Reusing the Model and Components of an IIR Study for Perceived Effects of OCR Quality Change

Kimmo Kettunen
School of Humanities, Finnish Language and Cultural Research
University of Eastern Finland
Joensuu, Finland
kimmo.kettunen@uef.fi

Heikki Keskustalo
Faculty of Information Technology and Communication Sciences
Tampere University
Tampere, Finland
heikki.keskustalo@tuni.fi

Birger Larsen
Department of Communication and Psychology
The Faculty of Humanities,
Aalborg University
Aalborg, Denmark
birger@hum.aau.dk

Tuula Pääkkönen
University of Helsinki
The National Library of Finland
Mikkeli, Finland
tuula.paakkonen@helsinki.fi

Juha Rautiainen
University of Helsinki
The National Library of Finland
Mikkeli, Finland
juha.rautiainen@helsinki.fi

ABSTRACT

Historical newspapers are increasingly accessed digitally for different purposes both by professional and lay users. These ever-growing historical collections are usually formed by utilizing Optical Character Recognition (OCR), which may introduce noise to the texts. This subsequently leads to compromised information retrieval (IR) performance and user understanding. The effect of OCR noise on IR performance has been studied earlier by utilizing artificially degraded OCR quality texts (see, e.g., [2, 15]), test collection containing documents with authentic low OCR quality [12], or by gathering end-user impressions [23]. However, it remains challenging to measure how the user's subjective perception is affected by the amount of OCR noise remaining in the documents. Recently, the National Library of Finland has set up an experimental system which allows studying this issue. The system allows presenting each underlying historical document as two alternatives – either based on the baseline OCR quality, or on the new, improved OCR quality. This set up facilitates studying the effects of OCR quality changes on the user's subjective perception of the document.

Following Gäde et al. [8] we describe in this paper the research design, infrastructure, and research data utilized in a recent user experiment of Kettunen et al. [19] entailing thirty-two test subjects performing simulated work tasks [4] and discuss the prospects of reuse of the experimental components of the study. So far, the system has been used in one experiment in which the subjects performed simulated tasks. However, the research design and its general model could be utilized in the future to study the effects of OCR quality on professional settings entailing historians performing naturalistic phases of their research tasks.

CCS CONCEPTS

Document representation Users and interactive retrieval Evaluation of retrieval results Task models Search interfaces

KEYWORDS

OCR quality, Interactive Information Retrieval, Evaluation, Simulated Work Task, Historical newspaper collections, User Study, Resource reuse

ACM Reference format:

Kimmo Kettunen, Heikki Keskustalo, Birger Larsen, Tuula Pääkkönen and Juha Rautiainen. 2022. Reusing the Model and Components of an IIR Study for Perceived Effects of OCR Quality Change. In *Proceedings of Third Workshop on Building towards Information Interaction and Retrieval Resources Re-use (BIIRR 2022). ACM, New York, NY, USA, 7 pages.* https://doi.org/XXX

1 Introduction

It is well known that OCR noise present in digitized historical documents disturbs end user perception of documents. However, this impinging on the desired access is difficult to study [6, 27]. In this paper, we describe a research design intended to allow studying this issue and discuss the model and its components from the point of view of reuse in research.

Digitized historical newspaper collections are produced increasingly in different parts of the world and their usage is expected to increase in the future. Access to information in these collections is valuable to various stakeholders such as professional historians, journalists, teachers, and ordinary citizens. To create these collections, the visual image of the original printed

newspapers is transformed automatically into a stream of text via Optical Character Recognition (OCR). The quality of the resulting text varies, but usually it contains at least some misinterpreted characters, which potentially cause problems for both information retrieval (IR) and user understanding. The effects of OCR quality on IR performance have been addressed by utilizing artificially degraded OCR texts [2, 15], laboratory-style IR experiments [12], and considering the user viewpoint in the context of usability in historical research [23-24, 26]. However, controlled studies in which real test subjects are used to study the effects of varying OCR quality on users are rare. The only exception we are aware of is a recent study by Kettunen et al. [19]. In this study, the subjective relevance feedback of the test subjects (N=32) was collected via a search system interface of the National Library of Finland while the OCR quality of the historical newspaper clippings was systematically varied. The same underlying document was presented for the user randomly based on two alternative versions, which differed in OCR quality. The users did not know about OCR quality differences in the documents.

In the following we use the resource type classification presented in Gäde et al. [8] and divide the resources of the retrieval system to three types: 1) research design, 2) research infrastructure, and 3) research data. In Gäde et al. research design is defined "as methods and techniques used to collect and analyse empirical data". Research infrastructure is defined mainly in relation to the technical infrastructure that is needed to carry out an interactive information retrieval study. Research data can be broadly defined as "any data that has been collected, observed, generated or created during or as results of the research process" [8]. Gäde et al. discuss the notion of reusability with regards to different current research articles. After some discussion they define a broad sense of reusability for the IIR community: reuse is "use of research data, research design or infrastructure for more than an individual purpose".

In the remainder of this paper, we describe resources used in the OCR quality user study and consider them from the point of view of adapting them for reuse. Section 2 discusses the issues of research design and its reuse, Section 3 discusses issues related to infrastructure and its reuse, and Section 4 discusses the research data with its reuse. Section 5 concludes the paper by discussion and conclusions.

2 Research Design

2.1 General Description

The user experiment of Kettunen et al. [19] entailed recruiting test persons to perform simulated work tasks to retrieve and evaluate historical newspaper articles of varying OCR quality. During the experiment (NewsEye query hackathon 2021) thirty-two (32) test subjects used a web search engine interface to search articles from the Finnish historical newspaper Uusi Suometar (1869–1918) based on some 86,000 original pages auto-segmented into 1.45 million articles called clippings. Thirty (30) topics of Finnish and world history for the period 1871–1918 were used for the search tasks. Thirty fixed queries were formed by the authors to allow controlled experiments regarding the set of documents retrieved for the user for her subsequent subjective evaluations. A simulated work task description was utilized in the experiment to explain the user tasks (the background story) and how to use a graded relevance scale (0-3) to express the usefulness of the observed newspaper article in the light of the user task [4].

Two different types of tasks were performed by the test persons. In the first type each test subject performed six pre-formulated queries (fixed queries) and evaluated the relevance of the top-10 documents retrieved. In the second task in a separate query session the test subjects could freely formulate their own queries based on the topic description and utilize background information gathered from the web. The interactive retrieval system balloted six formerly unseen topics for each user selected from the set of 30 topics for both query sessions. The retrieval system contained two versions of the same newspaper clippings data – with baseline OCR quality (old), and improved OCR quality (new). In all cases the query engine retrieved and ranked the clippings based on the improved OCR quality. However, the query environment balloted between the two OCR quality versions of the same underlying document presented for the user, while the user did not know about the quality difference. Subsequently we could evaluate the effect of OCR quality changes on the subjective perceptions based on the relevance evaluations expressed by the test subjects. Based on all search sessions of the 32 test subjects, 3893 user evaluations were gathered. The fixed query experiment and its results are reported in [19].

The research question in [19] was: Does the variation of OCR quality in historical newspaper clippings affect their perceived usefulness, as measured by the test subjects' graded relevance assessments in a simulated work task situation?

In accordance with Gäde et al. [8] following items of the model belong to research design: topics and queries, simulated tasks, recruitment of participants of the study, analysis methods, evaluation measures and significance tests. These are described next.

2.2 Topics and Queries

Search topics and queries are an important part of all (I)IR experiments [5, 15-16]. Thirty topics of Finnish and world history for the period 1871–1918 were created for the search task. The topics were created using history timelines from two popular history encyclopaedias: *Suomen historian pikkujättiläinen* [28] ('A small encyclopaedia of Finnish history') and *Maailmanhistorian pikkujättiläinen* [29] ('A small encyclopaedia of world history'). Final topic descriptions were based on Finnish Wikipedia articles related to the topics. The topics cover the time frame of the historical collection of Uusi Suometar, beginning from the 1870s and ending in 1918. Topics cover both domestic and foreign news, but the demarcation line between foreign and domestic news is not

always sharp, and some topics could be classified as both. For the topics we also created 30 short pre-formulated queries which were used in the first query session of the experiment.

The topic descriptions can be shared via public repositories; the pre-formulated queries are reported in Kettunen et al. [19]. This component can be adapted by creating new topics and queries as necessary, acknowledging the type of target data and its expected use [5, 24]. In the historical newspaper context, it might also be advantageous if a professional historian would take part in topic creation.

2.3 Simulated Tasks

Simulated work tasks are frequently used in IIR experiments to provoke information needs for the task participants [4, 16]. Simulated task descriptions were used here to inform the test subjects that they use the information retrieval system of digitized newspaper clippings to write an article of historical events in Finland or the world in the time span of 1869-1918. Participants needed to evaluate results of the search in relation to this information need using graded evaluation scale of 0–3, which was briefly explained in the instructions. The background story and evaluation instructions used in [19] is replicated in Table 1 as a translation.

Background story Imagine that you are writing an article related to topics in history of Finland or world history at the end of 19th century or the beginning of 20th century. Evaluate quality of the clippings you get as search results. Evaluate the quality of each clipping from the viewpoint, how it helps you to proceed with your article writing. **Evaluation of the search results (graded relevance scale of 0-3)[1]** 3. The clipping deals with the topic very broadly and its information content corresponds well with the task. The clipping helps well in accomplishing your task. 2. The clipping deals partially with the task or touches it. The content of the clipping helps to some extent in accomplishing your task. 1. The clipping does not deal with the actual topic but helps to find better search terms and to limit the topic somehow. It helps indirectly in accomplishing your task. 0. The clipping is wholly off topic and does not even help to formulate new queries. This clipping brings no benefit in accomplishing your task.

Table 1. The background story and evaluation instructions given to the participants

[1] Note that the evaluation instructions advise the participant to consider how well the clipping helps in accomplishing the task described in the background story - thus going beyond topicality.

This component can be adapted by creating variations of the specific tasks described by Kettunen et al. [19] in corresponding settings, possibly with the help of professional historians.

2.4 Recruitment of Participants

IIR experiments need a large enough user group so that the results of the study are generalizable, and recruitment of suitable participants may be sometimes hard [16]. We were able to recruit 32 participants for the evaluation task – students from the courses Information Retrieval and Language Technology and Information Retrieval Methods at the Tampere University, Faculty of Information Technology and Communication Sciences and three teachers of information science. Choice was based mainly on the ease of getting a large enough group (at least 30) to perform the tasks.

The group consisted mainly of students, which is not optimal. A large enough group of historians would have been hard to recruit, however, as e.g., experience of Kumpulainen and Late [23] shows. They were able to find 13 participants who use Finnish historical newspapers in their work. Kumpulainen and Late examined how the researchers used digital newspaper collections in their work and what were the obstacles they encountered using the collections.

This component could be adapted, e.g., by utilizing professional historians as test persons. This would be valuable as the personal interests of users are important [25]. However, the number of historians available as test persons may be limited, which must be considered in the beginning when the research is initially planned.

2.5 The Data Collection Protocol

During search sessions the system's search interface was used to collect user-given relevance information to answer the research questions. The test subjects evaluated the retrieved documents based on the simulated work tasks. Each participant was guided to perform six tasks (topics), both with pre-formulated and self-formulated queries. Out of the 32 users' work with pre-formulated queries we got 1861 evaluations, and out of the self-formulated queries 2032 evaluations. The information we gathered from the query sessions is described in detail section 3.6.

In the experiment described in [19] the system interface collected document-level relevance feedback information given by the test subjects during search sessions to facilitate answering the research questions. The data collection protocol of the experiment was intentionally simplified. This component can be adapted to collect other desired information items, such as time stamps of the clicks or other user feedback.

2.6 Analysis Methods

The research question of the study was whether OCR quality differences - baseline versus improved (old and new) OCR - affects the perceived usefulness of the newspaper clippings. To answer this, the relevance scores given by test subjects for the perceived documents in a simulated work task were analysed.

User-given relevance scores expressing the usefulness of the historical documents with respect to the simulated task were at the focus of investigation in [19]. This component can be adapted by including other types of measurements, such as addressing time stamps of activities, or interview data. While in Kettunen et al. [19] test subjects focused on the informational contents of data items (i.e., newspaper articles), it is noteworthy that the simulated tasks can be modified to request the test subjects to focus on other specific aspects, e.g., to evaluate the quality of the clipping boundaries, or quality of the overall presentation interface itself.

2.7 Evaluation Measures and Significance Tests

The analysis focused on the mean average evaluation scores given for the observed documents based on the gain values (0-3) given by the test subjects. The main evaluation means was to compare gain values given for the baseline OCR and the improved OCR. Wilcoxon's signed rank test was used to study statistical significance of the observed differences. First, the effects of the OCR quality on relevance judgments at the level of the individual documents retrieved was compared. Moreover, the effects of OCR quality on the average of cumulated gain among the top-10 documents retrieved for the 30 topics was compared. In both cases the improved OCR quality was related to higher gain values, but the difference was statistically significant only at the level of individual documents ($p=0.002$, [7]). The difference was not statistically significant ($p=0.10$) based on the mean average gain values for 30 topics.

Graded relevance scale (0-3) and Wilcoxon's signed rank test for measuring statistical significance were used in evaluation of [19]. Graded relevance (gain value) was used as a measure of success [13]. Other evaluation measures can be used in addition to rank-based measures utilizing individual queries (see, e.g., [21: 191-215]). For example, measures allowing use of multi-query sessions [14], or time-based measures [1], would be suitable. This component can also be adapted by complementing it with other measurements, such as durations of interactions, click measures, and interviews of test subjects, together with appropriate significance tests.

3 Research Infrastructure

As research infrastructure we handle the following items in accordance with Gäde et al. [8]: data, OCR software, segmentation software, search software, user interface and user log.

3.1 Target Data

The target data collection consists of the whole history of one newspaper, Uusi Suometar, between the years 1869 and 1918, entailing 86 000 pages and about 306.8 million words [20]. Uusi Suometar was at the time of its publication one of the most important Finnish language newspapers in Finland, where newspapers were published in two languages, Finnish and Swedish.

The data of the collection is out of copyright and could be thus published. The data of Uusi Suometar is historically realistic data of the period and of importance for real library users, too, according to user statistics of NLF. However, Finnish as target language limits its reuse per se. Subsequently, there is a need to adapt this component by selecting different target data. In adapting this component, the critical properties of the target data need to be considered, e.g., the size of the collection, and its timeliness.

3.2 OCR Software and OCR Quality

The baseline OCR for Uusi Suometar was performed using a series of ABBYY FineReader® products. Improved optical character recognition for the whole history of Uusi Suometar was achieved with Tesseract v.3.0.4.01. Improvement to the earlier quality in recognition of words is approximately 15% units as a mean over the whole period. On average 83% of the words of the re-OCR'ed newspaper data were recognized with automatic morphological analyzers, and the recognition rate varied from ca. 78 to 88% over the 49 years. For the baseline Optical Character Recognition, the mean word recognition rate was 68.2% [20].

The negative effects of poor OCR quality have been pinpointed during various information activities with historical digital newspapers [23-24]. Poor OCR quality has been noted also by digital humanists e.g., in Pfanzelter et al. [26] and Jarlbrink and Snickars [11]. The OCR software used in [18] is representative of library production system quality. In adapting the software components, utilizing state-of-the-art OCR helps justify the realism and subsequent validity of the subjective relevance assessments gained by the test persons.

3.3 Segmentation Software

In the original digitization of Uusi Suometar article structure is not present on the pages. Articles or clippings that were extracted automatically from the pages of Uusi Suometar with a trained machine learning model of software PIVAJ [9–10] were used for the user study. The search index contained the title and the textual contents of the specific article area taken from the OCR of ALTO XML of the whole page, either from the original OCR page or from the re-OCR'ed page. The size of the original OCR quality index is 9.82 Gb and the size of the re-OCR'ed index 9.04 Gb. Both indexes contain 1 459 068 clippings.

The ability to segment the image of newspaper page into separate articles is not yet a pervasive feature of historical collections [3], but demand for it is increasing [23]. In Kettunen et al. [19],

automated segmentation was produced with an experimental software, resulting in 1.47 million clippings ("articles"). However, noisy article boundaries are produced during the process, which may negatively affect the user perception. Subsequently, improving the segmentation is important. Note that the overall setting described in this paper fits well with studying this issue - comparing the effects of segmentation styles from the user viewpoint.

3.4 Search Software and Search Index

Participants of the evaluation task performed their task using the query engine Elastic search (https://www.elastic.co/), version 7.3.2, which is the background search engine of the library's presentation system. The index of the newspaper collection's database is lemmatized, i.e., it contains base forms of the words, which is crucial as Finnish is a highly inflected language [12]. The query engine always searched for the results of queries in the new optical character recognition version of the database and ranked the results according to these. A typical off-the-shelf search engine with embedding into the library's presentation system was thus used in the experiment.

In general, maintaining and reusing realistic software components is a hard problem [25]. In our case, e.g., we have utilized NLF's document presentation system, different OCR software, an experimental page segmentation software and a search engine. The query interface and user logging needed to be adapted into this environment. Adapting these components for reuse requires cooperation with the software maintainers or providers in practice.

3.5 The Query Interface

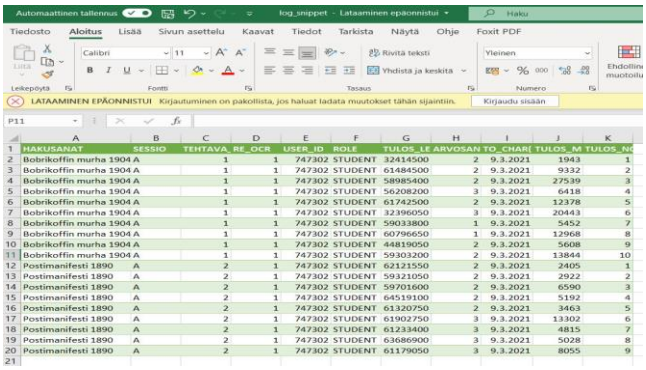
A simple web query interface was developed for the experiment. Six topics for search and evaluation were presented for each user in the query form, one at a time. Figure 1 shows the query interface after a pre-formulated query has been performed and 35 results retrieved, out of which 10 top results are shown for the user for evaluation. Topic is shown in the light blue rectangle, and the short query is under the topic description in pink text. Users were able to open the clippings and read the result documents and give their evaluations using the grades 0-3 with radio buttons, shown on the right corner of the purple rectangle.

Figure 1. Screenshot of the Query Interface

The search interface is considered as one of the most crucial parts of IIR experiments in different multiyear IIR tracks [25]. A simple search interface, specifically created for this use, was developed by the outsourced developer of the library. As such the interface could be reused as a model for similar tasks.

3.6 User Management and Interaction Logging

Transaction logging is one of the means for collecting data in an IIR evaluation [16]. Data out of the user sessions was collected to a query transaction log with the following data: A) query words B) session information C) number of the topic D) optical character recognition quality in the results (0 for the old and 1 for the new) E) user id F) role of the user (student or teacher) G) id number of the result clipping H) user-given evaluation result on the scale of 0-3 I) date and time of the session J) size of the clipping in characters K) rank (1-10) of the result clipping in the result list. Figure 2 depicts the query log as an Excel sheet.



HAKUSANAT	SESSIO	TEHTAVA	RE_OCR	USER_ID	ROLE	TULOS_LE	ARVOSAN	TO_CHAR(	TULOS_M	TULOS_NU
Bobrikoffin murha 1904	A	1	1	747302	STUDENT	32414500	2	9.3.2021	1943	1
Bobrikoffin murha 1904	A	1	1	747302	STUDENT	61484500	2	9.3.2021	9332	2
Bobrikoffin murha 1904	A	1	1	747302	STUDENT	58985400	2	9.3.2021	27539	3
Bobrikoffin murha 1904	A	1	1	747302	STUDENT	56208200	3	9.3.2021	6418	4
Bobrikoffin murha 1904	A	1	1	747302	STUDENT	61742500	2	9.3.2021	12378	5
Bobrikoffin murha 1904	A	1	1	747302	STUDENT	32396050	3	9.3.2021	20443	6
Bobrikoffin murha 1904	A	1	1	747302	STUDENT	59033800	1	9.3.2021	5452	7
Bobrikoffin murha 1904	A	1	1	747302	STUDENT	60796650	1	9.3.2021	12968	8
Bobrikoffin murha 1904	A	1	1	747302	STUDENT	44819050	2	9.3.2021	5608	9
Bobrikoffin murha 1904	A	1	1	747302	STUDENT	59303200	2	9.3.2021	13844	10
Postimanifesti 1890	A	2	1	747302	STUDENT	62121550	2	9.3.2021	2405	1
Postimanifesti 1890	A	2	1	747302	STUDENT	59321050	2	9.3.2021	2922	2
Postimanifesti 1890	A	2	1	747302	STUDENT	59701600	2	9.3.2021	6590	3
Postimanifesti 1890	A	2	1	747302	STUDENT	64519100	2	9.3.2021	5192	4
Postimanifesti 1890	A	2	1	747302	STUDENT	61320750	2	9.3.2021	3463	5
Postimanifesti 1890	A	2	1	747302	STUDENT	61902750	3	9.3.2021	13302	6
Postimanifesti 1890	A	2	1	747302	STUDENT	61233400	3	9.3.2021	4815	7
Postimanifesti 1890	A	2	1	747302	STUDENT	63686900	3	9.3.2021	5028	8
Postimanifesti 1890	A	2	1	747302	STUDENT	61179050	3	9.3.2021	8055	9

Figure 2. Screenshot of the Query Log

Minimal information about the users is collected in the transaction logs, and no GDPR issues arise from this. However, session lengths, duration of interactions, clicks and other such measures could be added to the collected data.

4 Research Data - User Sessions

The results of the experiment are research data that was accumulated in the search sessions. From the search sessions of one user group 3983 evaluations for query results were collected using the log format described in section 3.6. From the point of view of the reuse, the gathered data from user sessions is of high value, as it enables the analysis of the query sessions from many viewpoints. So far only basic analysis of the results with pre-formulated queries have been reported in Kettunen et al. [19].

Moreover, evaluation measures can be tuned to reflect the point of view of specific users modelled. We utilized gain values (0, 1, 2, 3) and top-10 documents retrieved, but other weighting schemes, such as (0, 0, 0, 1) or (0, 1, 10, 100), and thresholds for top-N documents retrieved (e.g., N=3 or N=5) could be used, to reflect different user preferences. Moreover, the discount factor can be used to model user preferences via DCG-based measures [13].

The result data does not include any personal information, that would hinder its publication. Thus, it is reusable.

5 Discussion and Conclusion

Kettunen et al. [19] showed that the simulated work task model offers a suitable paradigm for this kind of experiment. Although the results were achieved with one language, in one specific collection and with one user group, the method and the model are generalizable to any language and can be evaluated with further users and different collections.

In this paper we have described the elements of the user study from the viewpoint of resource type classification and reusability of the components of the developed model. It could be seen that even if the general model developed is reusable, the specific components of it cannot be reused easily. Even if there are not many proprietary parts in the overall system, the intertwining and combining of several components makes system's reuse outside of The National Library of Finland hard. Some of the components - target data, topics and pre-formulated queries, and the resulting research data - however, could be made publicly available and reused, if needed. Some of the components - simulated tasks, analysis, and evaluation methods - on the other hand, are general working methods used in IIR. They can be either quite easily reused as such or remodified for new studies.

To sum up, the research design presented is reusable as a whole. However, the adaptation of each individual component of the research setting must be considered when the reuse scenario is planned. Some components of the design, such as recruiting of participants, may be relatively straightforward to replicate in a new study; adapting other components, such as planning user tasks which allow focusing on, e.g., selected user activities [22] or modifying software, may be more laborious and often require specific expertise. Still, we hope this study helps understanding the challenges of reuse of the research design described as a whole and approaching the steps of adaptation to avoid unnecessary duplication of efforts.

ACKNOWLEDEGMENTS

This work was part of the NewsEye project, which has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 770299. Faculty of Information Technology and Communication Sciences of the Tampere University took part in the arrangement of the query sessions and evaluation of the results as part of the Project EVOLUZ (#326616) financed by the Academy of Finland.

The query environment was implemented by Evident Ltd. (https://evident.fi/).

REFERENCES

[1] Feza Baskaya, Heikki Keskustalo and Kalervo Järvelin. 2012. Time drives interaction: simulating sessions in diverse searching environments. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '12). Association for Computing Machinery, New York, NY, USA, 105–114. DOI: https://doi.org/10.1145/2348283.2348301.

[2] Guilherme Torresan Bazzo, Gustavo Acauan Lorentz, Danny Suarez Vargas and Viviane P. Moreira. 2020. Assessing the Impact of OCR Errors in Information Retrieval. In Jose J. et al. (eds) Advances in Information Retrieval. ECIR 2020. Lecture Notes in Computer Science, vol 12036. Springer, Cham. DOI: https://doi.org/10.1007/978-3-030-45442-5_13.

[3] Melodee Beals, Emily Bell, with contributions by Ryan Cordell, Paul Fyfe, Isabel Galina Russell, Tessa Hauswedell, Clemens Neudecker, Julianne Nyhan, Sebastian Padó, Miriam Peña Pimentel, Mila Oiva, Lara Rose, Hannu Salmi, Melissa Terras and Lorella Viola, The Atlas of Digitised Newspapers and Metadata: Reports from Oceanic Exchanges, Loughborough, 2020. DOI: 10.6084/m9.figshare.11560059.

[4] Pia Borlund. 2000. Experimental Components for the Evaluation of Interactive Information Retrieval Systems. Journal of Documentation, 50 (1), 71-90. DOI: https://doi.org/10.1108/EUM0000000007110.

[5] Chris Buckley and Ellen M. Voorhees. 2000. Evaluating evaluation measure stability. In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '00). Association for Computing Machinery, New York, NY, USA, 33–40. DOI: https://doi.org/10.1145/345508.345543.

[6] Guillaume Chiron, Antoine. Doucet, Mickael Coustaty, Muriel Visani and Jean-Philippe Moreux. 2017 Impact of OCR Errors on the Use of Digital Libraries: Towards a Better Access to Information. In 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), 2017, pp. 1-4, DOI: 10.1109/JCDL.2017.7991582.

[7] W. Bruce Croft, Donald Metzler and Trevor Strohman. 2010. Search Engines. Information Retrieval in Practice. Pearson.

[8] Maria Gäde, Marijn Koolen, Mark Hall, Toine Bogers and Vivien Petras. 2021. A Manifesto on Resource Re-Use in Interactive Information Retrieval. In Proceedings of the 2021 Conference on Human Information Interaction and Retrieval (CHIIR '21). Association for Computing Machinery, New York, NY, USA, 141–149. DOI: https://doi.org/10.1145/3406522.3446056.

[9] David Hebert, Thomas Palfray, Stephane Nicolas, Pierrick Tranouez and Thierry Paquet. 2014. PIVAJ: displaying and augmenting digitized newspapers on the web experimental feedback from the "Journal de Rouen" collection. In Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage (DATeCH '14). Association for Computing Machinery, New York, NY, USA, 173–178. DOI:https://doi.org/10.1145/2595188.2595217.

[10] David Hebert, Thomas Palfray, Stephane Nicolas, Pierrick Tranouez and Thierry Paquet. 2014. Automatic article extraction in old newspapers digitized collections. In Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage (DATeCH '14). Association for Computing Machinery, New York, NY, USA, 3–8. DOI:https://doi.org/10.1145/2595188.2595195.

[11] Johan Jarlbrink and Pelle Snickars. 2017. Cultural heritage as digital noise: nineteenth century newspapers in the digital archive. Journal of Documentation 73: 1228-1243. DOI: 10.1108/JD-09-2016-0106.

[12] Anni Järvelin, Heikki Keskustalo, Eero Sormunen, Miamaria Saastamoinen and Kimmo Kettunen. 2016. Information retrieval from historical newspaper collections in highly inflectional languages: A query expansion approach. J. Assoc. Inf. Sci. Technol. 67, 12 (December 2016), 2928–2946. DOI: https://doi.org/10.1002/asi.23379.

[13] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. ACM Trans. Inf. Syst. 20, 4 (October 2002), 422–446. DOI: https://doi.org/10.1145/582415.582418.

[14] Kalervo Järvelin, Susan L. Price, Lois M. L. Delcambre and Marianne Lykke Nielsen. 2008. Discounted cumulated gain based evaluation of

multiple-query IR sessions. In Proceedings of the IR research, 30th European conference on Advances in information retrieval (ECIR'08). Springer-Verlag, Berlin, Heidelberg, 4–15. DOI: https://doi.org/10.1007/978-3-540-78646-7_4.

[15] Paul B. Kantor and Ellen M. Voorhees. 2000. The TREC-5 confusion track: comparing retrieval methods for scanned text. Inf. Retrieval 2(2), 165–176.

[16] Diane Kelly. 2009. Methods for Evaluating Interactive Information Retrieval Systems with Users. Foundations and Trends in Information Retrieval 3 (1-2), 1-224.

[17] Kimmo Kettunen, Teemu Ruokolainen, Erno Liukkonen, Pierrick Tranouez, Daniel Antelme and Thierry Paquet. 2019. Detecting Articles in a Digitized Finnish Historical Newspaper Collection 1771-1929: Early Results Using the PIVAJ Software. In Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage (DATeCH2019). Association for Computing Machinery, New York, NY, USA, 59–64. DOI: https://doi.org/10.1145/3322905.3322911.

[18] Kimmo Kettunen, Tuula Pääkkönen and Erno Liukkonen. 2019. Clipping the Page – Automatic Article Detection and Marking Software in Production of Newspaper Clippings of a Digitized Historical Journalistic Collection. In A. Doucet et al. (Eds.), TPDL 2019, LNCS 11799, 356–360.

[19] Kimmo Kettunen, Heikki Keskustalo, Sanna Kumpulainen, Tuula Pääkkönen and Juha Rautiainen. 2022. OCR quality affects perceived usefulness of historical newspaper clippings – a user study. In Giorgio Maria Di Nunzio et al. (Eds.), IRCDL2022, 18th Italian Research Conference on Digital Libraries.

[20] Mika Koistinen, Kimmo Kettunen and Jukka Kervinen. 2020. How to Improve Optical Character Recognition of Historical Finnish Newspapers Using Open Source Tesseract OCR Engine – Final Notes on Development and Evaluation. In Vetulani Z., Paroubek P., Kubis M. (eds) Human Language Technology. Challenges for Computer Science and Linguistics. LTC 2017. Lecture Notes in Computer Science, vol 12598. Springer, Cham. DOI: https://doi.org/10.1007/978-3-030-66527-2_2

[21] Robert R. Korfhage. 1997. Information Storage and Retrieval, Wiley, New York.

[22] Sanna Kumpulainen, Heikki Keskustalo, Boyang Zhan and Kostas Stefanidis. 2020. Historical reasoning in authentic research tasks: Mapping cognitive and document spaces. Journal of the Association for Information Science and Technology, 71(2): 230-241. DOI: https://doi.org/10.1002/asi.24216.

[23] Sanna Kumpulainen and Elina Late. 2021. Struggling with digitized historical newspapers: Contextual barriers to information interaction in history research activities. *Journal of the Association for Information Science and Technology*, 1–13. DOI: https://doi.org/10.1002/asi.24608.

[24] Elina Late and Sanna Kumpulainen. 2020. Interacting with digitised historical newspapers: understanding the use of digital surrogates as primary sources, Journal of Documentation (ahead-of-print). DOI: 10.1108/JD-04-2021-0078.

[25] Vivien Petras, Marijn Koolen, Maria Gäde and Toine Bogers. 2019. Experiences with the 2013-2016 CLEF interactive information retrieval tracks. 29-36. In 2019 CHIIR Workshop on Barriers to Interactive IR Resources Re-use, BIIRRR 2019. DOI: https://doi.org/http://ceur-ws.org/Vol-2337/paper5.pdf

[26] Eva Pfanzelter, Sarah Oberbichler, Jani Marjanen, Pierre-Carl Langlais and Stefan Hechl. 2021. Digital interfaces of historical newspapers: opportunities, restrictions and recommendations. Journal of Data Mining and Digital Humanities. DOI: 10.46298/jdmdh.6121.

[27] Myriam C. Traub, Jacco van Ossenbruggen and Lynda Hardman. 2015. Impact Analysis of OCR Quality on Research Tasks in Digital Archives. In S. Kapidakis, C. Mazurek C., M. Werla (Eds.), Research and Advanced Technology for Digital Libraries. TPDL 2015. Lecture Notes in Computer Science, vol 9316. Springer, Cham, 2015. DOI: 10.1007/978-3-319-24592-8_19.

[28] Seppo Zetterberg (Ed.). 1989. Suomen historian pikkujättiläinen, WSOY.

[29] Seppo Zetterberg (Ed.). 1988. Maailmanhistorian pikkujättiläinen, WSOY.