Data Descriptor

The China Historical Christian Database: A Dataset Quantifying Christianity in China from 1550 to 1950

Alex Mayfield 1,2,*, Margaret Frei 1,3,*, Daryl Ireland 1,4,*,† and Eugenio Menegon 1,3,4,†

1 Center for Global Christianity and Mission, Boston University, Boston, MA 02215, USA; alex.mayfield@asbury.edu (A.M.); mrauch2@bu.edu (M.F.); emenegon@bu.edu (E.M.)
2 Department of Social Science and History, Asbury University, Wilmore, KY 40390, USA
3 Department of History, Boston University, Boston, MA 02215, USA
4 School of Theology, Boston University, Boston, MA 02215, USA
* Correspondence: dri@bu.edu
† These authors contributed equally to this work.

Abstract: The era of digitization is revolutionizing traditional humanities research, presenting both novel methodologies and challenges. This field harnesses quantitative techniques to yield groundbreaking insights, contingent upon comprehensive datasets on historical subjects. The China Historical Christian Database (CHCD) exemplifies this trend, furnishing researchers with a rich repository of historical, relational, and geographical data about Christianity in China from 1550 to 1950. The study of Christianity in China confronts formidable obstacles, including the mobility of historical agents, fluctuating relational networks, and linguistic disparities among scattered sources. The CHCD addresses these challenges by curating an open-access database built in Neo4j that records information about Christian institutions in China and those that worked inside of them. Drawing on historical sources, the CHCD contains temporal, relational, and geographic data. The database currently has over 40,000 nodes and 200,000 relationships, and continues to grow. Beyond its utility for religious studies, the CHCD encompasses broader interdisciplinary inquiries including social network analysis, geospatial visualization, and economic modeling. This article introduces the CHCD’s structure, and explains the data collection and curation process.

Dataset: https://github.com/chcdatabase/data

Dataset License: CC BY 4.0 DEED

Keywords: open data; relational database; history; China; networks; geography; Christianity

1. Background and Summary

The age of digitization has offered both new methods and new challenges to traditional humanities research by rapidly expanding access to sources and introducing new modes of analysis [1]. In response, a field of analysis known as digital or computational humanities has sprung up, often dated from Roberto Busa, SJ’s lemmatized concordance of the Summa Theologica completed in sections between 1949 and 1980 [2]. Quantitative approaches to historical study, like cliometrics [3] or social network analysis [4], have produced innovative scholarship. This scholarship relies on the creation, curation, and availability of data sets about historical subjects. In the field of Chinese history, there has been a series of such projects, including the China Historical Geographic System [5], the Contemporary Chinese Village Gazetteer Data project [6], and the China Biographical Database [7]. Following in these footsteps, the China Historical Christian Database (CHCD) makes a rich historical, relational, and geographical dataset publicly available for researchers.

Analyses of the dynamic nature of Christianity in China must overcome large challenges: historical agents were highly mobile, relational networks were constantly in flux,
and historical resources remain disparate and are written in many languages. These challenges have led most studies of Christianity in China to focus on localized areas, smaller groups, or specific time frames. Broader, more empirical approaches appear all but impossible because it continues to be difficult to triangulate historical sources or organize them for large-scale analysis. The CHCD is a new resource that addresses these linguistic, geographic, and technical problems facing the study of Christianity in China by quantifying and visualizing the place of Christianity in modern China from 1550 to 1950. It provides users the tools to discover where Christian churches, schools, hospitals, orphanages, publishing houses, and the like were located in China, and it documents who worked inside those buildings, both foreign and Chinese. Collectively, this information creates spatial maps and generates relational networks that reveal where, when, and how Western ideas, technologies, and practices entered China. Simultaneously, it uncovers how and through whom Chinese ideas, technologies, and practices were conveyed to the West.

The CHCD integrates spatial, temporal, and relational data to provide a complex picture of Christianity in China. However, this dataset is useful well beyond the analysis of “religious” networks [8]. As Nicolas Standaert has observed, the heading of “Christianity in China” extends beyond the modern category of “religion”, encompassing intricate nexuses of cultural interactions, scientific exchanges, diplomatic relations, and ritual life [9]. Therefore, the CHCD is a valuable resource for anyone interested in Sinology, religious studies, or intercultural studies, as well as social network analysis, geospatial research, economic development, and more.

The CHCD project has three major goals: the creation of a geographic and relational database, the creation of a user-friendly online platform, and fostering partnerships between Chinese and Western universities. In this article, we will focus on the first goal as we introduce the data we have collected and made publicly available.

2. Data Description

The CHCD is a graph database recording geographic and relational connections of the people and institutions involved with Christianity in China. It utilizes the open-source and industry-leading graph database platform Neo4j. The database covers the time period between the arrival of Francis Xavier, the first Jesuit missionary to reach China, in 1552 CE and the establishment of the People’s Republic of China in 1949 CE. The data are strategically bilingual, including names and alternate names for entities in Romanized scripts, traditional Chinese characters and modern pinyin whenever possible. The schema for the database can be found below in Figure 1.

![Figure 1. Database schema for the CHCD.](image-url)
2.1. Node and Relationship Types in the CHCD

The CHCD contains six primary node types: Person, Corporate Entity, Institution, Publication, General Area, and Event. People nodes represent human beings. Institution nodes represent organizations that have a direct geographic footprint, like churches, hospitals, and schools. Corporate Entity nodes represent organizations that do not have a direct geographic footprint, for example an international religious order like the Society of Jesus (the Jesuits). Event nodes represent important events that have specific geographic locations, like a Christian conference or a court case. Publication nodes represent any sort of textual document, including newspapers, journals, magazines and books. General Area nodes represent a geographic region. Since people and publications cannot have a direct relationship with a geographic node, these general areas allow people and publications to be associated with a location even when we do not have an institution or an event to which they would otherwise be connected.

In addition to the six primary node types described above, the CHCD has six types of geographic nodes: village, township, county, prefecture, province, and country. These nodes describe a location in increasingly specific terms, from country most broadly to village most specifically. Only geography nodes contain geographic coordinates, and only Institution, Event, and General Area nodes can link to a geography node. This decision is driven by the desire to avoid redundancy and minimize the workload associated with updating historical markers. By controlling the connectivity between nodes, the CHCD ensures that updates and modifications can be made efficiently without the need for extensive rework. This approach contrasts with a direct person-to-geography model, which would require more relationships and make updates more cumbersome.

The CHCD has eight types of relationships: part of, present at, related to, involved with, located in, linked to, and inside of. These relationships are time-bound with start and end dates wherever the historical sources provide this information. Relationship types are generally descriptive, but are mostly distinct for the purpose of more efficient querying. For example, the related to relationship type simply connects two people nodes; it does not imply a genetic or familial relation.

2.2. Geographic System

Given the complexity of China’s historical geography, the CHCD has opted to use contemporary (as of 2009) political maps as its primary reference; these include administrative divisions of the People’s Republic of China and the Republic of China (Taiwan). Rather than attempting to map historical locations based on their original administrative units or place names, which could result in inconsistencies and require extensive time, the CHCD opts for a simplified approach: centroid points for each administrative unit in the modern map are used as reference points in the database. For more specific locations such as village names or street addresses, researchers manually locate them within the modern administrative geography. While this method risks imposing modern geographical boundaries onto historical contexts, it offers the advantage of allowing researchers to analyze spatial relationships across different historical periods using a consistent framework. The original place name is retained within the data as an attribute of the relationship between a primary node and a geographic node where appropriate.

2.3. Data Set Statistics

Tables 1 and 2 below provide basic descriptive statistics on the size of the dataset.
Table 1. Summary of node types in the dataset.

<table>
<thead>
<tr>
<th>Node Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Geographic Nodes</td>
<td>3366</td>
</tr>
<tr>
<td>All Primary Nodes</td>
<td>43,360</td>
</tr>
<tr>
<td>People</td>
<td>34,768</td>
</tr>
<tr>
<td>Institution</td>
<td>6600</td>
</tr>
<tr>
<td>Corporate Entity</td>
<td>1165</td>
</tr>
<tr>
<td>Publication</td>
<td>414</td>
</tr>
<tr>
<td>General Area</td>
<td>258</td>
</tr>
<tr>
<td>Event</td>
<td>155</td>
</tr>
</tbody>
</table>

Table 2. Summary of relationship types in the dataset.

<table>
<thead>
<tr>
<th>Relationship Type</th>
<th>Source</th>
<th>Target</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Relationships</td>
<td>Any node</td>
<td>Any node</td>
<td>216,158</td>
</tr>
<tr>
<td>Part_of</td>
<td>Corporate Entity, Event, Institution, People</td>
<td>Corporate Entity</td>
<td>85,701</td>
</tr>
<tr>
<td>Present_at</td>
<td>People</td>
<td>Event, General Area, Institution</td>
<td>98,439</td>
</tr>
<tr>
<td>Located_in</td>
<td>Event, General Area, Institution</td>
<td>Geography</td>
<td>19,701</td>
</tr>
<tr>
<td>Related_to</td>
<td>People</td>
<td>People</td>
<td>8315</td>
</tr>
<tr>
<td>Inside_of</td>
<td>Geography</td>
<td>Geography</td>
<td>3363</td>
</tr>
<tr>
<td>Involved_with</td>
<td>Corporate Entity, Event, General Area, Institution, People, Publication</td>
<td>Publication, General Area</td>
<td>629</td>
</tr>
<tr>
<td>Linked_to</td>
<td>Event, Institution</td>
<td>Event, Institution</td>
<td>10</td>
</tr>
</tbody>
</table>

3. Materials and Methods

3.1. Material Collection

Data were collected from 205 primary (i.e., from the historical period) and secondary (i.e., after the historical) sources. These materials were collected from a variety of archives and, when possible, scanned into a digital format for easier processing. Of these sources, a minority provided the bulk of the data points, notably The Directory of Protestant Missions in China (1874–1950), The Educational Directory for China (1895–1920), Joseph Dehergne’s Répertoire des Jésuites de Chine de 1552 à 1800 (1973), and Сергей Головин, Российская Духовная Миссия в Китае: Исторический очерк (2013). Source materials were in a variety of languages, including English, French, Russian, Latin, and Traditional Chinese. A complete list of sources is linked in the Supplementary Materials.

3.2. Data Collection

Early efforts at Optical Character Recognition and automated data collection proved untenable due to the multilingual and multifarious nature of the documentary collection. As such, it was decided that manual collection was the only feasible solution. Thus, after the creation of a base data model, the principal investigators assembled an international team of scholars, archivists, and students to process the documentary collection. This collection took place over a four-year period from 2020 to 2024, and the team consisted of over 60 people, each contributing a varied number of hours based upon availability and expertise. The process of collection itself was rudimentary. The project team pursued what it termed a
“corpus model”, in which it worked programmatically through related sources (a “corpus”) before moving onto a different group of materials. This “corpus model” ensured that the data collected were, in general, of similar quality and kind. An example of a corpus would be the biographies of Maryknoll fathers stored online through the Maryknoll Mission Archives [10]. Rather than research each father individually, data were taken directly from one website that provided similar kinds of information on each missioner. Customized spreadsheets were developed for each data corpus, and team members were trained to identify and record descriptive and relational data in a manner that was consistent with our data model. Regular oversight by senior team members and project managers ensured a relatively high quality of data throughout the data collection process.

3.3. Data Normalization and Cleaning

1. Identification and Consolidation of Geographic Place Names and Coordinates. Historical sources related to China used multiple Romanization systems and often referred to small or obsolete administrative divisions. This meant it was often difficult to assign a geographic location to data objects. For difficult-to-find locations, project team members manually researched historical place names and assigned geographic coordinates and unique identifiers where possible. Where it was not possible to verify exact locations, the next highest verifiable administrative division was assigned. For example, if a school was known to be in a given prefecture but its location at the county level could not be identified, the school’s location was tied to the prefecture level. Historical place names were retained within the data model.

2. Merging and Division of Data Objects. Historical sources can refer to individuals and institutions in a multitude of ways, making the immediate identification of data objects in various sources challenging. Due to editorial choices, grammatical errors, and non-standard spelling practices, the same individual might appear as H. Noble, Hector N. Noble, H. N. Noble, or H. Nable. Such cases were identified using Winpure Clean and Match data cleaning software. The consolidation and division of objects was then decided based upon related data points. For example, one could reasonably assume that a missionary recorded as H. Noble in 1855 was not the same H. Noble reported in 1945. After such decisions were made for data objects, unique identifiers were assigned.

Where possible, cleaned and normalized data were used to speed up data collection. For example, data on historical place names, coordinates, and uniquely identified objects were made available to team members so that the process of data collection became more standardized over time.

3.4. Data Noise

This process resulted in a dataset that is true to its source materials. Some noise was unavoidable, especially in relation to Person objects. This is due to the fact that historical sources were sometimes vague, and individuals may have been wrongly duplicated or merged inadvertently. As such, the data are most useful in the aggregate.

4. Usage Notes

The CHCD data are fully accessible and publicly available on Github. The CSV files are in the UTF-8 encoding format with “@” as the separator. Please note that data collection for the CHCD continues, so new data will be added in the future. As updates are made, new CSVs are uploaded to the CHCD Github account. In order to access the API for the Neo4j application directly, please reach out to the CHCD team directly in order to create an authorized account.

5. Conclusions

Spanning 1550 to 1950, the CHCD covers a pivotal period for Chinese and global history. By tracing entities across time and space, it provides unique and reliable historical data
for largescale, quantitatively informed research in sinology, religious studies, intercultural studies, economic modeling, East Asian history, and more. Likewise, its size, complexity, and historical context provides a wealth of opportunities for social network analysts and geospatial researchers. Compiled from a variety of sources in multiple languages from different repositories around the world, the CHCD provides one of the most comprehensive datasets on historical Christianity to date. Such data have the capacity to help refine and redefine our understanding of China’s past, as well as the importance of religion in global East–West relations.

**Supplementary Materials:** The following supporting information can be downloaded at: For extensive documentation of all data attributes, please refer to our online documentation guide found here: https://chcdatabase.github.io/data-documentation/ (accessed on 22 May 2024). For a longer description of the project, please refer to our white paper found on the CHCD website here: https://chcdatabase.com/. A complete list of data sources can be found here: https://chcdatabase.github.io/data-documentation/docs/6_sources/.

**Author Contributions:** A.M.: conceptualization, software, data curation, and writing (original draft). M.F.: data curation, project administration, software, visualization, and writing (original draft). D.I.: supervision, funding acquisition, and writing (original draft). E.M.: resources and writing (review and editing). All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Endowment for the Humanities, grant HAA-280992-21, the Hariri Institute’s Software and Application Innovation Lab (SAIL) at Boston University, the Boston University Center for Innovation in Social Science, the Institute on Culture, Religion and World Affairs (CURA), and by individual donors. No grants were used to cover publication costs.

**Data Availability Statement:** The original data presented are openly available in the China Historical Christian Database repository on Github at https://github.com/chcdatabase/data (accessed on 22 May 2024).

**Acknowledgments:** This work was accomplished thanks to materials and labor provided by the Boston University Undergraduate Research Opportunity Program (UROP), the Carleton College Archives, the Centro Científico e Cultural de Macau, the Institute for Advanced Jesuit Studies at Boston College, the Institute of Qing History at Renmin University, Monumenta Serica, L’Orientale Università degli studi di Napoli, the Passionist Historical Archives, the Ricci Institute for Chinese–Western Cultural History at Boston College, St. Olaf College, the School of Oriental and African Studies (SOAS) at the University of London, the Taipei Ricci Institute, Università degli studi di Perugia, Whitworth University, and the Yale Divinity School. A list of project team members can be found on our website: https://chcdatabase.com/project-team/ (accessed on 22 May 2024).

**Conflicts of Interest:** The authors declare no conflicts of interest.

**Abbreviations**

The following abbreviations are used in this manuscript:

- CHCD  China Historical Christian Database

**References**


**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.