‘ArchiteXt Mining’ project:
Developments and Adjustments since the 2017 Symposium in Kyoto

Ana ESTEBAN-MALUENDA[1]
Technical University of Madrid

Luis SAN PABLO MORENO[2]
ArchiteXt Mining Project (MINECO-ERDF)

Francisco FERNÁNDEZ RODRÍGUEZ[3]
ArchiteXt Mining Project (MINECO-ERDF)

Abstract

A little less than a year ago, we went to Kyoto to present the ‘ArchiteXt Mining’ project in the International Symposium ‘Architectural and Planning Cultures across Regions. Digital Humanities Collaboration Towards Knowledge Integration’, organized by the Center for Integrated Area Studies of the Kyoto University. We lived a few intense days in which we were able not only to explain our work and listen to the doubts and suggestions that it provoked, but also to listen to the presentations of other interesting works, with different approaches from ours and always oriented to the improvement of humanities studies.

In December 2017, we had another opportunity to expose our research in another seminar organized by the Dipartimento di Architettura e Studi Urbani of the Politecnico di Milano under the title ‘Mapping visions, discourses and theories. Journals as platform for architectural and urban Knowledge. A network of projects’. There we presented the project again, adding some of the adjustments we had made since Kyoto.

This paper aims to relate some of the reflections – and the consequent changes or improvements - that have derived from these meetings and the acquired experience from the last year. First of all, a more technical approach than the one carried on the encounters and the final definition of the methodology that is being used to calculate the similarity between texts will be described. After that, it will be developed a reflection that intends to deepen some of the main questions that were formulated in both meetings. As a final point, it will be shown the advances we have done in order to establish the basic structure of the website that will host the definitive online tool.

Keywords: Architectural Periodicals, Text Mining, Spanish Modern Architecture, Data Analysis.

‘ArchiteXt Mining. Spanish modern architecture through its texts (1939-1975)’ is a research project funded by the Government of Spain through the 2015 Call for ‘Excellence Projects’ of the Ministry of Economy and Competitiveness. Architext Mining started in 2016 and is scheduled to be completed by the end of 2018, although a one-year extension may be requested if we needed.

This project aims to explore a new viewpoint and look into the special features of Spanish modern architecture. Despite the success of the development of data analysis as a tool in different disciplines, the research on architectural theory has never made an efficient use of these technologies. The Spanish and international circumstances of the development of the Modern Architecture has been scrutinized through qualitative research which established a shared theoretical ground. It is time to face a new in-depth research based on the collection and analysis of objective data. In order to obtain this, we propose the application of text mining techniques to take advantage of the best data source in the field: architectural magazines.

Most of the scholars from the field of architectural history resort to magazines as a relevant source for research. But we continue to do it exactly as we did fifty years ago, that is, we go to the archives and revise page by page every issue in search of data. In the last few decades hundreds of partial indexes have been created in which the basic bibliographic data of the main articles are included, but they are incomplete and do not usually contain data referring to smaller sections, like the ‘news’ or ‘letters’. In any case, the vast amount of informa-

1. Tenured Associate Professor. Universidad Politécnica de Madrid (Spain). Principal investigator. ArchiteXt Mining project.
2. Data Scientist. BNP Paribas Spain. Data Scientist. ArchiteXt Mining project.
3. Data Scientist. BNP Paribas Spain. Data Scientist. ArchiteXt Mining project.

tion they contain makes them incomprehensible to the human mind. We need to develop better methods of collection and analysis of data, and there is a way to transform that enormous 'database' into a readable format that will allow machines to support us in assimilating everything they have.

ArchiteXt Mining (that is the acronym of Architectural Text Mining) proposes to use the most advanced techniques of data analysis for the creation of a new tool. This facilitates the work of all the researchers who use architectural magazines as a source of information. The current computer engineering possibilities allow us to raise something that was impossible up until now: to perform a global analysis of the contents of architectural magazines. We aim at creating a powerful database hosted in a public website accessible for the global scientific community. In this context, this project fulfills some e-Research objectives:

- An increase in the computerization degree of research processes, taking advantage of networking. In this way, the collaboration among researchers will increase not only from within a national sphere, but also internationally, through the implementation of a collaborative environment that will provide information to users while they will provide feedback through their proposals and their particular searches.

- The development of a tool that support every stage of the research process, from the data gathering, its processing and analysis, to the dissemination of the results.

- And what is of no less importance: the use of information visualization tools that give meaning to the large volumes of data that will be managed.

As a first step, this project focuses on the Spanish case as a pilot experience that foresees to reach a larger worldwide scale of research. In particular, this stage begins with the Spanish architecture magazines published during the dictatorship period (1939-1975). The Spanish architecture and its media are a well-known research field for the project team. Therefore, the evolution and the changes that the architecture and the architectural thought have passed throughout the decades that Francisco Franco ruled provide immense possibilities for contrasting the first and last moments of the studied period.

Regarding the magazines, we have already digitized the journals of the Institute of Architects of Madrid (*Revista Nacional de Arquitectura and Arquitectura*) and the one of the Institute of Architects of Barcelona (*Cuadernos de Arquitectura*). Apart from these, we proposed to complete this material with other important Spanish periodicals, like *Hogar y Arquitectura* and *Nueva Forma*. The initial aim was also to scan and digitize some European periodicals: *L'Architecture d'Aujourd'hui* (from France), *The Architectural Review, Architectural Design* (from Great Britain), and *Domus* and *Casabella* (from Italy). But due to the reduced budget of the project –compared to what we have applied for- these European sources will not be included in the first phase. Following budgetary definitions, we will only focus on the journals of the institutes of architects of Madrid and Barcelona. Despite this obligatory and drastic reduction, we believe that these sources will allow to cover a reasonably complete framework of the Spanish architectural panorama of the time.

With the general objective to develop new research methods for the architectural studies, this project aims at creating a database of the modern architecture information published in the Spanish media, open to the academic sphere that exceeds the traditional bibliographic data and contents additional values. This biblio-thematic database follows an initial classification according to the traditional formula, done by the members of the research group. However, it provides a lot of information that is not included in traditional bibliographic indexes. In addition to the bibliographic data of the text (title, author, journal, issue, year and pages), we have been recording another important data in terms of the text type (article, review, new…), a brief description of its topic, the description of the section where the text is included, data about the building, personality or event that is dedicated to, and so on. Getting access to this information is already a big step forward and provides the researchers a powerful tool to engage in the first quantitative analysis and searches that would help them to begin their studies. In this moment, we have four people engaged in the process of building the biblio-thematic database: two undergraduate students, who are writing down the data as well as doing all the digitalization and the processing related to the optical character recognition work, and two PhD junior researchers, who validate and review their work. The columns of data entry in the database have been created by the eight senior researchers of the project. They are professors affiliated to four Spanish universities: Universidad Politécnica de Madrid, Universidad de Salamanca, Universidad de Navarra and Universidad de Alicante. Besides

them, the research team includes also a professor of Politecnico di Milano, expert in architectural periodicals.

Just with the creation of this database, the project has been justified. But, in addition we aim to bring additional values. Here is where the text mining techniques come into play to apply different statistics techniques, and to obtain another kind of information from all the texts stored in our database. To do this, we count with two data scientists who are testing the possibilities of application of text mining techniques over the data collected at this database.

ArchiteXt Mining methodologies

Envisaging the automatization of printed text requires several steps of preparation. Generally, scanning physical documents is one of the most difficult tasks that we have to face. Magazines are stored in libraries and archives and we need to work right there in order to obtain the digitalized material that will become our final source of data.

In order to get a high-quality scanning of documents, periodicals, and magazines, specific scanner devices are required. For that, this project is counting with the support of the Library of the School of Architecture of the Universidad Politécnica de Madrid (ETSA-UPM). On the one hand, the ETSA-UPM Library maintains the collection of all the issues of the main Spanish architectural periodicals, but, in addition, they allow us to use their professional scanner to engage in the digitalization of materials. This device is a Metis EDS Alpha which provides us with superior quality images, easier use and high productivity (about 30 full scans per minute). This has allowed us to store the entire scanned collections of the following three magazines: *Revista Nacional de Arquitectura* (from 1941 to 1958, issues 1-204), *Arquitectura* (from 1959 to 1975, issues 1-197) and *Cuadernos de Arquitectura* (from 1944 to 1975, issues 1-111).

On the other hand, we have started the process of optical recognition character of every text with the Abby Finereader 14 software, that allow us for the automation of the digitalization process. In short, with the digitalization of the periodicals and the processing of optical character recognition we have obtained a significate extra for our biblio-thematic database, one important more step to be differentiated from the traditional ones. Once we have this material in an electronic format (txt, doc, rtf…) we will become able to compute this data.

The first task is to obtain the so-called DNA of the different texts and automatize our work with them. A Text Matrix Document (TMD, the text's DNA) is the text essence, a matrix that serves to analyze the presence and distribution of words in a text. For this purpose, it is necessary to remove the words without significance when they stand alone by themselves, these are denominated stop-words. Stop-words are articles, prepositions, conjunctions and those words that are irrelevant for the purposes of our analysis. These words are very important for connecting phrases and paragraphs in language, but they have a reduced impact in the analysis of the meaning of texts.

The basic unit of analysis is the magazine article, which is a text that can be associated to a particular title. In that sense, we consider that a long text, a brief news report or a book review are similar in order to obtain the TMD. This is one of the fundamental steps that we need to add to each metadata record. Having this info available in a standard format in relation to the rest of the metadata, we can propose searches, comparisons and other data treatments which processing is an impossible task for the human brain.

The basic task is to calculate a ranking of frequencies of words. This is an interesting exercise by itself because it provides the most frequent terms that appear in the text and therefore some clues about the significance and sense of the text. But the ranking of frequencies not only serves to obtain a comprehensive view of the text contents. It can be stored in our database and used for automatic treatment of the information, as the study of similarities between texts. Once the process of clarification of the DNA of two texts is completed, we can compare them and calculate a certain similarity percentage (SIM). This percentage provides us with objective criteria to find similar text series before reading them. Of course, it is not an exact science and it could lead to little mistakes. In any case, it helps us to reduce radically our search framework.

Regarding in deep the different methods that statistics provide for this exercise, we have tried to use three different techniques to calculate SIMs which, however, share a unique starting point: the TDMs. Overlapping them two by two, we can calculate some indicators already defined in the statistical literature[4].

The first one is the *summation of products of pondered frequencies range*, which is a sort of terms based on a scalar product of frequencies of

common words between two texts [d1 and d2]. All of them are pondered by the total number of words they have.

SIM(d1,d2) = ∑xiyi = x1y1 + x2y2+ x3y3+ …… + xnyn

In the previous formula, xi and yi are the pondered frequencies of the i-word [wi] in the documents d1 and d2, respectively. The pondered frequency of a certain word [wi] in one text is obtained by the quotient between the frequency of that word [c(wi,d1)] and the total number of words in the document |d1|.

xi = c(wi,d1)/|d1|

yi = c(wi,d2)/|d2|

The second indicator is the *product of summations of pondered frequencies range*. In fact, this indicator and the previous one looks at similar calculus in terms of statistics and use the same pondered frequencies. We need to notice the importance of pondering the results. It is also remarkable that most frequent words in these two indexes have more weight in the calculus.

SIM(d1,d2) = ∑xi∑yi = (x1 + x2+ x3+ …… + xn)X×(y1 + y2+ y3+ …… + yn)

On the contrary, the Jaccard Index works with the simple appearance of words, without a special focus on how many times the word occurs in the text. This index shows the cardinality of the intersection of both texts (d1 and d2) divided by the cardinality of their union. Here, it doesn't matter how frequent certain terms are, but their presence in an absolute ratio. We give less importance to how many times a word appears in a text giving chance to the fact that the words have been mentioned at least once.

SIM(d1,d2) = |d1 ∩ d2| / | d1 ∪ d2|

After calculating these three indexes in many couples of texts, we have concluded that they tend to maintain a similar organization between them. But balancing the results we have chosen Jaccard Index as the optimal index for various reasons. First of all, this one usually remains in the middle of the three indexes in terms of values. The second reason is that Jaccard only uses concepts of union and intersection of word sets, that is very easy to understand even for beginners in statistics. Finally, we prefer this index because it doesn't beneficiate high words frequencies' in opposition to those with low frequencies.

Values obtained with this method have been analyzed by the two data scientists of the project who have established a criterion that conclude the acceptable index values between two texts: beyond 20-25 per cent of similarity, we can consider they talk about similar topics.

Some questions solved in 2017

The International Symposium 'Architectural and Planning Cultures across Regions', organized by the Center for Southeast Asia and Area Studies of the Kyoto University in March 2017, was the first occasion that we had to present the methodology of analysis to an academic audience. From that discussion, there were three basic questions formulated to us:

1. How languages and translations can affect the SIM indexes calculus?
2. Under what aspects do we want to observe the distribution of a certain concept in a text?
3. How can we consider the existence of synonyms, polysemy words and dictionary occurrences?

Curiously, in every meeting we have attended in since Kyoto, these questions have been reoccurred in one way or another. Therefore, we have strived to answer them.

The first one alludes to the future difficulties that could arise when we try to reach across the Spanish borders and *compare texts written in different languages*. Practicing this possible situation, we have done several experiments of similarity between two texts written in the original language – in our case, in Spanish - and its translation in English. Using automatic translators (Google Translator or similar), the results of similarity reproduce a common pattern: the SIM index is higher with the texts translated into English than in the original language in Spanish. In addition, this rule is followed by the three different methods exposed before.

Our theory explaining this increase in similarity is that automatic translators simplify the collection of words used in texts, having a standard language to express the same idea. Probably in the original texts the presence of synonyms, polysemic words and other aspects of language give more wealth in

4. Feldman, Ronen; Sanger, James. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* Cambridge: Cambridge Univ. Press, 2007.
Colas, Fabrice; Brazdil, Pavel. On the Behavior of SVM and Some Older Algorithms in Binary Text
Classification Tasks. In: Sojka, Petr; Kopeček, Ivan; Pala, Karel. (eds) *Text, Speech and Dialogue* [9th International Conference, TSD 2006, Brno, Czech Republic, September 11-15, 2006. Proceedings]. Part of the Lecture Notes in Computer Science book series, vol 4188. Berlin, Heidelberg: Springer, 2006, pp. 45-52.

terms of lexicon, but this is interpreted by this tool decreasing the value of coincidence, and consequently the index value.

So, comparing texts written in several languages into a standard language, like English, must be considered carefully. We need to quantify the SIM decrease that is involved in the process of automatic translation of texts. In this moment, we have estimated that automatic translation processing forces the reduction of this SIM index by a 10-15 per cent. In this relation, we still have a long way to go, until we become able to answer questions related to rather the indexes suffer the same reduction in processes of translating into English from any language. Other question is which is the best method for comparing a text translated into English with another text originally written in that language. But, we are paving the way of this work in progress.

The second aspect which has been frequently addressed to us regards the definition of parameters that we have considered to observe the presence of word or ideas in a text and how can these parameters can be shown to the site users[5]. The first parameter we are going to consider is the *global frequency of a word* in a text. This is the basic approach we can use to evaluate the significance of words in a text, and is the main approach used by the graphic tools that we can use to show the *global frequency of a word* and which are very popular in the contemporary media. We are speaking about the classic bar graphs or the much more effective word-clouds, from which we can capture the main ideas in a blink of an eye.

Other parameters that we have taken into account is the *correlation between words*, that express which words are distributed positioned near to others, showing the strength of certain conceptual links. Graphically, we can express these relations by network maps or link graphs.

But usually the presence of a certain word is not continuous throughout a text. In that sense, *space and time* concepts have been revealed as fundamental parameters. We shall agree that it is not the same word that appears with constant presence but rather with intermittent presence, as it just appears in a certain moment of the text but with a high rate of repetition. We can visualize the distribution of a word throughout a text, by marking its position and frequency in a bubble-line. If we place in parallel the bubble-lines of the most frequent words, we can deduce in which of the three classical parts of the text (beginning, development and conclusion) a single word or a group of words appears. This is one of the tools which we are prouder of, because the visual impact of the concepts' organization gives a lot of information to the user in just a few seconds.

Finally, some of the symposium attendants inquired if the different writing styles of authors could affect the calculus of the similarity index between texts and asked how can computers detect the phenomena of synonyms or polysemic words. In order to solve this, we can store in our database lists of synonyms or polysemic words by establishing groups of words related in terms of significance. Usually, a language like Spanish, English or French has no more than 90.000 entries in a dictionary, so these lists represent a very short collection for a database. Obviously, we are not studying the writing style of the authors, but looking for similar concepts in the texts. By this reason, we do not care for the words themselves, but the significance of them.

Drawing the ArchiteXt Mining web site

The last task we were faced with during this year was the design of the web site where the tool is going to be hosted in. For the moment, we have just sketched the sections that it would be able to contain and which prior tools must be developed in order to implement it. Our aim is to create a simple interface, a Flat Web Design where the users can navigate and find the sections intuitively. Al least for now, it consists of three sections: Project, Results and Contact.

We'll start by the easiest. Obviously, the 'Contact' section allows for users the possibility of contact with the research team. More interestingly, the 'Project' section contains information about the objectives of the project, the research team, the partner institutions and the financing entities. Also, we pretend to include information and links to other research projects in Digital Humanities from within the Architecture field that is the focus of interest of the research team.

But, the main section is 'Results', which hosts a tool that allows for Boolean searches in the biblio-thematic database. A powerful search tool that will provide the title and bibliographic data of the articles that meet the criteria requested by users, who can search a certain word, a group of words or a syntagm that can be combined with operators

5. Yau, Nathan. Visualize this. The FlowingData Guide to Design, Visualization and Statistics. Indianapolis: Wiley, 2001.
Sinclair, Stéfan; Rockwell, Geoffrey. Voyant Tools. 2016. Available in: http://voyant-tools.org/

(or modifiers) as AND, NOT and OR to further produce more relevant results. The search will be done inside the biblio-thematic database as much as in the full text of the articles.

Once the articles have been selected, the web user will be able to list and print them or enter in each article to see more particular data. Also, we want to provide a bar graphs that show the word (or group of words) distribution over the years, the quantity of articles by year, and the position where a certain term appears. If the users access one particular article's page, besides the bibliographic data it can be obtained a word-cloud of the content and a bubble-line graph of the more frequent words in the article. The most remarkable search feature will be the similarity index with other articles stored in the database.

Besides the search tool, the 'Results' section will include a part devoted to show specific studies' development by the research team using the ArchiteXt Mining techniques. We intend to establish several patterns and differences not only between the magazines, but also between decades and between the topics of interest within the Spanish architecture or within foreign architecture. Another target of this project is to supply an objective list of texts that have set influential trends in Spanish architecture and those that, on the contrary, have been a mere reflection and continuation of the same. It is considered that this is a goal of great importance for the advancement of future research. On the other hand, we aim to establish rankings that indicate the importance of architects, buildings, critics and a considerable number of interests' variables for the researchers. This part will evolve in parallel to the research group access and use of the search engine.

However, right now it is possible to display other results that are been producing by the research team while the search tool is been refined, as articles, book chapters, reviews and so on. Among the most significant contributions are two book chapters of the Principal investigator directly related with the analysis and research in architectural periodicals, which are "*Arquitectura y lo demás*, Mexican periodical", included in Gutiérrez, Ramón; Gutiérrez Viñuales, Rodrigo (eds.). *Patrimonio y modernidad en Latinoamérica. Revistas de Arte y Arquitectura* (1940-1960). Bogota: Asociación de Amigos del Instituto Caro y Cuervo, 2017, and, much more related to this project, the book chapter entitled "Periodicals and the return to modernity after the Spanish Civil War", included in Peckham, Andrew; Schmiedeknecht, Torsten (eds.) *Modernism and the Professional Architecture Journal: Reporting, editing and reconstructing in post-war Europe*. London: Taylor&Francis/Routledge, 2018. Furthermore, the research team has produced a significant number of articles and congress papers. Also, three doctoral thesis of research group's member have been defended based on the architectural periodicals' research: *Mexico exports. Modern architecture in European and American periodicals (1950-1970)*, by Vanessa Nagel; *Exported Architectures. The diffusion of Spanish architectural production in the international panorama throughout the foreign periodicals (1949-1986)*, by Pablo Arza Garaloces, and *Ads&Arts&Architecture.* Arts&Architecture *advertising and the building of the South California architectures image (1938-1967)*, by Daniel Díez Martínez, who has just been award with the UPM PhD Special Prize. In 2018, we expect that two more doctoral thesis defenses from our research team. In any case, the contents of this section will expand throughout time.

Apparently, 2017 has been a year without significant results for the ArchiteXt Mining. However, it has been key to lay the foundations of the project's methodology. We still have a long way to go and we are almost sure that a one-year extension will be necessary. At present, we can affirm that we are ready to start the creation of the research tool and its testing phase. We do not yet know how long it should take. But, at least, at present we already know what we want and how to implement it.

Acknowledgments

'ArchiteXt Mining. Spanish modern architecture through its texts (1939-1975)' HAR2015-65412-P (MINECO/ERDF) is a research project funded by the Government of Spain through the 2015 Call for 'Excellence Projects' of the Ministry of Economy and Competitiveness (MINECO) and the European Regional Development Fund (ERDF).