Data Management and Cleaning with Open Refine

Tuesday Morning
MJC-BSANA-Data-Workshop-2024

Session outline:

1. Recap from Session 1
2. Data acquisition and integration
3. Data management
4. Data cleaning and preparation
5. OpenRefine tutorial
    a. Explore and evaluate the data to find inconsistencies.
    b. Clean & transform the data for analysis
    c. Undo/Redo & using operations templates
    d. Export data for further analysis or visualisation

Recap

1. Why is metadata important?
2. Types of databases
3. Examples of metadata schemas
4. Data validation
5. Controlled Vocabularies
6. VREs

Data acquisition and integration

- Find your data sources
- Evaluate and review - What is in there? What is not?
- Define ways to extract/collect/integrate the data
- Define a data management plan
  - Where is the data going to sit?
  - How are you going to analyse/query it?
  - What questions do you want to ask?
  - How are you going to visualise/interpret the results.

Data acquisition and integration

1. Identification of relevant data sources
2. Surveying of data sources
3. Data extraction

Find your data sources

Electronic Resources Useful for Byzantinists:

- Database of Byzantine Book Epigrams (https://www.dbbe.ugent.be/)
- Dumbarton Oaks Byzantine Seals Collection (https://www.doaks.org/resources/seals)
- Dumbarton Oaks Byzantine Coins Collection (https://www.doaks.org/resources/coins)
- Mapping Past Societies (https://darmc.harvard.edu/)
- Pinakes (https://pinakes.irht.cnrs.fr/)
- Prosopographie der mittelbyzantinischen Zeit Online (https://www.degruyter.com/database/pmbz/html?lang=de)
- Prosopography of the Byzantine World (https://pbw2016.kdl.kcl.ac.uk/)
- Pyle. A Gateway to Greek Manuscripts (http://pyle.it/)
- Suda On Line (https://www.cs.uky.edu/~raphael/sol/sol-html/)
- TLG (https://stephanus.tlg.uci.edu/)

Some Museums and Libraries with Important Byzantine Collections

- Bibliothèque nationale de France
- The British Library
- The British Museum
- Dumbarton Oaks
- Louvre
- Metropolitan Museum of Art

Data extraction

- Process of gathering and processing data from a range of sources (eg. databases, publications or APIs), and transforming it into a preferred machine-readable format such as JSON, CSV, and XML for further processing in the software of your choice.
  - Structured Data: highly organized and formatted
  - Unstructured Data: without a preferred data model or structure
  - Physical Sources: physical documents.
  - Digital Sources: data generated or stored in digital formats.

Data extraction: data export

Place: **Byzantium**

Hide filters

Sort by

Object name

On display only

Image only

Enter a keyword

Production date

Person/organisation (70 terms)

Place (28 terms)

Object name (11 terms)

Culture/period/dynasty (4 terms)

coin

1853,0512.1 | Greek | 3rdC BC-2ndC BC | Minted in: Byzantium

coin

1921,0110.7 | Greek | 4thC BC-3rdC BC | Minted in: Byzantium

coin

1947,0606.1246 | Greek | 4thC BC-3rdC BC | Minted in: Byzantium

coin

1866,1201.759 | Greek | 240BC-220BC | Minted in: Byzantium

coin

1866,1201.758 | Greek | 1stC | Minted in: Byzantium

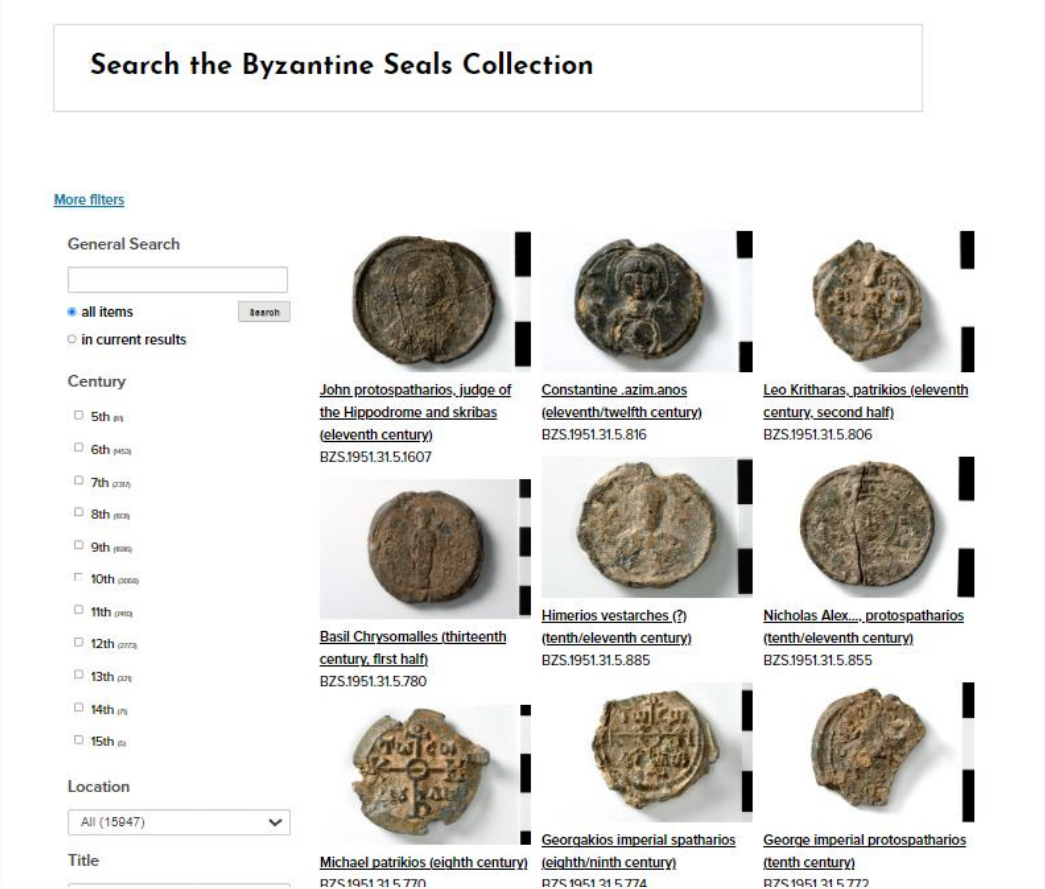
coin

1882,0506.3 | Greek | 3rdC BC-2ndC BC | Minted in: Byzantium

Data extraction: data export

Image	Object type	Museum r	Title	Denomina	Escapeme	Descriptic	Producer	School/st	State	Authority	Ethnic nar	Ethnic nar	Culture	Productio	Productio	Find spot	Materials	Ware
https://media.britishmuseum.org/me	coin	No: 1853,0512.1				Alloy coin.; Head.; Cornucopiae.							Greek	3rdC BC-2	Minted in: Byzantiur		alloy	
https://media.britishmuseum.org/me	coin	No: 1921,0110.7				Alloy coin.; Head.; Bull on dolphin.							Greek	4thC BC-3	Minted in: Byzantiur		alloy	
https://media.britishmuseum.org/me	coin	No: 1947,0606.1246				Alloy coin.; Head.; Bull on dolphin.							Greek	4thC BC-3	Minted in: Byzantiur		alloy	
https://media.britishmuseum.org/me	coin	No: 1866,1201.758				Alloy coin.; Head.; Obelisk.							Greek	1stC	Minted in: Byzantiur		alloy	
https://media.britishmuseum.org/me	coin	No: 1882,0506.3				Alloy coin.; Head.; Tripod.							Greek	3rdC BC-2	Minted in: Byzantiur		alloy	
https://media.britishmuseum.org/me	coin	No: 1866,1201.759				Alloy coin.; Head.; Trident.							Greek	240BC-220	Minted in: Byzantiur		alloy	
https://media.britishmuseum.org/me	coin	No: 1896,0601.28		reduced	tetradrachn	Silver coin.; Head.; Seated figure.							Greek	240BC-220	Minted in: Byzantiur		silver	
https://media.britishmuseum.org/me	coin	No: 1949,0411.243		triobol		Silver coin.; Forepart of bull on dolphin.; Trident.							Greek	387BC-340	Minted in: Byzantiur		silver	
https://media.britishmuseum.org/me	coin	No: HPB,p28.1.B		sigloi		Silver coin.; Bull on dolphin.; Incuse.							Greek	340BC-320	Minted in: Byzantiur		silver	
https://media.britishmuseum.org/me	coin	No: TC,p107.14.Lys				Silver coin.			Kingdom (	Lysimachus			Greek		Minted in: Byzantiur		silver	
https://media.britishmuseum.org/me	coin	No: 1910,1104.99				Silver coin.			Kingdom (	Lysimachus			Greek		Minted in: Byzantiur		silver	
https://media.britishmuseum.org/me	coin	No: 1920,0805.30				Copper alloy coin.; Head.; Trident.							Greek	240BC-220	Minted in: Byzantiur		copper alloy	
https://media.britishmuseum.org/me	coin	No: 1924,0708.11				Gold coin.			Kingdom (	Lysimachus			Greek		Minted in: Byzantiur		gold	
https://media.britishmuseum.org/me	coin	No: 1922,1115.16				Gold coin.			Kingdom (	Lysimachus			Greek		Minted in: Byzantiur		gold	
https://media.britishmuseum.org/me	coin	No: G.585				Silver coin.			Kingdom (	Lysimachus			Greek		Minted in: Byzantiur		silver	
https://media.britishmuseum.org/me	coin	No: 1841,B.425		triobol		Silver coin.; Forepart of bull on dolphin.; Trident.							Greek	387BC-340	Minted in: Byzantiur		silver	
https://media.britishmuseum.org/me	coin	No: RPK,p96F.2.Byz		diobol		Silver coin.; Bull on dolphin.; Incuse.							Greek	340BC-320	Minted in: Byzantiur		silver	
https://media.britishmuseum.org/me	coin	No: TC,p107.11.Lys				Silver coin.			Kingdom (	Lysimachus			Greek		Minted in: Byzantiur		silver	
https://media.britishmuseum.org/me	coin	No: 1935,0619.22				Silver coin.			Kingdom (	Lysimachus			Greek		Minted in: Byzantiur		silver	
https://media.britishmuseum.org/me	coin	No: 1949,0411.241		diobol		Silver coin.; Bull on dolphin.; Incuse.							Greek	340BC-320	Minted in: Byzantiur		silver	
https://media.britishmuseum.org/me	coin	No: 1923,1103.1				Silver coin.			Kingdom (	Lysimachus			Greek		Minted in: Byzantiur		silver	
https://media.britishmuseum.org/me	coin	No: BNK,G.214				Gold coin.			Kingdom (	Lysimachus			Greek		Minted in: Byzantiur		gold	
https://media.britishmuseum.org/me	coin	No: TC,p107.12.Lys				Silver coin.			Kingdom (	Lysimachus			Greek		Minted in: Byzantiur		silver	
https://media.britishmuseum.org/me	coin	No: 1911,0706.10				Silver coin.			Kingdom (	Lysimachus			Greek		Minted in: Byzantiur		silver	
https://media.britishmuseum.org/me	coin	No: 1866,1201.752		diobol		Silver coin.; Bull on dolphin.; Incuse.							Greek	340BC-320	Minted in: Byzantiur		silver	
https://media.britishmuseum.org/me	coin	No: 1896,0703.48		reduced	tetradrachn	Silver coin.; Head.; Seated figure.							Greek	240BC-220	Minted in: Byzantiur		silver	
https://media.britishmuseum.org/me	coin	No: 1911,0704.174				Silver coin.			Kingdom (	Lysimachus			Greek		Minted in: Byzantiur		silver	
https://media.britishmuseum.org/me	coin	No: 1910,1105.52				Silver coin.			Kingdom (	Lysimachus			Greek		Minted in: Byzantiur		silver	
https://media.britishmuseum.org/me	coin	No: 1921,0110.5				Copper alloy coin.; Head.; Standing figure.							Greek	240BC-220	Minted in: Byzantiur		copper alloy	
https://media.britishmuseum.org/me	coin	No: BNK,G.228				Silver coin.			Kingdom (	Lysimachus			Greek		Minted in: Byzantiur		silver	
https://media.britishmuseum.org/me	coin	No: 1921,0520.2				Copper alloy coin.; Head.; Seated figure.							Greek	3rdC BC-2	Minted in: Calchedo		copper alloy	
https://media.britishmuseum.org/me	coin	No: 1934,0523.3		tetradrachm		Silver coin.; Bull on dolphin, trident.; Incuse square.							Greek	387BC-340	Minted in	Excavated	silver	
https://media.britishmuseum.org/me	coin	No: RPK,Gre.42				Copper alloy coin.; Head of Apollo.; Tripod.							Greek	3rdC BC-2	Minted in: Byzantiur		copper alloy	
https://media.britishmuseum.org/me	coin	No: 1896,0601.27		tetradrachm		Silver coin.; Bull on dolphin.; Incuse square.							Greek	387BC-340	Minted in: Byzantiur		silver	
https://media.britishmuseum.org/me	coin	No: BNK,G.213				Gold coin.			Kingdom (	Lysimachus			Greek		Minted in: Byzantiur		gold	
https://media.britishmuseum.org/me	coin	No: BNK,G.221				Silver coin.			Kingdom (	Lysimachus			Greek		Minted in: Byzantiur		silver	
https://media.britishmuseum.org/me	coin	No: 1866,1201.860				Silver coin.			Kingdom (	Lysimachus			Greek		Minted in: Byzantiur		silver	

Data extraction: web scraping



Data extraction: web scraping

Python is popular for web scraping, some of the best libraries are:

- **BeautifulSoup** allows you to parse HTML and XML documents. Using API, you can easily navigate through the HTML document tree and extract tags, meta titles, attributes, text, and other content.
- **Requests** simple yet powerful Python library for making HTTP requests. It is designed to be easy to use and intuitive, with a clean and consistent API. With Requests, you can easily send GET and POST requests, and handle cookies, authentication, and other HTTP features.
- **Selenium** allows you to automate web browsers such as Chrome, Firefox, and Safari and simulate human interaction with websites. You can click buttons, fill out forms, scroll pages, and perform other actions. It is also used for testing web applications and automating repetitive tasks.

- **Pandas** allow storing and manipulating data in various formats, including CSV, Excel, JSON, and SQL databases. Using Pandas, you can easily clean, transform, and analyze data extracted from websites.

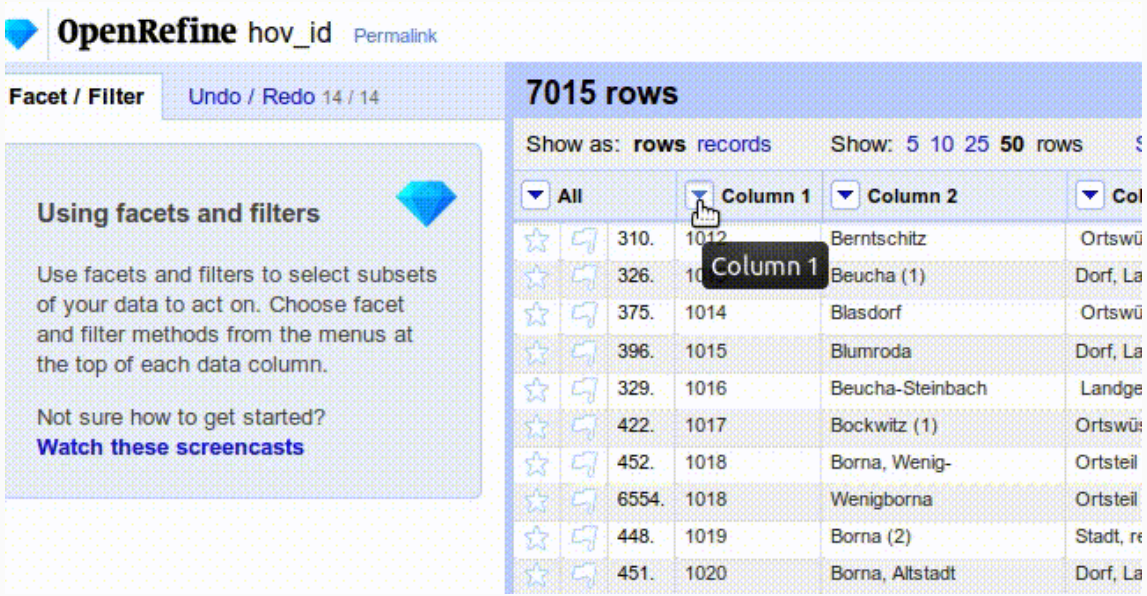
Edited from: https://nanonets.com/blog/web-scraping-with-python-tutorial/

Data extraction: text to data

Optical Character Recognition (OCR): method to convert an image of text into a machine-readable text format.

- Google Docs OCR free alternative offered by Google Drive to convert images to text. The process behind Google Docs OCR is nothing more than uploading images from which you need the data, into Google Docs, and exporting the data as a text format into your computer.
- FreeOCR is a Windows OCR application that uses the Tesseract OCR engine. It can handle PDF files and images.

Data management with OpenRefine



- Powerful tool to work with messy data: clean it, transform it from one format to another and extend it with web services and external data.

- It works by running a small server on your computer and tuning the search engine to execute the orders and interact with the tool.

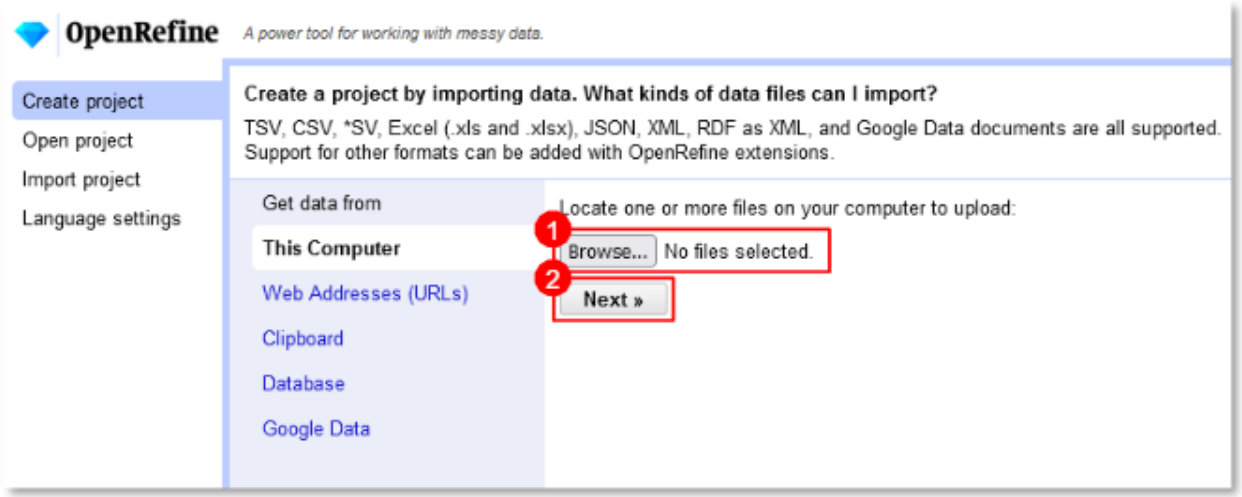
Create Project

- **Create Project:** loads a dataset into OpenRefine. There are various supported formats, and you can import data in different ways:
  - This Computer: Select a file stored on your local machine
  - Web Addresses (URLs): Import data directly from an online source
  - Clipboard: Copy-paste your data into a text field
  - Google Data: Enable access to a Google Spreadsheet or Fusion Table.

Create Project

On the browser page, import the *.CSV file provided and create a project.

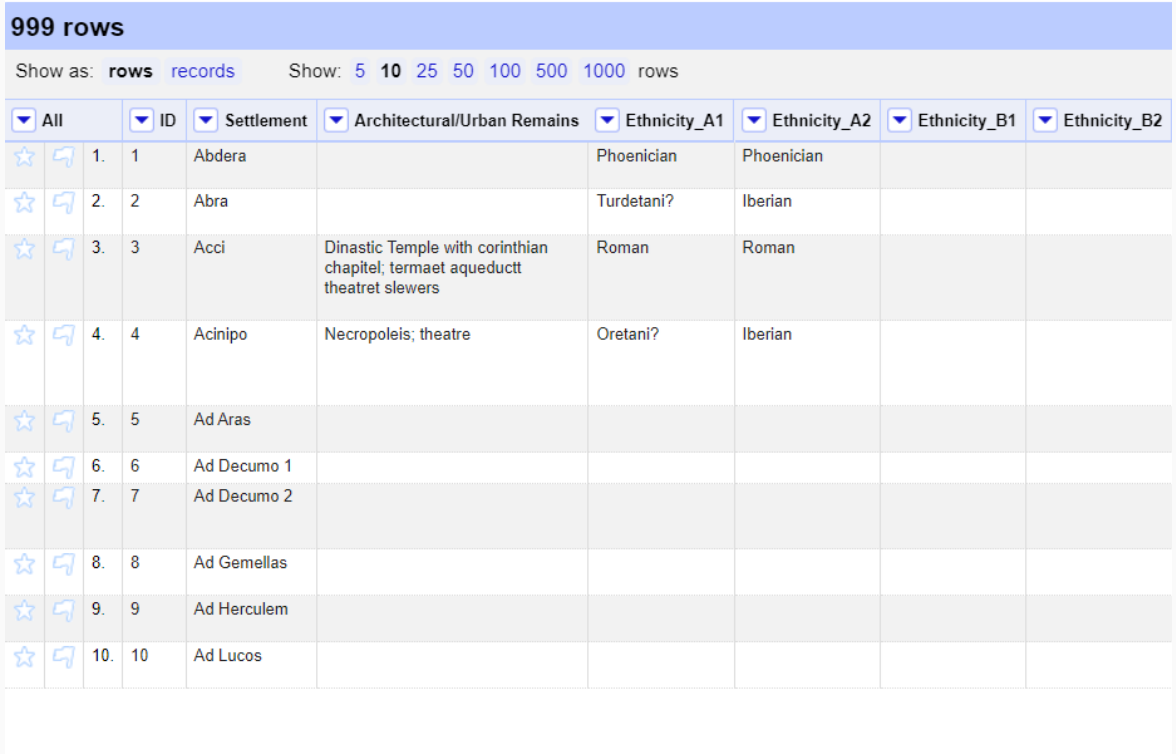
A subset of data for archaeological settlements and sculptures from the Roman province of Baetica.



Open Project

- **Open Project:** This option helps you go back to an existing project created during a former session.
- **Import Project:** directly import an existing OpenRefine project archive.

Explore your data



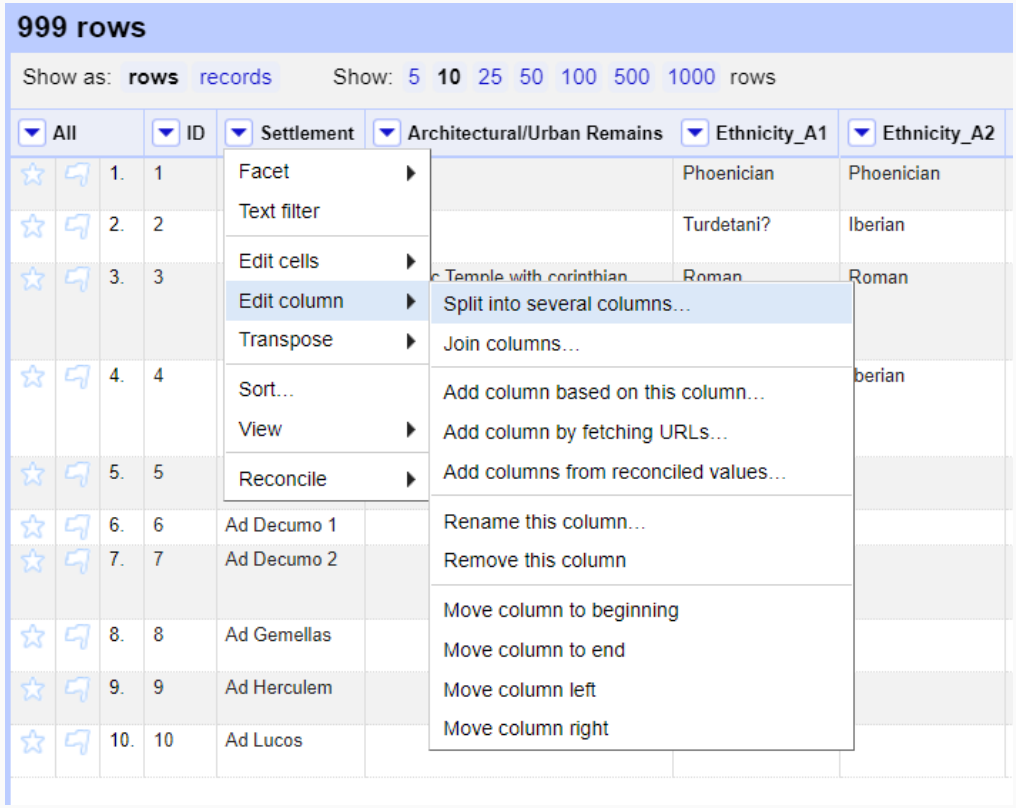
999 rows

Show as: **rows** records Show: 5 **10** 25 50 100 500 1000 rows

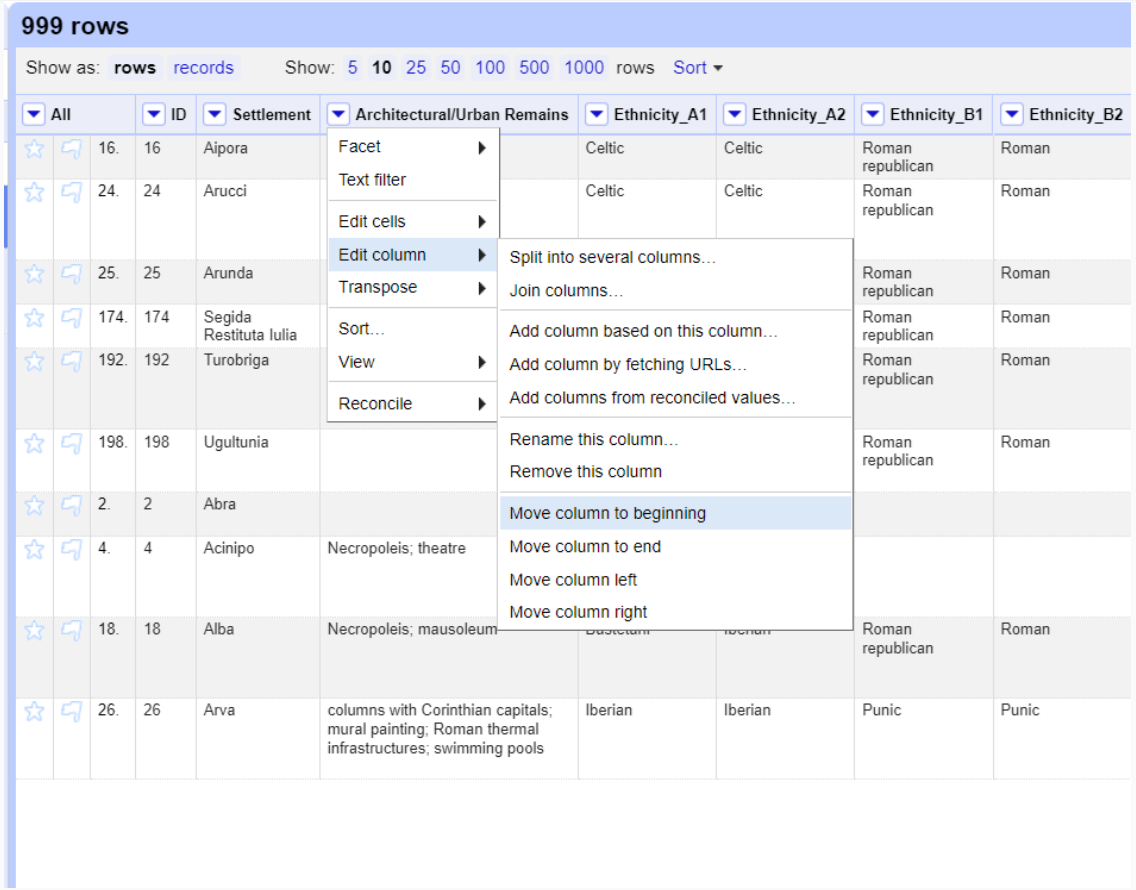
All		ID	Settlement	Architectural/Urban Remains	Ethnicity_A1	Ethnicity_A2	Ethnicity_B1	Ethnicity_B2
1.	1	Abdera		Phoenician	Phoenician		
2.	2	Abra		Turdetani?	Iberian		
3.	3	Acci	Dinastic Temple with corinthian chapitel; termaet aqueductt theatret slewers	Roman	Roman		
4.	4	Acinipo	Necropoleis; theatre	Oretani?	Iberian		
5.	5	Ad Aras					
6.	6	Ad Decumo 1					
7.	7	Ad Decumo 2					
8.	8	Ad Gemellas					
9.	9	Ad Herculem					
10.	10	Ad Lucos					

- Number of rows
- Display options
- Column headers and menus
- Cell contents

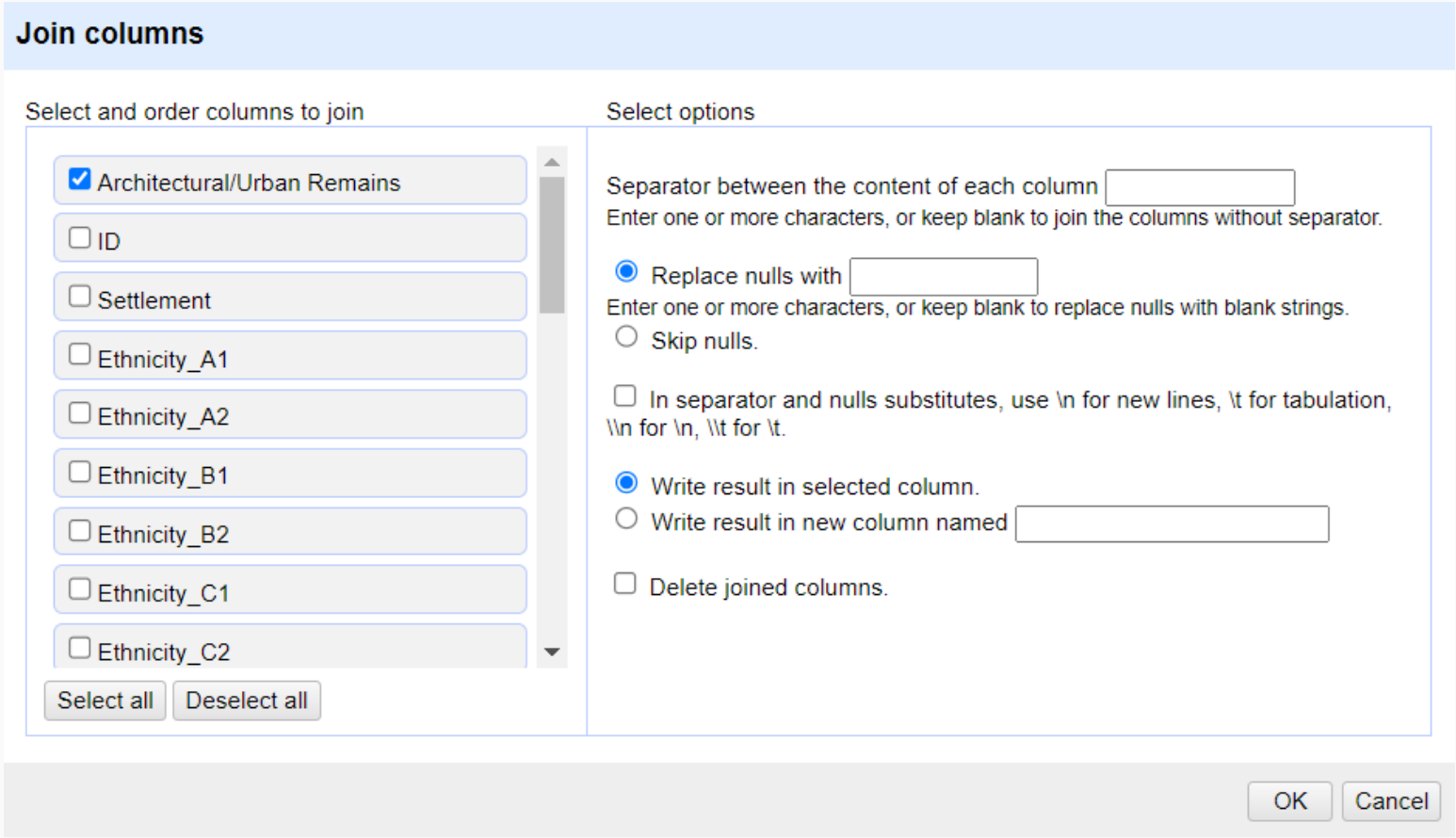
Manipulating columns



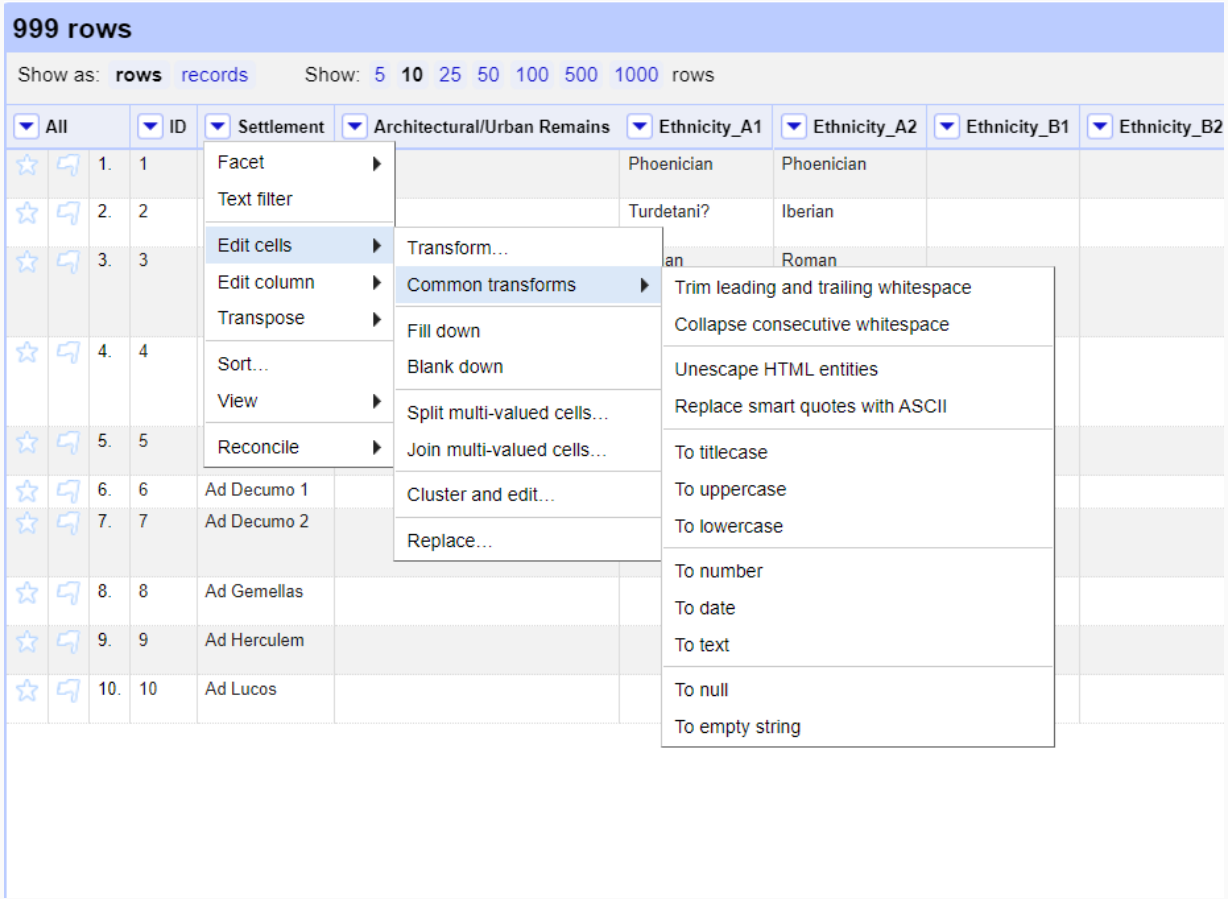
Move columns around



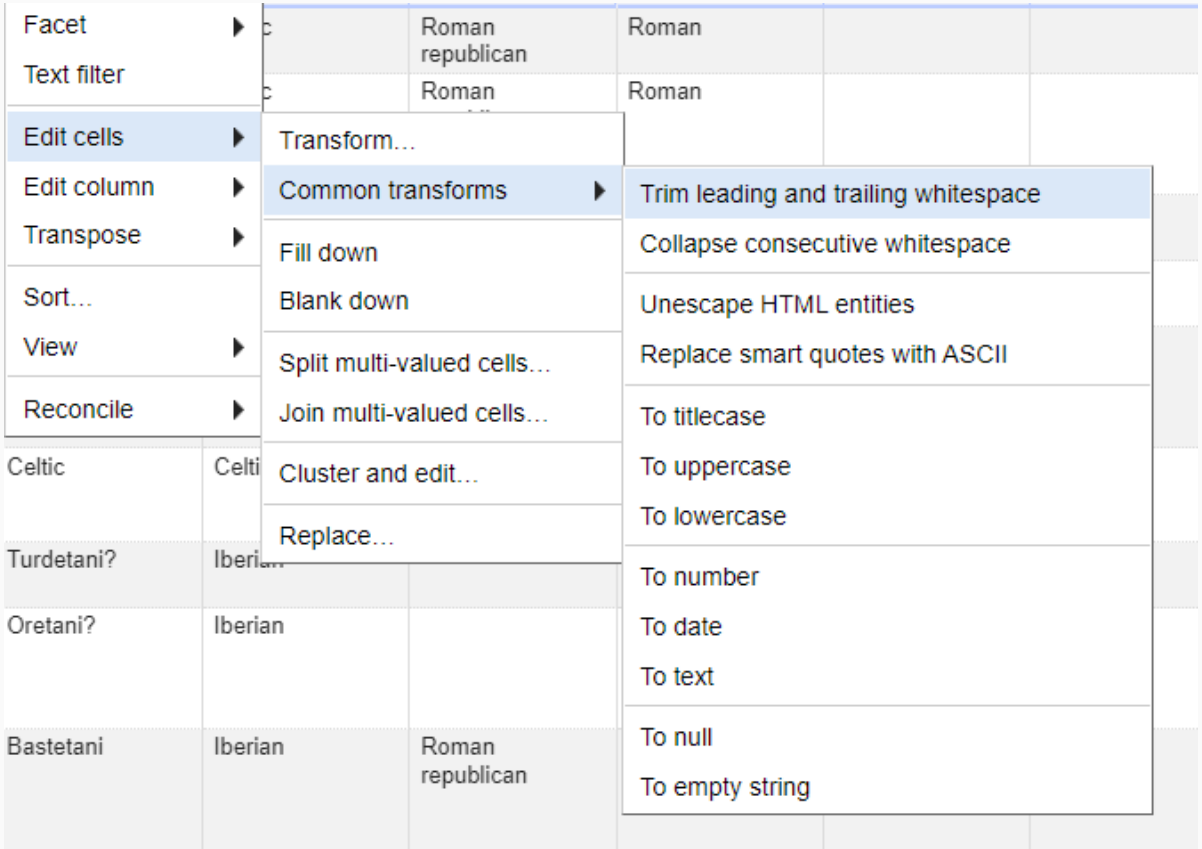
Merge several columns into one



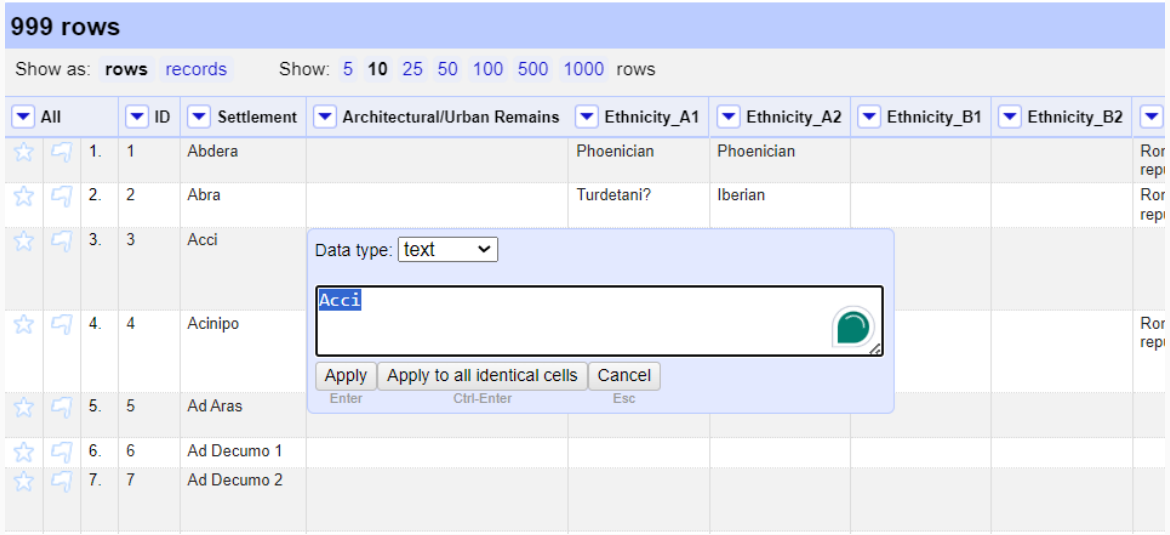
Manipulating cells



Manipulating cells



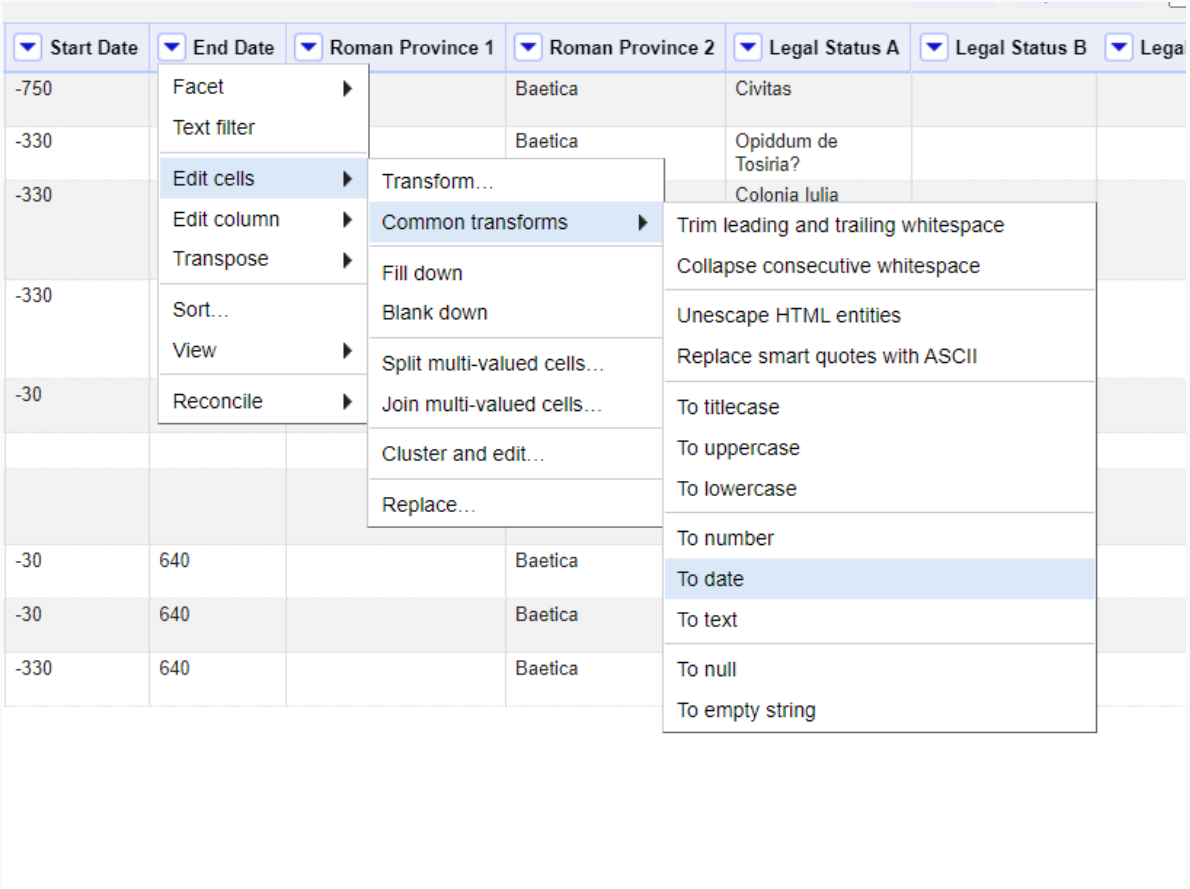
Manipulating cells - data types



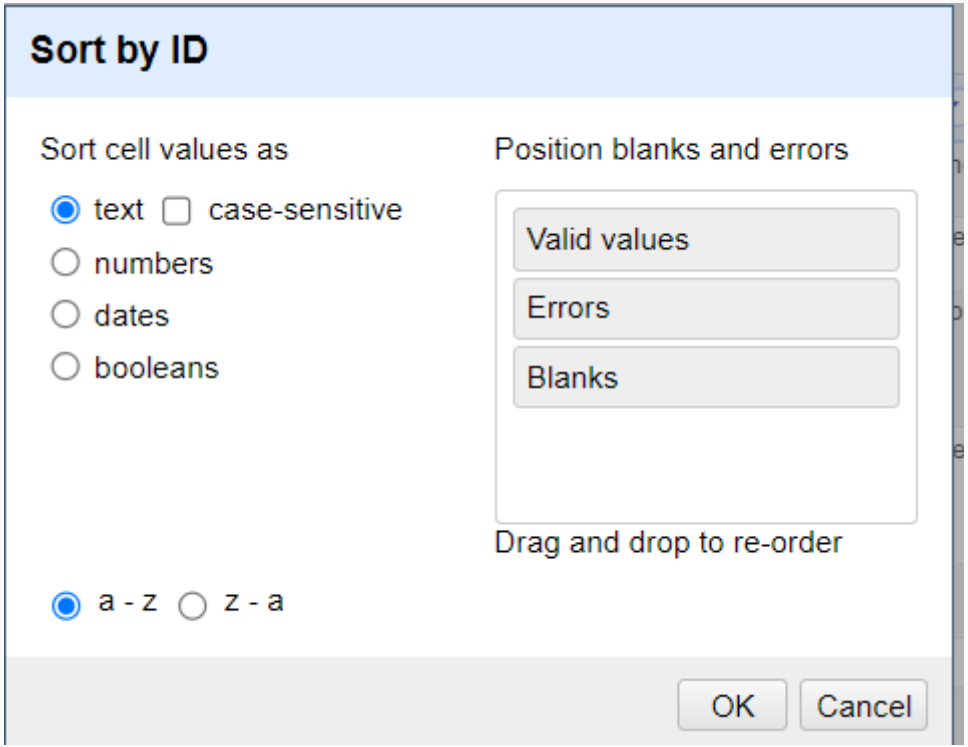
The data types supported are:

- string (one or more text characters)
- number (one or more characters of numbers only)
- boolean (values of “true” or “false”)
- date (ISO-8601-compliant extended format with time in UTC: YYYY-MM-DDTHH:MM:SSZ)

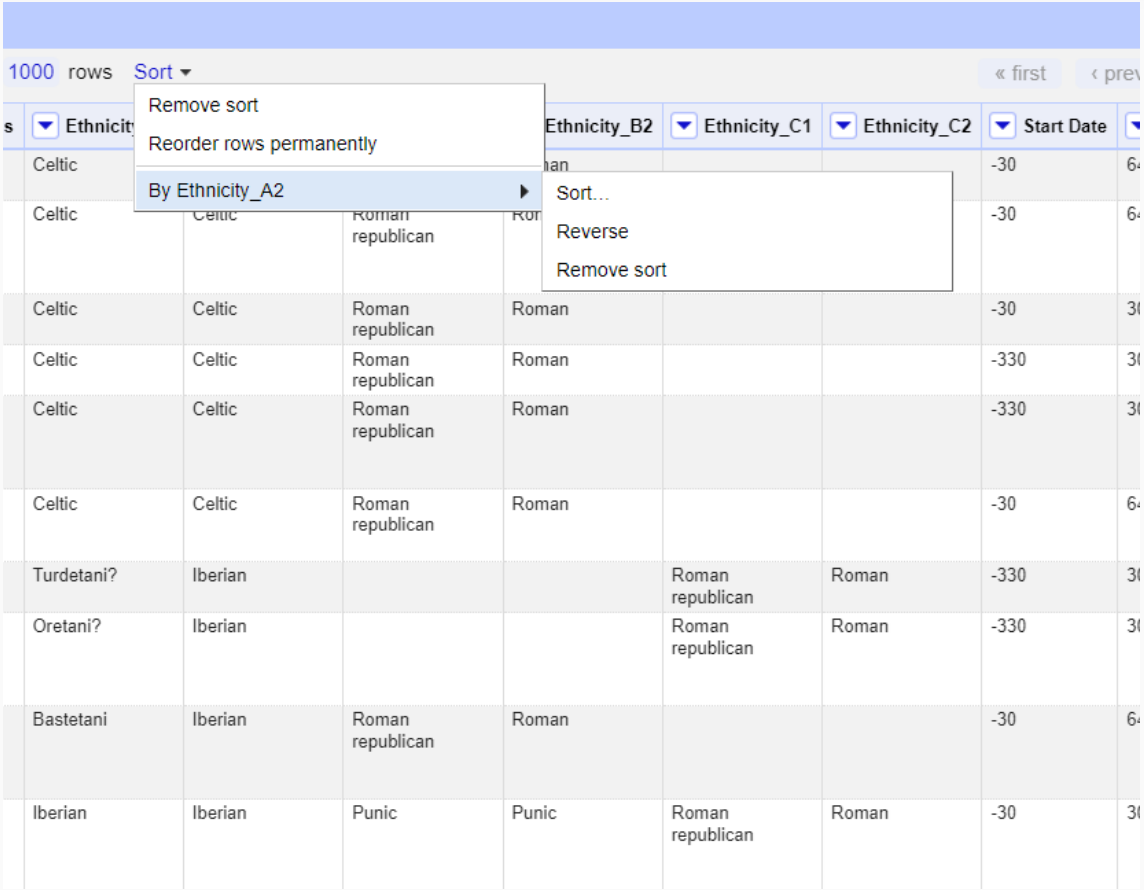
Dates


Start Date
End Date
Roman Province 1
Roman Province 2
Legal Status A
Legal Status B
Lega
-750
Baetica
Civitas
-330
Baetica
Opiddum de Tosiria?
-330
Colonia Iulia
-330
-30
-30
640
Baetica
-30
640
Baetica
-330
640
Baetica
Facet
Text filter
Edit cells
Edit column
Transpose
Sort...
View
Reconcile
Transform...
Common transforms
Fill down
Blank down
Split multi-valued cells...
Join multi-valued cells...
Cluster and edit...
Replace...
Trim leading and trailing whitespace
Collapse consecutive whitespace
Unescape HTML entities
Replace smart quotes with ASCII
To titlecase
To uppercase
To lowercase
To number
To date
To text
To null
To empty string

Sorting Data

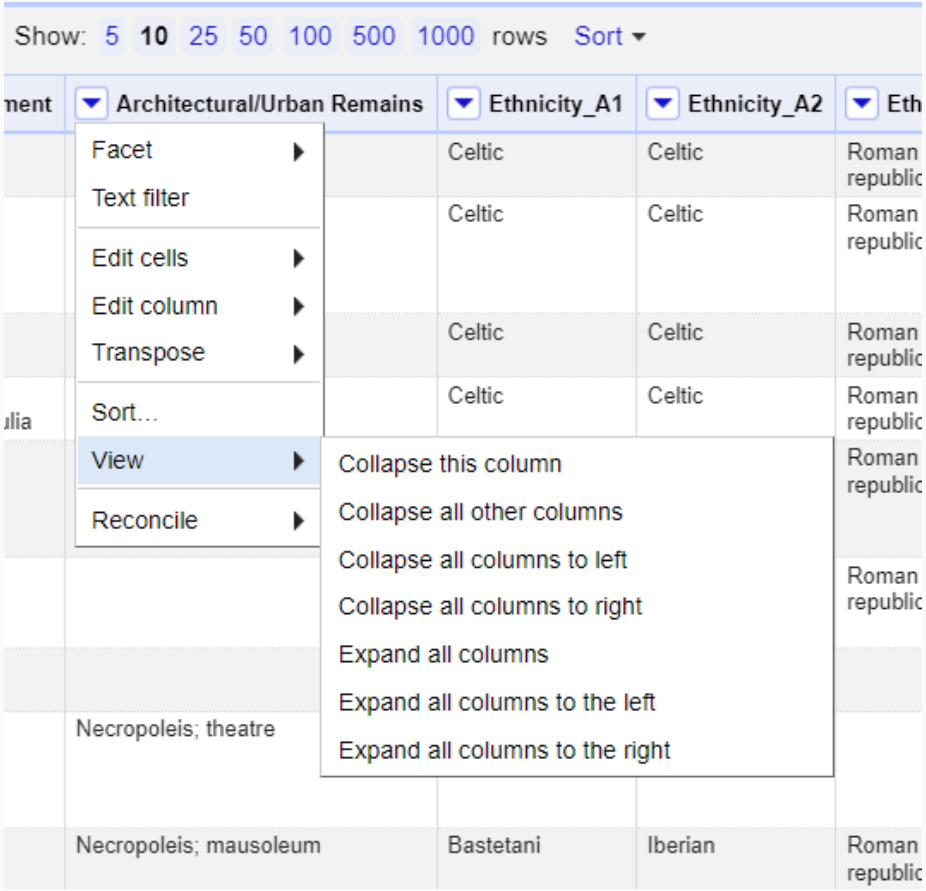


Sorting Data

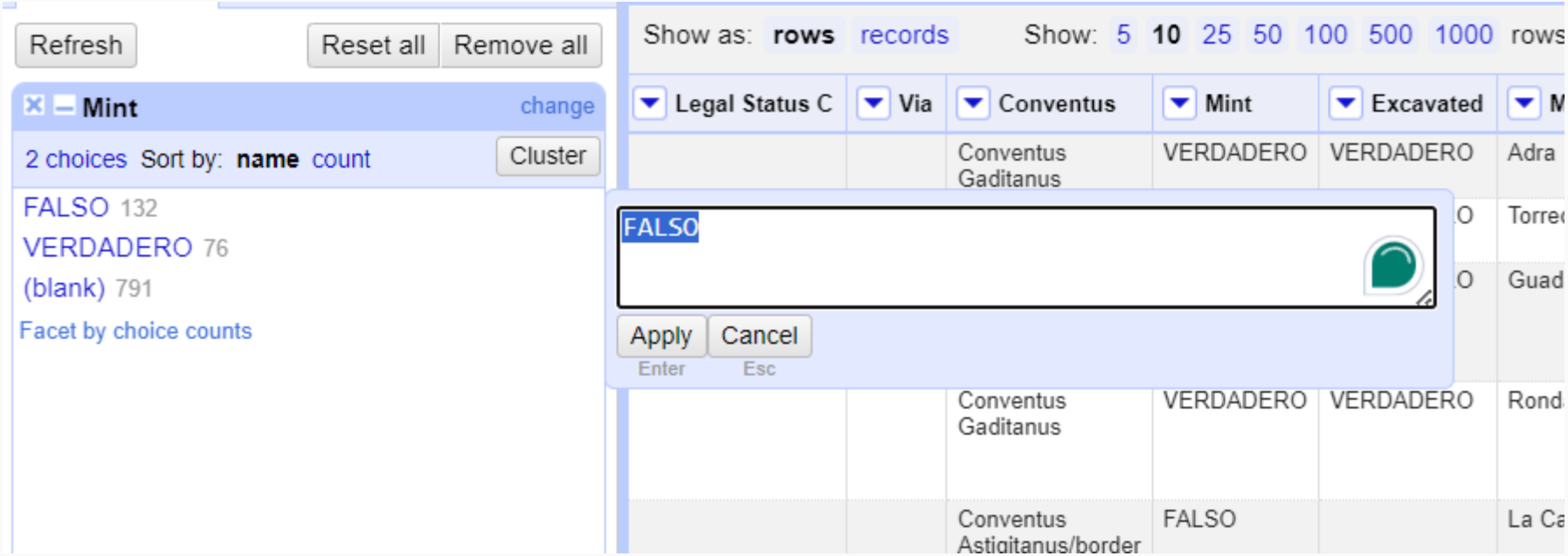


s	Ethnicity	Ethnicity_A2	Ethnicity_B1	Ethnicity_B2	Ethnicity_C1	Ethnicity_C2	Start Date	
	Celtic			an			-30	6
	Celtic	Celtic	Roman republican	Rom			-30	6
	Celtic	Celtic	Roman republican	Roman			-30	3
	Celtic	Celtic	Roman republican	Roman			-330	3
	Celtic	Celtic	Roman republican	Roman			-330	3
	Celtic	Celtic	Roman republican	Roman			-30	6
	Turdetani?	Iberian			Roman republican	Roman	-330	3
	Oretani?	Iberian			Roman republican	Roman	-330	3
	Bastetani	Iberian	Roman republican	Roman			-30	6
	Iberian	Iberian	Punic	Punic	Roman republican	Roman	-30	3

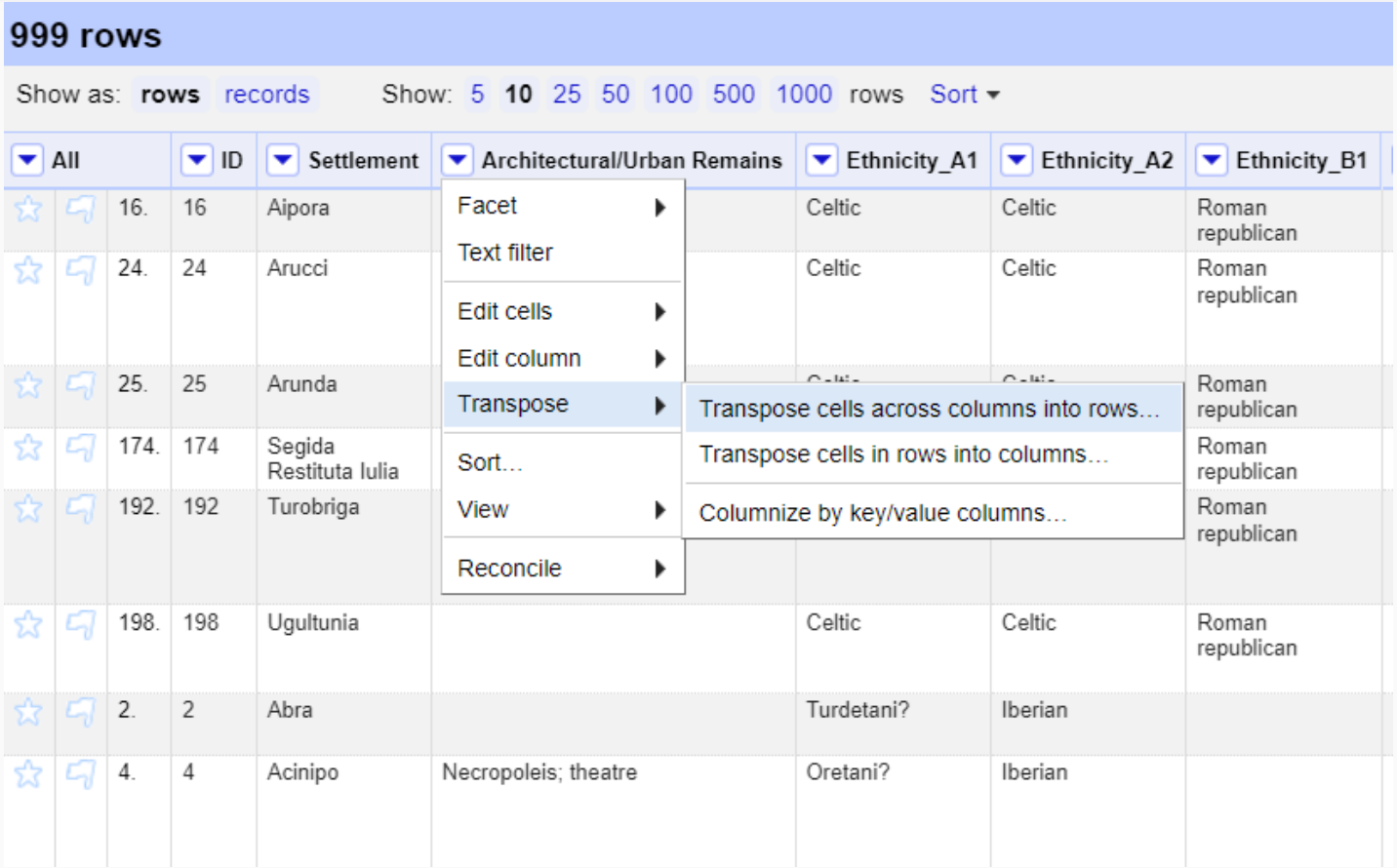
Control Data view



Faceting Data



Transposing cells across columns



Cluster and edit similar cells

Cluster and edit column "Ethnicity_A1"

Find groups of different cell values that might be other representations of the same thing. For example, "New York" and "new york" likely refer to the same concept and just differ by capitalization, and "Gödel" and "Godel" probably refer to the same person. Find out more...

Method Key collision Keying function Daitch-Mokotoff **2** clusters found

Cluster size	Row count	Values in cluster	Merge?	New cell value
2	26	• Turdetani (24 rows) • Turdetani? (2 rows)	☐	Turdetani
2	4	• Oretani (3 rows) • Oretani?	☐	Oretani

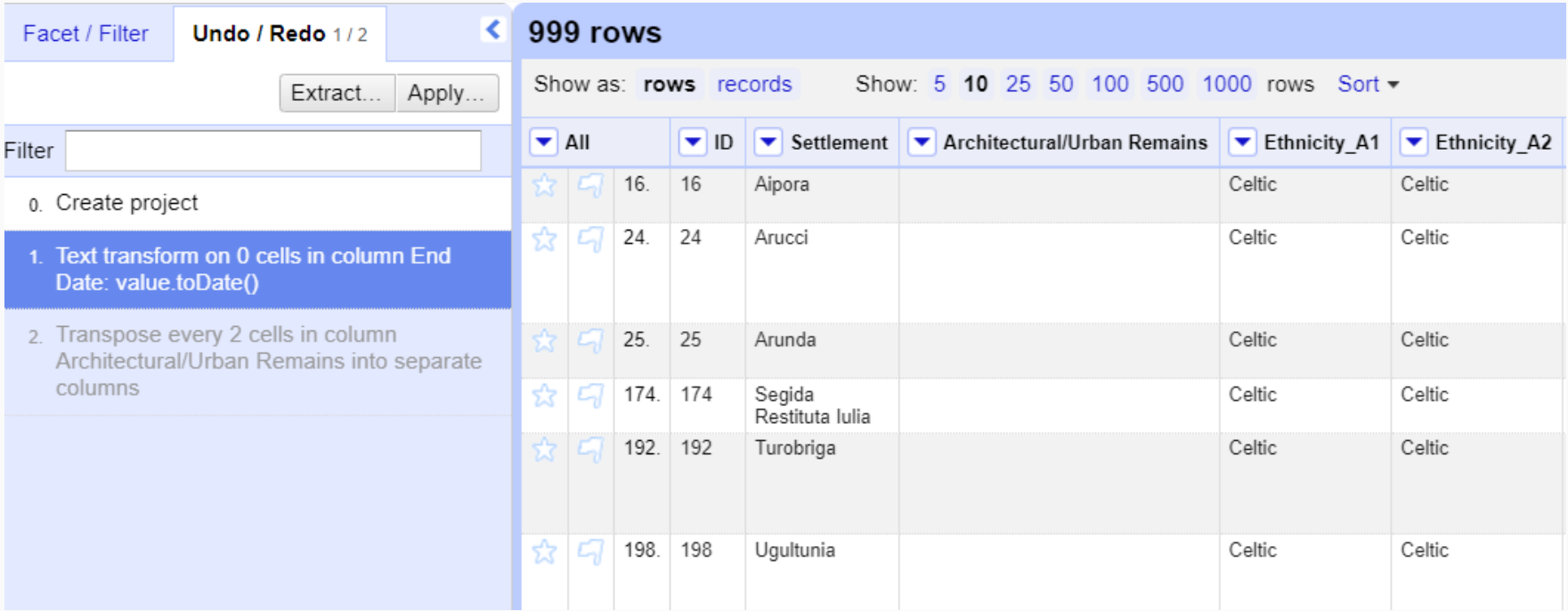
 # Rows in cluster

4 — 26

Average length of choices

7.5 — 9.5

Using the project History



Exercise

1. Change the order of rows and columns.
2. Edit cell contents within a particular column
3. Change cells type.
4. Transform rows into columns, and columns into rows
5. Add new columns based on existing data, with fetching new information.
6. Convert your rows of data into multi-record rows.