



DIGITUS II



We reconstruct history unleashing the information trapped in ancient handwritten documents using cutting-edge text recognition and retrieval technologies. In the last decades, archives devoted huge economic efforts in digitizing their documentary collections. Our technology provides an efficient access to that valuable information to scholars, and disseminate it to the general public.



NEEDS

- There is a massive amount of information in ancient manuscripts that cannot be easily accessed.
- Archives made important economic efforts in digitizing their collections. This allows remote access to the documents, but the image format is not convenient for content search.
- The only way to seize those documents is to manually transcribe them, which is prohibitively expensive and time consuming.
- A huge volume of information remains unexplored due to the high processing costs.

SOLUTIONS

- We provide an **efficient**, **scalable** and **affordable** service for automatically transforming manuscript images into searchable content.
- We do it by combining bleeding-edge handwriting recognition algorithms, based on Deep Learning architectures, with information retrieval methodologies.
- Once the data is unleashed, we envision new innovative data mining services for vertical markets.



TEAM



Dr. Marçal Rossinyol

PhD in Computer Science. Associate postdoc researcher at CVC. Two European Marie-Curie fellowships. 7+ years experience in tech. transfer projects.



Juan Ignacio Toledo

BSc and MSc in Computer Science with 15+ years of experience in IT. Expert in Deep Learning for text recognition. Holds an industrial PhD fellowship.



Pau Riba

BSc in Maths and Computer Science and MSc in Computer Vision, all with honors. PhD student with FPU fellowship. Expert on Information Retrieval.

INNOVATIVE AND DIFFERENTIATING FEATURES

Our value proposition is to offer an **efficient** solution for manuscript transcription, that is **scalable** to large document collections. It is also **affordable** in comparison to manually processing this data and is easy to use for non-technical users.

In the recent years, with the explosion of Deep Learning architectures the whole Artificial Intelligence research field has been revolutionized by the outstanding performances in difficult tasks. Specifically, the use of such technology in the domain of handwritten text recognition has swept away the rest of classical state-of-the-art techniques. We will use this technology as an entry barrier to set the difference with the rest of competitors.

Bred within the Document Analysis Group of the Computer Vision Center, one of the most prominent groups in this research area worldwide, our research team has acquired a strong background and a valuable know-how in the processing of ancient handwritten documents.

In the recent years, we have participated in several projects funded by either the European Commission or private companies designing new tools for the automatic processing of historic collections. These projects provided us with the needed experience to develop an innovative generic solution for large scale information extraction suited for any kind of historic cultural asset.

DEVELOPMENT STATUS

We evaluate the current status of our solution by using the Technology Readiness Level (TRL) measure proposed by NASA. It is a measure used to assess the maturity level of a given technology. In a scale ranging from 1: *Basic principles observed*, to 9: *System proven in real operational environment*, we currently are at the last research phase **TRL 4: Technology validated in lab environment.**

Through our diverse collaborations, we have tested our solution in a vast amount of heterogeneous collections of ancient manuscripts from all across Europe, validating the genericity and scalability of our algorithms.

FUTURE STEPS

In order to meet our objectives we will focus our imminent efforts in **development tasks** in order to have as soon as possible a **minimum viable product** that we can start to showcase and test in relevant operational environments.

Following the lean paradigm, we will have several build-measure-learn feedback loop iterations with a set of lead users that will act as early adopters of the technology.

INTELLECTUAL AND INDUSTRIAL PROPERTY

Our IP and copyright are protected by a trade secret strategy. We offer a *Software as a Service* solution, so reverse engineering is avoided. The integrated libraries from third parties have BSD or Apache 2.0 open source licenses, that are permissive. We will study if a software patent is needed if we start operating in the US.

TARGET MARKET AND COMPETITON

Just in Catalonia, there are more than 300 archives storing 840 linear km. of documents from the 9th century to the present. The mission of public archives is not only to preserve their documentary holdings, but also to disseminate their content. Making document images available online, has drastically increased their visibility evidencing the public interest in this information. Offering searching capabilities will allow further advances in this direction. Several companies, such as *FamilySearch*, *Ancestry*, *MyHeritage*, etc. offer a manual processing of document's content, but this is costly and time consuming compared to our automatic solution.

FINANCIAL NEEDS

With an initial investment of 190.000€, we will cover the foreseen research and development expenses for next year. Those expenses include the team salaries, the hiring of a programmer analyst and the acquisition of computation material resources.

This amount comes from CVC's own funding and a number of ongoing private and public project funds.

FINANCIAL PROJECTIONS

	2018	2019	2020	2021
Revenues	59,900€	149,900€	334,800€	644,800€
Expenses	84,500€	109,000€	275,000€	335,000€
Gross margin	-24,600€	40,900€	59,800€	309,800€
Number of costumers	2	5	10	18

ALIANCES, GOALS AND OTHERS

This project counts with the support of the CVC and the UAB. We have a strong background in tech. transfer and FP7 projects on the topic of handwritten recognition, collaborating in the last years with an important number of private companies such as *ITESOFT*, *Qidenus*, *Caixabank*, *Scytl...*

While those international partners are key players in document capture and processing solutions, we also have strong relationships with academic institutions, such as *CED*, and local archive networks, as *XAC*, that will act as lead users providing valuable feedback.

We are also participating in the Market Assessment Program with *EADA Business School* and *ACCIÓ* to help us shape our business model.