

EmDiglt 2.0 Data Schema


EmDiglt data has been organized into separate sheets, based upon EmDiglt 1.0 (accessible [here](#)).

The project uses the term “early modern” as the core corpus consists of itinerary books published between 1545-1750. An itinerary is defined as a book using the characteristic route table format to list places sequentially, as opposed to narrative prose found in other types of travel guides and accounts.


Caveats: The database was compiled through a combination of automated and hand-checked transcription and parsing. Individual references should be checked against the original sources for accuracy.

Overview Sheet

This is the primary sheet for the database. The **rows** represent **lines**, defined by single, coherent references, such as page numbers, headers, or observations on a given location. Lines as presented in the data are, in most cases, equivalent to the printed line of text found in the original source, however words or thoughts that originally spanned multiple lines have been combined. Lines retain the original sequence of reading order in which text appeared in the printed itinerary, however only text deemed itinerary tables has been converted into the data. The **columns** describe different **line characteristics**

The following features derive from  EmDigIt Itinerary Processing Notebook (IT).ipynb		
filename	A string filename that corresponds to the original XML file, which are themselves derived from the Transkribus alto files. Each unique filename corresponds to a different image file taken from the PDF of the printed itinerary.	1422689_0003_55312848.xml
itinerary_name	A string itinerary name corresponds to that assigned within EmDigIt 1.0 . This is generally the initials of an identified author or publisher followed by the four digit year of publication for the given edition.	OC1623
page_region_line	A string that provides a unique identifier for a line by combining the page, region, and line column values with letter abbreviations for ease of human interpretability. Note that pages, regions, and lines are assigned a sequence number by their appearance in the dataset and may not contain all pages, regions, and lines of the original printed text.	P1R8L3 (i.e. “the third line in the eighth region of the first page in the dataset”)
page	An integer representing the sequence of the page of a given line. Each electronic page corresponds to a single page of the original printed text.	1

region	An integer representing the sequence of the region of a given line on a given page. Regions are separated when either a page or region_type value has changed.	8
line	An integer representing the sequence of the given line within a given region, on a given page.	3
line_type	A string categorical variable describing the function of a given text line. Possible values include: page-number, chapter-heading, route-header, location, prose, and sum-distance. The line_type prose is used as a catch-all for text lines that do not fulfill the other given purposes and often represents author observations, such as on scenery, alternative routes, or mileage conversions.	page-number
route_description	The content value taken from the last occurring route-header. This represents the fact that lines representing locations, prose, and sum-distance are generally organized under a line with line_type value route-header. These are often distinguished visually in the original text through the use of font styles, capitalization, or other symbols.	Poste da Roma à Genova.
content	A string representing the text transcribed from an associated line of text in the original itinerary. Minimal tidying has been performed.	a San Nicolò p.1
cleaned	A string representing the lowercase location name found within the content value of a given line. Only provided for lines with line_type location.	san nicolò
The following columns only provide values for line_type= "location" unless otherwise noted.		

id	A number representing the unique identifier of a location from the Gazetteer sheet to which the value in column cleaned has been fuzzy matched and hand verified.	350
Geoname, geonameID, Location_Lat, Location_Lng, state, country_code	See Gazetteer sheet.	
distance	A number representing a distance if provided in the content value for a given line. Only provided for lines with line_type location, route-header, or sum-distance. Some interpretation of symbols has been performed where symbols such as an asterisk (*) or prose terms (such as “dieci” or “mezza”) were used to indicate distance values. If multiple distance values are provided, they are separated by the pipe symbol and correspond in order to the multiple units provided in the unit column.	1.5
unit	A string representing the unit modifying the distance value. Only provided for lines with line_type location, route-header, or sum-distance. Examples include: leagues, posts, miles, and hore. If multiple distance units are provided, they are separated by the pipe symbol and correspond in order to the multiple values provided in the distance column.	miles posts
The following entries derive from  EMDigIT Data Flagging Notebook.ipynb		
revised_distance	A number to one decimal point representing a calculated Euclidean distance in kilometers from prior location based upon the	13.8

	value of distance and the conversion dictionaries . Calculated based upon coordinates of current and prior line with line_type location.	
bearing	A number representing the last known bearing of direction of travel. Calculated based upon coordinates of current and prior line with line_type location.	-70.46956116
approx_coordinates	Coordinate value in format Latitude, Longitude. Intended to provide an approximation when no firm match has been made in Gazetteer based upon current itinerary. In this case, approx_coordinates will be based upon last known bearing and revised_distance. If Location_Lat and Location_Lng are available by matching to gazetteer, however, approx_coordinates will present the same values as those columns in coordinate format.	43.03453863769901, 12.80546179866986
Flag	Boolean value reflecting whether a given line is a lower confidence match. "TRUE" indicates that the given location reconciliation should be used only with caution and is in need of user review.	TRUE
alternative_location, alternative_match, alternative_geoname	These columns present alternative values for columns match_id, geoname, and geonameId when there is no known Location_Lat or Location_Lng but it is possible to calculate approx_coordinates. The value from approx_coordinates is compared to all values found in Gazetteer and the closest possible reconciliation is suggested.	FALSE
bounds_test	Flips "bounds_test" to True if the Latitude and Longitude for a given	TRUE

	location fall outside the minimum or maximum established for a given route_description. This method flags the most locations, resulting in the most false positives.	
dist_test	Flips "dist_test" to true where the Latitude and Longitude for a given location is more than 30km from those of location n-1.	FALSE
dist_test2	Flips "dist_test2" to true where the Latitude and Longitude for a given location is more than 30km from those the approximated coordinates. NB: TBD if 30km is the best cutoff.	TRUE
state_test	Flips "state_test" to true if the value for "state" for location n is NOT equal to "state" for location n-1 OR location n+1.	TRUE
state_test2	Makes a list of all states that occur within a given route_description and flips "state_Test" to true when location n does not match any of them	FALSE
The following entries derive from the EmDigIt Feature Tagging.ipynb		
content_no_location	Remaining string of content value once substring from column "cleaned" has been removed	"21 a , castello miglia 8"
features	Regular expression matching of content_no_location value to dictionary of types of physical sites	"Castle"
descriptors	Regular expression matching of content_no_location value to dictionary of common adjectives or adverbial phrases	"Good"
categories	Regular expression matching of content_no_location value to dictionary of dictionaries as well as	"Economic"

	selections of features and description	
--	---	--

GIS Sheet

The sheet GIS represents a modified form of Overview for ease of use with ArcGIS. Only lines with line_type location and known or approximated coordinate values are kept. Location_Lat and Location_Lng now include both approximated and known coordinates, as indicated by the new column approx_coords set to boolean TRUE (coordinates are approximated) or FALSE (coordinates are known).

Palladio Sheet

The sheet Palladio represents a modified form of Overview for ease of use with Stanford Palladio. Only lines with line_type location and known or approximated coordinate values are kept. Location_Lat and Location_Lng have been combined where matched into matched_coordinates, and all_coordinates includes both the matched and approximated value in a single column. The new column prior_coordinates has been added to allow for ease of point-to-point mapping.

Gazetteer Sheet

The **rows** represent **locations** as distinguished by unique references that are largely reconciled to coordinates. When coordinates are unknown but references are similar in name and/or occur in a similar context (such as a given route sequence, or state), they have been reconciled to a single row.

id	A number representing the unique identifier of a location	1772
Location_Name_Standardized	A string representing the most common name for a given location. Value taken from geoname where available (first letter capitalized), or most commonly occurring variant in Location_Name (all lowercase). May contain special characters.	A Coruña
geonameID	Corresponds to Geoname documentation .	3119841
Location_Lat	Corresponds to Geoname documentation .	43.37135
Location_Lng	Corresponds to Geoname documentation .	-8.396
state	Corresponds to Geoname documentation .	Galicia
country_code	Corresponds to Geoname documentation .	ES
Location_Name	String values representing lowercase name variants as found and reconciled from the cleaned column of the corresponding itinerary Overview sheet, as well as others from the project. Variants may be the result of original variation or common OCR errors. Individual values separated by the pipe symbol.	corvigna a coruña
The following columns may not be present, and are generated and used during the process of disambiguation and geo reconciliation		
Disambiguation_Need	Boolean value indicating whether	TRUE

ed	variants in Location_Name can be found in multiple rows of gazetteer.	
Last_Updated	String representing last update to values in given row in format MM.DD.YY	11.8.23
Note	String value indicating editorial annotation.	"No match found."
Test_Flag	An experimental test that indicates where in Overview a given location appears to fall outside of route boundaries, as established by the first and last known locations.	Failed boundary test OC1623: P111R6L6

Conversion dictionaries

These derive from [EMDigIT Data Flagging Notebook.ipynb](#), where the most recent versions can be found.

The following dictionary is used to calculate revised_distance in modern kilometers based upon country_code and the units leagues, posts or miles. Conversion ratios are based upon a rationalization of explicit guidelines for mileage conversion provided in the itineraries.

```
country_dist = {
    'AT': {'leagues': 4.5, 'posts': 14},
    'BE': {'leagues': 5.25},
    'CH': {'miles': 13, 'hore': 8.75, 'posts': 11},
    'CZ': {'miles': 8.75, 'posts': 22.5},
    'DE': {'miles': 8.75, 'posts': 15.5, 'leagues': 6.5},
    'DK': {'miles': 8.75},
    'ES': {'miles': 8.75, 'leagues': 7, 'posts': 11},
    'FR': {'miles': 5.25, 'posts': 10.5, 'leagues': 6},
    'HU': {'miles': 13, 'posts': 26.25},
    'IT': {'miles': 1.75, 'leagues': 5.75, 'posts': 13.75},
    'LU': {'miles': 8.75, 'posts': 15.5, 'leagues': 6.5},
    'LV': {'miles': 8.75},
    'NO': {'miles': 5.85},
    'PL': {'miles': 8.75},
    'SE': {'miles': 8.75},
    'UK': {'miles': 5.25},
}
```

The following dictionary is preferred instead when data is comprehensive enough to allow for reliance on state as opposed to country_code and only the unit posts. Values are based upon the median distances calculated between matched locations within given regions in OC1623. When data was too sparse to allow for this calculation, the overall median for the country_code has been used instead.

```
state_dist = {
    'Carinthia': {'posts': 15.8},
    'Lower Austria': {'posts': 15.8},
    'Salzburg': {'posts': 12.2},
    'Styria': {'posts': 19.4},
}
```

```
'Tyrol': {'posts': 13.7},
'Upper Austria': {'posts': 16.7},
'Vienna': {'posts': 15.7},
'Vorarlberg': {'posts': 14.8},
'Brussels Capital': {'posts': 17.3},
'Flanders': {'posts': 9.7},
'Wallonia': {'posts': 13.2},
'Aargau': {'posts': 28.1},
'Basel-City': {'posts': 9.9},
'Basel-Landschaft': {'posts': 11.9},
'Bern': {'posts': 10.7},
'Fribourg': {'posts': 23.4},
'Geneva': {'posts': 14.7},
'Grisons': {'posts': 8.5},
'Lucerne': {'posts': 8},
'Saint Gallen': {'posts': 15.1},
'Schaffhausen': {'posts': 32},
'Schwyz': {'posts': 39.8},
'Solothurn': {'posts': 27.5},
'Turgau': {'posts': 18.4},
'Ticino': {'posts': 12.6},
'Uri': {'posts': 7.6},
'Valais': {'posts': 9.8},
'Vaud': {'posts': 10.4},
'Zug': {'posts': 21},
'Zurich': {'posts': 20.3},
'Central Bohemia': {'posts': 16.3},
'Jihočeský kraj': {'posts': 17.4},
'Karlovarský kraj': {'posts': 16.5},
'Liberecký kraj': {'posts': 8.6},
'Moravskoslezský': {'posts': 22.1},
'Olomoucký': {'posts': 10.3},
'Plzeň Region': {'posts': 15.6},
'Prague': {'posts': 22.2},
'South Moravia': {'posts': 20.6},
'South Moravian': {'posts': 9.7},
'Ústecký kraj': {'posts': 12.6},
```

```
'Zlín': {'posts': 49.9},
'Baden-Wurttemberg': {'posts': 17.9},
'Bavaria': {'posts': 15.7},
'Hesse': {'posts': 13.5},
'North Rhine-Westphalia': {'posts': 14.6},
'Rheinland-Pfalz': {'posts': 15.8},
'Rheinland-Pflaz': {'posts': 8},
'Saxony': {'posts': 8.9},
'Andalusia': {'posts': 13.3},
'Aragon': {'posts': 14},
'Asturias': {'posts': 11.8},
'Basque Country': {'posts': 10.1},
'Cantabria': {'posts': 11.8},
'Castille and León': {'posts': 10.5},
'Castille-La Mancha': {'posts': 12.2},
'Catalonia': {'posts': 11.6},
'Extremadura': {'posts': 13.1},
'Galicia': {'posts': 25.1},
'La Rioja': {'posts': 11.4},
'Madrid': {'posts': 10.5},
'Murcia': {'posts': 13.8},
'Navarre': {'posts': 13.6},
'Valencia': {'posts': 12.6},
'Auvergne-Rhône-Alpes': {'posts': 11.1},
'Bourgogne-Franche-Comté': {'posts': 8.3},
'Centre-Val de Loire': {'posts': 8.7},
'Grand Est': {'posts': 11},
'Hauts-de-France': {'posts': 12.2},
'Île-de-France': {'posts': 10.2},
'Normandy': {'posts': 7.9},
'Nouvelle-Aquitaine': {'posts': 11.9},
'Occitanie': {'posts': 11.8},
'Piedmont': {'posts': 10.6},
'Provence-Alpes-Côte d\'Azur': {'posts': 10.2},
'England': {'posts': 4.7},
'Budapest': {'posts': 9.9},
'Győr-Moson-Sopron': {'posts': 10.8},
```

```
'Komárom-Esztergom': {'posts': 27.9},
'Abruzzo': {'posts': 9.7},
'Aosta Valley': {'posts': 8.8},
'Apulia': {'posts': 11.9},
'Basilicate': {'posts': 8.5},
'Calabria': {'posts': 15.6},
'Campania': {'posts': 13.4},
'Emilia-Romagna': {'posts': 13},
'Friuli Venezia Giulia': {'posts': 25.7},
'Lazio': {'posts': 13.8},
'Liguria': {'posts': 8.3},
'Lombardy': {'posts': 13.1},
'Piedmont': {'posts': 12},
'Sicily': {'posts': 25.7},
'The Marches': {'posts': 11.2},
'Trentino-Alto Adige': {'posts': 13.8},
'Tuscany': {'posts': 13.6},
'Umbria': {'posts': 12.5},
'Veneto': {'posts': 14.5},
'Clervaux': {'posts': 21.2},
'Limburg': {'posts': 13.6},
'Greater Poland': {'posts': 9.4},
'Kujawsko-Pomorskie': {'posts': 15.5},
'Lesser Poland': {'posts': 7.5},
'Lower Silesia': {'posts': 8.1},
'Lubusz': {'posts': 29.6},
'Opole Voivodeship': {'posts': 10.7},
'Pomerania': {'posts': 11.8},
'Silesia': {'posts': 9.7},
'West Pomerania': {'posts': 24.5},
'Évora': {'posts': 10.8},
'Lisbon': {'posts': 13.9},
'Portalegre': {'posts': 21},
'Setúbal': {'posts': 22.5},
'Brezovica': {'posts': 5.6},
'Celje': {'posts': 23.9},
'Ljubljana': {'posts': 9.6},
```

```
'Maribor': {'posts': 18},  
'Slovenska Bistrica': {'posts': 11.6},  
'Slovenska Konjice': {'posts': 17.3},  
'Vransko': {'posts': 37.9},  
'Vrhnika': {'posts': 28.9},  
'Jesenice': {'posts': 41.4}  
}
```