Qualitative and Quantitative Methods in Libraries (QQML) 4: 143-– 152, 2015

Digital libraries as a test bed for evaluating the effectiveness of information searching in OCR-processed texts

John Catlow[1], Mirosław Górny[2], Rafał Lewandowski[3]
[1,2,3] Adam Mickiewicz University. Department of Information Systems

Abstract: The aim of this paper is to point out certain weaknesses of OCR with regard to the problem of information searching, and to describe the mechanisms involved. A methodology is presented for evaluating the effectiveness of information searching in OCR-processed texts. It is also shown to what extent relying exclusively on OCR techniques limits the possibilities of obtaining information from texts. It is indicated how and at what costs these limitations can be overcome by the use of keywords entered by a cataloguer. The research was conducted based on the resources and users of the Digital Library of Wielkopolska.
Keywords: OCR, Optical Character Recognition, Digital Library, Searching, Effeciveness of Information Searching, Historical Publications

1. Introduction

The resources of Polish digital libraries are largely of a historical nature. There are several reasons for this. Firstly, there is the desire to preserve the nation's heritage and enable access to it. We aim in particular to preserve for future generations those collections which are subject to gradual degradation. Also, many resources are not made available to readers physically because of the possibility that they might be further damaged. Converting them to digital form and making them available via the Internet enables them to be made accessible to a large number of readers. Another significant factor is undoubtedly the matter of copyright, which in the case of historical publications has generally expired. This means that such works can be converted to digital form and made available to readers much more easily, without the need for additional formalities and related costs.

A fundamental tool facilitating access to digital data is the index. We may distinguish two types of indexes used in digital libraries:

Received: 29.4.2014 / Accepted: 1.12.2014 ISSN 2241-1925

- indexes created manually by professional cataloguers;
- indexes created automatically from texts obtained via the process of optical character recognition (OCR).

2. Problems with OCR

A number of problems frequently arise during the process of optical recognition of historical documents. They include:

- Paper degradation – this is affected by the process of acidization of paper. At the Jagiellonian Library in Kraków, 82% of books and journals, a total of around 1.5 million items, are printed on acid paper.
- The low-density paper used in newspapers means that the content of the reverse side of the page sometimes becomes visible during scanning.
- Imprecise printing. Many newspapers, particularly regional ones originating from small towns, were printed on obsolescent machinery. The text could become illegible even a short time after printing. Further damage was done by degradation, causing discontinuities in the printed letters.
- Improper storage.
- Creasing of newspapers and books. This may be caused by paper acidity or changes in the humidity of the rooms where the items were stored.
- Disused fonts. Until recently there were difficulties with recognizing texts printed in historical fonts, such as Gothic. At present, some OCR systems are able to cope increasingly well with such fonts.
- Damage occurring during use. This may include the staining or tearing of pages, or wearing away of text.

3. Models of publications contained in digital libraries

Every digital resource can be analysed as consisting of a certain number of objects. The analysis can be made at several levels of detail: we may consider the characters, the words, or the phrases which make up a publication.

In considering digital libraries, we may focus on three publication models:

- The newspaper – this has a small volume and covers varied topics. The articles are short, and their content chiefly informational. An important factor for the user is the possibility of looking through search results quickly and easily, due to the small volume. Newspapers were often printed on inferior paper using obsolescent printing machines. It can therefore be expected that the quality of OCR will be lowest in this case.
- The magazine – this is more extensive than a newspaper, containing long articles, often on varied topics. In the search results, the user is usually interested only in a selected article. In the case of publications of this type, it is helpful to use segmentation to make the data browsing process easier. OCR is more effective with magazines than with newspapers.
- The monograph – this is the most extensive out of the publication models being considered. Depending on the subject (whether it covers multiple topic areas or only one) it may be helpful to use segmentation to make

browsing easier. Because of the print quality, this type of publication provides the highest OCR quality.

4. Tools to support data searching – indexes

Indexes are a popular tool for users of information systems. For the purposes of our present analysis we will consider two types:

- indexes created based on terms entered manually by a cataloguer;
- indexes created automatically from OCR-processed text.

Indexes created for the same resource (one publication or a larger collection) may differ significantly, and thus may be received differently by users.

An automatically created index takes an unordered form. It includes terms which were not correctly recognized during the OCR process, and may therefore be incomprehensible to users. This applies in particular to terms that were not correctly recognized by the OCR software because they were split between lines. In languages such as Polish there are many words that have the same root, but appear in different inflected forms. Finally, in the case of historical literature and publications, there appear many terms which are no longer used and will be unknown to users.

The index will be very extensive and will include many terms that users will not in fact use during the data search process.

A manually created index is more friendly to the user and will be free of the defects listed above. It will nonetheless contain fewer terms, and it may well happen that a user fails to find a desired phrase in a manually created index, even though it would be present in an index produced automatically.

An automatically created index can be transformed into a more user-friendly form, although this requires time and costs. Incorrectly recognized terms can be corrected, lemmatization and stemming can be applied, and a dictionary of synonyms can be attached. Such an operation would need to be performed whenever a new publication is added. It would therefore require the allocation of certain hardware, software and human resources.

It is also possible to generate automatic indexes in the form of a tag cloud, which indicates not only the terms themselves, but also their popularity (frequency of occurrence) in a particular resource (publication or full collection).

5. Differences in searching resulting from the use of OCR or manual indexing

Full-text searching usually generates a large number of results. Particularly in the case of general or popular terms, a marked difference can be seen in comparison with manual indexes. In the case of rarely encountered terms, it may be that they are omitted from a manually created index. In that case an index created automatically from OCR data offers a clear advantage, as it provides the only possibility of finding the desired term.

On the other hand, the large number of generated results makes searching through them difficult and laborious, and may discourage the user from

investigating the resources further. There may also be a large number of irrelevant results, which further discourages the user.
The smaller number of results obtained in the case of manual indexing offers the possibility that all of the items found will be read.
On the other hand, the person making the catalogues and descriptions of the content may fail to mark all of the significant materials which would appear in full-text searching.

6. Costs of creating an index using OCR and manually

The costs of creating indexes from OCR output can be significant. The text obtained directly from scanned materials may be highly imperfect. Its quality is affected by several factors:

- The quality of the source materials – in the case of Polish digital libraries these are often historical materials, with low paper and printing quality. The text obtained from such scanned materials is often very imperfectly recognized.
- OCR software – the performance of such programs can vary markedly. Also of importance may be the list of dictionaries used to support text recognition and the ability to add non-standard dictionaries. In the case of larger libraries, when extensive quantities of materials are processed using OCR, the speed of the process will also be significant.

The texts obtained will usually require correction – by automatic means, but also with the involvement of humans. There is a need for "manual" text correction, removal of flexional variants, and addition of synonyms. It is also necessary to remove "stop words" (words with little information content).

7. Time required for searching and evaluation of information

The list of publications found as a result of a search, particularly a full-text search, may be extremely long. There may be several thousand publications and hundreds of pages to look through. The time needed to familiarize oneself with the results of a search is often unacceptable to a user. The user has to look through all of the publications in order to find which of them are valuable (significant) for him or her, and which are not. The time required to investigate the results may differ greatly depending on the model of publication involved. Checking the results found in a newspaper, in view of its small volume, will take only a short time. In the case of a monograph, when a term occurs in many places in the text, full investigation of the text may take longer.

8. Tests

A series of tests was carried out on the digital data stored in the Digital Library of Wielkopolska. The publications stored in the library are manually indexed, and selected publications are also indexed automatically using texts produced by OCR.

Methodology

The search terms were found in publication descriptions, and the number of successful hits was recorded. After that the same search was carried out in full-text mode, based on data from the index produced by OCR processing.

The search terms were placed in quotes, so as to search for them in exactly the form in which they were entered in the search field.

The search terms used in the tests related to various subject areas.

Table 1. Numbers of results obtained by searching the Digital Library of Wielkopolska for terms being names of Polish cities, using automatically and manually generated indexes.

Term	Number of results from full-text searching	Number of results from searching publication descriptions	Publication description results as percentage of full-text results
"Poznań"	202,066	102,858	51.09%
"Kalisz"	7,388	648	8.78%
"Piła"	9,391	17	0.18%
"Konin"	3,062	142	4.64%
"Gniezno"	24,294	3,602	14.83%
"Leszno"	11,924	5,004	41.97%
"Bydgoszcz"	21,505	2,831	13.16%
"Toruń"	14,991	708	4.72%
"Gdańsk"	19,887	697	3.50%

The results of the tests in the "cities" category show that the number of hits in the case of full-text searching is significantly greater. In view of the origin of the collections (the Wielkopolska region) the numbers of hits for cities outside that region are noticeably lower.

In certain cases, a word being the name of a city also has other meanings. Some of the search results may thus be misleading. This applies to terms such as *Piła* (which also means "saw" in Polish), *Łódź* ("boat"), *Brzeg* ("edge"), *Łuków* (genitive form of "bows" or "arcs"), *Krosno* ("loom"), *Koło* ("wheel"), *Turek* ("Turk"), etc. This may be one of the reasons for the much greater number of results returned by full-text searching than by searching of manual descriptions (as in the case of "Piła"). It should be remembered that information about these cities may also be available under their German or even Latin names – such results may appear when a dictionary of synonyms is used.

Table 2. Numbers of results obtained by searching the Digital Library of Wielkopolska for terms being names of smaller towns in Poland, using automatically and manually generated indexes.

Term	Number of results from full- text searching	Number of results from searching descriptions of publications
"Wągrowiec"	15,313	3,891
"Luboń"	1,676	17
"Szamotuły"	13,250	3,002
"Rogoźno"	5,057	41
"Kościan"	8,226	510
"Gostyń"	6,331	331
"Jarocin"	7,925	4,644
"Pleszew"	6,639	3,372
"Krotoszyn"	9,415	1,353
"Wolsztyn"	6,374	2,996
"Tuchola"	1,513	7

There is a certain group of terms which were not included in any manually created descriptions of publications. This brings to light an advantage of automatic full-text indexing – without such an index, the terms in question would not have been found at all.

Table 3. Numbers of results obtained by searching the Digital Library of Wielkopolska for terms referring to institutions and events, using automatically and manually generated indexes.

Search term	Number of results from full- text searching	Number of results from searching descriptions of publications
"Gimnazjum w Wolsztynie" (Wolsztyn Middle School)	31	0
"Gimnazjum w Kaliszu" (Kalisz Middle School)	26	0
"Biblioteka w Lesznie" (Leszno Library)	6	0

“Dom Kultury w Kościanie” (Kościan Culture Centre)	4	0
“Kino w Lesznie” (cinema in Leszno)	6	0
“kapela dudziarska” (bagpipe band)	24	0
“turniej recytatorski” (public recitation competition)	38	0
“wystawa akwarystyczna” (aquarium exhibition)	0	0
“mecz piłkarski Polska Niemcy” (Poland–Germany football match)	14	0
“Unia Leszno” (name of a speedway club)	15	0
“Ostrovia Ostrów” (name of a speedway club)	43	0
“przegląd filmów amatorskich” (amateur film review)	4	0
“Parafia w Pleszewie” (Pleszew Parish)	2	0
“zawody wędkarskie” (angling competition)	42	0
“koncert muzyczny” (musical concert)	21	0
“koncert muzyki” (music concert)	1418	0

9. Conclusions

A digitized text is a file – sometimes large, sometimes small – containing a certain number of strings of characters. These strings form words which have some significance for the user. If we map them to the file addresses at which they occur, we can obtain a list of files in which those words appear. This is how searching in digital texts works. The user provides a specific word in a query, and obtains as a result a set of texts which contain that word. The problem is that words usually appear in many different texts, and the user is often required to look through large numbers of texts in order to find one of interest.

Of course, computer tools provide certain means of narrowing down one’s search results. A user may add other words to the query, or may define the position of the words in the text – for example by requiring that certain words appear within a defined distance of each other – or may also require that certain words appear with a certain frequency, and that certain others must not occur. This, however, is all that computerization can offer. The presence or absence of

a given word in a text, its position, the co-occurrence of other words and the frequency of occurrence of a given word are effectively all that an automated system can allow one to specify. (It may also take account of the different ways in which particular words are distinguished, but the user is not usually able to use such features in a query – they are more likely to be used by the system itself, which may apply various default procedures to try to determine the relevance of a result.)

Use of these techniques requires a significant amount of effort and experience on the user's part, as well as knowledge about the collections being searched. The effect may still be unsatisfactory (with an excessively large number of results being returned). It should also be borne in mind that the creation of such possibilities entails the high costs of building and operating a computer system, not to mention the problem of erroneous reading of certain words in the OCR process.

An automated solution may be complemented by a manual indexing system. The indexer uses his or her knowledge of users' needs and applies an appropriate strategy. Because this work is expensive, it is desirable to limit its application to a selected group of sources.

Lists of addresses and other types of list do not require manual indexing, as they contain sets of personal or other names that are relatively well ordered and well suited to being looked through manually by a user (in the case of printed works), and are relatively easy to process with OCR. For similar reasons, manual indexing is not applied as a rule to encyclopaedias, dictionaries and other such works of reference. Indexing of monographs is useful to a limited extent. This is because they usually concern a single subject, and contain extensive material that needs to be read as a whole. A description of their content that is adequate for a reader's needs is usually contained in the bibliographic description and in the table of contents, if this is included as a subcollection with purely informational functions.

An enormous problem is posed, however, by press publications (particularly daily newspapers). It may be assumed, with a certain degree of simplification, that manual indexing is a procedure that should be applied above all to newspapers and magazines.

Press publications have a different structure to that of other types of source. They consist of relatively short reports, covering varied topics and often having significant informational significance. The indexer must therefore act in a manner appropriate to the expected needs of users. We attempt to describe an appropriate procedure here. The indexer must pay attention to:

1. Events – the indexer should assess whether a given event may be of significance to a potential reader. It is not possible to lay down any hard guidelines for this, as much depends on the indexer's knowledge of the subject matter and of readers' information needs. Nonetheless, in regional newspapers for example, the indexer will normally omit events of international or nationwide importance, on the assumption that such articles

are merely reprinted from national newspapers, instead focusing attention on local events, which may not be recorded in any other sources.

2. Absent key words – sometimes the indexer may decide that a word might be used by a reader even though it is absent from the text. For example, if the text contains extensive descriptions of military activity in some area, but the name of that area is not given explicitly, the indexer will often deduce what area is being described, and include its name as a key word.
3. Deformed words – this may be an error resulting from font smudging, for example.
4. Proper names which deserve to be exposed for some reason.
5. Volume of text on a given topic – the indexer will pay attention to relatively extensive texts devoted to a particularly subject area, assuming that long texts by their nature may contain significant information, opinions, etc.
6. Exceptional information – sometimes a relatively minor reference, which to contemporaries may not have seemed important, may now be of much greater significance, for instance as a step in the development of technology – an example of this was a report of an offer by German firms to install video telephones between Łódź and Warsaw in 1936.

In order for manual indexing to be of help to a reader, the words entered by indexers must be given greater weight during searching. Indeed, this is a feature provided by the software used by many systems.

It is difficult to assess the economic effectiveness of manual indexing. This is not due to any difficulty in calculating the costs of the process (if an indexer can index approximately 100 pages of text daily, then he or she will complete approximately 2000 pages per month, and approximately 20,000 pages per year – taking the annual cost of employing such a person in Poland to be around 40,000 zloty, the cost of indexing one page comes out to approximately 2 zloty or 0.5 euro). The obstacle faced is that it is not possible to place a value on the results of this indexing work.

The effectiveness of searching should not be assessed merely in terms of a percentage of appropriate results. In the case of searches performed by users of digital libraries, a single good result is sometimes of enormous value. Hence even if the large number of words inserted by indexers seems not to play any role during searching, it is enough that they may cause just one extremely important document to be returned.

Statistically, it is theoretically possible for a comparison of effectiveness to be made between searching using automatically constructed indexes and searching using manual indexes. Nonetheless this requires analysis to be performed in the course of real searching, since otherwise there is no possibility of determining the relevance of the results – only the person performing the search is able to make such a determination.

A solution, up to a point, is to prepare tests according to appropriate scenarios. Then, however, the results may be unconsciously distorted by the person arranging the test. We do not have access to material that has been manually

indexed in full, and a degree of error is likely to be introduced when samples of library content are prepared for testing purposes.
As a consequence, the tests are reduced to identifying a certain set of sources being found on the basis of given words from automatically constructed indexes. That set is then analysed to identify in it the words which ought to be exposed. The size of samples depends on the size of the basic collection and on the query made. At present the size of the basic collections of some digital libraries (including the Digital Library of Wielkopolska) would appear to be entirely adequate.
It must nonetheless be remembered that as the size of the basic collection increases, the effectiveness of OCR techniques is reduced, because of the increase in the size of the result sets. In that case the importance of manual indexing increases.

References

Barański, Andrzej, dok. elektr. (2006). Kwaśny papier, http://www3.uj.edu.pl/alma/alma/82/09.pdf.

Bieniecki, Wojciech, dok. elektr. (2005). Analiza wymagań dla metod przetwarzania wstępnego obrazów w automatycznym rozpoznawaniu tekstu. http://wbieniec.kis.p.lodz.pl/research/files/05_Bronislawow_OCR.pdf

Droettboom, Michael, dok. elektr. (2003), Correcting broken characters in the recognition of historical printed documents, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.93.7866&rep=rep1&type=pdf

EDL Report on Digitisation in European National Libraries 2006-2012 (2008), http://www.cenl.org/docs/Report_digitisation_NLs.pdf

Eikvil, Line, dok. elektr. (1993). OCR - Optical Character Recognition. Report No. 876, 1993. http://www.nr.no/~eikvil/OCR.pdf

Hauser, W. Andreas, dok. elektr. (2007). OCR Postcorrection of Historical Texts. http://www.cip.ifi.lmu.de/~hauser/papers/histOCRNachkorrektur.pdf

Impact: Improving access to text: Concept, dok. elektr., http://www.impact-project.eu/about-the-project/concept/

Lin, Leo, dok. elektr. (2009). Improving Digital Library Support for Historic Newspaper Collections. http://researchcommons.waikato.ac.nz/bitstream/10289/3262/1/thesis.pdf

Tanner, S., Muñoz T., Ros P. H. dok. elektr. (2009), Measuring Mass Text Digitization Quality and Usefulness, D-Lib Magazine, http://www.dlib.org/dlib/july09/munoz/07munoz.html

Vamvakas G., Gatos B.,Stamatopoulos N., Perantonis S.J., dok. elektr. (2008). A Complete Optical Character Recognition Methodology for Historical Documents, The Eighth IAPR International Workshop on Document Analysis Systems, http://iit.demokritos.gr/~bgat/3337a525.pdf