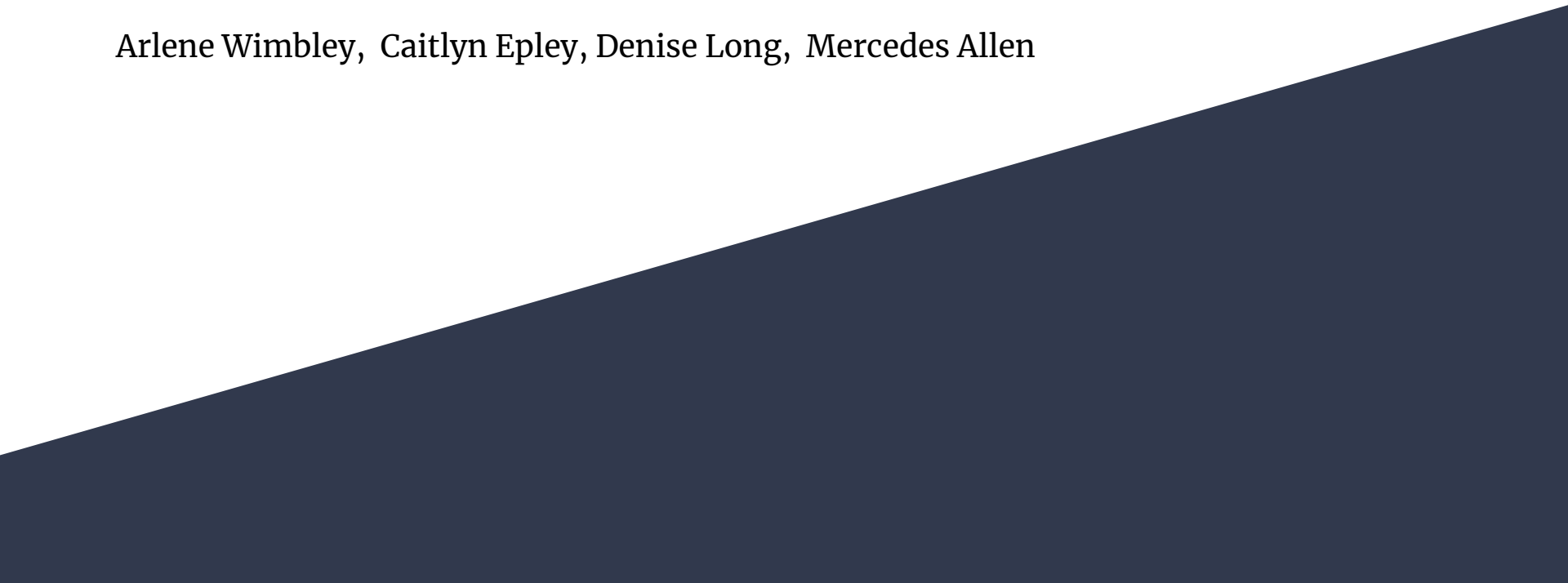


Library of Congress Relational Database

Arlene Wimbley, Caitlyn Epley, Denise Long, Mercedes Allen

A dark blue diagonal gradient bar that starts from the bottom left and extends towards the top right, covering the lower half of the slide.

What is the Library of Congress?



The Library of Congress is the largest library in the world, with millions of books, films and video, audio recordings, photographs, newspapers, maps and manuscripts in its collections. The Library is the main research arm of the U.S. Congress and the home of the U.S. Copyright Office.

Project Purpose

The purpose of our project was to transform a large, unwieldy flat-file of digital library items into a relational database to both save space and to form more meaningful relationships among the data.

Storing library data in a relational database, rather than a flat file, helps to prevent updating, deleting, and inserting errors by minimizing the number of times each data entry appears in the database.

Dataset of World Digital Library of Congress

<https://www.loc.gov/item/2020446966/>

- We condensed this dataset because it was too large to upload to GitHub.
- We removed some of the rows of data from the end of the file to make it smaller
- We also removed some of the 'less significant' columns (Ex: 'original language' was removed because we had another language column to represent that data)
- Our final dataset had 23,861 rows and 99 columns

Snippet from flat file

- This is what the file we started with looked like. It is too complicated to draw any meaningful conclusions from

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	
1	wdl_id	title	collection	description	wdl_url	Type of Item	Institution	Photograph	Date Crea	Subject	De Place	Dewey	De	Additional	Physical	D Creator	Publisher	Language	Artist	Cartograph	Note	Delineator	Author	Surveyor	Compiler	Translator	Contribut
2		1 Antietam, Maryland. At the out	https://ww	Prints, Phc	Library of C Gardner,	A 1862-10-0	1862-10-0	North Ame	973	Antietam,	1 negative : glass, wet collodion																
3		2 Chola Wor Frank and This photo	https://ww	Prints, Phc	Library of C Vargas, Ma	1911	1900/1923	Latin Ame	391	Portrait photographs Women																	
4		3 Maps of Ezo, Sakhalin This map v	https://ww	Maps	Library of Congress	1854	1854	East Asia >	912	1 color ma	Fujita, Ton Harimaya Japanese																
5		4 Warrior As Japanese f The Japan	https://ww	Prints, Phc	Library of Congress	1862	1862	East Asia > 769 792	Actors Co	1 print: woodcut, color ; 29.0 x 21 Japanese								Utagawa, Toyokuni, 1786-1865									
6		5 Manuscript Map of De This beaut	https://ww	Maps	Library of Congress	1764	1764	Latin Ame	912	Rivers	1 manuscript map : color ; 60 x 8 Spanish																
7		6 Map of a P. Henry Har Joan Vinck	https://ww	Maps	Library of Congress	1650	1650/1651	Latin Ame	912	Manuscrip	1 manuscript map : c Dutch Wee Dutch Spanish								Vinckeboons, Joan, 1617-1670								
8		7 Map of the World This late 1	https://ww	Maps	National Library of Br	1775	1775/1776	World	912	World map	Hand-col' Isle, Guil Tobias Cor																
9		8 House of k Japanese f The term u	https://ww	Prints, Phc	Library of Congress	1804	1804	East Asia > 306 769	Brothels C	1 print : woodcut, color ; 17.7 x 2 Japanese								Kitagawa, Utamaro, 1	From the series: SeirÅ ehon nunchÅ gyÅji : Yearly activities of the Green hou								
10		9 The Actor Japanese f The Japan	https://ww	Prints, Phc	Library of Congress	1818/1830	1818/1830	East Asia >	792	Kato, Kiyor	1 print: woodcut, color ; 38.2 x 2 Japanese								ShunkÅsai, HokushÅ, flourished 1810-1850								
11		10 An Actor Japanese f The Japan	https://ww	Prints, Phc	Library of Congress	1850	1849/1852	East Asia > 294 792	Actors Dir	1 print (2 sheets): woodcut, colo	Japanese								Utagawa, Kuniyoshi, 1798-1861								
12		11 Peony and Japanese f The Japan	https://ww	Prints, Phc	Library of Congress	1833	1833/1834	East Asia > 598 635	Canaries	1 print: woodcut, color ; 19.2 x 1 Japanese								Katsushika, Hokusai, 1760-1849									
13		12 Suffrage P. United Sta The suffra	https://ww	Prints, Phc	Library of Congress	4510	1910/1920	North Ame	323 324	Civil rights Civil rights demonstr	American English																
14		13 Girl's Day Japanese f The Japan	https://ww	Prints, Phc	Library of Congress	1726	1716/1736	East Asia >	394	Dolls Eati	1 print: woodcut ; 21.8 x 15.2 cer	Japanese							Nishikawa, Sukenobu, 1671-1751								
15		14 Thomas Je United Sta Thomas Je	https://ww	Prints, Phc	Library of Congress	1805	1800/1810	North Ame	973	Jefferson,	1 print : engraving								Saint-MÅ@min, Charles Balthazar Julien Fevret de, 1770-1852								
16		15 Two Jewish Frank and This photo	https://ww	Prints, Phc	Library of Congress	1900/1920	1900/1920	Middle Eas	391	Women																	
17		17 United Sta United Sta Construct	https://ww	Prints, Phc	Library of Congress	1834	1834	North Ame	743 975	Architectu	1 drawing : ink, watercolor, and v	English							Davis, Alexander Jackson, 1803-1892								
18		18 First Nerchinsk Regim The First N	https://ww	Books	Russian State Library	1907	1898/1906	East Asia >	355	Armies Ba	177 pages	Makovin, Å	Associatio	Russian													
19		19 7th War Loan. Now--Å C. C. Beall	https://ww	Prints, Phc	Library of Congress	1945	1945	East Asia > 332 940	Iwo Jima, E	1 photomechanical pl	U.S. Gover	English							Beall, C. C., 1892-1967								
20		20 A Chart of the Gulf Str This map, i	https://ww	Maps	Library of Congress	1786	1786	North Ame	551 911	Gulf Strear	21 x 26 cer	Poupard, J	American English														
21		21 A General Chart of the Captain Jo	https://ww	Maps	Library of Congress	1796-06-0	1796	Latin Ame	912	Caribbean	1 map : co Speer, Jos Robert Wil	English															
22		22 The Island and City of Jan Huygh	https://ww	Maps	National Library of Br	1595	1595	Central an	912	1 view ; 48	Linschoten, Jan Huyg	Portuguese															
23		23 A Journal c American The Ameri	https://ww	Books	Library of Congress	1783	1776/1778	World	910	Cook, Jam	208 pages : illustratio	Printed an	English														
24		24 A New Map of Nova Sc Thomas Je	https://ww	Maps	Library of Congress	1775	1775	North Ame	912	Atlantic Cc	1 map : co Jefferys, Ti	Robert Say	English														
25		25 A Plan of the Rosalij C France an	https://ww	Maps	Library of Congress	1776	1776	Latin Ame	912	1 color map, 125 x 92	centimeter	English															
26		26 A Voyage Down the Ar Perry McD	https://ww	Books	Library of Congress	1860	1856/1860	East Asia >	915	Amur River	390 pages	Collins, Pe D.	Appleto	English													
27		27 Abosko-B'v Geographi This card i	https://ww	Prints, Phc	National Library of Ru	1856	1856	Europe > F	914	Meeting of Frontiers																	
28		28 Aesop's Fc Lessing J. I This is the	https://ww	Books	Library of Congress	1479	1479	Europe	398	AesopÅcÅ 37, cxviii, 15 leaves :	Anton Sorg	German															
29		29 Jackie Rob United Sta Jack Roos	https://ww	Prints, Phc	Library of C Sandberg,	1954	1947/1957	North Ame	796	Baseball Baseball players Brooklyn	Dodge	English															

ETHICAL CONSIDERATION FOR DATA ENGINEERING – BEST PRACTICES

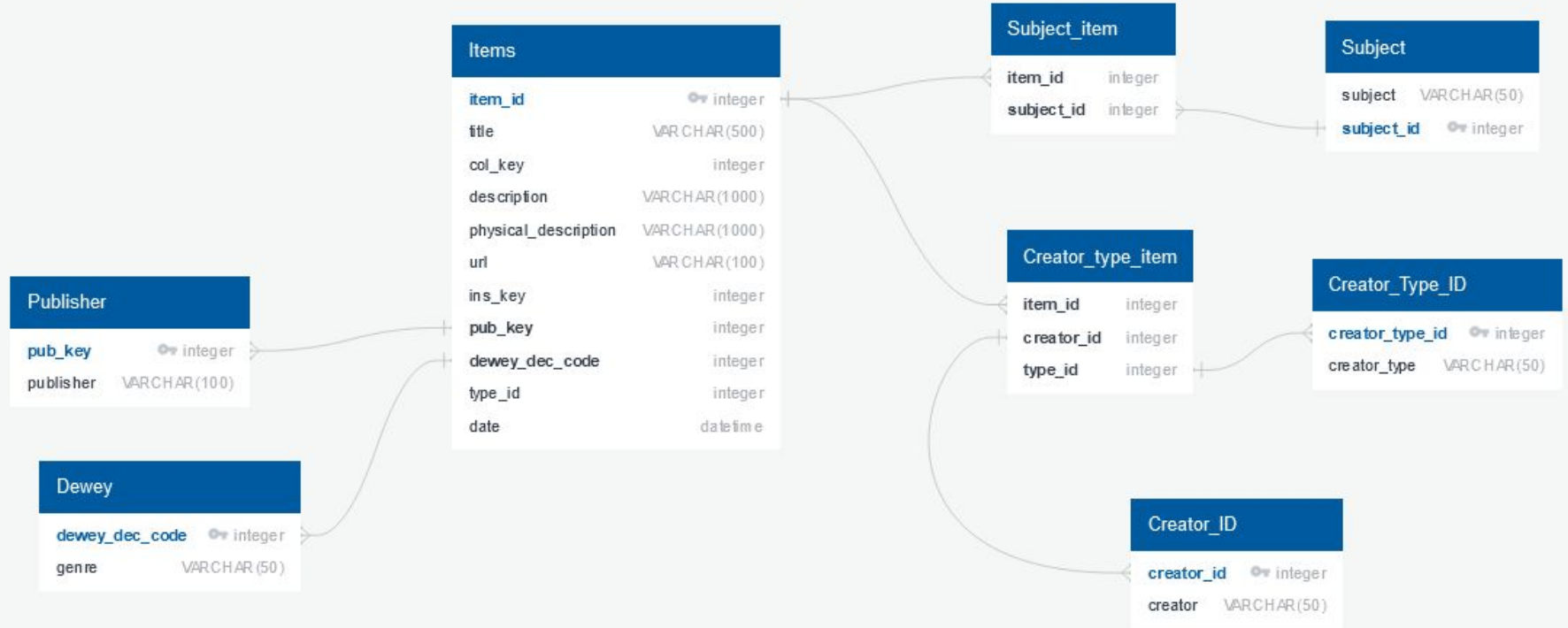
- Ownership
- Transparency
- Privacy
- Intention
- Outcomes

The dataset we used for our project did not contain any personally identifiable information (PII) other than the names of the creators of the items included in the dataset. However, since there wasn't any other PII (like contact information, addresses, or social security numbers) we did not take any special measures to protect the creator's information.

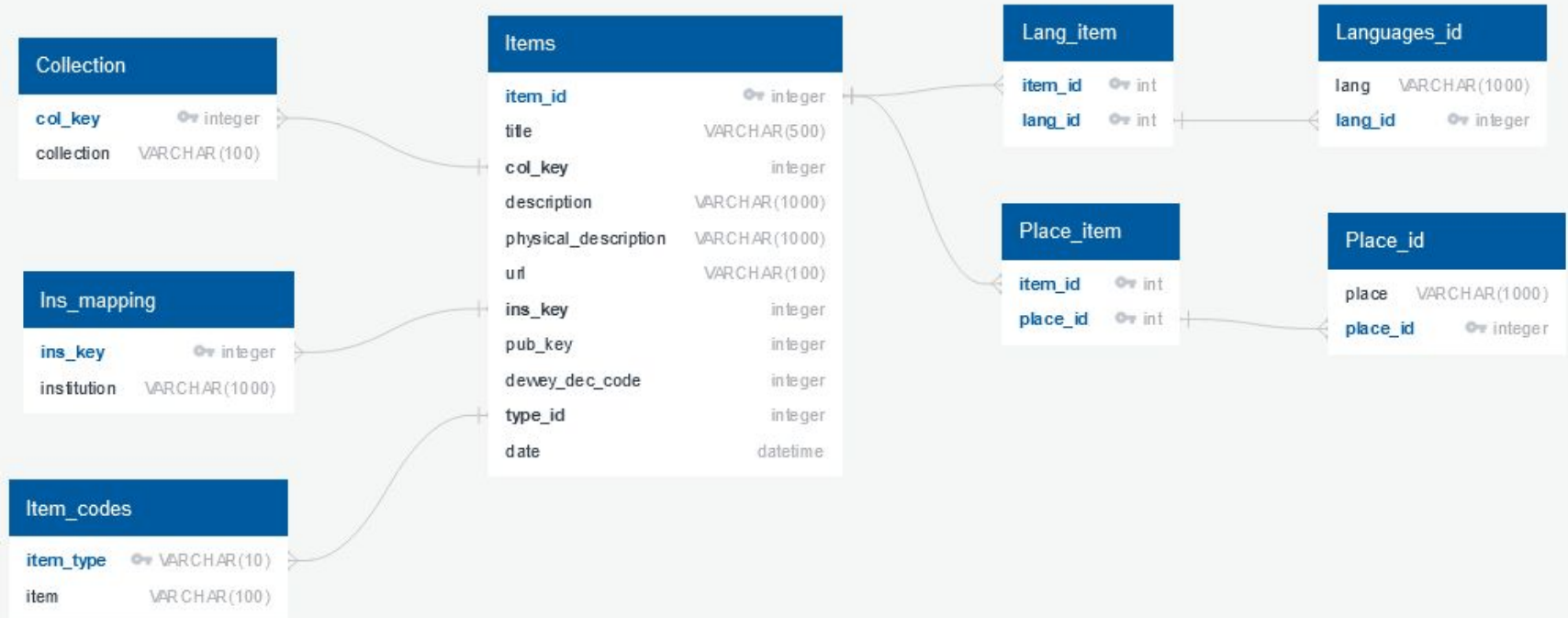
We chose to use a relational database because the data from our source file had some structure to it already. Each item had very similar characteristics and shared most of the features from the dataset, so it made more sense to use a relational database with a rigid structure rather than a non-relational database like NoSQL. To implement the relational database we used Postgres in the form of pgAdmin4 because everyone in the group had already used it before in class and there was no cost associated with it.

Database Choice & Reasoning

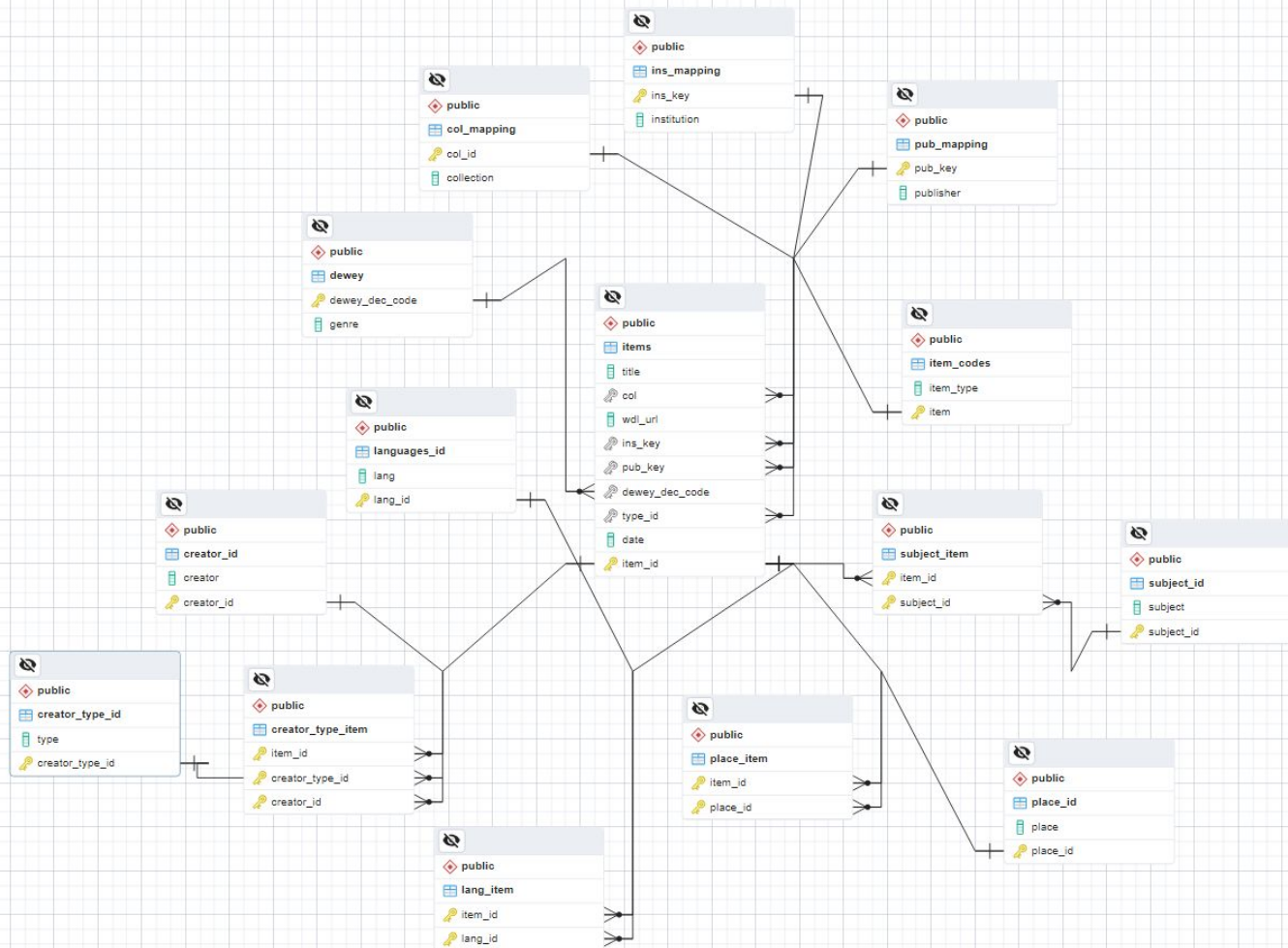
Simple ERDs (part 1)



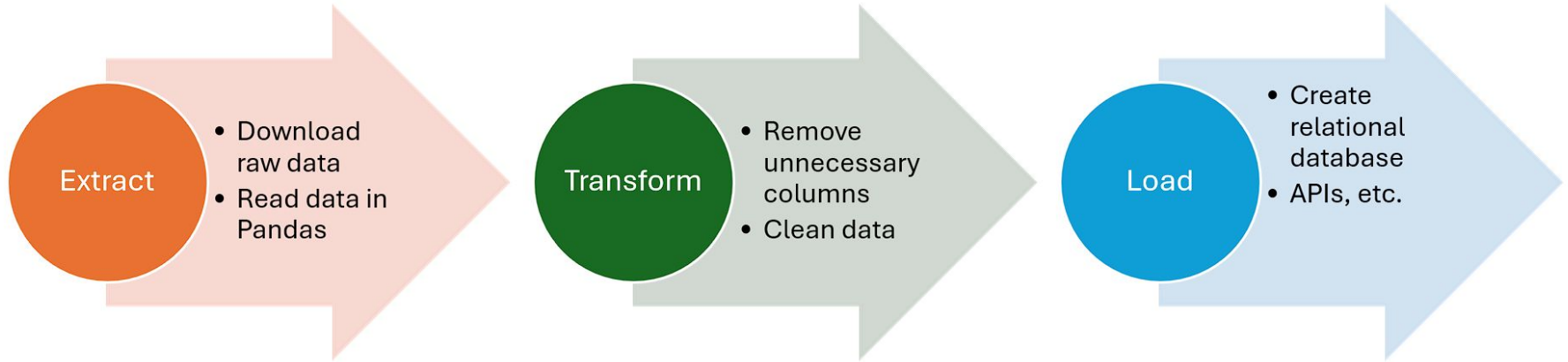
Simple ERDs (part 2)



ERD



ETL Process



Dask Python Library

- In keeping with the requirements of the project, DASK was the chosen library that we were not taught
- Chosen because it has the feel of working like Pandas
- Was able to read the .csv with DASK; instead of `pd.read` as in Pandas, it uses `dd.read`
- Was only able to get it to output the column headers and data types
- Time constraint in learning to fully incorporate it

ETL Workflows

- We needed to split the large flat file (the dataset from loc.gov) into relations that could be used in our database
- To do this, we used ETL workflows to read the dataset into Pandas Dataframes, manipulate it, and export the final product into corresponding .csv files
- We combined all of the creator types and their names into one table, rather than the multiple that were in the original file

	Subject	item_id	subject_id
0	Antietam, Battle of, Maryland, 1862	0	0
1	Antietam, Battle of, Maryland, 1862	11930	0
2	Generals	1011	1
3	Generals	1016	1
4	Generals	1017	1
...
113573	Popular culture	23859	4640
113574	Mohammad Daoud, Sardar, 1909-1978	11920	4641
113575	Mohammad Daoud, Sardar, 1909-1978	23850	4641
113576	Israel-Arab War, 1973	11929	4642
113577	Israel-Arab War, 1973	23859	4642

113578 rows × 3 columns

	Creator	item_id	type
0	Gardner, Alexander, 1821-1882	0	Photographer
1	Vargas, Max T., 1874-1959	1	Photographer
2	Sandberg, Bob	27	Photographer
3	Brumfield, William Craft, 1944-	31	Photographer
4	Ferrez, Marc, 1843-1923	41	Photographer
...
26749	Vahhāj, Sirāj al-Dīn	23850	Lead
26750	Vahhāj, Sirāj al-Dīn	23851	Lead
26751	Vahhāj, Sirāj al-Dīn	23852	Lead
26752	Vahhāj, Sirāj al-Dīn	23853	Lead
26753	Vahhāj, Sirāj al-Dīn	23854	Lead

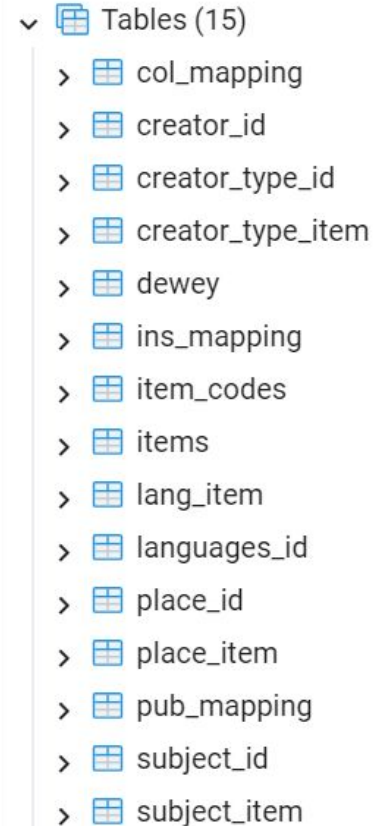
26754 rows × 3 columns

Final .csv files

- Dewey.csv
- col_mapping_new.csv
- creator_id.csv
- creator_type_id.csv
- creator_type_item.csv
- ins_mapping_new.csv
- item_codes.csv
- items.csv (Note: this file has '|' as the delimiter, rather than ',')
- items_csv (this is the same as the items.csv, except that this file has ',' as the delimiter)
- lang_item.csv
- languages_id.csv
- place_id.csv
- place_item.csv
- pub_mapping_new.csv
- subject_id.csv
- subject_item.csv

Implementing a Database in pgAdmin4

- To implement the database, we created SQL schema to create tables and then imported our final .csv files into the tables
- Once the data is imported, SQL statements can be used to view, manipulate, and edit the data



Ex Query

```
4 select title, subject from items i join subject_item si on i.item_id = si.item_id
5 left join subject_id s on s.subject_id = si.subject_id
```

Data Output

Messages

Notifications



	title character varying (1000)	subject character varying (1000)
69	Special Edition of Universal Current Events: Late...	Military camps
70	Special Edition of Universal Current Events: Late...	Military camps
71	View of the Encampment of the Corn Exchange ...	Military camps
72	View of the Encampment of the Corn Exchange ...	Military camps
73	Grand Review	Military camps
74	Grand Review	Military camps
75	President George Washington	Presidents
76	President Abraham Lincoln	Presidents
77	President Millard Fillmore	Presidents
78	President Jefferson Davis, Confederate States of...	Presidents
79	President George Washington	Presidents
80	President Abraham Lincoln	Presidents
81	President Millard Fillmore	Presidents
82	President Jefferson Davis, Confederate States of...	Presidents
83	Special Edition of Universal Current Events: Late...	Tents

Flask API

- To store the data for future use, we implemented a Flask API that displays the data in Json format via a browser
- All of the tables are available through the hyperlinks under 'Available Routes'

Available Routes:

[/api/v1.0/col_mapping](#)

[/api/v1.0/creator_id](#)

[/api/v1.0/creator_type_id](#)

[/api/v1.0/creator_type_item](#)

[/api/v1.0/dewey](#)

[/api/v1.0/ins_mapping](#)

[/api/v1.0/item_codes](#)

[/api/v1.0/items](#)

[/api/v1.0/lang_item](#)

[/api/v1.0/languages_id](#)

[/api/v1.0/place_id](#)

[/api/v1.0/place_item](#)

[/api/v1.0/pub_mapping](#)

[/api/v1.0/subject_id](#)

[/api/v1.0/subject_item](#)

CONCLUSION

In summation, pertinent information was extracted from the Library of Congress database to draw aggregate findings and design a database useful for the future utilization.

- *ETL workflows were implemented.*
- *SQL, Pandas, JSON, Flask API*
- *Dask Library*
- *Python Scripts*
- *Jupyter Notebook*
- *Final data was uploaded into Postgres to create a relational Database to aid in making it more user friendly.*

Sources

- Dask documentation: <https://docs.dask.org/en/stable/>
- Dataset source: <https://www.loc.gov/item/2020446966/>