PAM Pokrajinski arhiv Maribor

Moderna arhivistika

Časopis arhivske teorije in prakse
Journal of Archival Theory and Practice

Letnik 1 (2018), št. 2 / Year 1 (2018), No. 2

Maribor, 2018

Moderna arhivistika

Časopis arhivske teorije in prakse

Journal of Archival Theory and Practice

Letnik 1 (2018), št. 2 / Year 1 (2018), No. 2

ISSN 2591-0884 (online)
ISSN 2591-0876 (CD_ROM)

Izdaja / Published by:
Pokrajinski arhiv Maribor / Regional Archives Maribor

Glavni in odgovorni urednik / Chief and Responsible editor:
Ivan Fras, prof., Pokrajinski arhiv Maribor, Glavni trg 7, SI-2000 Maribor,
telefon/ Phone: +386 2228 5017; e-pošta/e-mail: ivan.fras@pokarh-mb.si

Glavna urednica / Editor in chief:
mag. Nina Gostenčnik

Uredniški odbor / editorial board:

- dr. Thomas Aigner, Diözesanarchiv St. Pőlten, Avstrija
- dr. Borut Batagelj, Zgodovinski arhiv Celje, Slovenija
- dr. Bojan Cvelfar, Arhiv Republike Slovenije, Slovenija
- mag. Nada Čibej, Pokrajinski arhiv Koper, Slovenija
- Ivan Fras, Pokrajinski arhiv Maribor, Slovenija
- mag. Nina Gostenčnik, Pokrajinski arhiv Maribor, Slovenija
- dr. Joachim Kemper, Institut für Stadtgeschichte Frankfurt am Main, Nemčija
- Leopold Mikec Avberšek, Pokrajinski arhiv Maribor, Slovenija
- dr. Miroslav Novak, Pokrajinski arhiv Maribor, Slovenija
- dr. Rik Opsommer, Stadsarchief Ieper - Universiteit Gent, Belgija
- Darko Rubčić, Državni arhiv u Zagrebu, Hrvaška
- dr. Izet Šabotić, Filozofski fakultet Univerziteta u Tuzli, Bosna in Hercegovina
- mag. Boštjan Zajšek, Pokrajinski arhiv Maribor, Slovenija

Recenziranje / Peer review process:
Prispevki so recenzirani. Za objavo je potrebna pozitivna recenzija. Proces recenziranja je anonimen. / All articles for publication in the conference proceedings are peer-reviewed. A positive review is needed for publication. The review process is anonymous.

Lektoriranje / Proof-reading:
mag. Boštjan Zajšek, mag. Nina Gostenčnik

Prevajanje:
mag. Boštjan Zajšek (slovenščina), mag. Nina Gostenčnik (slovenščina, angleščina), Lučka Mlinarič (bosanščina, hrvaščina, srbščina)

Oblikovanje in prelom / Design and typesetting:
mag. Nina Gostenčnik

http://www.pokarh-mb.si/si/p/3/49/moderna-arhivistika.html

Prejeto / Received: 23. 03. 2018

1.03 Kratki znanstveni članek *1.03 Short Scientific Article*

COMPUTATIONAL ARCHIVAL SCIENCE

Hrvoje Stančić, Ph. D.
Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
hrvoje.stancic@zg.t-com.hr

Abstract:

The digitisation of archival materials and ingest of digitally born materials in digital archives has led to the possibilities of application of the big data analytical principles in the digital archives. The author explains the 5V characteristics of big data. He proceeds to define the concept of Computational Archival Science (CAS). Two CAS examples are given in order to illustrate the type of research that can be conducted in that area. Further, the author explains the prerequisites for engaging with CAS. Finally, suggestions on how archival institutions might get involved in CAS activities are given.

Key words:

digital preservation, computational archival science, CAS, big data, NLP, network analysis, visualisation

Izvleček:

Računalniška arhivistika

Digitalizacija arhivskega gradiva in zajem izvorno digitalnega gradiva v digitalni arhiv sta pripeljala do možnosti aplikacije analitičnih načel masovnih podatkov v le-teh. Avtor v prispevku pojasni 5 značilnosti masovnih podatkov in nadaljuje z definicijo koncepta računalniška arhivistika. Za ilustracijo tipov raziskav, ki bi lahko bile opravljene na tem področju, predstavlja dva primera. Nadalje avtor razloži predpogoje, ki so potrebni za raziskave na področju računalniške arhivistike ter na koncu poda predloge kako se lahko arhivske institucije vključijo v ta proces.

Ključne besede:

digitalno varstvo, računalniška arhivistika, masovni podatki, procesiranje naravnega jezika, analiza omrežja, vizualizacija

1. Introduction

Digitisation in the archives is not news any more. It has become a regular activity. For some archives, particularly smaller ones, ingesting digitally born materials in the digital archive is still a challenge. However, this is changing as well. In this context the main activities are enabling the search and retrieval. In order to accomplish this, the archives are adding metadata to the digital materials for description and contextualisation purposes. To be able to successfully retrieve records, they have to stay preserved despite the inevitable and fast pace of changes of the information and communication technology. Therefore, the archives have defined best practices (Digital Preservation Coalition, 2015), listed preferred storage media, defined sustainability factors for file formats (Library of Congress, 2017) as well as developed and implemented records

management and digital archiving systems with the aim of (long-term) preservation of authentic, accurate and reliable records in their context. The archives have also defined their digital preservation policies. “Organisations are increasingly creating, using and storing records only in digital formats. Unlike analogue records, the digital records benefit significantly from assessment as early as possible for any preservation requirements. This presents a significant challenge for archives who will need to be able to identify, collect and manage the content of these records to ensure it remains authentic and accessible. The digital preservation policy provides a mandate under which an archive can oversee these processes and manage digital preservation.” (The National Archives, 2011) The archives today also actively communicate with the organisations producing digital archival materials in order to advise them on the proper procedures before the materials are transferred to the archives.

One of the activities that is still to be accomplished by most of the archives is proactive communication with cloud service providers offering cloud recordkeeping or cloud-based archiving services regarding setting up the services according to the archival requirements and expectations. (Stančić, Rajh, & Milošević, 2013) Another trend that is still to be tackled by the archives is the big data aspect of the archived materials. Further, analysis and discussion will focus on this aspect of challenges facing digital archives.

2. Archives and big data challenges

Oxford dictionary defines big data as “extremely large data-sets that may be analysed computationally to reveal patterns, trends, and associations, especially relating to human behaviour and interactions”[1]. The term *big data* is usually to be found in the context of real-time or near real-time data ingest in the records management solutions or digital archives and its usage in the business decision-making process. However, the digital archives are also dealing with large amounts of records that share some of the same characteristics the big-data systems have.

The fact that one is dealing with the big data is usually described using the 5Vs definition referring to 1) volume, 2) velocity, 3) variety, 4) veracity, and 5) volatility. The *volume* characteristic speaks to the fact that the digital archive is dealing with large amount of records or large data-sets for which it needs specialised (archival) software that can handle the volume of data being ingested. The *velocity* characteristic describes the rate at which the archive is receiving new records. For example, in case of digital archives capturing social media feeds this could be a real-time activity while in the case of web-archiving the velocity could range from once a day to once a month or once a year. The *variety* characteristic describes the number of different file formats the digital archive is preserving. The more file formats the more difficult it will be to perform long-term preservation. Taking again the example of web-archiving, the digital archive may wish to perform an inventory of harvested file formats and identify their versions (e.g. PDF/A-1, PDF/A-2, or PDF/A-3). This would make the required preservation actions much easier to accomplish later on. The *veracity* characteristic speaks about accuracy, confidence, or trust in the data, i.e. is the data correct. For illustration, is the same information in the records (e.g. a person’s title) always represented (e.g. abbreviated) in the same way and is it referring to the same person. Finally, the *volatility* characteristic describes how long the records need to be kept after which they became irrelevant and can be deleted. From the archival perspective this characteristic is closely connected with the retention and disposition plan.

[1] *Oxford dictionary, s.v. big data, https://en.oxforddictionaries.com/definition/big_data.*

Big data concept refers more on how the data, documents or records are used than on their sheer volume. Therefore, they need to be interpreted properly, or they need to be made interpretable by the users. This means that it is no longer enough to merely describe a record but also to semantically enrich it by identifying and tagging important information. It also means that certain records or tags could be connected to other open data resources. For example, one can identify and tag a person's name in a record and link it to an open source biographic database. Further, one can identify and tag the position held by that person and tag a date of the record. Later on, users could find all persons holding a position in certain time span and find their biographies. This is just a very straightforward example. Much more could be accomplished by application of different approaches from the newly defined field of computational archival science.

3. Computational Archival Science (CAS)

Computational Archival Science (CAS) can be defined as "an interdisciplinary field concerned with the application of computational methods and resources to large-scale records/archives processing, analysis, storage, long-term preservation, and access, with aim of improving efficiency, productivity and precision in support of appraisal, arrangement and description, preservation and access decisions, and engaging and undertaking research with archival material". (Computational Archival Science, 2016) The term appears for the first time in December 2016 at the IEEE Big Data 2016 Conference "Computational Archival Science: Digital Records in the Age of Big Data" held in Washington D.C. Many different topics fit under the CAS umbrella. The idea is to apply a cross-disciplinary approach to the analysis of digitised and born-digital archival materials.

3.1. Examples of CAS applications

How different analytical methods can be applied to archival materials will be firstly shown at the example of the Enron e-mail archive. CAS engages text- and data-mining applications to analyse records. It also applies sentiment analysis and opinion mining techniques in order to achieve insight into attitudes and opinions of the individuals. In relation to the e-mail archives, CAS applies natural language processing (NLP) and network analysis methods in order to detect relationships between people and determine their type (e.g. formal, informal relationship).

The Enron e-mail dataset "was collected and prepared by the CALO Project (A Cognitive Assistant that Learns and Organizes). It contains data from about 150 users, mostly senior management of Enron, organized into folders. The corpus contains a total of about 0.5M messages. This data was originally made public, and posted to the web, by the Federal Energy Regulatory Commission during its investigation". (Cohen, 2015) "There are many types of social networks that could be mined from the data. For example, with named entity recognition techniques in place, one could identify people and organizations mentioned within an e-mail body and infer social links based on textual proximity. (In the case of direct communication analysis) the inferred social network is based simply on e-mail communication, that is, an e-mail from one person to another is initially treated as a directed edge from sender to recipient" while setting focus on "Enron company business or strategy". (Heer, 2004) The analysis (partly shown in Figure 1) gave insight of who knew what and when, who reported to whom (was there an exchange of messages or they were unidirectional), and which individuals at which level of the decision-making process were the most influential ones (e.g. receiving messages from

many different employees indicated a person being either a hub or an authority in the network).

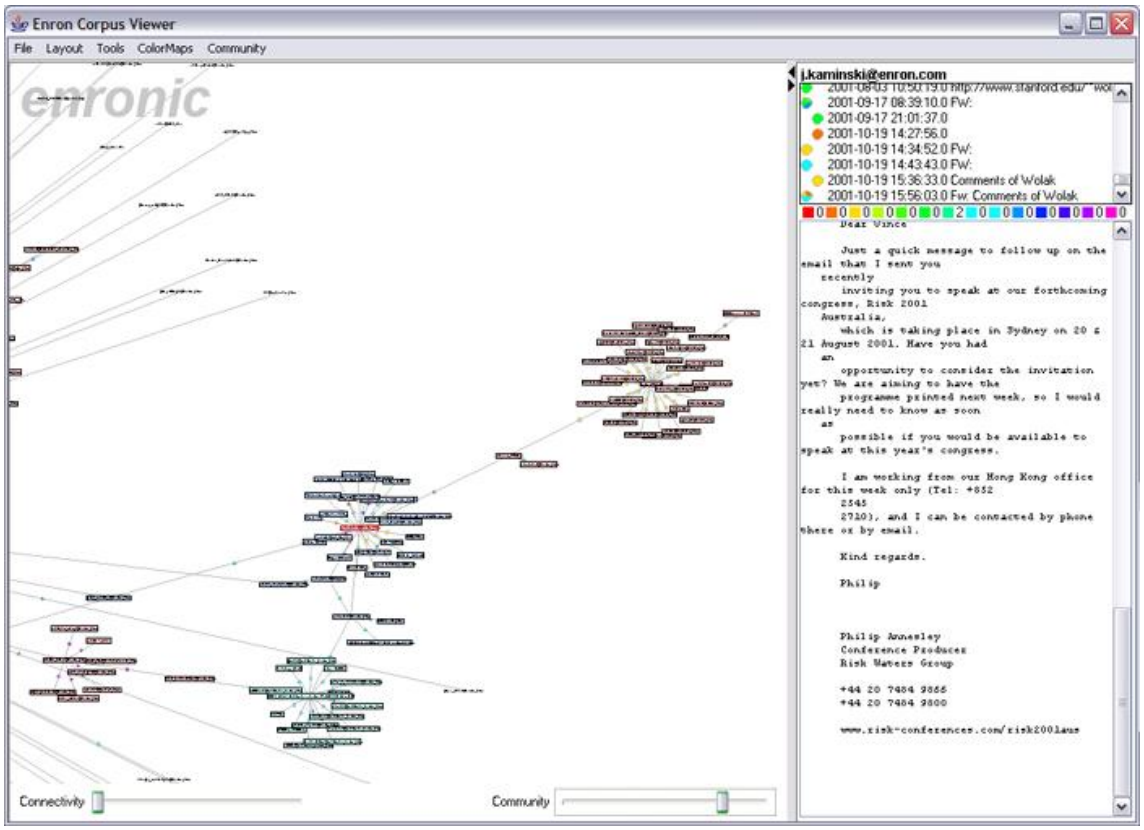

Figure 1. Analysis of the Enron e-mail corpus – automatically detected communities (Heer, 2004)

Another example of analysis that could be found as part of the CAS is mapping and visualisation of urban inequality in Richmond, Virginia, USA. The information from the area information records, including description of terrain, inhabitants (including fields: type, estimated annual family income, foreign-born (nationality), negro (yes/no) etc.), buildings (including fields: type, type of construction, average age etc.), history (range of sales or rental values and when the peak value occurred), occupancy (in percentage – of land, dwelling units and home owners), sales and rental demand, estimated availability of mortgage funds, trend of desirability in the following 10-15 years, and confidential remarks were digitised and analysed. The analysis involved geo-mapping of the digitised information to the geographical maps of the time. By analysing the ratio between various factors, e.g. quality of the buildings, their location and average income of the people living there, it was possible to visualise the dynamics of the population, determine the move of the wealthier population towards the suburbs etc. Part of the visualisation is showed in Figure 2.

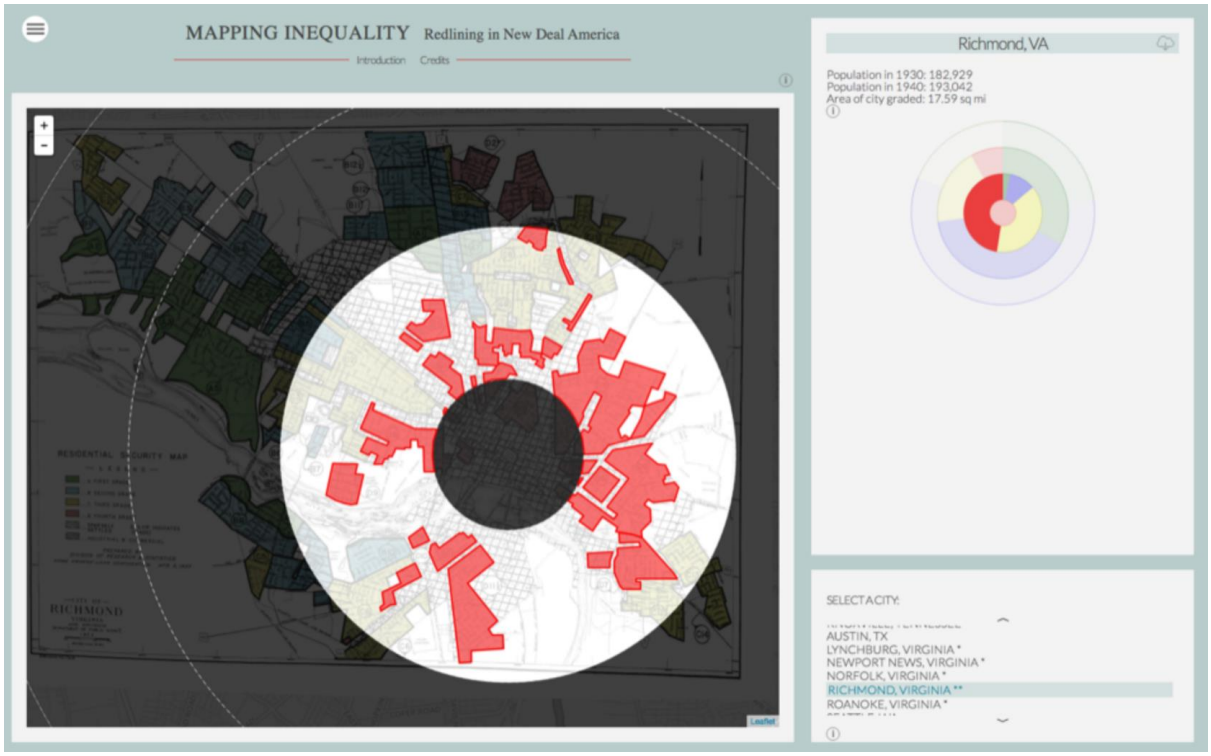

Figure 2. Mapping inequality in Richmond (Nelson)

The results of the CAS research detailed here are just two of many possible examples. They show how different analytical and visualisation methods could be applied to the archival records.

4. CAS prerequisites

In order to apply any kind of advanced analytical methods, enough materials need to be digitised or collected in the digital form at the first place. Next, they need to be described, semantically enriched and made publically available. Also, there is a need for an information infrastructure to support, often computationally intensive, requirements of analytical applications. More importantly, there is a need for knowledge and expertise to perform the analysis that often calls for cooperation of “traditional” and “modern” archivists, i.e. those more knowledgeable of the records and their contents and those more knowledgeable of the analytical solutions. It also requires interdisciplinary approach involving researchers and specialists from several different fields.

The aim of the computational archival science, taking into account the identified prerequisites, is to establish the new generation of archival and business-oriented e-services. Those services are aiming to enable re-use of the information stored in the records and large data-sets, i.e. to enable anyone with enough skills to connect to those resources through the developed APIs and to create new values through the development of yet another layer of new services – either for profit, for tourism, for heritage purposes, or for any other purpose imagined.

5. Conclusion

It still remains to be seen if the newly coined term Computational Archival Science will stick or not. It is also a question whether this is a new scientific field or just an application of IT tools and methods to the constantly broadening aspects of archival science. I would argue for the latter. However, CAS is a catchy term that may stay on for a while. It can also help in promoting archival science and making the records more readily usable.

Taking into account all that has been said, institutions responsible for the preservation of national heritage should take an active course toward formulation of their strategies aiming to ingest and preserve digitally born materials. It would require not only building internal capacities, both in terms of infrastructure and personnel, but also engaging in the outward-oriented activities. These could be realised through the education of wider public and key stakeholders about the activities of the archival institutions, their online services as well as on the possibilities of API connection and data, documents or records re-use. That way the big data aspect of the archival holdings might be brought to the individual users in a more understandable way (e.g. through visualisations or trend detections). By enabling easier access to the archival holdings, more quality research might be conducted. Also, by enabling the API connections the archives might gain more important role in providing access to the valuable resources, in their re-use and in creation of the new values.

POVZETEK

RAČUNALNIŠKA ARHIVISTIKA

dr. Hrvoje Stančić
Filozofska fakulteta, Univerza v Zagrebu, Hrvaška
hrvoje.stancic@zg.t-com.hr

V uvodu avtor podaja misel, da je digitalizacija v arhivih že redna dejavnosti enako kot je zajem izvorno digitalnih dokumentov v digitalni arhiv. Z namenom svetovanja o pravilnih postopkih pred prevzemom digitalnega gradiva v arhiv, le-ti aktivno sodelujejo z organizacijami, ki digitalno gradivo ustvarjajo.

Avtor zagovarja tezo, da lahko velike količine arhivske gradiva obravnavamo tudi z vidika masovnih podatkov. Podaja razlage petih značilnosti masovnih podatkov – obsega, hitrosti, raznolikosti, verodostojnosti in časa hrambe. Trdi, da pri konceptu masovnih podatkov ne govorimo o količini podatkov, dokumentov ali zapisov, ki jih hranimo ali arhiviramo, ampak o načinu kako jih uporabiti.

Avtor na podlagi razloženih načel definira računalniško arhivistiko kot interdisciplinarno področje, ki se ukvarja z aplikacijo računalniških metod in virov pri obdelavi, analizi, hrambi, dolgoročnem varstvu obsežnih zapisov/arhivskega gradiva in dostopu do njega. Cilj je izboljšanje učinkovitosti, produktivnosti in natančnosti pri vrednotenju, urejanju in popisovanju, podajanju odločitev o varovanju in dostopu do gradiva ter uporabe in izvajanja raziskav z arhivskim gradivom. Podaja dva primera raziskav, ki temeljita na računalniški arhivistiki – prva uporablja metode procesiranja naravnega jezika (ang. natural language processing – NLP) in analize omrežja za ugotavljanje relacij med ljudmi in določanje njihovega tipa pri obdelati e-poštnega seta podatkov Enron, in druga, ki prikazuje mapiranje in vizualizacijo urbane neenakosti v

Richmondu v državi Virginia, ZDA. Na osnovi predstavljenega, avtor definira predpogoje za računalniško arhivistiko.

Na koncu avtor razglablja ali je računalniška arhivistika nova znanost ali samo aplikacija IT orodij in metov na področju arhivistike (in podaja argumente za slednje). Zaključuje s prihodnjimi dejanji, ki bodo potrebni za ukvarjanje z dejavnostmi računalniško arhivistike in razloži pozitivne učinke, ki jih lahko imajo za širšo družbo.

Sources and bibliography

Cohen, W. W. (2015, May 8). *Enron Email Dataset.* Retrieved March 16, 2018, from https://www.cs.cmu.edu/~enron/

Computational Archival Science. (2016). The “COMPUTATIONAL ARCHIVAL SCIENCE (CAS)” Portal. Retrieved March 16, 2018, from http://dcicblog.umd.edu/cas/

Digital Preservation Coalition. (2015). Digital Preservation Handbook. (2nd Edition). Retrieved March 16, 2018, from https://dpconline.org/handbook

Heer, J. (2004, September). *Exploring Enron. Visualizing ANLP Results.* Retrieved March 16, 2018, from https://homes.cs.washington.edu/~jheer//projects/enron/v1/

Library of Congress. (2017). Sustainability of Digital Formats: Planning for Library of Congress Collections. Retrieved March 16, 2018, from https://www.loc.gov/preservation/digital/formats/index.shtml

Nelson, R. K. (n.d.). *Visualizing Urban Inequality.* Retrieved March 16, 2018, from https://drive.google.com/file/d/0B9kwFSGeIVm8Z2NkS0Z0cU5OLWs/view

Oxford dictionary, s.v. big data, https://en.oxforddictionaries.com/definition/big_data

Stančić, H., Rajh, A., & Milošević, I. (2013). "Archiving-as-a-Service". Influence of Cloud Computing on the Archival Theory and Practice. In L. Duranti, & E. Shaffer (Ed.), *The Memory of the World in the Digital Age: Digitization and Preservation* (pp. 108-125). Vancouver: UNESCO. Retrieved March 16, 2018, from http://ciscra.org/docs/UNESCO_MOW2012_Proceedings_FINAL_ENG_Compressed.pdf

The National Archives. (2011). Digital Preservation Policies: Guidance for archives. Retrieved March 16, 2018, from http://www.nationalarchives.gov.uk/documents/information-management/digital-preservation-policies-guidance-draft-v4.2.pdf