



# CLARIN European Demonstrator – C4 Use Case

Lothar Lemnitzer

Berlin-Brandenburgische Akademie der  
Wissenschaften

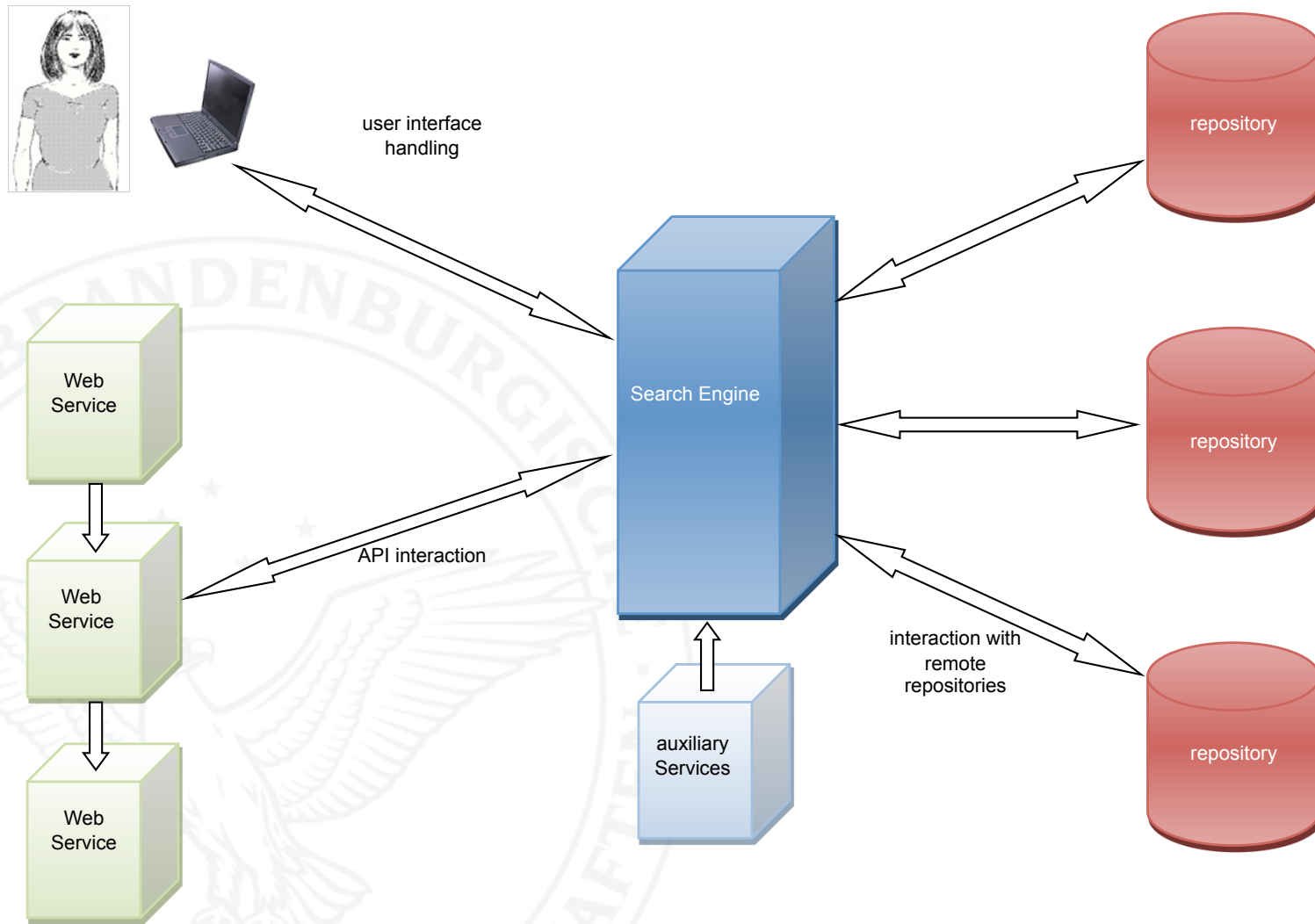
# Ziele eines CLARIN Demonstrators

---

- It should be a demo case involving various European partners, i.e. it should operate crossing country and institutional boundaries.
- It should exhibit extensibility to a certain extent to new functionality and be suitable to be integrated by other projects.
- It should demonstrate the integration of distributed resources and services which is central to the CLARIN mission.

(European Demonstrator Proposal, Sept. 2009)

# Architektur





# Use Case mit D-SPIN Beteiligung

---

Workshop in Amsterdam Januar 2010

Vorschlag, einen C4-Korpus Use Case zu modellieren, wurde zur weiteren Ausarbeitung vorgeschlagen

Vorschlag von: BBAW (Zimmer), AAC (Durco)

- Arbeitstitel: ***C4WebLicht*** - 'C4 Corpus for *WebLicht*'



# Bestandteile

---

## C4-Korpus:

- Korpus regionaler Varietäten des Deutschen (Standardhochdeutsch, Schweizer Deutsch, Österreichisch, Südtiroler Deutsch)
- Pro Varietät 20. Millionen Token (z.T. geplant)
- Metadaten
- Annotation (Part of Speech, Lemma)
- Verteilt auf Standorte, über gemeinsame Suchmaschine abfragbar
- <http://chtk.unibas.ch/korpus-c4/search> (DDC)



# Bestandteile

---

## Weblicht:

- NLP Werkzeuge und flexibel zu definierende Toolchain (s)
- Gemeinsames Input-Textformat, gemeinsame Output-Textformate



# Geplanter Workflow

---

1. Definition Teilkorpus über MD-Abfrage
2. Extraktion von Daten mit einer Suchmaschine (DDC?)
3. Konversion dieser Daten in das Weblicht Textformat
4. Annotation dieser Daten (syntaktische Analyse)
5. Ausgabe von Strukturbäumen
6. (weitere Ausgabeformate?)



# Szenarien / Fragen

---

1. „(keinen) Sinn machen“ vs. „(keinen) Sinn haben“ im Standardhochdeutschen und im Schweizerdeutschen
2. Unterschiede in der PP-Auswahl bei Verben der Bewegung/Lokalverben (gehen / sich befinden) im Schweizer vs. Standardhochdeutschen
3. Weitere???



1. “It [the use case] should exhibit extensibility to a certain extent to new functionality and be suitable to be integrated by other projects. “

→ Ist das wirklich der Fall?

2. Verhältnis der WebLicht Architektur zur CLARIN Planung

3. Wenige Beteiligte, z.Z. Dt/Österreich



# Chancen

---

- Use case ist mit relativ einfachem Aufwand zu realisieren
- Einige der Komponenten (verteilte Daten, gemeinsamer Zugriff über WS)
- Linguistische Beispiele / Forschungsfragen dürften ohne große Verrenkungen umzusetzen sein

Vielen Dank...

# Für Ihre Aufmerksamkeit

<http://www.vronk.net/wiki/index.php/>

[CLARIN EDC C4W](http://www.vronk.net/wiki/index.php/CLARIN_EDC_C4W)

(offenes Wiki für den Use Case)