ArchiteXt Mining: primjena tekstualne analitike na istraživanje moderne arhitekture

ArchiteXt Mining: Applying Text Analytics to Research on Modern Architecture

PRETHODNO PRIOPĆENJE
Predan: 2.6.2019.
Prihvaćen: 18.7.2019.
DOI: 10.31664/zu.2019.105.07
UDK: 050+72.036(460):[004.62

SAŽETAK
ArchiteXt Mining: Spanish modern architecture through its texts (*1939–1975*) istraživački je projekt koji je financirala Vlada Španjolske putem poziva za „projekte izvrsnosti" Ministarstva gospodarstva i konkurentnosti 2015. godine. Projekt ima za cilj istražiti novo gledište i razmotriti posebnosti moderne španjolske arhitekture. Unatoč sve većem uspjehu primjene analize podataka kao alata u nizu disciplina, istraživanja na području teorije arhitekture nikada se nisu najefikasnije koristila ovim tehnologijama. Španjolske i međunarodne okolnosti razvoja moderne arhitekture pomno su razmatrane kvalitativnim istraživanjem, koje je uspostavilo opće teorijske osnove. Sada je vrijeme za započinjanje novoga dubinskog istraživanja na temelju objektivnih podataka. Da bismo odgovorili na ovaj izazov, predlažemo primjenu tehnika „rudarenja teksta" (engl. *text mining*) kako bi se iskoristili najbolji izvori podataka na ovom području: arhitektonska periodika. Svrha je stvoriti snažnu bazu podataka koja će biti javno dostupna znanstvenoj zajednici na internetu. Dakle, ovaj projekt ispunjava nekoliko ciljeva e-istraživanja: olakšati informatizaciju istraživanja podataka, podržati svaku fazu prikupljanja podataka i upravljati analizama velikih podataka uz pomoć posebnih alata.

KLJUČNE RIJEČI
arhitektonska periodika, rudarenje teksta (engl. *text mining*), moderna španjolska arhitektura, analiza podataka

PRELIMINARY PAPER
Received: June 2, 2019
Accepted: August 20, 2019
DOI: 10.31664/zu.2019.105.07
UDC: 050+72.036(460):[004.62

ABSTRACT
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. This project aims to explore a new viewpoint and look into the special features of Spanish modern architecture. Despite the increasing success of using data analysis as a tool in a variety of disciplines, research on architectural theory has never made the most efficient use of these technologies. The Spanish and international circumstances of modern architecture development have been scrutinized through qualitative research, which has established a shared theoretical ground. It is now time to start a new in-depth research based on objective data. To address this challenge, we propose the application of text mining techniques to take advantage of the best data source in the field: architectural periodicals. The purpose is to create a powerful database hosted on a public website for the scientific community. Thus, this project fulfils several e-Research objectives: to facilitate the computerization of data research, to support every stage of data collection, and to manage big data analyses with the help of specific tools.

KEYWORDS
architectural periodicals, text mining, Spanish modern architecture, data analysis

Ana Esteban-Maluenda
Laura Sánchez Carrasco

Arhitektonski fakultet Politehničkog sveučilišta u Madridu /
School of Architecture, Technical University of Madrid

Luis San Pablo Moreno

Nezavisni istraživač, Madrid /
Independent researcher, Madrid

•

Razvoj informacijsko-komunikacijskih tehnologija tijekom druge polovice 20. stoljeća duboko je transformirao društvo i oblike stvaranja znanja. U 21. stoljeću teško je i razmišljati o znanstvenim istraživanjima ne uzimajući u obzir digitalni aspekt. U svojoj digitalnoj dimenziji svijet u kojem živimo sve se više oslanja na podatke, u kojima možemo pronaći vrijedne informacije, identificirati obrasce ponašanja i vizualizirati informacije u velikim razmjerima. Digitalni alati i tehnike česte uporabe, kao što su skladištenje ili rudarenje podataka (engl. *data mining*) uspješno se hvataju ukoštac s tom masovnom analizom podataka, stvarajući tako gotovo neograničene mogućnosti istraživanja u smislu masovne obrade informacija.

Text mining ili tekstualna analitika pojmovi su za procese koji iz teksta izvlače visokokvalitetne informacije. Visokokvalitetne informacije obično se dobivaju osmišljavanjem obrazaca i trendova kroz primjenu sredstava poput statističkog učenja obrazaca. Jedan od najfascinantnijih dijelova procesa jest mogućnost da se otkrije znanje koje se kao takvo ne pojavljuje ni u jednom od tekstova koji se pojedinačno provjeravaju, nego nastaje kada se više njih poveže u skup. Dakle, to je analiza koja otkriva neprikazane zajedničke podatke (ili nestrukturirane podatke) u nekoj zbirci i nudi zaključke koje nije moguće dobiti tradicionalnim metodama. Ukratko, *text mining* je primjena analitike kako bi se izvuklo znanje iz teksta, tako da bude moguće odgovoriti na neko prethodno formulirano pitanje (opisni modeli) ili otkriti skrivene obrasce u skupini tekstova (prediktivni modeli). Nadalje, tekstualna analitika može se primijeniti za identificiranje određenih svojstava ili elemenata u dokumentu[1] ili za njegovu klasifikaciju kao cjeline.[2]

Labavo povezan skup postupaka koji se naziva *text mining* ili „tekstualna analitika", zajedno s blisko srodnim „rudarenjem podataka" (*data mining*), omogućili su istraživačima da pristupe tekstovima na nove načine.[3] Te vrste analitičkih procesa uvelike se primjenjuju u raznim znanstvenim i humanističkim disciplinama, s pozitivnim rezultatima. Prisjetimo se, na primjer, isusovačkog svećenika Roberta Buse, koji je 1946. počeo graditi *Index Thomisticus*, alat za pretraživanje teksta u korpusu djela Tome Akvinskog. Na početku svojeg projekta Busa se koristio prilično rudimentarnim alatima za analizu. Međutim, zahvaljujući znatnom razvoju koji su ove tehnike doživjele posljednjih desetljeća, najbolji rezultati ovog projekta postignuti su krajem 20. stoljeća.[4]

U arhitekturi je također bilo važnih pionira na području digitalne analize, kao što je Juan Pablo Bonta. Njegova knjiga *American Architects and Texts*[5] objavljena je 1996. kao rezultat dugotrajnog projekta koji je u početku financirala Zaklada Graham, a kasnije Sveučilište u Marylandu. Bonta je u svojem istraživanju radio s podacima navedenima u 380 tekstova o američkoj arhitekturi nakon 1815. Nagla smrt argentinskog arhitekta nakon objavljivanja knjige prekinula je njegovo golemo djelo koje su, međutim, danas očigledno nadmašili vodeći alati kao što je preglednik Google knjiga Ngram. Ovaj korisni alat omogućava nam da u nekoliko

1
Tseng *et al.*, „Mining Concept Maps from News Stories".
2
Bichindaritz, Akkineni, „Concept Mining for Indexing Medical Literature".
3
Higgins, „Reading and Non-Reading", 85.
4
Alarcón, Bernot, „Index Thomisticus by Roberto Busa and Associates".
5
Bonta, *American Architects and Texts*.

•

The development of information and communication technologies throughout the second half of the 20th century has deeply transformed the society and the forms of knowledge generation. In the twenty-first century, it is scarcely possible to think about scientific research without considering the digital aspect. In its digital dimension, the world that we live in increasingly relies on data, in which we can find valuable information, identify patterns of behaviour, and visualize information on a big scale. Digital tools and techniques of frequent use, such as data warehousing or data mining, have managed to address this massive data analysis and have thereby generated almost unlimited possibilities for research in terms of massive treatment of information.

Text mining or text analytics are processes that extract high-quality information from texts. High-quality information is typically derived through devising patterns and trends by using means such as statistical pattern learning. One of the most fascinating parts of the process is that it is able to discover knowledge that does not appear as such in any of the texts checked individually, but arises when a set of them are related. Thus, it is an analysis that detects the non-shown common data (or unstructured data) in a collection and offers conclusions that are not obtainable by traditional methods. In short, text mining is an analysis exercise used to extract knowledge from a text, so that it is possible to answer a previously formulated question (descriptive models) or to discover hidden patterns in a group of texts (predictive models). Furthermore, text analytics can be used to identify specific properties or elements in a document[1] or to classify it as an entity.[2]

The loose set of practices called "text mining" or "text analytics," along with their close cousin "data mining," have allowed researchers to approach texts in new ways.[3] These kinds of analytic processes have been largely used in a variety of scientific and scholarly disciplines, with positive results. For example, let us recall the Jesuit priest Roberto Busa, who in 1946 began to build the Index Thomisticus, a tool for performing text searches within the corpus of Aquinas' works. At the beginning of his project, Busa used fairly rudimentary analysis tools. However, thanks to the major development that these techniques have experienced in the last decades, the best results of this project were obtained at the end of the 20th century.[4]

In architecture, there have also been significant pioneers in digital analysis, such as Juan Pablo Bonta. He wrote the book *American Architects and Texts*,[5] published in 1996 as result of a long project, funded initially by the Graham Foundation and later by the University of Maryland. In his research, Bonta worked with data cited in 380 texts about American architecture since 1815. The sudden death of the Argentinian architect after the publication of this volume put an end to his vast work that, however, today is clearly surpassed by mainstream tools such as the Ngram Viewer of Google Books. This useful tool allows us to

1
Tseng *et al*.,"Mining Concept Maps from News Stories."
2
Bichindaritz, Akkineni, "Concept Mining for Indexing Medical Literature."
3
Higgins, "Reading and Non-Reading," 85.
4
Alarcón, Bernot, "Index Thomisticus by Roberto Busa and Associates."
5
Bonta, *American Architects and Texts*.

sekundi vizualiziramo ogromne količine stavki koje se odnose na bilo kojeg autora uključenog u knjige napisane tijekom 19. i 20. stoljeća. Uz iznimke, te se referencije s vremenom povećavaju jer je izdavačko tržište danas znatno šire nego što je bilo nekada.

Dakle, ako rast izdavačke industrije „kontaminira" kvantifikaciju referencija i čini takvu analizu nevrijedećom, gdje možemo dobiti informacije o evoluciji moderne arhitekture bez umnožavanja tih pogrešaka? Je li se moguće koristiti bazama podataka koje nam pružaju informacije o modernoj arhitekturi u određenom trenutku? Odgovor je da.

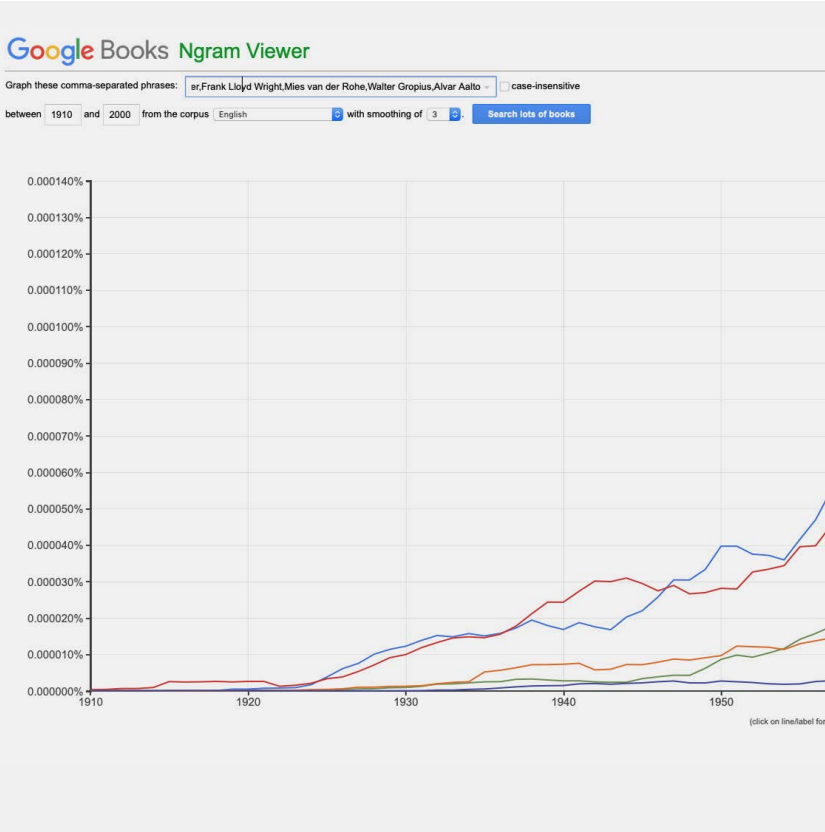
Arhitektura 20. i 21. stoljeća ima izvanrednu bazu podataka u kojoj su registrirani najvažniji pojmovi, događaji i zgrade, a to je arhitektonska periodika. Većina istraživača specijaliziranih za arhitekturu upotrebljava ih kao izvor. Nažalost, taj posao još uvijek se radi na isti način kao što se radio u posljednjih pedeset godina, što znači da još uvijek moramo ići u knjižnice i pregledavati sva izdanja stranicu po stranicu.

Posljednjih desetljeća izrađeni su brojni indeksi arhitektonske periodike kako bi se pomoglo istraživačima. No ti su indeksi nepotpuni i rijetko sadrže zapise o kraćim tekstovima (npr. vijestima). Velika količina informacija u časopisima ometa razumijevanje istraživača. Računalna podrška presudna je za pretvorbu ove velike baze podataka u čitljiv format koji se može lako analizirati.

SVRHA: PROJEKT ARCHITEXT MINING

ArchiteXt Mining (što je akronim za *architectural text mining*) predlaže upotrebu naprednih tehnika u analizi podataka, izrađujući alate za istraživače koji za svoj rad upotrebljavaju periodiku. Mogućnosti koje danas nudi računalni inženjering omogućuju nam da napravimo nešto što prije nije bilo moguće: da izvršimo globalnu analizu cjelokupnog sadržaja časopisa. Također, ArchiteXt Mining zamišljen je kao kolaborativni alat koji znanstvenoj zajednici pruža informacije i istodobno ih prima od svojih korisnika i istraživača.

Nipošto ne želimo sugerirati da takav alat može zamijeniti rad istraživača; umjesto toga, namjera nam je pomoći da lakše razumiju određeni tekst ili, što je mnogo složenije, velik broj tekstova paralelno. Vjerujemo da rad s više članaka istodobno eksponencijalno povećava potencijal alata i da može poslužiti kao izvor inspiracije. Pokazalo se da rezultati kvantitativne analize mogu stvoriti obrasce koji se ne mogu otkriti dok čitate članke jedan po jedan.[6] *Text mining* pomaže u asimiliranju tekstova, pojedinačno ili zajedno, što može unijeti neizvjesnost i zbrku u usporedbi s njihovom interpretacijom s pomoću programskih sredstava.[7] Praksa pažljivog čitanja ne može se zamijeniti algoritamskim sredstvima. Analitičke metode koje predlažemo samo otkrivaju pojave, poput veće ili manje pozornosti koja se obraćala na neku temu ili ideju tijekom desetljeća, ali ne mogu objasniti razloge tih promjena, jer to zahtijeva razumijevanje stručnjaka.

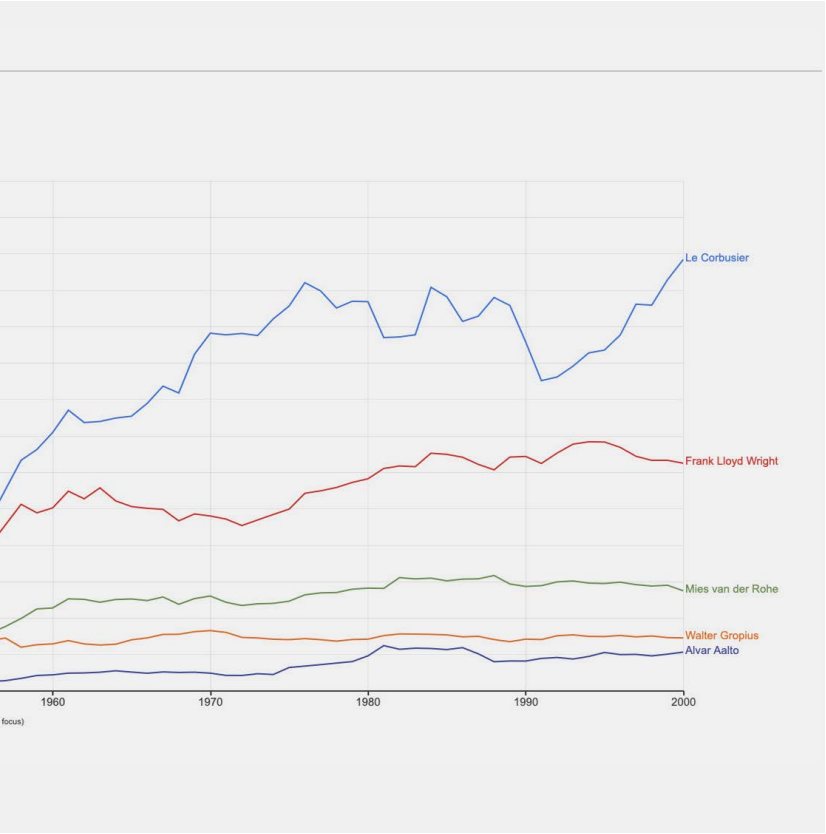


Slika 1. Ngram Viewer. Broj navoda između 1910. i 2000.: Le Corbusier, Frank Lloyd Wright, Mies van der Rohe, Walter Gropius i Alvar Aalto. (Izvor: https://books.google.com/ngrams) / Figure 1. Ngram Viewer. Number of cites between 1910 and 2000: Le Corbusier, Frank Lloyd Wright, Mies van der Rohe, Walter Gropius and Alvar Aalto. (Source: https://books.google.com/ngrams)

↑

6 Higgins, „Reading and Non-Reading".

7 Hotho, Nürnberger, Paaß, „A Brief Survey of Text Mining".



visualize in a few seconds huge amounts of items referring to any author included in books written during the 19th and 20th centuries. With exceptions, the references are growing with time, since nowadays the publishing market is much more extended than it used to be in the past.

So, if the growth of the publishing industry "contaminates" the quantification of references and makes this kind of analysis invalid, where can we obtain information about the evolution of modern architecture without multiplying these mistakes? Is it possible to use databases that can provide information about modern architecture at a certain point in time? The answer is yes.

Architecture of the 20th and 21st centuries has an extraordinary database where the most significant concepts, events, and buildings have been registered: architectural periodicals. Most researchers specialized in architecture use them as a source. Unfortunately, we are still doing this work the same way it has been done for the last fifty years, meaning that we still need to go to libraries and review all the issues page by page.

In the last decades, many architectural periodicals indexes have been built in order to help the researchers. But those indexes are incomplete and rarely include records about smaller texts (i.e. news sections). The big quantity of information in the periodicals hinders the researchers' understanding. Computer support is critical in transforming this large database into a readable format that can be easily analysed.

THE AIM: ARCHITEXT MINING PROJECT

ArchiteXt Mining (which is the acronym for Architectural Text Mining) proposes the use of advanced techniques in data analysis, building tools for researchers that use periodicals for their work. The possibilities offered today by computer engineering enable us to perform something previously impossible: to conduct a global analysis of the complete contents of the periodicals. Also, ArchiteXt Mining is conceived as a collaborative tool providing information to the scientific community while at the same time receiving it from its users and researchers.

We are certainly not trying to suggest that such a tool can replace the researchers' labour; instead, it is intended to help them understand more easily a particular text or, what is much more complicated, a large number of texts in parallel. We believe that working with many articles at the same time exponentially increases the tool potential and might even serve as a source of inspiration. It is demonstrated that the results of quantitative analysis can provide patterns that are not detectable while simply reading the articles independently.[6] Text mining helps to assimilate texts, individually or collectively, which may bring uncertainty and confusion as compared to their interpretation

6
Higgins, "Reading and Non-Reading."

Drugi aspekt koji valja istaknuti jest da je aktualni ArchiteXt Mining pilot-projekt, koji je nastao u Španjolskoj i usredotočen je na španjolsku arhitektonsku periodiku, ali s težnjom da preraste u nešto veće. U samo dvije godine testirane su različite analitičke metode, što nam pruža osnovu i dovoljno iskustva da se suočimo s novim i daleko ambicioznijim izazovima. Zahvaljujući projektu HAR2015–65412–P/MINECO–ERDF uspjeli smo uroniti u problematiku, demonstrirajući održivost projekta i uspostavljajući neke osnove od kojih možemo krenuti i predložiti uistinu inovativan projekt. Osim toga, ArchiteXt Mining ima ogroman kapacitet za rast, koji će stoga dovesti do eksponencijalnog povećanja njegova potencijala, dodatno potaknutog primjenom rečenih tehnologija na većoj količini podataka.

Kao prvi korak, ArchiteXt Mining usredotočio se na periodiku koja je objavljena za Francove diktature (1939.–1975.). Španjolska arhitektura i njezini mediji tijekom tih godina dobro su poznato polje istraživačkom timu. Stoga evolucija i promjene koje su se dogodile u arhitekturi tijekom desetljeća u kojima je Franco vladao zemljom pružaju ogromne mogućnosti u pogledu uočavanja suprotnosti između najranijih i kasnijih faza tog razdoblja.

Što se tiče periodike, već smo digitalizirali časopise madridskog Instituta za arhitekte (*Revista Nacional de Arquitectura* i *Arquitectura*) i barcelonskog Instituta za arhitekturu (*Cuadernos de Arquitectura*). U početku smo namjeravali dopuniti ovu građu drugim važnim španjolskim časopisima, kao što su *Hogar y Arquitectura* i *Nueva Forma*. Cilj projekta bio je uključiti i neke europske časopise, primjerice *L'Architecture d'Aujourd'hui* (Francuska), *The Architectural Review, Architectural Design* (Velika Britanija) te *Domus* i *Casabella* (Italija). No zbog smanjenog budžeta projekta—u usporedbi s onim za koji smo aplicirali—ovi europski izvori nisu razmatrani u prvoj fazi. Slijedom proračunskih smjernica usredotočili smo se samo na časopise koje objavljuju arhitektonski instituti u Madridu i Barceloni. Unatoč ovom nužnom i drastičnom smanjenju, vjerujemo da su nam ovi resursi omogućili da pokrijemo gotovo cijelu arhitektonsku panoramu Španjolske u danom razdoblju.

Na temelju navedenih časopisa razvijena je baza podataka objavljenih u španjolskim medijima, koja je stavljena na mrežnu stranicu otvorenog pristupa.[8] Ova bibliografsko-tematska baza podataka slijedi početnu klasifikaciju prema bibliografskim kriterijima, koju su ručno napravili članovi istraživačke skupine. Međutim, pruža daleko više informacija koje nisu uključene u uobičajene internetske indekse. Uz bibliografske podatke o tekstu (autor, naslov, časopis, broj, godina i stranice) bilježimo i druge važne podatke, kao što su vrsta teksta (članak, ogled, vijest), kratak opis teme, specifikacija dijela časopisa u kojem se nalazi tekst, podaci o građevini, osobi ili događaju o kojima je riječ i slično. Pristup ovim informacijama već je znatan korak naprijed, budući da daje istraživačima moćan alat za preliminarnu kvantitativnu analizu i pretraživanje koji će im pomoći da započnu istraživanje.

through the more programmatic means.[7] The practice of careful reading cannot be replaced by algorithmic means. The methods of analysis that we propose only detects issues such as the greater or lesser attention to a topic or an idea over the decades, but cannot explain the reasons behind those changes, which requires the experts' understanding.

Another aspect to highlight is that the current ArchiteXt Mining is a pilot project, which was born in Spain with a focus on Spanish architectural periodicals, but with the aspiration to grow into something larger. In just two years, different analysis methodologies have been tested, which gives us a background and sufficient experience to face new and much more ambitious challenges. Thanks to the project HAR2015–65412–P/MINECO–ERDF, we have been able to immerse ourselves in the matter, demonstrating the project's viability and establishing some bases from which to set out and propose a truly innovative project. In addition, ArchiteXt Mining has a huge capacity for volume growth that, as a consequence, will cause an exponential increase in its potential, boosted in turn as these technologies are applied to a greater amount of data.

As a first step, ArchiteXt Mining has been focusing on periodicals published during Franco's dictatorship (1939–1975). Spanish architecture and its media during those years are a well-known field for the research team. Therefore, the evolution and changes that occurred in architecture throughout the decades when Franco ruled the country provide immense possibilities in terms of attending to the contrasts between the earliest and later phases of that period. Regarding the periodicals, we have already digitized the journals of the Madrid Institute of Architects (*Revista Nacional de Arquitectura* and *Arquitectura*) and the Barcelona Institute of Architecture (*Cuadernos de Arquitectura*). We initially proposed to complement this material with other important Spanish periodicals, such as *Hogar y Arquitectura* and *Nueva Forma*. The aim of the project was also to include some European periodicals: *L'Architecture d'Aujourd'hui* (France), *The Architectural Review*, *Architectural Design* (Great Britain), *Domus* and *Casabella* (Italy). But due to the reduced budget of the project—compared to what we had applied for—these European sources have not been considered in this first phase. Following the budgetary guidelines, we have only focused on journals published by the architectural institutes of Madrid and Barcelona. Despite this compulsory and severe reduction, we believe that these resources have allowed us to cover a reasonably complete architectural panorama of Spain during the given period.

8
Esteban-Maluenda, San Pablo, „ArchiteXt Mining".

Već samo stvaranje ove baze podataka opravdalo je projekt, no želimo ostvariti i dodatnu vrijednost. Tu ulaze u igru tehnike rudarenja teksta, kako bi se primijenile različite statističke tehnike i dobile dodatne vrijednosti izvlačenjem različitih vrsta informacija iz tekstova pohranjenih u našoj bazi podataka.

ISTRAŽIVANJE: METODOLOGIJA PROJEKTA ARCHITEXT MINING

Prijevod tiskanih tekstova u strojno čitljiv jezik zahtijeva nekoliko koraka. Iako je u ovom trenutku količina digitaliziranih informacija dostupnih na mreži ogromna, to se ne odnosi na španjolske časopise o arhitekturi iz razdoblja nakon Građanskog rata. Kada je projekt započeo, bila je dostupna samo digitalizirana verzija časopisa *Cuadernos de Arquitectura*, koja k tome nije sadržavala neke relevantne rubrike poput kratkih vijesti. Dakle, prva stvar kojoj smo se posvetili bio je dugotrajni proces digitalizacije.

Za visokokvalitetne skenove časopisa potrebni su posebni skeneri. U tom pogledu projekt se oslonio na knjižnicu Arhitektonskog fakulteta Politehničkog sveučilišta u Madridu (ETSA-UPM). Srećom, ta knjižnica posjeduje sva izdanja glavnih španjolskih arhitektonskih časopisa. Nadalje, omogućili su nam upotrebu profesionalnog skenera za digitalizaciju izdanja, Metis EDS Alpha, koji se odlikuje slikama vrhunske kvalitete, jednostavnom upotrebom i visokom produktivnošću (oko 30 potpunih skenova u minuti). To nam je omogućilo da dobijemo kompletne skenirane zbirke sljedećih triju časopisa: *Revista Nacional de Arquitectura* (od 1941. do 1958., brojevi 1–204), *Arquitectura* (od 1959. do 1975., brojevi 1–197) i *Cuadernos de Arquitectura* (od 1944. do 1975., brojevi 1–111). Kada smo završili sa skeniranjem, pristupili smo optičkom prepoznavanju znakova (OCR) svakog članka s pomoću softvera Abby Finereader 14. Nakon što smo dobili zbirku u elektroničkom obliku (txt, doc, rtf...), mogli smo započeti s informatizacijom nestrukturiranih podataka, što je pravi cilj projekta.

7
Hotho, Nürnberger, Paaß, "A Brief Survey of Text Mining."
8
Esteban-Maluenda, San Pablo, "ArchiteXt Mining."

Based on these journals, a database of information published in the Spanish media has been developed and hosted on an open access website.[8] This biblio-thematic database follows an initial classification according to the bibliographic criteria, done manually by the members of the research group. However, it provides far more information that is not included in habitual online indexes. In addition to the bibliographic data on the text (author, title, journal, issue, year, and pages), we have been recording other important data in terms of text type (article, review, news), a brief description of the topic, specification of the section where the text is located, data about the building, personality, or event that it concerns, and so on. Getting access to this information is already a big step forward and provides the researchers with a powerful tool to engage in a first quantitative analysis and search that will help them to begin their studies.

The very creation of this database has justified the project, but we aim to bring in added value. It is at this point that the text mining techniques come into play, namely in order to apply different statistical techniques and to obtain additional values by extracting a different kind of information from the texts stored in our database.

THE RESEARCH: ARCHITEXT MINING METHODOLOGIES

The translation of printed texts into a machine-readable language requires several steps. Although at this moment the amount of digitized information available online is overwhelming, this does not apply to the Spanish architecture journals from the period after the Civil War. When the project started, the digitalized version of *Cuadernos de Arquitectura* was uniquely available, and it didn't contain some of the relevant sections, such as short news. So the first thing we tackled was the long process of digitalization.

In order to obtain high-quality scans of periodicals and magazines, specific scanners were required. In this regard, the project relied on the support of the Library of the School of Architecture at the Universidad Politécnica de Madrid (ETSA–UPM). Fortunately, the ETSA–UPM Library maintains the collection of all the issues of the main Spanish architectural periodicals. Furthermore, they allowed us to use their professional scanner to digitalize the issues: a Metis EDS Alpha, which provided us with superior quality images, simple usage, and high productivity (about 30 full scans per minute). This allowed us to obtain complete 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). Once all the scans had been made, we proceeded to the optical character recognition (OCR) of every article by using the Abby Finereader 14 software. When we had the collection in an electronic format (txt, doc, rtf...), we were able to begin the computerization of the unstructured data, the real goal of the project.

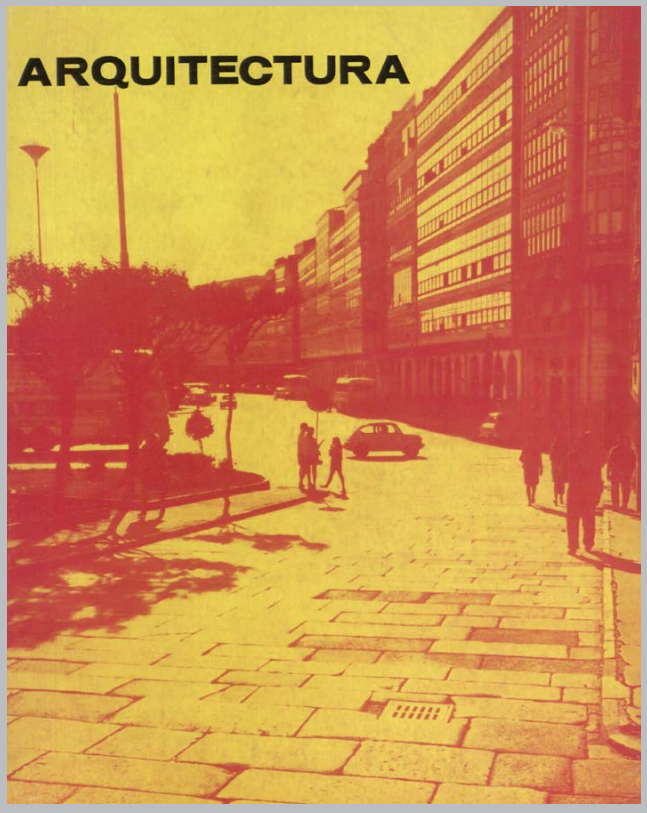
Slika 1. S lijeva na desno, naslovnice časopisa Instituta za arhitekturu u Madridu (*Revista Nacional de Arquitectura* and *Arquitectura*) i Barceloni (*Cuadernos de Arquitectura*). (Izvori: http://www.coam.org/es/fundacion/biblioteca/revista-arquitectura-100-anios; https://www.raco.cat/index.php/CuadernosArquitectura) / Figure 2. Some covers of the journals of the Institute of Architects of Madrid (*Revista Nacional de Arquitectura* and *Arquitectura*) and Barcelona (*Cuadernos de Arquitectura*). (Sources: http://www.coam.org/es/fundacion/biblioteca/revista-arquitectura-100-anios; https://www.raco.cat/index.php/CuadernosArquitectura)

↑ →


REVISTA NACIONAL DE ARQUITECTURA
Año XI
Núm 119
Noviembre 1951
ORGANO OFICIAL DEL CONSEJO SUPERIOR DE COLEGIOS DE ARQUITECTOS DE ESPAÑA
EDITADO POR EL COLEGIO OFICIAL DE ARQUITECTOS DE MADRID
226
183
86
70
43
27
Vaquero Turcios


APUNTES DE
milán
cuadernos de arquitectura
Publicación del Colegio Oficial de Arquitectos de Cataluña y Baleares · Núm. 67
1967


cuadernos de
arquitectura
nº 22
Publicación del Colegio Oficial de Arquitectos de Cataluña y Baleares


ARQUITECTURA

Prvi korak bio je dobivanje takozvane DNK različitih tekstova kako bi se s njima moglo automatski raditi. Text Matrix Document (TMD, DNK teksta) jest bit teksta, matrica koja analizira prisutnost i raspodjelu riječi u tekstu. U tu svrhu bilo je potrebno ukloniti riječi koje same po sebi ne pridonose tekstu, takozvane „stop riječi": članove, prijedloge, veznike i riječi koje su irelevantne za semantičku analizu. Ove su riječi vrlo važne za sintaksu jer povezuju rečenice, odlomke i drugo, ali njihov utjecaj na analizu značenja teksta vrlo je ograničen.

Naša osnovna jedinica analize bio je članak u časopisu, shvaćen kao tekst koji se može povezati s određenim naslovom. Pritom smo smatrali da su dugačak tekst, kratka vijest i ogled knjige slični u pogledu dobivanja TMD-a.[9] Kada ove podatke postavimo u odnos s ostalim metapodacima, možemo provoditi pretraživanja, usporedbe i druge vrste analiza podataka na velikim količinama tekstova, što bi bila nemoguća zadaća za ljudski mozak.

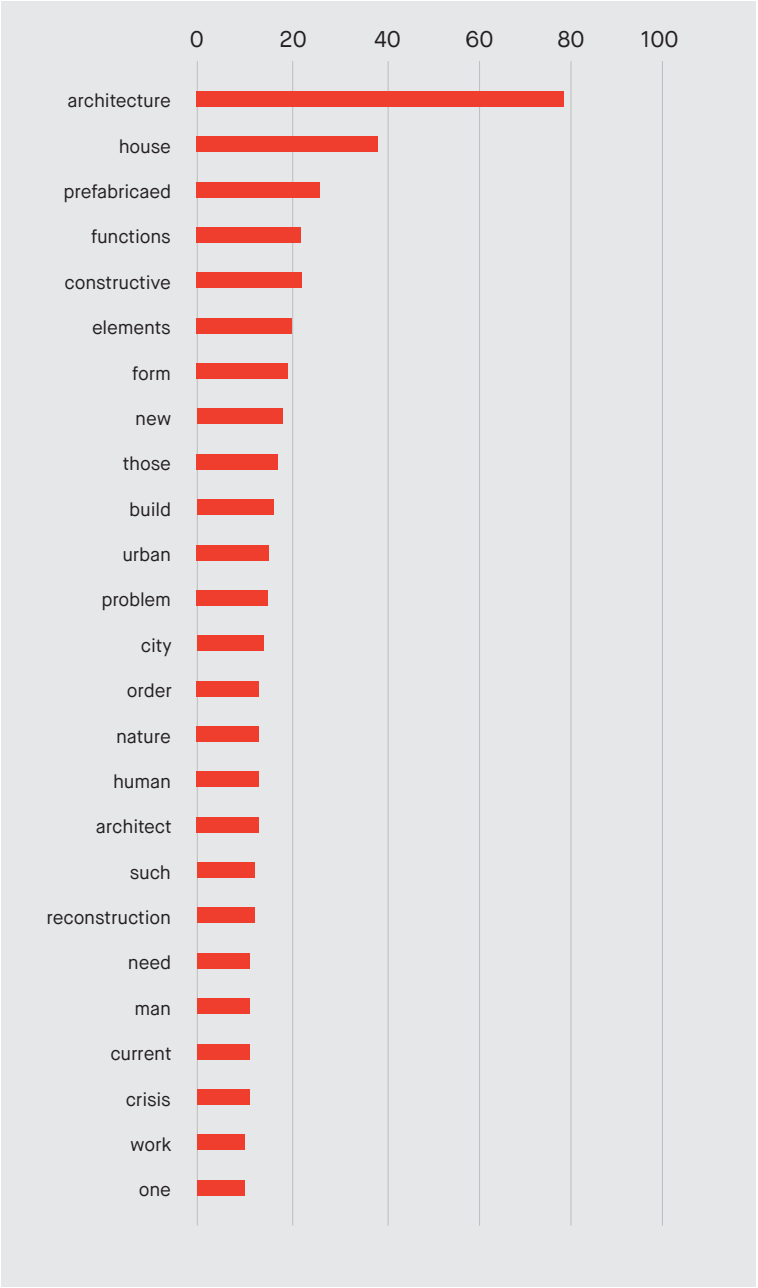
Digitalni tekstovi omogućuju da jezični sadržaji postanu sličniji „podacima", odnosno nešto što se može računalno obraditi kako bi se otkrila sirova svojstva teksta te kako bi se njima zatim moglo analitički i kreativno manipulirati.[10]

Na primjer, TMD nam omogućuje izračunavanje rangiranja učestalosti riječi. Ovaj postupak ukazuje na najčešće izraze koji se pojavljuju u tekstu i stoga nudi određeni ključ za njegovu temu.

Međutim, rangiranje učestalosti ne nudi samo sveobuhvatan prikaz sadržaja teksta. Može se pohraniti u našu bazu podataka i upotrebljavati za automatsku obradu informacija, poput ispitivanja sličnosti između tekstova. Nakon što smo dobili DNK dvaju tekstova, možemo ih usporediti i izračunati njihov indeks sličnosti (SIM). Ovaj postotak nudi nam objektivne kriterije za traženje niza sličnih tekstova prije nego što ih krenemo čitati. To nije egzaktna znanost i može dovesti do malih pogrešaka. Još uvijek je potrebno da istraživač naknadno analizira rezultate i pročisti popis sličnih članaka. U svakom slučaju, pomaže nam da radikalno smanjimo opseg pretraživanja.

Budući da je nemoguće pročitati sve, pa čak i značajniji podskup svega, bilo bi korisno imati način da pristupimo i onome što nismo pročitali ili da barem zaključimo što bi ondje moglo stajati. Alati za *text mining* nude istraživaču tekstova nadljudsku sposobnost da se na određenoj razini bavi tisućama tekstova odjednom i izvuče podatke koji bi inače zahtijevali cijeli život, da i ne spominjemo nevjerojatno veliku mogućnost pamćenja i kritičku oštrinu.[11]

Detaljno razmotrivši različite metode koje nam nudi statistika, testirali smo tri različite tehnike izračunavanja SIM-a, koje imaju jedinstveno polazište: TMD. Kombinirajući ih dvije po dvije, možemo izračunati neke pokazatelje koji su već definirani u statističkoj literaturi.[12] Prva je metoda zbrajanje proizvoda ponderiranog opsega frekventnosti, što je neka vrsta zbroja riječi koje dijele dva teksta [d_1 i d_2]. Sve su ponderirane ukupnim brojem riječi.



Grafikon 1. Grafikon frekvencija najučestalijih riječi u tekstu Alberta Sartorisa "Orientations in contemporary architecture". (Izvor: pripremili autori teksta) / Chart 1. Frequencies graph of the most frequent words included in the text "Orientations in contemporary architecture" by Alberto Sartoris. (Source: prepared by the authors)
↑

9
Kasnije ćemo pokazati da dužina teksta ipak utječe na rezultate nekih analiza.
10
Higgins, „Reading and Non-Reading", 86.
11
Isto, 87.
12
Jedan je od „klasika" literature o *text miningu* Feldman, Sanger, *The Text Mining Handbook*.

The first step was to obtain the so-called DNA of different texts so as to automatically work with them. A Text Matrix Document (TMD, the text's DNA) is the text essence, a matrix that analyses the presence and distribution of words throughout a text. For this purpose, it was necessary to remove the words that did not contribute to the text in themselves, the so-called stop-words, which include articles, prepositions, conjunctions, and words that are irrelevant in semantic analysis. These words are very important for the syntax because they connect phrases, paragraphs, etcetera, but their impact on the analysis of the texts' meaning is very limited.

Our elementary unit of analysis was the magazine article, understood as a text that can be associated to a particular title. In that sense, we considered that a long text, a piece of short news, or a book review were similar in terms of obtaining the TMD.[9] Having this information in relation to the rest of the metadata, we can conduct searches, comparisons, and other types of data analysis on large amounts of texts, which would be an impossible task for the human brain.

Digital texts allow language content to become more like "data," that is, something that can be processed computationally to reveal anew the brute properties of the text, and then manipulated in analytical and creative ways.[10]

For instance, the TMD allows us to calculate the ranking of a word's frequency. This process yields the most frequent terms that appear in a text and therefore offers some clues about its subject matter.

However, the ranking of frequencies does not only procure a comprehensive view of the content of the text. It can be stored in our database and used for the automatic treatment of information, such as similarity studies between texts. Once the DNA of two texts is obtained, we can compare them with each other and calculate their similarity index (SIM). This percentage offers us objective criteria to look for series of similar texts before reading them. This is not exact science and it can lead to small mistakes. It is still necessary that the researcher should subsequently analyse the results and refine the list of similar articles. In any case, it helps us to reduce our search scope radically.

Since it is impossible to read everything, or even a significant subset of everything, it would be helpful to have a means of accessing and, one hopes, making out the meaning of that which remains unread. The text mining toolbox offers the scholar of texts the super-human capacity to engage on some level with thousands of texts at once, and to draw out information that would otherwise require a lifetime of work, not to mention the amazingly prodigious powers of memory and critical acumen.[11]

9
It will be shown below that the length of texts does influence the results of some of the analyses.
10
Higgins, "Reading and Non-Reading," 86.
11
Ibid., 87.

$$SIM\ (d_1, d_2) =$$
$$\Sigma\ x_i y_i = x_1y_1 + x_2y_2 + x_3y_3 + \ldots\ldots + x_ny_n$$

U navedenoj formuli x_i i y_i su ponderirane frekventnosti i riječi $[w_i]$ u dokumentima d_1 i d_2. Ponderirana učestalost riječi $[w_i]$ u jednom tekstu dobiva se kvocijentom između frekventnosti te riječi $[c\ (w_i, d_1)]$ i ukupnog broja riječi u dokumentu $|d_1|$.

$$x_i = c\ (w_i, d_1)/|d_1|$$
$$y_i = c\ (w_i, d_2)/|d_2|$$

Druga tehnika koju smo testirali bila je izračunati rezultat zbrajanja ponderiranog područja frekventnosti. Ustvari su oba pokazatelja nadahnuta sličnim statističkim konceptima i upotrebljavaju ponderirane frekventnosti. Najčešće riječi imaju veću težinu u izračunu.

$$SIM\ (d_1, d_2) = \Sigma x_i\ \Sigma y_i =$$
$$(x_1 + x_2 + x_3 + \ldots + x_n) \times (y_1 + y_2 + y_3 + \ldots + y_n)$$

Suprotno tome, Jaccardov indeks radi s jednostavnom pojavom riječi, bez obzira na to koliko se puta pojavljuje u tekstu. Ovaj indeks pokazuje kardinalnost sjecišta dvaju tekstova (d_1 i d_2) podijeljenu s kardinalnošću njihove zajednice. U ovom slučaju važna je učestalost pojedinih izraza, a ne njihova prisutnost u apsolutnom omjeru.

$$SIM\ (d_1, d_2) = |d_1 \cap d_2|\ /\ |d_1 \cup d_2|$$

Nakon izračunavanja ovih triju indeksa na reprezentativnom uzorku tekstova zaključili smo da imaju tendenciju održavanja slične vrijednosti. Ipak, nakon što smo sravnili rezultate, Jaccardov indeks učinio nam se optimalnim iz nekoliko razloga. Kao prvo, obično ostaje u sredini triju indeksa u smislu vrijednosti. Drugi je argument da Jaccardov indeks upotrebljava samo pojmove zajednice i sjecišta skupova riječi, što je razumljivo čak i početnicima u statistici. Naposljetku, ovaj indeks ne preferira riječi visoke frekventnosti u odnosu na one s niskom. Vrijednosti dobivene ovom metodom analizirale su dvije stručnjakinje za obradu podataka na projektu, koje su uspostavile sljedeći kriterij: više od 20 do 25 posto sličnosti između dvaju tekstova znači da se bave sličnim temama.

Kako bismo dodatno poboljšali rezultate indeksa sličnosti, analizirali smo učestalost sinonima, polisemičnih riječi i rječničkih jedinica u tekstovima. Mogu li stilovi pisanja različitih autora utjecati na indeks sličnosti između tekstova? Kako bismo to sveli na najmanju moguću mjeru, u svoju bazu podataka možemo pohraniti popis sinonima ili polisemičnih riječi uspostavljanjem skupina riječi koje su povezane po značenju. Jezici poput španjolskog, engleskog ili francuskog obično sadrže više od 90 000 rječničkih jedinica pa ovi popisi predstavljaju vrlo kratku zbirku za bazu podataka. U svakom slučaju, ovaj način poboljšanja još ispitujemo.

Regarding in depth the different methods that statistics provides us with, we have tested three different techniques to calculate the SIMs, which, however, share a unique starting point: the TMDs. Overlapping them two by two, we can calculate some indicators already defined in statistical literature.[12]

The first method is the summation of the products of the weighted frequency range, which is a sort of sum of the words that are shared by two texts $[d_1$ and $d_2]$. All of them are weighted by the total number of words.

$$SIM\ (d_1, d_2) = \Sigma\ x_iy_i = x_1y_1 + x_2y_2 + x_3y_3 + \ldots\ldots + x_ny_n$$

In the previous formula, x_i and yi are the weighted frequencies of the i-word $[w_i]$ in the documents dı and d_2, respectively. The weighted frequency of a word $[w_i]$ in one text is obtained by the quotient between the frequency of that word $[c\ (w_i, d_1)]$ and the total number of words in the document $|d_i|$.

$$x_i = c\ (w_i, d_1)/|d_1|$$
$$y_i = c\ (w_i, d_2)/|d_2|$$

The second tested technique is to calculate the product of summations of the weighted frequency range. In fact, both indicators are inspired by similar statistical concepts and use weighted frequencies. The most frequent words have more weight in the calculation.

$$SIM\ (d_1, d_2) = \Sigma x_i\ \Sigma y_i =$$
$$(x_1 + x_2 + x_3 + \ldots + x_n) \times (y_1 + y_2 + y_3 + \ldots + y_n)$$

By contrast, the Jaccard Index works with the simple appearance of words, without considering how many times the word appears in the text. This index shows the cardinality of the intersection of both texts (d_1 and d_2) divided by the cardinality of their union. In this case, what matters is the frequency of certain terms, not their presence in an absolute ratio.

Ovdje treba napomenuti da ne želimo sugerirati da bi primjena ovih tehnika kvantitativne analize trebala zamijeniti kasniju kvalitativnu analizu. Iako ovaj projekt pokriva relativno malen vremenski raspon i stoga značenje pojmova i riječi ostaje manje ili više stalno, u sljedećim fazama bit će potrebno uzeti u obzir da su riječi i pojmovi živa konstrukcija, koja se s vremenom mijenja. TMD ne sadrži druge podatke osim učestalosti riječi i njihova položaja u tekstu, pa nakon što se pojmovi otkriju, potrebno je izvršiti kvalitativni pregled, čemu mogu pripomoći i druge tehnike rudarenja teksta, poput analize korelacije riječi, koja ulazi u trag odnosima riječi u tekstu te stoga pridaje veću važnost značenju nego pukoj učestalosti ili mjestu.

→

$$SIM\ (d_1, d_2) = |d_1 \cap d_2| / |\ d_1 \cup d_2|$$

After calculating these three indexes in a representative sample of texts, we concluded that they tended to maintain a similar value. Nevertheless, balancing the results, the Jaccard Index looks like the optimal index for several reasons. First of all, it usually remains in the middle of the three indexes in terms of values. The second argument is that the Jaccard only uses concepts of union and intersection of word sets, which is more understandable even for beginners in statistics. Finally, this index does not beneficiate high words frequencies' as opposed to those with low frequencies. Values obtained with this method have been analysed by two data scientists on the project, who have established a criterion: beyond 20–25 per cent of similarity between two texts implies that they deal with similar topics.

To further improve the results of the similarity index, we have analysed the incidence of synonyms, polysemic words, and dictionary entries in the texts. Could the different authors' writing styles affect the similarity index between texts? To minimize this, we can store in our database lists of synonyms or polysemic words by establishing groups of words related in terms of significance. Languages like Spanish, English, or French usually have more than 90.000 entries in a dictionary, so these lists represent a very short collection for the database. In any case, this is an improvement process that is on trial yet.

It should be noted here that our intention is not to suggest that the application of these techniques of quantitative analysis should replace a subsequent qualitative analysis. Although this project covers a relatively small time span and, therefore, the meaning of terms and words remains more or less constant, in the next stages it will be necessary to consider that words and concepts are living constructions, variable in time. The TMD does not contain more information than the frequency of words and their position in the text, so once the terms are detected, it is necessary to perform a qualitative review, which can be helped by other Text Mining techniques, such as the word correlation analysis, which detects relationships between words in the text and, therefore, assigns more importance to significance than the mere frequency or location.

→

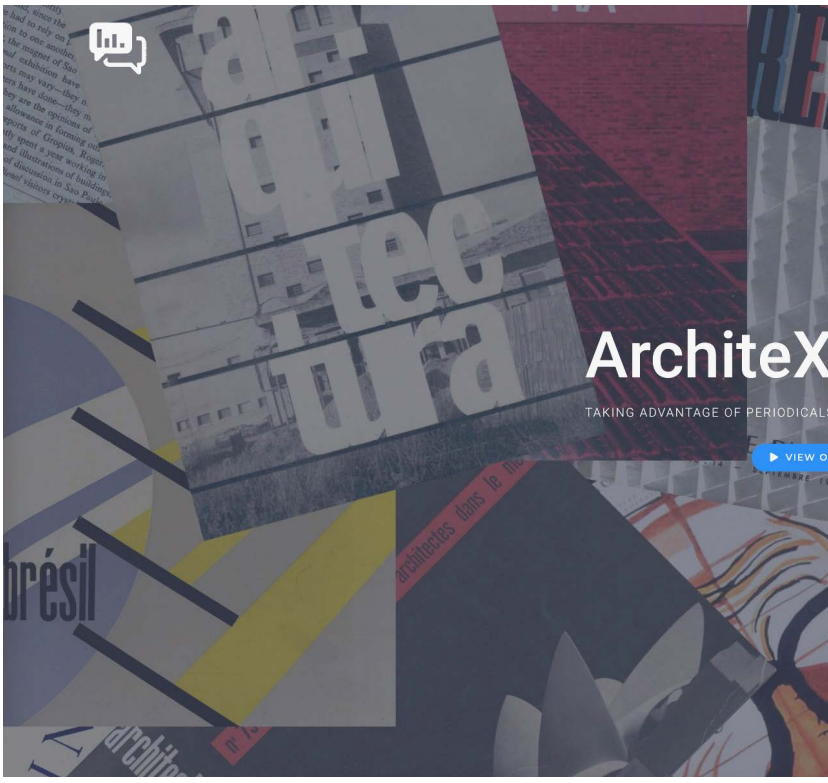
12
One of the "classics" in text-mining literature is Feldman, Sanger, *The Text Mining Handbook*.

SADAŠNJOST: MREŽNA STRANICA PROJEKTA ARCHITEXT MINING

Posljednji zadatak koji je pred nama jest dizajn mrežne stranice za *hosting* alata. Namjeravamo stvoriti mjesto susreta za istraživače koji rade s arhitektonskim časopisima. Dizajnirali smo jednostavno sučelje, tzv. Flat Web Design, gdje korisnici mogu testirati alate i stupiti u kontakt s istraživačkim timom.[13] Osim ovoga, mrežna stranica sadržavat će blog posvećen vijestima ili temama povezanima s istraživanjima koja upotrebljavaju časopise o arhitekturi ili se njima bave. Služit će i za prikupljanje podataka o drugim relevantnim projektima i za povezivanje s bazama podataka koje također mogu poslužiti kao resursi za istraživače.

U ovom se trenutku implementiraju prva dva alata koja će biti dostupna. Prvi omogućuje filtrirano pretraživanje u bibliografsko-tematskoj bazi podataka. Ovaj moćni alat za pretraživanje nudi bibliografske zapise onih članaka koji udovoljavaju kriterijima koje korisnik traži, a može se pretraživati jedna riječ ili sintagma u kombinaciji nekoliko podatkovnih polja. Korisnici će moći pokretati upite koji uključuju pretraživanje pojmova u jednom ili više dostupnih polja, u kojem će ih slučaju povezivati operator AND. Uz osnovno polje za pretraživanje—naslov članka—alat će pomoći korisniku da pronađe znakovni niz u drugim poljima poput „autora" ili „države". Upit se može dodatno ograničiti unosom naslova i broja časopisa, kao i godine izdanja. Nakon odabira članaka mrežni korisnik moći će ispisati ili preuzeti XLS ili PDF datoteku s rezultatima.

Drugi alat koji će biti dostupan izračunavanje je indeksa sličnosti između jednog članka i fiksnog broja drugih koji se nalaze u bazi podataka. Internetski korisnik moći će pokrenuti pretraživanje unosom bibliografskog zapisa članka koji želi usporediti ili povezati s nekim od drugih članaka koji su rezultat prethodne pretrage u alatu. U trenutku pokretanja upita korisnik će moći odrediti broj članaka koje želi dobiti. Ovaj bi alat mogao postati najkorisniji od svih koji su testirani, s obzirom na to da će korisnicima omogućiti da pronađu klastere sličnih članaka a da ih ne pročitaju te da tako mapiraju glavne teme koje su objavljene u časopisima. No još je uvijek potreban niz prilagodbi da bi se moglo smatrati da alat ispravno radi jer na uspjeh pretraživanja utječu čimbenici poput razlike u duljini članaka. U ovom trenutku radimo na otkrivanju postotka razlike u duljini tekstova koji omogućava dobivanje pouzdanih rezultata u pogledu sličnosti. Bez obzira na sve testove i prilagođavanja koja provodimo, pitamo se ima li konceptualno smisla uspoređivati kratke vijesti s opsežnim člancima. Osim našeg ispitivanja bit će potrebna implementacija beta-verzije alata što je prije moguće, kako bi korisnici mogli dati povratne informacije istraživačkom timu. U budućnosti bismo željeli moći zabilježiti sve upite korisnika na mreži i analizirati ih dok tražimo pogreške u alatu, ali za sada nam nedostaje tim koji bi bio zadužen za te zadatke, tako da povratne informacije korisnika mogu doći do nas samo putem kontaktnog formulara, gdje korisnici mogu



Slika 3. Početna stranica internetske stranice projekta ArchiteXt Mining. (Izvor: https://www.architextmining.es) / Figure 3. Homepage of ArchiteXt Mining web site. (Source: https://www.architextmining.es)
↑

13
„ArchiteXt Mining".

THE PRESENT: ARCHITEXT MINING WEBSITE

The last task we are facing is the design of a website to host the tools. We intend to create a meeting point for researchers working with architectural periodicals. We have designed a simple interface, a Flat Web Design, where the users can test the tools and get in touch with the research team.[13] Apart from this, the website will host a blog dedicated to the news or topics related to research with—and about—architecture magazines. It will also serve to compile information about other related projects and to link with databases that may also serve as resources for the researchers.

At this point, the first two tools that will be available are being implemented. The first one allows for filtered searches in the biblio-thematic database. This powerful search tool provides the bibliographic record of those articles that meet the criteria requested by the users, who can search for a single word or collocation in a combination of several database fields. Users will be able to run queries that involve the search of terms in one or more of the available fields, in which case they will be combined by the operator AND. In addition to the basic search field—the title of the article—the tool will help locate a character string in other fields such as "author" or "country." The query may be further limited by entering the journal title and issue, as well as the year of publication. Once the articles have been selected, the online user will be able to print them or download a XLS of PDF file with the results.

The second tool that is being implemented is the calculation of the similarity index of one article with regard to a fixed number of those contained in the database. The online user will be able to run the search by entering the bibliographic record of the article that he or she wants to compare or linking from any of the articles resulting from a previous search in the tool. At the moment of launching the query, the user will be able to determine the number of articles he or she wants to obtain. This tool may well become the most serviceable of all that have been tested, since it will allow the users to locate clusters of similar articles without reading them and thus map the major subjects published in the journals. However, it still needs a number of adjustments before it can be considered to work correctly, since the success of the search is influenced by factors such as the difference between the article lengths. At this moment, we are working on finding out what percentage of length difference between the texts allows us to obtain reliable results in terms of similarity. Regardless of all the tests and adjustments that are being made, the question we ask ourselves is whether it conceptually makes sense to compare short news with extensive articles. Besides our inquiry, it will be necessary to implement a beta version of the tool as soon as possible, so that the users can give feedback to the research team.

13
"ArchiteXt Mining."

priopćiti poteškoće na koje su naišli prilikom pokušaja postizanja koherentnih rezultata. Ako se ova mjera provede, dobiveni podaci upotrijebit će se samo za poboljšanje alata, jamčeći privatnost korisnika, koji će biti unaprijed upozoreni na ovu sekundarnu funkciju.

Kada oba alata budu na mreži i aktivna, pripremit ćemo drugu vrstu pretraživanja, koja će prikupljati nizove znakova u svakom polju baze podataka, pa i u cijelom tekstu članaka. Ovaj će alat imati istu snagu kao i velike tražilice poput Googlea i dijelit će iste poteškoće, budući da neće ograničavati pretraživanje na samo jedno polje, nego će raditi s cjelokupnim sadržajem baze podataka i članaka. Međutim, korisnici će morati znati kako pokrenuti preciznije pretraživanje i stoga te minimalne restrikcije i ograničenja.

Naposljetku, nastojimo poboljšati vizualizaciju rezultata. S jedne strane, članci će biti povezani sa svojim skenovima koji su dostupni na internetu (ako su dostupni). To je lako postići, ali trebat će razumno vrijeme. S druge strane, želimo poboljšati razumijevanje upita prevodeći ih u grafikone i grafike, kako prilikom razmatranja skupa članaka tako i kod analize određenog teksta.

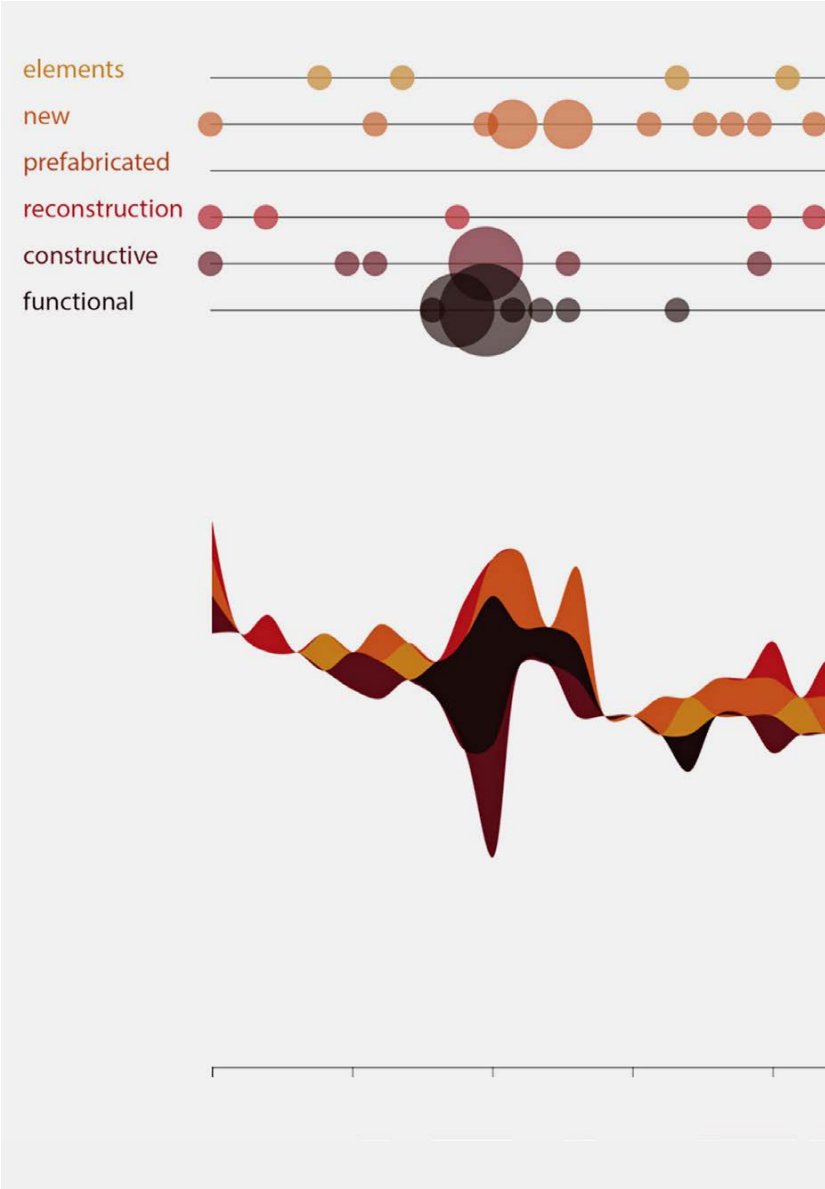
BUDUĆNOST: GLOBALIZACIJA PROJEKTA ARCHITEXT MINING

Tijekom prve faze projekta ArchiteXt Mining testirali smo mogućnosti alata za tekstualnu analitiku kada je u pitanju uvid u sadržaj časopisa o arhitekturi, polje na koje ova tehnika još nije primjenjivana. ArchiteXt Mining uveo nas je u sferu digitalne humanistike. Do neke nam je mjere pokazao snagu, ali i količinu teškoća s kojima ćemo se suočiti u sljedećim fazama.

Tehnike rudarenja teksta nisu u stanju voditi čitav istraživački proces, ali mogu nam pomoći da uštedimo znatnu količinu vremena i definitivno pridonose pronalaženju novih tema. Dodamo li statističku obradu podataka nestrukturiranim podacima prikupljenima iz časopisa, možemo stvoriti alat za analizu teksta visokih performansi.

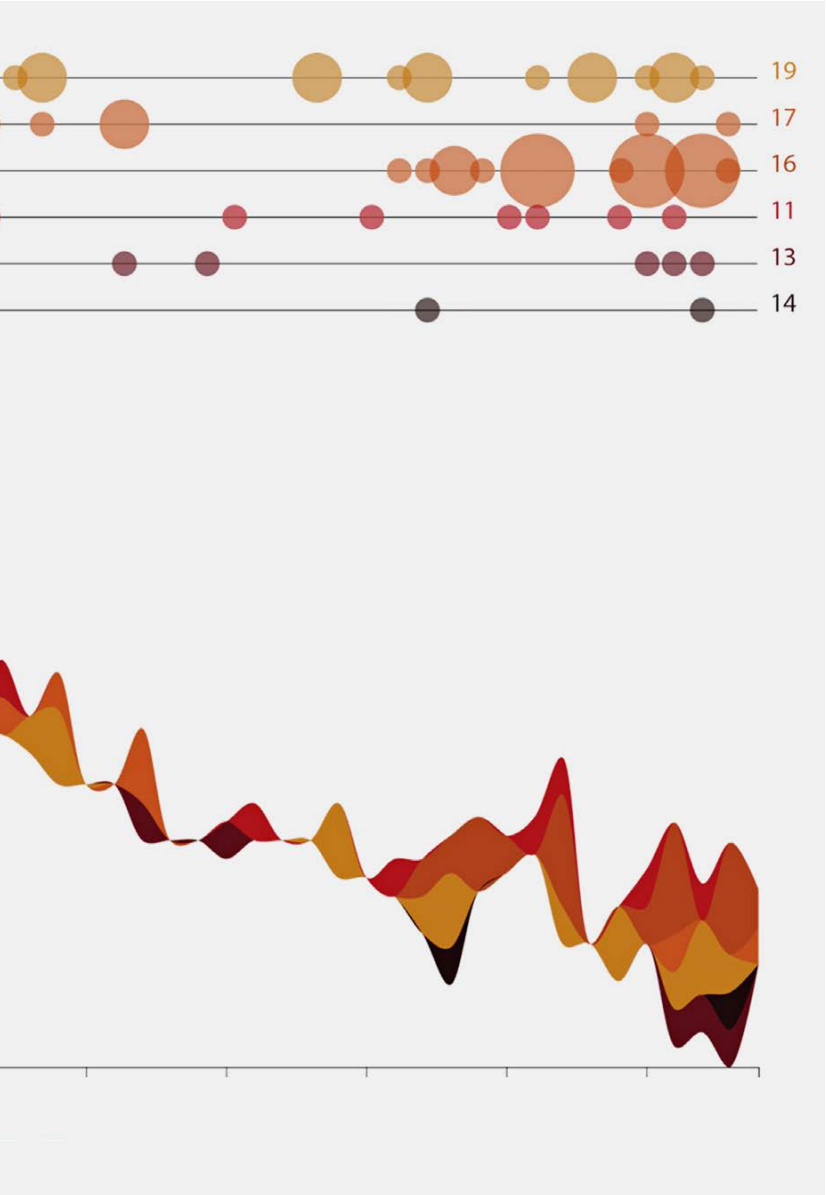
Želimo aplicirati za nova sredstva kako bismo nastavili projekt, šireći ga izvan granica Španjolske kako bismo ga pretvorili u globalni alat koji će olakšati istraživanje širenja arhitektonskih trendova i razmjene između država i kontinenata. U vezi s time, napravili smo korak unaprijed i testirali indeks sličnosti između dvaju tekstova napisanih na različitim jezicima. Kada se upotrebljava automatski prevoditelj, vrijednosti sličnosti reproduciraju uobičajeni obrazac: SIM indeks viši je kada se uspoređuju prevedeni tekstovi. Ovo je pravilo u skladu s trima metodama koje smo ranije predstavili.

To se čini sasvim logičnim s obzirom na to da automatski prevoditelji pojednostavljuju zbirku riječi koje se upotrebljavaju u tekstovima, koristeći se standardnim jezikom kako bi



Slika 4. Načini unapređenja vizualizacije ranijeg Sartorisova teksta. Mjehuričasti grafikon pokazuje učestalost pojave riječi kroz čitavu dužinu teksta, a strujni grafikon prikazuje relativnu frekvenciju riječi. U oba su grafikona akumulacijske točke vrlo jasne. (Izvor: pripremili autori teksta) / Figure 4. Some ways of improving the visualization of the same text by Sartoris. The Bubble Line Chart shows the word frequencies over the length of the article, and the Stream Graph expresses the relative frequencies of the words. In both graphs, the accumulation points are very clear. (Source: prepared by the authors)

↑



In the future, we would like to be able to record all the queries of the online users and analyse them while looking for errors in the tool, but for now we lack a team in charge of those tasks, so the users' feedback can only reach us through a contact form where they can communicate the difficulties they encountered when trying to obtain coherent results. If this measure is implemented, these data will be used only for improving the tool, guaranteeing the privacy of the users and previously warning them about this secondary use.

Once both tools are up and running, another type of search will be prepared that will collect a character string in every field of the database, even in the full text of the articles. This tool will have the same power as the large search engines such as Google and share the same difficulties, as it does not restrict the search to any field and works with the whole contents of the database and the articles. However, the users will need to know how to launch more accurate searches, hence the minimal restrictions and limitations.

Finally, we are trying to improve the visualization of results. On the one hand, the articles will be connected with their scans available online (in case they are available). This is an easy task to perform, but it will need a reasonable time. On the other hand, we expect to improve the comprehension of the queries by translating them into charts and graphics, both when considering a set of articles and when analysing a specific text.

THE FUTURE: GLOBALIZING ARCHITEXT MINING

During the first stage of ArchiteXt Mining, we have tested the possibilities of the text analytics tools when it comes to apprehending the contents of architecture journals, a field that has not been tackled yet by using these techniques. ArchiteXt Mining has introduced us to the digital humanities sphere. To some extent, it has shown us its power, but also the amount of difficulties that we will face at the stages to come.

Text mining techniques are unable to conduct the entire research process, but they can help us save a lot of time and definitely contribute to locating new topics. If we add the statistical treatment of information to the unstructured data collected from the journals, we can create a high-performance tool for text analysis.

We aim to apply for new funds to continue the project, extending it beyond the borders of Spain to turn it into a global tool that will facilitate the research on the dissemination of architectural trends and the exchange between countries and continents. In this regard, we have acted ahead and tested the similarity index between two texts written in different languages. When using automatic translators, the similarity values reproduce a common pattern:

izrazili istu ideju. Stoga treba pažljivo razmotriti usporedbu tekstova napisanih na nekoliko jezika sa standardnim jezikom, poput engleskog. Moramo kvantificirati povećanje SIM-a koje je uključeno u proces automatskog prevođenja tekstova s različitih jezika na engleski. Za sada smo procijenili da automatsko prevođenje sa španjolskog na engleski povećava SIM indeks za 10 do 15 posto. Svaki put kada u bazu podataka dodamo novi jezik, morat ćemo izračunati njegov specifični postotak redukcije.

Jedan od najtežih zadataka u ovoj prvoj fazi bila je digitalizacija časopisa. Zbog ograničenih sredstava projekta nismo mogli uzeti specijaliziranu tvrtku za ovaj posao, što je produljilo i zakompliciralo postupak. Uz to, nakon što smo završili s digitalizacijom triju časopisa, postali su gotovo u potpunosti dostupni na mreži, tako da smatramo da smo mogli uštedjeti mjesece rada i truda. Ovaj je projekt vjerojatno započeo nešto ranije nego što je trebao, zato što su sada neke važne zbirke časopisa o arhitekturi dostupne u otvorenom pristupu.[14] Uz to, ovo nije projekt digitalizacije časopisa, nego digitalne analize njihova sadržaja. Stoga je redoviti rad s već digitaliziranim zbirkama jedna od premisa koje smo postavili za buduće faze projekta. To je uobičajena praksa na polju rudarenja teksta u periodičkim medijima, gdje su provedeni zanimljivi eksperimenti s dinamičnom ekstrakcijom znanja.[15] Budući da radimo s časopisima iz druge polovice 20. stoljeća, ne možemo korisnicima ponuditi skenove stranica, ne samo zbog poteškoća koje donosi digitalizacija, nego i zbog autorskih prava. Stoga planiramo raditi sa zbirkama časopisa koji su već objavljeni u otvorenom pristupu, što će nam omogućiti i njihovo povezivanje s bibliografskim referencijama iz naše baze.

Drugi vrlo naporan zadatak bila je ručna obrada baze podataka i priprema članaka kako bi se automatizirao proces optičkog prepoznavanja znakova. Trenutačno imamo izvrsnu bazu podataka o temama objavljenima u španjolskim časopisima, kojom se možemo nastaviti koristiti godinama, ali shvatili smo da prosječnom korisniku ArchiteXt Mininga nije potrebna tako iscrpna baza podataka da bi dobio zanimljive rezultate. U tom smislu, drugi je od naših srednjoročnih ciljeva automatiziranje procesa, kako razvoja baze podataka tako i automatskog vađenja članaka i njihova sadržaja, što je već dobro funkcioniralo s novinama, koje su još složeniji medij jer svaka stranica sadrži daleko više referencija nego časopisna.[16]

Nedavno objavljeni članak „Content Analysis of 150 Years of British Newspapers"[17] primjer je koji dobro odražava i reafirmira neke od naših budućih ciljeva, s obzirom na to da tvrdi ne samo da je moguće otkriti obrasce skrivenih kulturnih promjena analizom opsežnog niza tekstova nego i da periodika sadrži obilje podataka koji se ne pojavljuju u knjigama, izvoru iz kojega se obično piše povijest.

the SIM index is higher when the translated texts are compared. This rule is consistent with the three different methods presented before.

This seems quite logical, given that the automatic translators simplify the collection of words used in texts, using a standard language to express the same idea. Therefore, comparing texts written in several languages with a standard language, such as English, must be carefully considered. We need to quantify the SIM increase that is involved in the process of automatic translation of texts of any language into English. For the moment, we have estimated that automatic translation from Spanish to English increases the SIM index by 10-15 per cent. Each time we add a new language to the database, we will have to proceed by calculating its particular percentage of reduction.

One of the hardest tasks at this first stage has been the digitalization of magazines. Due to the limited funds of the project, we could not delegate this work to a specialized company, which has lengthened and complicated the process. In addition, after finishing the digitalization of the three journals, they became almost completely available online, so we feel we could have saved months of work and effort. This project probably started a little sooner than it should have been the case, since currently some important collections of architecture journals are starting to be available in open access.[14] In addition, it is not a project of digitalizing journals, but of a digital analysis of their contents. Thus, working regularly with the already digitalized collections is one of the premises that we have set for the future stages of the project. This is a usual practice in the field of text mining in periodical media, where interesting experiments with dynamic extraction of knowledge have been carried out.[15] Because we work with magazines from the second half of the 20th century, we cannot provide the users with scans of the pages, not only because of the difficulties involved in digitalization, but also because of the copyright. Thus, we are planning to work with the collections of journals that are already published in open access, which will also allow us to connect them with the bibliographic references of our database.

14
Poveznice na neke od njih nalaze se na: „Repositories".
15
Aggarwal, Zhai, „An introduction to Text Mining".
16
Nørvåg, Oyri, „News Item Extraction for Text Mining in Web Newspapers".
17
Lansdall-Welfare *et al.*, „Analysis of 150 Years of British Periodicals".

Another of the most tedious tasks has been the manual elaboration of the database and the preparation of articles in order to automate the process of optical character recognition. Right now, we have an excellent database on topics published in Spanish journals, which we can continue to take advantage of for years, but we have realized that the common user of ArchiteXt Mining does not need such an exhaustive database to obtain interesting results. In this sense, another of our medium-term goals is to automate the processes, both the development of the database and the automatic extraction of articles and their contents, something that has already worked well with the newspapers, which is an even more complicated media type, since each page contains many more references than a magazine.[16]

Recently published, the article "Content Analysis of 150 Years of British Newspapers"[17] is one example that properly reflects and reaffirms some of our future objectives, since it argues not only that it is possible to detect patterns of hidden cultural changes by analysing extensive series of texts, but also that periodical publications contain plenty of data that do not appear in books, the source from which history is usually written.

.

ZAHVALE

„ArchiteXt Mining: Spanish modern architecture through its texts (1939–1975)" HAR2015–65412–P (MINECO/ERDF) istraživački je projekt koji je financirala Vlada Španjolske putem poziva za „projekte izvrsnosti" Ministarstva gospodarstva i konkurentnosti (MINECO) i Europskog fonda za regionalni razvoj (ERDF) 2015. godine.

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).

14
Some of them are linked on: "Repositories."
15
Aggarwal, Zhai, "An introduction to Text Mining."
16
Nørvåg, Oyri, "News Item Extraction for Text Mining in Web Newspapers."
17
Lansdall-Welfare *et al*., "Analysis of 150 Years of British Periodicals."

POPIS LITERATURE / BIBLIOGRAPHY

Aggarwal, Charu C., Zhai, ChengXiang. „An introduction to Text Mining", 1–10. U/In: *Mining Text Data*, ur./ed. Charu C. Aggarwal, ChengXiang Zhai. New York: Springer, 2012.

Alarcón, Enrique, Bernot, Eduardo. „Index Thomisticus by Roberto Busa and Associates". Dostupno na/Available at: http://www.corpusthomisticum.org/it/index.age (pristupljeno 16. svibnja 2019./last accessed on May 16, 2019).

„ArchiteXt Mining". Dostupno na/Available at: https://www.architextmining.es/ (pristupljeno 19. studenoga 2019./last accessed on November 19, 2019).

Bichindaritz, Isabelle, Akkineni, Sarada. „Concept Mining for Indexing Medical Literature". *Engineering Applications of Artificial Intelligence* 19/4 (2006.): 411–417.

Bonta, Juan Pablo. *American Architects and Texts: A Computer-Aided Analysis of the Literature*. Cambridge, MA: The MIT Press, 1996.

Esteban-Maluenda, Ana, San Pablo, Luis. „ArchiteXt Mining: Taking Advantage of Periodicals as an Architectural Database". Dostupno na/Available at: https://www.architextmining.es (pristupljeno 25. svibnja 2019./last accessed on May 25, 2019).

Feldman, Ronen, Sanger, James. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge: Cambridge University Press, 2007.

Higgins, Devin. „Reading and Non-Reading: Text Mining in Critical Practice", 85–100. U/In: *Top Technologies Every Librarian Needs to Know: A LITA Guide*, ur./ed. Kenneth J. Varnum. Chicago: American Library Association, 2014.

Hotho, Andreas, Nürnberger, Andreas, Paaß, Gerhard. „A Brief Survey of Text Mining". *LDV Forum* 20/1 (2005.): 19–62.

Lansdall-Welfare, Thomas, Sudhahar, Saatviga, Thompson, James, Lewis, Justin, FindMyPast Newspaper Team, Cristianini, Nello. „Analysis of 150 Years of British Periodicals", 1–9. U/In: *Proceedings of the National Academy of Sciences*, ur./ed. Kenneth W. Wachter. Berkeley, CA: University of California, Berkeley, 2017.

Nørvåg, Kjetil, Oyri, Randi. „News Item Extraction for Text Mining in Web Newspapers", 195–204. U/In: *International Workshop on Challenges in Web Information Retrieval and Integration*. Tokyo: IEE, 2005.

„Repositories", *ArchiteXt Mining*. Dostupno na/Available at: https://www.architextmining.es/links/links/ (pristupljeno 19. studenoga 2019./last accessed on November 19, 2019).

Tseng, Yuen-Hsien, Chang, Chun-Yen, Chang Rundgren, Shu-Nu, Rundgren, Carl Johan. „Mining Concept Maps from News Stories for Measuring Civic Scientific Literacy in Media". *Computers & Education* 55/1 (2010.): 165–177.