succeed

2nd Succeed Hackathon

10-11 April 2014, University of Alicante

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

First off: how do we roll?

Day 1 / 10 April 2014

12:00 - 12:30 Registration, coffee, meet & greet

12:30 - 13:15 Lunch

13:15 - 13:30 Background & Topics

13:30 - 14:00 Introduction of participants, forming of groups

14:00 - 17:00 Hacking time

17:00 - 17:30 Status round-up

19:00 - 21:00 Social dinner at *La Taberna de Tito*

Day 2 / 11 April 2014

09:30 - 10:00 Coffee

10:00 - 12:30 Hacking time

12:30 - 13:30 Lunch

13:30 - 16:00 Hacking time

16:00 - 17:00 Presentation of results

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

Background

- Libraries, archives, museums are digitising large quantities of (mainly historic) documents like books, newspapers, journals.
- To make these digital documents searchable, images must first be converted to electronic text with help of OCR (optical character recognition)
- Off the shelf OCR software is not suitable for processing historical documents – problems e.g. with old fonts, historic language variation, quality of paper originals
- 2009-2012: EU project IMPACT (IMProving ACcess to Text) (**www.impact-project.eu**)
- 2011: Official launch of the IMPACT Centre of Competence (**www.digitisation.eu**)
- 2013-2014: EU project SUCCEED (**www.succeed-project.eu**)

Tools, tools, tools

“I have this great tool, it does XYZ…

…if you call it with the right parameters…

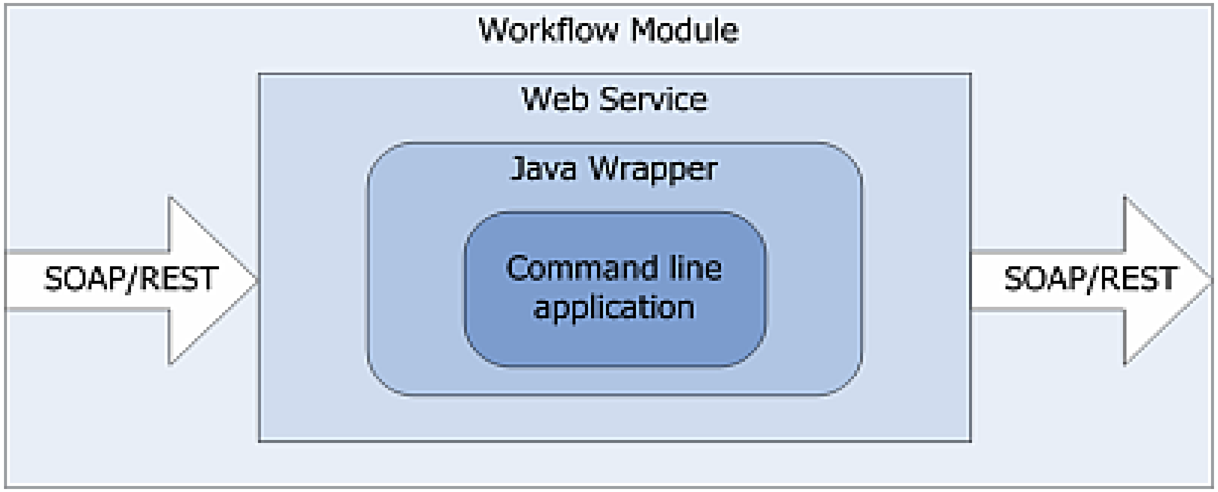
…and run it in this environment…

…with these minor tweaks…”

Succeed maintains an extensive list of tools for digitisation:
http://succeed-project.eu/publications/available-tools/index-succeed

IMPACT Platform

Interoperable, web-based platform for testing (combinations of) tools (workflows)

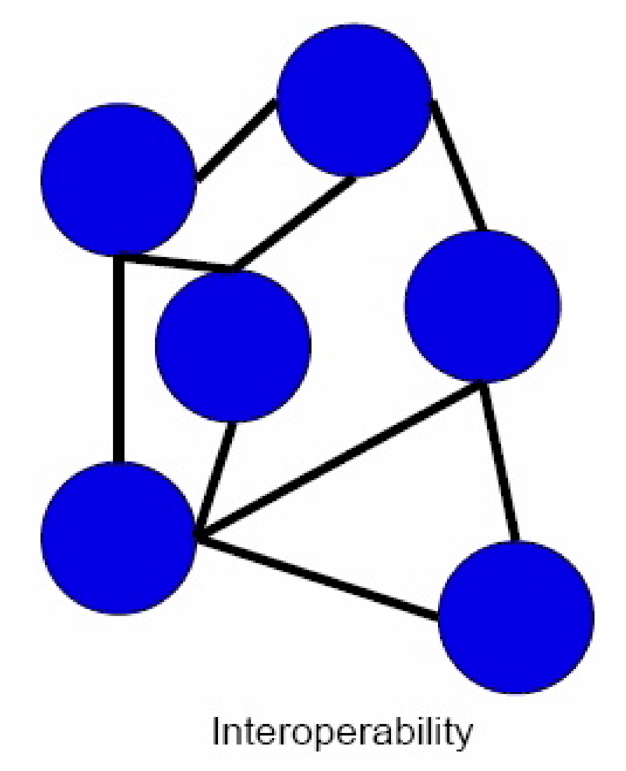


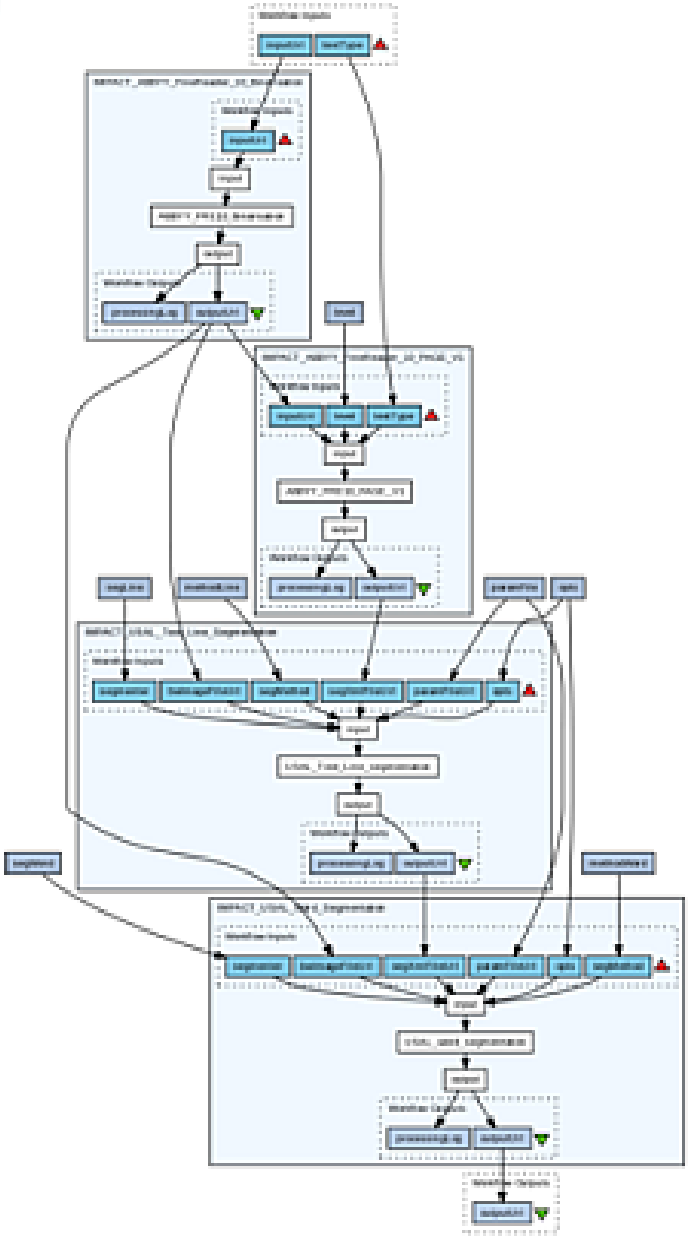
Software modules have been released as separate open source modules on **https://github.com/impactcentre**

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

Interoperability

VS





Components

The platform comprises of the following components/modules:

- **iif-toolwrapper**
  = a Java application for creating a web service wrapper project for command line tools

- **iif-wsdl-client**
  = a web application that can be used to test the operations of a SOAP web service

- **iif-generic-soap-client**
  = a Java library that can execute operations of an arbitrary SOAP web service

- **iif-taverna2-client**
  = a web application to remotely execute workflows on Taverna 2 Server

- **iif-resultsrepository**
  = a custom SOAP web service that stores files into a WebDAV repository

Bintray/Maven integration

All software binaries can be directly downloaded from
https://bintray.com/impactocr/maven

You can also integrate them into your Maven project by adding this to your pom.xml:

```
<repositories>
    <repository>
        <id>impactocr</id>
        <url>http://dl.bintray.com/impactocr/maven/</url>
     </repository>
</repositories>

<dependency>
    <groupId>eu.impact_project.iif.ws</groupId>
    <artifactId>generic-soap-client</artifactId>
    <version>0.7.0</version>
</dependency>
```

Action!

- Web Service Client
  http://succeed.kbresearch.nl/WS-Client/

- Taverna2 Workbench + myExperiment

- Workflow Client
  http://succeed.kbresearch.nl/dp/

More and more...

OCR evaluation tool
(**https://github.com/impactcentre/ocrevalUAtion**)
= compare ground truth with OCR result

PAGE generator
(**https://github.com/psnc-dl/page-generator**)
= generate training data for Tesseract OCR from PAGE xml

Franken+
(**https://github.com/idhmc-tamu/FrankenPlus**)
= create new training sets for Tesseract OCR

Format converter
(**https://github.com/subugoe/format-converter**)
Convert between different text formats, e.g. ALTO, TEI, FRXML

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

Last but not least...

Be creative and have fun coding and experimenting!

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