Co-funded by the
Creative Europe Programme
of the European Union

EUROPEAN DIGITAL TREASURES: MANAGEMENT OF CENTENNIAL ARCHIVES IN THE 21ST CENTURY

Activity 21 - Crowdsourcing

Final report

EXECUTIVE SUMMARY

This report has been prepared in the context of the Digital Treasures for Europe 21 activity, with the aim of presenting the work carried out in this subproject, summarising the results and lessons learned.

The aim of the activity was to work on active ageing; to show how retired people can contribute to and benefit from the Archives; and to use a joint method to identify, tutor and supervise new active elder users.

To implement the activity, the five participating national archives set a novel objective that went beyond the original project objectives and was significantly more complex. Combining the new technology of handwriting recognition (HTR) with the involvement of the silver generation was an ambitious undertaking but thanks to the effective coordination and the partners' commitment led to a long-term beneficial outcome.

It is important to note that it took place during a pandemic period and that it was a pilot test.

As key indicator for the activity 120 participants was expected, the project exceeded its original target of 120 participants with a total of 135 from 15 countries of three continents. Quantifying the impact was not easy, since it might result in the validation of metadata, but for the objectives it was more important to see the process and the level of satisfaction, enjoyment and gratification of the participants. Participants from the older generation were extremely positive about the experience and expressed their willingness to continue working together.

As a tangible output the quality of the HTR were good and the archives are in the process of being published. The indicator of the quality WER was varied in the different collections from 10.4 to 28.4 depending of the quality of the originals.

Impact of the pilot was multifold. Service delivery, organisational change, public awareness, training was equally mentioned by the partners distinguishing their short, medium and long-term effect.

- Such an activity should be especially examined as a possible way to **render their holdings more accessible**,

- especially because of **the use of technology** which enables searchability through handwritten text in digitised images.
- This activity needs some **organizational developments** for long term planning and new focuses on training.
- The involvement of persons from the general public could lead to **better awareness of the importance of the archives** within their societies. In this case, there was an emphasis on the involvement of persons from the silver generation which could help with attracting new users from that specific target group.
- This activity provides a **useful pastime for volunteers**, who are typically recently retired. The project provided equal opportunities for people living in smaller villages and larger towns, and even allowed volunteers from across the whole world to get involved. Archives contribute to giving new roles to retired people in society and helping them gaining new skills.
- By this pilot an incipient **community of volunteers has been created**, with a sense of belonging and confidence in the archives that can be developed in the médium-long term.

All the archivists agreed that this kind of cooperation is essential in order to be able to carry out similar van and technology programmes in the future.

IMPLEMENTATION

The activity started on 15th of January 2021 and concluded 30th of November thesame year. For the implementation, a collaboration between EDT and "The Pattern Recognition and Human Language Technology RESEARCH CENTER" of the Polytechnic University of Valencia and its spin-off tranSkriptorium was established. The five archives planned the implementation of the project together with the technology partner. This consisted of preparatory and validation phases and at the end of the project, the technology partner carried out the final validation.

1. Preparatory stage

Preparatory stage consisted of three steps and was completed by July 2021:

1.a Selection. Each archive selected a small portion of the collection to be indexed. The selection must have been representative of the issues to be found in the full collection, like quality of the images (degradation, bleed-through, writing style, scanning conditions illumination, resolution, layout of the images, set of characters. The amount of images for each archive must have consisted of 8000 handwritten or printed words. These images were sent to the technical partner to be reviewed.

1.b Producing initial ground truth and preliminary recognition. Archives must have annotated adequate quantity of words to support model building. Each archive was in charge of producing the GT for its selected representative images. This includes manually amending the baselines of each text line and producing the corresponding transcript. This process of GT preparation was produced using the Transkribus client, which provided accurate automatic line detection and comfortable options for manual transcription and baseline correction.

1.c Preliminary recognition. The selected images were sent to the technical partner for training and testing the indexing system. Followed by refinements and correction iterations, probabilistic index was produced for the full collection and a corresponding search server was developed for each archive. After the test period this server provided capabilities to allow silver generation volunteers to check, validate and/or amend any indexed word hypotheses they see fit.

In all five archives records were selected as useful data for family historians and silver generation persons looking for family roots.

Norway chose the National Population Register, established in 1960, mainly consist of registration cards of the permanent residents. The register cards give a lot of information that are highly appreciated by family historians. To create the training set, NAN has chosen 200 index cards, spread over several years from the time period of 1906-1914 with the representation of different handwritings. The Norwegian team received the first batch with automatically segmented pages on 29 March and finished working with them on 5 April 2021. The second batch was received on 5 April and finished by May.

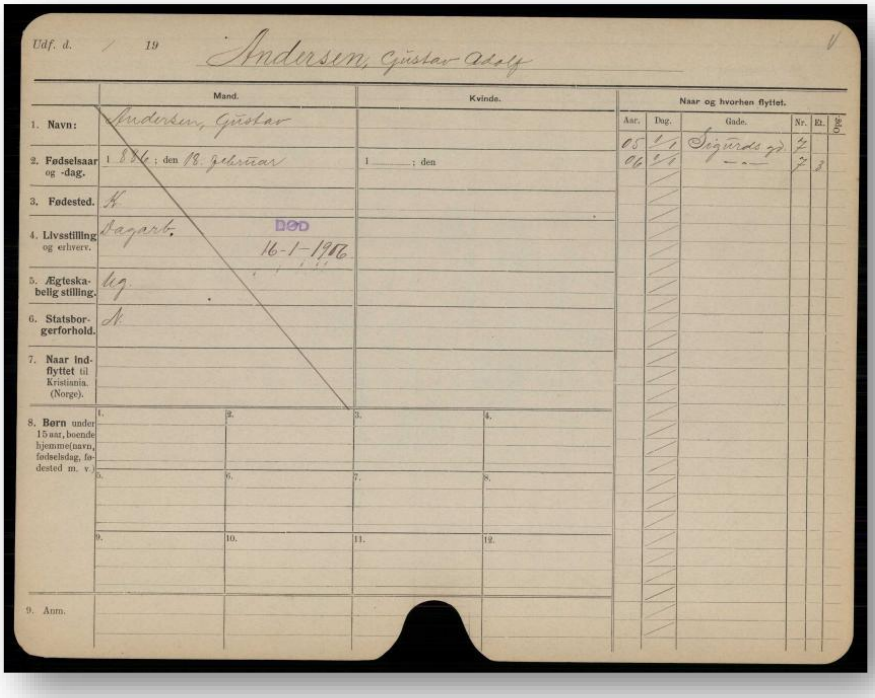

Figure 1. Example of text images from the EDT-Norway collection.

The DGLAB, Portugal documentation was the General Register of Charges of King John VI between 1792-1826 form the General Registry of Misericórdias. The purpose of the General Register of Misericórdias was to record all the misericórdias granted by the Crown, such as land grants, alcaidarias-mores, rents, jurisdictions, letters of recommendation, captaincies, offices, and justice and treasury positions, leases or privileges. 38 images were transcribed with 13370 words, which corresponds to 5% of a Book of 1754 pages. The GT was around 13370 words long.

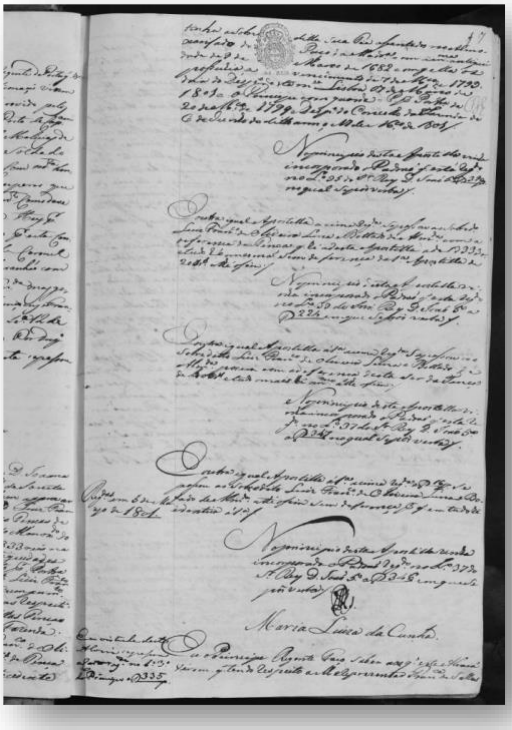

Figure 2. Example of text images from the Portugal collection.

In **Hungary**, National Census of 1828 was carried out to assess the tax capacity of the population. As it contains names and economic data and covers the entire territory of Greater Hungary, altogether 53 administrative regions and royal free towns, it is an invaluable source for family researchers. To create the training set 314 pages from 8 settlements had been chosen from 8 different regions in order to represent different handwriting styles in the training set. The Hungarian team received the automatically segmented pages on 11th March and finished working with them on 31st March 2021. Two archivists were transcribing and handling the technical questions, and an other specialist checked the transcriptions.

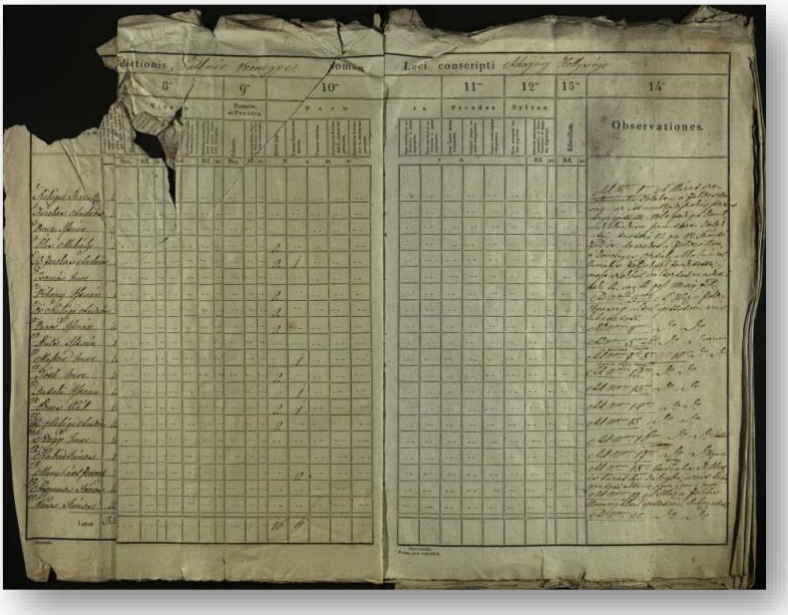

Figure 3. Example of text images from the Hungary collection.

The choice of the National Archives of **Malta** was from immigration records from the Customs Department Fonds, covering the periods of 1905-1966, having the name of persons and dates arriving to / departing from Malta in alphabetical order. For the initial Ground Truth, a representative sample of 51 images was selected. Two members of Malta staff made the transcription with tagging the data elements.

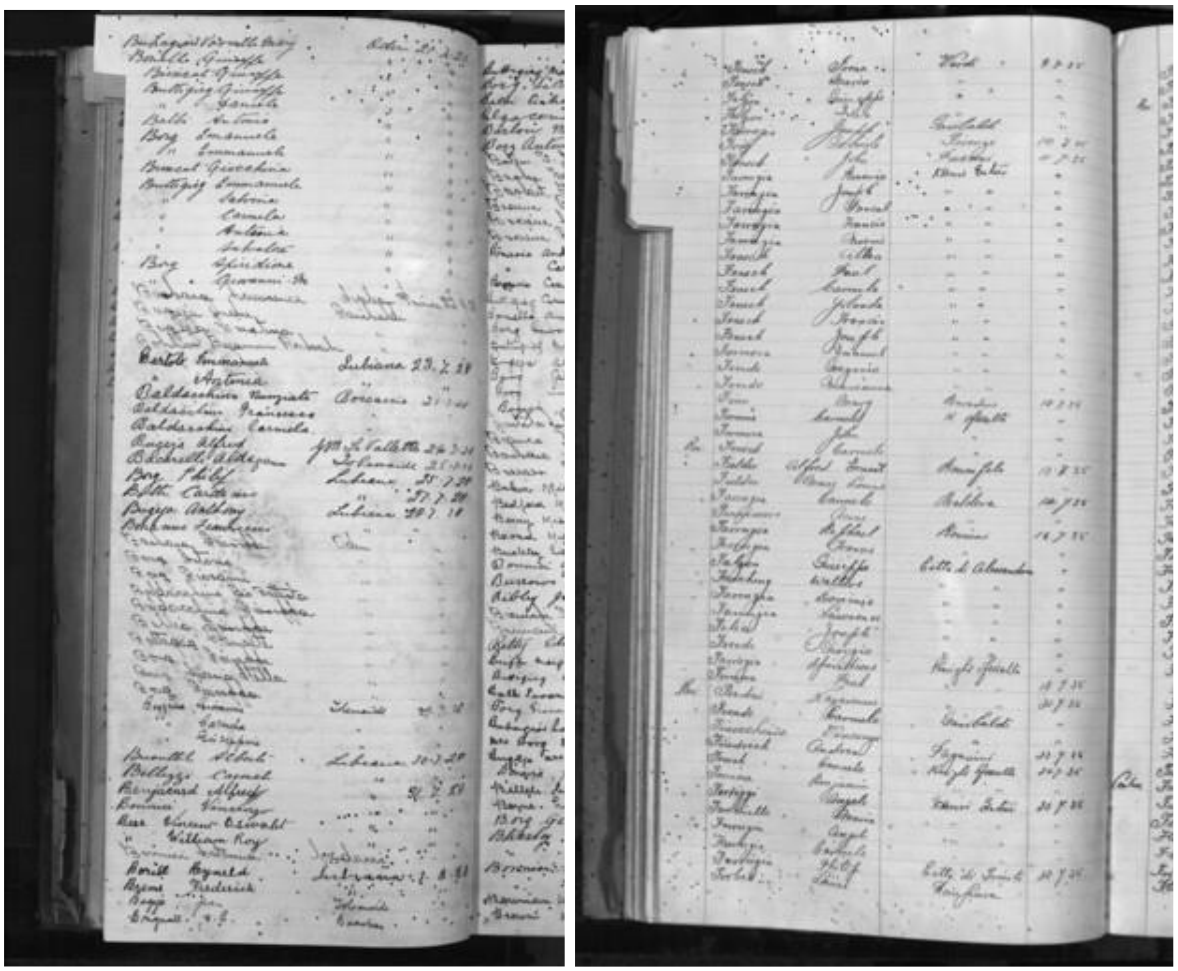

Figure 4. Examples of text images from the Malta collection.

Spain selected a documentation on Spanish emigration to Argentina. The collection is a series of Passport Issuance Register of the Consulate General of Spain in Buenos Aires between 1936 and 1939. A documentary series well suited for indexing and retrieving information. For the GT, the most representative 100 images were selected, and after an interim observation the first 70 images were annotated. DSGAE have tagged most important data too. Transcribing was made by Archival Workshop School: Two people from Archival Worskshop School, and one for DSGAE administrative staff, supervised by a senior archivist.

Figure 5. Examples of text images from the Spanish collection.

The automatic segmentation of the selected area worked out well for each team. In some cases names were left out from the segmentation and new baselines needed to be added or sometimes deleted. Regarding the length and curve of the baselines, we made corrections in most of the cases. It should be further observed whether segmentatin and baseline detection is really worth to do independently from Trankribus.

Communication between the teams and most importantly the support of tranSkriptorium helped getting through the initial difficulties and each team became more familiar with the platform. In Spain and Hungary transcription and paleography guidelines have been made. An MS Teams platform was established for sharing experiences.

Most frequent problems were how to handle unnecessary information, preprinted forms and abbreviations. Some IT help was needed for installing the client software, but altogether it was considerably easier than expected. Some issues came up when marking the baselines because of the vertical and irregular lines as well as tagging concepts to be retrieved.

Implementation was slightly different from the plan, since direct comparison was impossible; however, generally-speaking the activity was in two month delay.

For some archives, namely Norway, Portugal and Hungary the issue remains that the total size of the collection exceeded the 30,000 pages of processing capacity contracted. It is undoubtedly in the legitimate interest of the partners to apply the HTR model established to their entire collection selected.

2. Validation stage

Parallel with tranScriptorium model building, the community building and crowdsourcing preparation took place. Although it was more dependent from the structures, traditions and experiences of the partners, some consultations are being prepared with experienced institutions. We had two workshops: one with TNA UK and another one with National Archives of Norway, as these two national archives have a long tradition of successful crowdsourcing.

Recruited volunteers validated and amended words of the preliminary recognition. These words were stored and made available to the archivists through a specific web interface. Each validated or amended word was accompanied with complementary information such as the IP, the timestamp of the last time this word was visited.

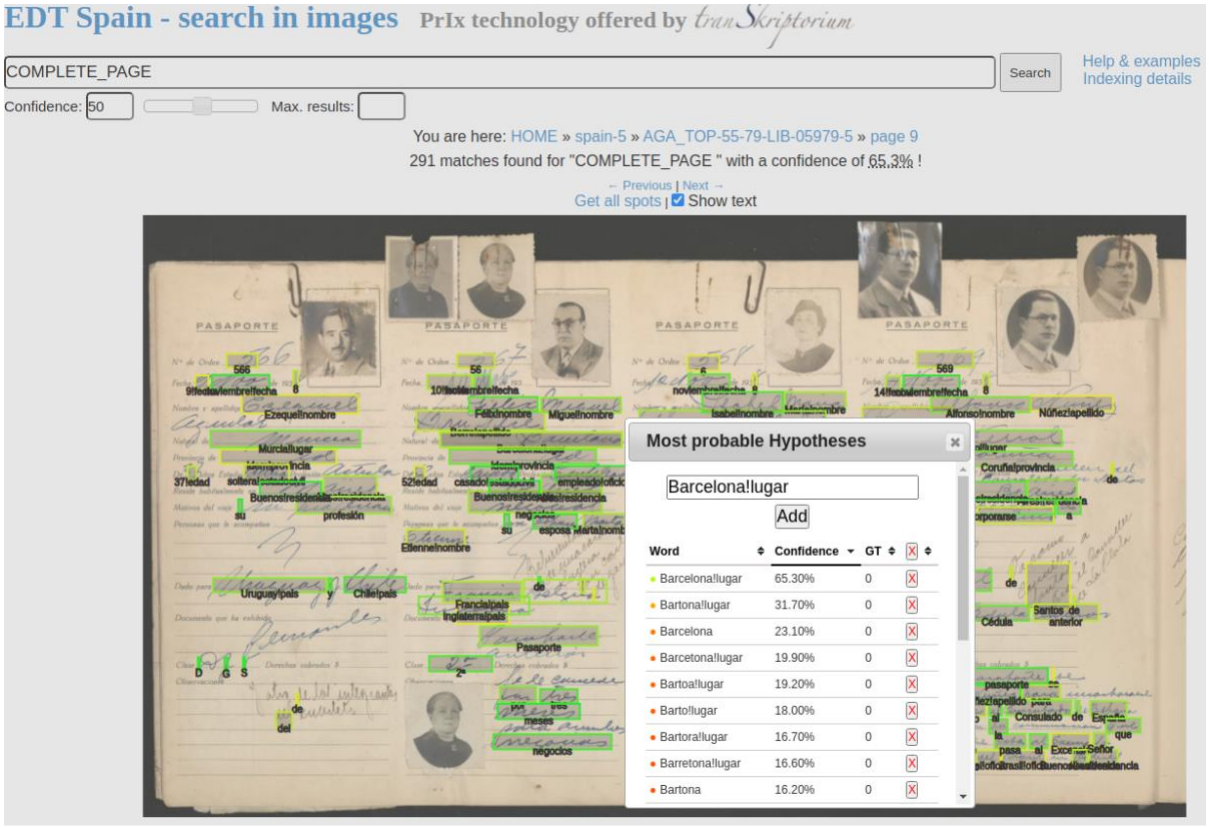

Figure 6. Crowdsourcing interface on the Spanish passport collection.

Indicators of the validation stage

The COVID epidemic has significantly affected the involvement of volunteers. The project initially planned to have 30 archives visits per partner on 5 occasions, but the epidemic limited the number of face-to-face meetings. Partly due to differences in national regulations and partly due to the capacity of the available premises, partners achieved different results.

There was also a wide diversity in the number of participants in the activity. As became clear during the implementation process, there are legal and social implications of involving volunteers. While in Spain there is no tradition of this form of participation, in Hungary there are over-regulated legal requirements for institutions, in Norway the involvement of volunteer indexers has long been part of archival work. Nevertheless, the project exceeded its original target of 120 participants with a total of 135.

Country	Number of volunteers
Malta	19 (2 UK, 1 FR, 1 AU
Hungary	70 (1 UK, 1 SK, 1 UA, 1 RO
Spain	16 (1 ECU, 1 ARG)
Portugal	25 (1 BR)
Norway	5
SUM	**135**

In its initial phase, the **National Archives of Malta** (NAM) issued two calls for volunteers on its Facebook page. Interested persons were directed to get in contact through the NAM's customer care email whereby more information on the activity was shared. These were divided into five groups and a one-hour long training session for each one was delivered via Zoom. These training sessions, that replaced the onsite visits in view of COVID-19 restrictions, included a presentation on the HTR system and step-by-step instructions about their task. In anticipation of these sessions, a set of usage guidelines was sent to each person; these were also to be used during the transcription process.

Each volunteer was allocated a number of pages to work on, with the maximum being 50 pages. A number of them managed to complete all 50 pages by the task's due date, and thus, they chose to proceed on the transcription of more text.

Continuous support was given to each participant throughout the duration of the activity. Such help consisted of the below:

- Feedback after completion of the first few pages.
- Volunteers were also offered the possibility to contact the National Archives of Malta in case of any difficulties through telephone/mobile, email and online meetings.

In occasion of the closing of the crowdsourcing activity, an online event took place on the 17th December 2021. All volunteers were invited to a meeting whereby appreciation towards their work was expressed.

During the programme the **National Archives of Hungary** involved 70 persons in the activity. Due to the pandemic situation and the fact that many volunteers from rural areas and abroad participated online, holding events that required personal presence was less popular. The opening and closing events were attended by 40 people each, though the classroom training and the exhibition tour as a reward attracted little interest.

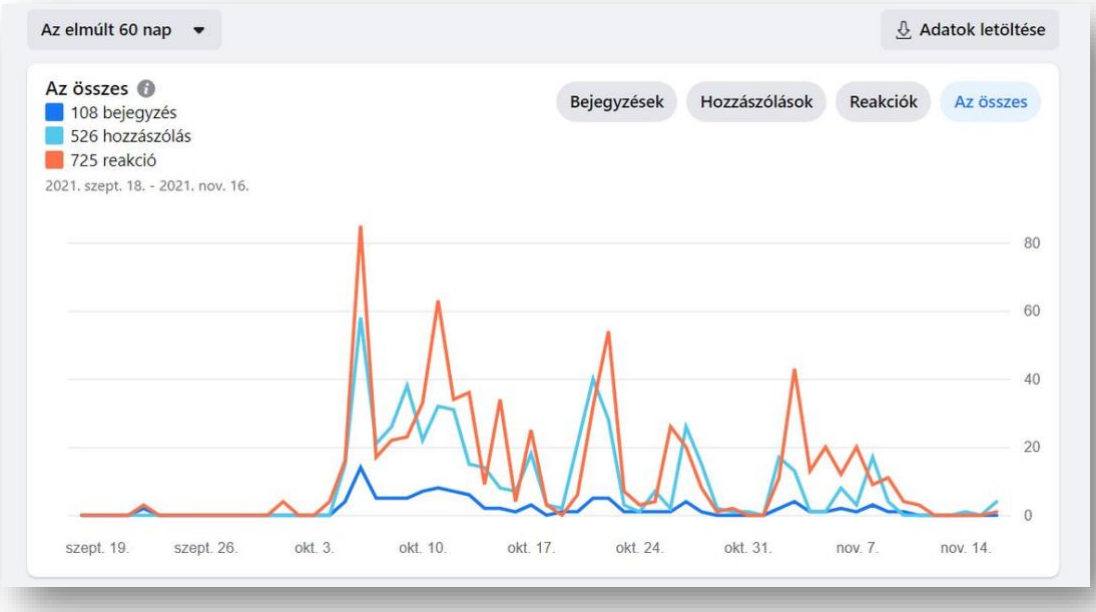

Figure 7. Screenshot of activity in the private Facebook group called Volunteers of the 1828 Census.

The initial activity in **Spain** consisted of a series of visits and workshops to national/state archives with a series of follow-ups and tutorials. The pandemic led to a reorientation of the activity. The Spanish State Archives had a total of 16 volunteers (14 from Spain, 1 from Ecuador and 1 from Argentina. 9 women and 7 men).

For the development of the activity, three online meetings and a face-to-face closing ceremony were held. In each meeting, a review of the state of development was carried out, doubts were solved, and new images were assigned. The closing ceremony count on a visit to the National Historical Archive, a presentation of results and a lunch with the participation of technicians from the Polytechnic University of Valencia.

In **Portugal** a total number of 25 volunteers were working on this project; 24 Portuguese and one Brazilian.

A total number of only 5 volunteers were working on the project in **Norway**. NAN announced for volunteers several times, but only five answered our request. We are working with several hundred volunteers in other projects, but for some reason they didn´t show interest in this project.

Difficulties in the validation stage

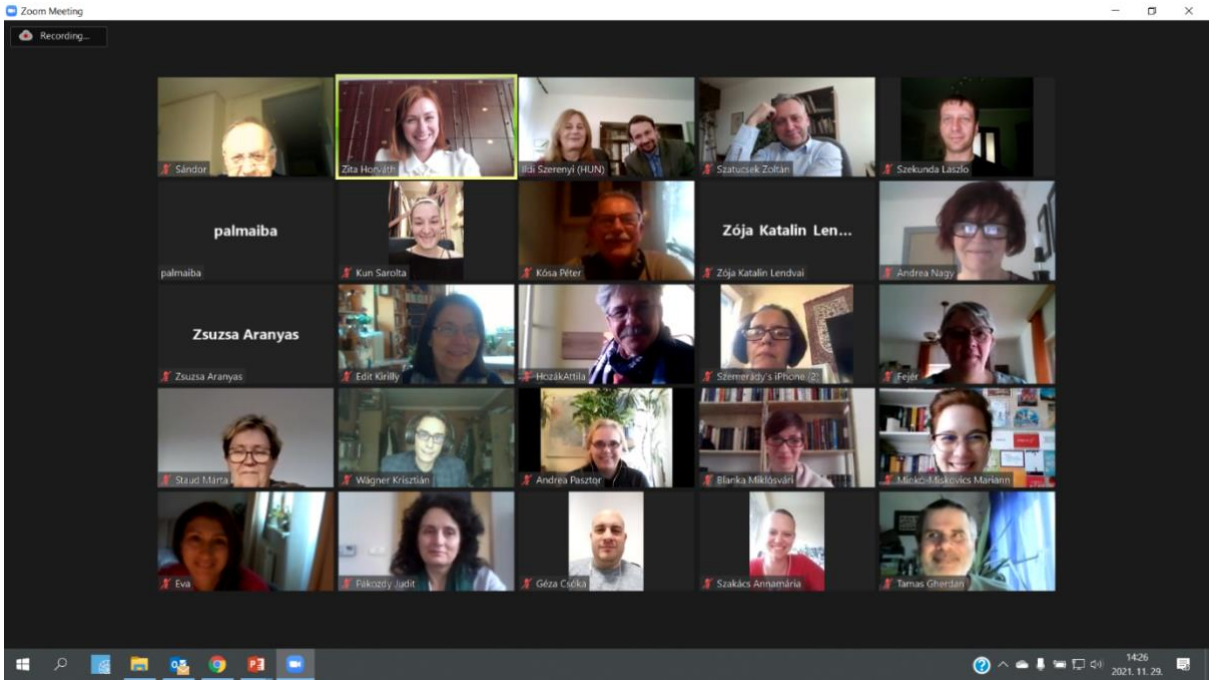

Figure 8. Closing event of the Hungarian validation stage.

Although the project met expectations both in terms of technical results and output indicators, the archivists had to overcome a number of difficulties and faced expected and unexpected challenges during the project.

- **Recruiting the volunteers.** As mentioned above, some partners had problems in contacting volunteers. This was partly due to the fact that all institutions except the Norwegian Archives had limited experience in this area and therefore their results varied widely. On the other hand, Malta, with a population of only half a million, managed to attract many contributors to the project, while in Hungary the large number of amateur family historians helped to overcome.
- **Level of IT literacy:** One requirement for participation in this activity was basic computer literacy. Nonetheless, level of IT knowledge varied across all volunteers and this considerably affected the pace and quality of the work. This particularly affected the silver generation as a priority target group for this exercise.
- **Technical difficulties:** Technical problems typical of IT based systems also made implementation difficult. The user interface used by the volunteers generally worked reliably, but there was not enough time for testing during the project, so some problems only came to light during the implementation phase.

  At the beginning of the exercise, a number of volunteers noted that the "Get All Spots" function was not working on several pages within the HTR system.

Another difficulty was the incorrect detection of lines (including the lack of detection of words within single lines). As mentioned in the technical part, warped pages and insect damage could have been a likely contributor. In such cases, participants could click on a specific spot to create a new bounding box whereby they could input the correct transcribed word. Volunteers also noticed that „#" for the ditto mark would appear as a hyphen in the "Most Probably Hypotheses List".

- **Difficulties with the identification of handwritten text**: Some records display a wide variety of handwriting and some of the volunteers found that it was difficult to decipher. Furthermore, the text included a lot of abbreviated words. A number of the volunteers were not used to this style of handwriting and therefore found it difficult to identify the full names. Due to this, in Malta a number of volunteers gave up the activity.

  The silver generation needed extra technical help to accomplish the programme. With a little care and assistance from both the archivists and the other younger volunteers, this generation has also fulfilled the task entrusted to them. Some of them also had a strong background in palaeography.

- **Organizational difficulties.** Due to the complexity of the project, the interdependent activities required tight schedules and concentrated use of resources. Several partners faced difficulties due to time constraints, lack of organisational resources to develop the activities and to follow up volunteers.

At the end of the validation of the project the technical partner provided the automatic transcript of each image with all the words corrected by silver generation voluteers, and made available all the four HTR-d collections through individual web interfaces.

EVALUATION

As part of the project, we felt it was important to study and analyse the results. This is based partly on feedback from the project's contributors throughout the project and partly on the responses of each partner to questionnaires after the project was completed.

Figure 9. Closing event of the Spanish presentation of results in the National Historical Archive.

Feedbacks from volunteers

Volunteer feedback was largely positive. Most of the volunteers replied that they found the activity to be interesting. Many mentioned that they were finding the work to be very enjoyable and offered to transcribe more pages. **In the case of all partners, volunteers stated that they were interested in participating in other future similar activities.** A number of volunteers commented that it was an educational experience, as users gained unique experiences while reading the documents.

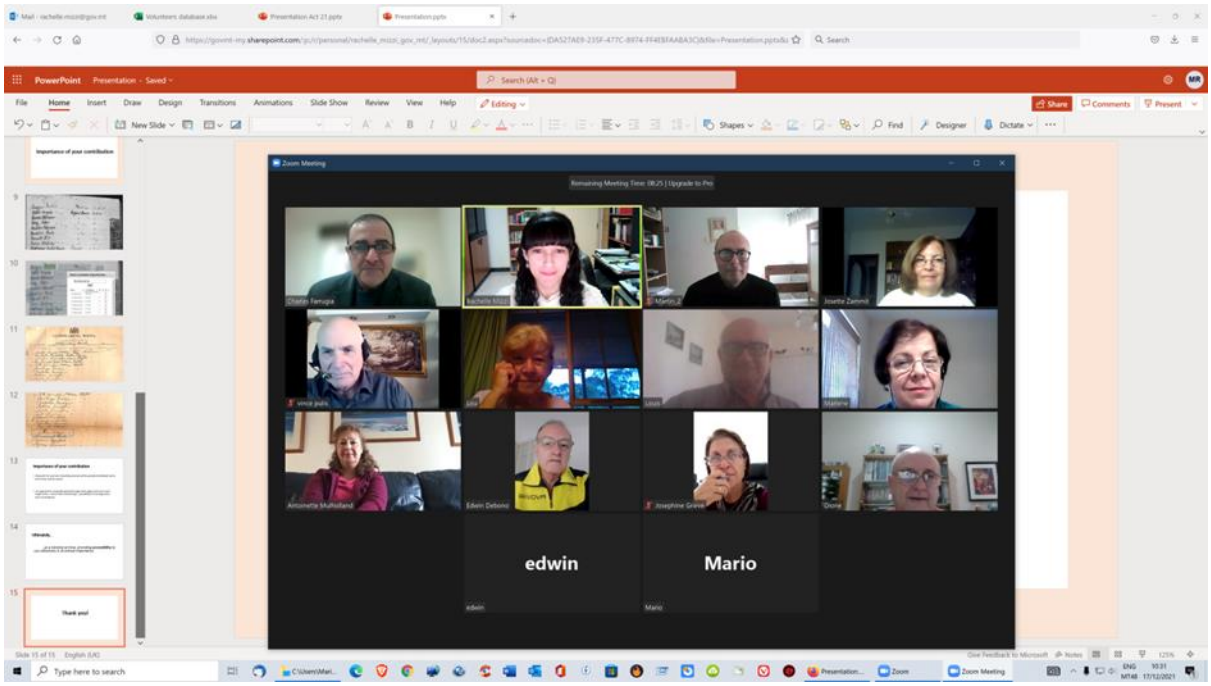

Figure 10. Closing event of the National Archives of Malta.

In Malta special comments were made on how coming across names of persons' and vessels' never heard before pique their curiosity and thus they would often proceed to look them up and read about their history. In the case of Spain and Hungary, there were many retired professionals and experienced genealogists. Both groups related closely to the archival world, so for them working with such an innovative technology as handwritten text recognition was very interesting. In addition, a very active knowledge exchange community was created among them.

On a personal level, a highlight was the rich cultural exchange due to the different profiles of the participants and the regional decentralization. The volunteers resided in different regions of the countries, the diaspora, expatriots, professionals and amateurs from other continents.

The motivations that led the volunteers to participate in the activity were very varied, as were their profiles, but, in general, the main objective of a crowdsourcing activity was achieved: generation of fidelity. The community of retired archivists/profesionals related with the archives that participated in this cultural volunteering experience declared their intention to participate in new activities that arise from the archives from now on.

On the other hand, several volunteers had chosen to pull out of the activity as they explained that it would have been more interesting for them if the choice of material was different; transcription of the same kind of entries felt too repetitive and boring. Furthermore, some volunteers commented that they found the software buggy and to be difficult to use. For some of them, it was not easy to understand how to operate the system while others noted various technical difficulties. The issue of a short time frame was in fact mentioned by several volunteers. Other factors contributed to this: other personal commitments (as

shown by both questionnaire and other volunteer additional comments) and varying levels of computer literacy.

In Hungary the final questionnaire was completed by 50 people who rated the programme on the overall programme scoring 4.82/5.00 points and the activity within the Facebook group scoring 4.47/5.00 points.

Archives' perspective

All archives reported that the project exceeded their expectations.

The direct objective of the activity was to involve the community of retired amateurs and professionals in the review and correction of the transcripts used in the automatic indexing and semantic tagging of the texts related to the selected records.

Regarding this **Malta, Hungary and Spain** fulfilled or exceeded the objectives. In Malta 545 pages (34 %) were transcribed out of a total of 1,458. In this regard, the collection of indexed images was quite large when taking into account the total number of participants and the specified time for activity completion. In Hungary the 70 active volunteers processed a total of 6,787 pages and 13,5740 names in the six weeks available. The oldest contributor was a 78-year-old lady who alone checked 309 pages. The Spanish collection consisted of 815 images, and although the number of volunteers was less than expected, the 815 images were reviewed with a double validation.

Regarding the number of pages the result fell short of expectations in **Portugal** because the General Registry of Mercy (1639-1949) was difficult to complete partly due to its size, partly due to the irregularity of the graphics and of the text stain itself. The time needed for the transcription of one book proved to be more than originally calculated. In contrast, in the case of **Norway** the handwriting was quite neat, and the information well organized on the cards, so the reason for not being able to fulfill the quantities was the the lack of volunteers. But the quality of the work they did was very good.

Thanks to the work of these volunteers, Spain develops a search engine embedded in the Iberoamerican Migratory Movements portal, Hungary publishes the collection in their AdatbazisokOnline service.

IMPACT

The benefits beyond the direct objectives were evident for all the participating archives. The main mission of the national archives is to provide accessibility to their country's documentary heritage. Therefore, it is important to analyse what worked and what didn't work in the project, and what impact the almost one year of activity had on these activities.

The method of combining HTR and crowdsourcing proved to be successful. It´s mainly lack of lower prices on the tools from the vendors that prevent us from upscaling the activity. In addition, some shortage of internal personnel resources to administer and follow it up.

Improvement of archival services

The pilot has demonstrated that handwritten text from digitised images could be made searchable. Instant retrieval of such information within digitised collections would be beneficial as it would result in significantly improved access to the archives holdings. In this way, users would save time as they would not have to manually go through the index to find the information needed.

It´s easy to see that this way of combining HTR and crowdsourcing has great potential to get more searchable data. We expect that the national archives will continue to experiment with HTR and also in combination with help from volunteers. This way of working is still in an early phase, we feel confident that we can expect great development in this field.

Changes in organizational activities

The continuation of this fruitful collaboration with volunteers needs reorganizing the insitutions' activities. Combination of HTR and crowdsourcing would require long-term planning in relation to identifying those fonds which could likely benefit from this technology.

Archives should also focus on the training aspect. Staff involved in the co-ordination of such projects would need to be adequately knowledgeable on the technology and the HTR system used; therefore, training would be necessary. In addition, the project has shown that there were varying levels of computer literacy among the volunteer group and this could have likely contributed to them having difficulties with using the HTR system. Therefore, while it is important for volunteers to be trained, archives should also ensure that they have a good understanding of IT. In this regard, expanding the volunteers' group to more diverse ages would be advantageous as younger persons are more likely to be acquainted with the technological world.

Public awareness

Such activities would help with increasing public awareness on the importance of archives. In the short term, volunteer feedback has positively shown that their involvement generated interest about archives. In the long term, if more similar projects are implemented in the future, this could likely help with contributing to

the archives visibility and better awareness on the role and value of a national archives within their respective societies and on the European level.

To maximize the impact of this activity, results will be reported through conferences, website news, podcasts and publications.

Impact on the society

This activity provides a **useful pastime for volunteers**, who are typically recently retired. The project provided equal opportunities for people living in smaller villages and larger towns, and even allowed volunteers from across the whole world to get involved.

The project showed that through these activities the silver generation is becoming more IT-competent, making it easier for them to use other general, online services.

Several contributors even expressed their appreciation towards the organisation at the closing event as it made them feel part of a team. In general, this activity made them feel part of a community, having directly contributed in its mission to provide access to those records which constitute the country's history. Such interest is reflected in the way that several of these volunteers want to further participate in other projects, joined tours of the archives and were also encouraged to conduct research at the archives.

Technological takeaways

The participating archives have gained a broad range of experience on what factors play an important role in making the combination of technology and volunteering successful.

Careful selection is crucial. In the case of the National Archives of Malta, the collection chosen to be reviewed was quite large and volunteers only managed to transcribe a small percentage of it.

It is important to focus on **better balancing the GT** to representing different hands.

Regarding the transcription, it would be essential to devise a system whereby the participants' **work would be cross-checked**. In Malta, Portugal and Hungary several volunteers had problems recognising the varying handwriting styles and abbreviated names. From time to time, volunteers asked for some of their work to be reviewed. Thus, several mistakes with the transcription were frequently identified. It is better to produce lesser amounts of highly accurate GT data than larger amounts of data of inferior quality.

It is also important to emphasize that the **archival work is complex**, requires more than palaeographic recognition and that, consequently, these projects are inseparable from a large investment in training, quality control, evaluation and monitoring. However, they could have a major impact on the social recognition, democratization, and visibility of archives in society.

All partners agreed and highlighted that **more collaboration** between the national archives is essential as it would further contribute to its progress and render it an effective tool which archives could turn to in order to enhance accessibility and visibility to their holdings.

Partners feel confident that these tools will be further improved in coming years and become essential in the future for the retrieval of archival information. The project highlighted some of the weaknesses of the technological solution. As there are several initiatives in the field of HTR, it remains to be seen which of them proves to be most successful.

New projects

In Portugal, the training and sharing of information and experiences, obtained during the Activity EDT21 Crowdsourcing, has already resulted in the release of a new project in Arquivo Distrital do Porto (Provas de Vidas/Proofs of Lives), so we believe that, in the short term, we already have an impact as far as service delivery and organizational change are concerned. Thanks to Crowdsourcing and Online Volunteering, ADP has got new tools to promote interaction and online cooperation. They are now working with several volunteers, including one local association, 5 Portuguese and 2 Brazilian individuals, and 4 face-to-face volunteers.

In Hungary, a new project is starting on 26th September, involving the indexing of the State Security files of the Ministry of the Interior seized by the Soviet army during the Second World War. For this project sixty volunteers returned from the EDT project.