OCR, Classification & Machine Translation (OCCAM)

Joachim Van den Bogaert, Arne Defauw, Frederic Everaert, Koen Van Winckel, Alina Kramchaninova, Anna Bardadym, Tom Vanallemeersch
CrossLang
Kerkstraat 106
9050 Gentbrugge
Belgium
{first.lastname}@croslang.com

Pavel Smrž, Michal Hradiš
Brno University of Technology
Božetěchova 2
612 00 Brno
Czech Republic
smrz@fit.vut.cz,
ihradis@fit.vutbr.cz

Abstract

The OCCAM project (Optical Character recognition, ClassificAtion & Machine translation), which runs from 2019 to 2021, and is carried out by CrossLang and Brno University of Technology, aims at integrating the CEF (Connecting Europe Facility) Automated Translation service with image classification, translation memories, optical character recognition, and machine translation. It will support the automated translation of scanned business documents (a document format that, currently, cannot be processed by the CEF eTranslation service) and will also lead to a tool useful for the digital humanities domain.

1 Introduction

The European Commission's Business Registers Interconnection System (BRIS) facilitates the access to information on EU companies and ensures that all EU business registers can communicate to each other electronically, in relation to cross-border mergers and foreign branches.[1] Its main task is to synchronize the information that is present within Members States' business registers.

The CEF has planned an integration of BRIS with the CEF eTranslation[2] Digital Service Infrastructure (DSI), to make draft translations of company information available via the European e-Justice Portal,[3] but a large volume of scanned documents would remain untranslated, because it consists of raw images that are not machine-readable.

A similar problem occurs in the digital humanities domain: while there are plenty of optical character recognition (OCR) frameworks available (both open source and proprietary), the need for OCR and translation within the digital humanities domain remains pressing. The European Newspaper Survey Report,[4] as conducted during the Europeana Newspapers project, revealed that access to twentieth-century content remains problematic, and only few libraries use OCR when scanning documents. At the same time, there is a growing interest in gaining multilingual access to cultural heritage resources.

2 Proposed solution for BRIS

Existing content within the member state databases will be leveraged to recognise, classify and translate legacy and new content. The presence of database links to scanned documents, and the template-like nature of administrative documents will be exploited to optimize OCR and translation. Since a *pipelined* (cascaded) implementation (i.e. an OCR step followed by a machine translation (MT) step) has the inherent risk of error accumulation, OCCAM proposes a more informed classification-based approach, as outlined in Figure 1, to:

[1] https://ec.europa.eu/cefdigital/wiki/pages/viewpage.action?pageId=46992657

[2] https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation

[3] https://e-justice.europa.eu/

[4] http://www.europeana-newspapers.eu/wp-content/uploads/2012/04/D4.1-Europeana-newspapers-survey-report.pdf

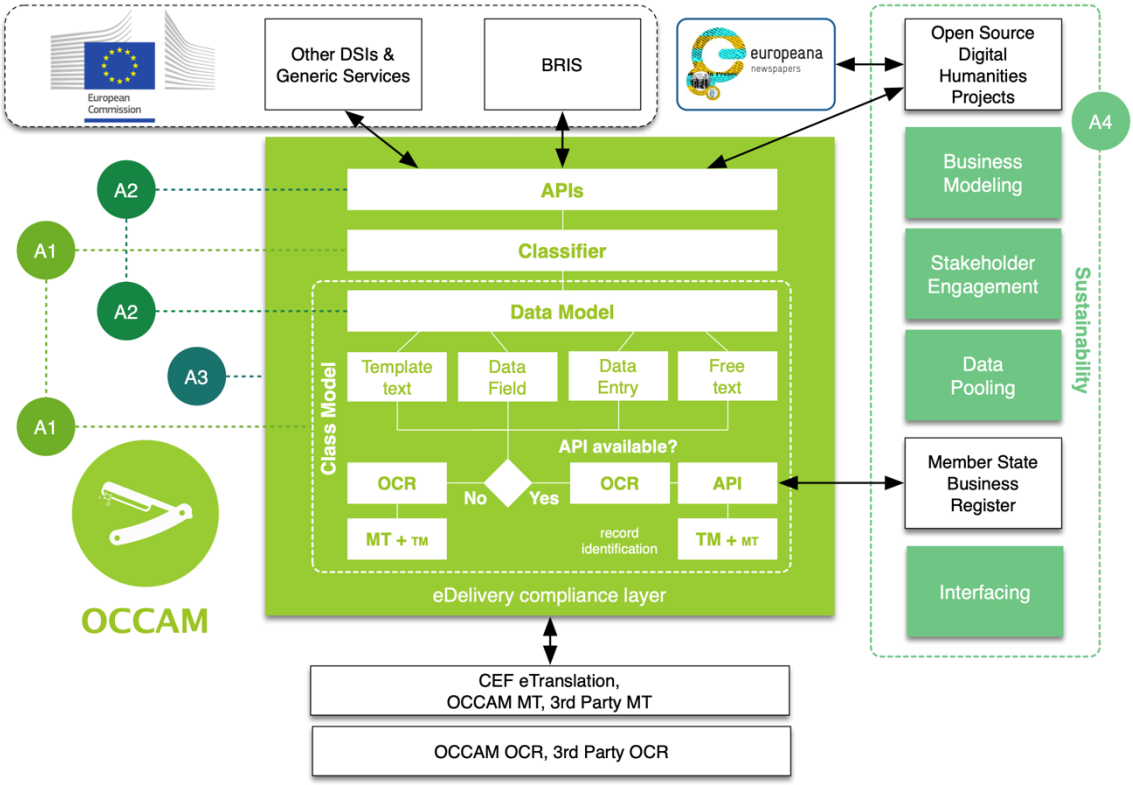

Figure 1: OCCAM system architecture

- recognise document types and link them to a corresponding data model (consisting of template text, data fields, data entries and free text);
- identify entities within documents and link them to corresponding entries in member state databases (e.g. by using OCR to recognise VAT numbers and retrieve the corresponding data from a national business register);
- retrieve translations from translation memoriess associated with the data model;
- use class-adapted OCR and MT for the remaining free text.

Brno University of Technology will provide the image classification and OCR tools, using the open source packages OpenCV,[5] Tesseract,[6] and TensorFlow,[7] and an in-house neural OCR system currently developed for the analysis of challenging historical documents in a project called PERO,[8] aimed at improving accessibility of cultural heritage. These tools are distributed under commercially-friendly (non-viral) licenses (3-clause BSD and Apache 2.0).

CrossLang will build MT engines using CEF eTranslation's MT system and its own Moses-based SMT (distributed under LGPL license) and OpenNMT-based (distributed under MIT license) systems. The developed MT systems will incorporate maned-entity recognition and terminology technology, and target the following language pairs: Dutch, French, German, and Czech into English.

The resulting OCCAM solution will be built as a reference implementation and made publicly available at the end of the project. The licenses of the used components will ensure that the implementation can be distributed freely, and adapted for use by business registers across Europe, after the project has ended.

For the digital humanities domain, a technology roadshow will be organised and tutorials will be published, to make researchers acquainted with the technology, so they can easily develop and adapt their own models.

Acknowledgement

OCCAM is funded by the EC's CEF Telecom programme (project 2018-EU-IA-0052).

[5] https://opencv.org/

[6] https://github.com/tesseract-ocr/tesseract

[7] https://www.tensorflow.org/

[8] http://www.fit.vutbr.cz/units/UPGM/grants/index.php.en?id=1165