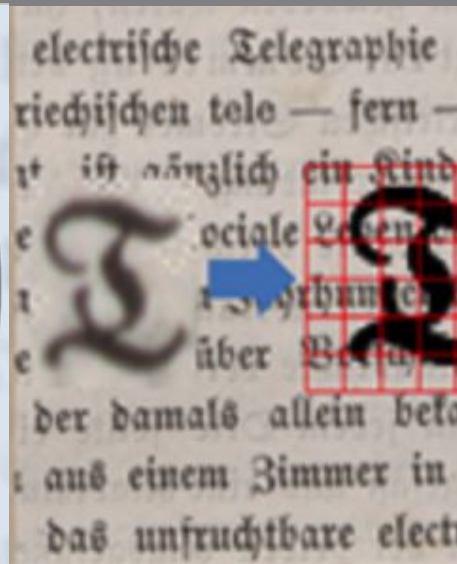


OCR-D Technische Systemarchitektur: Workflows, Repository, Schnittstellen

Ajinkya Prabhune (KIT) und Clemens Neudecker (SBB)

Institute for Data Processing and Electronics (IPE)



Einleitung: OCR-D

Das Projekt OCR-D unterscheidet 6 grundlegende Themenbereiche, die in die folgenden Module unterteilt sind:

- Bildvorverarbeitung
 - Layouterkennung
 - Textoptimierung
 - Modell-Training
 - Qualitätssicherung
 - Langzeitarchivierung und Persistenz
- OCR und OLR
- Daten- und Metadaten Repository
- Jedes dieser Module produziert und verwendet Daten und Metadaten
 - Als Daten betrachten wir sämtliche Dateien (Images, PDF, etc.) die entweder von einem Modul erzeugt oder verwendet werden (vom Bild über die Texterkennung bis hin zur Evaluation)
 - Metadaten sind hier diejenigen Informationen über das Modul und die Eingabe- bzw. Ausgabedateien, die in bestehenden Metadatenstandards kodiert werden können (z.B. METS, textMD, etc.)

Einleitung: OCR-D

- Jedem der Module ist ein eigenständiges Forschungsthema zugeordnet in dem entweder neue Module entwickelt oder bestehende technische Lösungen auf die besonderen Anforderungen hin angepasst werden
- Die aktuelle Situation hinsichtlich der Verbreitung von Werkzeugen für die OCR ist wie folgt:
 - Bestehende Werkzeuge (e.g. Tesseract, FineReader, OCRopus) werden eingesetzt, oder
 - Spezifische Eigenlösungen werden entwickelt und implementiert, oder
 - Eine Kombination beider Ansätze wird genutzt
- Die von jedem der Module erzeugten Daten (Ground Truth, Digitalisate und Metadaten) sollen zur Weiterverarbeitung für andere Module zur Verfügung stehen

Einschränkungen

- Aktuell werden oft an die lokalen Bedingungen in den Einrichtungen und Forschungsgemeinschaften spezifisch angepasste OCR-Prozesse eingesetzt, in denen
 - Ad hoc-Techniken für die Speicherung und Bereitstellung von Daten und Metadaten verwendet werden, sowie
 - Workflows nicht in standardisierten Formaten beschrieben werden (sondern typischerweise nur durch Skripte), und
 - Die verwendeten Werkzeuge, Softwarebibliotheken und Dienste nicht hinreichend dokumentiert (und für andere wiederverwendbar) sind

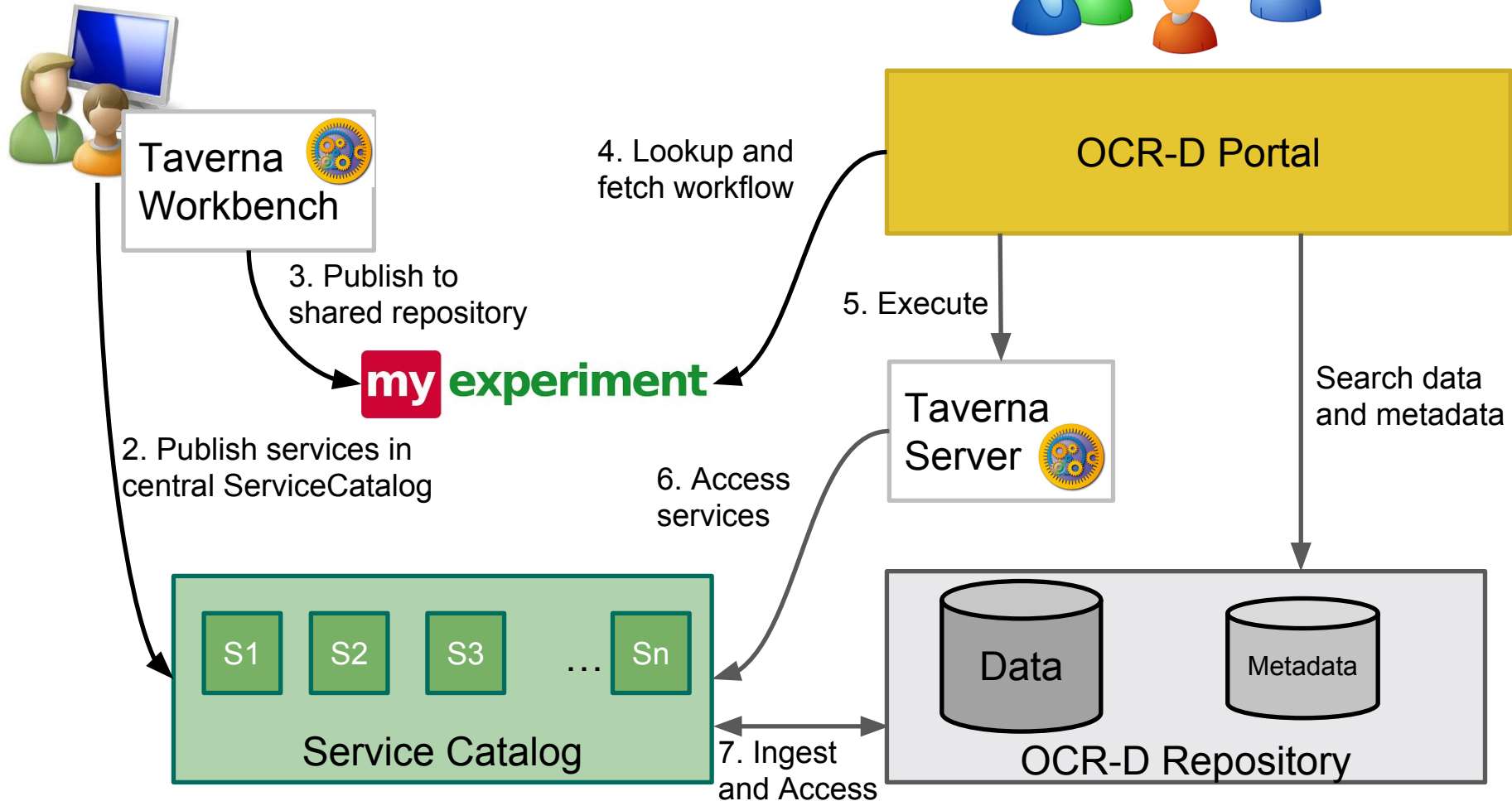
Ziele

- Das Design und die Entwicklung eines umfassenden OCR-D Frameworks sowie eines Repository mit mindestens den folgenden Funktionalitäten:
 - Langzeitarchivierung und Datenbereitstellung
 - Speicherung und Bereitstellung von (Prozess-)Metadaten
 - Service-Registry für die Veröffentlichung von OCR Diensten (Web Services) zur Erstellung von geeigneten Workflows
 - Integration mit einem Workflow Management System für die systematische Implementierung von modularen und reproduzierbaren Workflows (in-silico experimentation)
 - Web-Portal für die Bereitstellung von Daten und Metadaten sowie das Ausführen von Workflows

OCR-D Systemarchitektur

1. Create/deploy/test an OCR-D workflow locally

OCR-D community





-

Warum Workflows?

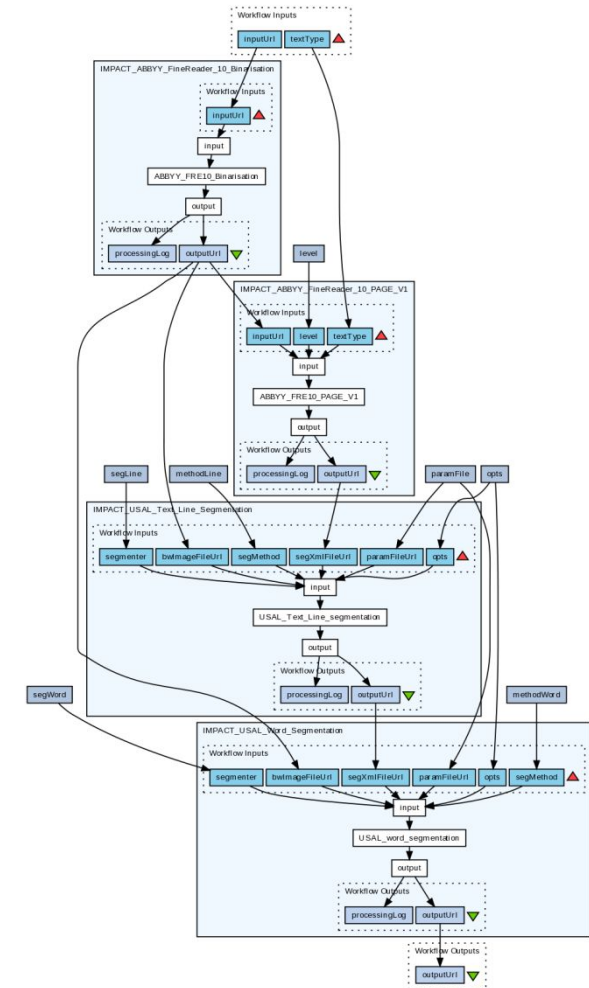
- Standardisierte Schnittstellen erlauben die flexible Kombination von einzelnen Knoten zu komplexeren Prozessen abhängig von kompatiblen Eingabe- und Ausgabeformaten
- Workflows sind dokumentierte und wiederverwendbare Interaktionen zwischen einzelnen Knoten



```

<Textline ID="P209_S70030" BPOS="70" VPOS="124" WIDTH="117" HEIGHT="20">
  <SP ID="P209_S70030" BPOS="117" VPOS="124" WIDTH="20">
    <string ID="P209_S70030" BPOS="117" VPOS="124" WIDTH="19" HEIGHT="19" CONTENT="aue" WC="0.76" CC="0.077">
      <SP ID="P209_S70030" BPOS="122" VPOS="124" WIDTH="21">
        <string ID="P209_S70030" BPOS="122" VPOS="124" WIDTH="20" HEIGHT="19" CONTENT="uue" WC="0.83" CC="0.067">
          <SP ID="P209_S70030" BPOS="131" VPOS="124" WIDTH="27">
            <string ID="P209_S70030" BPOS="131" VPOS="124" WIDTH="26" HEIGHT="19" CONTENT="perpetua" WC="0.77" CC="0.040507">
              <SP ID="P209_S70030" BPOS="137" VPOS="124" WIDTH="22">
                <string ID="P209_S70030" BPOS="137" VPOS="124" WIDTH="21" HEIGHT="19" CONTENT="ton" WC="0.77" CC="0.040507">
              </SP>
            </SP>
          </SP>
        </SP>
      </SP>
    </SP>
  </Textline>

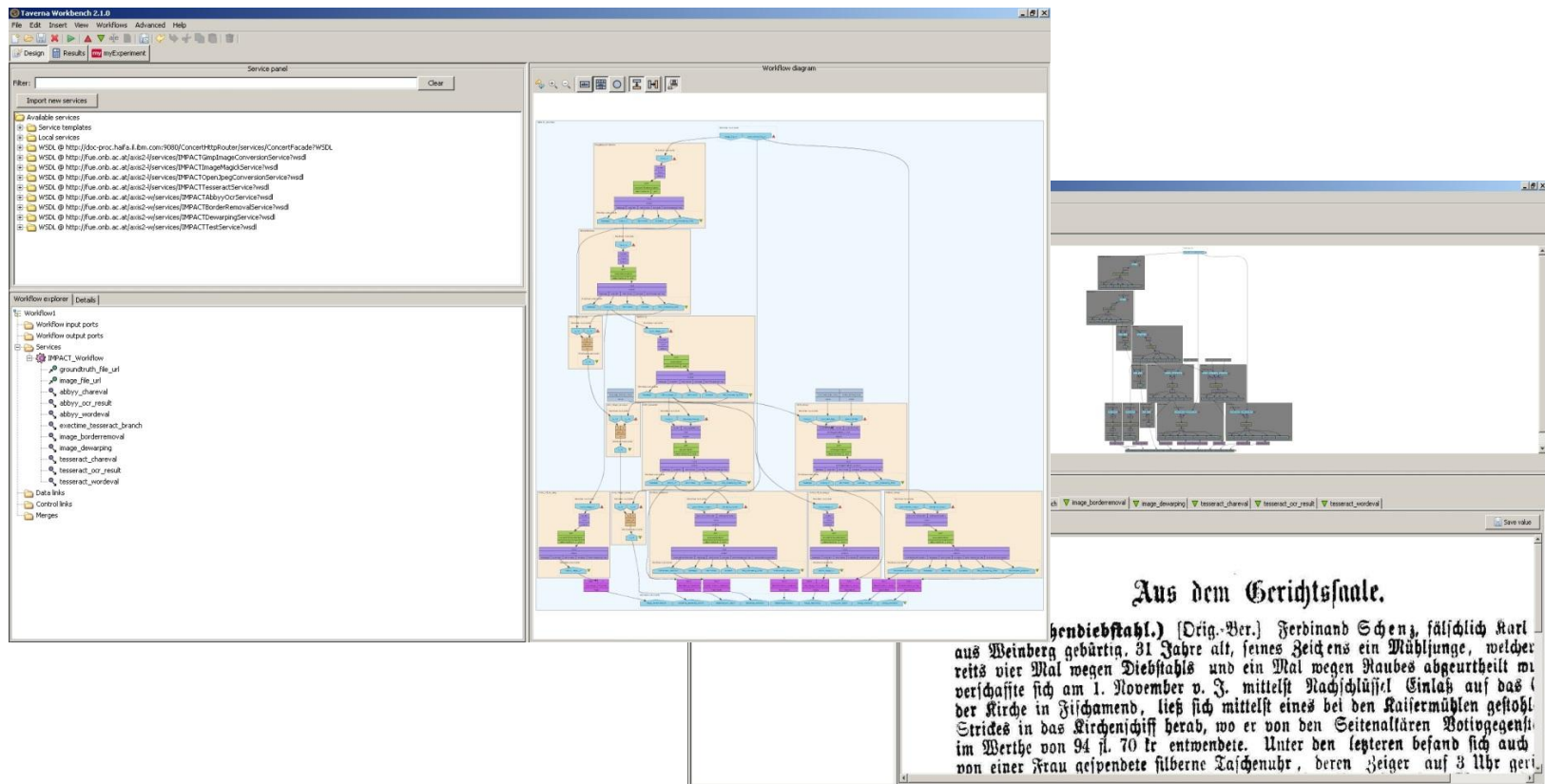
```





Workflowausführung (Lokal)

- *Taverna Workbench* ist ein Open Source Werkzeug mit einer GUI für das Design und die lokale Ausführung von Workflows



Workflowausführung (Remote)

- *Taverna Server* ist ein Open Source Werkzeug für das Ausführen von Workflows über das Web
- Es existieren Werkzeuge um aus einer Workflowbeschreibung dynamisch ein entsprechendes Webformular für *Taverna Server* zu erzeugen (und damit z.B. die Integration mit einer Webseite)

Please upload your workflow file:

Choose File
No file chosen

☒ Show input values, if available

Show input fields

Or login to MyExperiment and choose a workflow:

User:
Password:
Login

Group: IMPACT Centre of Competence
Workflow: Complex Abbyy FRE 9 OCR and Evaluati Details
Complex Abbyy FRE 9 OCR and Evaluation with Border Removal and Page Curl
Complex Experimental IMPACT Workflow
Complex Full Evaluation
Complex OCR Evaluation FRE9 FRE10 Tesseract2 (French) + Repository
Complex OCR Evaluation FRE9 vs CONCERT
Complex OCR Evaluation FRE9 vs FRE10 with IMPACT image enhancement
Complex Segmentation Combination
Helper Convert n PAGE-XML into TXT files
Helper Count characters in PAGE XML batch
Helper CSV to List Conversion
Helper Retain PRImA ID
Helper Timer
Helper URL to List Conversion
IMPACT ABBYY FineReader 10 Binarisation
IMPACT ABBYY FineReader 10 OCR
IMPACT ABBYY FineReader 9 OCR
IMPACT ABBYY FineReader 9 PAGE V3
IMPACT ALTO to Text Transformation
IMPACT FineReader
IMPACT Gimp PNG to TIF Conversion

☒ Show inp

Show inp

Operation: ocrByUrl
Apply OCR to input image file

dat (.dat file for external dictionary)
german.dat

input (URL reference to input image file)
http://kdemo.dnsalias.on view

format (Output format)
PAGEXML

script (Script of the document)
Gothic

languageItem

Bulgarian
Catalan
Croatian
Czech
Danish
Dutch
English
Spanish
Swedish
Turkish
Ukrainian

Show Results

Workflow Registry

- myExperiment* ist eine Web-Plattform für das Auffinden, Teilen und Wiederverwenden von Workflows sowie zum Austausch mit Experten



Taverna 2 **IMPACT IBM Concert UploadPagesUrl BasicWorkflow (v1)** [View](#)
[Download \(v1\)](#)
Created: 25/02/10 @ 09:30:10 | Last updated: 25/02/10 @ 09:36:12
Credits: Sven
License: Creative Commons Attribution-Share Alike 3.0 Unported License

Upload pages to the IBM Concert tool. It is required that you login to Concert and get a session key before applying this basic workflow (see indicate if the Alto XML result files should be returned as word, line, or character level. The Concert Web tool is available at <http://ibc-proc.haifa.ilb.com:5050/ConcertWebTool/>. Under construction, the upload does still not work correctly!

Ratings: 0.0 / 5 (0 ratings) | Versions: 1 | Reviews: 0 | Comments: 0 | Citations: 0
Viewed: 0 times | Downloaded: 0 times
This Workflow has no tags!

Taverna 2 **Helper Transform XML using XSLT Basic Workflow (v1)** [View](#)
[Download \(v1\)](#)
Created: 03/03/10 @ 14:11:23 | Last updated: 03/03/10 @ 14:30:14
Credits: Sven
License: Creative Commons Attribution-Share Alike 3.0 Unported License

The input XML document is converted using the input XSLT document to an output document which can be another XML file, HTML, or plain text, depending on the XSLT used in this workflow.

Ratings: 0.0 / 5 (0 ratings) | Versions: 1 | Reviews: 0 | Comments: 0 | Citations: 0
Viewed: 0 times | Downloaded: 0 times
This Workflow has no tags!

Taverna 2 **Abbyy OCR+Evaluation (v1)** [View](#)
[Download \(v1\)](#)
Created: 25/02/10 @ 11:01:15
License: Creative Commons Attribution-No Derivative Works 3.0 Unported License

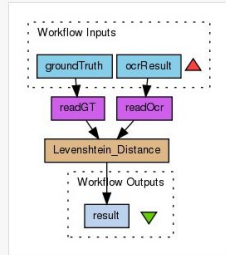
This workflow takes an image (URL) and a Ground truth file (URL) as input, processes the image in ABBYY FR and evaluates it against the Ground truth file.

Ratings: 0.0 / 5 (0 ratings) | Versions: 1 | Reviews: 0 | Comments: 0 | Citations: 0
Viewed: 0 times | Downloaded: 0 times

IMPACT OCR Evaluation by Levenshtein distance
Created: 2011-05-29 11:29:59 | Last updated: 2012-06-06 10:30:55
[Download Workflow](#) [Open in OnlineHPC](#)

The workflow calculates the Levenshtein distance of two input texts using the apache commons library and its `org.apache.commons.lang.StringUtils.getLevenshteinDistance(text1, text2)` method. In order to be able to use this workflow, the apache commons Java library (commons-lang-2.4.jar) needs to be made available to the Taverna Workbench by dropping it into {TAVERNA_HOME}\lib.

Preview



Download as scalable diagram (SVG)

Run


Run this Workflow in the Taverna Workbench...


Option 1:
Copy and paste this link into File > Open workflow location...
<http://www.myexperiment.org/workflows/2192/download?version=2>
[\[More Info \]](#)

Run this Workflow on the cloud with OnlineHPC...


Click the link below to visit OnlineHPC
<http://onlinehpc.com/workflows/editor?provider=myexperiment&workflowid=2192>
[\[More Info \]](#)

Workflow Type
Taverna 2


Uploader

IMPACT


License
All versions of this Workflow are licensed under:


Version 2 (latest) (of 2)
View version: 2 (latest)

Credits (1)
(People/Groups)
 IMPACT

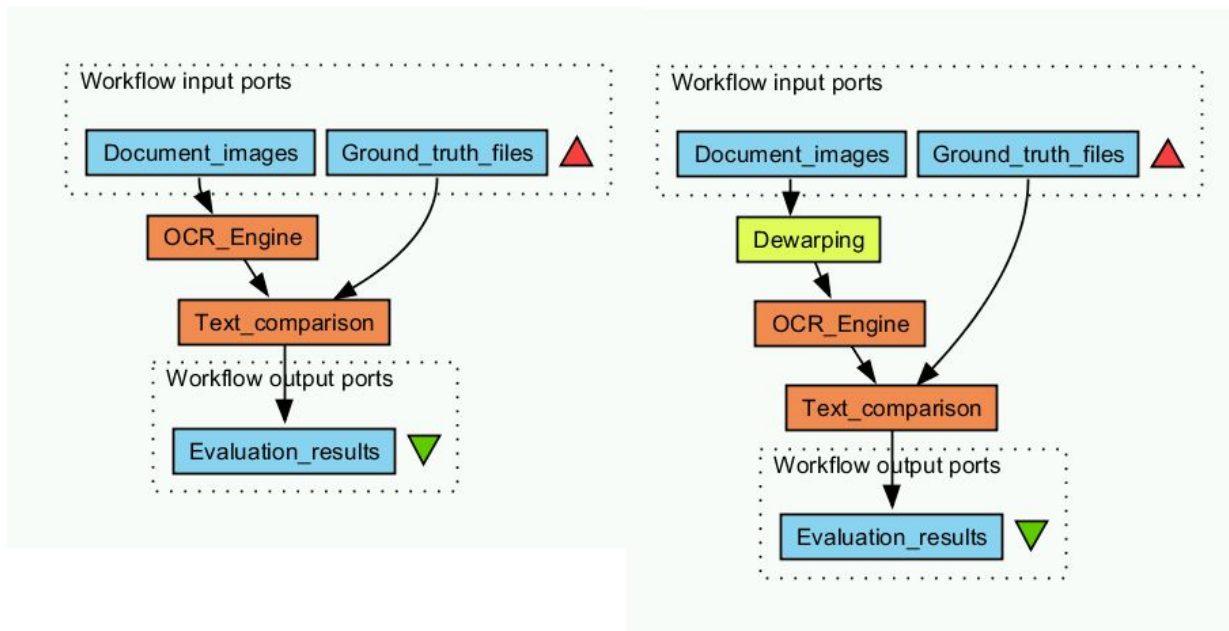
Attributions (0)
(Workflows/Files)
None

Tags (5)
 Uploader tags
evaluation | impact | levenshtein | OCR | text
Add Tags [\[+ \]](#)

Shared with Groups (1)
 IMPACT Centre of Competence

Vorteile

- Modulare, flexible, und dokumentierte Workflows können auf einfache Weise von anderen weiterverwendet werden
- Transparenz und Reproduzierbarkeit der Ergebnisse
- Sehr gut geeignet um den Einfluss einzelner Komponenten auf das Ergebnis eines gesamten Workflow evaluieren zu können



OCR-D Repository Services

Langzeitarchivierung und Datenbereitstellung

- *Data Storage and Access Service*: Stellt einfache Funktionen für den Dateneingest und den Download von Daten zur Verfügung. Daten können sowohl Eingabe- und Ausgabe eines Workflows oder auch Zwischenergebnisse einzelner (Sub-)Module sein. Zwei Varianten sind erforderlich:
 - Command-line client: Für Module, die lokal ausgeführt werden, wie z.B. die Erstellung eines Trainingsmodells, steht eine Kommandozeilenapplikation bereit um die Trainingsdaten und das Modell in das Repository hochzuladen
 - REST client: Ein REST-basierter Client der in einem Workflow zur automatisierten Speicherung in und dem Abruf von Daten aus dem Repository integriert werden kann

OCR-D Repository Services

MetadatenSpeicherung und Bereitstellung

- *Metadata Storage and Query Services*: Die OCR-D (Sub-)Module erzeugen bei ihrer Ausführung Metadaten
 - Metadaten zu OCR und OLR-Prozessen werden in PAGE XML modelliert und abgelegt während z.B. bibliographische Angaben, Strukturdaten usw. in ein OCR-D METS-Profil eingebettet werden
 - Zusätzliche Metadatenformate wie z.B. TEI und textMD werden auf ihre Relevanz für das OCR-D-Projekt untersucht
 - Um die Suche in Metadaten zu ermöglichen, werden diese für die Volltextsuche indiziert und entsprechende Suchmöglichkeiten implementiert

OCR-D Repository Services

Bereitstellung von Services für deren Verwendung in Workflows

- *Services Registry and Catalog*: Für die Auffindbarkeit und Wiederverwendung von OCR-D Services ist ein gemeinsames Verzeichnis (Service Registry) aller Services erforderlich
 - Sobald (Sub-)Module funktionierende Werkzeuge implementieren müssen diese mit ihren REST-Schnittstellen im ServiceCataloger registriert werden
 - Der ServiceCataloger stellt ein Verzeichnis für REST-basierte Services bereit und ist mit Taverna kompatibel
- *Auxiliary Services*: Services, die nicht direkt zu einem Modul gehören aber kleinere, hilfreiche Aufgaben implementieren
 - Von mehreren Modulen benötigte Funktionen und/oder externe Werkzeuge sollten nur einmal als gemeinsame und wiederverwendbare Services mit REST-Schnittstellen implementiert werden
 - Beispiele dafür sind Services für die Bildkonvertierung (beruhend auf ImageMagick/Graphicsmagick) oder XML-Transformation (mittels XSLT)

OCR-D Repository Services

- **Workflow-Integration:** Systematische Beschreibung, Wiederverwendung und Automatisierung von OCR-Prozessen
 - Workflows ermöglichen die Reproduzierbarkeit von Ergebnissen inkl. der automatischen Erfassung aller Provenienzinformationen
 - Eine Analyse der generierten Provenienzinformationen ermöglicht die gezielte Optimierung von Workflows und Ergebnissen
- **Workflow Registry:** Ein Repository zur Bereitstellung von Workflows die von den an OCR-D beteiligten Institutionen und externen Forschungsgruppen geteilt und wiederverwendet werden können
 - **myexperiment.org** ist ein weitverbreitetes Portal für die Online-Bereitstellung von Workflows in standardisierten Beschreibungssprachen wie bspw. Taverna Workflows in SCUFL

Modul-Schnittstellen

- Für jedes Modul existieren zwei Abstraktionsstufen:
 - Abstrakt: Beschreibung einer REST-Schnittstelle für das gesamte Modul
 - Detailliert: Abhängig von der Granularität eines Moduls kann dieses weitere Sub-Module enthalten die wieder über eigene REST-Schnittstellen verfügen
- REST-Schnittstelle für das Preprocessing Modul

REST-Interface: `DocumentImagePreProcessing`

HTTP Method: `POST`

Input Data: `[Mandatory]` Image file (with accepted formats) + `[Optional]` Configuration file

Input Parameter: `List of module-specific parameters serialized in XML`

Output Data: `Preprocessed Image (output data format)`

Output Metadata/Log: `A minimum level of log information + metadata describing the data, which is modeled in appropriate metadata standard`

Response Code: `Based on the exit code generated by the module, the appropriate HTTP response code is propagated`

Technische Gemeinsamkeiten

- Für jedes Modul müssen folgende Anforderungen erfüllt sein:

- Minimales Set an Logging (Provenienz)

Service name: <name of the module/sub-module>

Service version: <module version>

Service description: <textual description of the service>

Service URL: <location where the service is deployed>

Service Implementation: <implementation library name and version>

Input data URL: <URLs of the input data>

Input parameters: <list of input parameters>

Output file name and size: <name of the output file and size>

Exit code: <Enumeration of Exit codes, 0 = OK, any value > 0 corresponds to particular error code>

Service Processing Time: <time in milliseconds>

Service invocation Timestamp: <timestamp>

- Mapping zwischen exit codes und HTTP response codes

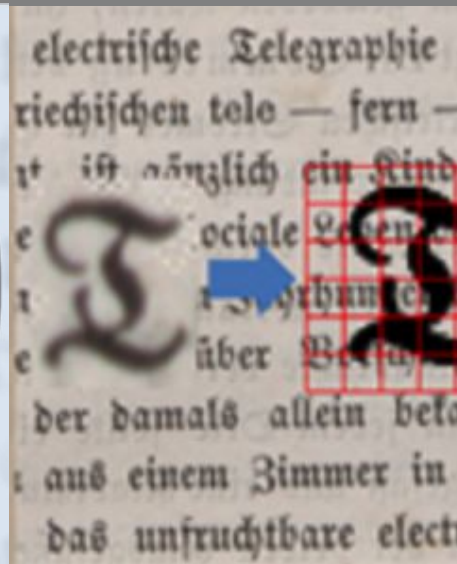
Exit code 0 → HTTP Response 200 Success

Exit code 1 → HTTP Response 500 Error

Danke für die Aufmerksamkeit! Fragen?

Ajinkya Prabhune (KIT) und Clemens Neudecker (SBB)

Institute for Data Processing and Electronics (IPE)



Was fehlt?

- OCR-D METS-Profil: Ein METS-Profil das eine Definition für die systematische Organisation sämtlicher, im OCR-D-Projekt verwendeter Metadatenformate bereitstellt. Relevante Metadatenformate sind z.B. PAGE XML, PREMIS und weitere.
- Einrichtung eines OCR-D Repository für die Langzeitarchivierung und Bereitstellung von Daten und Metadaten sowie für Trainings- und Referenzdatensets
- Ein abstrakter OCR-D Workflow der den Datenfluss zwischen den einzelnen Modulen sowie die Integration mit dem OCR-D Repository exemplarisch beschreibt