D2.4 Operational platform with additional tools integrated

Support Action Centre of Competence in Digitisation (Succeed)

December 4, 2014

Abstract

Deliverable 2.4 (*Operational platform with additional tools integrated*) is an online service and open-source platform which complements and enhances deliverable D2.3 (*Operational platform*, submitted in December 2013 and approved in February 2014). This report summarises the digitisation tools incorporated into the online platform, as well as the management of the releases of its open-source components.

In 2014, 13 tools have been integrated in addition to those initially available in the operational platform or subsequently added during Succeed's first year. The operational platform, currently provides 41 tools for text digitisation which can be executed and tested online without the need to install any software on the user's computer.

In addition to the aforementioned integration of tools, new versions of the platform have been released in order to improve its functionality, sustainability and interoperability.

Succeed is supported by the European Union under FP7-ICT.

Document information

Deliverable number	D2.4	Start: M1	Due: M22	Actual: M23
Deliverable name (DoW)	Operational platform with additional tools integrated			
Internal/External	External			
Activity type	SUPP			
Participant	UA, **KB**, INL, IAIS			
Estimated person months for this deliverable	10.00			
Dissemination level	PU (public)			

Document history

Revisions				
Version	Status	Author	Date	Changes
0.1	Draft document	Isabel Martínez	14/11/2014	
0.2	Initial version	Rafael C. Carrasco	26/11/2014	
1.0	Final version	Rafael C. Carrasco	04/12/2014	

Approvals				
Version	Date of approval	Name	Role	Signature
0.2	26/11/2014	Tomasz Parkola (PSNC)	Internal supervisor	

Distribution (this document was sent to)			
Version	Date of sending	Name	Role in project
0.2	26/11/2014	Tomasz Parkola	Internal supervisor
1.0	04/12/2014	Cristina Maier	Project Officer

Succeed is supported by the European Union under FP7-ICT and coordinated by Universidad de Alicante.

About this document

This document is a public report describing deliverable D2.4 of the Succeed project (FP7-ICT-600555).

Copyright statement

[1] http://creativecommons.org/licenses/by-sa/3.0/

Contents

1 Introduction

Deliverable 2.4 is part of Succeed WP2 (*Interoperability and infrastructure*) and meets the following objectives in the project's Description of Work:

O2.1 To provide and maintain the necessary technical infrastructure for the support action.

O2.2 To facilitate access to an interoperable platform showcasing digitisation tools and support the integration of new or existing tools created by international research groups.

In particular, D2.4 —together with its accompanying deliverable D2.3— is an output of the following tasks in WP2:

T2.1 Interoperable platform.

The objective of this task is to provide and maintain the interoperable service platform created during the IMPACT project (`www.impact-project.eu`) and to perform the necessary adaptations for the purposes of the support action.

T2.2 Tool integration.

The objective of this task is to integrate at least the five most relevant software applications that have been identified in WP3 into the interoperable service platform .

Deliverable 2.4 is an online service and open-source platform which complements and enhances deliverable D2.3 *Operational platform*, submitted in December 2013 and approved in February 2014. Due to the intangible nature of deliverable D2.4, this report only summarises the digitisation tools incorporated into the online platform and the management of the releases of its open-source components.

The interoperable platform has been deployed in and integrated into the the Impact Centre of Competence website (`www.digitisation.eu`) for a more effective dissemination and easier sustainability. For additional details on the architecture, deployment and licensing of the platform, please refer to D2.3 (published at `succeed-project.eu/publications`).

2 Summary of results

The operational platform is a software suite which was first created by the Impact project (2008–2011). The platform provides online access to a number of tools for digitisation which would otherwise work independently. In other words, the operational platform supports the interoperability of tools created by different institutions and simplifies the testing of these tools, working alone

or in combination with others. Due to its web-service orientation, users do not need to install, maintain or execute any software in their computers. This independence with regard to the user's system is a key feature of the platform which helps raise awareness of the methods and tools which are available for every digitisation step.

At the end of Impact, around 20 tools —listed in section 3— could be executed and tested with the platform. The platform has evolved in two directions since then: improving its usability and increasing the number of tools being supported. In 2013, 7 additional tools listed in section 4 were integrated and also key features such as advanced user management with security features —user profiles, personalised access control, usage quotas, etc) were implemented by the Succeed project. Section 5 lists 13 additional tools whose integration is the core of deliverable D2.4. It also briefly describes the integration of the Impact repository (images and ground truth collections) into the platform. A total of 20 new tools have thus been added to the platform by Succeed, far beyond the minimum number established in the project's Description of Work (DoW), which stated "at least five".

Section 6 describes how the new releases of the platform have been managed and the main improvements carried out after each version. Finally, section 7 describes the activities (*hackathons*) organised around the operational platform.

3 Initial set of tools

This section compiles the original list of tools which were active at the beginning of the project. Succeed worked both on the maintenance of these tools (updating the platform every time new versions of the tools were distributed by the providers) and the improvement of its usability (adding, for example, drag and drop capabilities to the website in order to ease tool usage).

NAME: Graphics Magick

TYPE: Image transformation

SHORT DESCRIPTION: Graphic Magick provides a robust and efficient collection of tools and libraries which support the reading, writing, and manipulating of an image in over 88 major formats.

NAME: IMPACT OpenJPEG

TYPE: Image transformation

SHORT DESCRIPTION: Conversion between JPEG2000 and TIFF image file formats with an implementation based on the OpenJPEG library.

NAME: IMPACT NCSR Border Removal Service

TYPE: Image Enhancement

Short description: Automatic detection and removal of black borders as well as noise regions from scanned document image files.

Name: IMPACT NCSR Geometric Correction Service

Type: Image Enhancement

Short description: Automatic correction of geometric distortions typically found in scanned document image files.

Name: NCSR Binarisation Service

Type: Image Enhancement

Short description: Image binarisation using an algorithm developed at NCSR.

Name: IMPACT Abbyy FineReader 10 Binarisation Service

Type: Image Enhancement

Short description: Image binarisation using Abbyy FineReader 10 technology

Name: IMPACT Abbyy FineReader 10 PAGE Segmentation Service

Type: Segmentation and document analysis

Short description: Segmentation of an input image file using Abbyy FineReader 10.

Name: IMPACT USAL Text line and Word Segmentation Service

Type: Segmentation and document analysis

Short description: Text line and word segmentation with an algorithm created by the University of Salford.

Name: IMPACT NCSR Character Segmentation Service

Type: Segmentation

Short description: Segmentation of words into characters.

Name: IMPACT ABBYY FineReader 10 OCR Service

Type: OCR engines

Short description: OCR on an input image file using Abbyy FineReader 10 technology.

Name: IMPACT FineReader 10 Service

Type: OCR engines

Short description: OCR on an input image file using IMPACT dictionaries with Abbyy FineReader 10 technology.

Name: IMPACT USAL Typewritten OCR Service

TYPE: OCR Engines

SHORT DESCRIPTION: OCR on a typewritten document image file.

NAME: IMPACT Tesseract 3.02 OCR Service

TYPE: OCR Engines

SHORT DESCRIPTION: OCR on an input image file using Tesseract 3.00 technology.

NAME: Gocr

TYPE: OCR engines

SHORT DESCRIPTION: GOCR is an OCR program, developed under the GNU Public Licence. It converts scanned images of text back to text files.

NAME: OCRad

TYPE: OCR Engines

SHORT DESCRIPTION: GNU Ocrad is an OCR (Optical Character Recognition) program based on a feature extraction method.

NAME: IMPACT NCSR OCR Evaluation Service

TYPE: Evaluation

SHORT DESCRIPTION: OCR evaluation by comparing text results from an OCR engine with ground truth text data.

NAME: IMPACT USAL Layout Evaluation Service

TYPE: Evaluation

SHORT DESCRIPTION: Evaluation of segmentation by comparing the results in PAGE format with ground truth according to predefined profiles.

NAME: IMPACT INL Word Evaluation Service

TYPE: Evaluation

SHORT DESCRIPTION: Word evaluation of OCR by comparing the results in PAGE format with ground truth.

NAME: IMPACT Iconv Encoding Conversion Service

TYPE: File Type and Encoding

SHORT DESCRIPTION: Conversion of character encoding using Iconv.

NAME: IMPACT USAL Ground Truth Normalisation Service

TYPE: File Type and Encoding

SHORT DESCRIPTION: Normalisation of ground truth in PAGE format according to predefined filter rules.

NAME: Xsltproc

TYPE: File Type and Encoding

SHORT DESCRIPTION: Xsltproc is a command line tool for applying XSLT stylesheets to XML documents. It is part of libxslt, the XSLT C library for GNOME.

4 Additional tools (first year)

The selection of tools integrated into the platform during the project —and summarised in this one and in the next section—, is based on the outcomes of task T3.1 (*Survey of tools*) which maintained a comprehensive and updated list of the tools available for digitisation. The selection also took into account the feedback provided by the libraries participating in task T3.3 (*Take-up support*). The following tools were added during the first year of the project (months 1 to 12).

NAME: mydec

TYPE: Digitisation platforms

SHORT DESCRIPTION: `mydec` is a service platform provided by Fraunhofer IAIS for the digitisation of documents including, for example, image processing, OCR and metadata enrichment. For easier maintenance, since the mydec service has been recently discontinued, the tools were later separately integrated into the platform, as listed in the next section.

NAME: Ocropus

TYPE: OCR engines

SHORT DESCRIPTION: Ocropus is an open source OCR and machine learning system for document analysis.

NAME: Kakadu

TYPE: Image transformation

SHORT DESCRIPTION: Kakadu is a commercial software library for the encoding and decoding of images in JPEG2000 format.

NAME: Texteval

TYPE: Evaluation

SHORT DESCRIPTION: Texteval is an alternative OCR evaluation tool from PRImA research.

NAME: ocrevalUAtion

TYPE: Evaluation

Short description: `ocrevalUAtion` is an alternative and open-source OCR evaluation tool created by the Universidad de Alicante.

Name: Mallet

Type: Other

Short description: Mallet is a popular software framework for machine learning.

Name: Gamera

Type: Segmentation and document analysis

Short description: Gamera is a Python framework for building document analysis applications.

5 Additional tools (second year)

This section compiles the tools which have been integrated into the operational platform within the extent of deliverable D2.4 (months 13 to 22).

Name: Fraunhofer Newspapers Segmenter & Korrektor

Type: Segmentation and document analysis

Short description: The Korrektor is a manual post-correction tool for automatically processed newspaper scans. By loading the result XML files into the software, it is possible to correct automatically detected layout elements, texts and other properties. The scanned documents are displayed in two separate windows to allow for a detailed inspection. Results can be edited using context menus, drag and drop and keyboard shortcuts.

Name: XML to text

Type: File type and encoding

Short description: Conversion of XML files to raw text.

Name: Taverna 2 Server Client

Type: Other

Short description: The Taverna 2 Server Client can be used to execute workflows (sequences of operations) created with Taverna 2.

Name: Gimp Image Conversion

Type: Image transformation

Short description: GIMP is a raster graphics editor used for image retouching and editing, free-form drawing, resizing, cropping, photo-montages, converting between different image formats, and other specialised tasks.

NAME: Exiftool

TYPE: Image transformation

SHORT DESCRIPTION: ExifTool is a free software program for reading, writing, and manipulating image, audio, and video metadata. It is platform independent, available as both a Perl library and command-line application. ExifTool is commonly incorporated into different types of digital workflows and supports many types of metadata including Exif, IPTC, XMP, JFIF, GeoTIFF, ICC Profile, Photoshop IRB, FlashPix, AFCP and ID3, as well as the manufacturer-specific metadata formats of many digital cameras.

NAME: Cuneiform

TYPE: OCR engines

SHORT DESCRIPTION: CuneiForm is a software tool for OCR. It was originally developed at Cognitive Technologies and, after a few years with no development, released as freeware on December 12, 2007. The kernel of the OCR engine was released under the open source BSD licence at the beginning of April 2008.

NAME: FR10 to PAGE XML exporter

TYPE: File type and encoding

SHORT DESCRIPTION: Exports FR10 format to PAGE XML.

NAME: Deskewer (mydec)

TYPE: Image enhancement

SHORT DESCRIPTION: The alignment of input images is corrected in order to improve the detection of structures and text.

NAME: Layout analysis (mydec)

TYPE: Segmentation and document analysis

SHORT DESCRIPTION: It automatically detects the layout and identifies the text regions (paragraphs, headings, etc) contained in the scanned page.

NAME: Color binarize (mydec)

TYPE: Image enhancement

SHORT DESCRIPTION: Color binarize separates letters from the background. Grey-scale images are converted to binary. It can be calculated for the separation either for the entire image or for each pixel of the optimal contrast.

NAME: Deshadow (mydec suite)

TYPE: Image enhancement

Short description: This tool reduces noise caused by transparencies in the physical document.

Name: Cutouts

Type: OCR engines

Short description: Cutouts supports preparation of the proper training material for the OCR system.

Name: INL Lemmatizer

Type: Language processing

Short description: Tagger and lemmatizer for historical Dutch.

Name: INL NERT

Type: Language processing

Short description: It performs recognition and tagging of named entities (persons, locations and organisations) in a text file.

Name: JHOVE2

Type: File type and encoding

Short description: The JHOVE2 project generalises the concept of format characterisation to include identification, validation, feature extraction, and policy-based assessment.[2]

Name: Stanford NER

Type: Language processing

Short description: Stanford NER is a tool that can mark and extract named entities (persons, locations, organisations or even titles) from a text file. It uses a supervised learning technique, which means it has to be trained with a manually tagged training file before it is applied to a different text.

Apart from the tools incorporated in 2014, a viewer has been implemented which simplifies access to the open data-sets created by the Impact project: 250,000 high-resolution images and 30,000 ground truth files. The service uses the protocol (API) provided by the University of Salford. For example, images can be browsed, previewed and searched for using key words (language, institution, title and publication year).

[2] https://confluence.ucop.edu/download/attachments/3932229/Abrams_a70_pdf.pdf?version=1.

6 Release management

Release management is the process through which software is managed from the development phase to its publication and distribution stage. Software products typically run in an ongoing development, testing, and release cycle. The operational platform maintained by Succeed has been scheduled to produce periodic versions and the history of new releases in the reporting period (months 13 to 24) is summarised below. These releases are updated on GitHub.[3]

Version number: 0.7

Date of release: March, 2014

Main improvements:

- Timeout in generic SOAP-client dependency. The web service client produced timeouts when executing highly time-intensive processing requests (e.g. Typewritten OCR). The fix required updating the generic SOAP client.
- Problems with the output file name. The parameter OutputFileName was not working correctly when the tool sent output to the standard output stream. This problem has been solved in release 0.7.

Version number: 0.8

Date of release: November, 2014

Main improvements:

- Removal of temporal files is made optional in order to provide additional information during the debugging phase.
- The container website communicates to the platform both the user name and password of users who are already logged-in; no new log-in is therefore needed to access the tools and services after having registered with the website.

Version number: 0.9

Date of release: December, 2014

Main improvements:

- Enhanced support for tool input. A custom drag and drop application has been implemented in order to support the optional uploading of the input data.

[3] `https://github.com/impactcentre/interoperability-framework/releases`

- New releases are automatically published as Maven artefacts on bintray, a free online service for binary code distribution.
- The style-sheet to be employed is established in a configuration file, making it simpler to maintain a consistent look-and-feel when the platform is deployed on a particular website.

A new release is planned for January 2015 which will include, at least, enhanced support for testing. Test units automatically detect mistakes and bugs caused by software evolution. The new version will provide a comprehensive suite of test units meant to facilitate the quality control in future releases.

7 Developers' Workshops

Two developers' workshops (D2.1 and D2.2) were organised during the Succeed project as effective instruments to provide training and hands-on experience on the operational platform, as well as to collaborate with external tool developers who might be willing integrate their tools into the platform.

The first of these workshops was held at the KB, in the Hague in September 2013. The second one was held on 10-11 April, 2014 at the Universidad de Alicante. Some interviews were recorded during this workshop, and they are available at `https://www.youtube.com/user/theimpactproject`. A blog post with detailed information about the outcomes of this hackathon was additionally published at `http://digitisation.eu/blog` in the Succeed section (`http://www.digitisation.eu/blog/succeed-2nd-hackathon`).

8 Conclusions

During this period 13 tools have been integrated in addition to those initially available on the operational platform or subsequently added during Succeed's first year. At present, the operational platform provides 41 tools for text digitisation which can be executed and tested online without the need to install any software on the user's computer.

In addition to the aforementioned integration of tools, three new versions of the platform have been released in order to improve its functionality, sustainability and interoperability.

Participant libraries have been using the platform during the validation performed in WP3 to test commercial tools (such as FineReader 10 or PageCurl) without the need to apply for individual licenses. The workload during these trials and also with virtual requests (submissions which were automatically generated) were observed with a specific monitor that showed the stability of the platform under operation.