Collections as Data in Texas Digital Library Repositories:

Practical Recommendations for Use

Karla Roig Blay, Anne Morgan, and Katy Tuck

School of Information, University of Texas at Austin

Collections as Data in Texas Digital Library Repositories:

Practical Recommendations for Use

This document provides practical recommendations for the use of Collections as Data in Texas Digital Library (TDL) repositories. The document presents recommendations for how to undertake a Collections as Data project using the Texas Conference on Digital Libraries (TCDL) Proceedings as a case study. The recommendations were written by three graduate students at the University of Texas at Austin's School of Information as part of a semester-long group project with TDL. The purpose of this document is to increase understanding and improve methods of using Collections as Data for TDL members.

What is Collections as Data?

In "On a Collections as Data Imperative," Thomas Padilla (2017, p. 1) describes Collections as Data as "reframing all digital objects as data," defined as "ordered information, stored digitally, that is inherently amenable to computation." In a sense, then, Collections as Data is a "mindset" (Wittmann et al., 2019, p. 51): viewing collections of digital objects in terms of discrete pieces of information that can be analyzed and used in meaningful ways. On a more practical level, Collections as Data involves the application of computational methods in order to analyze and use collections of digital objects meaningfully. These computational methods include text mining, data visualization, mapping, image analysis, audio analysis, and network analysis (Padilla et al., 2019, May 20b, p. 2).

The *Always Already Computational: Collections as Data* project, which began in 2016, produced guidance for institutions interested in working with Collections as Data. The project's deliverables include the Santa Barbara Statement on Collections as Data, which outlines foundational Collections as Data principles, such as commitments to shared documentation, open access, and transparency. *Always Already Computational* also produced a set of 50 recommendations intended to provide practitioners with initial steps to take in order to begin supporting Collections as Data at cultural heritage institutions. In 2018, *Always Already Computational* was succeeded by a new project, *Collections as Data: Part to Whole*, which has focused on developing models for the implementation and use of Collections as Data.

Using collections as data is an example of a digital humanities project. The field of digital humanities (DH) broadly refers to a certain methodological approach involving the application of technology tools to digital (or digitized) scholarship in order to analyze, visualize, interpret, and present this information in new ways. It is characterized by an emphasis on collaboration, interpretation, openness, interoperability, and innovation and can be applied to multiple disciplinary fields toward a variety of purposes. Examples of DH methods and uses include “text analysis, data mining, visualization, modeling and simulation, geospatial analysis and mapping, multimedia storytelling, information design, network analysis, interface design, and mark-up…[and applying] these tools to humanistic questions.” (“About Us - Digital Humanities,” n.d.). Many of these methods overlap with the CAD methods of inquiry mentioned above so using collections as data falls squarely into this category of digital humanities.

What are the Benefits and Uses of Using Collections as Data?

The following are potential uses of Collections as Data:

- Discover patterns in data
- Present data in a new, interesting way
- Attract new users and drive site traffic
- Create compelling visualizations of datasets to tell a story about your collection(s)

These potential uses point to two main benefits of using Collections as Data: increased useability and visibility of collections. Collections as Data moves “beyond traditional use,” allowing “more flexible access” to collections (Wittmann et al., 2019, p. 49).

In addition, Padilla discusses how a Collections as Data approach allows for new ways of understanding and deriving meaning from digital objects, including the various contexts in which they are embedded (2017, p. 1-2). Padilla also emphasizes Collections as Data's active engagement with ethical issues, such as inclusivity and bias (2016). According to Elizabeth Russey Roke, Collections as Data approaches apply the key archival principles of transparency, documentation, and provenance, which enhance understanding of digital objects (2019, July 16).

TDL Repositories and the TCDL Proceedings

Texas Digital Library (TDL) and its member repositories use DSpace, an open source repository application. The Texas Conference on Digital Libraries (TCDL) proceedings are stored in the TDL DSpace repository. This collection includes presentations, posters and other materials from the annual Texas Conference on Digital Libraries. The proceedings were selected for use as a case study because the project contact at TDL indicated an interest in having the group work with this collection.

Process

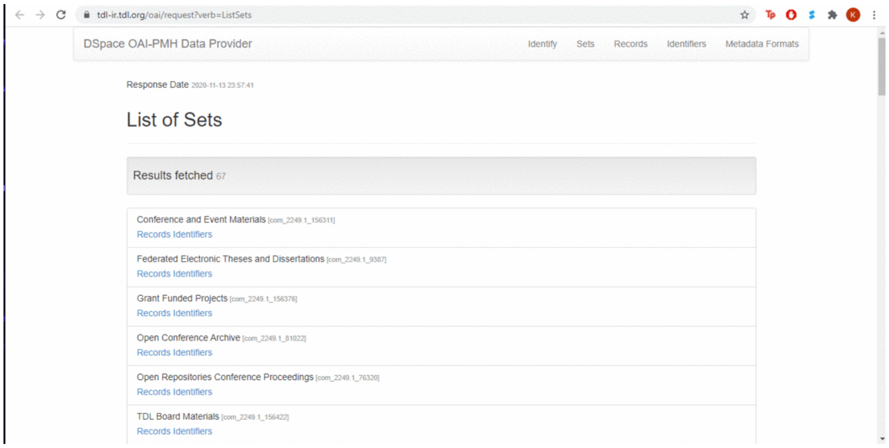
The task presented to our group was the following: collect metadata from the TCDL proceedings, clean the data, and analyze it computationally. We approached our Collections as Data project by dividing it into a three-part process: metadata extraction, metadata cleaning, and metadata visualization. We divided the work equally and worked together in weekly group meetings over Zoom to walk through each step of the process so we could learn together. We also broke down the written work into sections so each member could contribute equitably. Karla Roig has previous experience working with the LLILAS Benson collections on a Digital Humanities project and was able to contribute her knowledge of this subject to enhance our understanding of steps we needed to take. In addition, Karla's sister, Aleshka Blay, provided the group with information on how to clean the data in Excel.

Note: We were able to extract metadata from the OAI-PMH website of the TDL's Dspace repository but this process is not universal. If an archive or institution does not use this protocol, or uses another Application Programming Interface (API) that can be queried to find metadata, there are many different metadata scraping tools available (please review our list of resources and tools below).

Data Collection

The data collection process is divided into steps denoting the process we undertook to extract the desired metadata from the OAI PMH DSpace data portal. The process for extracting and cleaning the data was broken down into three main stages: downloading the XML metadata files, importing the uncleaned tabular data in Excel, and cleaning and editing the tabular data in Excel. This section shows a series of screenshots as a guide for the metadata extraction process. There are more detailed instructions in our "Step by Step Process for Extracting and Cleaning metadata" document.

Step 1: Select desired collection from the DSpace OAI-PMH portal, then select the "records" for this collection. In the List of Records page, right click on the website and select "View Page Source" to see the XML HTML source.



tdl-ir.tdl.org/oai/request?verb=ListSets

Response Date 2020-11-13 23:57:41

List of Sets

Results fetched 67

Conference and Event Materials [com_2249.1_156311]
Records Identifiers

Federated Electronic Theses and Dissertations [com_2249.1_9387]
Records Identifiers

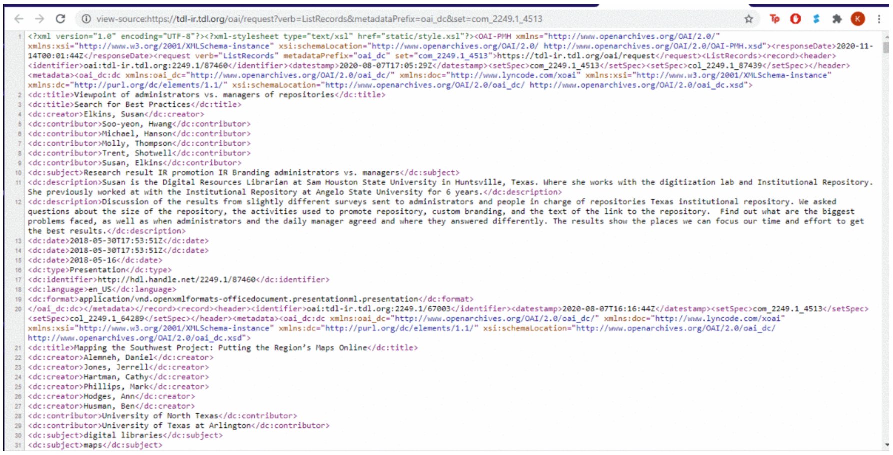
Grant Funded Projects [com_2249.1_156376]
Records Identifiers

Open Conference Archive [com_2249.1_81022]
Records Identifiers

Open Repositories Conference Proceedings [com_2249.1_76320]
Records Identifiers

TDL Board Materials [com_2249.1_156422]
Records Identifiers

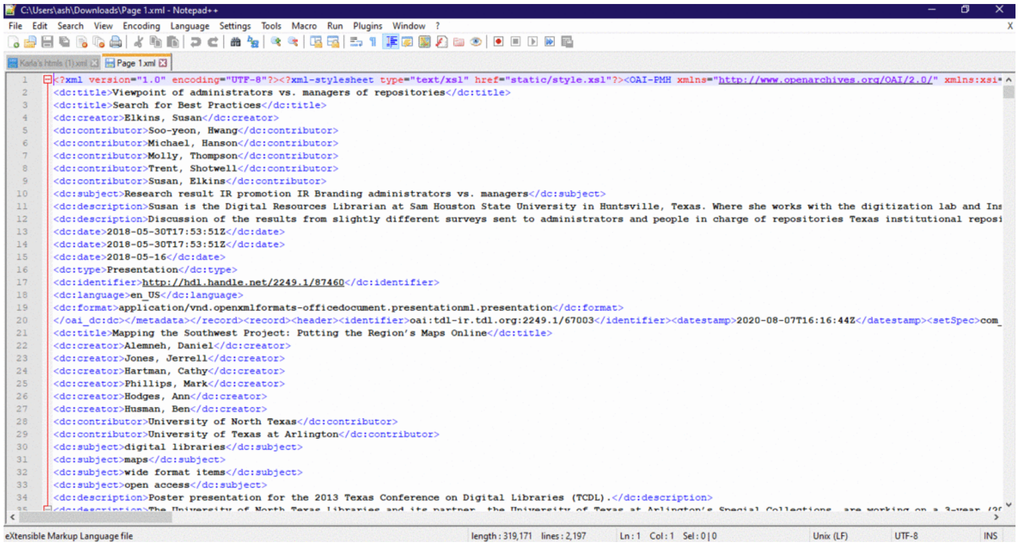
Step 2: Save this page onto your local drive by right clicking on it and selecting “Save as.” This page only shows the first one hundred items of the collection, since the OAI-PMH portal divides it as such, so we repeated the process until we captured all the data.



view-source:https://tdl-ir.tdl.org/oai/request?verb=ListRecords&metadataPrefix=oai_dc&set=com_2249.1_4513

```
<?xml version="1.0" encoding="UTF-8"?><?xml-stylesheet type="text/xsl" href="static/style.xsl"?><OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd"><responseDate>2020-11-
14T00:01:44Z</responseDate><request verb="ListRecords" metadataPrefix="oai_dc" set="com_2249.1_4513">https://tdl-ir.tdl.org/oai/request</request><ListRecords><record><header>
<identifier>oai:tdl-ir.tdl.org:2249.1/87460</identifier><datestamp>2020-08-07T17:05:29Z</datestamp><setSpec>com_2249.1_4513</setSpec><setSpec>col_2249.1_87439</setSpec></header>
<metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:doc="http://www.lyncode.com/xoai" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:dc="http://purl.org/dc/elements/1.1/" xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/ http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
<dc:title>Viewpoint of administrators vs. managers of repositories</dc:title>
<dc:title>Search for Best Practices</dc:title>
<dc:creator>Elkins, Susan</dc:creator>
<dc:contributor>Soo-yeon, Hwang</dc:contributor>
<dc:contributor>Michael, Hanson</dc:contributor>
<dc:contributor>Molly, Thompson</dc:contributor>
<dc:contributor>Trent, Shotwell</dc:contributor>
<dc:contributor>Susan, Elkins</dc:contributor>
<dc:subject>Research result IR promotion IR Branding administrators vs. managers</dc:subject>
<dc:description>Susan is the Digital Resources Librarian at Sam Houston State University in Huntsville, Texas. Where she works with the digitization lab and Institutional Repository.
She previously worked at with the Institutional Repository at Angelo State University for 6 years.</dc:description>
<dc:description>Discussion of the results from slightly different surveys sent to administrators and people in charge of repositories Texas institutional repository. We asked
questions about the size of the repository, the activities used to promote repository, custom branding, and the text of the link to the repository.  Find out what are the biggest
problems faced, as well as when administrators and the daily manager agreed and where they answered differently. The results show the places we can focus our time and effort to get
the best results.</dc:description>
<dc:date>2018-05-30T17:53:51Z</dc:date>
<dc:date>2018-05-30T17:53:51Z</dc:date>
<dc:date>2018-05-16</dc:date>
<dc:type>Presentation</dc:type>
<dc:identifier>http://hdl.handle.net/2249.1/87460</dc:identifier>
<dc:language>en_US</dc:language>
<dc:format>application/vnd.openxmlformats-officedocument.presentationml.presentation</dc:format>
</oai_dc:dc></metadata></record><record><header><identifier>oai:tdl-ir.tdl.org:2249.1/67003</identifier><datestamp>2020-08-07T16:16:44Z</datestamp><setSpec>com_2249.1_4513</setSpec>
<setSpec>col_2249.1_64289</setSpec></header><metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:doc="http://www.lyncode.com/xoai"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:dc="http://purl.org/dc/elements/1.1/" xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
<dc:title>Mapping the Southwest Project: Putting the Region’s Maps Online</dc:title>
<dc:creator>Alemneh, Daniel</dc:creator>
<dc:creator>Jones, Jerrell</dc:creator>
<dc:creator>Hartman, Cathy</dc:creator>
<dc:creator>Phillips, Mark</dc:creator>
<dc:creator>Hodges, Ann</dc:creator>
<dc:creator>Husman, Ben</dc:creator>
<dc:contributor>University of North Texas</dc:contributor>
<dc:contributor>University of Texas at Arlington</dc:contributor>
<dc:subject>digital libraries</dc:subject>
<dc:subject>maps</dc:subject>
```

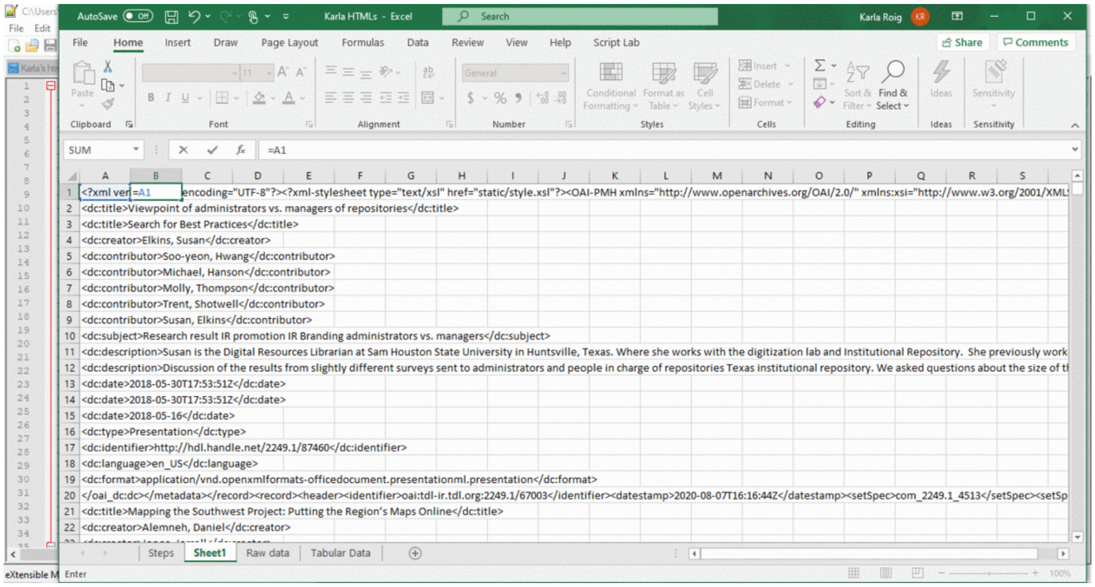
Step 3: Open xml file with Notepad++, then copy and paste it into Microsoft Excel using the Text Import Wizard.


C:\Users\ash\Downloads\Page 1.xml - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
Karla's htmls (1).xml
Page 1.xml

```
<?xml version="1.0" encoding="UTF-8"?><?xml-stylesheet type="text/xsl" href="static/style.xsl"?><OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:xsi=
<dc:title>Viewpoint of administrators vs. managers of repositories</dc:title>
<dc:title>Search for Best Practices</dc:title>
<dc:creator>Elkins, Susan</dc:creator>
<dc:contributor>Soo-yeon, Hwang</dc:contributor>
<dc:contributor>Michael, Hanson</dc:contributor>
<dc:contributor>Molly, Thompson</dc:contributor>
<dc:contributor>Trent, Shotwell</dc:contributor>
<dc:contributor>Susan, Elkins</dc:contributor>
<dc:subject>Research result IR promotion IR Branding administrators vs. managers</dc:subject>
<dc:description>Susan is the Digital Resources Librarian at Sam Houston State University in Huntsville, Texas. Where she works with the digitization lab and Ins
<dc:description>Discussion of the results from slightly different surveys sent to administrators and people in charge of repositories Texas institutional reposi
<dc:date>2018-05-30T17:53:51Z</dc:date>
<dc:date>2018-05-30T17:53:51Z</dc:date>
<dc:date>2018-05-16</dc:date>
<dc:type>Presentation</dc:type>
<dc:identifier>http://hdl.handle.net/2249.1/87460</dc:identifier>
<dc:language>en_US</dc:language>
<dc:format>application/vnd.openxmlformats-officedocument.presentationml.presentation</dc:format>
</oai_dc:dc></metadata></record><record><header><identifier>oai:tdl-ir.tdl.org:2249.1/67003</identifier><datestamp>2020-08-07T16:16:44Z</datestamp><setSpec>com_
<dc:title>Mapping the Southwest Project: Putting the Region's Maps Online</dc:title>
<dc:creator>Alemneh, Daniel</dc:creator>
<dc:creator>Jones, Jerrell</dc:creator>
<dc:creator>Hartman, Cathy</dc:creator>
<dc:creator>Phillips, Mark</dc:creator>
<dc:creator>Hodges, Ann</dc:creator>
<dc:creator>Husman, Ben</dc:creator>
<dc:contributor>University of North Texas</dc:contributor>
<dc:contributor>University of Texas at Arlington</dc:contributor>
<dc:subject>digital libraries</dc:subject>
<dc:subject>maps</dc:subject>
<dc:subject>wide format items</dc:subject>
<dc:subject>open access</dc:subject>
<dc:description>Poster presentation for the 2013 Texas Conference on Digital Libraries (TCDL).</dc:description>
```

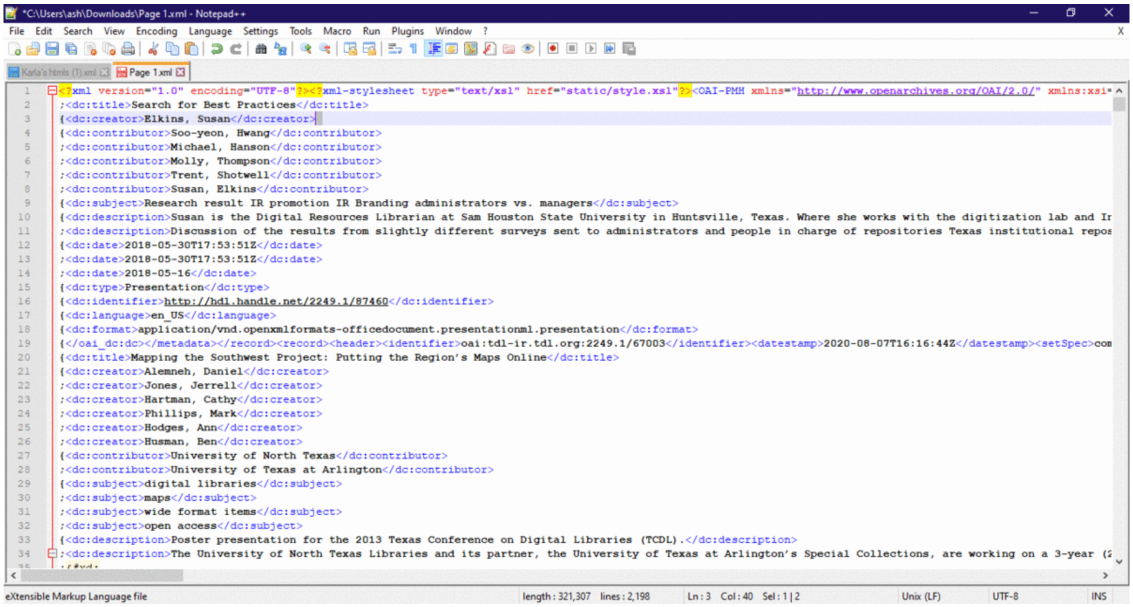
eXtensible Markup Language file
length : 319,171 lines : 2,197
Ln : 1 Col : 1 Sel : 0 | 0
Unix (LF)
UTF-8
INS

Step 4: Create a formula to add a ";" each time an element is repeated, and a "{" each time a new element begins. Once applied, this formula divides the elements into the corresponding columns. Apply this formula to all remaining cells on Column B then select the entire column and paste it into Notepad++.


AutoSave Off
Karla HTMLs - Excel
Search
Karla Roig
File Home Insert Draw Page Layout Formulas Data Review View Help Script Lab
Share
Comments
Clipboard Font Alignment Number Styles Cells Editing Ideas Sensitivity
Steps Sheet1 Raw data Tabular Data
Enter

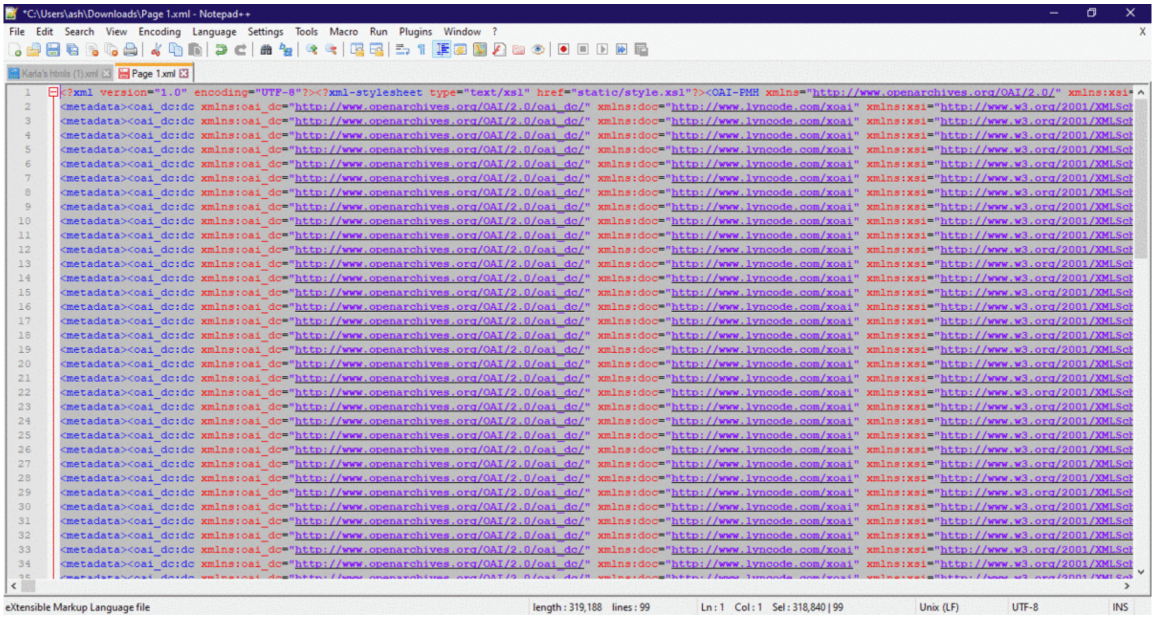
```
SUM    =A1

1  <?xml ver | =A1 | encoding="UTF-8"?><?xml-stylesheet type="text/xsl" href="static/style.xsl"?><OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:xsi="http://www.w3.org/2001/XML
2  <dc:title>Viewpoint of administrators vs. managers of repositories</dc:title>
3  <dc:title>Search for Best Practices</dc:title>
4  <dc:creator>Elkins, Susan</dc:creator>
5  <dc:contributor>Soo-yeon, Hwang</dc:contributor>
6  <dc:contributor>Michael, Hanson</dc:contributor>
7  <dc:contributor>Molly, Thompson</dc:contributor>
8  <dc:contributor>Trent, Shotwell</dc:contributor>
9  <dc:contributor>Susan, Elkins</dc:contributor>
10 <dc:subject>Research result IR promotion IR Branding administrators vs. managers</dc:subject>
11 <dc:description>Susan is the Digital Resources Librarian at Sam Houston State University in Huntsville, Texas. Where she works with the digitization lab and Institutional Repository.  She previously work
12 <dc:description>Discussion of the results from slightly different surveys sent to administrators and people in charge of repositories Texas institutional repository. We asked questions about the size of t
13 <dc:date>2018-05-30T17:53:51Z</dc:date>
14 <dc:date>2018-05-30T17:53:51Z</dc:date>
15 <dc:date>2018-05-16</dc:date>
16 <dc:type>Presentation</dc:type>
17 <dc:identifier>http://hdl.handle.net/2249.1/87460</dc:identifier>
18 <dc:language>en_US</dc:language>
19 <dc:format>application/vnd.openxmlformats-officedocument.presentationml.presentation</dc:format>
20 </oai_dc:dc></metadata></record><record><header><identifier>oai:tdl-ir.tdl.org:2249.1/67003</identifier><datestamp>2020-08-07T16:16:44Z</datestamp><setSpec>com_2249.1_4513</setSpec><setSp
21 <dc:title>Mapping the Southwest Project: Putting the Region's Maps Online</dc:title>
22 <dc:creator>Alemneh, Daniel</dc:creator>
```

Step 5: Move all individual rows into a single one by deleting all the "invisible enters," or line breaks. Then create line breaks by pressing "enter" before each iteration of "<metadata>" in order to divide the data into 100 rows, one for each complete metadata record.


*C:\Users\ash\Downloads\Page 1.xml - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
Karla's htmls (1).xml
Page 1.xml
eXtensible Markup Language file
length : 321,307 lines : 2,198
Ln : 3 Col : 40 Sel : 1 | 2
Unix (LF)
UTF-8
INS

```
<?xml version="1.0" encoding="UTF-8"?><?xml-stylesheet type="text/xsl" href="static/style.xsl"?><OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:xsi=
;<dc:title>Search for Best Practices</dc:title>
{<dc:creator>Elkins, Susan</dc:creator>
{<dc:contributor>Soo-yeon, Hwang</dc:contributor>
;<dc:contributor>Michael, Hanson</dc:contributor>
;<dc:contributor>Molly, Thompson</dc:contributor>
;<dc:contributor>Trent, Shotwell</dc:contributor>
;<dc:contributor>Susan, Elkins</dc:contributor>
{<dc:subject>Research result IR promotion IR Branding administrators vs. managers</dc:subject>
{<dc:description>Susan is the Digital Resources Librarian at Sam Houston State University in Huntsville, Texas. Where she works with the digitization lab and Ir
;<dc:description>Discussion of the results from slightly different surveys sent to administrators and people in charge of repositories Texas institutional repos
{<dc:date>2018-05-30T17:53:51Z</dc:date>
;<dc:date>2018-05-30T17:53:51Z</dc:date>
;<dc:date>2018-05-16</dc:date>
{<dc:type>Presentation</dc:type>
{<dc:identifier>http://hdl.handle.net/2249.1/87460</dc:identifier>
{<dc:language>en_US</dc:language>
{<dc:format>application/vnd.openxmlformats-officedocument.presentationml.presentation</dc:format>
{</oai_dc:dc></metadata></record><record><header><identifier>oai:tdl-ir.tdl.org:2249.1/67003</identifier><datestamp>2020-08-07T16:16:44Z</datestamp><setSpec>com
{<dc:title>Mapping the Southwest Project: Putting the Region’s Maps Online</dc:title>
{<dc:creator>Alemneh, Daniel</dc:creator>
;<dc:creator>Jones, Jerrell</dc:creator>
;<dc:creator>Hartman, Cathy</dc:creator>
;<dc:creator>Phillips, Mark</dc:creator>
;<dc:creator>Hodges, Ann</dc:creator>
;<dc:creator>Husman, Ben</dc:creator>
{<dc:contributor>University of North Texas</dc:contributor>
;<dc:contributor>University of Texas at Arlington</dc:contributor>
{<dc:subject>digital libraries</dc:subject>
;<dc:subject>maps</dc:subject>
;<dc:subject>wide format items</dc:subject>
;<dc:subject>open access</dc:subject>
{<dc:description>Poster presentation for the 2013 Texas Conference on Digital Libraries (TCDL).</dc:description>
;<dc:description>The University of North Texas Libraries and its partner, the University of Texas at Arlington’s Special Collections, are working on a 3-year (2
```

Step 6: Import the data from Notepad++ into Excel once more using the Text Import Wizard, delimiting the data by the “;” that was added earlier.

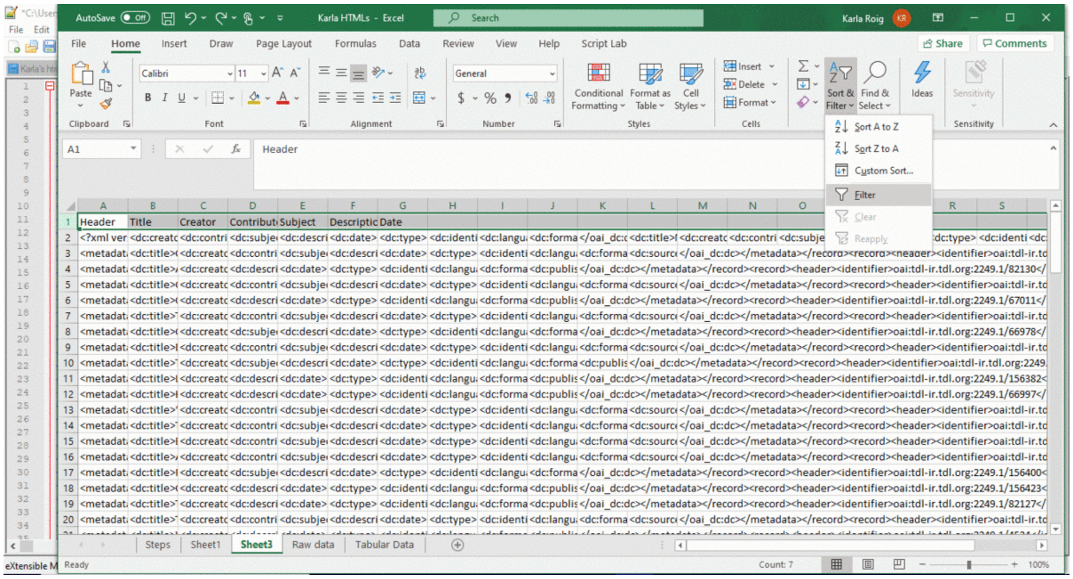

*C:\Users\ash\Downloads\Page 1.xml - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
Karla's htmls (1).xml
Page 1.xml
eXtensible Markup Language file
length : 319,188 lines : 99
Ln : 1 Col : 1 Sel : 318,840 | 99
Unix (LF)
UTF-8
INS

```
<?xml version="1.0" encoding="UTF-8"?><?xml-stylesheet type="text/xsl" href="static/style.xsl"?><OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:xsi=
<metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:doc="http://www.lyncode.com/xoai" xmlns:xsi="http://www.w3.org/2001/XMLSch
<metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:doc="http://www.lyncode.com/xoai" xmlns:xsi="http://www.w3.org/2001/XMLSch
<metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:doc="http://www.lyncode.com/xoai" xmlns:xsi="http://www.w3.org/2001/XMLSch
<metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:doc="http://www.lyncode.com/xoai" xmlns:xsi="http://www.w3.org/2001/XMLSch
<metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:doc="http://www.lyncode.com/xoai" xmlns:xsi="http://www.w3.org/2001/XMLSch
<metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:doc="http://www.lyncode.com/xoai" xmlns:xsi="http://www.w3.org/2001/XMLSch
<metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:doc="http://www.lyncode.com/xoai" xmlns:xsi="http://www.w3.org/2001/XMLSch
<metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:doc="http://www.lyncode.com/xoai" xmlns:xsi="http://www.w3.org/2001/XMLSch
<metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:doc="http://www.lyncode.com/xoai" xmlns:xsi="http://www.w3.org/2001/XMLSch
<metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:doc="http://www.lyncode.com/xoai" xmlns:xsi="http://www.w3.org/2001/XMLSch
<metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:doc="http://www.lyncode.com/xoai" xmlns:xsi="http://www.w3.org/2001/XMLSch
<metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:doc="http://www.lyncode.com/xoai" xmlns:xsi="http://www.w3.org/2001/XMLSch
<metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:doc="http://www.lyncode.com/xoai" xmlns:xsi="http://www.w3.org/2001/XMLSch
<metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:doc="http://www.lyncode.com/xoai" xmlns:xsi="http://www.w3.org/2001/XMLSch
<metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:doc="http://www.lyncode.com/xoai" xmlns:xsi="http://www.w3.org/2001/XMLSch
<metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:doc="http://www.lyncode.com/xoai" xmlns:xsi="http://www.w3.org/2001/XMLSch
<metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:doc="http://www.lyncode.com/xoai" xmlns:xsi="http://www.w3.org/2001/XMLSch
<metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:doc="http://www.lyncode.com/xoai" xmlns:xsi="http://www.w3.org/2001/XMLSch
<metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:doc="http://www.lyncode.com/xoai" xmlns:xsi="http://www.w3.org/2001/XMLSch
<metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:doc="http://www.lyncode.com/xoai" xmlns:xsi="http://www.w3.org/2001/XMLSch
<metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:doc="http://www.lyncode.com/xoai" xmlns:xsi="http://www.w3.org/2001/XMLSch
<metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:doc="http://www.lyncode.com/xoai" xmlns:xsi="http://www.w3.org/2001/XMLSch
<metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:doc="http://www.lyncode.com/xoai" xmlns:xsi="http://www.w3.org/2001/XMLSch
<metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:doc="http://www.lyncode.com/xoai" xmlns:xsi="http://www.w3.org/2001/XMLSch
<metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:doc="http://www.lyncode.com/xoai" xmlns:xsi="http://www.w3.org/2001/XMLSch
<metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:doc="http://www.lyncode.com/xoai" xmlns:xsi="http://www.w3.org/2001/XMLSch
<metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:doc="http://www.lyncode.com/xoai" xmlns:xsi="http://www.w3.org/2001/XMLSch
<metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:doc="http://www.lyncode.com/xoai" xmlns:xsi="http://www.w3.org/2001/XMLSch
<metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:doc="http://www.lyncode.com/xoai" xmlns:xsi="http://www.w3.org/2001/XMLSch
<metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:doc="http://www.lyncode.com/xoai" xmlns:xsi="http://www.w3.org/2001/XMLSch
<metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:doc="http://www.lyncode.com/xoai" xmlns:xsi="http://www.w3.org/2001/XMLSch
<metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:doc="http://www.lyncode.com/xoai" xmlns:xsi="http://www.w3.org/2001/XMLSch
<metadata><oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:doc="http://www.lyncode.com/xoai" xmlns:xsi="http://www.w3.org/2001/XMLSch
```

Data Cleaning

This section shows a series of screenshots as a guide for the metadata cleaning process. There are more detailed instructions in our “Step by Step Process for Extracting and Cleaning metadata” document.

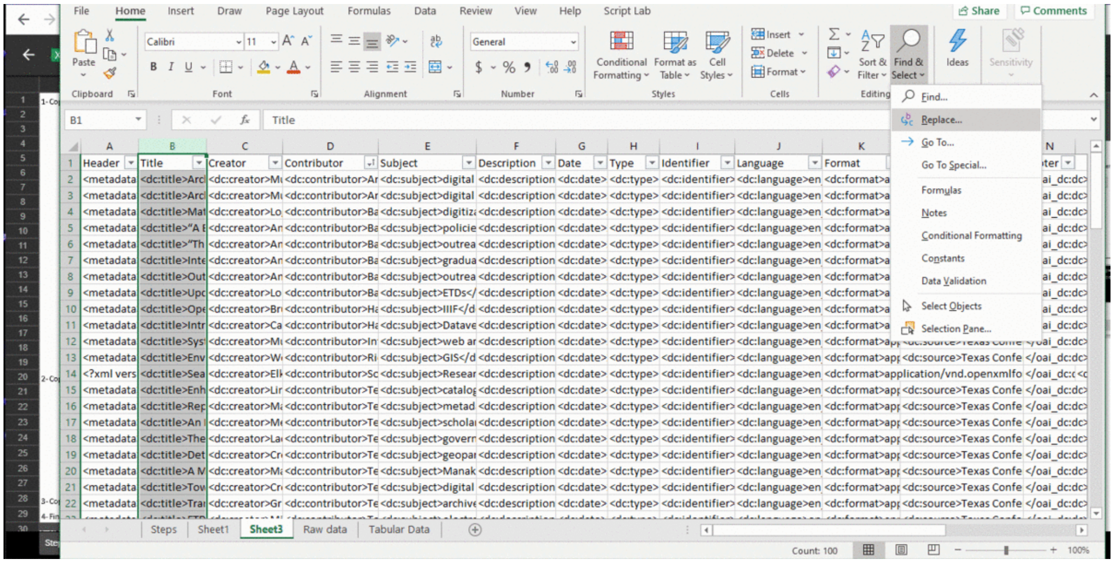
Step 7: To begin cleaning the data, create a new row at the top of the Excel document to add the titles to each column. Apply filters to column headings in order to sort and edit multiple cells containing specific elements at the same time and start shifting all elements to the right that don’t correspond to the column subject heading until every element is in the correct place.



A1 Header

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Header	Title	Creator	Contribut	Subject	Descriptic	Date													
2	<?xml ver	<dc:creatc	<dc:contri	<dc:subje	<dc:descri	<dc:date>	<dc:type>	<dc:identi	<dc:langu	<dc:forma	</oai_dc:c	<dc:title>I	<dc:creatc	<dc:contri	<dc:subje			dc:type>	<dc:identi	<dc:
3	<metadat	<dc:title>(	<dc:creatc	<dc:contri	<dc:subje	<dc:descri	<dc:date>	<dc:type>	<dc:identi	<dc:langu	<dc:forma	<dc:sourc	</oai_dc:dc></metadata></record><record><header><identifier>oai:tdl-ir.td							
4	<metadat	<dc:title>/	<dc:creatc	<dc:descri	<dc:date>	<dc:type>	<dc:identi	<dc:langu	<dc:forma	<dc:publis	</oai_dc:dc></metadata></record><record><header><identifier>oai:tdl-ir.tdl.org:2249.1/82130</									
5	<metadat	<dc:title>(	<dc:creatc	<dc:contri	<dc:subje	<dc:descri	<dc:date>	<dc:type>	<dc:identi	<dc:langu	<dc:forma	<dc:sourc	</oai_dc:dc></metadata></record><record><header><identifier>oai:tdl-ir.td							
6	<metadat	<dc:title>(	<dc:creatc	<dc:descri	<dc:date>	<dc:type>	<dc:identi	<dc:langu	<dc:forma	<dc:publis	</oai_dc:dc></metadata></record><record><header><identifier>oai:tdl-ir.tdl.org:2249.1/67011</									
7	<metadat	<dc:title>T	<dc:creatc	<dc:contri	<dc:subje	<dc:descri	<dc:date>	<dc:type>	<dc:identi	<dc:langu	<dc:forma	<dc:sourc	</oai_dc:dc></metadata></record><record><header><identifier>oai:tdl-ir.td							
8	<metadat	<dc:title>(	<dc:creatc	<dc:subje	<dc:descri	<dc:date>	<dc:type>	<dc:identi	<dc:langu	<dc:forma	</oai_dc:dc></metadata></record><record><header><identifier>oai:tdl-ir.tdl.org:2249.1/66978</									
9	<metadat	<dc:title>I	<dc:creatc	<dc:contri	<dc:subje	<dc:descri	<dc:date>	<dc:type>	<dc:identi	<dc:langu	<dc:forma	<dc:sourc	</oai_dc:dc></metadata></record><record><header><identifier>oai:tdl-ir.td							
10	<metadat	<dc:title>T	<dc:creatc	<dc:subje	<dc:descri	<dc:date>	<dc:type>	<dc:identi	<dc:langu	<dc:forma	<dc:publis	</oai_dc:dc></metadata></record><record><header><identifier>oai:tdl-ir.tdl.org:2249.								
11	<metadat	<dc:title>I	<dc:creatc	<dc:descri	<dc:date>	<dc:type>	<dc:identi	<dc:langu	<dc:forma	<dc:publis	</oai_dc:dc></metadata></record><record><header><identifier>oai:tdl-ir.tdl.org:2249.1/156382<									
12	<metadat	<dc:title>I	<dc:creatc	<dc:descri	<dc:date>	<dc:type>	<dc:identi	<dc:langu	<dc:forma	<dc:publis	</oai_dc:dc></metadata></record><record><header><identifier>oai:tdl-ir.tdl.org:2249.1/66997</									
13	<metadat	<dc:title>'	<dc:creatc	<dc:contri	<dc:subje	<dc:descri	<dc:date>	<dc:type>	<dc:identi	<dc:langu	<dc:forma	<dc:sourc	</oai_dc:dc></metadata></record><record><header><identifier>oai:tdl-ir.td							
14	<metadat	<dc:title>T	<dc:creatc	<dc:contri	<dc:subje	<dc:descri	<dc:date>	<dc:type>	<dc:identi	<dc:langu	<dc:forma	<dc:sourc	</oai_dc:dc></metadata></record><record><header><identifier>oai:tdl-ir.td							
15	<metadat	<dc:title>I	<dc:creatc	<dc:contri	<dc:subje	<dc:descri	<dc:date>	<dc:type>	<dc:identi	<dc:langu	<dc:forma	<dc:sourc	</oai_dc:dc></metadata></record><record><header><identifier>oai:tdl-ir.td							
16	<metadat	<dc:title>/	<dc:creatc	<dc:contri	<dc:subje	<dc:descri	<dc:date>	<dc:type>	<dc:identi	<dc:langu	<dc:forma	<dc:sourc	</oai_dc:dc></metadata></record><record><header><identifier>oai:tdl-ir.td							
17	<metadat	<dc:title>I	<dc:creatc	<dc:subje	<dc:descri	<dc:date>	<dc:type>	<dc:identi	<dc:langu	<dc:forma	</oai_dc:dc></metadata></record><record><header><identifier>oai:tdl-ir.tdl.org:2249.1/156400<									
18	<metadat	<dc:title>(	<dc:creatc	<dc:descri	<dc:date>	<dc:type>	<dc:identi	<dc:langu	<dc:forma	<dc:publis	</oai_dc:dc></metadata></record><record><header><identifier>oai:tdl-ir.tdl.org:2249.1/156423<									
19	<metadat	<dc:title>T	<dc:creatc	<dc:descri	<dc:date>	<dc:type>	<dc:identi	<dc:langu	<dc:forma	<dc:publis	</oai_dc:dc></metadata></record><record><header><identifier>oai:tdl-ir.tdl.org:2249.1/82127</									
20	<metadat	<dc:title>T	<dc:creatc	<dc:contri	<dc:subje	<dc:descri	<dc:date>	<dc:type>	<dc:identi	<dc:langu	<dc:forma	<dc:sourc	</oai_dc:dc></metadata></record><record><header><identifier>oai:tdl-ir.td							

Step 8: Continue cleaning the data by deleting the header and footer of each element (in this case each element is wrapped in Dublin Core XML) using the find and replace tool. Use this same tool to continue cleaning up additional errors in the data.



B1 Title

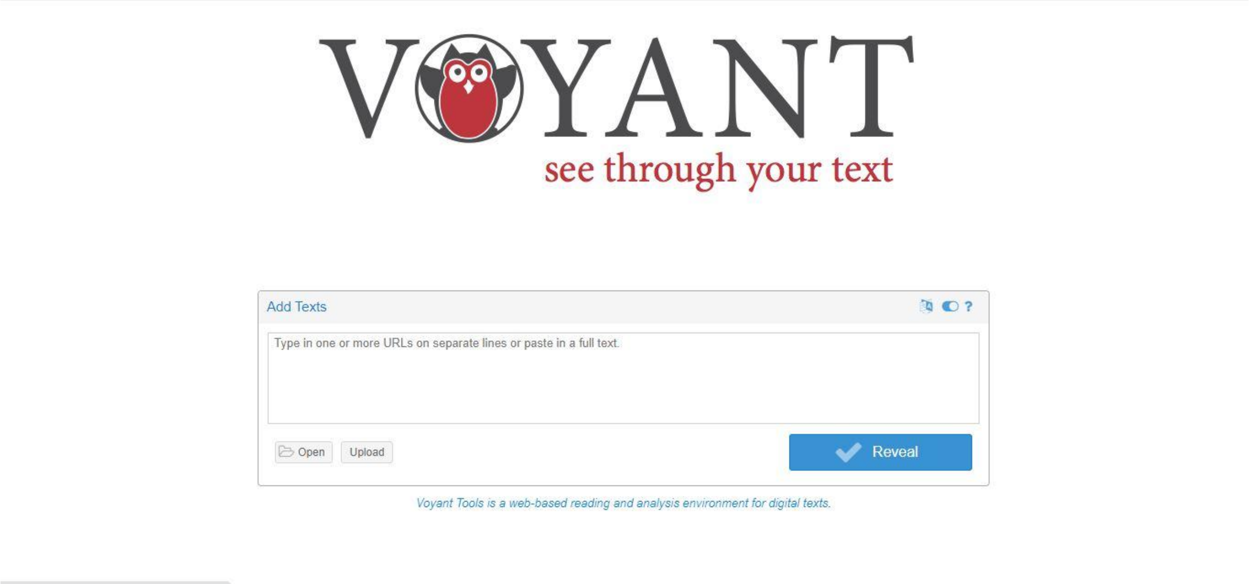
	A	B	C	D	E	F	G	H	I	J	K		N
1	Header	Title	Creator	Contributor	Subject	Description	Date	Type	Identifier	Language	Format		ter
2	<metadata	<dc:title>Arc	<dc:creator>M	<dc:contributor>Ar	<dc:subject>digital	<dc:description	<dc:date>	<dc:type>	<dc:identifier>	<dc:language>en	<dc:format>a		ai_dc:dc>
3	<metadata	<dc:title>Arc	<dc:creator>M	<dc:contributor>Ar	<dc:subject>digital	<dc:description	<dc:date>	<dc:type>	<dc:identifier>	<dc:language>en	<dc:format>a		ai_dc:dc>
4	<metadata	<dc:title>Mat	<dc:creator>Lo	<dc:contributor>Ba	<dc:subject>digitiz	<dc:description	<dc:date>	<dc:type>	<dc:identifier>	<dc:language>en	<dc:format>a		ai_dc:dc>
5	<metadata	<dc:title>"A E	<dc:creator>An	<dc:contributor>Ba	<dc:subject>policie	<dc:description	<dc:date>	<dc:type>	<dc:identifier>	<dc:language>en	<dc:format>a		ai_dc:dc>
6	<metadata	<dc:title>"Th	<dc:creator>An	<dc:contributor>Ba	<dc:subject>outrea	<dc:description	<dc:date>	<dc:type>	<dc:identifier>	<dc:language>en	<dc:format>a		ai_dc:dc>
7	<metadata	<dc:title>Inte	<dc:creator>An	<dc:contributor>Ba	<dc:subject>gradua	<dc:description	<dc:date>	<dc:type>	<dc:identifier>	<dc:language>en	<dc:format>a		ai_dc:dc>
8	<metadata	<dc:title>Out	<dc:creator>An	<dc:contributor>Ba	<dc:subject>outrea	<dc:description	<dc:date>	<dc:type>	<dc:identifier>	<dc:language>en	<dc:format>a		ai_dc:dc>
9	<metadata	<dc:title>Upc	<dc:creator>Lo	<dc:contributor>Ba	<dc:subject>ETDs</	<dc:description	<dc:date>	<dc:type>	<dc:identifier>	<dc:language>en	<dc:format>a		ai_dc:dc>
10	<metadata	<dc:title>Ope	<dc:creator>Br	<dc:contributor>Ha	<dc:subject>IIIF</d	<dc:description	<dc:date>	<dc:type>	<dc:identifier>	<dc:language>en	<dc:format>a		ai_dc:dc>
11	<metadata	<dc:title>Intr	<dc:creator>Ca	<dc:contributor>Ha	<dc:subject>Datave	<dc:description	<dc:date>	<dc:type>	<dc:identifier>	<dc:language>en	<dc:format>a		ai_dc:dc>
12	<metadata	<dc:title>Sys	<dc:creator>M	<dc:contributor>In	<dc:subject>web ar	<dc:description	<dc:date>	<dc:type>	<dc:identifier>	<dc:language>en	<dc:format>ap		</oai_dc:dc>
13	<metadata	<dc:title>Env	<dc:creator>W	<dc:contributor>Ri	<dc:subject>GIS</d	<dc:description	<dc:date>	<dc:type>	<dc:identifier>	<dc:language>en	<dc:format>ap	<dc:source>Texas Confe	</oai_dc:dc>
14	<?xml vers	<dc:title>Sea	<dc:creator>Ell	<dc:contributor>Sc	<dc:subject>Resear	<dc:description	<dc:date>	<dc:type>	<dc:identifier>	<dc:language>en	<dc:format>application/vnd.openxmlfo		</oai_dc:c <c
15	<metadata	<dc:title>Enh	<dc:creator>Lir	<dc:contributor>Te	<dc:subject>catalog	<dc:description	<dc:date>	<dc:type>	<dc:identifier>	<dc:language>en	<dc:format>ap	<dc:source>Texas Confe	</oai_dc:dc>
16	<metadata	<dc:title>Rep	<dc:creator>M	<dc:contributor>Te	<dc:subject>metad	<dc:description	<dc:date>	<dc:type>	<dc:identifier>	<dc:language>en	<dc:format>ap	<dc:source>Texas Confe	</oai_dc:dc>
17	<metadata	<dc:title>An I	<dc:creator>M	<dc:contributor>Te	<dc:subject>schola	<dc:description	<dc:date>	<dc:type>	<dc:identifier>	<dc:language>en	<dc:format>ap	<dc:source>Texas Confe	</oai_dc:dc>
18	<metadata	<dc:title>The	<dc:creator>La	<dc:contributor>Te	<dc:subject>goverr	<dc:description	<dc:date>	<dc:type>	<dc:identifier>	<dc:language>en	<dc:format>ap	<dc:source>Texas Confe	</oai_dc:dc>
19	<metadata	<dc:title>Det	<dc:creator>Cr	<dc:contributor>Te	<dc:subject>geopa	<dc:description	<dc:date>	<dc:type>	<dc:identifier>	<dc:language>en	<dc:format>ap	<dc:source>Texas Confe	</oai_dc:dc>
20	<metadata	<dc:title>A M	<dc:creator>M	<dc:contributor>Te	<dc:subject>Manak	<dc:description	<dc:date>	<dc:type>	<dc:identifier>	<dc:language>en	<dc:format>ap	<dc:source>Texas Confe	</oai_dc:dc>
21	<metadata	<dc:title>Tow	<dc:creator>Cr	<dc:contributor>Te	<dc:subject>digital	<dc:description	<dc:date>	<dc:type>	<dc:identifier>	<dc:language>en	<dc:format>ap	<dc:source>Texas Confe	</oai_dc:dc>
22	<metadata	<dc:title>Tra	<dc:creator>Gr	<dc:contributor>Te	<dc:subject>archive	<dc:description	<dc:date>	<dc:type>	<dc:identifier>	<dc:language>en	<dc:format>ap	<dc:source>Texas Confe	</oai_dc:dc>

Data Visualization: Using Voyant Tools and Tableau Public

Data Visualization Using Voyant Tools

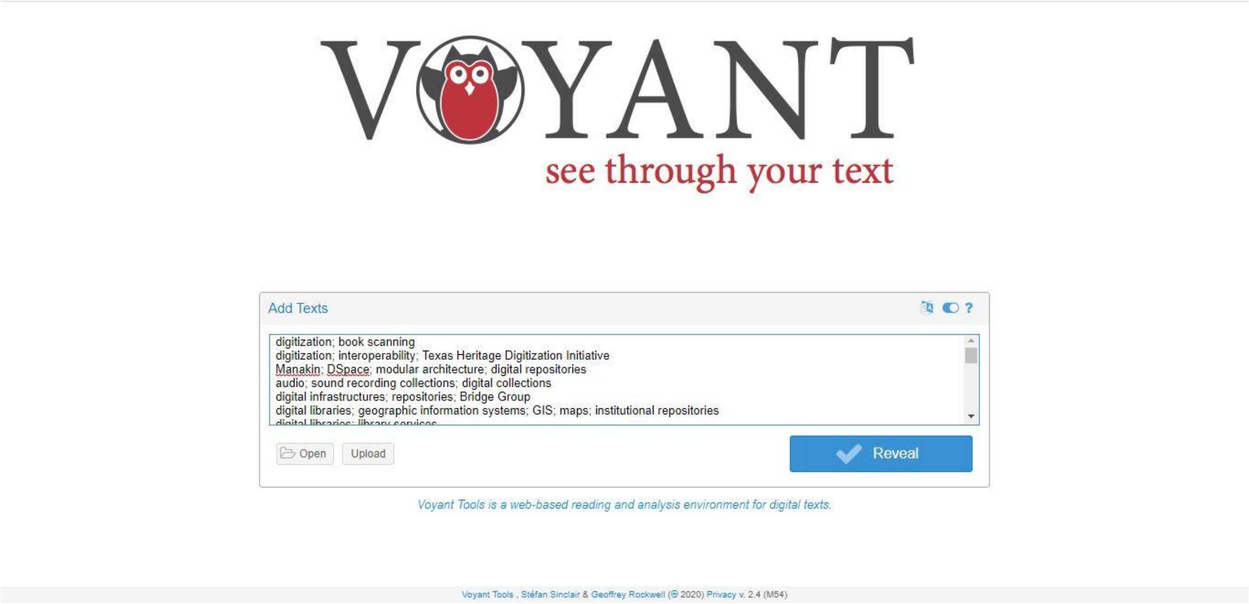
The following steps describe the data visualization process using the cleaned "Subject" metadata from the TCDL proceedings.

1. Go to the Voyant Tools home page.


VOYANT
see through your text
Add Texts
Type in one or more URLs on separate lines or paste in a full text.
Open
Upload
Reveal

Voyant Tools is a web-based reading and analysis environment for digital texts.

2. Open an existing corpus, upload a file from a computer, or copy/paste text into the text box. In this case, the data was copy/pasted. Click the “Reveal” button.

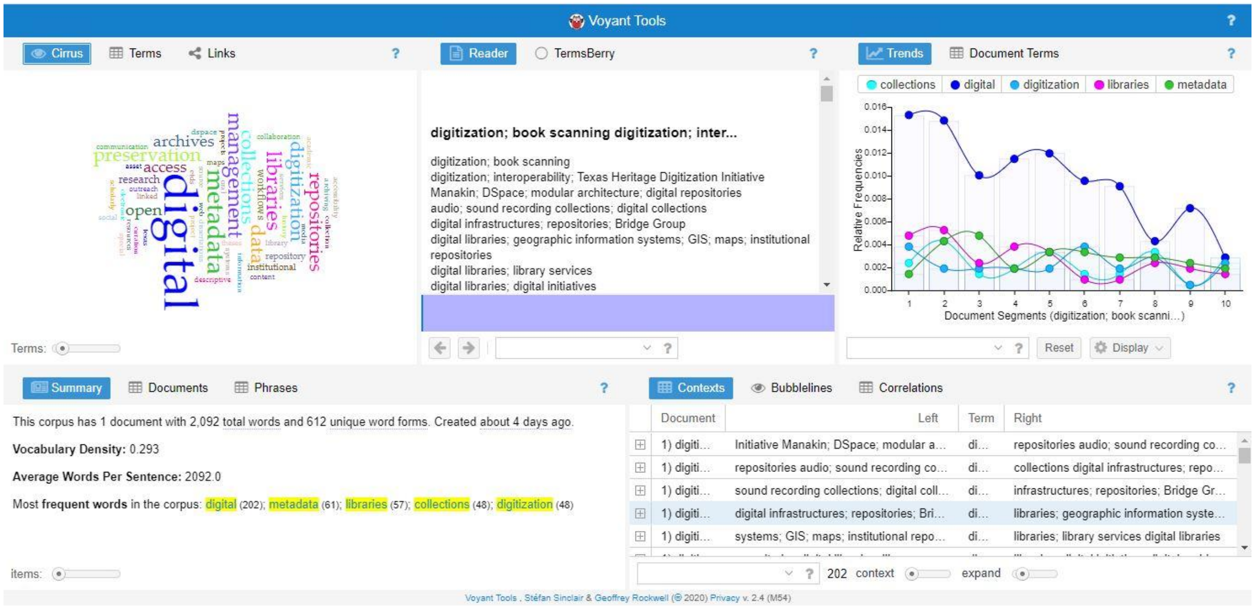


digitization; book scanning
digitization; interoperability; Texas Heritage Digitization Initiative
Manakin; DSpace; modular architecture; digital repositories
audio; sound recording collections; digital collections
digital infrastructures; repositories; Bridge Group
digital libraries; geographic information systems; GIS; maps; institutional repositories
digital libraries; library services

Voyant Tools is a web-based reading and analysis environment for digital texts.

Voyant Tools , Stéfan Sinclair & Geoffrey Rockwell (© 2020) Privacy v. 2.4 (M54)

3. Multiple text analyses, using multiple tools, should be displayed on the screen in panels. Click on the “?” buttons in each panel for more information about each text analysis tool. Hover your mouse over the “?” and click the windowpane icon to choose different tools for each panel.



digitization; book scanning digitization; inter...

digitization; book scanning
digitization; interoperability; Texas Heritage Digitization Initiative
Manakin; DSpace; modular architecture; digital repositories
audio; sound recording collections; digital collections
digital infrastructures; repositories; Bridge Group
digital libraries; geographic information systems; GIS; maps; institutional repositories
digital libraries; library services
digital libraries; digital initiatives

This corpus has 1 document with 2,092 total words and 612 unique word forms. Created about 4 days ago.

Vocabulary Density: 0.293

Average Words Per Sentence: 2092.0

Most **frequent words** in the corpus: digital (202); metadata (61); libraries (57); collections (48); digitization (48)

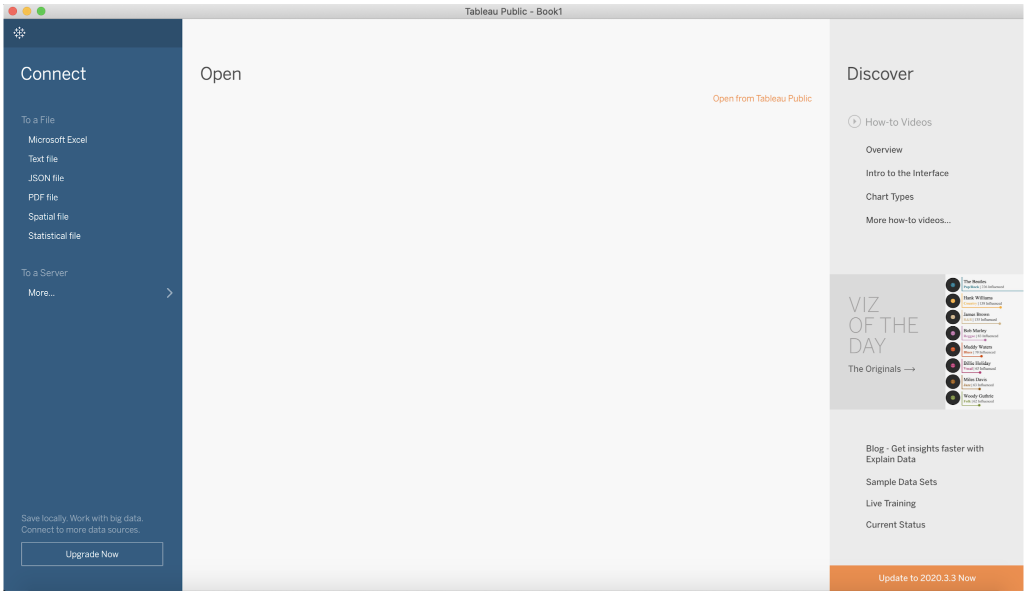
	Document	Left	Term	Right
⊞	1) digiti…	Initiative Manakin; DSpace; modular a…	di…	repositories audio; sound recording co…
⊞	1) digiti…	repositories audio; sound recording co…	di…	collections digital infrastructures; repo…
⊞	1) digiti…	sound recording collections; digital coll…	di…	infrastructures; repositories; Bridge Gr…
⊞	1) digiti…	digital infrastructures; repositories; Bri…	di…	libraries; geographic information syste…
⊞	1) digiti…	systems; GIS; maps; institutional repo…	di…	libraries; library services digital libraries

Voyant Tools , Stéfan Sinclair & Geoffrey Rockwell (© 2020) Privacy v. 2.4 (M54)

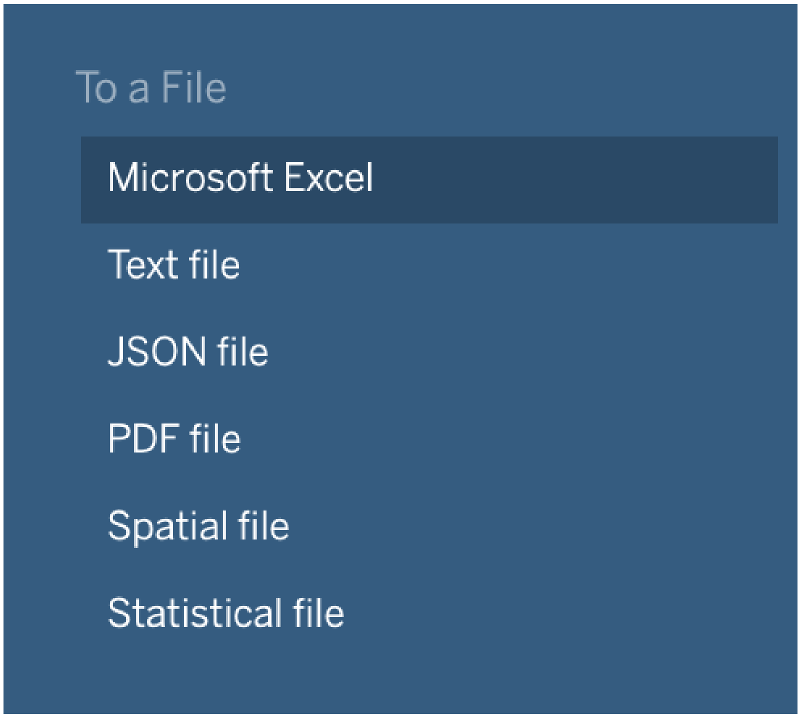
4. Voyant Tools' "Getting Started" page has more information about how to work with Voyant, as well as a list of text analysis tools.

Data Visualization Using Tableau Public

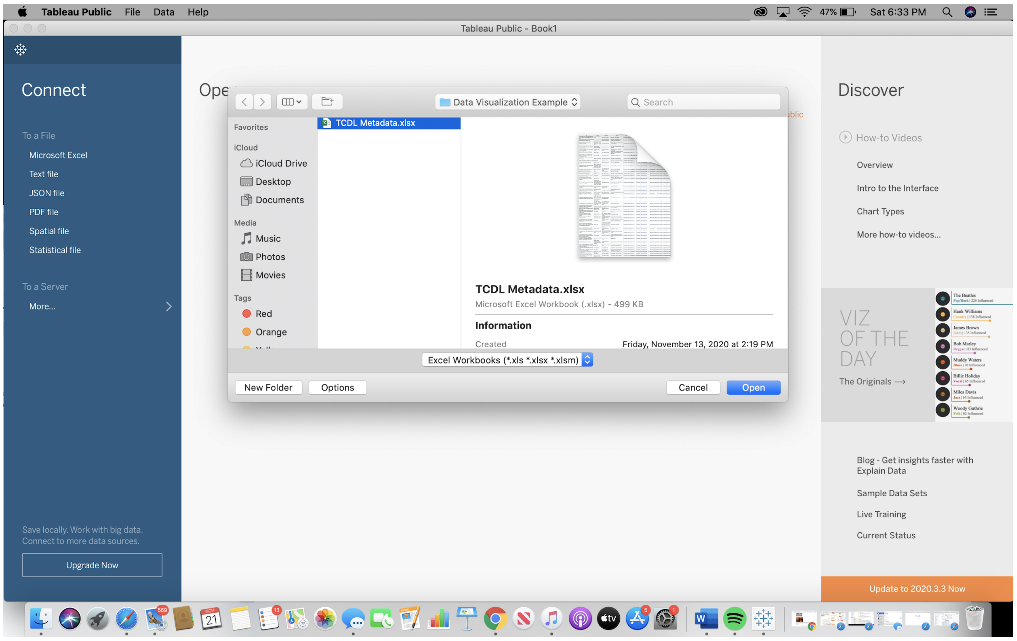
1. Open the Tableau Public application on your desktop.



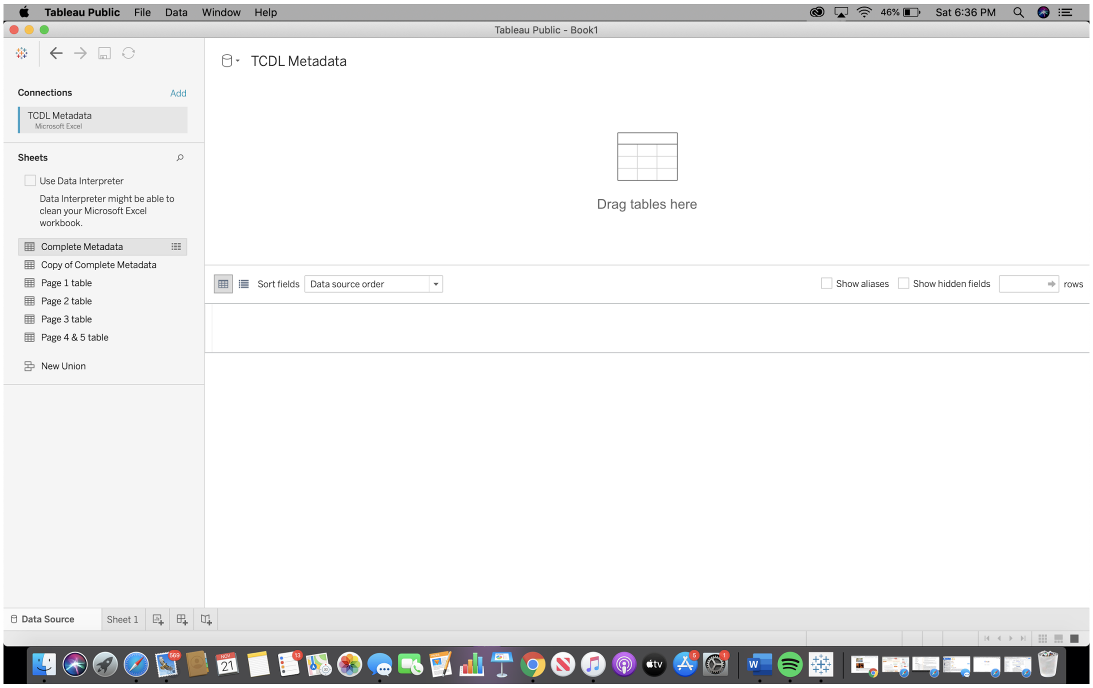
2. Select the type of file format of your dataset and you will be prompted to select a file from your computer.

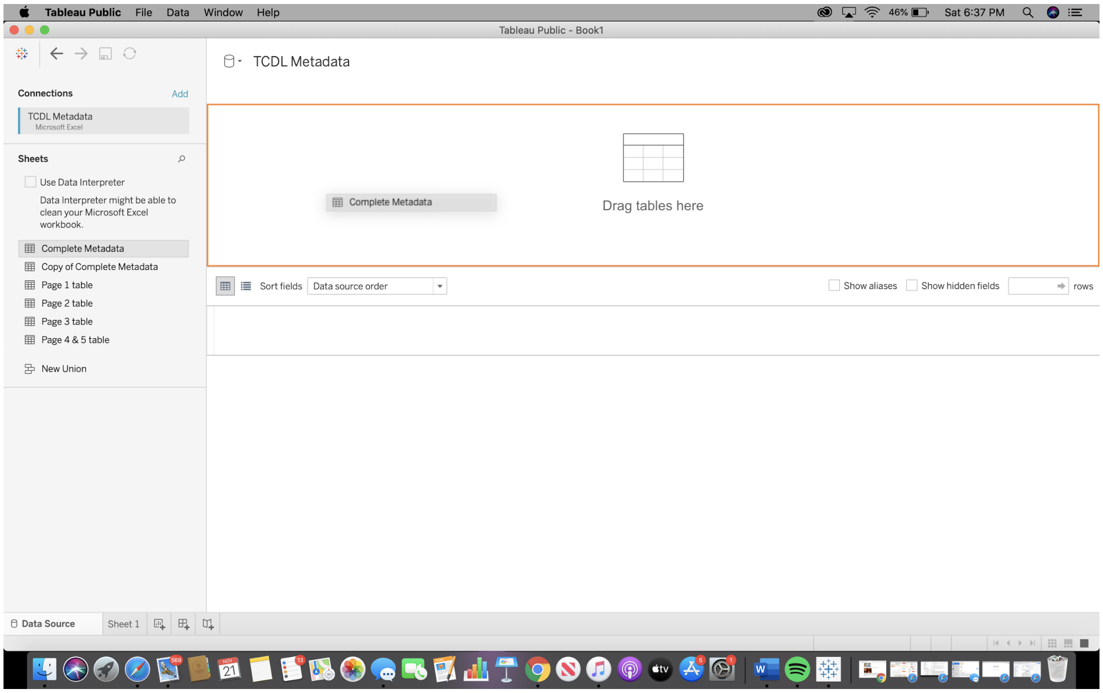


3. Navigate to the file containing your dataset on your hard drive and select “Open.” This action will open the data within a Tableau Public workbook.

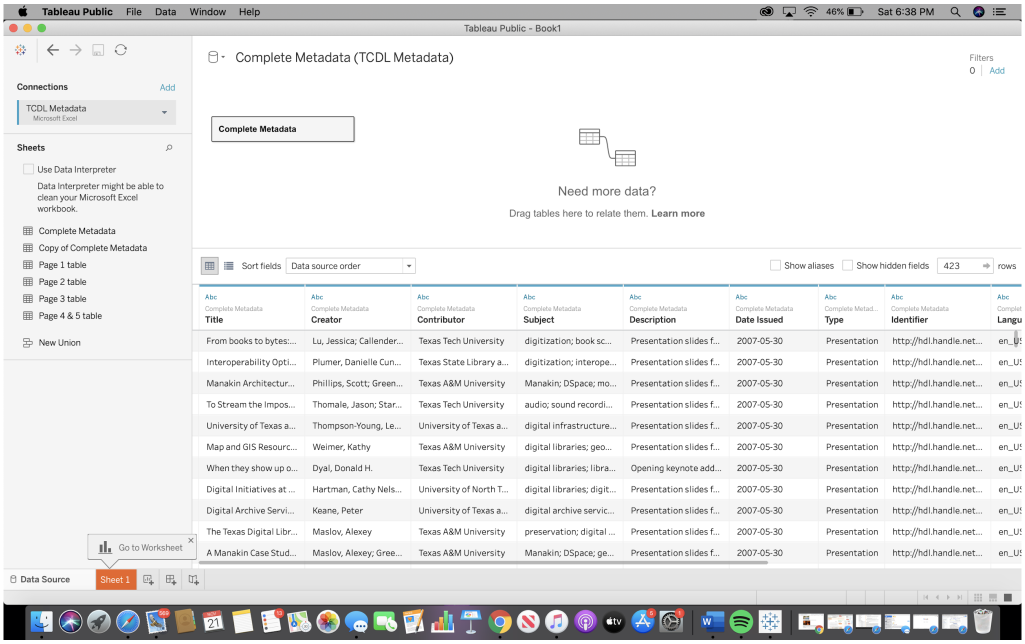


4. Select the worksheet you want to use on the left-hand side under the “Sheets” header and drag and drop it into the “Drag tables here” area.





5. Open a new worksheet by selecting “Sheet 1” in the bottom left-hand corner.



Complete Metadata (TCDL Metadata)

Complete Metadata

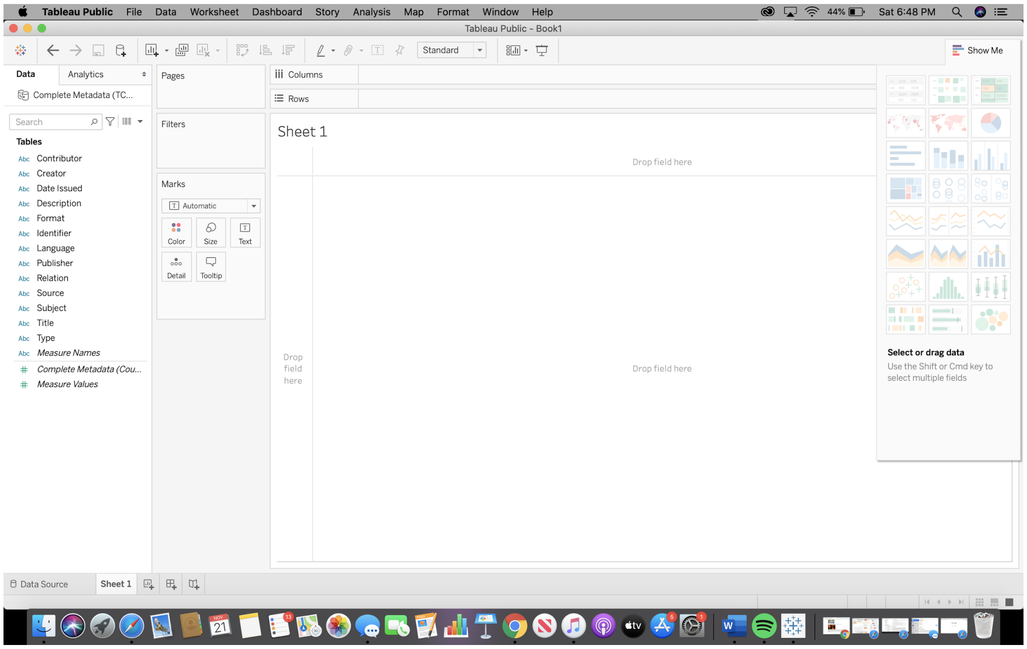
□ Use Data Interpreter

Data Interpreter might be able to clean your Microsoft Excel workbook.

Sort fields: Data source order □ Show aliases □ Show hidden fields 423 rows

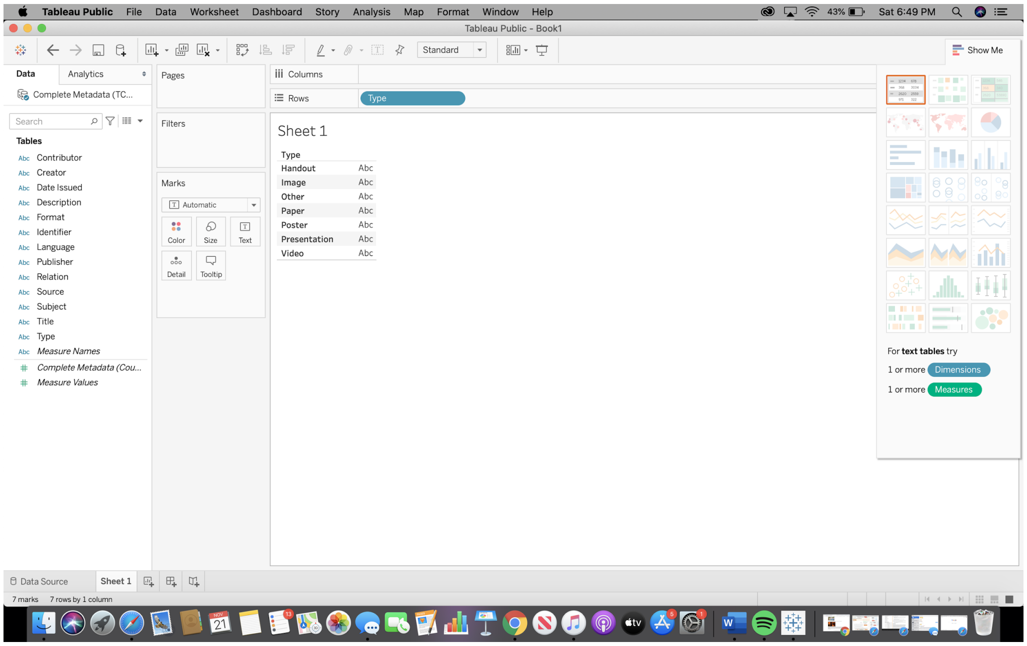
Title	Creator	Contributor	Subject	Description	Date Issued	Type	Identifier	Langu
From books to bytes:...	Lu, Jessica; Callender...	Texas Tech University	digitization; book sc...	Presentation slides f...	2007-05-30	Presentation	http://hdl.handle.net...	en_US
Interoperability Opti...	Plumer, Danielle Cun...	Texas State Library a...	digitization; interope...	Presentation slides f...	2007-05-30	Presentation	http://hdl.handle.net...	en_US
Manakin Architectur...	Phillips, Scott; Green...	Texas A&M University	Manakin; DSpace; mo...	Presentation slides f...	2007-05-30	Presentation	http://hdl.handle.net...	en_US
To Stream the Impos...	Thomale, Jason; Star...	Texas Tech University	audio; sound recordi...	Presentation slides f...	2007-05-30	Presentation	http://hdl.handle.net...	en_US
University of Texas a...	Thompson-Young, Le...	University of Texas a...	digital infrastructure...	Presentation slides f...	2007-05-30	Presentation	http://hdl.handle.net...	en_US
Map and GIS Resourc...	Weimer, Kathy	Texas A&M University	digital libraries; geo...	Presentation slides f...	2007-05-30	Presentation	http://hdl.handle.net...	en_US
When they show up o...	Dyal, Donald H.	Texas Tech University	digital libraries; libra...	Opening keynote add...	2007-05-30	Presentation	http://hdl.handle.net...	en_US
Digital Initiatives at ...	Hartman, Cathy Nels...	University of North T...	digital libraries; digit...	Presentation slides f...	2007-05-30	Presentation	http://hdl.handle.net...	en_US
Digital Archive Servi...	Keane, Peter	University of Texas a...	digital archive servic...	Presentation slides f...	2007-05-30	Presentation	http://hdl.handle.net...	en_US
The Texas Digital Libr...	Maslov, Alexey	Texas A&M University	preservation; digital ...	Presentation slides f...	2007-05-30	Presentation	http://hdl.handle.net...	en_US
A Manakin Case Stud...	Maslov, Alexey; Gree...	Texas A&M University	Manakin; DSpace; ge...	Presentation slides f...	2007-05-30	Presentation	http://hdl.handle.net...	en_US

6. You can create visualizations by dragging and dropping the fields on the left-hand side onto the canvas (where it says “Drag fields here”) or into the “Rows” or “Columns” areas.



Tables
- Contributor
- Creator
- Date Issued
- Description
- Format
- Identifier
- Language
- Publisher
- Relation
- Source
- Subject
- Title
- Type
- *Measure Names*
- *Complete Metadata (Cou...*
- *Measure Values*

7. For this example, we want to create a visualization of how many different item types are in the TCDL Proceedings Collection (2007 to 2020), so we drag the “Type” field on to the canvas.



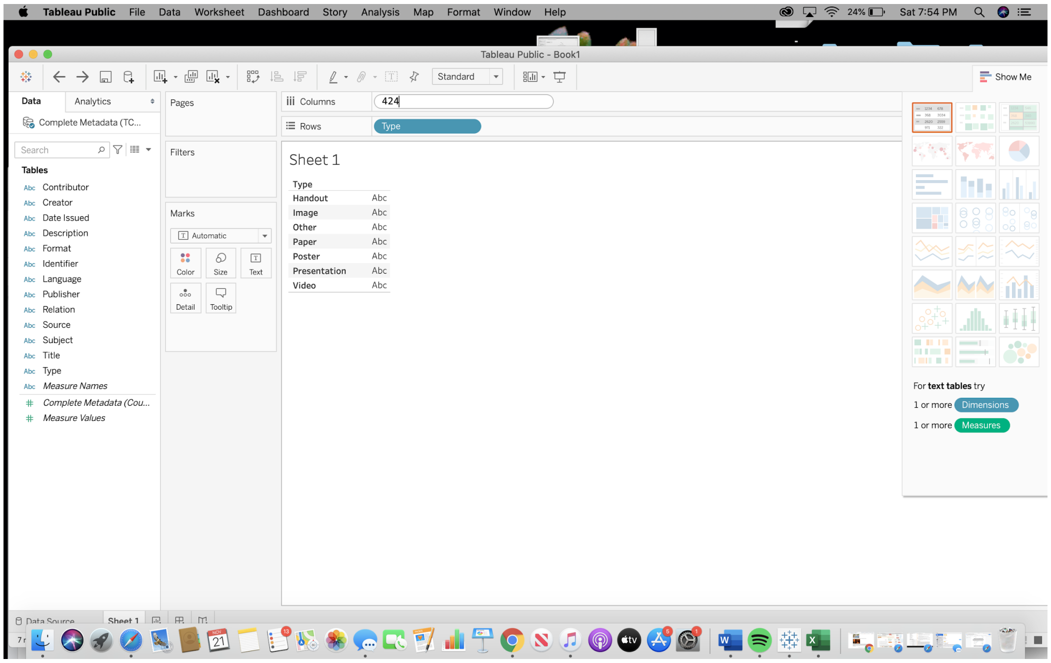
Columns

Rows: Type

Sheet 1

Type	
Handout	Abc
Image	Abc
Other	Abc
Paper	Abc
Poster	Abc
Presentation	Abc
Video	Abc

8. Next, we entered “424” to the number of columns field because that is the amount of individual items that are in the TCDL Proceedings Collection.



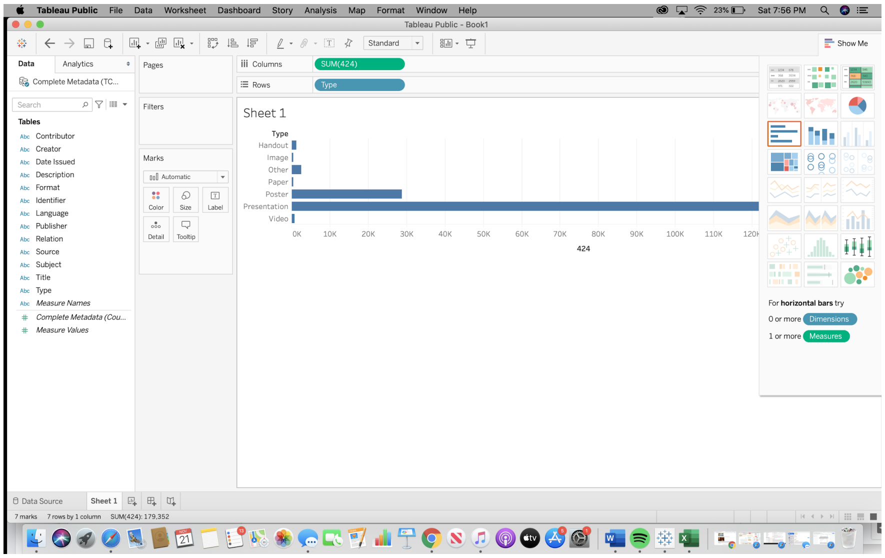
Columns: 424

Rows: Type

Sheet 1

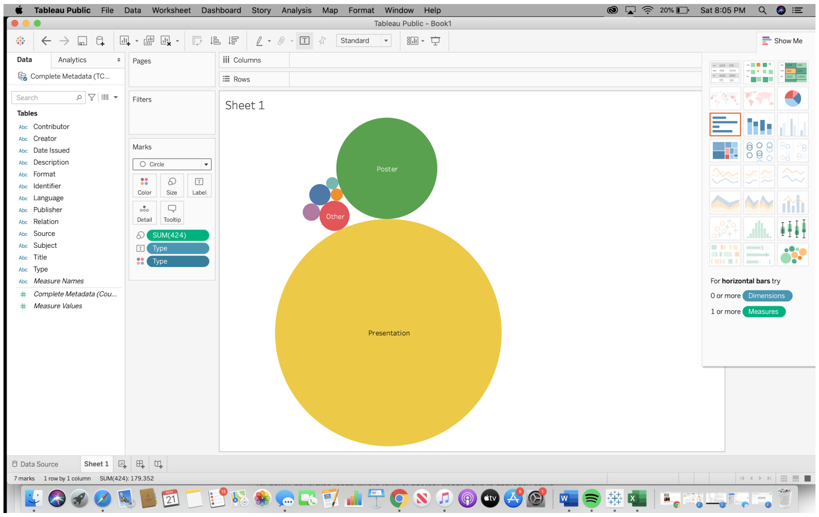
Type	
Handout	Abc
Image	Abc
Other	Abc
Paper	Abc
Poster	Abc
Presentation	Abc
Video	Abc

9. Once you hit “enter,” Tableau Public will generate a visualization of the proportion of each item “type” relative to the whole of the collections.


Columns: SUM(424)
Rows: Type

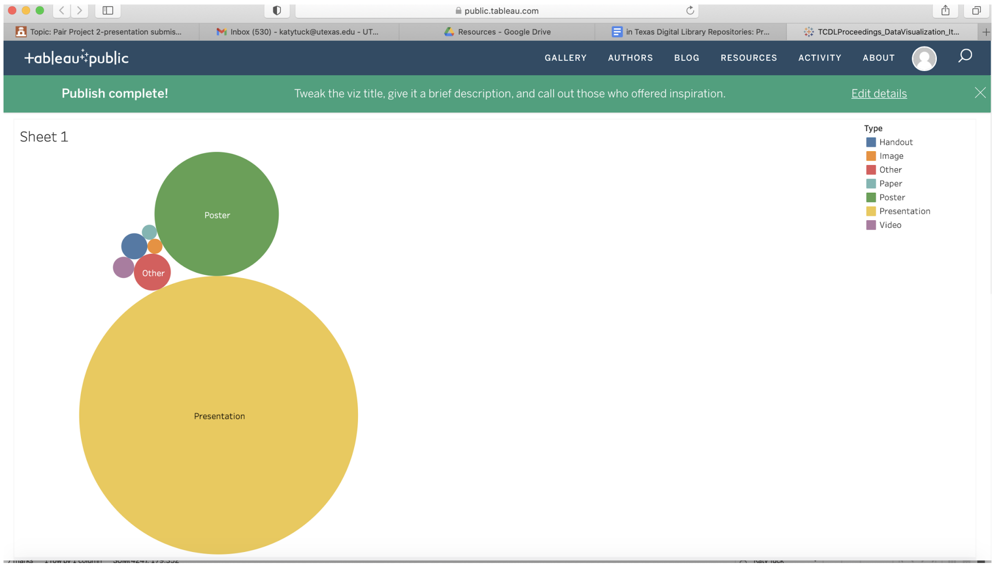
7 marks 7 rows by 1 column SUM(424): 179,352

10. On the right-hand side, you can select what kind of visualization you would like to use for your data. Tableau Public automatically indicates which graphical representations would be best suited for your visualization. We picked a simple “packed bubbles” representation to show the breakdown of item types within the collection. As you can see below, the majority of items in the collection are presentations, with posters as the next largest type group.


Marks: Circle
SUM(424)
Type
Type

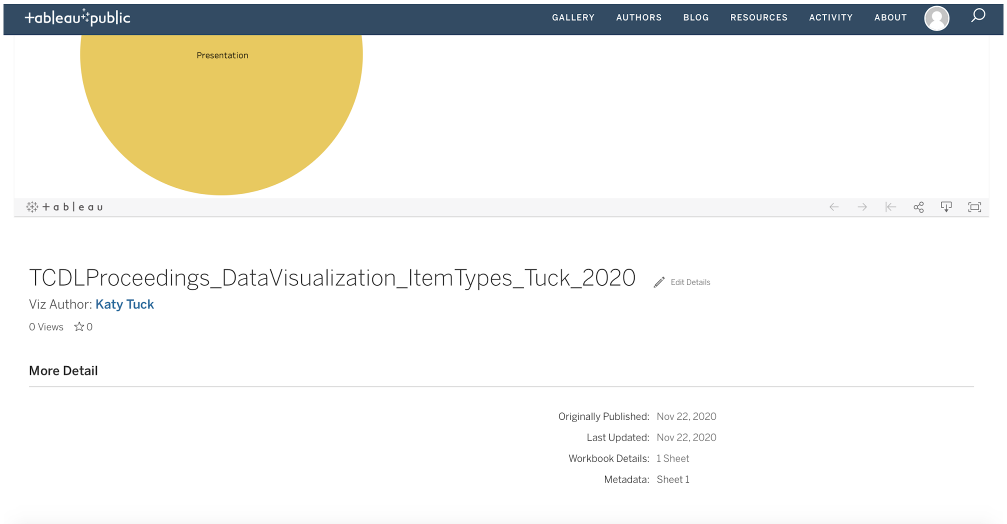
7 marks 1 row by 1 column SUM(424): 179,352

11. Login to your Tableau Public account (creating one if you need to) to access your data visualization (or “viz” as they are called by Tableau Public).



Publish complete! Tweak the viz title, give it a brief description, and call out those who offered inspiration. Edit details

12. Download your data visualization by pressing the download icon in the bottom right-hand corner of the image.



TCDLProceedings_DataVisualization_ItemTypes_Tuck_2020

Viz Author: Katy Tuck

0 Views ☆0

More Detail

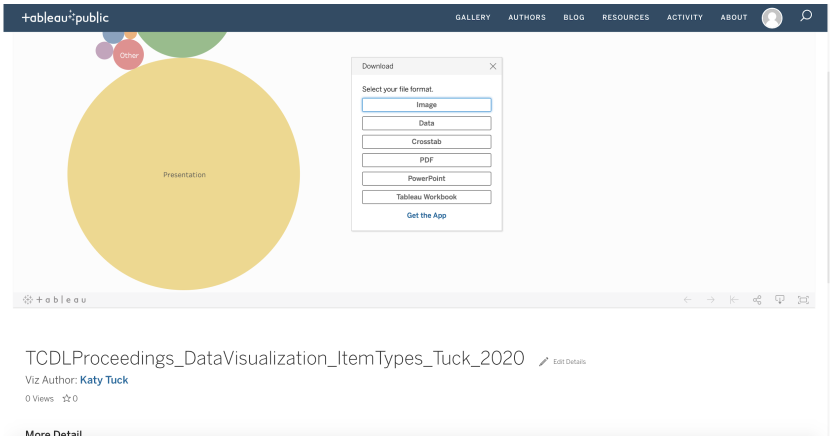
Originally Published: Nov 22, 2020

Last Updated: Nov 22, 2020

Workbook Details: 1 Sheet

Metadata: Sheet 1

13. Choose how you would like to download your visualization. For this example, we chose "image."



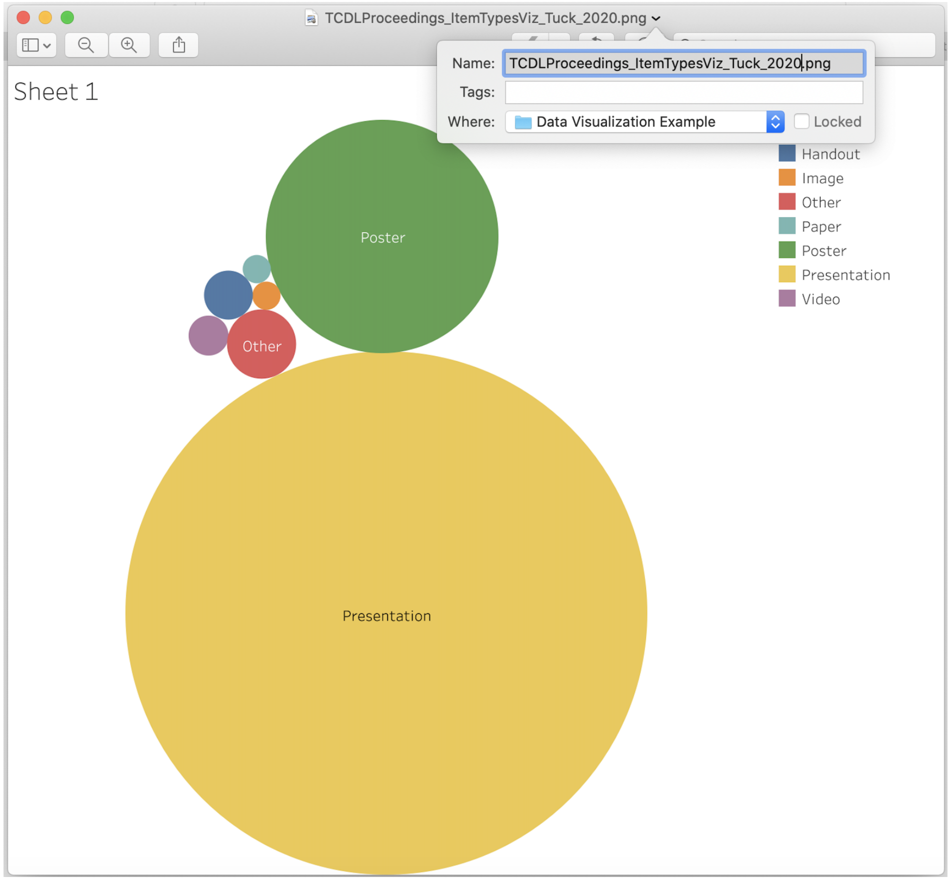
TCDLProceedings_DataVisualization_ItemTypes_Tuck_2020

Viz Author: Katy Tuck

0 Views ☆0

More Detail

14. Now you can save the downloaded png image on your desktop for later reuse on a website or for any other use!



TCDLProceedings_ItemTypesViz_Tuck_2020.png

Name: TCDLProceedings_ItemTypesViz_Tuck_2020.png

Tags:

Where: Data Visualization Example

□ Locked

Initial Steps: Starting a Collections as Data Project

- Choose a collection you would like to explore.
- Ask the question: Why do you want to analyze this collection computationally?
- Weigh the potential costs and benefits of undertaking the project.
- Identify the format of your data (XML, TIFF, PDF, Excel, etc.) and the level of technical expertise required to work with the format. For example, working with XML files may have a steeper learning curve than working with Excel files.
- Determine whether your data needs to be cleaned.
- Select the appropriate tools to analyze your data. Different stages of analysis (collection, cleaning, etc.) may require different tools. Open access tools are preferred. Consult this document's List of Tools.
- Once the data is cleaned, you can select and download a data visualization software (for the purposes of this project, we chose to work with two popular free applications: Voyant Tools and Tableau Public).

Recommendations

- Start where you are and break the process into small, manageable tasks so that the work doesn't feel overwhelming.
- Preserve the contexts of data. This includes documenting, and making available to others, data's origins, as well as how data is analyzed and used over time. For example, a README plain text file records basic information about data and can help preserve its context. Recommendation no. 34, 50 Things --- Always Already Computational (Padilla et al., 2019, May 20a, p. 4).
- Provide open access to data and documentation. For example, data and documentation can be made available through a public Github repository.

- Convert your data into a format that doesn't require a lot of technical expertise to work with. This increases accessibility, long-term preservability, and institutional collaboration.
- Seek out digital humanities and digital scholarship practitioners in the field and reach out to IT professionals within your institution for collaboration. "Network with people who work with data and have the skills or knowledge you need to get your work done." Recommendation no. 18, 50 Things --- Always Already Computational (Padilla et al., 2019, May 20a, p. 3).

Conclusion

The practical recommendations for use provided in this document are intended to be a model for TDL members interested in working with Collections as Data. The methods utilized in this project are not universally applicable. Data collection, cleaning, and visualization methods will vary according to the needs of each Collections as Data project.

Glossary

- **Digital Humanities:**

  The field of digital humanities (DH) broadly refers to a certain methodological approach involving the application of digital technology tools to scholarship in order to analyze, visualize, interpret, and present this information in new ways. It is characterized by an emphasis on collaboration, interpretation, openness, interoperability, and innovation and can be applied to multiple disciplinary fields toward a variety of purposes. Examples of DH methods and uses include "text analysis, data mining, visualization, modeling and simulation, geospatial analysis and mapping, multimedia storytelling, information design, network analysis, interface design, and mark-up…[and applying] these tools to humanistic questions" ("About Us - Digital Humanities," n.d.).
- **Texas Digital Repository (TDR)**

“The **Texas Data Repository** is a platform for publishing and archiving datasets (and other data products) created by faculty, staff, and students at Texas higher education institutions. The repository is built in an open-source application called Dataverse, developed and used by Harvard University.

The repository is hosted by the Texas Digital Library, a consortium of academic libraries in Texas with a proven history of providing shared technology services that support secure, reliable access to digital collections of research and scholarship. For a list of TDL participating institutions, please visit: http://tdl.org/member.” (Mumma, C.C., n.d.).

- **Dataverse**

  “Dataverse is an open source web application to share, preserve, cite, explore, and analyze research data. It facilitates making data available to others, and allows you to replicate others' work more easily...A Dataverse repository is the software installation, which then hosts multiple virtual archives called Dataverses. Each dataverse contains datasets, and each dataset contains descriptive metadata and data files (including documentation and code that accompany the data). As an organizing method, dataverses may also contain other dataverses” (King, G., n.d.).

- **Datasets**

  A dataset is a structured collection of related data that can be analyzed computationally (example: collection metadata in a tabular spreadsheet) (“What are the differences between data, a dataset, and a database?” n.d.).

- **Data Visualization**

  The process of creating graphical representations of data, including maps, charts, and graphs, which can be used to visually illustrate, identify, and analyze data (Camm, 2017, p.473).

- **DSpace**

“DSpace is an open source repository application that allows you to capture, store, index, preserve and distribute your digital material including text, video, audio and data” (Hollister, V., 2011).

- **Tableau Public**

  A free software platform for creating visual representations useful for data storytelling and analysis (“Digital Tools - Digital Humanities,” 2019).

- **Voyant Tools**

  Voyant Tools is an open-source application for text editing, analysis, and visualization.

- **OpenRefine**

  An open-source application for cleaning and refining data and data sets.

List of Tools

Data Extraction, Editing, and Cleaning Tools

- **Notepad++**

  An open-source text and source code editor tool for MS Windows.

- **Atom**

  An open-source text and source code editor tool for Mac OS, MS Windows, and Linux.

- **Microsoft Excel**

  Spreadsheet program for organizing, analyzing, and visualizing data in tabular format. Can use to clean and sort datasets using filters, find and replace, and formula functionality.

- **OpenRefine**

  An open-source application for cleaning “messy” data and transforming large datasets into other formats (https://openrefine.org).

- **Python**

An open-source programming language that can be used for data management, analysis, and visualization. Scripted and reproducible workflows can be developed to run on large datasets (https://guides.lib.utexas.edu/data-and-donuts/schedule).

- **Beautiful Soup**

  This tool is used for scraping data found on the web. It is based on Python and it pulls the information from HTML and XML encoded data.

- **Metadata Extraction Tool**

  Open-source tool developed by National Library of New Zealand that can be used for extracting metadata from a wide range of file formats (http://meta-extractor.sourceforge.net)

- **EMET (Embedded Metadata Extraction Tool)**

  Tool for extracting metadata embedded in JPEG and TIFF files.

- **ExifTool**

  Tool for extracting and editing metadata in image, audio, and video files.

- **Metadata++**

  Freeware software for extracting and editing metadata from image, audio, video, text, and other file formats.

- **DigiPres Commons - Tools**

  A comprehensive list of links to tools for a variety of functions that can be used for your institution's CAD project undertaking on the DigiPres Commons website (see resource and annotated bibliography list below).

Data Visualization Tools

- **Voyant Tools**

  A web-based, open-source application for text editing and analysis; useful for digital humanities scholars. (https://voyant-tools.org/docs/#!/guide/about)

- **Tableau Public**

  Free data visualization software application that enables a user to create interactive visualizations for data storytelling. (https://public.tableau.com/en-us/s/resources).

- **Gephi**

  Used for creating network visualizations of data.

- **R**

  This is an open source programming language used for visualizing data and performing statistical analysis.

- **ArcGIS**

  This is an online visualization tool for geographic information that maps, shares and analyzes data.

- **StoryMapJS**

  This is a free tool that allows you to visualize data using maps, images and textual information.

- **TimelineJS**

  This tool is also a free online website that lets you create timelines based on visual and textual data.

Annotated Bibliography

About Us - Digital Humanities. (2019, February 25). Retrieved November 21, 2020, from https://dh.ucla.edu/about/

UCLA has a Digital Humanities Program and their “About Us” page offers a helpful and condensed definition of DH practices and methods.

“All tools in detail.” (n.d.). DigiPres Commons. Retrieved November 21, 2020, from http://www.digipres.org/tools/all-tools/

List of links to a comprehensive list of community-owned digital preservation tools that can be sorted by function and used for many of the steps needed for embarking on a Collections as Data project (including tools for metadata extraction and scraping, data cleaning, and data visualization, among others).

Camm, J., Fry, M., & Shaffer, J. (2017). A practitioner’s guide to best practices in data visualization. *Interfaces (Providence)*, *47*(6), 473–488. https://doi.org/10.1287/inte.2017.0916

This resource is a tutorial about best practices for those interested in using data visualization and introduces some of the key concepts.

Candela, G., Sáez, M. D., Escobar Esteban, Mp., & Marco-Such, M. (2020). Reusing digital collections from GLAM institutions. *Journal of Information Science*. https://doi.org/10.1177/0165551520950246

This article provides a guide to creating Collections as Data from different datasets of cultural heritage institutions. The authors discuss the best practices to make data discoverable and accessible by looking at several examples.

Clement, T., & Carter, D. (2017). Connecting theory and practice in digital humanities information work. *Journal of the Association for Information Science and Technology*, *68*(6), 1385–1396. https://doi.org/10.1002/asi.23732

A study of digital humanities practices and current trends in methodology, goals, and research which includes interviews and surveys of graduate programs currently using DH practices and explores their motivations for doing so.

DigiPres Commons. (n.d.). Retrieved November 21, 2020, from https://www.digipres.org/

A website containing community-owned digital preservation resources. Compiles tips for getting started, a place to submit and look up answers to digital preservation questions, tools, a guide to understanding file formats, and preservation requirements and solutions.

Digital Tools - Digital Humanities. (2019, May 15). Retrieved November 21, 2020, from https://dh.ucla.edu/digital-tools/

A list of Digital Humanities tools on the UCLA website containing descriptions and links to the websites for the tools. The page includes short descriptions and links for the following tools: OpenRefine, Salty, Vectr, Carto, Tableau, Balsamiq, and Voyant.

Hendrigan, H. (2019). Mixing digital humanities and applied science librarianship: Using Voyant Tools to reveal word patterns in faculty research. *Issues in Science and Technology Librarianship*, (91). https://doi.org/10.29173/istl3.

This article represents another case study on how Voyant Tools can be a useful tool for data visualization and further research on collections. This paper explores the use of this tool with regards to subject librarians' work.

Hollister, V. (2011, December 02). What is DSpace? - DSpace KnowledgeBase. Retrieved November 21, 2020, from https://wiki.lyrasis.org/pages/viewpage.action?pageId=25467341

A wiki page written by Valorie Hollister on the Lyrasis website introducing DSpace with information about its uses, benefits, how it works, and how to install it, as well as software documentation and other helpful information.

King, G. (n.d.). About. Retrieved November 21, 2020, from https://dataverse.org/about

The "About" page for the Dataverse Project explains what a Dataverse is, the mission and goals of the project, history, and collaborators (institutions and individuals).

Kirschenbaum, M. (2012). What is digital humanities and what's it doing in English departments? In M. K. Gold (Ed.), *Debates in the Digital Humanities* (pp. 1-2). Minneapolis, MN: University of Minnesota Press. https://dhdebates.gc.cuny.edu/read/40de72d8-f153-43fa-836b-a41d241e949c/section/f5640d43-b8eb-4d49-bc4b-eb31a16f3d06#ch01

Mumma, C. C. (n.d.). About - Texas Data Repository user documentation. Retrieved November 21, 2020, from https://texasdigitallibrary.atlassian.net/wiki/spaces/TDRUD/pages/289144946/About.

TDL wiki page written by Courtney Mumma outlining reasons to deposit data into TDR, what can be deposited and what cannot be deposited in the repository, and an explanation with visuals of how dataverses work.

Neely, L., Luther, A., and Weinard, C. Cultural collections as data: Aiming for digital data literacy and tool development. *MW19: MW 2019*. Published February 1, 2019. Consulted November 18, 2020. https://mw19.mwconf.org/paper/cultural-collections-as-data-aiming-for-digital-data-literacy-and-tool-development/

This article discusses Collections as Data in a museum and cultural institutions context. It also mentions the importance of open data in collections and how having data accessible is a crucial part of working with Collections as Data.

Oglesby, N. (n.d.). Digital humanities tools and resources: Home. Retrieved November 21, 2020, from https://guides.lib.utexas.edu/digitalhumanities

University of Texas Libraries' LibGuide for digital humanities tools and resources with links to useful resources.

Padilla, T. (2016). “Collections as Data: Conditions of possibility.” In Collections as Data: Stewardship and Use Models to Enhance Access. https://www.thomaspadilla.org/2016/09/29/possibility/

Padilla’s closing keynote for the 2016 Library of Congress conference *Collections as Data: Stewardship and Use Models to Enhance Access* describes the ways in which Collections as Data promotes agency, empowerment, and ethics in working with data.

Padilla, T. (2017). On a collections as data imperative. UC Santa Barbara. https://escholarship.org/uc/item/9881c8sv

In this article, Padilla describes Collections as Data’s potential to fundamentally change the way libraries and users engage with collections. Padilla also provides a concise definition of data in a Collections as Data context: “ordered information, stored digitally, that are amenable to computation.”

Padilla, T. (2020). Always Already Computational - Collections as Data. Retrieved November 21, 2020, from https://collectionsasdata.github.io/

Website of Always Already Computational project (which ran from 2016 to 2018) with links to the final report and findings. The project researched approaches to applying computational methods to collections from cultural heritage institutions.

Padilla, T., Allen, L., Frost, H., Potvin, S., Russey Roke, E., & Varner, S. (2019, May 20a). 50 Things --- Always already computational: Collections as data. Zenodo. http://doi.org/10.5281/zenodo.3066237

This document is intended to provide practitioners with initial steps to take in order to begin supporting Collections as Data at cultural heritage institutions. The recommendations provide a starting point for institutions such as TDL to develop their own recommendations for working with Collections as Data.

Padilla, T., Allen, L., Frost, H., Potvin, S., Russey Roke, E., & Varner, S. (2019, May 20b). Santa Barbara statement on collections as data --- Always already computational: Collections as data (Version 2). Zenodo. http://doi.org/10.5281/zenodo.3066209

The Santa Barbara statement is one of the deliverables of the Always Already Computational project. The document outlines ten foundational Collections as Data principles, including commitments to shared documentation, open access, and transparency.

Padilla, T., Allen, L., Frost, H., Potvin, S., Russey Roke, E., & Varner, S. (2019, May 22). Final report --- Always already computational: Collections as data (Version 1). Zenodo. http://doi.org/10.5281/zenodo.3152935

The *Always Already Computational: Collections as Data* project's final report summarizes the project's Collections as Data framework, impacts, findings, and areas for further investigation. The final report's appendices provide the project deliverables, including the 50 recommendations document and the Santa Barbara statement on Collections as Data.

Padilla, T., Allen, L., Varner, S., Potvin, S., Russey Roke, E., Frost, H., & Heppler, J. (2017, March 09). Serendipitous Collections as Data. Retrieved November 21, 2020, from https://collectionsasdata.github.io/serendipity/

"Serendipitous Collections as Data" is a resource from the Already Always Computational site. It is a tool that randomly generates ideas (in the form of quotes) about collections as data and one can refresh the page to view a new idea. The ideas are drawn from partners of the project, the Collections as Data National Forum writings, and collections as data position statements from the site. This tool was coded by Jason Heppler.

Paradise, L. (2015). When you find out what digital humanities is, will you tell me? *The Serials Librarian 69*(2), 194-203. https://doi.org/10.1080/0361526X.2015.1036198

An essay that acknowledges the nebulous definitions of DH as a field and the confusion that can cause, while attempting to define what it means in terms of historical research. The essay covers digital archives for born digital content, the digital presentation of scholarly content, and working with historical materials in a digital environment (as well as the controversy that can cause).

Rethinking Collections as Data. (2019). In *Open a GLAM Lab* (1st ed.). https://doi.org/10.21428/16ac48ec.f54af6ae

This website provides an introduction on how to assess new collections, from thinking about copyright, accessibility, and reusability. It also explains some useful concepts regarding the description of collections.

Russey Roke, E. (2019, July 16). Collections as data. *bloggERS*. https://saaers.wordpress.com/2019/07/16/collections-as-data/

Russey Roke is a co-investigator of the *Collections as Data: Always Already Computational* project. In this blog post, Russey Roke provides a brief overview of Collections as Data. She states that Collections as Data approaches are archival and enhance access to archival material, and advises practitioners to “just start” working with their collections as data.

Shensky, M. (2020). Data & Donuts: Schedule. Retrieved November 21, 2020, from https://guides.lib.utexas.edu/data-and-donuts/schedule

University of Texas Libraries LibGuide page for the Data & Donuts series, which features links to recordings of webinars on data-related topics including using Python for data visualization, mapping data with R, and GIS operations. Compiled by Michael Shensky.

Sinclair, S., & Rockwell, G. (2016). Voyant Tools. Retrieved November 21, 2020, from https://voyant-tools.org/docs/

Open-source software for text editing, analysis, and visualization that can be used for DH data projects.

Women Techmakers (Directors). (2018, October 24). What are APIs and how to use them in 60 seconds! [Video file]. Retrieved November 21, 2020, from https://www.youtube.com/watch?v=BRAMRUxfiXY

Video from the Women Techmakers group introducing the concept of application programming interfaces (APIs) and how they are used.

Welcome! (n.d.). Retrieved November 21, 2020, from https://openrefine.org/

Official website for OpenRefine tool for cleaning and transforming data with videos covering the following topics: “Explore Data,” “Clean and Transform Data,” and “Reconcile and Match Data.” Website also includes the software documentation.

What are the differences between data, a dataset, and a database? (n.d.). Retrieved November 21, 2020, from https://www.usgs.gov/faqs/what-are-differences-between-data-a-dataset-and-a-database?qt-news_science_products=0

United States Geological Survey (USGS) webpage that contains a helpful description of the distinction between data, datasets, and databases.

Wittmann, R., Neatrour, A., Cummings, R., & Myntti, J. (2019). From digital library to open datasets: Embracing a “collections as data” framework. *Information Technology and Libraries*, *38*(4), 49-61. https://doi.org/10.6017/ital.v38i4.11101

Wittmann et al. speak of a “collections as data mindset,” which is a helpful way of conceptualizing Collections as Data (p. 51). The authors examine five case studies that use a variety of Collections as Data approaches, such as text mining, GIS, and topic modeling.