READ

Recognition and Enrichment of Archival Documents

D8.1. Open Innovation Forum

Report for Period 1

Günter Mühlberger (UIBK),
Markus Diem (CVL), Fabian Hollaus (CVL),
Stefan Fiel (CVL)Florian Kleber (CVL), ,

Distribution: Public

http://read.transkribus.eu/

READ
H2020 Project 674943

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 674943

Project ref no.	H2020 674943
Project acronym	**READ**
Project full title	**Recognition and Enrichment of Archival Documents**
Instrument	H2020-EINFRA-2015-1
Thematic Priority	EINFRA-9-2015 - e-Infrastructures for virtual research environments (VRE)
Start date / duration	01 January 2016 / 42 Months

Distribution	Confidential
Contractual date of delivery	31.12.2016
Actual date of delivery	29.12.2016
Date of last update	23.12.2016
Deliverable number	D8.1.
Deliverable title	Open Innovation Forum P1
Type	Report
Status & version	Final
Contributing WP(s)	All WPs
Responsible beneficiary	UIBK
Other contributors	All beneficiaries
Internal reviewers	
Author(s)	Günter Mühlberger
EC project officer	Martin Majek
Keywords	Innovation

Table of Contents

Executive Summary

This paper describes results achieved in *Task 8.1 Open Innovation Forum*. Apart from raising awareness for innovation in general, we are able to report about an innovative idea for a "ScanTent" which was indeed not foreseen in the work plan of the project but which may have significant impact on some of the working areas of the READ project.

1. Introduction

"Innovation" is one of the five key concepts of the READ project. From the DoW:

> One of the main characteristics of innovation is its disruptive nature. This means that innovation always has an impact on the "original plan", that it "knocks out" well established concepts and that it bears a much higher risk, since the greater the innovation is, the less one can estimate if it will be successful at the end. Innovation therefore contradicts by its nature to a strict project plan, since it breaks up conventional patterns and already taken decisions. (DoW p152)

Work package 4 Service Innovation and *Task 8.1. Open Innovation Forum* are the two fields where innovation "resides" in the project.

2. Open Innovation Forum

Innovation was a dedicated subject of sessions at the Kick-off meeting in Marburg (01/2016), as well as the All Staff Meeting in Passau (09/2016). Several ideas where discussed, we mention here two:

- Monolingual translation
  - As a matter of fact, many historical texts from early modern periods or from the middle-ages are not well understood by a majority of readers.
  - The idea would be to use machine translation to increase the understanding of historical documents for a broader audience. Historical texts would be translated into modern texts in an automated way by using the same machine translation technology as it is applied for inter-language translation. Several partners have experience in language technologies, namely machine translation. These resources could be used.

- Crowd-funding for automatic transcription
  - UIBK experienced a strong interest from family historians towards HTR. The expectation from this user group is that any kind of script can be automatically transcribed with a satisfying result.
  - The idea would be to start a crowd-funding campaign where money is collected to produce more training data for the HTR engines in order to fulfil the expectation from above.

Both ideas are worth to be taken into account in more detail, but both require a significant amount of extra resources and were therefore not pursued. In contrast the innovative idea for a ScanTent is directly along the work plan of READ and therefore much easier to integrate.

3. Innovation: ScanTent

3.1. Transkribus ScanREAD (DocScan)

In *Task 5.6. Crowd-Scanning: ScanREAD* an application for smart phones is developed with which users are enabled to use their mobile phone as document scanner and to upload and process the images directly in the READ/Transkribus platform.

The main motivation for this app was the observation that humanities scholars and family historians are often taking pictures in archives due to the high costs connected with on-demand scanning services, e.g. several EUR per copy (1) are sometimes charged by archives. The Transkribus platform could take benefit from this and provide a central repository for images taken by such users. This would strengthen the position of Transkribus and increase the benefit of the platform.

The original plan was to build this app on top of a commercial product, namely the FineScanner App from ABBYY. Due to the fact that the Computer Vision Lab from Vienna was working on very similar technology (camera imaging, sharpness calculation, edge detection, etc.) it was more suitable to use the CVL technology instead of an external provider. This turns out now to be a great advantage since the app can be configured in more detail to the requirements coming from the ScanTent.

This decision was taken in early 2016. A first prototype of ScanREAD app with the working title “Transkribus DocScan” has been released in late 2016 (cf. D5.14). Further development will take place in 2017.

3.2. ScanTent

During first tests of the DocScan app it turned out that there are two challenges which need to be tackled when it comes to “scanning” of archival documents with a smart phone:

(1) Documents are often bound and cannot be opened in the same way as a modern book. Taking a picture requires therefore two hands for fixing the document and somehow flattening the pages.

(2) Sufficient light is often not given in a usual reading room. Turning on the flash of the smart phone is no solution since other users are of course disturbed but also the quality of the images is suffering from flash.

Therefore, it was obvious to think about a device where

(1) the smart phone can be fixed so that the user can use both hands for fixing the document and where

(2) an extra source of light will be available.

Though both principles are fulfilled in typical Overhead Scanners the main innovative aspect is to take the concept of a “dome tent” (cf. figure 1) and to model it into a “ScanTent” which is used as a scanning device for smart phones. (Of course the tent is built on a much smaller scale than a usual dome tent.) Figure 1 shows also a sketch of the “ScanTent”.

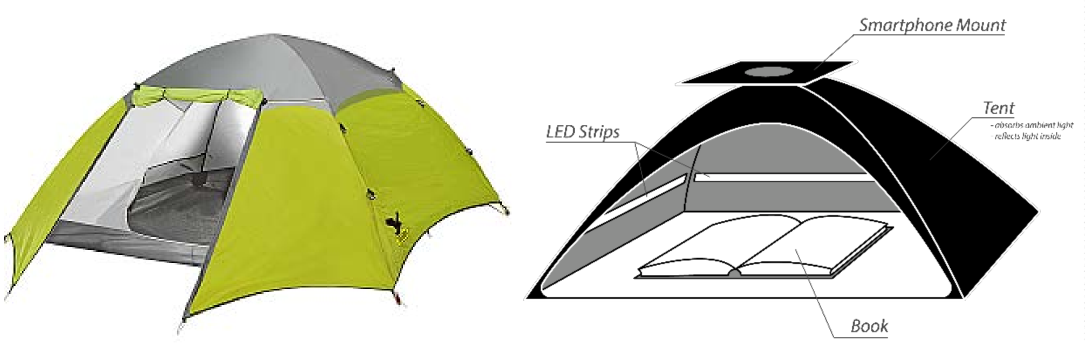


Figure 1 Typical “dome tent” with its characteristic form (here from the Italian company Salewa) and a sketch of the planned ScanTent

The main features of such a “ScanTent” are:

(1) Portability
    a. It is light, consists just of textile and some sticks and is demountable. Users will be able to take it with them to an archive, or to use it at home but to put it away after usage.
(2) Fixed camera
    a. The mobile phone can be placed on top at a platform and therefore the user gets his hands free.
(3) Extra light
    a. An extra light source (LED lights) can be amounted at the side panels. A textile is chosen which reflects light but does not allow light to pass through so that other users in the archive are not disturbed.
(4) Adapted software
    a. The DocScan app is directly accommodated to the ScanTent setting. This means that pictures can be taken in an automated mode tailor made for scanning bound documents.

During 12/2016 CVL and UIBK developed a first prototype where we used several products available on the market.

- A “softbox” from Neewer which is built very similar to a “dome tent” and comes already with the suitable textile panels
- An LED light source which can be glued to the side panels
- A simple extension mechanism based on the application of some extra stripes
- A cardboard for fixing the smart phone on top of the ScanTent

Based on these devices the costs for assembling a ScanTent are currently significantly below 100 EUR.

First tests concerning the quality of the images showed excellent results:

- The resolution of a typical smart phone (e.g. Samsung Galaxy S5) is sufficient for documents even as large as A3 and excellent for documents with a size of A4 (e.g. a typical book)
- The user is able to sit at a table and to place the document carefully
- Bound documents can be fixed with both hands, pages can be turned quickly
- The image quality benefits strongly from the extra light source and makes the user completely independent from any lights available

- Even glossy paper and especially photos can be taken without or a minimum of reflections
- Specific software features of the DocScan app support quick production
- A simple mechanism allows to adjust the distance of the platform with the smartphone so that small and large documents can be scanned with the same device

The following pictures provide a first impression of this prototype.

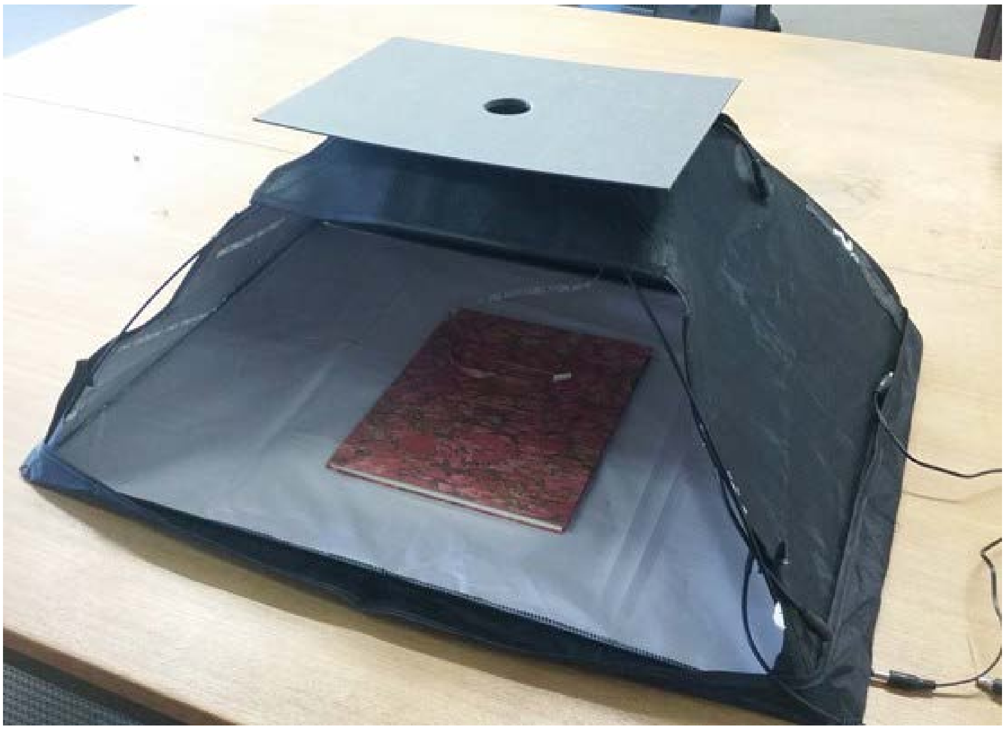
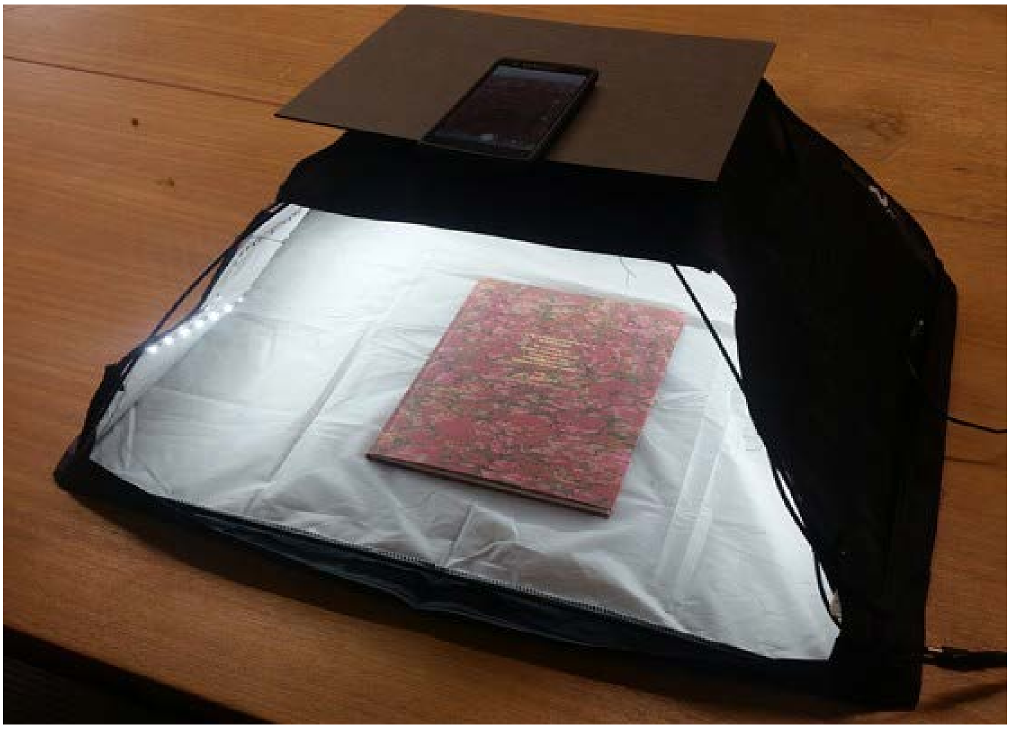


Figure 2 ScanTent - prototype



- Figure 3 ScanTent - prototype – lights turned on

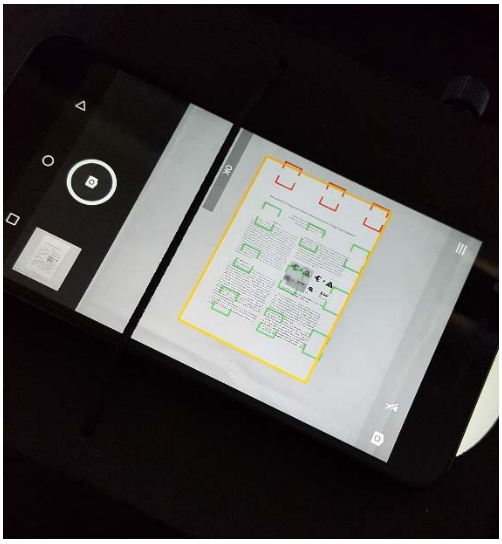


Figure 4: User Interface of the DocScan App with the detected page and focus values

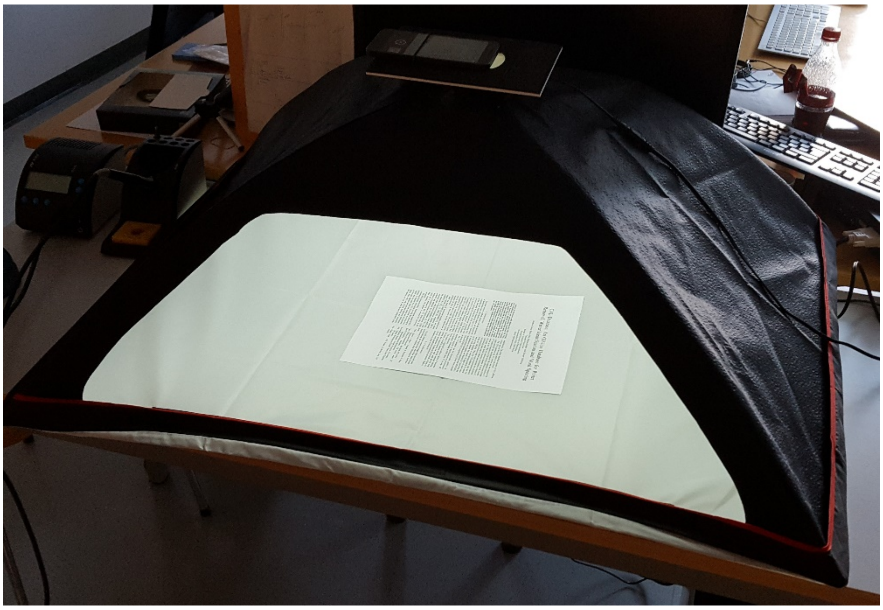
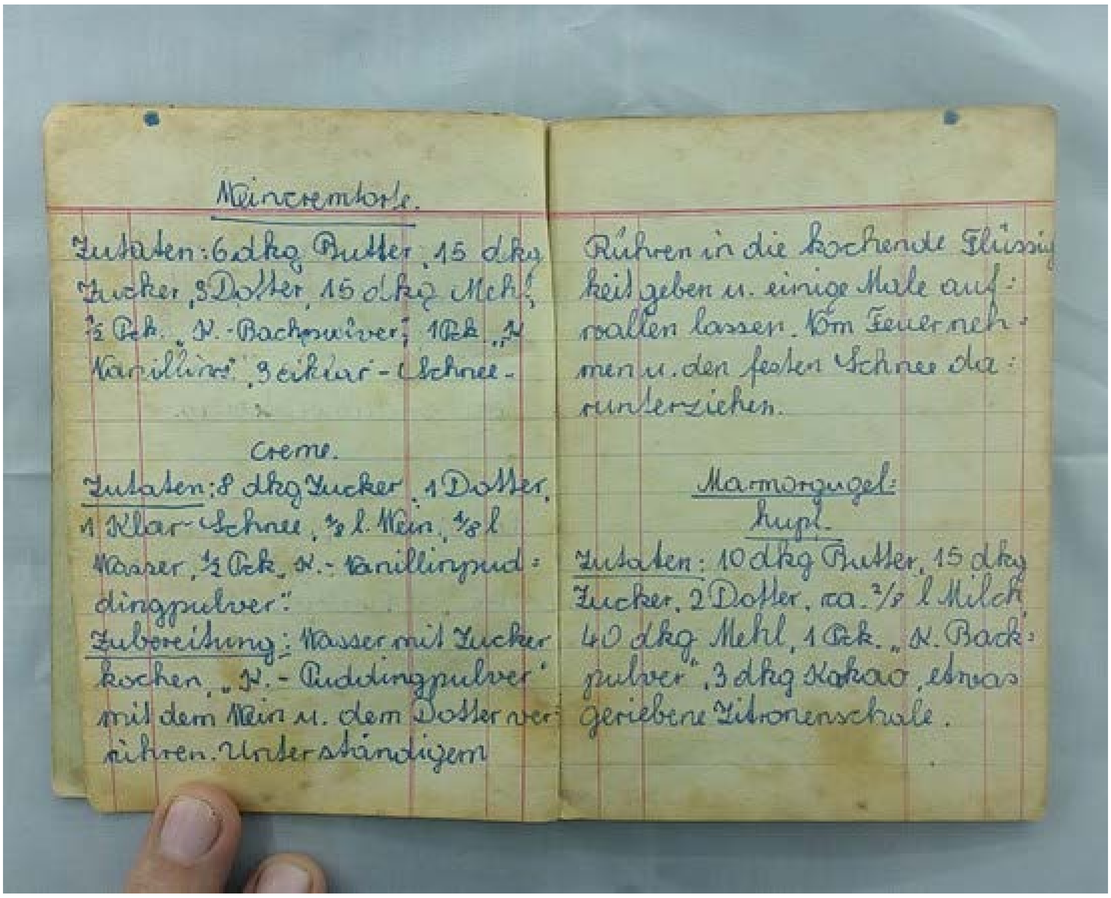


Figure 5: ScanTent – prototype with 3D plotted mount for the smartphone

The following images where taken with a Samsung Galaxy S5. Resolution is about 300 ppi and fully sufficient for any further processing, such as Text Recognition, Writer Identification or similar processes such as face and object recognition.



Weincremtorte.

Zutaten: 6 dkg Butter, 15 dkg
Zucker, 3 Dotter, 15 dkg Mehl,
½ Pck. „Dr.-Backpulver", 1 Pck. „Dr.
Vanillinz.", 3 Eiklar-Schnee.

Creme.

Zutaten: 8 dkg Zucker, 1 Dotter,
1 Klar-Schnee, ⅛ l. Wein, ⅛ l
Wasser, ½ Pck. „Dr.-Vanillinpud:
dingpulver".
Zubereitung: Wasser mit Zucker
kochen, „Dr.-Puddingpulver"
mit dem Wein u. dem Dotter ver:
rühren. Unter ständigem

Rühren in die kochende Flüssig:
keit geben u. einige Male auf:
wallen lassen. Vom Feuer neh:
men u. den festen Schnee da:
runterziehen.

Marmorgugel:
hupf.

Zutaten: 10 dkg Butter, 15 dkg
Zucker, 2 Dotter, ca. ⅜ l Milch
40 dkg Mehl, 1 Pck. „Dr. Back:
pulver", 3 dkg Kakao, etwas
geriebene Zitronenschale.

Figure 6 Cooking book from the 1950ies

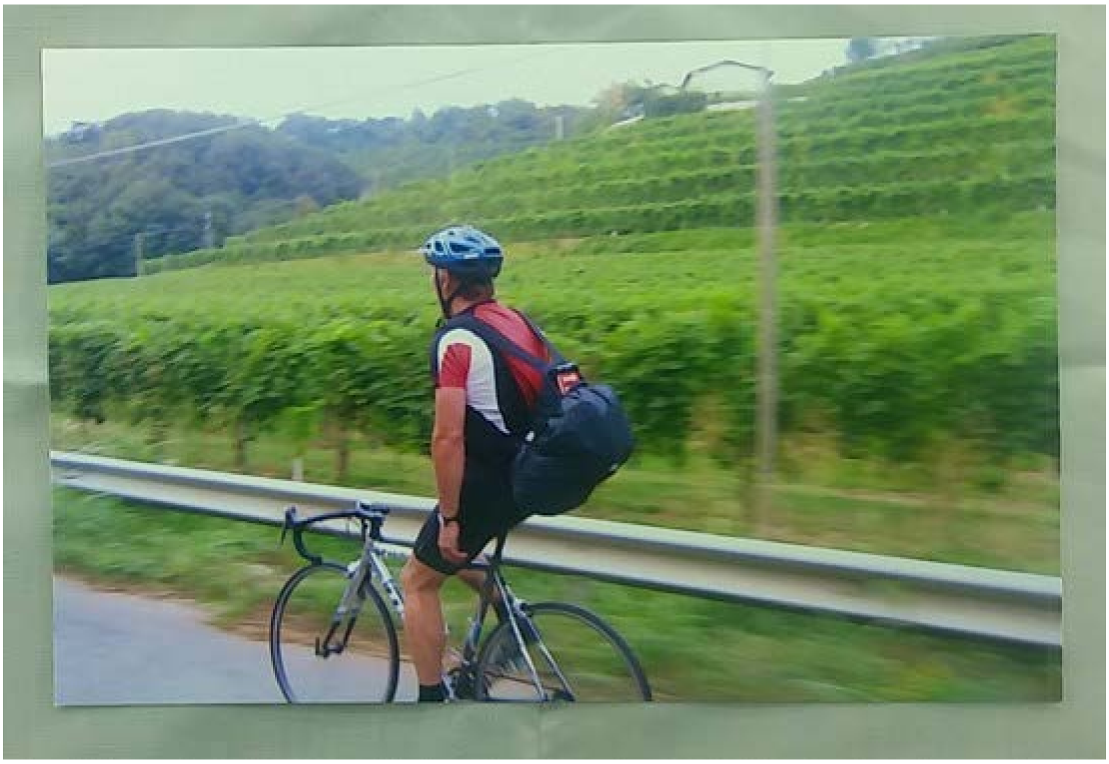


Figure 7 Picture taken from a photo

Based on these encouraging results we will continue our tests in 2017 and come up with a reproducible prototype for further testing in real world situations.

3.3. Market considerations

Given that the production of the ScanTent can be done for a reasonable price we see a considerable market size for this device. Three user groups are from our point of view potential customers.

3.3.1. Humanities scholars and family historians

This group is actually working with historical documents. They are already using their mobile phone in an archive but would like to do this in a more systematic way in order to scan large amounts of documents and to work with them at home and/or in the Transkribus platform. More or less every user who regularly visits archives is a potential customer of the ScanTent. We estimate that these are tens-of-thousands of users in Europe.

This observation is emphasized by the fact that Google announced just in these days the release of its PhotoScan app which is tailored for exactly the setting described above: that people are using their smart phone to digitize their personal photo collection. The ScanTent will enable them to do it in a much better quality.

Figure 8 Google PhotoScan marketing site

3.3.2. Archives and memory organisations

Nevertheless, many archives in Europe are not in favour of users who take pictures with their smart phone and therefore simply restrict or forbid the usage of smart phones in their facilities.

From our point of view this constraint may even increase the business opportunities for the ScanTent. The READ/Transkribus platform will be able to offer two interesting extensions:

Firstly, the ScanTent comes with DocScan and therefore all images are uploaded to the Transkribus platform.

Of course this can be done in a way that also the archive gets a copy of the images taken by the user. In this way the digital collection grows according to the real needs of users – without any further involvement of the archive.

As a special feature which we plan to include is a QR code recognition in DocScan so that the images can be directly sent to the right directory within the archive. When handing over a bulk of documents the archives just would need to print out the QR code with the identification number of the box and the user would need to take a picture of such a QR code when starting his scanning batch in order to directly place the images in the right order of the archive. Other users will of course benefit from the already taken images and be able to add images or to resort them.

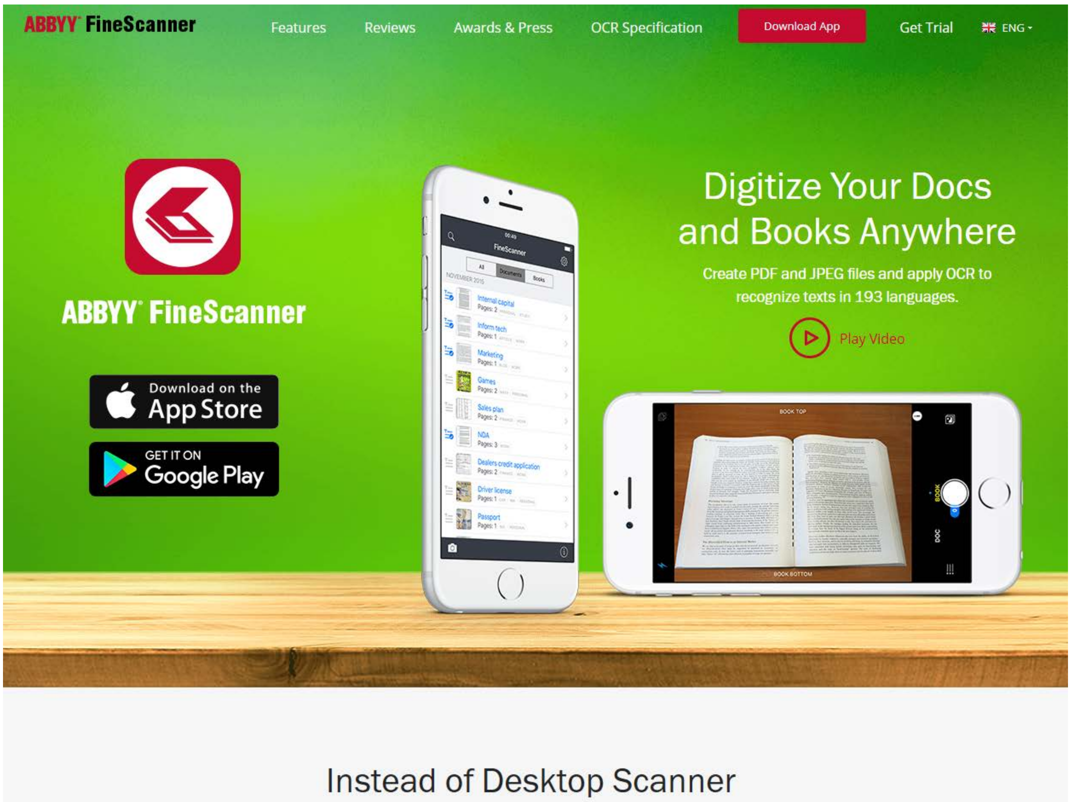
Secondly DocScan can also be used as an accounting device. The number of images taken by the user can be counted and an automated billing system can be implemented. Again the archive would not need to invest anything but receive some income from the users. Actually this is for some archives one of the strongest arguments that they get some income from digitisation-on-demand orders, money which is not bound to yearly budgets and therefore of higher value.

We believe that the two arguments together – reuse of images taken by the user and a small copy fee – will convince many archives which are nowadays sceptical towards the use of smartphones in their premises to conclude service contracts with the Transkribus platform providers.

3.3.3. Library users and students

The third group who may be interested in the ScanTent are library users and especially researchers and students from the humanities. Whereas in technical and life sciences nearly all papers are already available online in the humanities and social sciences still a majority of scientific papers are printed as books. Copying of books is therefore still a daily task for researchers and students. To use overhead scanners from libraries is one solution but often connected with costs. Putting a bound book on an office or flatbed scanner is a time consuming effort.

Again the smart phone comes into the game. The Finescanner app from ABBYY which is the world’s largest company for Optical Character Recognition reflects this situation as it can be seen from the following screenshot:



Digitize Your Docs and Books Anywhere

Create PDF and JPEG files and apply OCR to recognize texts in 193 languages.

Instead of Desktop Scanner

Figure 9 FineScanner from ABBYY

The FineScanner app directly addresses people who are scanning books. Given that by using the ScanTent the speed is much higher and the quality better especially for glossy paper and narrow bound books it will be attractive to this kind of users as well. It also has to be mentioned that – as long as not a complete book is scanned – the scanning of books is not restricted by copyright – as long as users are doing this on their own and are not distributing the scans.

3.3.4. IPR and existing patents

A similar idea such as the ScanTent is subject of a US patent announced in late 2014. The “Imaging Apparatus” is described in the following way:

> The present invention provides a system to facilitate the imaging of objects using a portable computing device equipped with an imaging system invention providing a means to stably position a portable computing device in an elevated position above a surface to enable the rear facing camera of said device to image objects positioned thereunder. The invention further provides illumination systems and lenses to facilitate imaging.

Further investigations will be made to check the details of this patent application but significant differences are obvious so that a conflict is from our point of view not given. Nevertheless, this patent again underlines that the idea of a scanning device for smart phones can be seen as an interesting field of innovation.

3.4. Outline of a business plan for the Transkribus ScanTent - DocScan

The corner stones of a business plan specifically for the ScanTent / DocScan application are:

- Outsource mass production, ordering and delivery of the ScanTent as far as possible

- Investigate options for a crowd-sourcing campaign (Kick-starter, Indigogo) to start mass production
- Keep mass production costs low, e.g. below 50 EUR, so that the market price is not above 100 EUR. Receive a revenue for every tent which is sold.
- Offer contracts to archives regulating the usage of the images in their own repositories as well as offering accounting/billing service. Receive service fees partly per installation, partly based on turnover.
- Use the READ project and Transkribus platform to market the ScanTent.

Further actions towards the ScanTent and the business plan shall be taken in *Task 3.2. Business plan implementation* as well as in Task *8.1. Open Innovation Forum*.

7. References

1. Abbyy FineScanner

    http://www.finescanner.com/

2. Google PhotoScan

    https://www.google.com/photos/scan/

3. Imaging Apparatus by Richard B. Murphy

    US 20150332129 A1

    http://images1.freshpatents.com/imageviewer/20150332129-p20150332129

    https://www.google.com/patents/US20150332129