A Database Model for Social History

Historical Data Grinder and the Transylvanian Society of 19th and 20th Centuries

Angela Cristina Lumezeanu

The HDG database has the possibility to accommodate historical information independent of the chronological period, geographic area or thematic range to which reference is made, which also makes it crowd-source friendly and suitable as a teaching and experimental tool.

Angela Cristina Lumezeanu
Ph.D. candidate, Doctoral School Population Studies and History of Minorities, Center for Population Studies, Babeş-Bolyai University, Cluj-Napoca.

Premises and Motivation

Amid the rise of new historiographical orientations in the last two decades, and also due to the rapid advance in information technology, the historiography of Transylvania at the present time benefits from a multitude of digital and biographical information search tools. The most advanced of these tools is the Historical Population Database of Transylvania (HPDT, http://hpdt.ro:4080/), a historical population database that includes over five hundred thousand people. Most of them are inhabitants of the Transylvanian rural area from 1780–1914[1] and the database was built based on the information provided by the parish registers.[2] Other small databases, presently in different stages of development, are dedicated to some socio-professional categories or specific research topics: the middling clergy, invalids, widows and orphans from the First World War, testaments etc.

The massive digitization of the nineteenth and twentieth-century periodicals (both historical and contemporary), and to a limited extent of the historical studies, allows the rather fast identification of biographical information about people who until recently were unknown and thus facilitates the work of researchers focused on Transylvania in the period 1848–1948.[3] Coexisting with these digital solutions there are also a relatively high number of biographical and prosopographic instruments (books, dictionaries, various lists),[4] unfortunately accessible only in printed format or, at best, in electronic versions made available by the authors.

Easy access to information does not mean, however, any ease when it comes to data collection methods, or to data processing and analysis, namely, the areas of research which are time consuming and often dictate the scale and depth of the research process. At this stage more problems are generated both by the diversity of sources and by the unrelated work of several historians on the same source. In the absence of a unitary system of information gathering and storage, the easy access to the sources is effective only on the level of individual research, as each historian builds their own work systems, according to their interests and knowledge. Even when historians are working together these systems may be incompatible with the ones used by their colleagues and require considerable time and effort in harmonizing the datasets, or even repeated turns to the source.[5]

Another problem of working with historical sources is the lack of a digital methodology. Until 2014 there were no historical databases in Romania whose building architecture followed a rigorous methodology and construction principles on which the Western literature has agreed in the last decades.[6] That does not imply that Romanian historians did not work in a meticulous manner, but only that they did not follow a common set of good practices, which led to a very heterogeneous outcome of their work. As a result, to standardize the existing prosopographical lists or biographical collections required, first and foremost, the development of a harmonization methodology based on a common framework to host the multitude of typological sources.

Taking these conditions into consideration, we are proposing a database model whose main objective is to accommodate information from any area of historical research; in our opinion, the model would prove more suitable for integrating the information that has been organized so far in multiple databases or independently built spreadsheets. The concept and model of the database have already been circulating for about a decade in the Western literature in the field of population studies under the name of IDS—Intermediate Data Structure. The need to build an homogeneous system for storing information on historical populations has occurred because, in the last 40 years, a multitude of population databases have been developed independently in the West, containing tens of

millions of records.[7] In order to be able to analyze and compare the data they collected, they proceeded to its integration in a common framework which ensures the unity of the data.[8]

The database model presented herein is also indebted to the Mosaic project for the idea of gathering and harmonizing datasets of various origin, and most importantly, for highlighting the large quantity of data still in spreadsheet format and its widespread use among historians.[9]

In Transylvania, in the last two decades, researchers followed the tradition of publishing biographical and prosopographic collections in printed form, which raised similar problems of data organizing however on a much smaller scale.[10] It is very possible that, in the absence of a common framework, these instruments will not be interconnected in the next decade, which would be a major disadvantage for the Romanian historiography in terms of complying with the requirements of current research.

At the present moment, the local practice of historical research also has to be taken into account. Despite some initiatives of historical database building, there are actually very few historians in the Romanian scientific milieu who have at their disposal a rigorously constructed tool. The great majority of them are not familiarized with database construction, nor have they interacted with a relational database or any other kind of databases. Most of them are still working with datasets preserved in spreadsheets (ms Excel/Libre Office Calc) or non-related ms Access tables. A rigorous methodology of data cleaning and standardization is rarely applied, while linkage is being done exclusively manually. Furthermore, most of the datasets they are using are stored locally on their own computers, or circulate by means of personal interactions in a small circle. They are not subjected to peer review, nor introduced into a wider access circuit and, in particular, cannot be associated with other similar sets, likely to contain identical or highly similar historical entities but with different attributes and values.[11]

The abovementioned premises also open the path to overlaps and duplications in data recording. Multiple historians can create datasets which overlap or complete each other to a certain degree. Sometimes they are extracted from different sources covering the same event/area/timespan, without the possibility of interconnecting the datasets, cross-checking them, or linking various common entities, for purely technical reasons, from the poor understanding of database functionality to more or less incompatible differences in the different databases/datasets architecture. Following the general historical research methodology and practices, Romanian historians would generally be able to successfully compare the results of their researches, but to a far lesser degree the datasets these results were based on.

Therefore, the realities of the Transylvanian research environment require the development of a tool that would be able to interconnect the information from multiple databases and also accommodate and harmonize a variety of datasets for historians with different interests, knowledge and objectives.

Making such an instrument would also be an incentive for the historians to make available their datasets, in exchange for access to a much wider amount of interconnected information. Last but not least, in order to make this instrument both useful and attractive for researchers less familiar with technology, it would need to function both as an aggregator of information from existing collections/databases and as a means to manually insert the information from any historical sources.

The Theoretical Model

The development of a conceptual model that brings together in a common framework the information found in multiple and varied sources from modern and contemporary Transylvania, some digitized, started in 2016. The initial aim was to study how different groups, mainly political ones, emerged and interacted at the level of the Romanian upper classes in Hungary and Transylvania during the second half of the nineteenth century, by researching the business, social, and kinship relationships of their members, previously subjected only to limited approaches.[12]

The sources that cover such a social network are numerous and varied. If best practices were to be applied in building a conventional database, a first version of it should be a faithful replica of the sources, or at least should allow the reconstruction of the original source content at any time, if need be, and this would imply the creation of a custom-made table for each type of source. However, there are instances when sources, even if belonging to the same typological category, present so much diversity and so many minor differences that they require the development of a very complicated database architecture in an attempt to fully cover the ever increasing number of variables. Then, there is also the question of connecting the information with the already existing databases.

Considering these factors, we believe that the Entity-Attribute-Value (EAV) model responds much better to the needs of the project mentioned above and takes better into account the specificity of historical information.[13] The EAV model is also best suited for the computer knowledge of Romanian historians.

In relational databases the conventional way of defining attributes is by using one column for each attribute. This is useful when the number of attributes is

not large and they can be applied to a vast majority of entities. However in the situation described above the sources are so diverse and the attributes numerous and specific to one source that we cannot apply the classical model. In most cases we would end up with a large number of columns, most of them with empty values. To solve this problem we applied the Entity-Attribute-Value (EAV) model. The EAV model was first used and defined in bio-medicine and it is considered the best solution when dealing with heterogeneous data with continuous changes of attributes, as Prakash Nadkarni and his team contend in their study.[14]

In the EAV model only the non-empty values are preserved into the database. Each attribute-value pair describes a single attribute of a given entity. These tables are usually characterized as being "long" and "skinny" because there are multiple rows describing the same entity and they do not use too many columns.[15]

At the basis of EAV model is the so called "row modeling"—a standard component of a database design set. Row modeling means that all the facts related to an entity are recorded in a single table across the database containing a column for the entity, one for the attribute and another one for the value. Each attribute of the same entity is recorded in a separate row.[16]

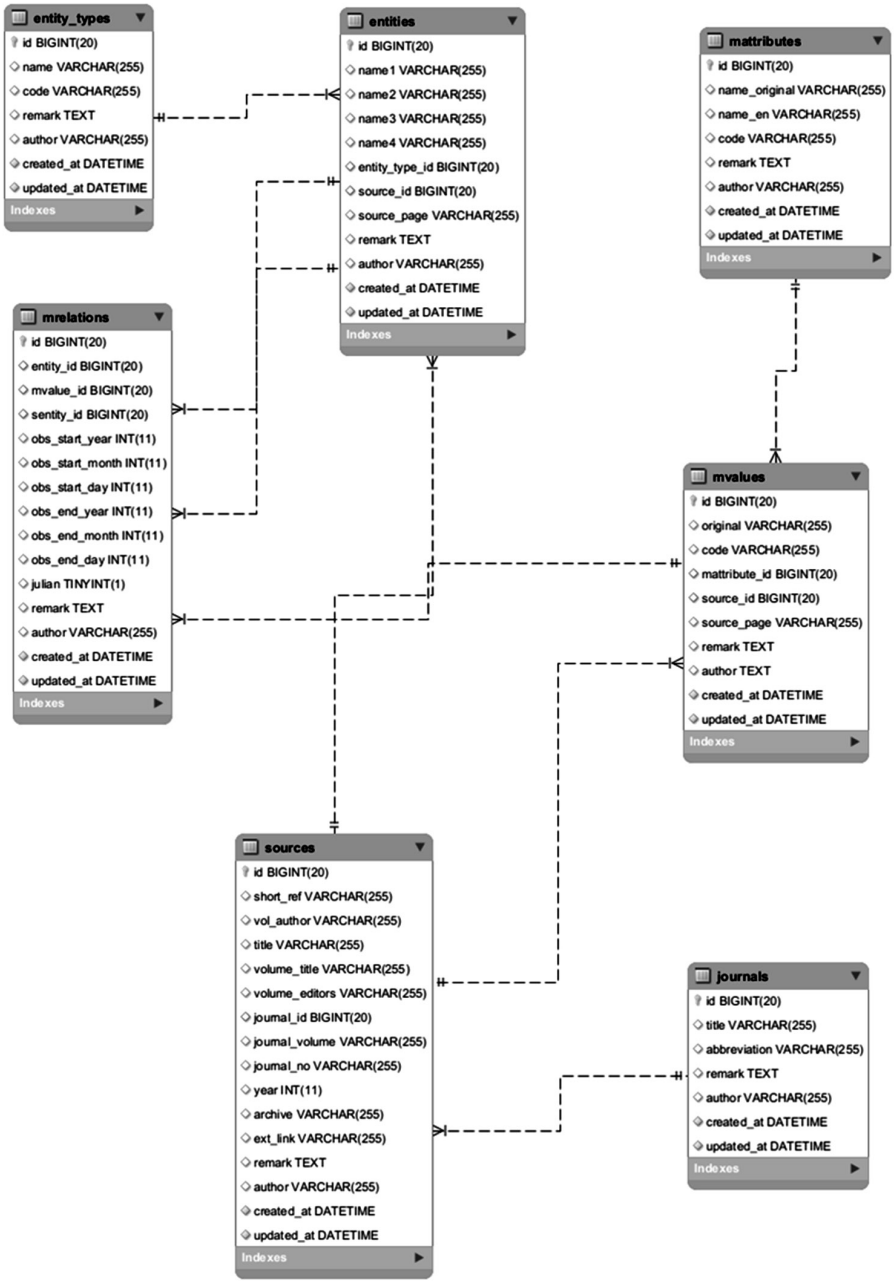
The EAV design generalizes the "row modeling" standard using different tables for entities, attributes and values. It is more efficient when dealing with large classes of attributes and entities, when we need to differentiate between data-types (string, integer, date-time, decimals etc.), or when the database is in constant change in order to adapt to new attributes.[17]

The main advantage of the EAV system is its flexibility. The simplest form of this design contains only three tables for entity, attribute and value, and the metadata table. Adding a new attribute to the entity does not require creating new columns and redesigning the database by the programmer. It is a simple task of adding another row to the attribute table, which can be done by the researcher. The data format is very clean, the relations are very clear and it is much more suitable when new attributes need to be stored in the database on a regular basis.[18]

For the historian, the main challenge is understanding how the EAV design works. In order to insert the data, the historian has to decompose and assign the information from the primary source according to the specificity of the EAV model. We created a model named Historical Data Grinder (HDG) where the information is decomposed according to the EAV design (see the Diagram below).

With complex and diverse types of sources and information, the simple model of three tables was not enough. We also added a few tables created in the conventional relational model that help in dealing with large classes of entities and with the relationships between them.

DIAGRAM. DESIGN MODEL OF MIXED-SCHEMA IN HDG



SOURCE: Historical Data Grinder database architecture.

This *mixed-schema design* is often used when we have complex entities and relationships that cannot be separated in just three tables, but the core of the database is still the EAV tables.[19]

For HDG any given historical information, independently of the type of source that provides it, can be ground into these basic elements and subsequently reordered and integrated into a simple structure of variables. The historian's job is to carefully assign each type of historical data to the aforementioned three categories, to correctly transpose the historical relations into the database and to add the timestamp and the source to this aggregate.[20]

Entities may be of various types: beings (human or animals, real or imaginary), places (geographical or administrative units, localities, etc.), institutions or any other physical or social entity whose characteristics or activity are recordable, measurable, or can be ordered (e.g. objects).

They are managed through the *Entity* table that contains the main facts about the entity, such as name and alternatives names, source, source_page, and remarks. Each entity has a unique identifier, assigned by the computer, which is to be used across the database to describe it. The entities from the HDG can be organized in several classes found in the *EntityType* table, such as Locality, Individual, Document etc. This is of a mixed design and it helps in organizing the diversity of entities.[21]

Values are the qualitative or quantitative characteristics of entities. A value may be the same for a virtually infinite number of entities. Values cover a broad spectrum, from a person's eye color to the geographical extension of an administrative unit. Professions, body measurements (respectively, the technical specs of a machine, or the area of an administrative unit), pathological details (including the cause of death), school results, all fall into this category.

Attributes are a more versatile category. On the one hand, they can group large sets of values and thus define the relation between an entity and a value. For example, "Profession" is an attribute that defines the relationship between a particular individual (the entity) and the type of activity he performs at a given time (the value = the professional title). On the other hand, they can define the relation between two entities. "Residence" or "Birthplace" are attributes that link an individual (entity no. 1) to a particular topographic location (entity no. 2), whether it is a country, a county, a sub-county, a city, a neighborhood, a street, a house, etc. It is possible for the same attribute to link an entity to both a value and another entity. "Education" is an attribute that links a person with the different values defining his or her position in a schooling process, from school results to its school status, but it can also link the same person to a specific educational facility (school, university, etc.). "Health" is an attribute that binds a

person to all pathological manifestations recorded throughout her/his life (values), but it can also link the respective person to a hospital (entity).

Because of this particularity we have two types of relationships. One is the normal EAV relation connecting the entity with the value through an attribute. The second follows the connection between two different entities, and in this case we have a conventional relational design.

Chronological consistency is provided by a timestamp marking the moment in time, or the period of observation of a relationship. The timestamp is associated directly in the relations table, and only through these with entities or values, because a source captures a relation between an entity and a value, or between two entities, for a very specific period: *ante quem*, *post quem*, or in between.

A distinct set of tables allows for the information to be linked to the source, including common hyperlinks, if the source is available online. It was not integrated in the EAV model because the easy access to the primary source is very important to historians, and in this way it is ensured that every row of the EAV table has a direct connection to it.

Each table that records entities, attributes, and values is provided with a manual coding system that does not necessarily involve the existence of a codebook when inserting information into the database. In this way, the variety of the information that enters the database is in no way restricted, the codebooks for all three types of variables can be easily developed alongside data gathering and, more importantly, can be changed at any time without affecting the structure, the functionality or the coherence of the database. It can also be added at a later time. Thus, in the current structure, the base can theoretically accept any kind of historical information, from the reconstruction of family relations in small circles or in local or geographical clusters, to the reproduction of the evolution of administrative changes, with their modifications in time and even with the related population if desired (the latter in turn divided into categories, as needed).

The history of social relations, public administration systems, medical history, military history, political history, or the history of science and technology can be also integrated with this database model. In addition, they can all be integrated at the same time, by different people, not necessarily sharing a common research goal, each working on their own narrow subject, and it only needs a supervisor to harmonize, standardize and code the data. If need be, of course, multiple coding systems can be implemented concomitantly, one for users with individual research interests and another one for the administrator in charge of standardization. In this way, anyone can use the database according to their own interests, necessities and expectations, making it ideal for crowd sourced data gathering as well as teaching activities.

Operating Mode

There are two methods of inputting information into the HDG: by massive ingestion of previously processed information from spreadsheet files, or by manual input. For the first method the data have to be harmonized and highly standardized, since the general intent is to eliminate redundancy.

The combination of the two methods makes it possible to retrieve data from almost any relatively structured electronic system file, from encoded text to spreadsheets, while at the same time allowing public access and manual data entry—the latter essential for manual linkage.

Envisioned Benefits

The HDG database has the possibility to accommodate historical information independent of the chronological period, geographic area or thematic range to which reference is made, which also makes it crowdsource friendly and suitable as a teaching and experimental tool.

It has an increased ability to act both as an aggregator, and as a stand-alone database.

Having a simple design, the costs for construction and maintenance are very low; the simplicity of the architecture also makes it very easy to replicate and use offline, which in its turn opens the way for collaboration and eases future datasets integration. It is also very easy to transfer the data to XML format.

As we mentioned above, the HDG has flexibility in relation to the sources: instead of adapting the database and the number of variables to each new source (which means constant financial investment and horizontal database expansion), it suffices to adapt the attribute, entity, and value tree codebooks. It is obviously much easier and cheaper to adapt the codebook than the database architecture.

Based on the reconstructed biographies, the HDG also offers the possibility of better understanding the relationship between the family environment and an individual's social trajectory. This type of interconnection provides not only more complete biographic information but it also opens up the path to a deeper level of analysis based on a larger set of variables. Questions concerning the influence of family life on one`s studies and career or the underlying reasons of university dropout rates may be answered by such more detailed knowledge.

For its great flexibility, the EAV model gives up a few advantages. A major disadvantage is that the retrieval of the information in complicated relations is less efficient than from relational databases. The information is very fragmented,

so working with large quantities of data requires a lot of join tables and programming skills. This is why the historian has to decide based on a very good knowledge of the sources when the EAV model is to be used. If the data is heterogeneous, with numerous attributes and with new ones often needed, and they do not apply to every entity, then the EAV model might be the best solution. It is simple and efficient, does not occupy a lot of physical space and can be easily adapted to the needs of the research.

Conclusions

THE ENTITY-Attribute-Value (EAV) model responds much better to the needs of the historians who use various sources and need to integrate them into a complex analysis. Flexibility is the main advantage proposed by the EAV model, while the main challenge is understanding how the EAV design works and how to join multiple data into a coherent model that allows complex analyses. It can support multiple projects with broad subjects in a single build form, thus making it accessible for a large community of researchers.

❑

Notes

1. Angela Lumezeanu, "Insights into Designing and Building a Historical Population Database," *Romanian Journal of Population Studies* 12, 2 (2018): 77–98.
2. Vlad Popovici,"Parish Registers from Transylvania—Sources for the History of Medicine (Late 18th–Early 20th Centuries)," *AMHA: Acta medico-historica Adriatica* 13, 2 (2015): 287–302; Bogdan Crăciun, Crinela Elena Holom, and Vlad Popovici, "Historical Population Database of Transylvania: Methodology Employed in the Selection of Settlements and Micro Zones of Interest," *Romanian Journal of Population Studies* 2, 9 (2015): 17–30; Elena Crinela Holom, Oana Sorescu-Iudean, and Mihaela Hărăguş, "Beyond the Visible Pattern: Historical Particularities, Development, and Age at First Marriage in Transylvania, 1850–1914," *The History of the Family: An International Quarterly* 23, 2 (2018): 329–358.
3. Among the largest collections covering Transylvania: https://www.arcanum.hu/en/; http://dspace.bcucluj.ro/; http://www.transilvania100plus.ro/index; http://www.digibuc.ro/.
4. Cornel Sigmirean, *Istoria formării intelectualităţii româneşti din Transilvania şi Banat în epoca modernă: Studenţi români la universităţi din Europa Centrală şi de Vest* (Cluj-Napoca: Presa Universitară Clujeană, 2000); id., *Intelectualitatea ecleziastică: Preoţii Blajului (1806–1948)* (Târgu-Mureş: Ed. Universităţii Petru Maior, 2007);

id., *Formarea elitelor militare ale Imperiului Austro-Ungar: Studenţi transilvăneni la Academia Militară 'Ludovika' din Budapesta* (Târgu-Mureş: Ed. Universităţii Petru Maior, 2011); Vlad Popovici, *Studies on the Romanian Political Elite from Transylvania and Hungary (1861–1918)* (Cluj-Napoca: Mega, 2012); Mirela Popa-Andrei et al., eds., *Canonici, profesori şi vicari foranei din Biserica Română Unită (1853–1918): Dicţionar* (Cluj-Napoca: Mega, 2013); Judit Pál, Vlad Popovici, Andrea Fehér, and Ovidiu Emil Iudean, eds., *Parliamentary Elections in Eastern Hungary and Transylvania (1865–1918)* (Berlin: Peter Lang, 2018).

5. Although they used an overlapping sample of people, namely the priests of the Romanian Greek Catholic Church, extracted from similar archive records, the methods employed by C. Sigmirean (*Intelectualitatea ecleziastică*) and distinctly by the team of the project directed by M. Popa-Andrei (*Canonici, profesori*), as well as their way of presenting the results of the research, are completely different.
6. Kees Mandemakers and Lisa Dillon, "Best Practices with Large Database on Historical Populations," *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 1, 37 (2004): 34–38.
7. For an extensive approach regarding the existing European population databases we recommend browsing the list compiled for the European Historical Population Samples Network, at https://ehps-net.eu/databases.
8. George Alter, Kees Mandemakers, and Myron P. Gutmann, "Defining and Distributing Longitudinal Historical Data in a General Way through an Intermediate Structure," *Historical Social Research/Historische Sozialforschung* 3 (109), 34 (2009): 78–114.
9. For a detailed presentation on Mosaic database creation and the main issues the creators had to deal with, see Mikołaj Szołtysek and Siegfried Gruber, "Mosaic: recovering surviving census records and reconstructing the familial history of Europe," *The History of the Family* 21, 1 (2016): 38–60.
10. See note 4.
11. As an example, we mention two volumes that include the lists of victims and material losses among the Romanians in Transylvania during the 1848–1849 Revolution, as compiled by the Greek Catholic and Orthodox Churches. They use various archive records, but the comparisons are difficult, since the data is not integrated. See D. Suciu, coord., Alexandru Moraru, Iosif Marin Balog, Diana Covaci, Cosmin Cosmuţa and Loránd Mádly, eds., *Revoluţia transilvană de la 1848–1849: Date, realităţi şi fapte reflectate în documente bisericeşti ortodoxe* (Bucharest: Asab, 2011) as well as Dumitru Suciu, coord., Alexandru Moraru, Iosif Marin Balog, Diana Covaci, Vlad Popovici, Loránd Mádly, and Cosmin Cosmuţa, eds., *Războiul naţional de la 1848–1849 reflectat în documente bisericeşti greco-catolice* (Cluj-Napoca: Argonaut, 2014).
12. Vlad Popovici, "Family relations and group mobilization within the Romanian political elite in Transylvania (1861–1900)," *Transylvanian Review* 21 (2013), suppl. no. 2, *Economic and Social Evolutions at the Crossroads of the World-System,* eds. Iosif Marin Balog, Rudolf Gräf, and Cristian Luca: 107–118; I. M. Balog, "The Clergy's Involvement in the Romanian Credit System from Transylvania during the Late

Nineteenth and the Early Twentieth Centuries. Case Study: the Greek-Catholic Clergy," in *Recruitment and Promotion among the Romanian Greek-Catholic Ecclesiastical Elite in Transylvania (1853–1918): A Collection of Studies*, eds. M. Popa-Andrei et al. (Cluj-Napoca: Mega, 2014).

13. Prakash M. Nadkarni, Luis M. Arenco, Roland Chen, Emmanouil Skoufos, Gordon Shepherd, and Perry Miller, "Organization of Heterogeneous Scientific Data Using the EAV/CR Representation," *Journal of the American Medical Informatics Association* 6 (1999): 478–493.
14. Ibid.
15. Robert Raszczynski, "The EAV data model," https://inviqa.com/blog/eav-data-model, accessed 23.04.2019.
16. Valentin Dinu and Prakash Nadkarni, "Guidelines for the effective use of entity-attribute-value modeling for biomedical databases," *International Journal of Medical Informatics* 76, 11–12 (2007): 769–779.
17. Ibid.
18. Ibid.; Prakash Nadkarni and L. Marenco, "Database Architectures for Neuroscience Applications," in *Neuroinformatics*, ed. Chiquito Joaquim Crasto, Methods in Molecular Biology, 401 (Totowa, New Jersey: Humana Press, 2007), 37–52.
19. Dinu and Nadkarni.
20. A brief presentation of HDG's structure, focusing on the usefulness of the database for prosopographical research, rather than the actual conceptual model, has been published by Vlad Popovici and Rada Varga, "Building Life Courses and Explaining Life Choices with the Help of Digital Prosopography," *Studia Universitatis Babeş-Bolyai: Digitalia* 63, 2, 3–4 (2018): 55.
21. Nadkarni et al., "Organization."

Abstract

A Database Model for Social History: Historical Data Grinder and the Transylvanian Society of 19th and 20th Centuries

The study presents a new model for building a historical database, namely, the Historical Data Grinder (HDG). It is based on the EAV design model developed in bio-medicine and it offers some advantages for the historian especially when dealing with a large variety of heterogeneous sources. The database is very flexible and has a simple architecture, with fewer tables and relations when compared to a relational database. The HDG database has the possibility to store any kind of historical information with no limitations regarding the time period, geographic area or thematic range to which reference is made, and is also suitable as a teaching and experimental tool.

Keywords

EAV design model, historical population database, Historical Data Grinder, Transylvania, heterogeneous sources