Submitted on: 11/08/2014

Semantic analysis of the user queries in the Croatian Historical Newspapers Portal log files

Sofija Klarin Zadravec
Croatian Institute for Librarianship, National and University Library in Zagreb, Zagreb, Croatia
E-mail address: sklarin[at]nsk.hr

Abstract:

A historical newspapers digital library contains the basic elements and relations of a digital library, but also some special features and functions necessary for processing its content - huge amount of digitised pages with complex granularity and large amount of (historical) text. Descriptive metadata of newspapers are usually available on the title level only and the full-text search depends on the accuracy of OCR as well as orthographic and semantic issues of historical text. The characteristics of the historical newspapers digital library content pose a challenge for information retrieval and give rise to the following questions: How do users search digitised historical newspapers?, How do they formulate their search queries?, What topics are they looking for?, How do they deal with historical text issues?. Data stored in the historical newspapers digital library search logs can provide some of the answers and help to improve information access. The paper reports on the results of the semantic analysis of the Croatian Historical Newspapers Portal user queries.

Keywords: historical newspapers, search queries, log files analysis, semantic analysis.

Introduction

Building a historical newspapers digital library poses many strategic, technical and practical challenges. Digitised newspapers management and historical newspapers digital library services are mostly determined by the *complex content* (i.e. structure, layout and language), the *requirements and resources of cultural heritage institutions* (i.e. service providers) and *users' needs*. Unlike ILS, building a digital library system requires new types and levels of metadata to enable the organisation of digital content, to document the processes and the context of a digital object (i.e. digitisation, rights, provenance etc.) and to improve retrieval

and access. The historical newspapers digital library surely needs a metadata standard ”that will work for more specific and at more granular levels of descriptions to provide granular levels of access services.“[1] To overcome the limitations of the MARC format, Allen and Schallow proposed ”metadata relevant to the image processing and to the historians who will use the collection“ [2] including *page metadata*, *layout metadata*, *text objects metadata*, *graphical objects metadata* etc. New administrative and structural metadata schemes support management and presentation of complex and compound objects but there is room for improvement in the area of information retrieval of the historical text. Full text search of digitised newspapers can provide an extremely large number of results. Search quality depends on the accuracy of OCR and orthographic and semantic characteristics of the historical text. For instance, while searching for a term *woman* in the *Croatian Historical Newspapers Portal*[3], a user should be aware of its variant orthographic forms (i.e. *žena* and *xena*). Diachronically related forms that are considered synonyms pose another linguistic obstacle in the *name* and *place* searching (i.e. *Split, Spljet, Spalato; Zadar, Zara, Jadera*). It is obvious that ”Common full-text search tools can only be applied successfully by users who are able to formulate queries with (a) knowledge of historical language and (b) insight in the relevant time spam from which the words have evolved.“[4] To enhance the retrieval of historical texts, large scale digitisation projects are applying and developing a range of specific tools, including scanned image enhancement tool[5], Optical Layout Recognition – OLR, Optical Character Recognition – OCR, automated identification of segments and genres[6], extracting metadata from OCRed text, text mining[7], Named Entity Recognition – NER, linking of contemporary search terms to their historical equivalents[8] etc., as well as

[1] Han, Myung-Ja. Metadata with levels of description: new challenges to catalogers and metadata librarians. // International Federation of Library Associations, World Library and Information Congress, Helsinki, 2012. P.3. Available at: http://conference.ifla.org/past-wlic/2012/80-han-en.pdf

[2] Allen, Robert B.; John Schallow. Metadata and data structure for the historical newspapers digital library. // Proceedings of the 8th international conference on Information and knowledge management CIKM 99. Available at: http://boballen.info/PAPERS/META/meta.pdf

[3] Portal Stare hrvatske novine = Croatian Historical Newspapers Portal. Available at: http://dnc.nsk.hr/newspapers/Default.aspx

[4] De Jong, Francisca; Henning Rode; Djoerd Himjestra. Temporal language models for the disclosure of historical text. Available at: http://eprints.eemcs.utwente.nl/7266/01/db-utwente-433BCEA2.pdf

[5] Impact project (Improving Access of Text). Available at: http://www.impact-project.eu/

[6] Allen, Robert B.; I. Waldstein; W. Zhu. Automated processing of digitized historical newspapers: identification of segments and genres. // International Conference on Asian Digital Libraries, Hanoi, Vietnam, 2008. Pp. 380-387. Available at: http://boballen.info/PAPERS/NewsGenres.pdf

[7] Allen, Robert. Improving access to digitized historical newspapers with text mining, coordinated models, and formative user interface design. // IFLA Newspaper Section Meeting, New Delhi, February 2010. Available at: http://boballen.info/RBA/PAPERS/IFLA2010/iflaDelhi.pdf

[8] De Jong, Francisca; Henning Rode; Djoerd Himjestra. Ibid.

crowdsourced OCR text correction. As the Ahonen and Hyvönen's article[9] announced, future development in this area will probably focus on semantic web techniques implementation – adding semantic metadata to the historical text and publishing cultural content in the semantic web. The techniques applied in the field of Information Extraction (IE) and Natural Language Processing (NLP) can also be used to better understand users' interaction with the digital library of newspapers.

Prior researches of newspapers digital library users

According to Jansen and Booth "Understanding what the user is searching for and providing this content is at the heart of designing successful Web search applications."[10] The studies that investigated research topics in the humanities and social sciences can provide a broader context for understanding the historical newspaper digital library users. A. Jones extensively explored the types of content that are the subject of interest for historians and social scientists.[11] Petras, Larson and Buckland[12] give a review of the studies investigating users' queries in the humanities (M. Bates et al.[13], H. Tibbo[14], T. Gill[15]) and identify the following search categories: *biography* (*person*), *chronology* (*period or event*) and *geography* (*place*). These categories are also represented in the *Functional Requirements for Bibliographic Records* (1998)[16] as entities of interest for users of bibliographic records (i.e. *concept, object, event, place, work, expression, manifestation, item, people, family, corporate body*). Historical newspapers are not only an important research source for scholars but also for other user groups. Recent studies in the newspapers digital library users domain (e.g. California Digital Newspaper Collection, Cambridge Public Library, Utah Digital Newspapers, National Library of New Zealand, National Library of Australia) identified

[9] Ahonen, Eeva; Eero Hyvönen. Publishing historical texts on the semantic web: a case study. Available at: http://www.seco.tkk.fi/publications/2009/ahonen-hyvonen-historical-texts-2009.pdf

[10] Jansen, Bernard J.; Danielle Booth. Classifying Web queries by topic and user intent. CHI 2010: Work-in-Progress, April 14–15, 2010, Atlanta, GA. Available at: http://faculty.ist.psu.edu/jjansen/academic/jansen_user_intent.pdf

[11] Jones, Alison. The many uses of newspapers.
Available at: http://dlxs.richmond.edu/d/ddr/docs/papers/usesofnewspapers.pdf

[12] Petras, Vivien; Ray R. Larson; Michael Buckland. Time period directories: a metadata infrastructure for placing events in temporal and geographic context. // Joint Conference on Digital Libraries - JCDL Workshop, 2006. Available at: http://metadata.sims.berkeley.edu/tpdJCDL06.pdf

[13] Bates, Marcia; Deborah N. Wilde; Susan Siegfried. An analysis of search terminology used by humanity scholars: The Getty Online Searching Project Report Number 1. // Library Quarterly 61,1(1993), 61-82.

[14] H. R. Tibbo. Abstracts, online searching, and the humanities: an analysis of the structure and content of abstracts of historical discourse. PhD thesis, University of Maryland, 1989.

[15] Gill, Tony. Building semantic bridges between museums, libraries and archives: The CIDOC Conceptual Reference Model. // First Monday: peer reviewed journal on the Internet. 9, 5(May 2004). Available at: http://firstmonday.org/issues/issue9_5/gill/index.html

[16] Functional Requirements for Bibliographic Records, Final Report / IFLA Study Group on the Functional Requirements for Bibliographic Records. München : K.G. Saur, 1998. (UBCIM Publications, New Series ; v. 19). Available at: http://www.ifla.org/VII/s13/frbr/frbr.pdf

"genealogists and family historians who are 50+ years of age" as the main user group of the digitised newspapers collections.[17] Republished on the Internet, digitised newspapers reach a broader audience that consists of "physically absent, geographically dispersed and culturally unbounded reader populations".[18] A. Smolczewska Tona stated that "investigating the current and potential user information needs and behavior in the digital context" requires an appropriate data collection method.[19] Web log analytics is considered as an unobtrusive method of collecting data that documents user interaction with the system and its content. Available website transaction log files analyses (e.g. Jansen, Spink and Saracevic, 2000[20]; Spink et al., 2001[21]; Jansen and Spink, 2006[22] etc.) are useful sources of information on Internet users' search habits. The results of the studies confirm that Web queries are short and that the classification schemes of *informational*, *navigational*, and *transactional* queries vary across topics.[23] What are the characteristic of digital libraries users' search queries? Digital libraries transaction log files analyses (like for instance Jones, Cunningham, McNab, Boddie, 2000[24]; Europeana, 2011[25], Han, Jeong and Wolfram, 2014[26] etc.) are not available in great number. There are only a few studies that address this issue in a topic-specific database such as historical newspapers digital library. A. Smolczewska Tona analysed log files of the 19th century newspapers website (CaNu XiX project, City Library in Lyon) in 2010.[27] The aim of her research was to determine the content of user queries and use the results to enhance the website interface. Web query classification (CQ) - a method of user queries classification into

[17] Zarndt, Frederic; Brian Geiger; Robert Stauffer; Alyssa Pacy; Meredith Palmer; Joanna DiPasquale. Digital collections: if you build them, will they visit? // International Federation of Library Associations, World Library and Information Congress, Singapore, 2013. Available at: http://www.dlconsulting.com/wp-content/uploads/2013/10/2013-IFLA-Satellite-Zarndt-et-al-Marketing-cultural-heritage-digital-collectionsedt1.pdf

[18] Smolczewska Tona, Agnieszka. Combining web analytics and computational linguistics to enhance access to digital libraries: a case study. // Biblioteki, informacja, książka: interdiscyplinarne badania i praktyka w XXI wieku, 7(2010), p. 264.

[19] Smolczewska Tona, Agnieszka. Ibid. P. 268.

[20] Jansen, B. J.; A. Spink; T. Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the Web. // Information Processing and Management 36(2000), 207−227.

[21] Spink, Amanda; Dietmar Wolfram; B. J. Jansen; Tefko Saracevic. Searching the Web: the public and their queries. // Journal of the American Society for Information Science and Technology 52, 3(2001), 226-234.

[22] Jansen, Bernard J.; Amanda Spink. How are we searching the world wide web? A comparison of nine search engine transaction logs. // Information Processing and Management. 42(2006), 248-263.

[23] Jansen, Bernard J.; Danielle L. Booth; Amanda Spink. Determining the informational, navigational, and transactional intent of Web queries. // Information Processing and Management: an International Journal archive. 44, 3(2008), 1251-1266.

[24] Jones, S., S. J. Cunningham; R. McNab; S. Boddie. A transaction log analysis of a digital library. // International Journal on Digital Libraries, 3(2000), 152-169.

[25] Europeana: an evaluation of users, usage and information seeking behaviour derived from web-server log-files (October 2009-April 2011). Available at: http://ciber-research.eu/download/20110821-M3.1.2_eConnect_LogAnalysis.pdf

[26] Han, Hyejung; Wooseob Jeong; Dietmar Wolfram. Log analysis of academic digital library: user query patterns. // iConference 2014 Proceedings. Pp. 1002-1008.

[27] Smolczewska Tona, Agnieszka. Ibid.

predefined target categories can also provide information on user's intent as the "the need behind the query".[28]

A study of the user search queries in the Croatian Historical Newspapers Portal

The analysis of digitised newspapers user queries was conducted at the National and University Library in Zagreb (NSK) in 2012.[29] It was assumed that semantic analysis of the queries in the Croatian Historical Newspapers Portal would contribute to better understanding of the users' information needs, and offer a basis for more precise decisions in the enhancement of the newspapers digital library system, especially in the application of NER. The analysis of user queries was a part of a more extensive study on the functional granularity concept in the digital library information organisation.[30] Three entities (*content*, *institution* and *user*) have been considered as main factors that affect metadata granularity in the newspapers digital library. Among them, the users of the newspapers digital library were surely a less known category.[31] It is not just a local problem – user's interaction with a newspapers digital library system is not well explored at the moment.

Croatian 19th century newspapers at NSK

To find out more about the NSK's 19th century newspapers collection, the study of formal characteristics of the newspapers, as described in their bibliographic records, was conducted in 2012. A sample of 225 bibliographic records was used for the analysis.[32] The results of the study indicate that the majority of newspapers were published weekly (28.57%) and daily (13.39%)[33], almost half of the titles (49.02%) were published in Zagreb and the dominant formats were 45-50 cm (33%) and 31-40 cm (27%). There are 73 horizontal links between bibliographic records that relate a newspaper title with its bibliographic family (the most numerous (68%) are sequential relations). The study shows that 225 Croatian 19th century newspapers were published in 7 languages (86% of the titles are published in Croatian) using

[28] Broder, Andrei. A taxonomy of web search. // SIGIR Forum. 36, 2(2002). Available at: http://www.cis.upenn.edu/~nenkova/Courses/cis430/p3-broder.pdf. Broader classify web queries according to their intent into 3 classes: *navigational*, *informational* and *transactional*.

[29] The CaNu XiX study analyses log files of the local newspapers website, while the NSK study explores national newspapers collection log files. Also, the methods are different (automatic text analysis vs. manual examination of terms) as well as the period covered by the analyses (six months vs. one year).

[30] Klarin Zadravec, Sofija. Koncept funkcionalne granularnosti u organizaciji informacija digitalne knjižnice = A concept of functional granularity in digital library information organisation: doktorski rad. Zagreb : Sveučilište u Zagrebu, Filozofski fakultet, 2012.

[31] Previous studies of the Croatian historical newspapers were focused on the analogue newspapers collections and their users – mostly by scholars in the humanities and social sciences.

[32] The analysis was not meant to provide full insight into the Croatian 19th century newspapers publishing history.

[33] The frequency of the 49 tiles (21.28%) is *unknown* or *irregular*.

both Latin (including *Fraktur*) and Cyrillic scripts. A very small number of newspapers in the sample are illustrated (5.33%). The analysis also included other characteristics of the newspapers, such as format changes, types of available reproductions, number of unique copies, etc. Data gathered in the research are used in the decision-making process of the newspapers digitisation project.[34]

Croatian Historical Newspapers Portal

National and University Library in Zagreb launched the *Croatian Historical Newspapers Portal* in 2010. In 2011 it became a cooperative portal, gathering seven Croatian heritage institutions and 111 digitised newspaper titles issued from 1789 to 1940.

Fig. 1. *Historical Croatian Newspapers Portal title page*

The number of digitised pages (300 000) is the result of the available resources of the heritage institutions involved in newspapers digitisation. The Portal of historical newspapers drew more unique Web visitors (20 000 in 2013[35]) than any other digital content of NSK. The majority of visitors came from Croatia and neighbouring countries, spending about 4 minutes per visit on the Portal.[36] The user interface has a set of features that are designed to search and browse digitised newspapers (i.e. a search box on the title page[37], a list of digitised titles, a calendar, an interactive map of Croatian counties with links to the local newspapers metadata, etc.). The federation of digital collections on the Portal has determined the metadata granularity. *Title-level metadata* are linked with *collection-level metadata* of the

[34] Only the completely digitised collection will enable more precise studies on the Croatian 19th century newspapers publishing history.
[35] Source: Google Analytics.
[36] Ibid.
[37] A full text search box is also available on each digitised newspapers title page.

contributing institutions' newspapers collection and with *digitisation project metadata*. There is a negligible number of the *article-level metadata* and *graphic objects metadata* in the system. *Page-level metadata* are linked with digitised pages and their OCRed text. The lack of *issue-level metadata* has a negative effect on the exchange of metadata with other systems (e.g. Europeana). After five years of existence, the user interface needs significant improvement (i.e. new design, new newspapers viewer, new features – browsing newspapers by topic map etc.).

Semantic analysis of the search queries

To create conceptual framework for the *Croatian Historical Newspapers Portal* users' query categorisation, basic categories were chosen from the previous user studies[38], FRBR, CIDOC CRM and FAST. CIDOC CRM (2008) was consulted as "a formal ontology intended to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information."[39] FAST (Faceted Application of Subject Terminology) vocabulary includes eight distinct categories or facets: *personal names*, *corporate names*, *geographic names*, *events*, *titles*, *time periods*, *topics* and *form/genre*.[40] Other relevant digital collections transactional log file analyses and users' query categorisations (S. Jones et al.[41], O. Zavalina[42], A. Smolczewska Tona[43]) were also consulted. Before the semantic analysis of search query terms, six categories were established: *person*, *event*, concept, object, *time*, and *place*. During the categorisation process, two categories (*concept* and *object)* were merged in one (*topic*) and a new category (title) was added. The semantic analysis and categorisation of user queries were conducted using a sample of transactional log file that covered a one-year period (Jan 21, 2010 – Jan 21, 2011).[44] In the one-year period 12 227 unique visitors visited the home page of the Portal. A small percentage (18.14%) of visitors left the web site without any activity. 9.25% users searched the content of the Portal using a search box on the title

[38] Petras, Vivien; Ray R. Larson; Michael Buckland. Ibid.

[39] Definition of the CIDOC conceptual reference model / editors Nick Crofts et al. ; produced by the ICOM/CIDOC Documentation Standards Group, continued by the CIDOC CRM Special Interest Group. Version 5.0. December 2008. Available at: http://www.cidoc-crm.org/docs/cidoc_crm_version_5.0_Dec08.pdf

[40] FAST - Faceted Application of Subject Terminology. Available at: http://oclc.org/research/activities/fast.html

[41] Jones, S.; S. J. Cunningham; R. McNab; S. Boddie. A transaction log analysis of a digital library. // International Journal on Digital Libraries 3(2000), 152–169.

[42] Zavalina, Oksana. Collection-level user searches in federated digital resource environment. // Proceedings of the 70th ASIS&T Annual Meeting (Milwaukee WI, Oct. 18-25, 2007). Available at: http://hdl.handle.net/2142/8983

[43] Smolczewska Tona, Agnieszka. Ibid.

[44] The study sample consists of 365 text files documenting user queries on the portal home page within its (simple) full-text search option. The IP addresses in the log file (1219 unique IP addresses) have been marked with the serial numbers in order not to compromise the privacy of users of the portal. To protect the privacy of the users, the content of the queries (e.g., name of the person as subjects of user queries) will not be publicly available since there is a possibility of infringement of privacy.

page. The transaction log file was manually processed to extract all the search query strings – a total of 6843 queries.[45] The majority of queries (71.23%) contained one term.[46] The number of unique queries[47] was 1422. Unique queries were examined and classified manually. Data was coded using previously identified categories and, during the coding procedure, verification and correction were conducted. The meanings of terms in the queries were checked using relevant literature, including dictionaries, linguistic portals, Wikipedia and the content of the Portal. To check the data identified as "title", the *Historical Croatian Newspapers Portal* and NSK ILS were used. The constraints of the study were the following: a) incomplete sample of the users search terms – due to technical issues, log files of the simple and advanced search on the *search web page* of the Portal were not preserved and analysed[48], b) possible mistakes in categorisation due to polysemy of the search terms.[49]

Findings and discussion[50]

Figure 2 shows the distribution of unique queries within six categories: *person* (31.15%), *place* (24.75%), *topic* (24.40%), *title* (17.79%), *date* (1.33%) and *event* (0.77%).[51]

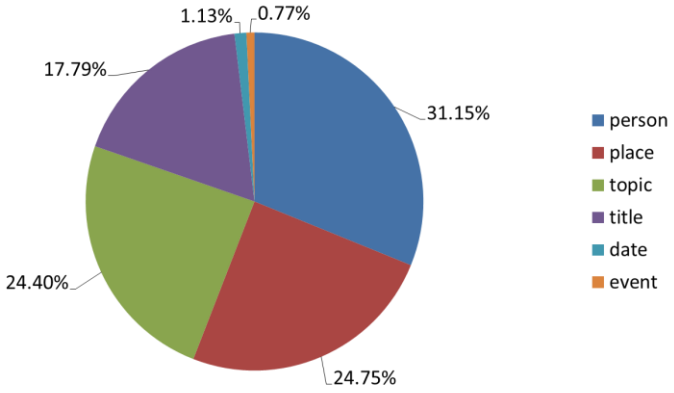

Fig. 2. *Distribution of unique queries across categories*

[45] Spink, Amanda et al., Ibid.: *Query*: a set of one or more search terms; it may include advanced search features, such as logical operators and modifiers.
[46] Spink, Amanda et al., Ibid.: *Term*: any unbroken string of alphanumeric characters entered by a user.
[47] Spink, Amanda et al., Ibid.: *Unique queries* are all *differing* queries entered by one user in one session; the differing queries could be modifications of the previous query or entirely new queries.
[48] The analysed queries come from the simple search feature on the title page.
[49] Each query was assigned to one category only.
[50] Complete results of the study are available in: Klarin Zadravec, Sofija. Koncept funkcionalne granularnosti u organizaciji informacija digitalne knjižnice = A concept of functional granularity in digital library information organisation: doktorski rad. Zagreb : S. Klarin Zadravec, 2012.
[51] Each category consists of a number of subcategories, not shown here.

The distribution of the search terms shows the domination of 4 categories (*person*, *place*, *topic* and *title*) while 2 categories (*date* and *event*) are represented in a small percentage.

The proportion of category *person* (31.15%) shows that the majority of the users prefer *name searching* while starting their search on the portal. Most of the queries in this category are Croatian surnames and a few names may refer to the people from public, political and cultural life of Croatia. The results are in accordance with the growing interest of Croatian users for family history and genealogical research in archives[52] as well as with the results of recent studies of the digital newspapers libraries.[53]

The proportion of the category *places* in the queries distribution is 24.75%. Geographical names are assigned to several subcategories (i.e. *state*, *town*, village, *river*, *island*, *mountain* etc.) that show users' interest in local toponyms. It should be noted that a part of the sample represents toponyms related to the neighbouring countries. A small number of the query terms were *historical place names* (e.g., *Abbazia*, *Sissek*). This indicates the existence of a subgroup of users who adjust the search term to the historical text database.

The category *topic* (24.40%) consists of two FRBR categories – *concept* and *object*. It is a category that covers a broad range of topics. Some of the queries in this category were *archaisms* (e.g., *etažba*, *mudroskup*, *terčalac*, *bilinska hrana*). This confirms the existence of a subgroup of users who adjust the search term to the historical text database. A few contemporary expressions were used as search queries (e.g., *marketing communications*, *privatisation*, *e-business, NLO*[54]) in this category. It indicates that some users are not familiar with the content of the Portal (i.e. historical material).

A considerable amount (17.79%) of titles in search queries reflects users' interest in the *title* searching. Among 253 unique queries in this category, the largest group of queries consists of *serials* (242) while other queries are identified as titles of other types of publications/works (*book* (7), *law* (2), *poem* (1), *film* (1)). Several titles of contemporary newspapers and journals were among titles of serials found in the queries (e.g., *Feral Tribune*, *Modra lasta*, *Plavi oglasnik*, *Start, Večernji list 2000. godina* itd.). This also indicates that some users are not familiar with the content of the Portal (i.e. type of material).

[52] Ćosić, Stjepan. Hrvatska traži povrat 765 fondova i zbirki. // Hrvatsko slovo, March 27, 2009.
[53] Zarndt, Frederic; Brian Geiger; Robert Stauffer; Alyssa Pacy; Meredith Palmer; Joanna DiPasquale. Ibid. P. 3.
[54] In Croatian: *marketinške komunikacije, privatizacija, e-poslovanje.*

The results of the analysis show low level of *event* (0.77%) and *date* (1.13%) searching. It is possible to assume that low interest in the *date* searching is caused by the availability of semi-structured *date* searching in the calendar on the title page of the Portal, even though it does not bring full coverage of *time* as a subject in the text of the newspapers. Without more detailed information of the users, it is not possible to interpret the results relating to the *event* searching. Among a small number of queries relating to *events*, there are some contemporary events (i.e., *the fall of the Berlin Wall*, *the Zagreb rocket attacks*[55], *9.11.2001 world trade center*).

Conclusion

The article brings a broad summary of the *Croatian Historical Newspapers Portal* users' query analysis. From the sample of queries collected in one-year period, the domination of 4 search query categories (*person*, *place*, *topic* and *title*) is evident while 2 categories (*date* and *event*) are represented in a small percentage. The results also show that the searching habits of Portal users are similar to other Internet users (a small number of search terms in a query) and to some extent the specific habits of a group of researchers of historical sources (represented semantic categories, customization of the search terms to the historical material). The content of the Portal users' search queries is *informational*, even though, in this context, the content of the category *title* can be understood as *navigational* or *transactional*, indicating users' need to obtain the precise resource – a digitised newspapers' title page.

The categorisation of the search terms in the queries of the *Croatian Historical Newspapers Portal* was just a first step in understanding what the users are searching for. Further research [56] and other methods should provide more precise information on their real information need. As Petras, Larson and Buckland pointed out: "Chronological, geographical and biographical data lend themselves naturally to being connected: an event is associated with a place, a time and potentially with particular people; places are associated with different events and people; and individual people are also associated (in a variety of ways) with different places and events. One can foresee a plethora of relations and possible search questions that a truly interconnected system should be able to answer".[57]

[55] In Croatian: *pad berlinskog zida, raketiranje zagreba 1995.*
[56] The use of contemporary expression in the queries may indicate the users' interest for current Croatian publications in digital form but also their insufficient knowledge of the content of the Portal and, probably, inadequate marketing of the digital newspapers collection.
[57] Petras, Vivien; Ray R. Larson; Michael Buckland. Ibid.

To paraphrase Jensen and Booth[58], providing better access and retrieval of the historical text is at the heart of designing successful historical newspapers digital library. It is expected that continued user query analysis will contribute to better understanding of the Portal users' motivations and searching habits and lead to the improvement of the digital newspapers system functions. Implementation of NER, especially for the most common search terms categories (*people*, *place*), can provide functional semantic granularity and enhance the quality of the search.

References

Allen, Robert B.; I. Waldstein; W. Zhu. Automated processing of digitized historical newspapers : identification of segments and genres. // International Conference on Asian Digital Libraries, Hanoi, Vietnam, 2008. Pp. 380-387. Available at: http://boballen.info/PAPERS/NewsGenres.pdf

Allen, Robert B.; John Schallow. Metadata and data structure for the historical newspapers digital library. // Proceedings of the 8th international conference on Information and knowledge management CIKM 99. Available at: http://boballen.info/PAPERS/META/meta.pdf

Allen, Robert B. Improving access to digitized historical newspapers with text mining, coordinated models, and formative user interface design. // IFLA Newspaper Section Meeting, New Delhi, February 2010. Available at: http://boballen.info/RBA/PAPERS/IFLA2010/iflaDelhi.pdf

Ahonen, Eeva; Eero Hyvönen. Publishing historical texts on the semantic web: a case study. Available at: http://www.seco.tkk.fi/publications/2009/ahonen-hyvonen-historical-texts-2009.pdf

Bates, Marcia; Deborah N. Wilde; Susan Siegfried. An analysis of search terminology used by humanity scholars: The Getty Online Searching Project Report Number 1. // Library Quarterly 61,1(1993), 61-82.

Broder, Andrei. A taxonomy of web search. // SIGIR Forum. 36, 2(2002). Available at: http://www.cis.upenn.edu/~nenkova/Courses/cis430/p3-broder.pdf

Ćosić, Stjepan. Hrvatska traži povrat 765 fondova i zbirki. // Hrvatsko slovo, March 27, 2009.

Definition of the CIDOC conceptual reference model / editors Nick Crofts et al. ; produced by the ICOM/CIDOC Documentation Standards Group, continued by the CIDOC CRM Special Interest Group. Version 5.0. December 2008.
Available at: http://www.cidoc-crm.org/docs/cidoc_crm_version_5.0_Dec08.pdf

De Jong, Francisca; Henning Rode; Djoerd Himjestra. Temporal Language Models for the Disclosure of Historical Text. Available at: http://eprints.eemcs.utwente.nl/7266/01/db-utwente-433BCEA2.pdf

Europeana : an evaluation of users, usage and information seeking behaviour derived from web-server log-files (October 2009-April 2011). Available at: http://ciber-research.eu/download/20110821-M3.1.2_eConnect_LogAnalysis.pdf

[58] Jansen, Bernard J.; Danielle L. Booth. Ibid.

FAST - Faceted Application of Subject Terminology.
Available at: http://oclc.org/research/activities/fast.html

Functional Requirements for Bibliographic Records, Final Report / IFLA Study Group on the Functional Requirements for Bibliographic Records. München : K.G. Saur, 1998. (UBCIM Publications, New Series ; v. 19). Available at: http://www.ifla.org/VII/s13/frbr/frbr.pdf

Gill, Tony. Building semantic bridges between museums, libraries and archives: The CIDOC Conceptual Reference Model. // First Monday: peer reviewed journal on the Internet. 9, 5(May 2004). Available at: http://firstmonday.org/issues/issue9_5/gill/index.html

Han, Hyejung; Wooseob Jeong; Dietmar Wolfram. Log analysis of academic digital library: user query patterns. // iConference 2014 Proceedings. Pp. 1002-1008.

Han, Myung-Ja. Metadata with levels of description: new challenges to catalogers and metadata librarians. // International Federation of Library Associations, World Library and Information Congress, Helsinki, 2012. Available at: http://conference.ifla.org/past-wlic/2012/80-han-en.pdf

Impact project (Improving Access of Text). Available at: http://www.impact-project.eu/

Jansen, B. J.; A. Spink; T. Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the Web. // Information Processing and Management 36(2000), 207−227.

Jansen, Bernard J.; Amanda Spink. How are we searching the world wide web? A comparison of nine search engine transaction logs. // Information Processing and Management 42(2006), 248-263.

Jansen, Bernard J.; Danielle Booth. Classifying Web queries by topic and user intent. CHI 2010: Work-in-Progress, April 14–15, 2010, Atlanta, GA.
Available at: http://faculty.ist.psu.edu/jjansen/academic/jansen_user_intent.pdf

Jansen, Bernard J.; Danielle L. Booth; Amanda Spink. Determining the informational, navigational, and transactional intent of Web queries. // Information Processing and Management: an International Journal archive. 44, 3(2008), 1251-1266.

Jones, Alison. The many uses of newspapers.
Available at: http://dlxs.richmond.edu/d/ddr/docs/papers/usesofnewspapers.pdf

Jones, Steve; Sally Jo Cunningham; Rodger McNab; Stefan Boddie (2000). A transaction log analysis of a digital library. // International Journal on Digital Libraries 3(2000), 152–169.

Klarin Zadravec, Sofija. Koncept funkcionalne granularnosti u organizaciji informacija digitalne knjižnice = A concept of functional granularity in digital library information organisation: doktorski rad. Zagreb : Sveučilište u Zagrebu, Filozofski fakultet, 2012.

Petras, Vivien; Ray R. Larson; Michael Buckland. Time period directories: a metadata infrastructure for placing events in temporal and geographic context. // Joint Conference on Digital Libraries - JCDL Workshop, 2006. Available at: http://metadata.sims.berkeley.edu/tpdJCDL06.pdf

Portal Stare hrvatske novine = Croatian Historical Newspapers Portal. Available at: http://dnc.nsk.hr/newspapers/Default.aspx

Spink, Amanda; Dietmar Wolfram; B. J. Jansen; Tefko Saracevic. Searching the Web: the public and their queries. // Journal of the American Society for Information Science and Technology 52, 3(2001), 226-234.

Smolczewska Tona, Agnieszka. Combining web analytics and computational linguistics to enhance access to digital libraries: a case study. // Biblioteki, informacja, książka: interdiscyplinarne badania i praktyka w XXI wieku, 7(2010), 264-278.
Available at: http://skryba.inib.uj.edu.pl/wydawnictwa/e07/n-tona.pdf

Tibbo, H. R. Abstracts, online searching, and the humanities: an analysis of the structure and content of abstracts of historical discourse. PhD thesis, University of Maryland, 1989.

Zarndt, Frederic; Brian Geiger; Robert Stauffer; Alyssa Pacy; Meredith Palmer; Joanna DiPasquale. Digital collections: if you build them, will they visit? // International Federation of Library Associations, World Library and Information Congress, Singapore, 2013. Available at: http://www.dlconsulting.com/wp-content/uploads/2013/10/2013-IFLA-Satellite-Zarndt-et-al-Marketing-cultural-heritage-digital-collectionsedt1.pdf

Zavalina, Oksana. Collection-level user searches in federated digital resource environment. // Proceedings of the 70th ASIS&T Annual Meeting (Milwaukee WI, Oct. 18-25, 2007). Available at: http://hdl.handle.net/2142/8983