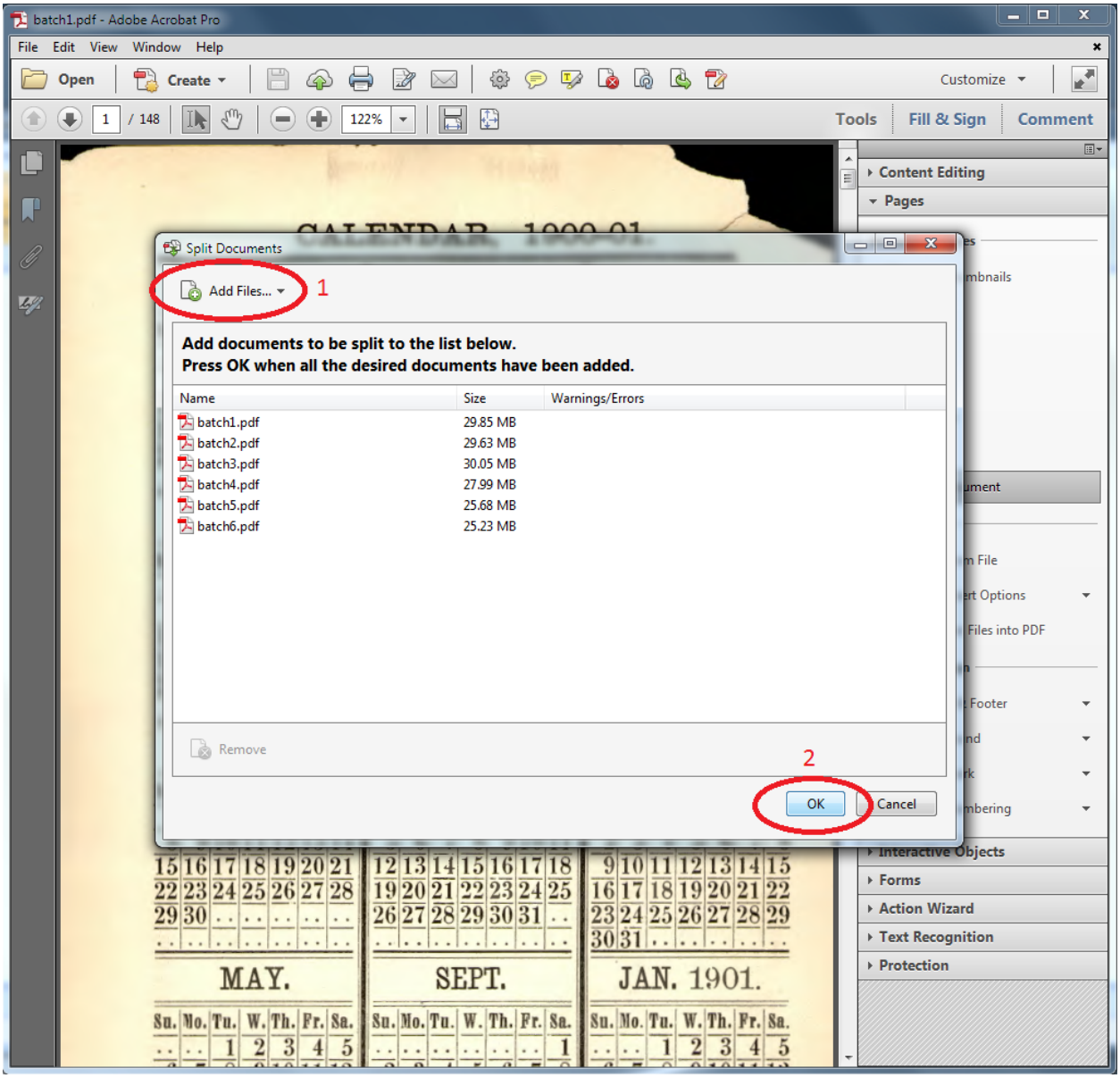
Open the “Pages” tab (1), select “Split Document” (2), set the maximum number of pages to have in each new file (3), select “Output Options” (4, see below for option configuration), select “Apply to Multiple” (5, see below output option configuration for details), and finally “OK” (6), at which point the new files will be created:



Choose the folder to store the upload files (1), press “OK” to confirm and return to above picture:

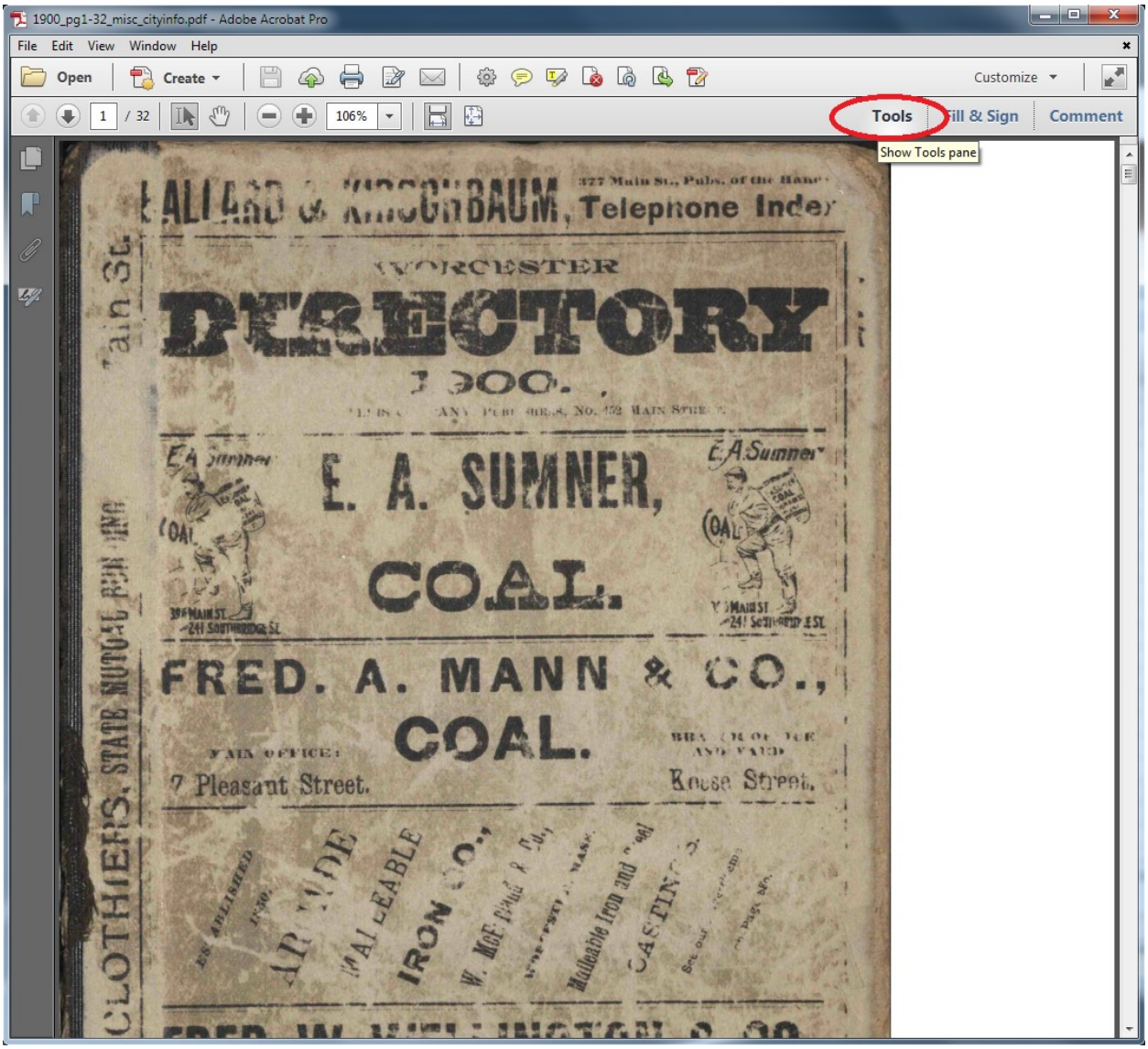


Select all files to apply the splitting to (for directory it should be all of the batches) (1), press “OK” to continue here and in the next dialog window:

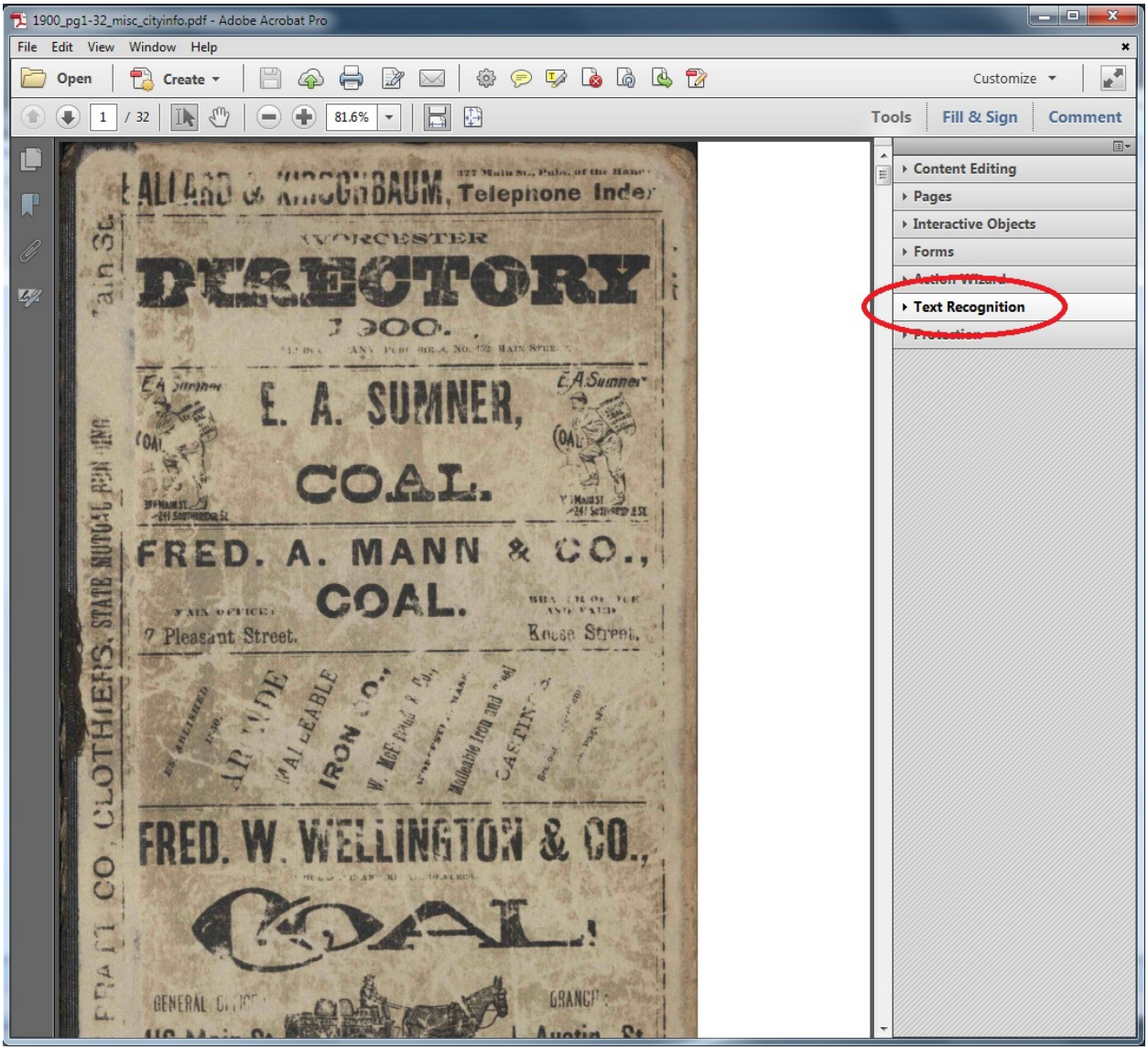


## Running OCR on the split files

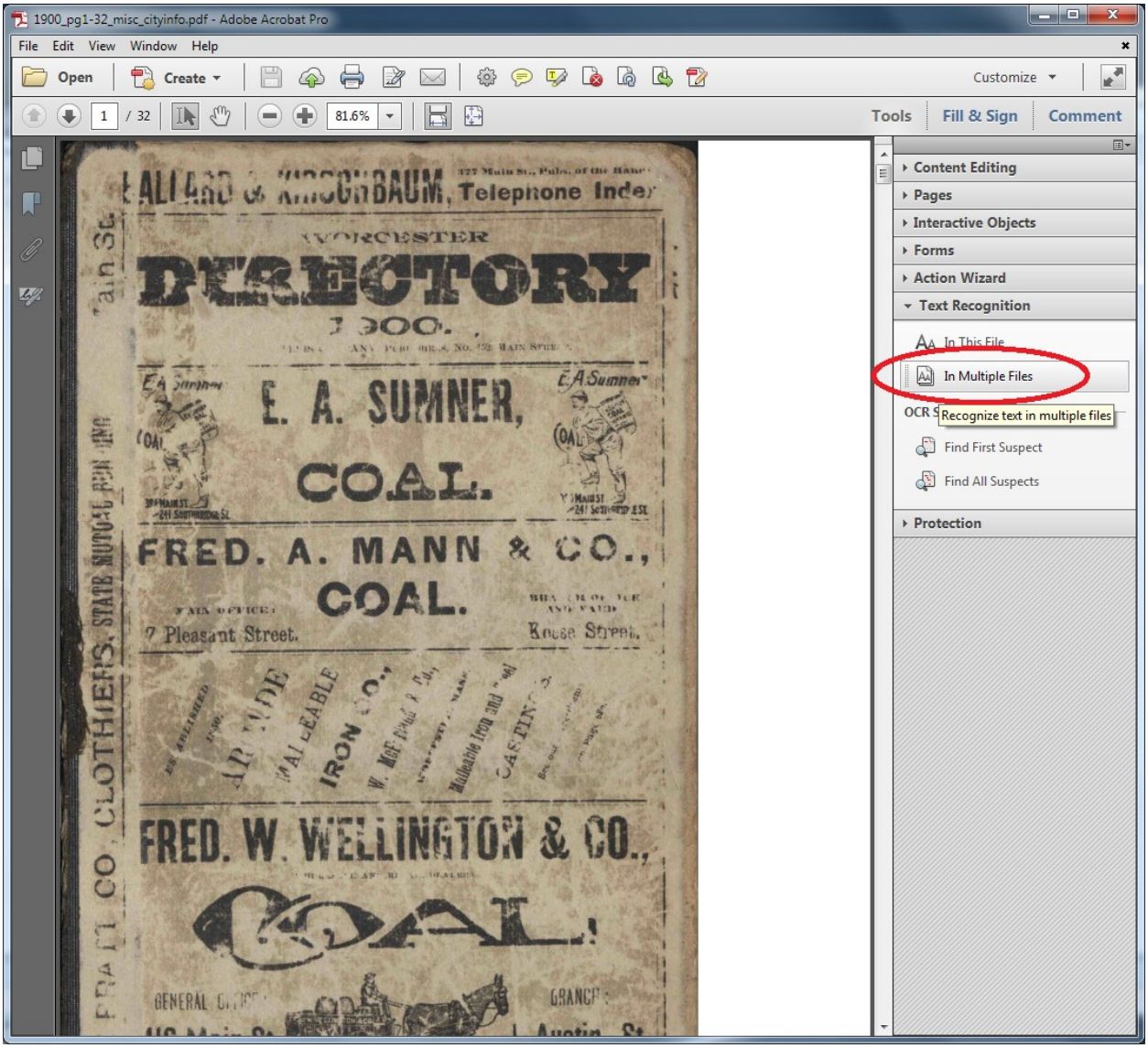
Open “Tools” menu:



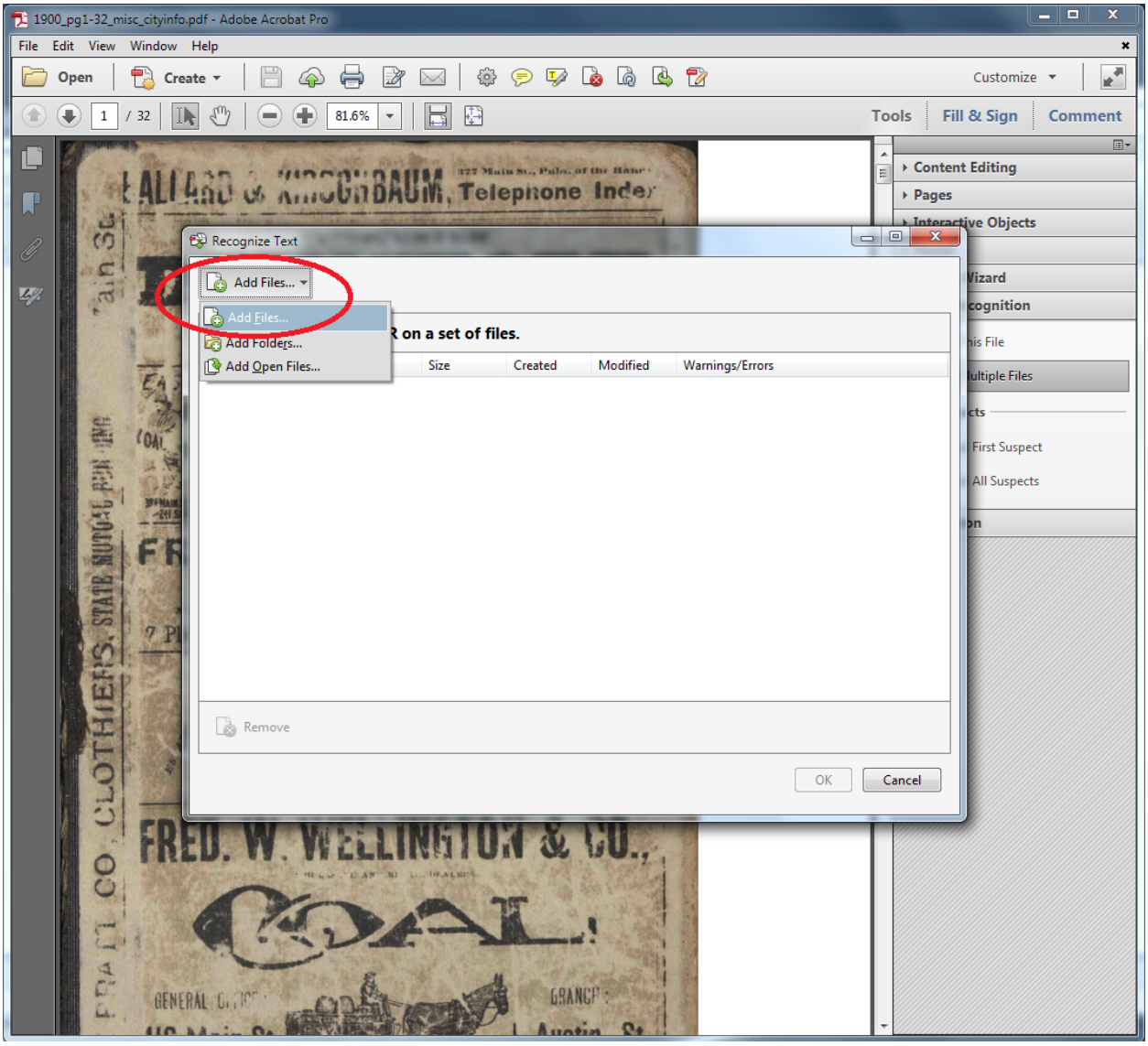
Expand the “Text Recognition” tab:



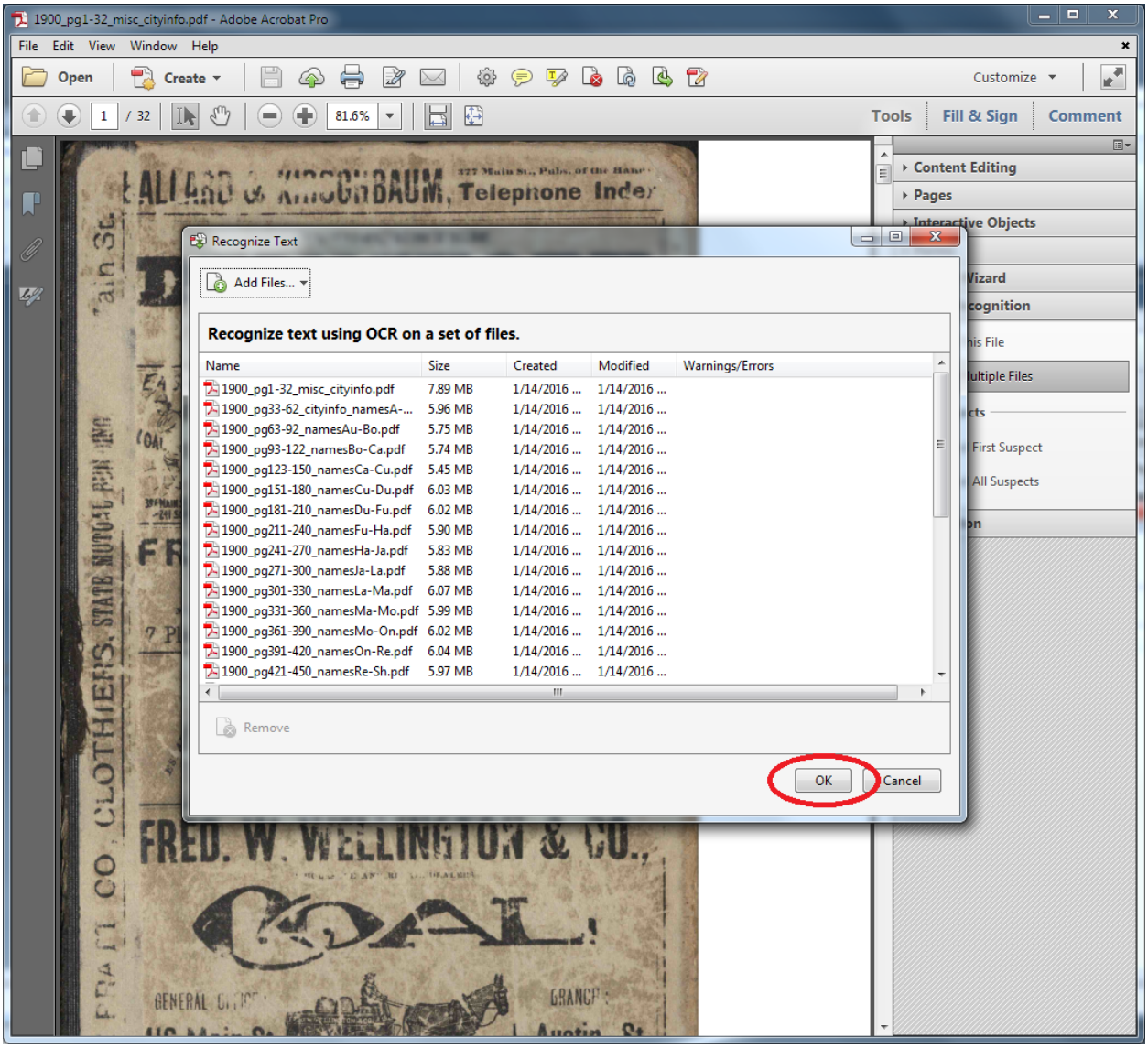
Select “In Multiple Files”:



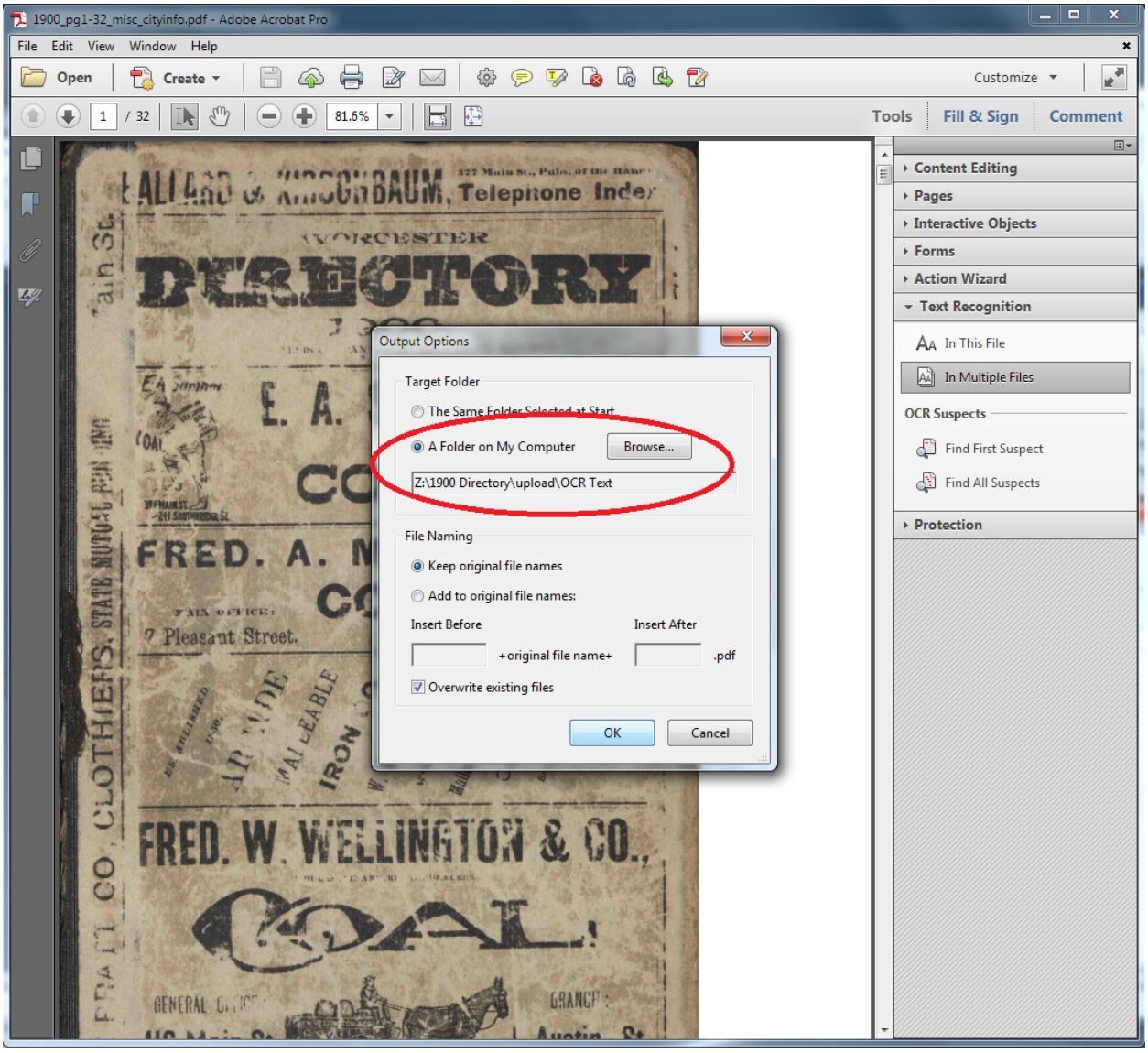
Add files to run OCR on (should be all of the split files that make up the directory):



Once all files are added, press “OK” to continue:

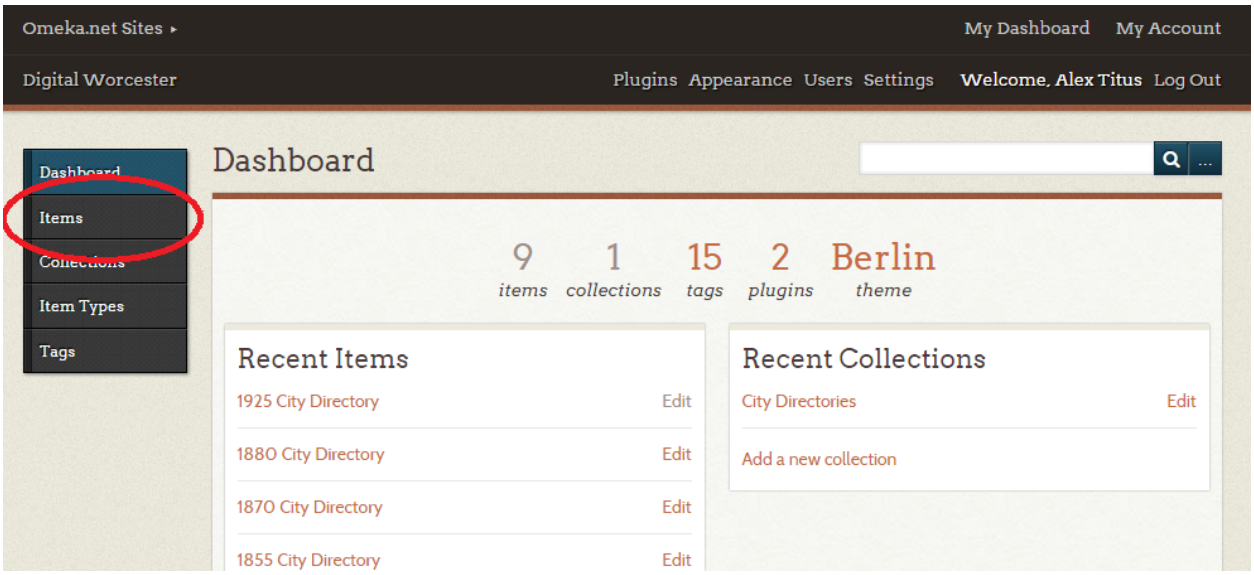


Select the same folder that the split documents are in, leaving “Keep original file names” and “Overwrite existing files” selected, press “OK” to begin OCR (and take a coffee break):

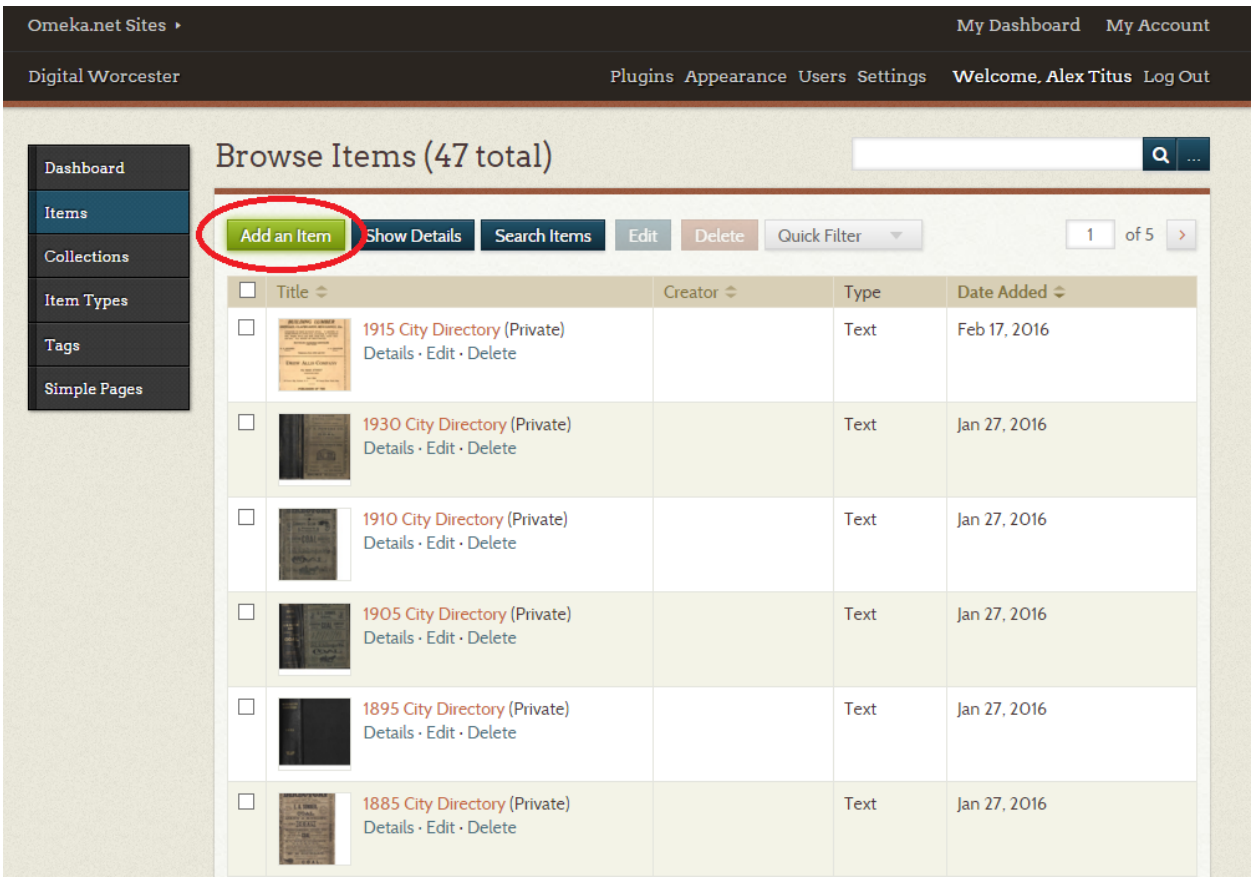


## Creating Item and Uploading Files in Omeka

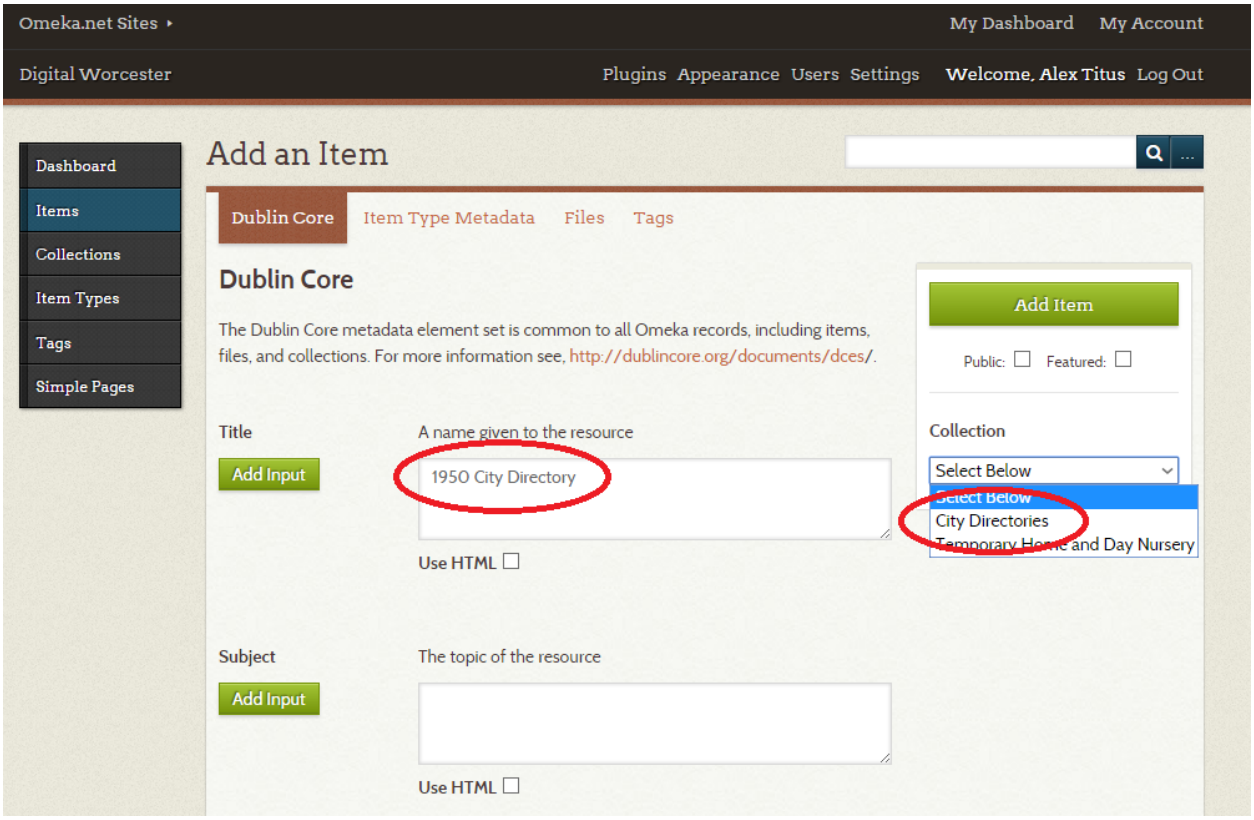
From dashboard, select Items:



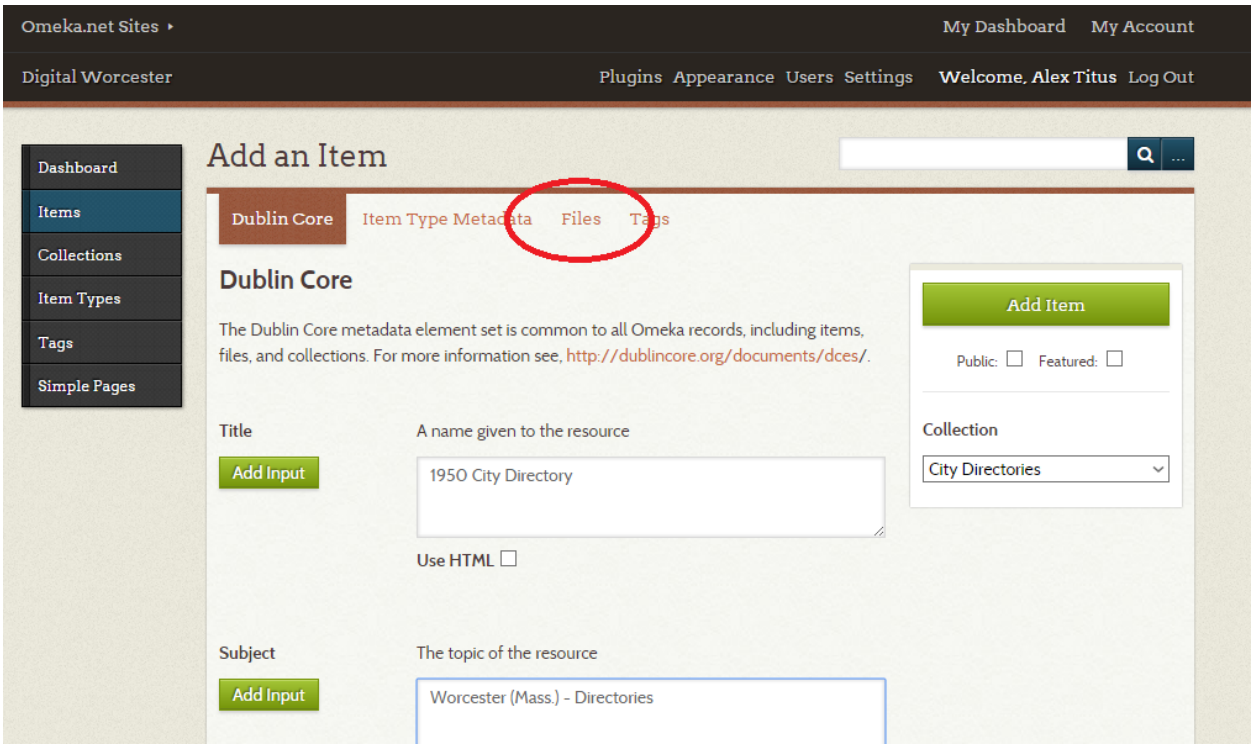

Select “Add an Item”:



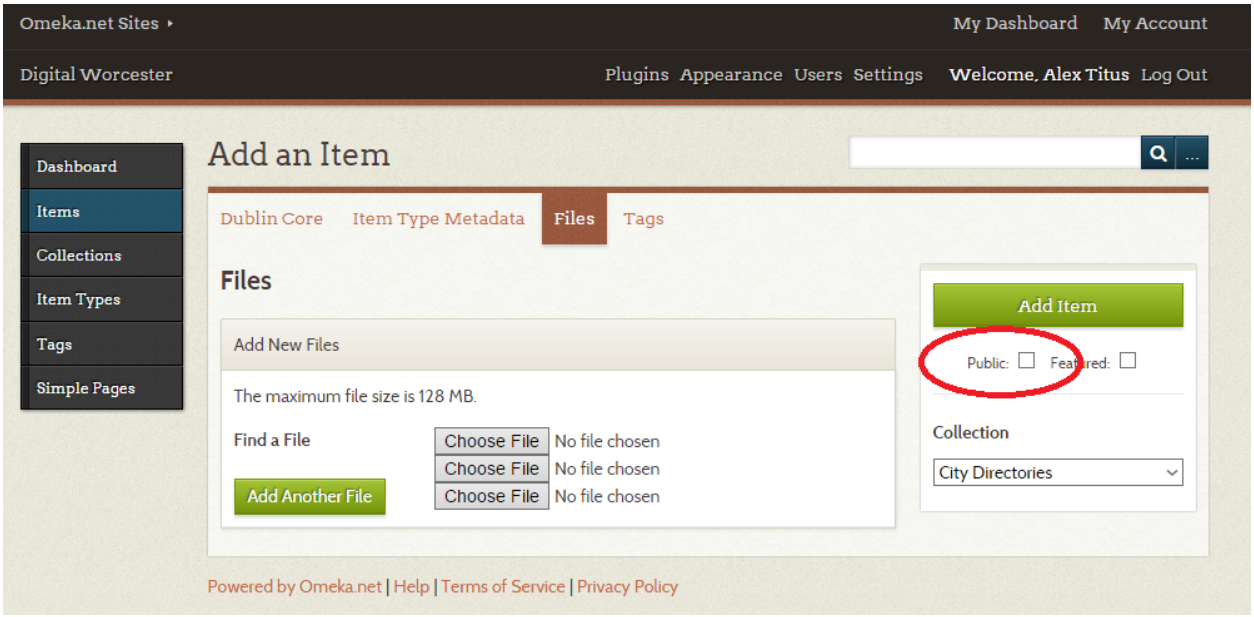
Fill in Metadata fields and select the Collection the Item will belong to:



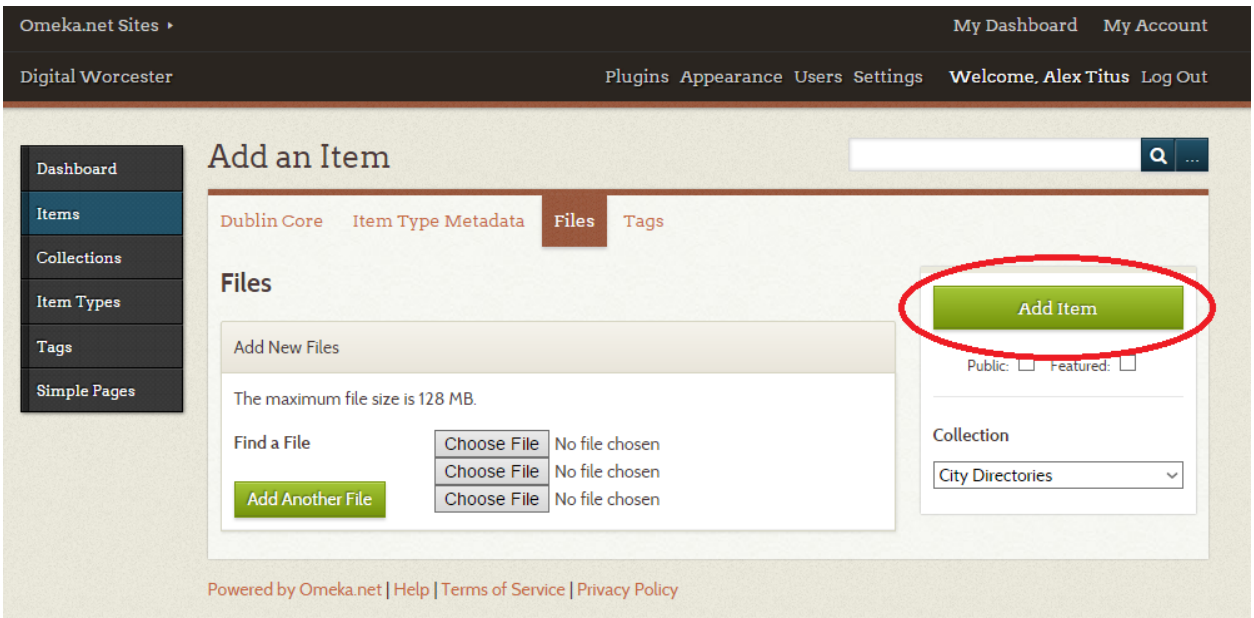

Select “Files” to add files to the item:



Select files, using “Add Another File” to upload multiple files at once (upload file size, which is different from the shown maximum file size, applies and will limit how many files you can upload at once; if/when that happens, just edit the newly created item and add the additional files); upload won’t begin until “Add Item” is selected; check the “Public” box (circled) to make the item accessible to the public from the website:




Select “Add Item” to begin file upload and create new item:



# Appendix C: Metadata


Available Metadata Fields in Omeka and Values to Use

Dublin Core Extended Fields for Item (Single Directory)

* Indicates values currently used

| |
|---|
| Name of Directory (YYYY City Directory)* |
| Worcester (Mass.) - Directories* |
| Page count* |
| Printing Company |
| Nothing |
| Printing Company or the WHM |
| Year of Directory* |
| Nothing |


Based on examples in Figure 2-2

| |
|---|
| Nothing |
| PDF* |
| English* (or eng/en-US) |
| Text* |
| ISBN or WHM Identifier |
| Worcester, Massachusetts* (or n-us-ma) |
| Nothing |
| Nothing |

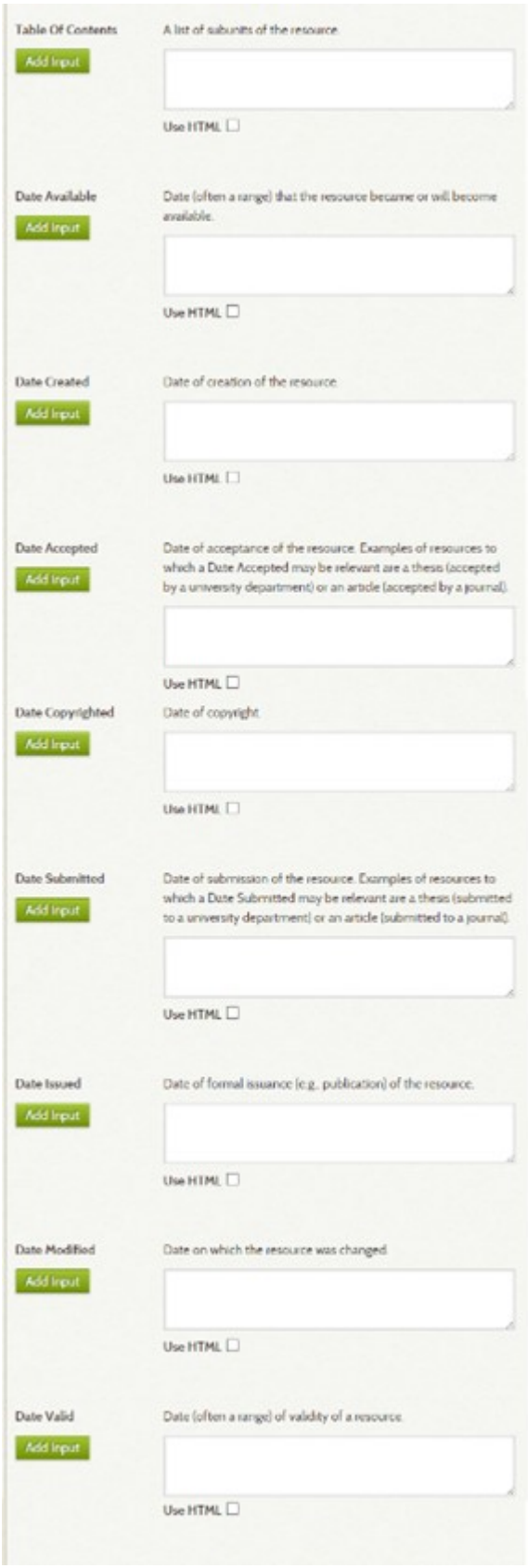
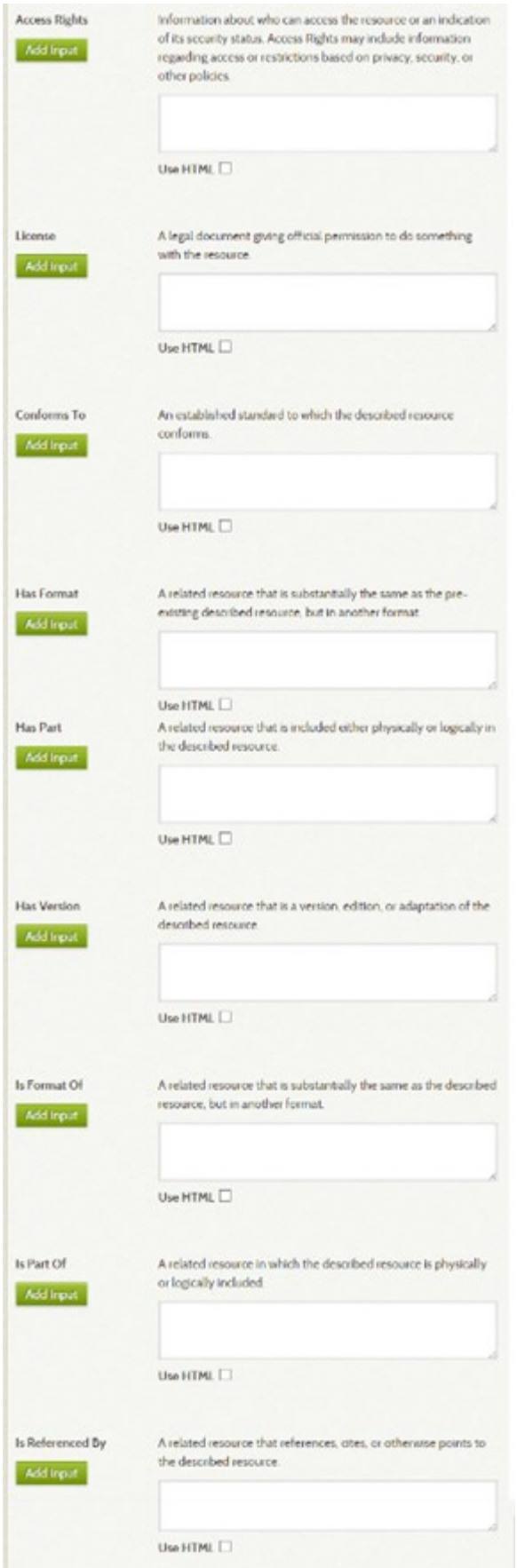


Table Of Contents
A list of subunits of the resource.
Add Input
Use HTML
Date Available
Date (often a range) that the resource became or will become available.
Add Input
Use HTML
Date Created
Date of creation of the resource.
Add Input
Use HTML
Date Accepted
Date of acceptance of the resource. Examples of resources to which a Date Accepted may be relevant are a thesis (accepted by a university department) or an article (accepted by a journal).
Add Input
Use HTML
Date Copyrighted
Date of copyright.
Add Input
Use HTML
Date Submitted
Date of submission of the resource. Examples of resources to which a Date Submitted may be relevant are a thesis (submitted to a university department) or an article (submitted to a journal).
Add Input
Use HTML
Date Issued
Date of formal issuance (e.g., publication) of the resource.
Add Input
Use HTML
Date Modified
Date on which the resource was changed.
Add Input
Use HTML
Date Valid
Date (often a range) of validity of a resource.
Add Input
Use HTML

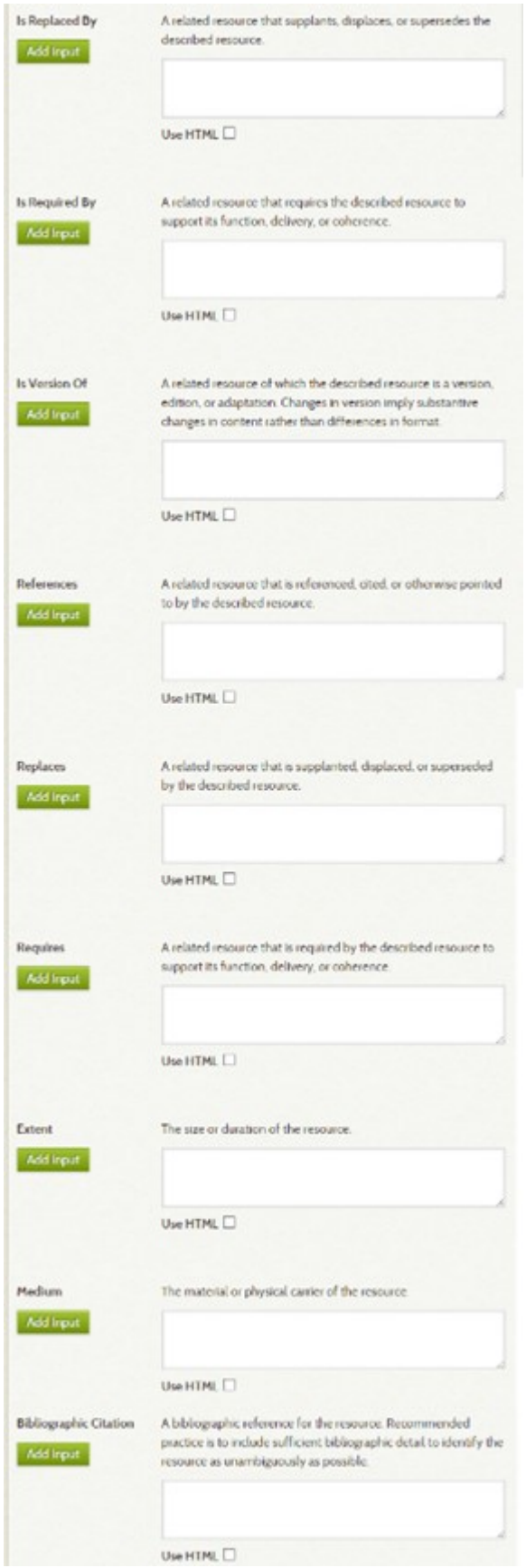
| |
|---|
| Nothing |
| Date of publication |
| Date of upload* |
| Nothing |
| Year directory was published* |
| Nothing |
| Date published or when site is made public |
| Nothing |

Year of Directory or Nothing



| |
|---|
| Nothing (public access) |
| License document or Nothing |
| Dublin Core, LoC, or Nothing |
| Nothing |
| Nothing |
| Nothing |
| Nothing |
| Nothing |

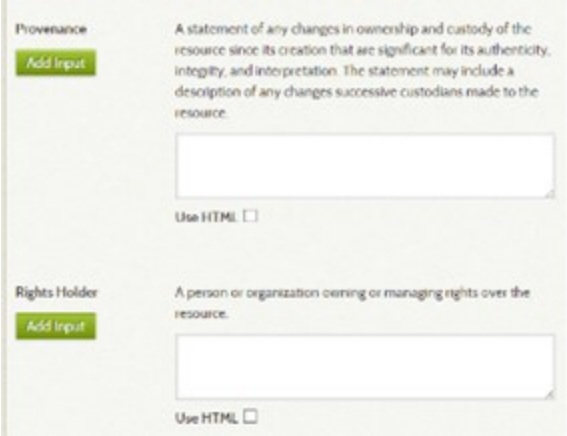
| |
|---|
| Nothing |


Is Replaced By
Add Input
A related resource that supplants, displaces, or supersedes the described resource.
Use HTML
Is Required By
Add Input
A related resource that requires the described resource to support its function, delivery, or coherence.
Use HTML
Is Version Of
Add Input
A related resource of which the described resource is a version, edition, or adaptation. Changes in version imply substantive changes in content rather than differences in format.
Use HTML
References
Add Input
A related resource that is referenced, cited, or otherwise pointed to by the described resource.
Use HTML
Replaces
Add Input
A related resource that is supplanted, displaced, or superseded by the described resource.
Use HTML
Requires
Add Input
A related resource that is required by the described resource to support its function, delivery, or coherence.
Use HTML
Extent
Add Input
The size or duration of the resource.
Use HTML
Medium
Add Input
The material or physical carrier of the resource.
Use HTML
Bibliographic Citation
Add Input
A bibliographic reference for the resource. Recommended practice is to include sufficient bibliographic detail to identify the resource as unambiguously as possible.
Use HTML

| |
|---|
| Nothing |
| Nothing |
| Nothing |
| Nothing |
| Nothing |
| Nothing |
| Page count, dimensions |
| Paper |
| Based on chosen format (MLA, APA, etc.) |


Spatial Coverage
Spatial characteristics of the resource.
Add Input
Use HTML
Temporal Coverage
Temporal characteristics of the resource.
Add Input
Use HTML
Accrual Method
The method by which items are added to a collection.
Add Input
Use HTML
Accrual Periodicity
The frequency with which items are added to a collection.
Add Input
Use HTML
Accrual Policy
The policy governing the addition of items to a collection.
Add Input
Use HTML
Audience
A class of entity for whom the resource is intended or useful.
Add Input
Use HTML
Audience Education Level
A class of entity, defined in terms of progression through an educational or training context, for which the described resource is intended.
Add Input
Use HTML
Mediator
An entity that mediates access to the resource and for whom the resource is intended or useful. In an educational context, a mediator might be a parent, teacher, teaching assistant, or care-giver.
Add Input
Use HTML
Instructional Method
A process, used to engender knowledge, attitudes and skills, that the described resource is designed to support. Instructional Method will typically include ways of presenting instructional materials or conducting instructional activities, patterns of learner-to-learner and learner-to-instructor interactions, and mechanisms by which group and individual levels of learning are measured. Instructional methods include all aspects of the instruction and learning processes from planning and implementation through evaluation and feedback.
Add Input

| |
|---|
| Nothing |
| Nothing |
| Nothing |
| Nothing |
| Nothing |
| Nothing |
| Nothing |
| Nothing |
| Nothing |
| Nothing |




| |
|---|
| Nothing |
| Depends on Rights value, could be WHM, publishing company, or Nothing |

## File (1 Batch of Scans) Dublin Core Extended


File metadata is different from Item metadata only in the Title and Description fields (shown below)

^ Indicates different value than Item metadata

| |
|---|
| Name of File^ |
| Worcester (Mass.) - Directories |
| Indicate that it is a file^ |
| Printing Company |
| Nothing |
| Printing company or WHM |
| Year of Directory |
| Nothing |

# Appendix D: Data Management Policy

Our data management policy will fit our data needs in the following ways. The location of storage is a shared WPI research drive directory. We will be working with PDF scans of books, which will range from 100 megabytes to 1 gigabyte per book, depending on its size and page count. The files can be accessed by team members only at the moment, as a security precaution. The storage will need to be maintained until all directories are scanned. The data will be organized into subdirectories, with subdirectories having a constant naming scheme of

YYYY  Directory, where YYYY is the year expressed with four digits, (e.g. 1865 Directory). WPI servers are backed up to an off--campus location, in case of disaster, so the data is well protected  against loss. Long--term data access will be provided via Professor Cullon and the WHM.