# Development of Information Extraction Algorithms from Photographs

**Challenge Owner:** TOELT

Website URL: http://toelt.ai/

TOELT is a research company based in Switzerland that focuses on research on Artificial Intelligence applied in different sciences. It works bringing together infrastructure and technology partners (as NVIDIA) and research entities to work together on impactful research projects, bringing to the game research and artificial intelligence advanced expertise.

**Context**

Many companies struggle to optimise their processes and contracts because lots of information about clients, partners and processes are often stored in documents (PDFs, images, scans, etc.). Extracting text is not always easy, especially if the structure of said text plays a role. OCR methods do not always work, as scan quality, place on paper and many more factors influence the results that are normally not good enough for a wide-spread use of such methods. Sometimes information is not only stored in text, but in small images and in their position.

TOELT is collaborating in the construction and exploration of a new historical dataset on the Italian interwar labour market in partnership with Dr. Aurora Iannello, researcher in economic and social history at the Fondazione Luigi Einaudi in Turin. The purpose of the study is to perform an analysis of the social security cards stored at the archives of the National Institute for Social Security (INPS) and the Einaudi Foundation.

**Challenge**

The cards record the working history of Italian employees since 1920, when the subscription to the national system of social insurance became mandatory. A sample of several hundred private-sector employees active in the 1920s and 1930s will be selected and digitized from the INPS archives of Vercelli, Turin, Rome and Naples.

The goal of this challenge is, by using this dataset, the development of new AI Based algorithms for the extraction of information, in this case, from images.

**Expected demonstration:**

This project has the main aim of developing new algorithms that will be able to be used to extract information from PDFs and images. The goal is not to build a better OCR (Optical Character Recognition) software, but build an AI based software that will understand the context of the texts and images and translate that into a digital format. Multiple objectives will be considered in increasing order of difficulty. For example, extract number and position of stamps based on the colour and pattern, extract specific information about persons (as birth date, etc.), train models to recognize handwritten text from the beginning of last century, and so on.

**Available datasets:**

A dataset of 5000 images of several hundred workers will be available. The privacy concerns will be addressed together with possible groups that will be selected for the challenge and with INPS (which owns the cards). The ultimate goal is to make all the data available and free to use for research purposes to everyone.