Original scientific paper
http://doi.org/10.24867/JGED-2020-2-005

Analysis of quality factors for digitization process of old books

ABSTRACT

The digitization process of old books is one of the measures to preserve the cultural heritage of mankind and to make them accessible to a large audience. The task of creating a high quality digital copy of an old book is quite complex and it depends on many factors. Thus, the analysis of the factors that determines the quality of the digitization process has been done, the models of relationships between the factors have been constructed and their priority has been established with the help of the method of factors ranking using the hierarchical representation of relationships between them in the form of graphs, the calculation of the corresponding weight coefficients and the expert survey with its interpretation in a fuzzy form.

KEY WORDS
old book digitization, quality, multilevel model, fuzzy sets

Olena Tsimer
Vyacheslav Repeta
Ihor Myklushka

Ukrainian Academy of Printing,
Faculty of Publishing, Printing and Information Technologies,
Lviv, Ukraine

Corresponding author:
Vyacheslav Repeta
e-mail: vreneta@yandex.ua

First recieved: 25.03.2020.
Accepted: 04.05.2020.

Introduction

In the process of obtaining a digital copy of an old book, there are difficulties that require the adjustment of its technological operations, in particular the general condition of pages, the opening degree of the book block, the heterogeneity of the page background and changes in the optical density of the text and illustrative information due to the aggressive influence of different environments in the process of the old book storing. They form the range of optical densities for each page and require adjusting the light distribution over the recorded area.

Specialized equipment is used for the old book digitization, the application of which does not involve the fastest possible process implementation but obtaining a high quality digital copy. It is known that such equipment is currently installed in many scientific libraries of Ukraine (Kotsyuba, 2012; Anon, 2019; Kotsyuba, 2019). The following companies present it on the market: Zeutschel GmbH, Image Access GmbH, Atiz Innovation Co., Ltd, TREVENTUS Mechatronics GmbH, SMA Electronic Document GmbH, ELAR, i2S SA. Each manufacturer offers several types of devices for old book scanning that differ in format, book placement principle, digital camera resolution, and different innovative approaches .

The operation of any recording system is characterized by digital noise. Digital noise is a defect in an image that results in statistically inevitable randomly scattered pixels of random colour and brightness over the area of the recorded image that does not match the original. It is clear that the value of digital noise varies depending on the condition of the original, the parameters of the scanning device, the principles of image processing software (Crowley, 2016). The result of optimal selection of the scanning process factors and its modes is the characteristic feature of a digital copy.

It is known that the software has been tested for performing technological operations with digital images and the expediency of using the following applications

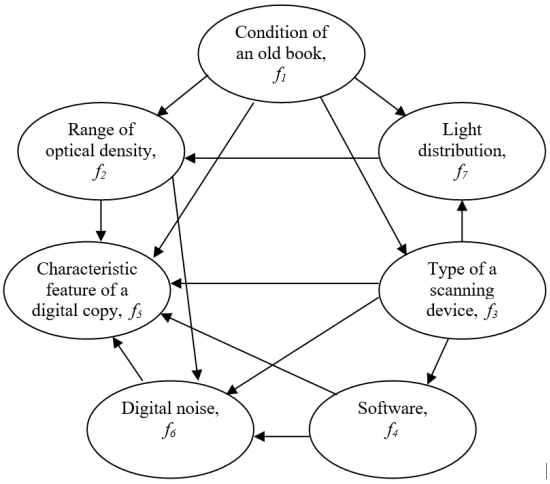
has been established: ACDSee, Book Restorer, Scan Kromsator Scan Tailor, Adobe Photoshop, Easy Scan Plus.

The objective of the work is to determine the priority of the factors that determine the quality of the digitization process of old books for their subsequent digital restoration. The following factors have been identified in the analysis of the digitization technology of old books:

- the condition of an old book (f_1);
- the range of optical density (f_2);
- a type of a scanning device (f_3);
- the software (f_4);
- the characteristic feature of a digital copy (f_5);
- the digital noise (f_6);
- the light distribution (f_7).

Methodology

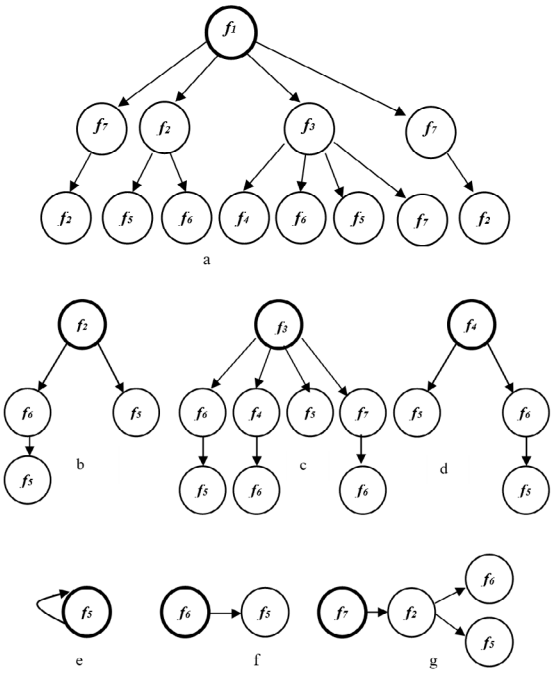
For the analysis of the process factors, the method has been used, which has shown its versatility to solve the problems of factors priority of different processes (Senkivskyy, Pich & Melnykov, 2013; Repeta, Senkivsky & Piknevych, 2014). This method, in contrast to the method of hierarchies analysis and solving the reachability matrix and pairwise comparison of factors (Pikh & Senkivskyy, 2013), takes into account both direct influences and dependencies between factors and indirect or mediate, that is, those that pass through another factor. The method of ranking distinguishes between the types of influences and the dependencies of factors by giving different weights to each of them, and the conducted analysis does not lead to the placement of two or more factors at the same level of the hierarchy.



» **Figure 1:** *An oriented graph of factors for the digitization process of old books*

Based on the graph (Figure 1), we construct for each of the factors hierarchical trees of their relationship with other factors, taking into account the influences of both types – direct and mediate, that is, indirect, passing through another factor.

After analysing the graphs (Figure 2), we calculate the total weight values of the direct and indirect influences of factors and their integral dependency on other factors. To do this, we introduce the following symbols. Let k_{ij} be the number of influences (i = 1 – direct, i = 2 – indirect) or dependencies (i = 3 – direct, i = 4 – indirect) for the j-th factor (j = 1,...,n); w_i is the weight of the i-th type. For calculations, we take the following conditional values for the weight coefficients in conditional units: w_1 = 10, w_2 = 5, w_3 = -10, w_4 = -5.



» **Figure 2:** *Graphs of multilevel hierarchical relationships for factors of the digitization process of old books (a – g)*

We denote the total weight values of all types of relationships of quality factors by P_{ij}.

We use the following formulas for calculations:

$$P_{ij} = k_{ij} w_i \ \ (i = 1,2,3,4; \ \ j = 1,...,n) \tag{1}$$

where n is a number of the factor.

For our oriented graph (Figure 1) taking into account (1), we get:

$$P_{Fj} = \sum_{i=1}^{4} \sum_{j=1}^{7} k_{ij} w_i \tag{2}$$

It is clear that, in the absence of one factor for one of the relationship types, its corresponding value k_{ij} in the expression (2) will be zero. The presented formula is the basis for getting weight values of factors ranking, taking into account different types of relationships between

them. For the formation of the number of influences and dependencies of factors (Table 1), we define direct influences for each of them, the number of which is fixed by the coefficients k_{1j}. "Ways of dependency" provide with obtaining the coefficients k_{3j} in a similar way. The combined consideration of indirect influences or dependencies of a factor (i.e. influence or dependency through other factors) determines the coefficients k_{2j} and k_{4j}.

Table 1

Directions of influence and ways of dependency of factors for the digitization process of an old book

Characteristics of the factor	Number of factor *j*						
	1	2	3	4	5	6	7
Impacts	4	2	4	2	0	1	1
Dependencies	0	2	1	1	5	3	2

It should be noted that $P_{3j} < 0$ and $P_{4j} < 0$ since according to the given initial conditions $w_3 < 0$ and $w_4 < 0$. So, to bring the total weight values of the factor with the lowest priority to zero and the rest to a positive value, we transform the formula (2) into:

$$P_{Fj} = \frac{1}{5}\left(\sum_{i=1}^{4}\sum_{j=1}^{7} k_{ij} w_i + S_j\right), \quad (3)$$

$$S_j = max|P_{3j}| + max|P_{4j}|. \quad (4)$$

The specified values are added in each row to the sum of the values in the columns P_{1j}, P_{2j}, P_{3j} and P_{4j}. Finally, we obtain the resultant weight of the factor, which is the basis for establishing the factor rank r_j, which is equivalent to the priority of its influence on the digitization process of old books.

At the next stage, a questionnaire has been formed according to these factors, in which experts (employees of Vasyl Stefanyk National Science Library, Ukrainian Academy of Printing, Lviv Printing College of Ukrainian Academy of Printing) have been asked to evaluate the importance of the factors influence on the digitization process of old books. And it is necessary to indicate how the digitization process will improve, according to the questions asked:

1. How does the condition of the old book affect the quality of the digitized copy?
2. Is the range of optical densities of the original essential during its digitization?
3. How will the selection of a scanning device affect the quality of the old book digitization?
4. How does the software affect the quality of the received digital copy?
5. Does the effect of digital noise have a significant impact on the quality of the digitization process?
6. How does the light distribution affect the quality of the digital copy?

The questionnaire form is filled in when answering six questions (Table 2).

Table 2

The form of the questionnaire, which is proposed to the experts

No.	Number of factor j				
	0	25	50	75	100*
1					
2					
...					
n					

*scale values 0-100% indicate the level of improvement of the process of digitizing the old print.

Each expert interprets these answers individually, that is, it takes a fuzzy form. Therefore, it is desirable not only to select an acceptable answer, but also to provide a quantitative assessment. To do this, the expert is suggested a so-called "soft" form of quantitative interpretation of answers when the expert has to give not one but several quantitative assessments. For the expert orientation, the set of possible quantitative assessments takes on values, for example, from 0 to 100% (Baranov & Ptushkin, 2004). In addition, the expert is asked to give a degree of certainty to each of such assessment that the selected quantitative assessment would be correct. The degree of certainty is quantitatively characterized by the verbal-numerical scale by Harrington (Saaty, 1980), which is presented below (Table 3).

Table 3

The verbal-numerical scale by Harrington

Confidence level	Value
very high	0.8 – 1.0
high	0.63 – 0.8
average	0.37 – 0.63
low	0.20 – 0.37
very low	0.0 – 0.20

Thus, the task of the expert is to select the answer in the form of Table 2, which is offered to the expert when answering the question, and to evaluate it on the verbal-numerical scale by Harrington. For example, this is one answer to the question: "Does the effect of digital noise have a significant impact on the quality of the digitization process?" in Table 4.

Table 4

The expert's answer to the given question

Possible values of the parameter changes, %				
0	25	50	75	100
0.8	1.0	0.5	0.3	0.1

In the presented questionnaire: 0 is the lowest degree of certainty, 1 is the highest degree of certainty.

According to the table, the membership function has the following form:

$$\mu(u) = [0.8; 1.0; 0.5; 0.3; 0.1]$$

The example in the table above points to the following:

- the influence of digital noise on the digitization process is 25% (the degree of certainty is 1.0);
- there is slightly less certainty (0.8) that such an influence will be absent at all.

The membership function of the generalized thought is determined by the formula:

$$M(U_i) = min\,[(\mu_1(u_i),(\mu_2(u_i),\ldots,\ (\mu_n(u_i)\ ] \quad (5)$$

The result of the expert survey is the maximum value of the function:

$$u^*_i = argmax\mu_i(u_i) \quad (6)$$

Results

The calculations (according to formula 3), we form Table 5 to establish the factors ranks. As it can be seen from the table, max $|P_{3j}|$ = 50; max $|P_{4j}|$ = 30. The factor with the highest value of P_{Fj} has highest rank.

Table 5

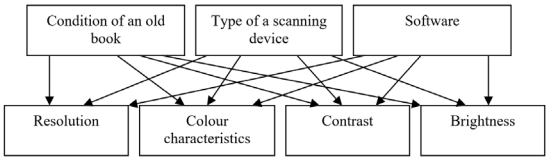
Calculated data of factors ranking for the digitization process of old books

Number of factors j	k_{1j}	k_{2j}	k_{3j}	k_{4j}	P_{1j}	P_{2j}	P_{3j}	P_{4j}	P_{Fj}	Factor rank r_j
1	4	4	0	0	40	20	0	0	28	1
2	2	1	2	2	20	5	-20	-10	15	4
3	4	1	1	0	40	5	-10	0	23	2
4	2	1	1	1	20	5	-10	-5	18	3
5	0	0	5	6	0	0	-50	-30	0	7
6	1	0	3	4	10	0	-30	-20	8	6
7	1	1	2	1	10	5	-20	-5	14	5

According to the received results, factors with the highest rank are (Figure 3): the condition of an old book, a type of a scanning device, the software. Accordingly, these factors have the greatest influence on such parameters of the resulting digital image as resolution, brightness, contrast and colour characteristics, which determine the prerequisites for the next digital copy processing.

At the second stage the subsequent answers to the *i*-th question is summarized in Table 6. If the expertise level of the experts is the same, then the overall fuzzy assessment will be obtained at the intersection of fuzzy sets belonging to the experts' answers.

The membership function quantifies this assessment according to the fuzzy set intersection rule.



» **Figure 3:** *Influence of priority factors for the digitization process of an old book on parameters of a digital copy*

Table 6

Experts' answer to the given question

Experts	Possible values, %				
	0	25	50	75	100
1	0.8	1.0	0.5	0.3	0.1
2	0	0.7	1.0	0.5	0.2
3	0	0.8	0.6	0.4	0.2
4	0	1.0	0.4	0.2	0.1
5	0	0.95	0.7	0.5	0.25
6	0	1.0	0.75	0.2	0
7	0.4	1.0	0.8	0.7	0.6
8	0	1.0	0.8	0.6	0.5
9	0	0.9	0.7	0.5	0.25
10	0	0.8	0.5	0	0

Ten representatives from the expert group have answered the above question (see Table 6) in the form of a fuzzy set M with the corresponding membership function μ:

M1=0,8/0+1,0/25+0,5/50+0,3/75+0,1/100	μ1(u) = [0.8; 1; 0.5; 0.3; 0.1];
M2=0/0+0,7/25+1,0/50+0,5/75+0,2/100	μ2(u) = [0; 0.7; 1; 0,5; 0.2];
M3=0/0+0,8/25+0,6/50+0,4/75+0,2/100	μ3(u) = [0; 0.8; 0.6; 0.4; 0.2];
M4=0/0+1,0/25+0,4/50+0,2/75+0,1/100	μ4(u) = [0; 1; 0.4; 0.2; 0.1];
M5=0/0+0,95/25+0,7/50+0,5/75+0,25/100	μ5(u) = [0; 0.95; 0.7; 0.5; 0.25];
M6=0/0+1,0/25+0,75/50+0,2/75+0/100	μ6(u) = [0; 1; 0.75; 0.2; 0];
M7=0/0+1,0/25+0,8/50+0,6/75+0,5/100	μ7(u) = [0; 1; 0.8; 0.6; 0.5];
M8=0/0+1,0/25+0,8/50+0,6/75+0,5/100	μ8(u) = [0; 1; 0.8; 0.6; 0.5];
M9=0/0+0,9/25+0,7/50+0,5/75+0,25/100	μ9(u) = [0; 0.9; 0.7; 0.5; 0.25];
M10=0/0+0,8/25+0,5/50+0/75+0/100	μ10(u) = [0; 0.8; 0.5; 0; 0]

The answer for the influence of digital noise is: [0; 0.7; 0.4; 0; 0].

Thus, u^* = 0.8 correspond to 25 % influence on the process quality. The answers to other questions are shown in Table 7.

This survey has revealed that the condition of an old book and the light distribution are of the highest importance, experts have shown for them with certainty u^* = 1 that these factors determine the digitization quality by 75 and 100%, respectively.

Table 7

Results of the survey according to the suggested questionnaire

Question	Maximum value of the function, u^*				
	0	25	50	75	100
1	0	0	0.5	0.7	1.0
2	0	0.2	0.6	0.8	0.6
3	0	0	0.5	0.9	0.7
4	0	0	0.9	0.8	0.5
5	0	0.7	0.4	0	0
6	0	0.4	0.5	1	0.8

Conclusions

According to the analysis by the method of ranking, the following factors have been found to have the highest priority in the digitization process of old books: the condition of the old book, a type of a scanning device, and the software. The survey conducted in the second stage has confirmed the priority of such factor as "the condition of the old book". In addition, the importance of the factor "light distribution" has been shown, for which experts have revealed with certainty $u^* = 0.8$-1 that this factor determines the quality of the digitization process of an old book by 75-100%, although this factor has occupied the fifth position during ranking. It should also be noted that the factor "light distribution" in the construction of the relationship model between the factors depends on the condition of the old book and the type of a scanning device, which reduces its importance. The third most important factor ($u^* = 0.9$) is the factor "a type of a scanning device", which is expected to achieve 75% of the process improvement. The factor "software" was next in priority and experts have indicated with certainty $u^* = 0.9$ that selecting it correctly could improve the digitization process by 50%. The two methods of the factor analysis used have allowed getting almost identical results that complement each other. The next stage of our study will be the research of the influence of priority factors by means of fuzzy logic.

References

Baranov, L. & Ptushkin, A. (2004) *Fuzzy sets in an expert survey*. Avaliable from: http://www.isras.ru/files/File/4M/19/Ptushkin.pdf [Accessed 24st December 2019].

Crowley (2016) *Determining the Best Method for Scanning Bound Materials*. Avaliable from: https://www.thecrowleycompany.com/determine-best-method-scanning-bound-materials [Accessed 27st December 2019].

Kotsyuba, E.Y. (2019) Technological Features of Digitization of Valuable and Unique Documents of the Scientific Library. *The Place and Role of Libraries in the Formation of the National Information Space: Proceedings of the International Scientific Conference, 21-23 October 2014, Kyiv, Ukraine.* Avaliable from: http://conference.nbuv.gov.ua/report/view/id/276 [Accessed 23st December 2019].

Kotsyuba, V. (2012) *How to Digitize History.* Avaliable from: https://ms.detector.media/web/online_media/yak_otsifrovuyut_istoriyu [Accessed 21st December 2019].

Pikh, I. & Senkivskyy, V. (2013) *Information technology for modeling publishing workflows.* Lviv, Ukrainian Academy of Printing.

Anon (2019) *Presentation of the new resource "Treasures of Ukraine: digital collection".* Avaliable from: http://odnb.odessa.ua/img/novini_2019/2559/pro-proekt.pdf [Accessed 21st December 2019].

Repeta, V., Senkivsky, V. & Piknevych, S. (2014) Calculation of the importance of quality factors in braille application process on labels by screen UV-varnishes. *Journal of Graphic Engineering and Design*. 5 (2), 5-8.

Saaty, T. (1980) *The analytic hierarchy process: planning, priority setting, resource allocation*. New York, McGraw-Hill.

Senkivskyy, V., Pich, I. & Melnykov, O. (2013) The method of ranking factors influencing the quality and manufacturing processes. *Printing and Publishing.* 61-62 (1-2), 33-41.