DELIVERABLE SUBMISSION SHEET

To: Marcel Watelet (Project Officer)

DG Communications Networks, Content & Technology
Creativity Unit
EUFO 01/196A- European Commission
L-2920 Luxembourg

From: Support Action Centre of Competence in Digitisation

Project acronym: Succeed Project Number: 600555

Project Manager: Rafael C. Carrasco Jiménez

Project Coordinator: Universidad de Alicante

The following deliverable:

Deliverable title: Report on validation criteria and procedures for tools and resources

Deliverable number: D3.3

Deliverable date: M11

Partners responsible: IAIS

Status: ☒ Public ☐ Restricted ☐ Confidential

is now complete

☒ It is available for your inspection

☒ Relevant descriptive documents are attached

The deliverable is:

☒ A document

☐ A website Url:

☐ Software

☐ An event

☐ Other

Sent to Project Officer:	**Sent to functional mailbox:**	**On date:**
Marcel.Watelet@ec.europa.eu	CNECT-ICT-600555@ec.europa.eu	20/12/2013

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

D3.3 Report on validation criteria and procedures for tools and resources

Succeed

20/12/2013

Abstract This deliverable is part of WP3. This work package will support the validation of digitization tools, linguistic tools and resources created by either commercial parties or research and development programs and their transference for exploitation in libraries and other cultural heritage organizations. The objective of this deliverable is to define validation parameters and procedures for the implementation of tools in productive environments. The criteria are then used to validate each tool (or group of tools in the case of interconnected tools that operate together).The evaluation criteria are not defined per tool, but per task which will be carried out by using a tool.

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

Document information

Deliverable number	D3.3	Start: 7	Due: 11	Actual: 12
Deliverable name	Report on validation criteria and procedures for tools and resources			
Internal/External	External			
Activity type	R			
Participant	UA, INL, **IAIS**, PSNC			
Estimated person months	1 PM			
Dissemination level[1]	PU			

Document history

Revisions				
Version	**Status**	**Author**	**Date**	**Changes**
0.1	Draft	Sebastian Kirch (IAIS)	13-11-2013	
0.2	Draft	Bob Boelhouwer (INL)	11-12-2013	
0.3	Draft	Jesse de Does (INL)	15-12-2013	
0.4	Draft	Katrien Depuydt	16-12-2013	
1.0	Final	Katrien Depuydt	20-12-2013	Final version

Approvals				
Version	**Date of approval**	**Name**	**Role in project**	**Signature**
0.4		Isabel Martínez	Internal review	
1.0		Isabel Martínez	Internal review	OK

Distribution			
This document was sent to:			
Version	**Date of sending**	**Name**	**Role in project**
0.1	13-11-2013	Katrien Depuydt	WP3 co-lead
0.4	17-12-2013	Isabel Martinez (UA), Rafael Carrasco (UA)	Internal review
1.0	20-12-2013	Isabel Martínez (UA), Marcel Watelet (EC)	

[1] PU Public; RP Restricted to other programme participants (including Commission Services); RE Restricted to a group specified by the consortium (including Commission Services); CO Confidential, only for members of the consortium (including the Commission Services)

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

Table of Contents

1. Introduction 4

2. Methodology 5

3. Overview on libraries and selected tools 6

3.1. External Libraries 6
3.2. Internal Libraries 6

4. Evaluation Tasks and Implementation Plans 7

4.1. Usability Evaluation 7
4.2. Image Processing and Enhancement Evaluation 10
4.3. Layout Analysis Evaluation 15
4.4. Text recognition (OCR) Evaluation 18
4.5. Evaluation of OCR Evaluation tools 21
4.6. Text processing: Named Entity Recognition and Resolution Evaluation 24
4.7. Text Processing: NE Linking (Resolution) 28
4.8. Text Processing: Linguistic annotation tools / tools for manual annotation and verification 30
4.9. Text Processing / NLP Tools / Lexicon as a Web Service 32
4.10. Miscellaneous tools: JHOVE2 34

1. INTRODUCTION

This deliverable is part of WP3. This work package will support the validation of digitization tools, linguistic tools and resources created by either commercial parties or research and development programs and their transference for exploitation in libraries and other cultural heritage organizations. In particular, a selection of the tools compiled in the Survey of Tools[2] (deliverable 3.1) will be made available to the community and the partners involved will provide assistance for the adaptation of the tools to specific domains and languages as well as training in the usage of tools.

The objective of this deliverable is to define validation parameters and procedures for the implementation of tools in productive environments. The criteria are then used to validate each tool (or group of tools in the case of interconnected tools that operate together). The evaluation itself and the corresponding test scenarios will be worked out in cooperation with libraries, based on an analysis of library requirements, and will be summarized in a work plan for each library.

The evaluation criteria will not be defined per tool, but per task which will be carried out by using a tool. Therefore, this deliverable is organized as follows: Firstly, the methodology for compiling this deliverable is described. Secondly, an overview of the tasks and corresponding tools that have been selected by the libraries is given. Thirdly, per task an implementation plan for the evaluation and an evaluation form is presented.

[2] Overview on available tools for text digitisation: http://succeed-project.eu/publications/availabletools/index-succeed.

2. METHODOLOGY

The production of evaluation material can be summarized in five consecutive steps. The dates when each step was performed are given in parentheses.

1. Document Structure (July 2013)

During a workshop held at Fraunhofer IAIS on the 9th of July, quality criteria per tool type were discussed as well as the usability criteria. It was decided convert the usability criteria into a template, to be used for all tools, and to produce per task an evaluation form for quality assessment. For the latter, it was decided to produce an example evaluation form for Text Recognition

2. Tool Selection (August – October 2013)

Before the production of evaluation material could be started, the participating libraries had to choose which tools they wanted to evaluate within the Succeed project. Libraries were supposed to select the tools by the end of September. However it took some libraries as long as end of October to select the tools, which is the main reason why this deliverable was slightly delayed.

3. Distribution of Work (October 2013)

The production of evaluation forms was done by INL and IAIS according to the experience and knowledge the partners already had with the selected tools.

4. Creation and Compilation of Evaluation Material (October – December 2013)

Per task/tool type, separate evaluation forms were produced as well as a usability evaluation form, each accompanied by an implementation plan. The final versions of these documents were then aggregated by the WP3 leaders to produce this deliverable.

5. Distribution of Evaluation Material (December 2013)

After the evaluation material had been finalized it was distributed by the partners to the libraries using the respective tools.

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

3. OVERVIEW ON LIBRARIES AND SELECTED TOOLS

3.1. External Libraries

Library	Country	Selected Tools
Wielkopolska Biblioteka Cyfrowa	Poland	- Scan Tailor - JHOVE2 - Image Magick
General Historical Library of Salamanca	Spain	- Gimp - Omnipage
Wroclaw University Library	Poland	- Scan Tailor - Tesseract OCR
University Library of Bratislava	Slovak Republic	- Scan Tailor - ImageMagick
National Library of Finland	Finland	- Newspaper segmentation - Korrektor - Document Deskewer
Library of the University of Granada	Spain	- Scan Tailor - Alchemy API
University Library of Leuven	Belgium	- OCR: Abbyy FRE - NERT
University Library of Antwerp	Belgium	- NE Attestation tool, - NLTK (NE), - Stanford (NE)
University Library of Darmstadt	Germany	- Newspaper segmentation - Korrektor - Document Deskewer

3.2. Internal Libraries

Library	Country	Selected Tools
Biblioteca Virtual Miguel de Cervantes	Spain	- Abbyy FRE - Geometric correction: Page Curl - COBaLT - Lexicon as Webservice
Bibliotèque nationale de France	France	- DBPedia Spotlight - Evaluation Tool for OCR - Lexicon as Webservice
Koninklijke Bibliotheek	Netherlands	- Lexicon as Webservice - NLTK - NERT
The British Library	United Kingdom	- Lexicon as Webservice

4. EVALUATION TASKS AND IMPLEMENTATION PLANS

The criteria for evaluation will be based on technical validity, scientific performance measures (as applicable in the context of content holding institution's data), and measures of the productivity and cost effectiveness obtained with the application of the tool. Therefore, the libraries have to fill out two evaluation forms per tool. One evaluation form is about the usability of tools and covers measures of the productivity and cost effectiveness. The other evaluation form is specific to the tool type and includes parameters on the technical validity and scientific performance measures.

The preferred evaluation processes and the corresponding evaluation forms are described in the following sections.

We encourage the libraries to publish the evaluation datasets, preferably under CC by-nc-sa license.

4.1. Usability Evaluation

4.1.1. Introduction

In addition to the technical aspects of the tool evaluation, libraries are asked to assess the usability of each tool in their respective environment. Usability aspects include on one hand the time and effort it takes to install, configure and integrate a tool into the existing digitization workflow. On the other hand, costs and the general handling of tools are very important aspects of usability in the library context. The libraries are asked to describe usability aspects in plain text.

4.1.2. Implementation Plan

The usability evaluation is performed in two steps:

1. After installation and configuration

For a tool to be used in a production environment it is usually installed, configured and integrated into the existing digitization workflow. To have an overview on the time, effort and skills needed for this process, libraries are asked to describe the process and the staff that performed each step (question 1).

2. After the technical evaluation

After evaluating the technical aspects of the tools, libraries are asked to give a statement about the usability and expected costs of using the tool (questions 2-4).

4.1.3. Evaluation Form

The following form summarizes the criteria used to validate the usability of tools. They are based on the measures of the productivity and cost effectiveness obtained with the application of the tool. The form has to be filled out by each library for each tool they are evaluating.

1. To use a digitization tool in a production environment, it has to be installed, configured and integrated into the existing digitization workflow. For each of these steps, please describe

- the step-by-step process for installation, configuration and integration including any problems you encountered
- the staff you used for performing each step (e.g. technical, developer, domain expert, librarian, …)
- the effort (in PM) it took to perform each step

1.1 Installation

Process description	
Skills/staff needed	
Effort per type of staff	
Effort Succeed staff	(completed by Succeed partner)

1.2 Configuration

Process description	
Skills/staff needed	
Effort per type of staff	
Effort Succeed staff	(completed by Succeed partner)

1.3 Integration

Process description	
Skills/staff needed	
Effort per type of staff	
Effort Succeed staff	(completed by Succeed partner)

1.4 Please indicate for which of the above mentioned steps input from the tool provider was requested.

(completed by Succeed partner)

2. Please provide a calculation of the costs for licensing and support contracts for the tool for the evaluation period.

3. Please provide a calculation of the costs for licensing and/or support for the tool if you would continue using the tool in a production environment.

4. Please give a statement about the usability of the tool (including positive and negative aspects) from a library point of view.

5. Please state which improvements would be necessary for you to implement the tool in a productive environment.

6. Please give a statement about the usability of the documentation. Please describe shortcomings if applicable.

4.2. Image Processing and Enhancement Evaluation

4.2.1. Introduction

Improving the quality of scanned images can serve two different purposes: It is either done to enhance the visual appearance of images when viewed by humans, or it is done to enhance the quality for post-processing steps such as OCR and layout analysis. Depending on the use case, different tools or settings have to be applied to optimize the image processing result for a particular purpose or material.

Common software tools used to enhance the visual appearance of images are tools for deskewing, contrast enhancement or border adjustment. The overall goal is to transform scanned images in a way that results in sharp and readable text, clear images and a white background. Additionally, borders and page sizes are usually adjusted to be the same size for every page to improve the viewing experience for a set of pages. These requirements can result in different processing parameters applied to different regions of an image in order to have letters rendered with very high contrast compared to images or photos which require much less contrast.

Image enhancement for post-processing purposes usually involves tools for deskewing, noise removal and binarisation. However the parameters used for such tools depend very much on the intended use case. For example if the goal is to improve OCR results by applying image enhancement tools, the optimal parameters might vary for different OCR engines. Additionally, parameters might have to be adjusted for different data sets or even individual pages within a given data set.

Due to the different possible use cases there are two possible evaluation processes. Libraries have to select the appropriate evaluation process depending on their intended use case: Either for improving the visual appearance of images or to improve post-processing results.

- Image Enhancement to Improve Post-Processing Results

The most important quality criteria are those implied by the following post-processing steps. In most use cases this will be the quality of OCR results. Consequently, the evaluation process requires a comparison of post-processing results with and without applying image enhancement steps. For OCR, the word or character error rate should decrease after integrating image enhancement tools.

- Image Enhancement to Improve Visual Appearance

In opposition to image enhancement for post-processing purposes, it is very difficult to find consistent and measurable quality criteria for the visual appearance of images. An evaluation can only be performed by selecting random samples and comparing the processed image with the expected outcome.

4.2.2. Implementation Plan

The preferred evaluation process for image enhancement tools can be described in four consecutive steps. It can be adapted depending on the specific requirements and use cases of a library.

1. Selection of Evaluation Dataset

The evaluation performed in Succeed can only cover a limited data set. Therefore, it is very important to choose a set that is representative in some way to the majority of the material of a library.

2. Description of the Status Quo

After selecting the dataset, the library should describe the problems and challenges they are currently facing with this material and what exactly they want to improve. In particular it must be made clear if the purpose of image enhancement is to improve the visual appearance or to improve post-processing results. Ideally, this description includes sample images with a detailed description of the identified challenges.

3. Evaluation for Improving Post-Processing Results
   a. Development of Ground Truth

      To compare post-processing results before and after image enhancement was integrated, it is necessary to have some form of ground truth material for selected pages of the evaluation data set. This allows comparing automatic processing results with a manually created result that is 100% correct.

   b. Evaluation against Ground Truth

      The digitisation.eu platform offers tools to compare automatic processing results with ground truth and determine quality measures such as the word error rate (e.g. to improve OCR). Libraries should make use of these tools to check whether the image enhancement tool they integrated improved the post-processing results. This evaluation should be supervised by Succeed WP3 participating members.

c. Publication of Evaluation Dataset (optional)

The evaluation dataset consisting of ground truth, images and post-processing results should be published if possible. In this way the measurements are reproducible and new methods can be benchmarked and compared to existing approaches.

4. Evaluation for Improving visual appearance

a. Preparation of Example Images

Since and automatic comparison of visually enhanced images is not possible, libraries should provide a form of "visual ground truth" to clarify what their expected results should look like. This could be accomplished by using a manual image editing software such as Gimp or Photoshop to prepare a set of sample images.

b. Comparison of Results

The automatic processing results should then be compared to the manually created images to identify shortcomings of the tools that were used. For each of the manually created images the libraries should describe where the algorithms failed to meet their expectations and were they performed satisfactory.

4.2.3. Evaluation Form

The following form summarizes the criteria used to validate image processing tools. They are based on the technical validity and scientific performance measures. The form has to be filled out by each library for each image processing tool they are evaluating.

1. Describe the target collection for which image processing has been evaluated (specify period, language, text type, font type (Gothic/Roman/…).

2. Describe the image processing tool you have been evaluating

Tool	Version	Subversion/build number	Customizations

3. Describe the input and output format, for instance: JPEG, TIFF, PNG, …

Input	Output

4. Describe the subset of your collection you use for evaluation, and the way in which you arrived at the selection (This might be a random selection consisting of a certain amount of pages, or a balanced selection by text type, date or other parameters)

5. Specify your use case scenario. Did you evaluate this image processing tool to…

(A) … enhance the visual appearance of images for display?	
(B) … enhance image quality to improve OCR results?	
(C) … other (please describe)	

6. Evaluation results

- If you chose (A) or (C):

  6.1 Please describe the areas where you want to improve the visual appearance of your material (noise removal, deskewing etc.)

  6.2 For each relevant subset of your evaluation set (for instance by selected date, title, font type), please provide visual examples of the original image, a custom prepared version of how you expected the image to look like (e.g. use Gimp or Photoshop to prepare this image), the image processing result and your assessment/opinion of the results:

Set	Image Samples		Assessment
		Original Image(s)	
		Envisioned Image(s)	
		Result Image(s)	

- If you chose (B):

  6.3 Describe the OCR engines you are using

Engine	Version	Subversion/build number	Customizations

6.4 Did your evaluation use ground truth? yes/no

6.5 If you answer to the previous was "yes", please describe

- Format of ground truth (Page XML; Alto, plain text)
- Size of the ground truth collection in pages

Format	Number of Pages

6.6 In which way did you perform the evaluation?

☐By comparing OCR with ground truth and measuring character error rate

☐By comparing OCR with ground truth and measuring word error rate

☐By counting errors in the output

☐By comparing OCR from different engines

6.7 Which evaluation tools have you been using?

Tool	version

6.8 For each relevant subset of your evaluation set (for instance by selected date, title, font type), describe:

1. Amount of pages
2. Amount of words
3. Estimated word error rate without/with the use of the enhancement tool
4. Estimated character error rate without/with the use of the enhancement tool

			Without Image Processing		With Image Processing	
Set	Pages	Words	Word Error Rate	Character Error Rate	Word Error Rate	Character Error Rate

7. Will you allow publication of your evaluation dataset (images)? (yes/no)

If so, what type of license do you envisage?

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

4.3. Layout Analysis Evaluation

4.3.1. Introduction

Layout analysis is about automatically identifying structural elements in scanned documents by detecting the logical units that make up the layout of a page such as articles, headings, captions images or tables. With this information at hand it is possible to generate a better user experience when working with digitized documents. Newspapers for example can contain several articles on a single page that might not be thematically related (e.g. on the title page of a newspaper). By identifying individual articles it is possible to cluster these articles by topic, propose related article to the user and integrate these articles into existing content management systems. Additionally, article segmentation can support search by allowing users to search only within certain layout units such as image captions or headlines.

Compared to other automatic processing steps such as OCR, layout analysis is a rather difficult task to perform automatically. That is due to the fact that e.g. newspaper layouts can be quite complex or they may change over time. Therefore many layout analysis tools include a manual post-correction step in which they support users in the correction process. As a result the evaluation can be based on either the comparison of the automatic results with previously created ground truth material or the time and effort it takes to manually correct the automatic results.

4.3.2. Implementation Plan

The preferred evaluation process for layout analysis tools can be described in four consecutive steps. It can be adapted depending on the specific requirements and use cases of a library.

1. Selection of Evaluation Dataset

The evaluation performed in Succeed can only cover a limited data set. Therefore it is very important to choose a set that is representative in some way to the majority of the material of a library.

2. Development of Ground Truth

To evaluate the results of the automatic layout analysis processing, it is necessary to have some form of ground truth material for selected pages of the evaluation data set. This allows comparing automatic processing results with a manually created result that is 100% correct.

3. Evaluation against Ground Truth

The digitisation.eu platform offers tools to compare layout analysis results with ground truth and determine quality measures. This evaluation should be supervised by Succeed WP3 members.

4. Publication of Evaluation Dataset (optional)

The evaluation dataset consisting of ground truth, images and post-processing results should be published if possible. In this way the measurements are reproducible and new methods can be benchmarked and compared to existing approaches.

4.3.3. Evaluation Form

The following form summarizes the criteria used to validate layout analysis tools. They are based on the technical validity and scientific performance measures. The form has to be filled out by each library for each layout analysis tool they are evaluating.

1. Describe the target collection for which text recognition has been evaluated (specify period, language, text type, font type (Gothic/Roman/…).

2. Describe the layout analysis tools you have been evaluating

Tool	Version	Subversion/build number	Customizations

3. Describe the output format, for instance: Proprietary XML, PAGE, ALTO, …

4. Describe the subset of your collection you use for evaluation, and the way in which you arrived at the selection (This might be a random selection consisting of a certain amount of pages, or a balanced selection by text type, date or other parameters)

5. Did your evaluation use ground truth? (yes/no)

6. If you answer to the previous was “yes”, please describe

- Format of ground truth (Page XML; Alto, plain text)
- Size of the ground truth collection in pages

Format	Number of pages

7. In which way did you perform the evaluation?

☐By correcting the results and measuring time and effort

☐By comparing the results from different layout analysis tools

8. Which evaluation tools have you been using?

Tool	version

9. Will you allow publication of your evaluation dataset (images, OCR and ground truth)? (yes/no)

If so, what type of license do you envisage?

10. For each relevant subset of your evaluation set (for instance by selected by date, title, font type), describe:

1. Amount of pages
2. Number of layout elements on page
3. Estimated error rate
4. Estimated time for correction per page

Set	Pages	Error Rate	Time for Correction

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

4.4. Text recognition (OCR) Evaluation

4.4.1. Introduction

OCR is the process extracting machine-readable text from document images. Evaluation of OCR can be done in a number of different ways:

1. The preferred way is to develop ground truth transcriptions for an evaluation dataset, and evaluate by measuring character error rate and/or word error rate, using an OCR evaluation tool (available in the Impact Centre of Competence evaluation platform, http://www.digitisation.eu/tools/demonstrator-platform/).
2. Alternatively, errors may be counted on a number of output pages. This does not require development of ground truth, but the complete process has to be repeated if more than one (run of an) OCR engine is to be evaluated.
3. When ground truth is not available, comparison (using a text comparison tool) of the output of one or more engines may yield useful insights.

OCR evaluation, and indeed development of innovative approaches to OCR itself, benefits enormously from publication of well-structured evaluation sets. This is why we strongly encourage libraries to publish their evaluation data.

4.4.2. Implementation Plan

The preferred procedure from the point of view of Succeed is as follows:

1. Selection of an evaluation dataset

Needless to say, the evaluation dataset should be representative for the collection you are interested in.

2. Development of ground truth for the evaluation dataset

Ground truth can be developed in a number of ways: either as plain text, or (preferably if required resources are available) as XML containing location coordinates of at least text regions on the page. A tool for producing (PAGE) XML ground truth is the Aletheia tool, which is on the tools shortlist of Succeed.

3. Evaluation in the digitisation.eu evaluation platform, supervised by Succeed wp3 members

4. Publication of evaluation dataset (optional).

The evaluation dataset, consisting of ground truth, OCR and images, can be published. In this way the measurements are reproducible, and new methods can be benchmarked and compared to existing approaches

4.4.3. Evaluation Form

The following form summarizes the criteria used to validate OCR tools. They are based on the technical validity and scientific performance measures. The form has to be filled out by each library for each OCR tool they are evaluating.

1. Describe the target collection for which text recognition has been evaluated (specify period, language, text type, font type (Gothic/Roman/…).

2. Describe the OCR engines you have been evaluating

(include versions and, if possible, minor versions like SDK build number. List different applications of the same engine, i.e. with different customizations for instance relating to language, dictionary or character set, as separate options)

Engine	Version	Subversion/build number	Customizations

3. Describe the OCR output format, for instance: ALTO / Abbyy XML / plain text / RTF

4. Describe the subset of your collection you use for evaluation, and the way in which you arrived at the selection

(This might be a random selection consisting of a certain amount of pages, or a balanced selection by text type, date or other parameters)

5. Did your evaluation use ground truth? yes/no

6. If your answer to the previous question was “yes”, please describe

- Format of ground truth (Page XML, ALTO, plain text)
- Size of the ground truth collection in pages

Format	
Number of pages	

7. In which way did you perform the evaluation?

☐By comparing OCR with ground truth and measuring character error rate

☐By comparing OCR with ground truth and measuring word error rate

☐By counting errors in the output

☐By comparing OCR from different engines

8. Which evaluation tools have you been using?

Tool	version

9. Evaluation by comparison of ground truth and OCR proceeds by means of alignment of GT and OCR. This is more difficult if the page layout is complex (columns, tables, etc). Does your evaluation dataset contain pages where reading order is not immediately obvious (answer "yes" if pages deviate from standard single-column book format)? Describe the complex layouts if applicable and provide examples.

10. Will you allow publication of your evaluation dataset (images, OCR and ground truth)? (yes/no)

If so, what type of license do you envisage?

11. For each relevant subset of your evaluation set (for instance selected by date, title, font type), describe:

1. Amount of pages
2. Amount of words
3. Estimated word error rate per engine
4. Estimated character error rate per engine

set	pages	words	Tesseract CER	Tesseract WER	Finereader CER	Finereader WER

4.5. Evaluation of OCR Evaluation tools

4.5.1. Introduction

Evaluation of OCR can be done in a number of different ways. From the point of view of reproducibility and objectivity, the preferred way is to *develop ground truth transcriptions* for an evaluation dataset, and evaluate by measuring character error rate and/or word error rate. An essential prerequisite for this is a tool which compares the OCR with the ground truth transcription, and collects statistics about the amount of errors, both in terms of characters (*Character Error Rate, CER*) and words (*Word Error Rate, WER*). Most evaluation tools proceed by *alignment* of ground truth and OCR, allowing to precisely locate the discrepancies between GT and OCR. An alternative evaluation measure which gives a rough indication of word error rate without considering the location of the words is the *bag-of-words* error rate. This may be used when complex layouts render alignment problematic.

OCR evaluation, and indeed development of innovative approaches to OCR itself, benefits enormously from publication of well-structured evaluation sets. This is why we strongly encourage libraries to publish their evaluation data.

4.5.2. Implementation Plan

The preferred procedure from the point of view of SUCCEED is as follows:

1. Selection of an evaluation dataset
2. Development of ground truth for the evaluation dataset
3. Installation of the evaluation tool
4. Evaluation of dataset, using the evaluation tool
5. Evaluation of the evaluation tool itself
6. Publication of evaluation dataset, and evaluation results by the tool (optional).

4.5.3. Evaluation Form

1. Describe the evaluation set for which text recognition has been evaluated (specify amount of pages. period, language, text type, font type (Gothic/Roman/…).

2. Describe the OCR engines you have been evaluating

(include versions and, if possible, minor versions like SDK build number. List different applications of the same engine, i.e. with different customizations for instance relating to language, dictionary or character set, as separate options)

Engine	Version	Subversion/build number	Customizations

3. Describe the ground truth file format, for instance: ALTO / Abbyy XML / PAGE XML / plain text. Specify whether the GT contains coordinates for the locations of text regions / text lines / words / characters on the page.

4. Describe the OCR output format, for instance: ALTO / Abbyy XML / PAGE XML / plain text. Specify whether the OCR contains coordinates for the locations of text regions / text lines / words / characters on the page.

5. Describe the subset of your collection you use for evaluation, and the way in which you arrived at the selection

(This might be a random selection consisting of a certain amount of pages, or a balanced selection by text type, date or other parameters)

6. Which evaluation tool have you been using?

Tool	version

7. Evaluation by comparison of ground truth and OCR proceeds by means of *alignment* of GT and OCR. This is more difficult if the page layout is complex (columns, tables, complex newspaper layouts, etc).

Does your evaluation dataset contain pages where reading order is not immediately obvious (answer "yes" if pages deviate from standard single-column book format)? Describe the complex layouts if applicable, provide examples, and explain how you dealt with the problem.[3]

[3] If the evaluation tool supports bag-of-words error rate, a signification discrepancy between this and the word error rate is an indication of alignment problems caused by complex layout.

8. Does the evaluation tool support the XML file format of your OCR and Ground truth? If so, did you use this option or convert to plain text before running the evaluation?

9. Will you allow publication of your evaluation dataset (images, OCR, ground truth, and evaluation report)? yes/no

If so, what type of license do you envisage?

10. For a subset of the evaluated pages, manually inspect the OCR errors detected by the tool and compare with your own findings, obtained in a different way (possibly using another evaluation tools). Do your criteria for error detection agree with the ones implemented in the tool? Does the tool provide you with satisfactory options to influence the criteria uses (for instance, your GT contains long s, but the OCR uses plain s, which you consider correct, or you are not interested in case and/or accent differences)

11. For each relevant subset of your evaluation set (for instance selected by date, title, font type), describe:

1. Amount of pages
2. Amount of words
3. Estimated word error rate per engine as estimated by the tool
4. (If supported by the tool) Bag-of-words error rate
5. Estimated character error rate per engine as estimated by the tool

set	pages	words	Tesseract CER	Tesseract WER	Finereader CER	Finereader WER

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

4.6. Text processing: Named Entity Recognition and Resolution Evaluation

4.6.1. Introduction

Named Entity Recognition aims at detecting elements from a text and classify these into predefined categories such as persons, organizations, locations. Such information can be used to facilitate search by end users, but can also be used to supplement lexicons for OCR in order to enhance text recognition.

4.6.2. Implementation Plan

Since the algorithms used in NER systems, even in state-of-art techniques, are brittle, the systems have to be finely tuned to a limited domain. This is often possible when the NER system is based on statistical methods, but it means that an extensive set of training material needs to be produced in order to build a model suitable for the domain.

Therefore, the implementation of a NER system typically involves the following steps:

1. **Production of a training/test corpus.** A representative subset of the target domain is selected for the manual of semi-automated production of ground truth data. One part of that material is used to train the system, and another part is used to evaluate the accuracy of the recognition.
2. **Building the model using the training material.** If necessary, the annotated data is converted to the format required for the training function, and the model is produced.
3. **Testing the quality of the model.** Recognition is performed on the test material without annotations. The result is compared with the manually annotated test material. Some systems provide a function for automatically testing the quality. If the quality is below expectation, it might be necessary to extend the training material.
4. **Integrate software in the workflow.** Decide how to deploy the software (e.g. as service or as standalone), and harmonize input and output formats.

4.6.3. Evaluation procedure

The performance of NE recognition is usually defined in terms of *precision*, *recall* and *F1* score[4] (harmonic mean of precision and recall). A typical NE evaluation result table would give results per NE category as well as an overall result, and could look like the following example:

[4] For an explanation, refer to http://en.wikipedia.org/wiki/Precision_and_recall, and http://en.wikipedia.org/wiki/F1_score

	precision	recall	F1
Overall	0.853	0.838	0.845
Location	0.83	0.729	0.776
Organisation	0.516	0.251	0.339
Person	0.867	0.917	0.896

The evaluation platform for Succeed does not currently support NER evaluation. However, we will ensure all NER evaluation in the project is done according to the same criteria. We strongly encourage you to publish the evaluation material.

Hence, the steps for NER evaluation are similar to those of the other tools with measurable performance indicators:

1. **Selection of an evaluation dataset**
   Needless to say, the evaluation dataset should be representative for the collection you are interested in.
2. **Development of gold standard data[5] for the evaluation dataset**
   If you have developed special training data to tune the NER system, this is usually a held-out portion of the training data. Otherwise, tools for rapid development of gold standard data are available, like the NE attestation tool selected for Succeed.
3. **Evaluation, supervised by Succeed wp3 members**
   You can carry out your own evaluation, but please consult WP3 members about the procedure.
4. **Publication of evaluation dataset (optional).**
   The evaluation dataset, consisting NE-tagged text, can be published. In this way the measurements are reproducible, and new methods can be benchmarked and compared to existing approaches.

4.6.4. Evaluation Form

The following form summarizes the criteria used to validate tools for Named entity Recognition. They are based on the technical validity and scientific performance measures. The form has to be filled out by each library for each tool they are evaluating.

1. Data

1.1. Describe the data used in this procedure: specify period, type, etc.

[5] In this document, we refer to correctly transcribed text as "ground truth", and to verified linguistic annotations as "gold standard" data

1.2. What language or languages is/are used in this collection?

1.3 List the different types of NE's you annotate, if possible with a pointer to annotation guidelines you followed

2. Preparation

2.1. Did you produce your own training and test material? If not, continue with question

2.2. What software did you use to create the training/test corpus?

2.3. How large was the training/test corpus?

2.4. How was the training/test corpus selected?

2.5. How many person-hours did it take to produce the training/test corpus?

2.6. How long did it take the software to build the new model?

2.7. Did you test the performance of the recognition? If yes, what was the result?[6]

	precision	recall	F1
overall			
location			
organisation			
person			

[6] The predefined table is intended as an example. Of course, when you annotated different NE categories, the row headings would be different.

3. Usage

3.1. (In case of NERT) Did you use the spelling variation reduction module?

3.2. (In case of NERT) Did you need to change the spelling variation rules?

3.3. Were the input/output options sufficient for your purpose?

4.7. Text Processing: NE Linking (Resolution)

4.7.1. Introduction

Named Entity Linking (resolution) is linking named entities in a text to an authority file, or other description of that entity, for instance a wikipedia article or DBPedia database entry. NE linking may be performed after NE recognition, in which case the input consists of NE tagged text, or the NE linking software does its own recognition, in which case untagged text can be fed as input.

4.7.2. Implementation Plan

1. Production of test corpus (optional). In order to estimate the quality of the linking, a manually annotated test corpus needs to be produced.
2. Testing the quality of the linking using a test corpus (optional). The system annotates the test material and the difference with the manual annotation is calculated.
3. Integrate software. The service may either be part of a document processing workflow or to be integrated in a web application if it the linking is performed on the fly.

Here, as in the case of NE recognition, precision, recall and F1 score are the standard technical evaluation criteria. Hence, the steps for a technical evaluation are the same:

1. Selection of an evaluation dataset
2. Development of gold standard data for the evaluation dataset
   This consists of text, with Named Entities annotated and linked to the reference database. Usually, some XML format is used for this purpose.
3. Evaluation, supervised by Succeed WP3 members
   You can carry out your own evaluation, but please consult WP3 members about the procedure.
4. Publication of evaluation dataset (optional).
   The evaluation dataset, consisting of an NE-linked text corpus, can be published. In this way the measurements are reproducible, and new methods can be benchmarked and compared to existing approaches.

4.7.3. Evaluation Form

1. Data

1.1. Describe the data used in this procedure: specify period, type, etc.

1.2. What language is used in this collection?

2. Preparation

2.1. Did you produce your own test material? If not, continue with question 2.6

2.2. What software did you use to create the test corpus?

2.3. How large was the test corpus (in number of words)?

2.4. How was the test corpus selected?

2.5. How many person-hours did it take to produce the test corpus?

2.6. Did you test the performance of the recognition? If yes, what was the result? Please specify precision / recall / F1 if possible.

precision	recall	F1

2.7. (If NE-tagging is not an integral part of the system) What system for NE-tagging did you use?

4.8. Text Processing: Linguistic annotation tools / tools for manual annotation and verification

4.8.1. Introduction

For many tasks in computational linguistics, it is necessary to produce gold standard data, consisting of text with of manually verified linguistic annotation. This type of resource is needed for the evaluation of NLP software, as well as for the training of statistical approaches to NLP tasks. Special tools have been developed to assist users in this labour intensive work. In the SUCCEED project, we have selected tools for annotation of corpus material with lemma and part of speech and for NER gold standard data production.

4.8.2. Implementation Plan

An implementation plan for the evaluation of an annotation tool has the following steps:

1. **Select a corpus for annotation**
   The selected corpus should be described in terms of size (number of tokens) and type of material (period, language, text type, region of origin)
2. **Determine which additional data and software may support the annotation process; get hold of such resources if possible**
   For instance, annotation with part of speech and lemma may be assisted by a computational lexicon or an automatic tagger lemmatizer; annotation of named entities may be prepared by running a NE tagger. These preparatory steps do not depend on the tool to be evaluated; nevertheless, please record which data and tools were selected for the preparation, and how much time was spent on these tasks if possible
3. **Install the annotation tool, load corpus data and auxiliary data**
   Ease of installation, quality of the provided documentation, and time spent on the installation process should be recorded. If assistance by the tool producer is necessary, please record this. Also record the type of expertise which was necessary for installation and loading of data
4. **Train annotators to work with the software**
   It should be recorded how much time is needed for an annotator to become proficient with the tool, whether the supplied documentation was satisfactory.
5. **Annotate the corpus**
   During the annotation process, keep track of the amount of tokens annotated and the time spent on the task. User experiences should be recorded.

4.8.3. Evaluation Form

1. Data

1.1. Describe your collection in terms of number of documents, period, text type, etc.

1.2. What language or languages is/are used?

1.3. Specify the corpus size (number of tokens)

2. Preparation, auxiliary data and software

2.1. Describe the auxiliary data and software you used for the annotation task

2.2. Specify the type of staff used, and the amount of time spent in preparing the corpus for annotation

3. Training

3.1 How long did it take before annotators were sufficiently acquainted with the interface?

4 . Production

4.1. How many annotators have been working on the project?

4.2. How many tokens were the annotators able to process per hour?

4.3 What was the total amount of tokens processed? In the case of the CoBaLT lexicon building tool, also specify the number of distinct word forms and lemmata in the resulting lexicon

4.4 Please describe the recorded annotator experiences

4.9. Text Processing / NLP Tools / Lexicon as a Web Service

4.9.1. Introduction

This tool is a way to enable deployment of a computational lexicon as a web service. One of the possible uses is enhancement of information retrieval. A query for a word form can be expanded to include a number of closely related other word forms, like inflected forms and historical variant spellings. The web service can also be used to reduce inflected forms and variant spellings to a standard lemma form. The service has a simple REST interface.

4.9.2. Implementation Plan

1. Select a use case for deploying the lexicon content in an application
2. Integrate the lexicon web service in your application
3. Determine a testing plan, either using your application, or independent of your application, or both.
   The service may be tested for
   a) Correctness (are the responses delivered by the web service correct in the sense that they correspond to the actual lexicon content)
   b) Performance (is the service responsive; which amount of queries per time unit can it handle?)
   c) Usefulness (Does the web service enhance the application into which it has been integrated?)
      Although it will be difficult to quantify the usefulness of the service, it will be possible to describe the advantages to the user.
   d) Interoperability aspects (Does the web service give you the degree of access to the lexical data you need; is the API satisfactory for your application or would you prefer a different type of access (asynchronous, SOAP, Linked open Data, ….)
4. Carry out test plan, recoding experiences

4.9.3. Evaluation Form

1. Use case

1.1. Describe your use case: what kind of application uses the api (e.g. lemmatizer, presentation software, search engine), and what are the benefits you expect from deploying the lexicon?

1.2. Which lexicon(s) have you used?

1.3. Which functions have you used (get_lemma, get_wordforms, expand)?

2. Tests

2.1. Is the implementation **correct** according to your experiences? If not, record errors

2.2. Is the implementation **responsive** enough according to your experiences? If not, record errors, and describe, as exactly as possible, the test scenario which cause the breakdown or unacceptable delay

2.3. In what way has the end user experience been enhanced?

3. Interoperability aspects

3.1. Does the service give you access to the data you need, or does it hide useful information from you? Please describe missing functionality if applicable

3.2 Does the API satisfy you, or would you prefer a different type of access? Please explain your preferred scenario if applicable.

4.10. Miscellaneous tools: JHOVE2

4.10.1. Introduction

JHOVE2 is open source software for format-aware characterization of digital objects. Characterization can be thought about in two ways. First, it is information about a digital object that describes that object's character or significant nature and that can function as a surrogate for the object itself for purposes of much preservation analysis and decision making. Second, characterization is the process of deriving this information. This process has four important aspects: identification, validation, feature extraction, and assessment. JHOVE2 analyzes digital objects with these questions:

- What is it? (Identification)
- What about it? (Feature extraction)
- What is it, really? (Validation)
- So what? (Assessment)

The JHOVE2 project generalizes the concept of format characterization to include identification, validation, feature extraction, and policy-based assessment. The target of this characterization is not a simple digital file, but a (potentially) complex digital object that may be instantiated in multiple files.[7]

4.10.2. Implementation Plan

The preferred evaluation process for JHOVE2 can be described in four consecutive steps. It can be adapted depending on the specific requirements and use cases of a library.

1. **Selection of Evaluation Dataset**
   The evaluation performed in Succeed can only cover a limited data set. Therefore it is very important to choose a set that is representative in some way to the majority of the material of a library.

2. **Evaluation**
   Before your start evaluating JHOVE2 you should formulate your use case, expectations and your input data. The result of the evaluation should then be to what extend the tool met your expectations and how it helped to achieve your use case.

3. **Publication of Evaluation Dataset (optional)**
   The evaluation dataset consisting of images and processing results should be published if possible. In this way the measurements are reproducible and new methods can be benchmarked and compared to existing approaches.

[7] https://bitbucket.org/jhove2/main/wiki/Home

4.10.3. Evaluation Form

The following form summarizes the criteria used to validate JHOVE2. They are based on the technical validity and scientific performance measures.

1. Describe the target collection for which the tool has been evaluated (specify period, language, text type, …)

2. Describe the subset of your collection you use for evaluation, and the way in which you arrived at the selection (This might be a random selection consisting of a certain amount of pages, or a balanced selection by text type, date or other parameters)

3. Will you allow publication of your evaluation dataset (images, OCR and ground truth)? (yes/no)

If so, what type of license do you envisage?

4. Which of the following JHOVE2 functionalities did you evaluate?

4.1 ☐ Format Identification

4.1.1 Describe your use case: what do you need the format identification for?

4.1.2 Describe the input format: which files did you use and why?

4.1.3 How satisfied were you with the results? Did the tool support your use case?

4.2 ☐ Feature Extraction

4.2.1 Describe your use case: which features or properties did you want to extract?

4.2.2 Describe the input format, for instance: TIFF, JPEG2000, …

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

4.2.3 How satisfied were you with the results? Did the tool support your use case?

4.3 ☐ Validation

4.3.1 Describe your use case: what did you want to validate?

4.3.2 Describe the input format, for instance: TIFF, JPEG2000, …

4.3.3 How satisfied are you with the results? Did the tool support your use case?

4.4 ☐ Assessment

4.3.1 Describe your use case: what was the purpose of your assessment?

4.3.2 Describe the input format, for instance: TIFF, JPEG2000, …

4.3.3 How satisfied are you with the results? Did the tool support your use case?

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