

tranScriptorium

D7.3: Exploitation Plans (M36)

Günter Mühlberger, UIBK

Distribution: Public

tranScriptorium

ICT Project 600707

Deliverable 7.3 (January, 2016)



Project funded by the European Community
under the Seventh Framework Programme for
Research and Technological Development



Project ref no.	ICT-600707
Project acronym	tranScriptorium
Project full title	tranScriptorium
Instrument	STREP
Thematic Priority	ICT-2011.8.2 ICT for access to cultural resources
Start date / duration	01 January 2013 / 36 Months
Distribution	Public
Contractual date of delivery	D7.3: December, 31, 2015
Actual date of delivery	D7.3: 12. January, 2016
Date of last update	
Deliverable number	D7.3
Deliverable title	D7.3: Exploitation Plans
Type	Report
Status & version	D7.3: Final
Number of pages	22
Contributing WP(s)	All
WP / Task responsible	Günter Mühlberger
Other contributors	All
Internal reviewer	Joan Andreu Sánchez
Author(s)	Günter Mühlberger (editor),
EC project officer	José María del Águila Gómez
Keywords	Exploitation Plans

The partners in **tranScriptorium** are:

Universitat Politècnica de València - UPVLC (Spain)
 University of Innsbruck - UIBK (Austria)
 National Center for Scientific Research "Demokritos" - NCSR (Greece)
 University College London - UCL (UK)
 Institute for Dutch Lexicology - INL (Netherlands)
 University London Computer Centre - ULCC (UK)

For copies of reports, updates on project activities and other **tranScriptorium** related information, contact:

The **tranScriptorium** Project Co-ordinator
 Joan Andreu Sánchez,
 Universitat Politècnica de València
 Camí de Vera s/n. 46022 València, Spain
jandreu@dsic.upv.es
 Phone (34) 96 387 7358 - (34) 699 348 523

Copies of reports and other material can also be accessed via the project's homepage: <http://www.transcriptorium.eu/>

Table of Contents

- Executive Summary 4
- 1. Products..... 4
 - 1.1. Data Sets..... 4
 - 1.2. HTR and Word Spotting Tools 5
 - 1.3. Document Image Analysis and Language Modelling Tools 6
 - 1.4. Interfaces and Platforms 8
 - 1.5. Software as a Service..... 9
- 2. Service platform..... 10
 - 2.1. User driven approach 10
 - 2.2. Service catalogue..... 11
 - 2.3. Service extensions in 2016 14
- 3. Business model 15
 - 3.1. Background and general considerations 15
 - 3.2. Sources of income 15
- 4. Market size 18
 - 4.1. Overview..... 18
 - 4.2. Competitors..... 19
- 5. Governance model 20
 - 5.1. General considerations..... 20
 - 5.2. Association or foundation 20
 - 5.3. Companies (limited by shares) 21
 - 5.4. Cooperatives..... 21
- 6. Appendix: Technology Readiness Levels of the EU Commission..... 22

Executive Summary

The report shows that there are two main approaches to exploit the results of the tranScriptorium project: Firstly a product based approach and secondly a service based approach. The main advantage of the Software as a Service (SAAS) approach is that it fits much better to the general environment of the project partners and the target groups. Especially the benefits for the target groups as well as the synergies which can be gained by a service platform are discussed in more detail.

Due to the fact that the technology of the project will form the basis of a Virtual Research Environment set up as part of the H2020 Project READ (Recognition and Enrichment of Archival Documents) this exploitation plan will be realized from 2016 onwards as a research infrastructure for archives, libraries, humanities scholars, computer scientists, volunteers and the public in general.

A first outline of this exploitation plan has been presented to the public as working paper in December 2014 at the HistoInformatics Workshop in Barcelona.

1. Products

In the following we briefly describe the products from the tranScriptorium project, their general purpose and their potential user groups. We will use the Technology Readiness Levels of the European Commission to indicate the technological maturity of the tools and systems.

Based on this overview we will then go into more detail concerning the potential of those products which are from our point of view especially interesting for exploitation.

1.1. Data Sets

Name	tS Data Sets (Ground Truth) – WP2
Provider/Owner	Content providers (outside of the project) and all partners
Description	tS produced data sets with altogether nearly 2000 pages in Dutch, English, German and Spanish. If we assume that the commercial production of one page will cost about 20-50 EUR per page the monetary value of this dataset is about 60.000 – 80.0000 EUR.
Technological Readiness Level	TRL8. System complete and qualified Data sets are highly standardized (PAGE format, metadata about the alignment and transcription process), tested and used in real environments (research competitions, HTR training).
Potential usage scenarios	Research competitions, training of HTR engines.
Constraints	Some content providers have not accepted the Open Data concept, therefore making available some of the data sets may require registration/authorization. The data sets are not fully homogenous since different annotation and transcription rules where applied.
Target groups	Research groups, companies.
Potential for exploitation	High potential mainly for academic purpose, since Ground Truth data are very valuable for research and development. Once these data sets are free for download it is to expect that various research

	groups will use them.
Monetarisat	Very low, since by definition these sets will be Open Data and are created to foster research in HTR.

1.2. HTR and Word Spotting Tools

Name	HTR Engine
Description	The HTR technology consists of several modules and is freely available as Open Source. It can be trained and used to recognize (“transcribe”) handwritten documents.
Provider/Owner	UPVLC
Technological Readiness Level	TRL6. Technology demonstrated in relevant environment The system is stable and has been used for training and decoding (recognition) of ten-thousands of pages.
Potential usage scenarios	Training of HTR models, recognition of handwritten documents with the support of an internal language model
Constraints	The HTR engine is a set of scripts and libraries, not a Software Development Kit (SDK) with a full featured Application Programming Interface (API). Documentation is available but on a very basic level. The engine is only partially optimized for performance. Updates may require higher effort due to internal dependencies. The HTR engine is directly depended on the quality of baselines which are generated by the tool chain (image pre-processing, layout analysis).
Target groups	Computer scientists, software integrators such as computing centres or companies.
Potential for exploitation	Low to medium. The effort to set up the complete package must not be underestimated. A complete system for document management, image processing, etc. is needed.
Monetarisat	Low. The package is available as Open Source and it is part of Transkribus where the integrated version of the HTR engine can be used for free.

Name	Keyword Spotting Application I (Query by String)
Description	The KWS QbS application is an innovative and user-friendly way to search collections of HTR recognized documents. Details are described in the respective deliverables from WP3.
Provider/Owner	UPVLC
Technological Readiness Level	TRL6. Technology demonstrated in relevant environment
Potential usage scenarios	This application is especially interesting to archives which have digitized large collections of handwritten documents and want to

	provide their users with an effective way to search them.
Constraints	The main constraint is that KWS is directly dependent on the quality of the HTR process and therefore also on all related constraints (see above).
Target groups	Archives, libraries, editors of digital editions
Potential for exploitation	Low to medium. Though the application itself is of course interesting to archives and libraries there are many dependencies which make it very hard for a potential user to exploit exactly this product.
Monetarisaton	Low. Archives are still very suspicious on the benefits of full-text search. Expectations towards the accurateness are often too high with respect to the State-of-the-Art.

Name	Keyword Spotting Application II (Query by Example)
Description	The KWS QbE module enables users to search within handwritten collections without any need to train a model or to carry out any layout analysis.
Provider/Owner	NCSR
Technological Readiness Level	TRL 5. Technology validated in relevant environment
Potential usage scenarios	Use case 1: To search document collections for “similar” words, respectively “images of words”. Use case 2: To support users in transcribing or deciphering hard-to-read documents by providing suggestions and similar word images.
Constraints	The technology works well for very small datasets, but requires high computing resources for real-world collections consisting of thousands of pages.
Target groups	Archives and humanities scholars.
Potential for exploitation	Low respectively it is still unclear if the technology will proof its usability in an operational environment.
Monetarisaton	Low, unclear.

1.3. Document Image Analysis and Language Modelling Tools

Name	Platform for text block, line and word segmentation as well as image enhancement tools
Description	A platform which combines several tools which analyse a page image and provide the coordinates (zones, regions) of text blocks, lines, baselines and words.
Provider/Owner	NCSR, UPVLC
Technological Readiness Level	TRL 7. System prototype demonstration in operational environment

Potential usage scenarios	<p>Use case 1: These tools are needed as pre-processing tools for the HTR engine and play therefore a decisive role for the overall quality of the recognition and indexing process. They are tools in an HTR processing chain, respectively can be used by expert users to find out the optimal parameters for processing documents.</p> <p>Use case 2: Humanities scholars who want to create a digital edition which is based on a strong linking of text and image are also highly dependent on such tools (otherwise the complete segmentation is done manually).</p>
Constraints	<p>Block and line segmentation are currently split into two modules with no interaction. Since the block analysis is tuned to rather simple one-column documents a large number of archival documents cannot be processed with the tools or needs manual input. Moreover vast amounts of archival documents come as complex tables which cannot be analysed with the current DIA tools. This is currently one of the strongest constraints for processing large amounts of documents with HTR or KWS QbS.</p>
Target groups	Technology providers, research groups.
Potential for exploitation	Low. Similar tools – though in many cases of lower quality – are available by many other research groups. Many technology providers have their own “home-made” tool set at hand.
Monetarisaton	Low

Name	Language Server (LS)
Description	The Language Server is a module which has been built to increase the accurateness of the HTR engine by utilizing external language resources.
Provider/Owner	INL
Technological Readiness Level	TRL 4. Technology validated in lab
Potential usage scenarios	The main use case is to support users who are actually working with the HTR engine as described above.
Constraints	This product is designed as an internal “support tool” and offers therefore very specific services.
Target groups	Technology integrators who want to set up their own HTR processing system based on the modules provided by the tranScriptorium project.
Potential for exploitation	Low
Monetarisaton	Low

1.4. Interfaces and Platforms

Name	TSX Crowd-sourcing Interface
Description	A web-interface which provides a simplified interface for transcribing handwritten documents with or without the support of an HTR engine.
Provider/Owner	ULCC/UCL
Technological Readiness Level	TRL 6. Technology demonstrated in relevant environment
Potential usage scenarios	TSX can be used to set up a crowd-sourcing interface with a minimum of effort.
Constraints	TSX is an interface to the Transkribus platform and therefore dependent on the availability of the Transkribus services. A publishing component is missing so that users are able to display their digital editions to a public audience.
Target groups	Archives, humanities scholars.
Potential for exploitation	Medium to high. Many archives and libraries are interested to involve users in a dedicated digital environment which deals with text correction or transcription. Also humanities scholars may involve students or volunteers with this interface.
Monetarisaton	Low. The software is Open Source and dependent on the Transkribus platform.

Name	Transkribus (Transcription and Recognition Platform)
Description	A comprehensive platform which allows to manage the whole transcription process and which integrates most of the tools developed in tranScriptorium. Features include document management (upload, processing, download, export), user management (registration, roles, groups), document and image editing (segmentation, transcription, annotation), document and image processing (block, line segmentation, applying HTR models), HTR processing, CATTI based correction, a wide variety of export formats (METS, PAGE, ALTO, TEI, PDF, Excel, RTF) provision of web-services and serving the TSX Crowd-sourcing as well as the Transkribus expert interface.
Provider/Owner	UIBK, all partners.
Technological Readiness Level	TRL 7. System prototype demonstration in operational environment
Potential usage scenarios	(1) Production of Ground Truth for research purposes. (2) Manual transcription of handwritten documents with semi-automated support for block- and line segmentation (=basis for digital editions of documents) (3) Automation assisted transcription of handwritten documents (if HTR model is available).

Constraints	<p>Training (learning) process for the HTR engine needs to be organised off-line and requires manual input by UIBK</p> <p>All constraints listed for the DIA tools, the HTR engine, etc. are also constraints for using Transkribus in a fully productive environment.</p>
Target groups	<p>(1) Research groups and companies who are interested to produce Ground Truth in a standardized way.</p> <p>(2) Humanities scholars who are interested in transcribing handwritten documents with a close alignment of text and image and in a standardized environment. Main benefit is that the output (TEI, PDF) fits to the general research environment.</p> <p>(3) Content holders (archives, libraries,...) who want to provide their users with a simple and highly standardized tool for the transcription of handwritten documents. Main benefit is that the output (METS/ALTO) also fits to their digital library systems.</p>
Potential for exploitation	High potential. TRANSKRIBUS is currently the only comprehensive transcription and recognition platform which offers not only a large number of features but also the integration of DIA and HTR tools in a regular way.
Monetarisaton	Medium to high. Thousands of archives, libraries, humanities scholars and volunteers should have potential interest on the services provided by the platform.

1.5. Software as a Service

The list above shows in an impressive way the tool sets and products created by the tranScriptorium project partners. Nearly all of these products go far beyond what can be expected by basic research and are validated or demonstrated in relevant or operational environments.

With the Transkribus platform an integrated approach has been performed so that users are able to access most of the tools also directly via a Graphical User Interface.

In contrast to the software as a product based approach it became rather clear during the project that a the concept of “Software as a Service” is much more appropriate to our situation.

Main reasons to follow this approach:

- **Heterogeneity of the user groups and services**
HTR and related technologies is not only interesting for archives or technology providers who want to recognise and index large amounts of digitised documents, but also for humanities scholars and the public. Also the portfolio of products and services is heterogenic and makes it difficult to form one product out of it. In contrast a service platform is able to turn this heterogeneity into a benefit and to offer specific interfaces and services to a large variety of user groups.
- **Synergy through centralized data storage**
One of the main reasons to go for a SAAS model was the fact that HTR and related modules need to be trained before they can be applied to a given document set. If such training would take place locally by customers using the software as a product, these training data would get “lost”, i.e. would not be reusable for other users. In a centralized platform not only the training data itself but also the user data are available for further improvement of all

software tools (not only HTR, but also DIA or correction interfaces). Though all data in the platform a private it is obvious that machines can process these data without infringing copyright or personal rights.

- **Easier maintenance and dynamic development of the software**

Since there is only one central instance of the several software modules which needs to be maintained the environment can be optimized for each software module. E.g. UIBK is already running Linux and Windows servers with very specific configurations for each of the modules. Such an environment makes it much easier for research groups at universities and research centres to deploy their software and to make it available via the platform. Also updates and service extensions can be much easier organised since the release cycles can be much shorter.

The main decision which was taken in year 2 and 3 of the project is therefore to follow the SAAS approach and to run a service platform after the end of the project.

2. Service platform

2.1. User driven approach

The creation of a SAAS Platform was triggered by the fact that we have to deal with several heterogenic user groups. Actually we have identified mainly four target groups involved in transcribing and recognizing handwritten documents.

In short these four target groups are:

1. Content providers

Archives, libraries, museums and similar memory organisations are hosting thousands of kilometres of handwritten documents. They will need systems with which they can organise the recognition and transcription process of large amounts of documents in an effective and highly standardized way.

2. Humanities scholars

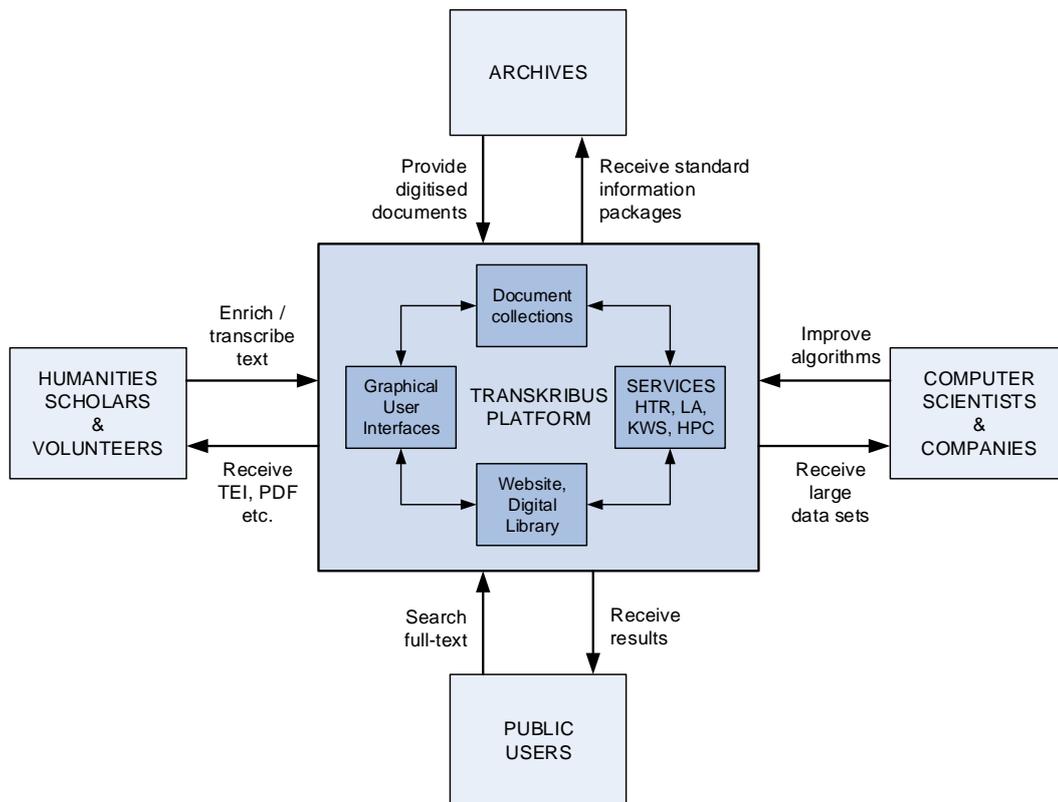
This is the target group which is mainly concerned with the transcription of handwritten documents. Either by acting as editors for scholarly editions, or by transcribing some pieces of documents as part of a larger research effort, this target group is especially interested in the content of the documents and professionally engaged with searching and transcribing these documents.

3. Technology and research providers

This target group, in our case represented mainly by UPVLC, NCSR and INL are highly dependent on the availability of data sets in general and reference data sets specifically. The more data are available the higher is the chance that the state-of-the-art in HTR can be exceeded. As all other groups these data should be as standardized as possible in order to minimize the effort of adapting tools and algorithms to specific document collections. The same is true for commercial technology providers.

4. Volunteers and public users

As Transcribe Bentham has demonstrated thousands of people are willing to contribute to the transcription of handwritten documents. Simple to use interfaces and clear task descriptions are one prerequisite for this target group.



The services of the platform need to be designed in a way that they not only support the needs of these four groups but also promote synergies between these target groups which are otherwise not possible.

2.2. Service catalogue

The Transkribus platform is dedicated to provide a rich service portfolio for the transcription, indexing and publishing of historical documents. The service catalogue of Transkribus consists of the following main modules:

- **Specific modules and services**
All these services are directly connected with the task of processing historical documents with the services developed in the tranScriptorium project. They comprise the services described in section 1, as there are DIA, HTR, CATTI, KWS and the Language Server.
- **Services connected with the Graphical User Interfaces**
In order to make the services available to users several Graphical User Interfaces were developed in the project. First of all the Transkribus expert GUI, the TSX crowd sourcing interface and the web interface as well as a Wiki site. Also a number of demonstrators.
- **Basic platform services**
These services comprise generic modules as they are typical for any kind of platform: document, user and job management tools, export services and web services.

The services are described in detail in the next section.

2.2.1. Specific modules and services

Specific services directly connected with the research work carried out in tranScriptorium are the following:

- **Generate training data for Handwritten Text Recognition**
Users are enabled to create training data with which the HTR engine can be trained. the number of pages is depending on the type of documents, scripts, language, alphabet, but also the layout and the image quality. In some cases only 50 pages may be sufficient to train a good optical model of a script, in other cases some hundreds or even more pages will be needed. To generate such training data is a simple process and is supported by the Document Image Analysis tools and – if e.g. a first model is already available – also by the correction and editing tools (CATTI Service, HTR Suggestions service).
- **Improve an HTR model with additional language data**
Since the HTR engine is strongly dependent on the dictionary the accurateness of the recognition process can be improved dramatically with extra language data. As a rule of thumb every word which is does not appear in the Language Model but appears in the text which is processed causes another error. In other words: If the Out of Vocabulary (OOV) rate is e.g. 20% than WERs go up to 40% even if the Optical Model of the HTR engine is well trained. The Language Server allows therefore users to add more language data to their HTR model.
- **Apply a trained model to a given document**
Once an HTR model is available the user is able to run the recognition for other pages or documents. As it has been shown in the demonstrators the Word Error Rate is between 15% and 40 or 50%. The WER is especially dependent on the availability of a good language model which requires extra language resources (= similar texts of the same period or language). The correction of the automatically recognized text will be in some cases require less user input than the actual typing of the text. Nevertheless in cases where the WER is high, the benefit which can be gained is not that high.
- **Search in pages which were indexed by the HTR engine**
The main benefit of an HTR model is less in speeding up the transcription of a text, but in the possibility to apply it to larger amounts of pages and documents and to use this index for full-text searching. There are mainly two methods for this: Firstly a conventional full-text search performed on the first-best index provided by the HTR engine. The main advantage is this has a very positive effect on the processing time needed for decoding (recognition). Secondly the Keyword Spotting engine (KWS-QbE) enables users to benefit from the internal options considered by the HTR model and is therefore clearly an advantage over the conventional full-text search.
- **Use Keyword Spotting QbE methods to search within documents**
As already described above the KWS QbE allows users to search within documents purely based on the graphical features of a word (or a similar graphical component). It is to expect that this service supports users when transcribing hard-to-read documents since they are enabled to compare similar words in the same document.

2.2.2. Services connected with the Graphical User Interfaces

In order to be able to communicate with the Transkribus platform currently three user interfaces are available. All of them will be improved and new features will be added in 2016.

- **Manage all services with an expert client (Transkribus)**

This expert interface is a platform independent JAVA tool with which users are able to process most of the services offered by the platform. Transkribus can be downloaded from the website, a comprehensive Wiki site is available as user guide. The software is free and will be released as Open Source during 2016. Nevertheless users need to register in the platform for download and for processing documents in the cloud (Transkribus server). Currently more than 2600 users are registered in the platform.
- **Involve users via a Crowd-Sourcing Interface**

All documents which are in the Transkribus cloud can also be accessed via the TSX interface which is a web based interface for transcribing documents. In this way the effort for archives and libraries, but also humanities scholars to set up a crowd-sourcing project on the basis of their own documents is minimized. Volunteers just need to register in the Transkribus platform and are immediately enabled to contribute to the transcription of a document.
- **Publish documents (digital edition)**

The TSX module will be extended in 2016 towards a real publishing interface. This means that owners of a document will be able to “publish” their documents in digital form, i.e. to make the document accessible to the public. Users of published documents need not to register or login but the documents are free for anyone.
- **Transkribus Wiki**

The Transkribus Wiki is used as a central information platform and user guide for the service catalogue and the user interfaces of the platform. It is currently maintained in English (and some pages in German), during 2016 it will be extended to other languages as well.

2.2.3. Basic Platform Services

All services and interfaces described above are specific for the Transkribus platform. But as a platform also services are offered which are common to any platform. From our point of view these basic platform services are highly important and will play a key role in the future.

- **Access, upload and process documents in the cloud**

The Transkribus platform has a document management component with which documents (page images) can be uploaded, processed with the various tools and services and also exported in a wide variety of formats for integration into other software tools. In the future it is planned to connect the platform with repository systems so that archives and libraries can make their documents easily available via the Transkribus platform. This would make it much easier for users to access and collect specific documents in which they are interested in.
- **Manage users and their roles**

Though we are committed to the concept of an open platform (Open Source, Open Content, Open Access) it is important to understand that a platform needs to respect that users have a strong interest in the privacy of their activities. By default all documents in Transkribus are therefore private, only the owner of the document (i.e. the person or institution who uploads or makes a document available in the cloud) has access to a document. The user management component allows a user to add other registered users to his collection.
- **Activate jobs in the platform**

Currently jobs – with the exception of the ABBYY OCR engine which is also included in Transkribus – are limited to single pages, e.g. to run the DIA tools on one page. In the future users will also be able to apply jobs to larger amounts of pages or documents.

- **Export documents in a wide variety of formats**

It is one of the most important objectives of the service platform to support users that they can use and reuse their documents in many ways. Therefore already now a number of formats are provided such as: PDF (for exchange, reviewing, reading), RTF (for simple editing, reviewing, exchange), METS (for professional use in archives and libraries), PAGE (for computer scientists and professional use), ALTO (for use in repository systems), Excel (for reviewing and working with the documents).

- **Access documents and services via web services**

One of the biggest advantages of the platform is the possibility to access documents and services via web services. In other words: Most services which are triggered by human beings via a Graphical User Interface can also be triggered via a machine readable web service interface. Also in this case the user management is applied so that most web services require an authentication at the Transkribus servers.

2.3. Service extensions in 2016

This service catalogue will be extended in the READ Project (see below). The most important services which will come as prototype implementations during 2016 are:

- **Automatic Writer Identification and Retrieval**

Independently of any HTR the script of a specific writer can be trained and retrieved with good accuracy. AWI enables therefore users to search large collections of documents for specific “hands”, e.g. documents of famous persons. AWI can also be used to cluster documents for the HTR training process and to improve therefore the creation of HTR models.

- **Table- and Forms Recognition**

As already indicated above this is one of the most important service extensions since a vast majority of documents in archives is not the “single-column running text” document, but tables including administrative data ordered often by person names, dates, numbers. E.g. all the church registers from the 18th century onwards are typically complex tables and forms. Apart from a generic service which shall be able to process such tables and to find rows, columns and cells, also a template based service will be available where users can define specific tables which are used as model for the following documents.

- **Text to Image Matching tool**

This service allows to match an available transcription with the corresponding image. This will not only ease the manual transcription of text but also enlarge the basis of training data since so many transcriptions are already available in digital form or in printed form, but without any detailed linking between image and text (which is the prerequisite to train the HTR engine).

- **ScanApp and Mobile Correction App**

A mobile app will support users to take pictures of documents within archives and to directly upload them to the Transkribus platform. In this way users become more independent from the digitisation activities in archives and are even able to contribute to the digitisation of documents. With the Mobile Correction App expert users can get involved in reviewing “hard-to-read” words.

- **e-Learning Module**

The e-Learning service is designed to support students and volunteers in learning to decipher historical handwriting and to produce a scientifically correct transcription with Transkribus.

- **Publishing interface**

This is an important component for humanities scholars and archives who wish to publish documents which were transcribed with Transkribus. The service will be part of web interface.

3. Business model

3.1. Background and general considerations

In January 2015 the tranScriptorium partners (except INL) together with a number of other research institutions and archives have applied for an H2020 project in the e-Infrastructure call for Virtual Research Environments with a proposal called READ (Recognition and Enrichment of Archival Documents). The proposal received very favourable reviews and is therefore granted with 8,2 mill. EUR. It will start on 1.1.2016 and run for 3,5 years. Among research in HTR and related technologies the project will also continue to maintain and further develop the service platform Transkribus with the final aim to turn it into a self-sustained service in 2019.

Part of the project is therefore to deliver a business plan and to turn this plan into a working model during the course of the project. Main corner stones of this business plan will be the following:

- **Deliver an Open Platform**

Open Access, Open Source, Open Research Data, Open Content - this list indicates that tranScriptorium and READ are part of the Digital Agenda of the European Commission which is based on the assumption that publicly funded research also should in return be available to the public with no (or a minimum) of restrictions. All the services of the Transkribus platform shall therefore be free to all users.

- **Promote a collaborative platform**

One of the main advantages of the platform is that synergies can be unlocked which are otherwise hard to reach or impossible to gain. To promote the idea of collaboration among the user groups is a key point and in this way the platform shall serve as “mediator” between heterogenic user groups

- **Generate income by services**

In addition to the free services a “freemium” model will be applied. This means that above certain thresholds such as number of documents, number of users or number of jobs and processes service fees will be charged to the users.

- **Involve users wherever possible**

As an Open Service Platform it will be a key success factor that the platform is anchored within its target communities as a “valuable” player. Users need to get the chance to actually get involved in the design and further development of services.

3.2. Sources of income

3.2.1. Public funding

In order to be able to provide the services from the Transkribus service catalogue for free to archives, libraries, humanities scholars, computer scientists and volunteers a basic infrastructure will be

needed. Given that the services are of value to the user groups and that the Transkribus platform operators are able to clearly show the usage it can be expected that public bodies are willing to finance these free services. We see mainly three public bodies which are from our point of view predestined:

- European Commission
- National Ministries of Sciences
- National Funding agencies

For all three bodies the following arguments can be applied:

- **Free, open, community-driven infrastructure services** are an excellent way to promote collaboration among different target groups which are funded anyway by the public (archives, libraries, universities, research institutions) and to benefit from synergies.
- An Open Platform is also an excellent tool to **promote “Openness” and “Sharing” of research data**, in other words for the overall objectives of these public bodies.
- An Open Platform with free services lowers the entrance hurdle to take up these services which – on the longer run – will lead to better interoperability, reusability of data and therefore contribute to **standardization** in general.

Public funding will not only comprise liquid money, but also the availability of other publicly maintained platforms. We mention here:

- **HPC Clusters**
Many Transkribus services require high computing resources. HPC clusters are on the other hand maintained by many universities or research centres in order to foster research.
- **Repository, storage and backup facilities**
In order to run a central platform large amounts of storage as well as repository and backup facilities are needed which are typically provided by academic computing centres financed by ministries and funding agencies.
- **General computing services**
There are a number of basic services (databases, user validation and authentication) which are necessary to run an open platform and which could be utilized.

3.2.2. Community funding

In a very similar way also the community itself, such as archives, libraries but also humanities and computing departments of universities and research centres should have a vital interest that an open platform for the processing historical documents will be sustained on the long term. The advantages for these communities are similar to above:

- **Transparency**
As an open platform users are able to directly review the performance of the tools and services. The entrance hurdle is minimal and requires no extra costs or fees. If an archive is interested to see the results of e.g. HTR processing on its own documents it is enabled to carry out tests, to measure the performance, to get support from the platform operators or other users, to compare its results with similar projects etc. Also the software pieces are available and can be deployed and run in other environments.
- **Efficiency**
Instead of the need to set up own services, to deploy software packages and maintain a complete infrastructure the open platform concept allows archives, libraries and humanities scholars to start immediately with a complete system without any need for local

infrastructures. Uploading or making available document images is the only requirement to get started.

- **Synergy**

The more archives, libraries, humanities scholars or volunteers are using the platform for their own purposes, the more data and models are available, which means that archives which join the platform will have a higher chance that similar documents were already processed and that therefore models are existing which can be applied to their own documents.

- **Standardization**

Also this argument has been mentioned already but indeed it is also of eminent importance for the archives, libraries and humanities sector. The real benefit of digitisation is not enjoyed as long as access is provided on a local basis (archive by archive, digital edition by digital edition) but once users are enabled to access documents and data across archives, libraries and digital editions.

All these arguments should be appropriate to convince representatives of archives, libraries, and humanities departments to support the platform in a regular way:

- **Monetary support**

Here we think mainly on (small) membership fees where hundreds of archives, libraries and universities demonstrate with their membership fee that they are committed to an open platform dedicated to the processing of historical documents.

- **In-kind support**

If archives and libraries are directing their users to the platform and recommend to make use of it than the platform will become more and more popular which again will result in more documents, data and models.

3.2.3. Service fees

The third pillar besides public and community based funding will be service fees. Whereas all services are in principal free to all users, it is obvious e.g. above a certain threshold some services will require some kind of (monetary) remuneration. Service fees will very likely be dependent on the following criteria:

- **Number of documents in the platform**

For several reasons the image files need to reside in the Transkribus platform which requires storage and backup facilities. For smaller archives this service is interesting per se, since they will very likely not be able to run their own systems. Since storage and backup facilities are available at many public universities and research centres and since image files of historical documents are e.g. in contrast to video data, rather “light”, the threshold will be rather high, e.g. some ten-thousands or even hundred thousands of files.

- **Number of users within a collection**

The TSX crowd-sourcing interface is an important service within the platform and is of course available to every document owner. Nevertheless if hundreds or even thousands of users are involved by one document owner it is clear that the infrastructure needs to be prepared for such a project and that therefore service fees will be required.

- **Computing power**
HPC clusters provide their services in “core hours” and indeed also in the Transkribus platform the processing of large amounts of page images will be a crucial component. E.g. processing a page with the current HTR module may require 2 hours for one page (on a 8 core server) which means that e.g. processing 10.000 pages will sum up to more than 2 years of processing time. Involving a HPC cluster were 1000 cores instead of 8 can be used, will reduce this to 6 days, but if we assume that we process e.g. 100.000 pages it again goes up to 60 days.
- **Training and support fees**
In an Open Platform concept users are enabled to carry out all tasks on themselves. Nevertheless in many cases it might be much more efficient to involve the operators of the Transkribus platform and to benefit from training courses and ongoing project support in order to make the whole process more efficient.
- **Customization fees**
As it is usual with a SAAS platform also specific adaptations can be offered to the customers. E.g. currently the processing of music sheets does not play any role in the Transkribus platform. Given that an archive would be interested to adapt the interface towards this task this customization could be realized by the platform service team. The extension would then also become available to all other users of the platform as well.

4. Market size

4.1. Overview

The following table provides a short overview on expected market size with respect to the different user groups.

Market participants	Number on European level	TRANSKRIBUS Services	Specific measures for involvement	Source of income
Humanities scholars, volunteers, public users	Some 50.000 scholars, some hundred thousand of potential volunteers, public users and family historians	Standard platform services open to everyone Most of the services are targeting for this group: Expert user interfaces (Transkribus, Image Retrieval), crowd-sourcing applications (TSX, mobile versions), e-Learning application	Information campaigns tailored to humanities scholars and their associations Approaching users of archives involved in the project directly	No monetary income, but highly valuable “in-kind contributions” such as documents, data, reputation, number of users
Archives, libraries, memory organisations	Some 1000 large archives, some ten-thousands of small and medium archives and special	Standard platform services, as well as specific services tailored to the needs of an archive or collection.	TRANSKRIBUS is already in contact with several archives and collection holders.	Project based fees and/or Service Level Agreements already during the course of the project

	collections			
Humanities scholars and computer scientist groups applying for research grants	Several hundreds of research groups applying for grants per year, either in the humanities or in computer science	Run the platform and create special services based on TRANSKRIBUS platform	Workshops at conferences (Digital Humanities,) Information Days from archives	Service provider / project partner for or share on project grants
Companies	A handful of start-up companies in the domain of HTR, document processing, etc.	Document and data provision	Model agreements	Fees for data and/or in-kind contributions
Funding agencies	Some 100 funding agencies on national and regional level in Europe	Standard platform services, as well as specific services for research projects	Personal contacts with German Science Funds and Austrian Science Funds	Project based grants and/or service level agreements

4.2. Competitors

To our best knowledge there is currently no service platform or Virtual Research Environment available worldwide which would be able to offer services similar to those of TRANSKRIBUS. As we have seen from our market estimations it is also very unlikely that a start-up company will have the capacity to set up a similar platform acting on a European level. Of course there is a handful of companies which are coming mainly from the Computer Vision, Intelligent Character Recognition and Document Processing domain, such as PLANET (Germany), or A2IA (France). We see these **companies** more **as cooperation partners**, than as competitors. TRANSKRIBUS will offer them a “showroom” for their products and in this way foster further research and development.

To provide a full picture of the market it will also be necessary to have a short look to the global situation of “book digitisation”. In October 2004 Google announced that it will digitize 15 million historical books until 2015 (actually it already reached in 2013 the 30 million milestone) and make it available for free on the Internet. At this time the impact of this decision could barely be assessed. Today we see a clearer: On the one hand more books were digitized and are now available for full-text search on the Internet than without the engagement of Google. For the individual user this is a clear benefit. But on the other hand due to the contractual agreements which Google offered the libraries, today billions of master images are in the hand of a private company, whereas the libraries only get “library copies” with a significantly reduced quality. From a research point of view this is highly questionable and a clear drawback since many operations can only be performed on the (high-quality) master images.

If Google decides to also enter the field of archives and offer them the same deal – free digitisation and provision of secondary digital surrogates vs. access to archival collections – it is obvious that the market for TRANSKRIBUS would change dramatically. Nevertheless we are confident that we would be able to adapt our current model and to find attractive strategies to answer this challenge adequately.

5. Governance model

5.1. General considerations

The governance model plays an important role for implementing the business model. The following considerations can be done already now:

- Public involvement
The governance model should reflect that the envisaged platform has a strong public component: It will provide free services which are especially interesting to other publicly funded organisations (archives, libraries, universities) and to the public in general.
- User and community driven approach
The governance model should also reflect that the success of the platform is directly connected with its popularity among the user groups which means that the users should be involved in a substantial way and not only as customers of free or paid services.
- Service driven approach
In contrast to the first two criteria it should also be part of the governance model that all services will require monetary support and that service fees will play a vital role and also reflect the usage of the platform among the various user groups.

If we look at various governance models we will see mainly three models which can be taken into account. These three models mainly vary in the way they treat the “public” aspect within their governance model. The three models are:

- Associations and foundations
- Companies
- Cooperatives

In the following section we will briefly discuss these options in more detail.

5.2. Association or foundation

An association reflects best the “public” aspect of the platform. A group of voluntary organisation would form this association with the common purpose to run this infrastructure. Associations are typically not-for-profit organisations, often with a tax exemption. The formal requirements are rather low, usually no capital is needed. One of the best known foundations is e.g. the Apache Software Foundation. It was formed in 1999 as a 501(c)3 non-profit charity organisation with the following objectives:

- *provide a foundation for open, collaborative software development projects by supplying hardware, communication, and business infrastructure*
- *create an independent legal entity to which companies and individuals can donate resources and be assured that those resources will be used for the public benefit*
- *provide a means for individual volunteers to be sheltered from legal suits directed at the Foundation's projects*
- *protect the 'Apache' brand, as applied to its software products, from being abused by other organizations (cf. <http://www.apache.org/foundation/how-it-works.html>)*

Several EU funded project followed this approach. Foundations are e.g. the following organizations:

- Europeana Foundation

A portal which provides access to digitized collections via a centralized metadata search.

- Open Preservation Foundation
A membership organisation which provides long-term preservation services.
- ARROW Association
A membership organisation which runs a platform for simplifying the investigations on out-of-commerce books.

5.3. Companies (limited by shares)

In contrast to these organisations which stress their public mission university spin-offs usually are stressing the for-profit aspect and therefore organised as “companies”. In general companies need some share capital, company accounts and some registered office. The main purpose is to increase the shareholder value. Spin-off companies play an important role for universities since in many cases they will keep some shares in exchange to supporting such spin-off companies at their starting phase.

Though there are some kinds of “public” companies (Public limited company, Gemeinnützige GmbH) these forms are rather seldom used.

There are thousands of examples of spin-offs from universities, nevertheless in our specific domain – archives, libraries, processing of historical documents with HTR – to our best knowledge no examples are known.

5.4. Cooperatives

An popular governance model which has a long tradition especially in central Europe is the cooperation. It combines public and private (for-profit) aspects and can be seen as a combination of a foundation and a company. With the concept of “Shareconomy” cooperatives are now discussed as an innovative business model which fits well to some of the requirements of the new economy. And indeed, cooperatives offer a number of advantages compared to the legal entities mentioned above. Cooperatives are not geared toward maximizing profits, but focus on optimizing its service for the co-op members who are also shareholders of the co-op. However, the co-op acts in a commercial manner with its members, its services are completely offset and subject to supply and demand. The members are by no means compelled to purchase a certain service from "their" co-op, provided this service can be bought cheaper on the market. In this way co-ops manage to find a balance between the effectiveness of a private company and the collaborative spirit among communities following the same mission.

Additional special features of co-ops include their direct democratic constitution as well as the legal coverage of a member in the event of bankruptcy. Austrian cooperative law, for instance, ensures that in the event of bankruptcy, no more than twice the amount of the purchased share certificate is owed. Also mentioned is that the share certificates in a cooperative, contrary to other legal forms, can never become an item of speculation as the sale can only ever occur based on the nominal value thereof, but not based on the actual value of a cooperative. It will be an important task in WP3 Network and Business Development to make deeper investigations into this matter and to prepare an informed decision.

6. Appendix: Technology Readiness Levels of the EU Commission

Technology readiness levels (TRL), HORIZON 2020 – WORK PROGRAMME 2014-2015 General Annexes, Extract from Part 19 - Commission Decision C(2014)4995.

Technology Readiness Level	Description
TRL 1.	basic principles observed
TRL 2.	technology concept formulated
TRL 3.	experimental proof of concept
TRL 4.	technology validated in lab
TRL 5.	technology validated in relevant environment (industrially relevant environment in the case of key enabling technologies)
TRL 6.	technology demonstrated in relevant environment (industrially relevant environment in the case of key enabling technologies)
TRL 7.	system prototype demonstration in operational environment
TRL 8.	system complete and qualified
TRL 9.	actual system proven in operational environment (competitive manufacturing in the case of key enabling technologies; or in space)