Title	CLARIN Switchboard Specification
Version	1.02
Author(s)	Claus Zinn, Marie Hinrichs, Emanuel Dima, Dieter van Uytvanck
Date	2015-12-02
Status	Draft
Distribution	Public
ID	CE-2015-0684

1. Planning

Month /Stage	2015				2016									
	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	July	Aug	Sep	Oct
Pre-dev														
Alpha														
Beta														
RC														
Stable														

* **Pre-development: Sept - Nov 2015**
  - create development plan
* **Alpha stage: Dec 2015 – Jan 2016**
  - define metadata for describing applications
  - ask selected CLARIN centers to provide application metadata for beta testing (Weblicht will provide metadata for WebLicht easy-chains for initial testing)
  - implement *demo-upload* website
  - create an icon for use in resource discovery application (like HelpDesk icon)
* **Beta stage: Feb - April 2016**
  - fully functional Switchboard prototype, including REST API
  - integration of selected CLARIN applications
  - general call to all CLARIN centers to provide metadata and integrate their applications
* **Release Candidate: May - Aug 2016**
  - bug fixing
  - integration of additional applications
  - update documentation
* **Final stable version: Sept - Oct 2016**

- final code release
- final documentation release

Alpha:

- A test application (a simple, Javascript-based *demo-upload* website) will be implemented, where users can upload their language resource, and a corresponding CMDI file describing the resource. Alternatively, they can provide relevant metadata information themselves. Once all information is provided, a switchboard icon appears, which users can invoke to get access to the switchboard. In addition, a textfield to enter a URL with a pointer to the resource will be provided.
  This demo-upload website is primarily supporting the development of the switchboard. But it will also be made available to interested CLARIN members to gather feedback early in the project (and before the other three main access points are adapted to receive their switchboard icon plugin).

- The metadata format for describing applications to the switchboard will be defined (see Application Registry).

- Once an initial metadata format for the applications is specified, it will be circulated among selected CLARIN partners, with a call to register their applications with the switchboard service. Candidates for integration are (non-exhaustive list):

    - @PhilosTEI: http://ticclops.clarin.inl.nl/philostei/
    - Weblicht: http://weblicht.sfs.uni-tuebingen.de
    - LAP: http://lap.hpc.uio.no/
    - TTNWW: http://yago.meertens.knaw.nl/apache/TTNWW/
    - Some more tools from diverse countries:
        - TEI-compliant tools
        - OCR tools
        - Speech recognition tools

- A subset of the easy chains offered through WebLicht will be made available for the switchboard. This involves:
    - Creating workflow metadata for the easy chains already available in WebLicht.
    - Changing WebLicht to handle the new input parameters.
- Create a Switchboard icon that can be used by calling applications – similar to the HelpDesk icon currently used.

Beta:

- Prototypes for Profiler, App Registry, and Matcher will be ready for testing. The GUI as well as the underlying REST-based API is implemented.
- Integration of selected CLARIN applications will be completed
- Switchboard prototype will be made available to all CLARIN members with a general call to integrate applications

Release Candidate:

- All applications / workflow engines that responded to the general call will be integrated with the switchboard.
- Bug fixing and any necessary adjustments to components will be carried out
- Document Switchboard and its components

Stable:

- The switchboard will be user-tested and made available to the CLARIN community.
- Documentation released for publication on the CLARIN webpage

2. Objective

The goal of constructing the Language Resource Switchboard (Task 2.3.1) to help users to connect language resources with applications that can process them. For example to perform linguistic analyses (for text data), OCR (for scanned images), or speech recognition (for sound recordings) on data found in the VLO, FCS, or VCR.
The switchboard's main purpose is to match resources with applications that can process them to perform some kind of analysis or data transformation. It is targeted at non-expert users who will be able to invoke the switchboard on CLARIN sites where resources can be found: the CLARIN Virtual Language Observatory, the CLARIN Federated Content Search, and the CLARIN Virtual Collection Registry.

Note that the switchboard is a technical vehicle to match language resources (described by CMDI metadata) with applications that are registered to the switchboard via an application metadata description. Note that the switchboard is not a workflow composition engine (like e.g. WebLicht), and also that the switchboard is not storing or maintaining any data in connection with the matching service (analysis results, provenance data *etc.*).

2.1. Use Case Example

Consider the scenario where a linguist identifies a resource, a Dutch text corpus for example, through the CLARIN Virtual Language Observatory or the CLARIN Federated Content Search. Given the resource (consisting of one or more digital objects) the user can now call for the switchboard to identify all applications that can process or analyze the resource in question. For this, the switchboard uses:

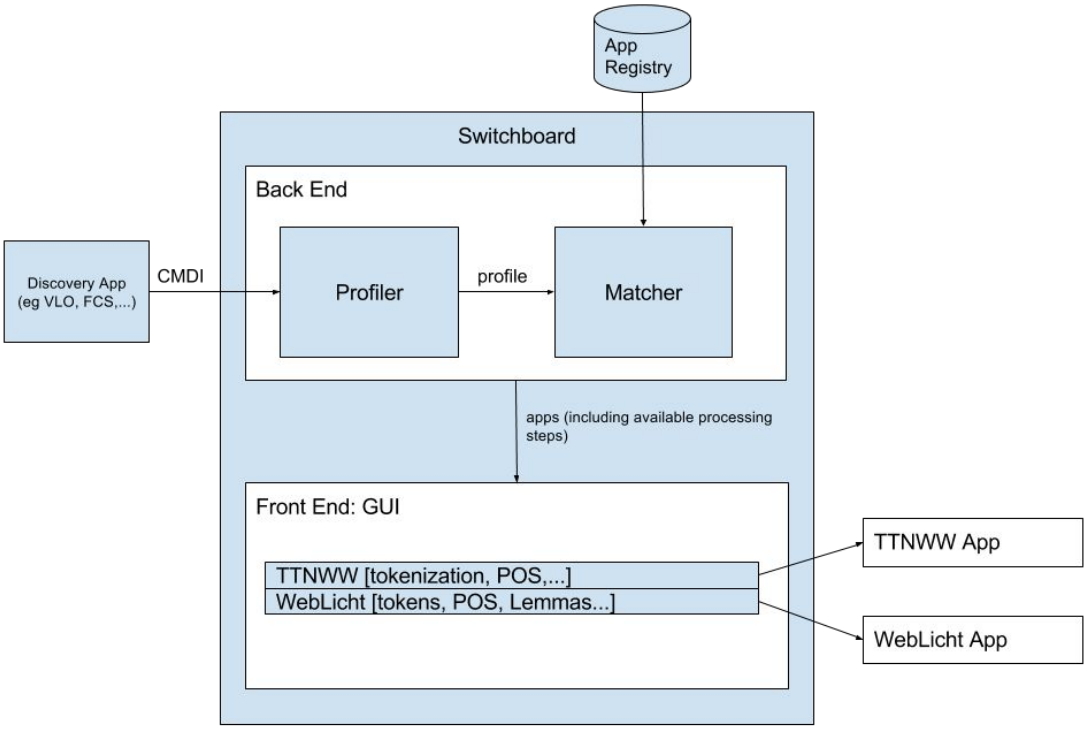
- the CMDI metadata description of the resource, and the resource itself
- the metadata that describes all the applications the switchboard is aware of (which are registered with the switchboard)
- an algorithm that matches resource metadata with the application metadata

The switchboard's *back-end* is complemented by its *front-end,* which

- displays a list of applications that can process the data, along with a list of possible analyses for each application

- forwards the user to the selected application's web page (URL), together with a reference to the resource in question (link to the resource' metadata)

3. Switchboard Architecture



3.1. Resource Profiler: Identification of resource characteristics

The Resource Profiler extracts from the CMDI metadata description of a given resource all information relevant for the identification of applications which are able to process the resource in question. This component has to cope with:

- a wide variety of different CMDI profiles where relevant information might be found at different locations in the metadata (interpretation of the schemas)
- a wide variety of metadata values (interpretation of the value domain of a given data descriptor)
- possibly, metadata information that is erroneous or missing

As a future enhancement, the profiler may need to resort to additional information to identify a resource's mime type (and other relevant information). For this, it may be necessary to implement a *mime type detector mechanism.*

3.2. Application Registry

The switchboard requires a table that lists each application together with all relevant metadata to inform the *Matcher.* This includes:

- the title of the application
- an English description in prose about the application's capabilities
- a controlled vocabulary description of the processing step(s) the application offers, *e.g.,* "tokenization", "part-of-speech-tagging", "optical character recognition")
- a language identification (using ISO 639-3) for which the processing step(s) are available
- input parameters (eg mime type, language, etc)
- the web address (URL) of the application, together with the mechanism to pass on the parameter (URL of the CMDI file for the resource, or URL to the resource in question)
- the invocation method: POST or GET
- Invocation cardinality: can the receiving end deal with a single file (single URL) or with multiple files (JSON with multiple URLs)
- contact person or support email
- screenshot + logo
- access information to the application (access to all, or to CLARIN members with Shibboleth login)

3.3. Matcher: The Resource - Application matching algorithm

The matching algorithm receives the input from the profiler and the Application Registry and returns:

- a list of all applications available for the resource in question
- attached to each application, the analyses that it can provide

The switchboard will offer a task-oriented view and a tool-oriented-view. In the task-oriented view, the user gets a list of tasks (e.g., OCR, speech recognition, part-of-speech tagging, named-entity recognition) (s)he could perform; in the tool-oriented view, the user gets a list of tools that can process the resource in question.
The matching algorithm might rank the applications based on usage frequency (stats collected on server or by Piwik).

3.4. UI : The User Interface

The switchboard offers a REST-based interface as well as a GUI-based interface. The REST-based interface receives as input parameters either:

- the CMDI file of the resource in question, and the resource itself (if necessary, for mimetype detection)
- a PID to the metadata of the resource
- a URL to a resource

The REST service returns as output parameter

- JSON file listing all applications and the respective analyses (or data transformations) that can be performed by each one.a

The Graphical User Interface is a standalone webpage and builds upon the REST-based API. The GUI webpage is accessible via a URL, the URL's parameter points to the CMDI description of the resource in question. The GUI is invokable from any webpage listing a linguistic resource, including:

- the CLARIN Virtual Language Observatory,
- the CLARIN Virtual Collection Registry, and optionally,
- the CLARIN Federated Content Search.

Once a resource has been identified (eg in one of the three CLARIN resources access pages), the resource will be connected with a *switchboard icon*, which the user can activate to get delegated to the switchboard GUI.

For a given resource, the switchboard displays all the applications that can accept the resource (including what types of analyses/data transformations can be performed). The controlled vocabulary mentioned in the description of the Application Registry will be used.

The user may need to have the possibility to override the metadata given in the CMDI file (or the metadata identified by, for instance, a mimetype detector).

4. Assumptions

All applications registered with the switchboard are encouraged to add provenance information to their analyses (supporting the reproducibility of the results).

With the demo upload app, it will be possible to implement a poor man's workflow engine. Research data derived from the application of a tool can be uploaded to the demo upload app; with appropriate metadata, follow-up tools are suggested, which in turn can be invoked, and the process can be repeated.