International Cartographic Association
Association Cartographique Internationale

Mining multiple sources of historical data: The example of a standardized dataset of medieval monasteries and convents in France

Adam Mertel [a,b] *, David Zbíral [b]

[a] *Masaryk University, Faculty of Science, Department of Geography, mertel.adam@mail.muni.cz*

[b] *Masaryk University, Faculty of Arts, Centre for the Digital Research of Religion, david.zbiral@mail.muni.cz*

* Corresponding author

Abstract: In this paper, we present a dataset of medieval monasteries and convents on the territory of today's France and discuss the workflow of its integration. Spatial historical data are usually dispersed and stored in various forms – encyclopedias and catalogues, websites, online databases, and printed maps. In order to cope with this heterogeneity and proceed to computational analysis, we have devised a method that includes the creation of a data model, data mining from sources, data transformation, geocoding, editing, and conflicts solving.

The resulting dataset is probably the most comprehensive collection of records on medieval monasteries within the borders of today's France. It can be used for understanding the spatial patterns of medieval Christian monasticism and the implantation of the official Church infrastructure, as well as the relation between this official infrastructure and phenomena covered in other datasets. We open this dataset, as well as scripts for mining, to the public (https://github.com/adammertel/dissinet.monasteries) and provide a map tool to visualize, filter, and download the records (http://hde.geogr.muni.cz/monasteries).

Keywords: Christian monasteries and convents, religious orders, spatiotemporal dataset, data mining, data integration

1. Introduction

Christian monasteries and convents were among the most important institutions of medieval Europe. Religious houses of different orders served as spiritual, economic, administrative, technological, and educational centres and played a crucial role in the formation of landscapes by draining wetlands, promoting the migration of people, and organizing space (Bond, 1989, 2001, 2003; Currie, 1989; Ferenczi, 2018; Milecka, 2012). Many traces of their presence and economic activity are still clearly visible in European landscapes (Farnedi & Togni, 2014).

For centuries rather than decades, Christian monasticism has been a prominent subject of historiography as well as encyclopedic efforts. Various encyclopedias, catalogues, datasets, and maps are available in printed (Ardura, 1993; Becking, 2005; Cocheril, 1976; Emery, 1962; Gerhards, 1998; Hadcock, Great Britain, & Ordnance Survey, 1950; Institut géographique national, 1995; Jedin, Latourette, & Martin, 1970; Moorman, 1983; Pelliccia & Rocca, 1974; Poras & Cottineau, 1935) or digital form (e.g., Wikipedia; the Digital Atlas of Roman and Medieval Civilizations). These sources present a comprehensive picture of Christian monasticism in medieval Europe. However, the available data suffer from significant limitations. For obvious reasons, data from printed encyclopedias cannot be directly used for spatial analysis and visualisation. Online resources, for their part, often come from particular, sometimes short-lived research projects, the geocoding of mostly older printed resources, and larger community projects involving many editors and formatting habits. It is relatively easy to localize religious houses but it is more difficult to quickly access data on them in a convenient form (Vauchez & Caby, 2003: 52-53).

The integration of dispersed and variously structured digital resources is a precondition for meaningful analytical use. With this goal, we developed a workflow

Proceedings of the International Cartographic Association, 2, 2019.
29th International Cartographic Conference (ICC 2019), 15–20 July 2019, Tokyo, Japan. This contribution underwent single-blind peer review based on submitted abstracts. https://doi.org/10.5194/ica-proc-2-85-2019 | © Authors 2019. CC BY 4.0 License.

consisting of mining, integrating, supplementing, and representing what is probably the most comprehensive spatiotemporal dataset of medieval religious houses within the boundaries of today's France, whether dissolved or still in existence. In this article, we present the workflow, draw more general implications of the method, and discuss various choices we made during the process. The first of these choices was to limit the area of interest so that it remained manageable. In this respect, we decided on the region covering today's Metropolitan France (excluding Corsica). The immediate motivation for this choice was that in a current project ("Dissident Networks Project / DISSINET", https://dissinet.cz), we plan to use a part of this dataset to inquire into the relation between the presence of official Church infrastructure and the places of religious dissent. On a more general level, our choice was driven by the fact that the regions which form today's France played a crucial role in the history of Christian monasticism; it is here, for instance, that some major orders and branches such as Cluniacs, Cistercians, Carthusians, and Dominicans were born and from where they spread. Nevertheless, the geographic scope of this project can easily be extended towards further regions and periods thanks to the scalability of our implementation of data integration.

The resulting dataset in the given context not only enriches our knowledge of medieval Christian monasticism, but can also serve to test particular hypotheses concerning the role of monasteries and convents in medieval economies and in medieval society, their spatial patterns, and their relation to demography and physical geography.

2. Background

The existence of a valid, comprehensive, and internally coherent dataset is a precondition for most research in the fields of computational history and the geohumanities. However, such datasets are still relatively rare. Even when data are available, they are often stored in a multitude of different sources, from which they need to be acquired (Devogele, Parent, & Spaccapietra, 1998), and in a variety of different forms (e.g., printed book, CSV table, SQL databases...), each of which requires a different approach with respect to mining and data integration.

Data integration has been a focus of research in information science for a long time, and is now gaining more and more attention also in the disciplines included under the umbrella term 'the digital humanities' (Oldman, de Doerr, de Jong, Norton, & Wikman, 2014). This is the case, for example, in archaeology, in which there is a need to integrate data from multiple sources to provide a broader view of the phenomena under scrutiny and to refine statistical models (e.g., Faniel, Kansa, Whitcher Kansa, Barrera-Gomez, & Yakel, 2013; Kintigh, 2006).

The process of data integration is divided into several steps (Gagnon, 2007). It begins with data collection, when unchanged data are stored, and data fusion, which determines consistent representation and resolves conflicts. The following operations are abstraction, supplementation, and aggregation. Naumann, Bilke, Bleiholder, & Weis (2006) recognize three steps – schema matching, duplicate detection, and data fusion, where each step resolves inconsistencies at a different level (schematic heterogeneity, duplicates, and data conflicts).

According to the character of the source, acquired data are often heterogeneous. Gagnon (2007) recognizes three kinds of heterogeneity: syntactic (differences between data models), structural (differences in structures), and semantic (differences in content). The vital part of the process of data integration is the method of resolving data conflicts that occur when the same entity is identified in more sources, these giving different values for the same variable. These conflicts can be resolved by ignoring the conflict and keeping both records, by avoiding the conflict (e.g., preferring one source over another), or by resolving the conflict through the application of an individual strategy (e.g., each attribute is filled from the source where the particular value is the most complete) (Bleiholder & Naumann, 2006).

The situation is even more specific in the integration of spatial sources, where we store spatial information. Coordinates provided in different data sources tend to be in different formats, mostly based on the medium of the source. Printed maps need to be scanned and georeferenced to retrieve spatial information; digital sources may need coordinate transformation and the generalization of cartographic content. Also, geocoding is essential in cases where the source gives spatial information in textual form (e.g., place name, or region name) but does not mention the coordinates directly. In cases when multiple sources are being integrated, it is crucial to note the certainty and precision level of such information. This additional geospatial attribute can then help in the process of finding duplicates and can guide further manual work on the dataset.

Beside coordinates, time is another vital component of most historical geodatabases. Even if temporal data types are a standard feature of most database systems nowadays, historical research often needs much more

sophisticated ways of handling time information. The problem is described by De & Deploige (2016), who argue that the time model should handle time uncertainty. Time information can take various forms: that of intervals, i.e. the terms post quem (limit after which) and ante quem (limit before which) for an event (e.g., erecting a monastery), but also that of uncertain dates without explicit mention of such limits in the relevant data sources (e.g., “around 1050”).

There are many available spatial historical data stored and presented in various forms, from printed atlases and tables accompanying books and articles to online maps and databases. These data are very valuable in their own right, but gain even more value if they can be used together. To achieve this goal requires the design of a suitable process of data integration – i.e., collecting, cleaning, and supplementing data from various sources, but also merging them into one dataset.

3. Method

To mine, transform, integrate, and post-process records of religious houses from the selected data sources, we created a workflow (see Figure 1) derived from the theoretical background presented in the previous section.

Preparation is the initial phase, whose crucial part is the creation of a unified data model to store attributes (see Figure 2) from heterogeneous sources. Since the history of Christian monasticism is dynamic (new orders emerge, old orders reform and create new branches etc.), one of the most valuable items of information is the list of religious orders that were active at the given place where the religious house was located. To standardize this list, we created a table of all major orders active within the borders of today’s France during the Middle Ages, and open-accessed it to other researchers (Zbíral, 2019). The table is classified according to branches of those orders and provides the dates of foundation and dissolution, category (monastic, eremitic, mendicant, military...), the existence or not of male and female communities in that order, and other information. Another important attribute that had to be standardized was temporal information. Time appears at different places in the dataset; we store time values for the presence of an order or branch but also for the status of the religious house (e.g., abbey or priory) and for noting down whether the community was a male or female one (see Figure 2). Just as we expected, a major challenge in parsing time information was the variation in the date values. To solve this, we designed a module that takes a time value in any usual form (e.g., 15th century, 1402-1548, around 1500) and returns a standardized object with “from” and “to” attributes, each with a post quem and ante quem value, and a note if needed. Finally, we compiled a list of data sources we chose to mine, described their characteristics, the expected issues with their mining, and the legal conditions of their use, and stored them with some additional values (e.g., if a dataset was explicitly a dataset of Cistercian monasteries, the order information was supplied for all records from this dataset).

Mining is the process of extracting content from the sources and retrieving all information possible. Each source needs a separate parser implementation that transforms the original according to the data model. We tried to incorporate the most relevant and valuable sources that we found, to achieve the highest possible completeness of the outcome. Two major sources that we used were the Digital Atlas of Roman and Medieval Civilizations (DARMC, http://darmc.harvard.edu/) and Wikipedia. The first is a Harvard University project that collects several spatial historical sources and displays them through a web map. The second, Wikipedia, is the most comprehensive open-sourced online encyclopedia, which has been proven to be a valuable source of data for further historical research (see Bhagavatula, Noraset, & Downey, 2013; Chasin, Woodward, Witmer, & Kalita, 2014; Hecht, Rohs, Schöning, & Krüger, 2011 or Hienert & Luciano, 2015). In the near future, we also plan to use printed maps and catalogues, which will have to be digitized manually or with the help of OCR.

A specific issue was the geolocation of the religious houses. In the case of Wikipedia, we could request the linked page of either the monastery or convent itself or at least the settlement in which (or close to which) they were situated, and derive the coordinates from there. More problematic are printed maps, where a precise georeference process is needed, and sources without coordinates, where we have to use a geocoder to obtain the most probable localisation.

The code for mining and processing sources is written in Node.js and TypeScript and uses various modules like Cheerio (https://github.com/cheeriojs/cheerio) for manipulation of the content of the website and Turf (https://github.com/Turfjs/turf) for geospatial operations. The mined records are stored in JSON format, which can be easily transformed into a CSV table or a GIS format.

Post-processing is a set of operations that are applied to the mined data. The first operation is supplementation, which means enriching the records with additional information from our table of orders and table of data sources. For example, if no explicit date for the foundation of a monastery is provided but the source dataset is that of religious houses between 370 and 740

A.D., we provide the year 740 as the latest possible date when that house was established (i.e., the terminus ante quem of its foundation). The next operation is filtering, which determines a subset of records by applying specified conditions. For this dataset, we filtered out records with an invalid geolocation or with a geolocation falling outside the boundaries of today's Metropolitan France. In addition, we filtered out records whose temporal information did not intersect with the period of 350-1500. Aggregation is the operation that identifies entities in the collected records, merges those which are, on the basis of specific conditions, evaluated as referring to the identical entity (e.g. the same monastery), and resolves possible conflicts. We implemented a strategy that prefers sources tagged as more reliable in case of conflict but allows the use of all values that are not in conflict. For every piece of information, we stored the data source, the date and time of mining, the version of the mining script, and all conflicting values, in order to allow the tracing of each value back to the original source. The last part of the post-processing phase is manual editing and validation. Here, we opted for building a database manager environment where researchers can edit the attributes of the processed records on the basis of additional research.

Presentation and analysis is an extension to the workflow that allows the visualization, sharing, presentation, and further analysis of the collected dataset. Some examples of the application can be found in the following section.

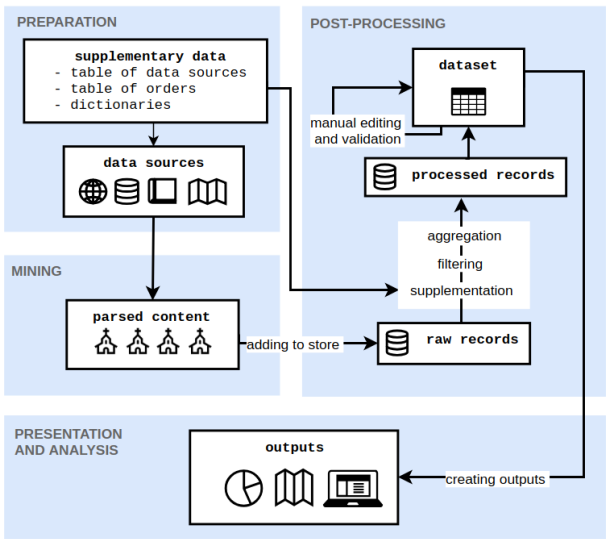

Figure 1. Workflow of data integration.


```
{
  "id": "1554905368516-c8tn3q",
  "names": [
    {
      "value": "Divielle",
      "primary": true,
      "long": false,
      "lang": "fr"
    }
  ],
  "link": "https://fr.wikipedia.org/wiki/Abbaye_de_Divielle",
  "geo": {
    "lat": 43.7392,
    "lng": -0.9244,
    "precision": 1
  },
  "statuses": [
    {
      "time": {
        "from": {"post": 1132, "ante": 1132},
        "to": {"post": 1209, "ante": false}
      },
      "id": 1
    }
  ],
  "orders": [
    {
      "time": {
        "from": {"post": 1132, "ante": 1132},
        "to": {"post": 1209, "ante": 1209}
      },
      "id": 15,
      "gender": "m"
    },
    {
      "time": {
        "from": {"post": 1209, "ante": 1209},
        "to": {"post": 1209, "ante": false}
      },
      "id": 3,
      "gender": "m"
    }
  ],
  "meta": {
    "timestamp": "Wed Apr 10 2019 16:08:01",
    "source": "3"
  }
}
```


Figure 2. Data model.

4. Results

The data we collected, integrated, supplemented, and automatically corrected following the workflow discussed above resulted in a comprehensive digital dataset of monasteries and convents extant between 350 and 1500 A.D. within the boundaries of today's Metropolitan France. It contains more than 2,500 individual records concerning medieval religious houses, ranging from Benedictine abbeys through Dominican convents to Templar commanderies. Every record stores the name and coordinates as well as order and branch, status (e.g., priory, abbey etc.), and gender of the community, including changes in these attributes through time. Some data are naturally missing in the records, which highlights the incomplete character of historical evidence and also the limits of the sources we mined. The most significant gaps concern the different phases in the history of a religious house. The data sources mostly reflect only the prevailing or latest status, order, branch, and gender, and ignore historical changes in these attributes. This is partly due to how demanding the collection of such data is, and partly to the prevailing format (a list or table rather than a database).

The collected data can be used in different ways. An obvious one is cartographic visualization that would

complement existing printed and online atlases. Among interactive web atlases, the Digital Atlas of Roman and Medieval Civilizations (DARMC) offers the broadest coverage of medieval religious houses (4617 records overall as of April 2019, of which 1389 refer to today's France) and forms a significant part of our dataset; however, DARMC data on monasteries and convents is quite incomplete (e.g., at the time of the mining, DARMC did not cover: Benedictines apart from the Cluniac branch, military orders, Carthusians, or Carmelites; etc.). In addition, the DARMC map application does not allow filtering, or the display of the number of houses of the different orders. Therefore, an integrated map of religious houses was created, which allows the filtering of records by order and branch, time span, gender, status, and data source.

The dataset allows scholars to explore the quantitative and spatiotemporal patterns of medieval Christian monasticism as well as to test particular hypotheses about it. For example, monks are known for the ideal of retirement "from the world", to the wasteland (i.e., to the "desert", which could be, for instance, the actual desert in Egypt or an inhospitable island in Ireland). This pattern is abandoned in the 13th century by mendicant orders such as Franciscans and Dominicans, who prefer urban settings and whose male branches engage in pastoral care. However, how deserted are the areas really chosen by traditional monasticism in medieval Europe from the perspective of both physical geography and population estimates? To take another example, Cistercians are well known for their preference for wetlands, which allowed them to deploy their remarkable technological skills in draining wetlands, managing water, and using waterpower (e.g., Bond, 1989; Raynaud & Wabont, 1998). However, did this really hold for most Cistercian monasteries, or is this "common knowledge" in fact based on several well-known but not necessarily representative examples? And is it really more typical of Cistercians than, for instance, Benedictines? In short, confirming or challenging some of the accepted knowledge about Christian monasticism requires a systematic and quantitative approach, for which a comprehensive dataset of religious houses is a precondition.

Finally, an integrated dataset of religious houses in medieval Europe allows their complex relations to society to be studied from the perspective of human geography. It has been argued that the tendency towards asceticism is positively correlated with the wealth of the surrounding society (Baumard, Hyafil, Morris, & Boyer, 2015). Is this pattern confirmed when we correlate the spatiotemporal dataset of medieval monasteries with proxies of wealth in particular regions of medieval Europe in particular periods? Other questions relate to religious houses as key parts of the official Church infrastructure. Does the presence of such infrastructure negatively correlate with occurrences of dissident religious cultures which entertain a high degree of tension with the official Church (e.g., Waldensians or Cathars)? A reasonably comprehensive spatiotemporal dataset of religious houses allows such questions to be addressed in a systematic way and thus helps conclusions based on anecdotic evidence to be avoided.

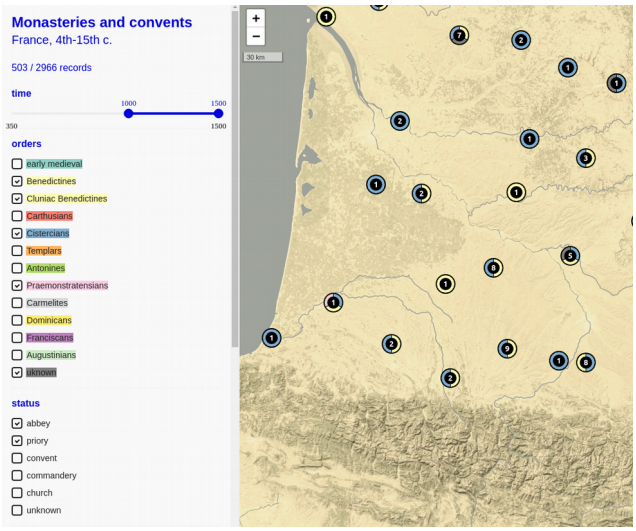

Figure 3. Map application.

5. Conclusions

The body of historical data available online is slowly but steadily growing, which, in turn, opens up unforeseen avenues for spatial analysis in this area of application. At the same time, attempts at using such data highlight their limits. The available data is often fragmentary, variously structured, not curated after the ends of the projects in which they originated, and often not geocoded.

In this paper, we presented the case of an aggregated and restructured dataset on medieval religious houses within the boundaries of today's Metropolitan France as a proof of concept for a data mining approach exploiting a broad range of online sources. Such an approach involves a workflow incorporating the cleaning and enriching of data by supplementing, filtering, and matching them; the detection of duplicates; and the resolution of conflicts between data sources. Despite all their limitations, such data can help verify (or otherwise) some of the accepted knowledge about Christian monasticism and its relations to space, as well as answer some intriguing questions which require quantitative analyses of a larger dataset of

religious houses. We plan to work further on the manual validation and supplementation of these data, as well as on broadening the geographic scope towards further areas, beginning with England, Lombardy and Tuscany.

6. Acknowledgements

The research presented in this paper is a part of the "Dissident Networks Project" (DISSINET, https://dissinet.cz) and received funding from the Czech Science Foundation (project No. GX19-26975X "Dissident Religious Cultures in Medieval Europe from the Perspective of Social Network Analysis and Geographic Information Systems"). We gratefully acknowledge this financial support.

7. References

Ardura, B. (1993). Abbayes, prieurés et monastères de l'ordre de Prémontré en France des origines à nos jours: dictionnaire historique et bibliographique. Nancy : Pont-à-Mousson: Presses universitaires de Nancy; Centre culturel des Prémontrés.

Baumard, N., Hyafil, A., Morris, I., & Boyer, P. (2015). Increased affluence explains the emergence of ascetic wisdoms and moralizing religions. Current Biology, 25(1), 10–15.

Becking, G. C. M. (2005). Zisterzienserklöster in Europa: Kartensammlung. Berlin: Lukas-Verlag.

Bleiholder, J., & Naumann, F. (2006). Conflict Handling Strategies in an Integrated Information System.

Bond, J. (1989). Water management in the rural monastery. In R. Gilchrist & H. Mytum (Eds.), The archaeology of rural monasteries (pp. 83–111). Oxford.

Bond, J. (2001). Monastic water management in Great Britain: a review. In G. Keevill, M. Aston, & T. A. Hall (Eds.), Monastic Archaeology: Papers on the Study of Medieval Monasteries (pp. 88–136). Oxford: Oxbow.

Bond, J. (2003). Monastic landscapes. Tempus.

Cocheril, M. (1976). Dictionnaire des monastères cisterciens I: Cartes géographiques. Rochefort, Belgique: Abbaye Notre-Dame de St-Remy.

Currie, C. K. (1989). The role of fishponds in the monastic economy. In The archaeology of rural monasteries (Vol. 203, pp. 173–184). Oxford.

De, T. G., & Deploige, J. (2016). Time modelling in digital humanities. Challenges posed by the development of a database of medieval charters. It - Information Technology, 58(2), 97–103.

Devogele, T., Parent, C., & Spaccapietra, S. (1998). On spatial database integration. International Journal of Geographical Information Science, 12(4), 335–352.

Digital Atlas of Roman and Medieval Civilizations. (2014). Retrieved April 8, 2019, from http://darmc.harvard.edu/

Emery, R. W. (1962). The friars in medieval France: a catalogue of French mendicant convents, 1200-1550. New York ; London: Columbia University Press.

Farnedi, G., & Togni, N. (Eds.). (2014). Monasteri benedettini in Umbria: alle radici del paesaggio umbro. Cesena: Regione Umbria ; Centro storico benedettino italiano.

Faniel, I., Kansa, E., Whitcher Kansa, S., Barrera-Gomez, J., & Yakel, E. (2013). The Challenges of Digging Data: A Study of Context in Archaeological Data Reuse. Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries, 295–304.

Gagnon, M. (2007). Ontology-based integration of data sources. 2007 10th International Conference on Information Fusion, 1–8.

Gerhards, A. (1998). Dictionnaire historique des ordres religieux. Paris: Fayard.

Hadcock, R. N., Great Britain, & Ordnance Survey. (1950). Map of monastic Britain: south sheet. Chessington, Eng.: Director-General of the Ordnance Survey.

Institut géographique national. (1995). France: abbayes et sites cisterciens. Paris: Institut géographique national.

Jedin, H., Latourette, K. S., & Martin, J. (1970). Atlas zur Kirchengeschichte: Die christlichen Kirchen in Geschichte und Gegenwart. Freiburg; Basel; Rom; Wien: Herder.

Kintigh, K. (2006). The Promise and Challenge of Archaeological Data Integration. American Antiquity, 71(3), 567–578.

Laszlo, F. (2018). Management of monastic landscapes: a spatial analysis of the economy of Cistercian monasteries in Medieval Hungary (Ph.D. thesis). Central European University, Budapest.

Milecka, M. (2012). Średniowieczne dziedzictwo sztuki ogrodowej klasztorów europejskich. Hereditas Monasteriorium, 1, 31–56.

Moorman, J. R. H. (1983). Medieval Franciscan houses. St. Bonaventure: Franciscan Institute Publications.

Naumann, F., Bilke, A., Bleiholder, J., & Weis, M. (2006). Data Fusion in Three Steps: Resolving

Inconsistencies at Schema-, Tuple-, and Value-level. IEEE Data Engineering Bulletin, 29(2), 11.

Oldman, D., de Doerr, M., de Jong, G., Norton, B., & Wikman, T. (2014). Realizing Lessons of the Last 20 Years: A Manifesto for Data Provisioning and Aggregation Services for the Digital Humanities (A Position Paper) System. D-Lib Magazine, 20(7/8).

Pelliccia, G., & Rocca, G. (Eds.). (1974). Dizionario degli istituti di perfezione (Vol. I–X). Roma: Edizioni Paoline.

Poras, G., & Cottineau, L. H. (1935). Répertoire topo-bibliographique des abbayes et prieurés (Vol. I–III). Mâcon ; Turnhout: Protat ; Brepols.

Raynaud, C., & Wabont, M. (1998). Réseaux hydrauliques de l'abbaye cistercienne de Royaumont (Asnières-sur-Oise, Val d'Oise). Actes Des Congrès de La Société d'Archéologie Médiévale, 6(6), 71–72.

Rasmussen, J. E. K. (2015). The foundation of Cistercian monasteries in France, 1098-1789: an historical GIS evaluation (M.A. thesis). Western Michigan University, Kalamazoo.

Vauchez, A., & Caby, C. (Eds.). (2003). L'histoire des moines, chanoines et religieux au Moyen âge: guide de recherche et documents. Turnhout: Brepols.

Wikipedia: the free encyclopedia. (n.d.). Retrieved April 8, 2019, from https://www.wikipedia.org/

Zbíral, D. (2019, April 7). Medieval religious orders of the Latin Church: Orders table. Retrieved April 7, 2019, from https://docs.google.com/spreadsheets/d/1oPjlu-9lAbADfbvpXnYJBqE5gDxuXWapq0k86jSj65g/edit?usp=sharing

Proceedings of the International Cartographic Association, 2, 2019.
29th International Cartographic Conference (ICC 2019), 15–20 July 2019, Tokyo, Japan. This contribution underwent single-blind peer review based on submitted abstracts. https://doi.org/10.5194/ica-proc-2-85-2019 |