Updated March 18, 2021

Corpus Synodalium Project Narrative

Corpus Synodalium consists of four components: a working repertory of local ecclesiastical legislation; the corpus of transcribed texts; the digital atlas of dioceses and provinces; and the online interface through which these elements are brought together for exploration and analysis.

Repertory:

The origins of the project lie in a working repertory of local ecclesiastical legislation that the project leader developed in the course of his dissertation research. The first version of this was posted online in December 2012 on a (now defunct) scholar.harvard.edu website; it listed 715 diocesan, provincial, and legatine statutes issued between 1274-1400, along with information about their date, place, classification type, and whatever printed edition or manuscript source had been consulted.

Further research by the project leader as well as helpful feedback from colleagues around the world has since expanded the repertory considerably. The repertory now contains nearly 2300 entries, mostly diocesan, provincial, and legatine statutes (as before), but with scattered other genres appearing as well (e.g. episcopal mandates, abbreviations, synodal *ordines*). Each entry includes up to thirty fields of associated metadata, which indicate date, place, classification type, issuing authority, language, manuscript source(s), printed edition(s), URLs for digitized online sources, relevant bibliography, and whether a full-text transcription has been added to *CoSyn*. In most cases, the listing of sources, editions, and bibliography are far from comprehensive, indicating only those sources used to generate the repertory information and text transcription (if any).

Text Corpus:

The creation of the text corpus began in July 2016. The sources fell into three categories for the purposes of producing transcriptions. The first category included those texts that existed in machine-readable editions, generally those from the nineteenth and twentieth centuries. Assuming the chosen text did not already exist in digital/machine-readable form, it was first scanned to PDF. For short texts, this was done manually; where entire volumes needed to be scanned, the task was outsourced to a company that specialized in high-volume non-destructive scanning. The resulting PDF was then run through ABBYY FineReader 12 to generate OCR (Optical Character Recognition) transcriptions. Through repeated hand-cleaning of the early transcriptions, the project leader trained his local instance of ABBYY FineReader 12 to recognize medieval Latin word-forms and recurring technical vocabulary. Each resulting transcription was then reviewed and formatted by one of the undergraduate or graduate research assistants recruited to the project, all of whom had prior training in Latin and could thus catch most of the errors. The project leader then reviewed each transcription as a final check.

The second category included printed texts that available OCR technology could not easily handle, whether because of unusual characters, frequent abbreviations, or other complicating factors. Most of the major early modern conciliar collections fell into this category, as did incunables and other early printings. Where digitized images were not already available online, new digital scans were ordered from the appropriate repositories. For most of these texts (amounting to ca. 650,000 transcribed words thus far), the project relied on the services of Gavin Robinson, a freelance UK-based transcriber who specializes in early modern English and Latin texts, followed again by a review by the project leader.

The final category included all transcriptions that had to be made from unprinted (i.e. manuscript) sources. Where digitized images were not already available online, new digital scans (or microfilm printouts) were ordered from the appropriate repositories. Some of these were transcribed by the project leader, and several colleagues also volunteered to produce transcriptions or set these as exercises for their paleography students. The rest were produced by a small team of European graduate students and postdoctoral researchers who had the necessary paleographical skills along with appropriate knowledge of local sources and writing conventions. Once each transcription was finished, it was sent to one of the research assistants to check for typos and systematize the formatting, after which the project leader then did a further review. This final review usually involved spot checks against the original manuscript rather than a systematic review of the entire transcription.

The formatting guidelines are archived along with the text corpus. In general, all editorial material (footnotes, critical variants, etc.) in the originating source was omitted from the transcriptions in order to ensure compliance with copyright restrictions. No attempt was made to systematize orthography or capitalization in this initial phase of transcription. Article numbering was standardized (i.e. Arabic rather than Roman numerals), and any changes to the source text (e.g. the removal of square/angle brackets) were indicated in a corresponding note. No other markup was done to the texts during the transcription phase.

As of March 2021, the project included approximately fourteen hundred reviewed transcriptions, ninety percent of which date between 1200-1400. Of these, roughly one thousand were generated using OCR technology; three hundred were manually retyped from printed sources; and one hundred were transcribed from manuscripts. The transcriptions and metadata were subsequently archived in the Stanford Digital Repository (DOI given above), along with the most up-to-date version of the repertory.

Digital Atlas:

The creation of the digital atlas of dioceses and provinces began in July 2016. The first phase involved scouring the Harvard Map Collection and the holdings of the Harvard University Libraries for print maps of medieval ecclesiastical jurisdictions, as well as any existing digital maps. Several hundred sources were consulted in this process, from which approximately one hundred sources were selected to serve as the basis for the boundary information. These were scanned to produce high-resolution images.

The next task was to establish the list of dioceses and provinces that existed between 1200-1500, along with suffragan lists and relevant dates for the foundation, suppression, union, and translation of sees. This was generated by collating information from standard reference works (e.g. Eubel's *Hierarchia catholica*), online resources (e.g. http://www.katolsk.no/; http://www.catholic-hierarchy.org/), and specialist studies of particular dioceses and regions. After excluding those jurisdictions for which no mapping information could be identified (particularly those in the Latin East and eastern Europe), we ended up with a list of ca. 730 dioceses and ca. 90 provinces. Each diocese and province was assigned an individual ID number (D###; P###). All of this information was organized in a .csv spreadsheet.

Members of the map team georeferenced the scanned map images using the ArcGIS Desktop suite, then extracted the boundary information for each jurisdiction. The diocese layer was completed first, and the provincial boundary files were then created by merging together the suffragan dioceses of each metropolitan. To accommodate shifting diocesan and provincial boundaries during the period in question, it was necessary to generate up to four different boundary files for a given jurisdiction.

Eight individual maps ended up being created: two composite maps of dioceses and provinces, and then separate diocese/province maps for the thirteenth, fourteenth, and fifteenth centuries. The two composite maps displays all dioceses or provinces that existed at any point between 1200-1500; they are therefore fictive amalgams insofar as this 'maximal' configuration never existed at any point during this period. For the century-maps, if the boundaries of a particular jurisdiction changed over the course of the century, the maps shows the longest-lasting boundaries within that period.

The initial version was released on March 10, 2021, with the files being made available via both the Stanford Digital Repository (*Earthworks* and *Searchworks*) and Harvard's *Digital Atlas of Roman and Medieval Civilization* (DARMC).

Online Interface:

The development of the online interface began in February 2018. At the recommendation of Dr. Katie McDonough, who was then serving as Academic Technology Specialist in the History Department at Stanford University, it was decided to use *Philologic4* as the starting framework. *Philologic* is an open-source online text search, retrieval, and analysis tool that was developed out of the ARTFL Project, a long-running cooperative enterprise between the University of Chicago and the French CNRS that focused on the creation of digitized French-language texts. All subsequent development work was carried out by Thawsitt Naing, an undergraduate at Stanford University.

To make the text corpus compatible with *Philologic4*, Thawsitt first created a Python script that encodes each transcription according to the platform's standards. The main element of this is a customized TEI header that contains the associated metadata. The script also automatically encodes each transcription to identify the incipits and explicits of each text, the section numbering, and the editorial information found in each transcription's accompanying Notes file. A separate script was also created to generate a "normalized" version of the corpus that standardizes variant orthographies (e.g. ae→e; j→i; v→u); users could therefore choose to work with either the raw or normalized corpus.

The standard *Philologic4* platform was subsequently modified in several ways to accommodate the project's specific needs. A customized landing page was created to allow for quick searching of texts by origin place or issuing date. Search fields and facets were expanded to make full use of the extensive metadata. A new feature was added by which search results could be exported either as a list of unique record identifiers or as a list of short-form bibliographic references.

In addition, Thawsitt created a simple homepage for the project (later hosted at http://www.corpus-synodalium.com), which included links to project documentation and the working repertory.

In October 2018, Thawsitt produced the first prototype of a online mapping tool, by which search results generated via our local instance of *Philologic4* could be visualized spatially. Thawsitt wrote a script to integrate the digital atlas boundary files with the associated information from the mapping metadata spreadsheet. (We have since included the most important metadata within the embedded attribute tables of the boundary files, but the tool was designed to integrate this from the separate .csv instead.) Another script was written to export results from *Philologic4* to the map tool.

New features continued to be added to the mapping tool over the coming years, including an option to display results using different color schemes; the ability to toggle between diocese- and province-level results; an option to superimpose diocese-level results as proportional circles overtop the province-level

display; and a pop-up window to display the texts and metadata associated with the hits from any given jurisdiction. All of the associated code is documented on the project's Github website (https://github.com/corpus-synodalium/mapping).

Technical Overview

The working repertory was created using FileMaker Pro 13, and subsequently exported to a .csv file for archiving purposes.

Every transcription in the text corpus was produced using Notepad or Wordpad using UTF-8 encoding and is saved as a separate .txt file. (Early instances of transcriptions encoded in Windows-1252 were retroactively converted to UTF-8 for consistency). File names are given in the form RecordID_OriginPlace_SortDate (e.g. 1474_Cologne_1360). The associated Notes file for each transcription is similarly saved as a .txt file, with _Notes added to the end of the filename (e.g. 1474_Cologne_1360_Notes).

The diocese and province boundary files were created using the ArcGIS Desktop suite, and are archived as both zipped shapefiles (.shp) and GeoJSONS (.geojson). The associated metadata was produced using Microsoft Excel 2013 and is archived as a comma-delineated spreadsheet (.csv). The Map Bibliography was produced using Microsoft Word 2013 and is archived as a PDF (.pdf)

The main project database operates on a local instance of *Philologic4* (https://artfl-project.uchicago.edu/philologic4), with customized elements written in JavaScript and Python. The homepage, database, and mapping websites are hosted on Namecheap (https://www.namecheap.com/), with the domains registered with Google Domains (https://domains.google/).

The online mapping tool was built using React (https://reactjs.org/), a Javascript library for building user interfaces; and Leaflet (https://leafletjs.com/), an open-source JavaScript library for interactive maps. It was deployed using Netlify (https://www.netlify.com/) and hosted at https://cosyn.app/. The GeoJSON files are compressed and hosted on the Amazon CloudFront Content Delivery Network (https://aws.amazon.com/cloudfront/).

The project Bibliography, Progress Report, and user guides were produced using Microsoft Word 2013 and are archived as PDFs (.pdf). The spatial representation of transcription completion progress was produced using ArcGIS Online, and is saved as a JPEG image (.jpg).