Ildikó, Szerényi[1]

INDEXING USING ARTIFICIAL INTELLIGENCE AND INVOLVING VOLUNTEERS AT THE NATIONAL ARCHIVES OF HUNGARY (NAH)

Abstract

Purpose: *The pilot project detailed in this article covers the text recognition of a set of hand-written 19th century archival documents by artificial intelligence and with the help of volunteers.*

Method/approach: *The work started at the archives with the creation of a handwriting recognition model. The paleographical volunteers were tasked with testing this model, validating and correcting the words transcribed by the algorithm.*

Results*: In addition to handwriting recognition, the greatest benefit of this pilot project in Hungary was the huge social activity it generated. The project created a cooperating community of volunteers that will continue to work together in the future. The database, which has a great importance to genealogists as well, will be made available to the public by the National Archives of Hungary in autumn 2022.*

Conclusions/findings: *The high number of volunteers and the huge amount of work they completed surpassed all expectations and was a pleasant surprise for the archival sector. Preparations for the next similar project using artificial intelligence have already started.*

Key words: *Artificial Intelligence, transcribing, volunteers, HTR, census, European Digital Treasures Project*

[1] Ildikó Szerényi, National Archives of Hungary (NAH), 1014 Budapest, Bécsi kapu tér 2-4, szerenyi.ildiko@mnl.gov.hu, IIAS member

1. INTRODUCTION

From January 2020, the Databases Online portal (www.adatbazisokonline.hu), available free of charge and operated by the NAH, provides researchers with a renewed, more user-friendly interface and it is supported by visual elements. At present the constantly expanding research site contains more than 3 million text pages, images and audio material. The database collection currently includes around 70 individual databases; however, the site is much more than just a set of databases. The common search interface with a renewed search engine allows you to run simple and complex searches quickly and easily, providing highly accurate results. Searches are based primarily on the finding aids and content extracts produced by archivists and in ever-increasing numbers, on indexed digitised documents. Until very recently, text-searchable databases have in most cases been created using OCR (Optical Character Recognition) technique in case of printed or typewritten archival material. The indexing of handwritten archival records until 2021 meant human processing, which due to limited available capacity, proved to be less feasible in quantitative terms.

Most of the documents preserved in Hungarian archives are handwritten, as the use of typewriters in state offices became widespread only at the beginning of the 20th century. The NAH were the first archives in Hungary to make use of the possibilities offered by artificial intelligence. This innovative project was implemented with the involvement of volunteers and social collaboration, which also highlights the novelty and creativity of the project.

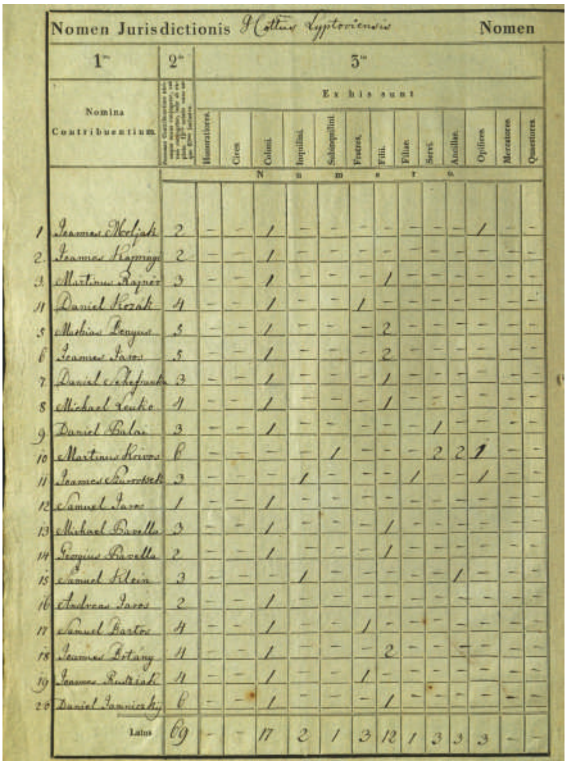
The use of artificial intelligence in a public collection environment is not entirely new in Hungary, as the NAH has previously used software processing and reading assistance, for example, to index the *Hungarian prisoners of the Soviet camps* (In Hungarian: Szovjet táborok magyar foglyai) database. However, the use of TRANSKRIBUS handwriting recognition software for this purpose can be considered an innovation. In Hungary, the Digital Humanities Centre of the Petőfi Literary Museum (PLM) – the department has since been integrated into the National Széchényi Library – was the first to start testing it. In January 2021, the staff of the Petőfi Literary Museum and the PLM Digital Humanities Centre held a workshop, where Eszter Mihály and Kata Szűcs, digital humanities experts of the PLM shared their experiences of using the TRANSKRIBUS software with the NAH staff.

The framework for the practical application of Handwritten Text Recognition (HTR) in archives was provided by the international project called *European Digital Treasures (www.digitaltreasures.eu)*, in which the national archives of Malta, Norway, Portugal and Spain collaborated.

2. ARCHIVAL DOCUMENTS USED

During the preparation phase, based on unifom criteria, each of the archives participating in the project decided on which documents they would process. The selection of archival records with a uniform structure, a uniform time frame and, as far as possible, a uniform handwriting was intended to help the efficiency of machine handwriting recognition. In addition, the needs of researchers were also a major consideration. Based on the selection criteria, in Malta emigration registers, in Norway address registers, in Portugal pardon applications, and in Spain passport registers were processed by volunteers under the guidance of archivists. In Hungary, the 1828 census (reference code: HU MNL OL N 26, official name: Census ordered on the basis of Act VII of 1827) was chosen for this purpose. The selection of the records was significantly influenced by the favourable reception of the 1715 and 1720 censuses, which had been processed previously by archivists at NAH.

The digital version of the 1828 census consists of about 170,000 image files and contains the census sheets of 53 administrative units (counties) and the free royal towns. The census was created in the 19th century to measure the taxpaying capacity of the population, and thus summarises names and economic data in tabular form. There were no uniform spelling rules and guidelines at the time, which, in addition to the handwritten recording of data, makes transcription even more difficult, and the peculiarities of the Hungarian language can also be observed in the documents.



Nomen Jurisdictionis I. Cottus Lyptoviensis Nomen

	1mo	2do	3tio												
			Ex his sunt												
	Nomina Contribuentium	[illegible]	Honoratiores.	Cives	Coloni	Inquilini	Subinquilini	Fratres	Filii	Filiae	Servi	Ancillae	Opifices	Mercatores	Quaestores
		N u m e r u s													
1	Joannes Moljak	2	–	–	1	–	–	–	–	–	–	–	1	–	–
2	Joannes Kapringi	2	–	–	1	–	–	–	–	–	–	–	–	–	–
3	Martinus Rajnér	3	–	–	1	–	–	–	1	–	–	–	–	–	–
4	Daniel Kozák	4	–	–	1	–	–	1	–	–	–	–	–	–	–
5	Mathias Bengus	5	–	–	1	–	–	–	2	–	–	–	–	–	–
6	Joannes Jaros	5	–	–	1	–	–	–	2	–	–	–	–	–	–
7	Daniel Schefranko	3	–	–	1	–	–	–	1	–	–	–	–	–	–
8	Michael Lenko	4	–	–	1	–	–	–	1	–	–	–	–	–	–
9	Daniel Balai	3	–	–	1	–	–	–	–	–	1	–	–	–	–
10	Martinus Krivos	6	–	–	–	–	1	–	–	–	2	2	1	–	–
11	Joannes Burovtsek	3	–	–	–	1	–	–	–	1	–	–	1	–	–
12	Samuel Jaros	1	–	–	1	–	–	–	–	–	–	–	–	–	–
13	Michael Pavella	3	–	–	1	–	–	–	1	–	–	–	–	–	–
14	Georgius Pavella	2	–	–	1	–	–	–	1	–	–	–	–	–	–
15	Samuel Klein	3	–	–	–	1	–	–	–	–	–	1	–	–	–
16	Andreas Jaros	2	–	–	1	–	–	–	–	–	–	–	–	–	–
17	Samuel Bartos	4	–	–	1	–	–	1	–	–	–	–	–	–	–
18	Joannes Botány	4	–	–	1	–	–	–	2	–	–	–	–	–	–
19	Joannes Rudziak	4	–	–	1	–	–	1	–	–	–	–	–	–	–
20	Daniel Jamniczky	6	–	–	1	–	–	–	1	–	–	–	–	–	–
	Latus	69	–	–	17	2	1	3	12	1	3	3	3	–	–

Picture 1: A page of the 1828 census (Reference code: HU MNL OL N 26)

3. THE VOLUNTEER PROGRAMME

In a crowdsourcing project, each member of the crowd contributes only a small part to the overall task, but as the many small parts come together to form a big whole, the end result adds significant value to the activity. The volunteer programme of the international project in Hungary lasted 6 weeks, from 1 October to 15 November 2021. The recruitment of volunteers started primarily on social media, namely on the Facebook page of the NAH, and it spread rapidly among genealogists through the world wide web with shares. By the time the institution published the call on its website and in the press, around a hundred people had already applied. The application was therefore closed within two or three days. Hungary coordinated the largest number of contributors among the participating countries. Such great interest was certainly also due to the fact that the task could be done online, at any time, from anywhere.

The NAH did not select the volunteers, all applicants were given the opportunity to participate in the project, but at the same time the volunteers had the opportunity to test their own skills. In the form of a quiz for beginners and advanced volunteers, visitors could test their ability to read the names of the census. They also filled in a questionnaire as part of the pre-registration process, answering questions such as whether they had experience in transcription and in genealogy. With this method the archives managed to reach people who were already familiar with archival research. Though there was no restriction on the level of education of the applicants, it can be stated that mostly qualified people with foreign language skills decided to apply. The opening event took place in the main building in Budapest, though the closing event could only be held online due to the pandemic situation. Both events were streamed live by the archives, and the recordings were also available for viewing. Each volunteer undertook to transcribe a minimum of 50 pages, that is to say 1000 names. A total of 70 applicants completed the programme, of which 24 asked for additional pages again and again.

For volunteers from different backgrounds, participating in the professional archival work was a double challenge: they had to prove themselves both in terms of information technology and palaeography. In terms of social background, it was mainly intellectuals and recent retirees with experience in genealogy and indexing who joined the work. The average age was 50.9 years. Looking at the geographical distribution, we find people living in both small villages and larger towns. Thus, the opportunity for flexible working was also taken advantage of by Hungarians living and working in different parts of the world. Hungarians from Transylvania and Transcarpathia, Hungarians living/working abroad from Italy, the United States, the United Kingdom and Sweden participated in the programme.

3.1 THE WORKFLOW

The work was started by two archivists who transcribed 400 pages of names from different municipalities on Transkribus. The data was sent to Transkriptorium, an IT company at the Technical University of Valencia in Spain. The Spanish computer scientists used the data to create a handwriting recognition model and ran it on a subset of the 1828 census, that is to say on about 30% of the census sheets. Once the NAH received this data from Spain, the 70 volunteers recruited earlier were also involved in the workflow, which at this point included a verification phase. Thus, with the involvement of human resources, the transcription carried out by the handwriting recognition software was checked, corrected or approved. The software, thus perfected, was finally run over the entire census file, creating the final result, a searchable, online publishable database of the 1828 census.

Volunteers worked on a designated, easy-to-use, user-friendly interface created by Transkriptorium, which had an automatic save function. The site was accessible to contributors with a username and password.

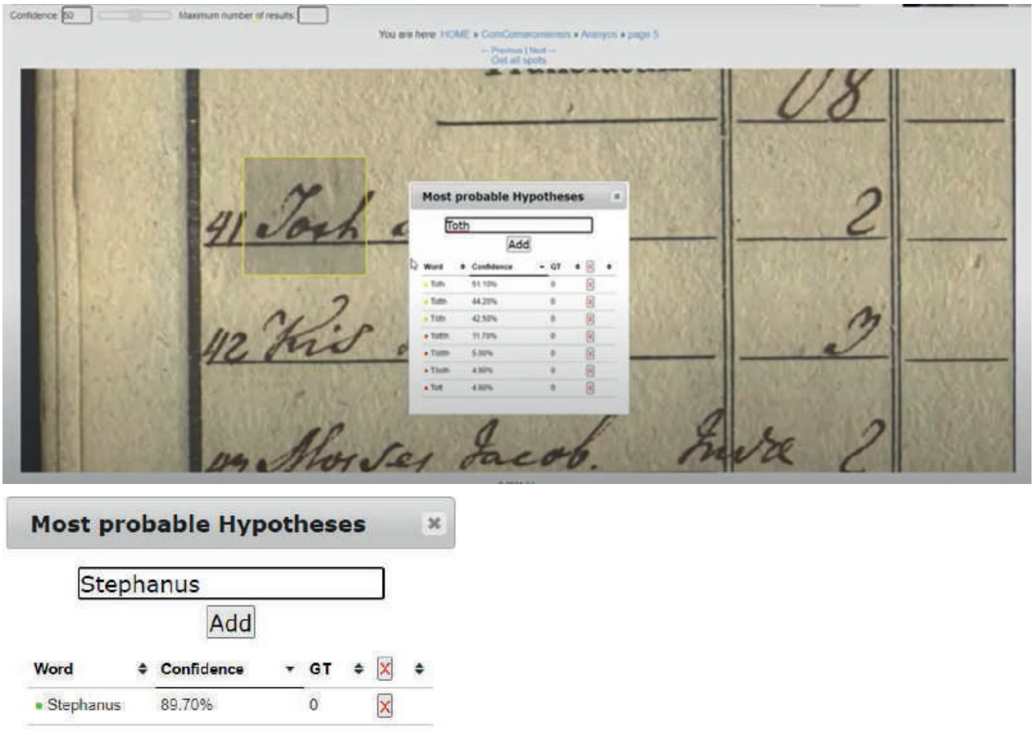
Many people may wonder why the census records of the free royal towns provided a large part of the records to be checked by the volunteers. The census rolls of the free royal towns are representative of the whole archival material, as they cover the whole territory of historical Hungary and their writing is sufficiently varied to serve as a sample for the whole documentary material. Volunteers were given the opportunity to choose the municipalities they wished to check, which was particularly encouraging. Most genealogists have a strong emotional attachment to their place of birth and their immediate surroundings, and are often very familiar with the names of their own geographical region. Volunteers navigated and worked in a folder structure that reflected the structure of the original records, and the archives were able to provide at least partial access for each person to check the names of the people in the settlements of their choice. Beyond the free royal towns, however, partial verification of the census records of certain counties was eventually carried out, thanks to the outstanding enthusiasm of some volunteers who, after completing the basic 50-page dossier, asked for and were given additional tasks. Extra work was typically undertaken by more experienced genealogists.

3.2 TRAINING THE ALGORITHM

The archives were eager to receive the processed documents back from Spain, which showed the extent to which the software could recognise handwritten 19th century names after 400 pages of training. The results were generally satisfactory, yet further improvements to the software were expected after the volunteer validation.

The transcribed words were surrounded by text boxes, which marked the area of the page to which the transcription belonged. The text boxes varied in colour, different colours indicated different probabilities of transcription. The colour codes were as follows: green represented high, yellow medium and red low probability. In the case of a grey text box, the algorithm did not suggest any transcription. The blue colour code became the colour of GT (Ground Truth), once a word was selected as the correct solution, the text box changed to blue, indicating that there was nothing more to do with the word and the solution was final.

By clicking on the text boxes, the transcription suggestions for that word were displayed, along with a percentage of how the algorithm rated the transcription of that word. The software did surprisingly well with transcribing common first names, often recognising words/names with a probability of 90-95%. In these cases, the volunteers' only task was to approve the solution offered. In more difficult cases, they had to look through a longer list and select the correct transcription, which could appear in fifth position or even later. In some cases, they had to provide the correct solution on their own. Volunteers had the opportunity to skip some of the more difficult words. The aim was not to find the solution to every word, but to make sure that the one they chose was the right one. Only then can the software be trained correctly.



Confidence: 50 Maximum number of results:

You are here HOME » ComComaromensis » Aranyos » page 5

← Previous | Next →
Get all spots

Most probable Hypotheses

Toth

Add

Word	Confidence	GT	☒
● Toth	51.10%	0	☒
● Toth	44.20%	0	☒
● Tóth	42.50%	0	☒
● Totth	11.70%	0	☒
● Tôth	5.00%	0	☒
● Tsoth	4.90%	0	☒
● Tot	4.80%	0	☒

Most probable Hypotheses

Stephanus

Add

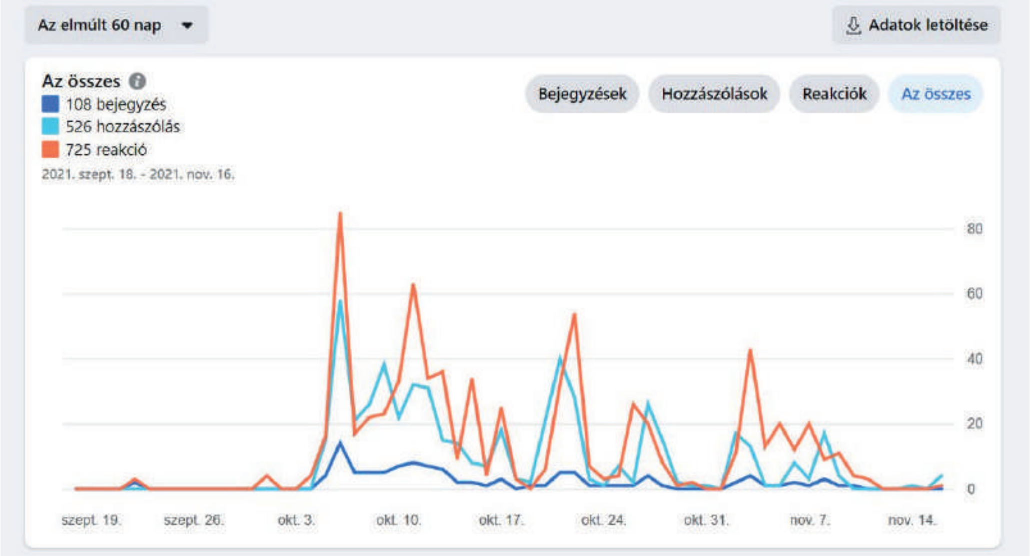
Word	Confidence	GT	☒
● Stephanus	89.70%	0	☒

Picture 2: Part of the platform provided by Transkriptorium for GT (Ground Truth) selection

The work of the volunteers was constantly monitored by the archivists on the online work platform, and once the work had started, they checked individually that everyone had understood the task. The work done by volunteers exceeded all expectations in terms of quantity. 70 active volunteers checked a total of 6,787 pages and 13,5740 names in the six weeks available. The oldest was a 78-year-old lady who alone checked 309 pages. After the work was completed, spot checks were carried out in the archives and after minor corrections, the new data was sent to Spain so that the Spanish IT specialists could run it on the full census once the model had been perfected.

3.3 SUPPORTING MATERIAL

The archives supported the work of the volunteers throughout the programme by providing them with professional support materials. How the platform works and the information about the tasks to be carried out by the volunteers were summarised in a Task Guide by Ildikó Szerényi. In addition, a *Paleographic Example Booklet (Paleográfiai Példatár (1828)*[2], – a *Help for reading and transcribing the names in the 1828 national census* – was also published by the NAH on its website, on a special subpage dedicated to the EDT project. Volunteers also received up-to-date, interactive help in a private Facebook group called *Volunteers of the 1828 Census*, where informal discussions, comments and playful posts were published. Over the six weeks, the group generated 108 posts, 526 comments and 725 reactions, which shows an extraordinary level of activity.



Picture 3: Screenshot of activity in the private Facebook group called Volunteers of the 1828 Census

3.4 FEEDBACK

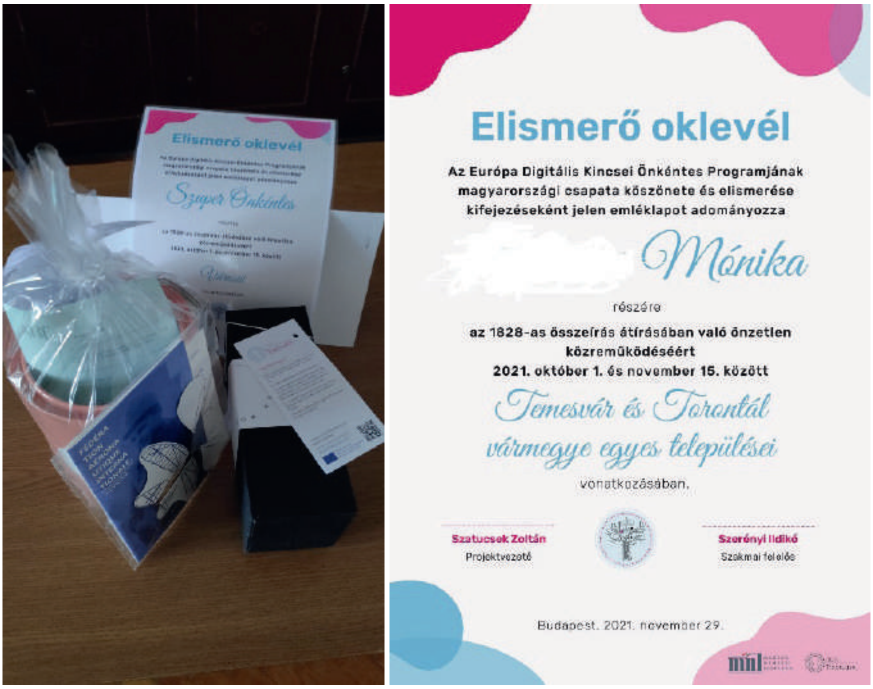
It is of utmost importance for the National Archives of Hungary to maintain a good relationship with the research community. To reach this goal, the archives asked volunteers for feedback on their experiences and opinions. The questionnaire was filled in by 50 people who rated the programme on a scale of 1 to 5, with the overall programme scor-

[2] Szerényi, I., Kántás, B., H. Németh, I., & Szatucsek, Z. (2021). Paleográfia példatár (1828). Az 1828. évi országos összeírásban szereplő nevek kiolvasásához és átírásához. (I. H Németh & Z. Szatucsek, Eds.). Budapest: Magyar Nemzeti Levéltár.

ing 4.82 points and the Facebook group activities scoring 4.47 points. The vast majority would like to participate in similar activities in the future.

3.5 PRIZES

As its name suggests, volunteering is essentially work done for no financial reward. However, the NAH from the beginning intended to honour the work of volunteers, since they had supported the cause of Hungarian culture in a selfless way, sacrificing their free time and performing high-quality work. Taking all this into account, at the end of the programme, the archives offered a DNA test, and historical magazine subscriptions as prizes among the volunteers who participated in the programme. These prizes were won by many. In addition, as a special prize, the 5 most hard-working and outstanding professionals were awarded with various souvenirs. Each participant received a commemorative card with their name on it.



Elismerő oklevél

Az Európa Digitális Kincsei Önkéntes Programjának
magyarországi csapata köszönete és elismerése
kifejezéseként jelen emléklapot adományozza

[redacted] Mónika

részére

az 1828-as összeírás átírásában való önzetlen
közreműködéséért
2021. október 1. és november 15. között

Temesvár és Torontál
vármegye egyes települései

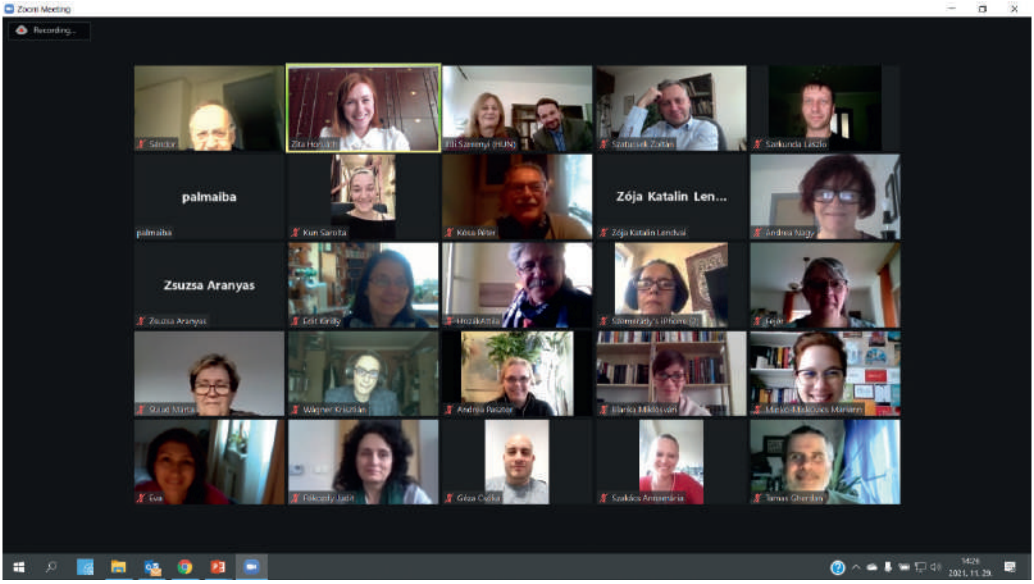
vonatkozásában.

Szatucsek Zoltán
Projektvezető

Szerényi Ildikó
Szakmai felelős

Budapest, 2021. november 29.

Picture 4: Gift pack and certificate of appreciation presented to volunteers



Picture 5: Screenshot of the Volunteer Closing Ceremony (Zoom meeting) dated 11. November 2021.

3. EXPERIENCES

Within this project, the NAH combined the new technology of artificial intelligence with crowdsourcing in an innovative way to serve multiple social goals simultaneously. As with all pilot projects, there were of course difficulties in indexing the 1827 national census. The uncertainty created by the pandemic was a challenge from the beginning. Basically, a team of people who were complete strangers to each other, had to work together in a team. Social media, however, was a great help, as the private Facebook group served as the main communication forum for volunteers and archivists from completely different places of residence to meet.

During the workflow, when volunteers validated the transcriptions made by the algorithm, although everyone worked in a folder designated to them, all users were able to access all pages. This situation could only have been changed by IT improvements, which would obviously have had budgetary implications. The experience gained from this work shows that some volunteers would have requested exclusive access to their folders, so the Archives will consider the feasibility of this in future projects.

In the European Digital Treasures project, the institutions in each participating country initially planned to employ between 20 and 25 volunteers per country, and the number of archival professionals involved in the project was set accordingly. However, the NAH soon faced a huge and unexpected over-subscription. As already mentioned, in the space of two or three days, around 100 people applied during the pre-registration process, of whom 70 completed the programme. The institution offered all applicants the opportunity to participate, and although the higher than expected number of volunteers was welcome, their coordination also meant an additional challenge for the archivists.

The archives were also surprised by the high number of applicants. The 'anytime, anywhere', individual time-scheduled approach emphasised in the call for applications certainly played a role in attracting a large number of interested people. At the same time, the popularity of genealogy in Hungary also explains the high level of interest from the

Hungarian side. The NAH has been developing its databases and services for many years, building its relations with the public, organising workshops, events and lectures at national level. Relying on the experience gained over the years and the appreciation of volunteering, 70 volunteers enthusiastically completed the programme. The greatest benefit of this pilot project from the Hungarian side was social activity.

As for the effectiveness of handwriting recognition, it can be stated as a conclusion, that an improvement could be shown for all the manuscript collections of the participating countries as a consequence of re-training the algorithm. However, this improvement was not always as significant as expected from the quantitative data. The question arises as to whether it is more efficient to rely on the large amount of data produced by crowdsourcing processes or to concentrate resources on producing higher quality GT. Considering all the results reported here, the latter may be the preferable alternative.

The project also considered equal opportunities aspects, as the task, which could be done from home and required practically only an active internet connection and a computer, could be done by anyone from anywhere, without any quantitative and temporal restrictions. The volunteers, mostly of senior age, included several people with mobility problems and one hearing impaired person who was able to communicate and work as a full member of the group thanks to the detailed information material published online and the virtual communication channel (mailing list, Facebook group, chat) that was constantly available.

Although the voluntary project was originally a national social initiative, it eventually extended beyond national borders, with the participation of many Hungarian citizens living abroad or permanently abroad. We can therefore say that the love for Hungarian culture and history, the desire to preserve the common past and to carry out some kind of useful community activity for this purpose united people regardless of their age and geographical location, so that by the end of the project not only a cultural and historical product was created, but also a real human community based on mutual support, which strengthened social cohesion and whose cooperation is expected to continue in the future.

REFERENCES

Kahle, P., Colutto, S., Hackl, G., & G. Muhlberger. (2017). Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. https://doi.org/10.1109/icdar.2017.307.

Plüss, R., & C. Sieber. (2020). Digitalisierungsprojekte des Staatsarchivs Zürich mit Einsatz von Machine-Learning-Verfahren. *ABI Technik 40*(3), 218–228. https://doi.org/10.1515/abitech-2020-2018.

Rabus, A. (2019). Recognizing Handwritten Text in Slavic Manuscripts: a Neural-Network Approach Using Transkribus. *Scripta & E-Scripta 19*, 9–32. https://www.ceeol.com/search/article-detail?id=793865.

Rothballer, M. (2021) *Transkribus. Erfahrungsbericht zu maschinellem Lernen und Handwritten Text Recognition in der Heimat- und Familienforschung*. https://archivalia.hypotheses.org/124394.

Tomić, M., Grzunov, L., & M. Dragija Ivanović. (2021). Crowdsourcing transcription of historical manuscripts: Citizen science as a force of revealing historical evidence from Croatian Glagolitic manuscripts. *Education for Information 37*(4), 443–464. https://doi.org/10.3233/efi-211555.

TYPOLOGY: 1.02 Review article