D-Spin & WebLicht

Thomas Zastrow

University Tübingen

thomas.zastrow@uni-tuebingen.de

Outline

- The D-Spin Project
- Clarin Federation
- WebLicht
- Further Work

The D-Spin Project

The D-Spin Project

- *eScience is about global collaboration in key areas of science and the next generation of infrastructure that will enable it.* (J. Taylor)

- The aim of D-SPIN along with CLARIN is to establish a *virtual research infrastructure* based on available language resources and tools
- D-Spin stands for *Deutsche Sprach Resourcen Infrastructure* (German Language Resource Infrastructure)
- D-SPIN is the German contribution to the European CLARIN-Projekt

The D-Spin Project intends:

- 7 to 9 centres are created
- These centres will work together within a resource-provider-federation and are embedded in the **Clarin Federation infrastructure** (AAI = Authentification and Authorization Infrastructure)
- German resources, data, and tools are gradually made available via **state-of-the-art-registries and web services**
- The regulatory framework is designed in such a way that researchers from DFN AAI institutions are able to merge and exchange data
- Simple workflow models and tools can be defined on these interoperable resources
- Various projects are accomplished along with humanists, in order to develop specific solutions and basic services
- Training sessions are carried out

The D-Spin Project

- Partners are:
  - Max Planck Institute for Psycholinguistics, Nijmegen
  - Department of Linguistics, Computational Linguistics, Tübingen (Coordinator)
  - IDS - Institute for the German Language, Mannheim
  - BBAW - Berlin-Brandenburgische Akademie der Wissenschaften, Berlin
  - ASV - Department of Computer Science, NLP Group, Leipzig
  - Comparative Linguistics, Frankfurt a.M.
  - DFKI - German Research Center for Artificial Intelligence, Saarbrücken
  - IMS - Institute for Natural Language Processing, Stuttgart
  - FB05 - Applied Linguistics and Computational Linguistics, Gießen

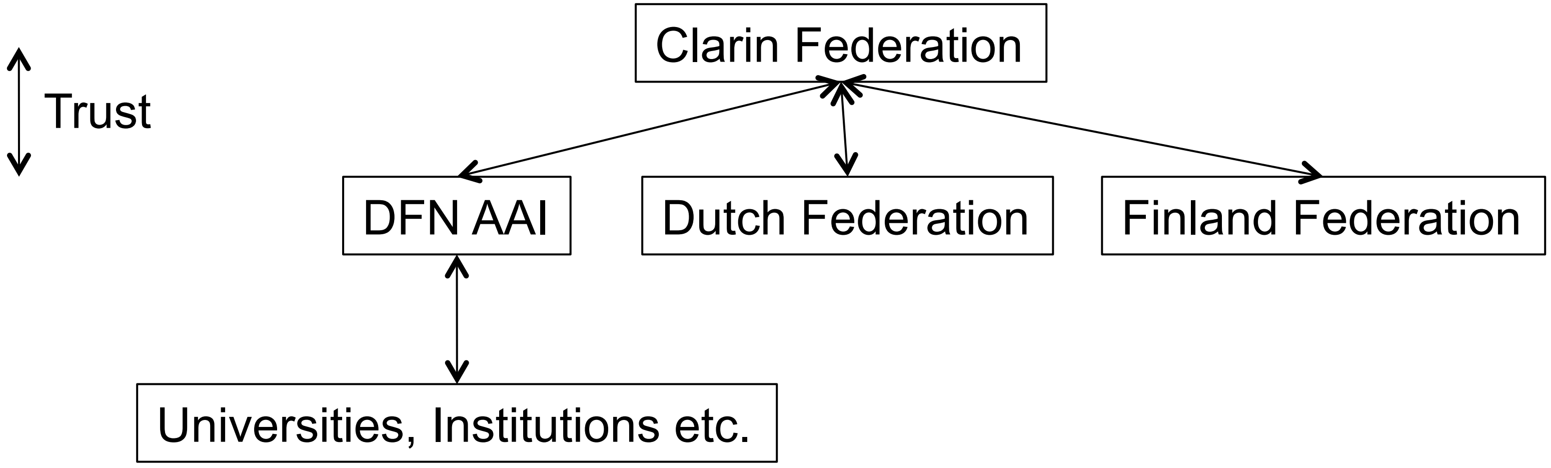
The Clarin Federation

The Clarin Federation

- The CLARIN Service Provider Federation consists of a set of formal agreements and some software technology specifications to authenticate and identify the researchers in a secure way when they work on distributed language resources and applications
- The initial CLARIN Service Provider Federation already gives more than one million researchers and students from 3 countries (Netherlands, Finland, Germany) access to resources and applications of the participating centers

The Clarin Federation

- The Clarin federation is based on the Single-Sign-On system *Shibboleth*
- Shibboleth is already implemented as a nationwide service in many countries



Trust

Clarin Federation

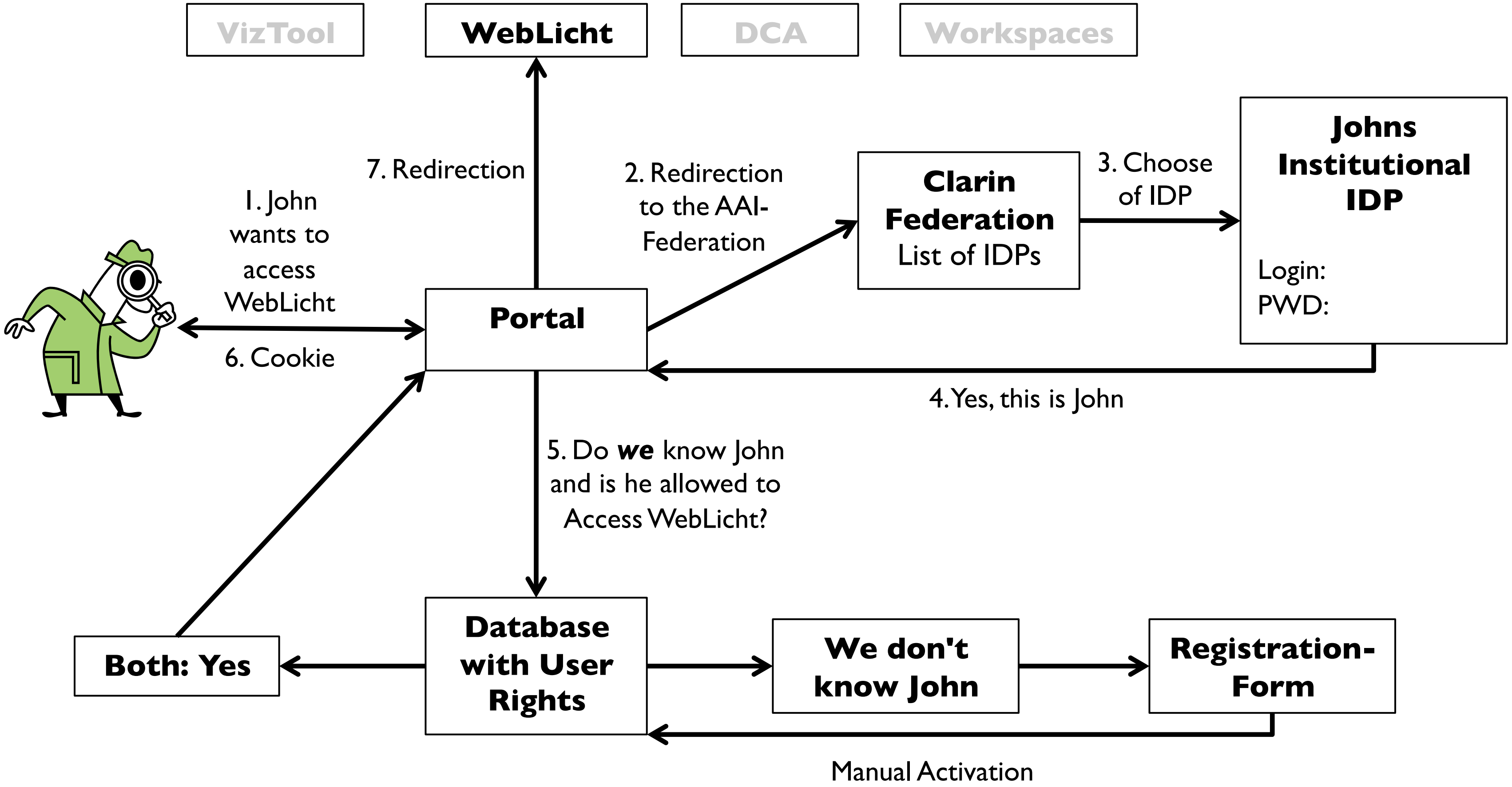
DFN AAI

Dutch Federation

Finland Federation

Universities, Institutions etc.

The Clarin Federation



WebLicht

WebLicht - Motivation

- Many linguistic resources (corpora, dictionaries, …) and tools (tokenizer, tagger, parser, …) are available
- Most of them are implemented to run on local machines. This can be inconvenient and error-prone
- One possible solution: *Make them available on the web!*

WebLicht - Motivation

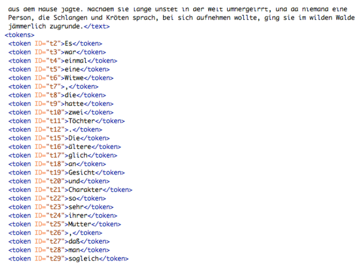
- For some kinds of LRT, its easy to put them online (make resources downloadable, offer search engines etc.)
- For other kinds, more effort is necessary (limiting access to resources, how to make tools online usable)
- Solution: a *Service Oriented Architecure (SOA)*
  - ➔ The end user needs just a browser: no more installation, configuration etc. of software is necessary

WebLicht - Architecture

- WebLicht is a *SOA* for building annotated text corpora
- Development started in October 2008
- WebLicht consists of the following components:
  - **Distributed Services:** offering functionality (resources & tools) over the (inter-)net. Implemented as webservices
  - **Repository:** stores metadata and technical information about the services
  - **User interface:** interacts with the user and combines services and information from the repository. Access still possible via scripts / programming code

WebLicht - Architecture

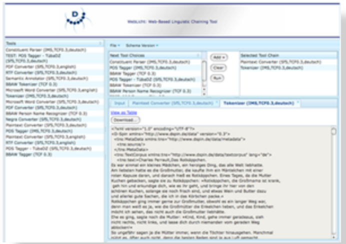
Stuttgart



```
 aus dem Hause jagte. Nachdem sie lange unstet in der Welt umhergeirrt, und da niemand eine
 Person, die Schlangen und Kröten sprach, bei sich aufnehmen wollte, ging sie im wilden Walde
 jämmerlich zugrunde.</text>
<tokens>
 <token ID="t2">Es</token>
 <token ID="t3">war</token>
 <token ID="t4">einmal</token>
 <token ID="t5">eine</token>
 <token ID="t6">Witwe</token>
 <token ID="t7">,</token>
 <token ID="t8">die</token>
 <token ID="t9">hatte</token>
 <token ID="t10">zwei</token>
 <token ID="t11">Töchter</token>
 <token ID="t12">.</token>
 <token ID="t15">Die</token>
 <token ID="t16">ältere</token>
 <token ID="t17">glich</token>
 <token ID="t18">an</token>
 <token ID="t19">Gesicht</token>
 <token ID="t20">und</token>
 <token ID="t21">Charakter</token>
 <token ID="t22">so</token>
 <token ID="t23">sehr</token>
 <token ID="t24">ihrer</token>
 <token ID="t25">Mutter</token>
 <token ID="t26">,</token>
 <token ID="t27">daß</token>
 <token ID="t28">man</token>
 <token ID="t29">sogleich</token>
```

Standard-conformant
Text Corpus Encoding

Tübingen

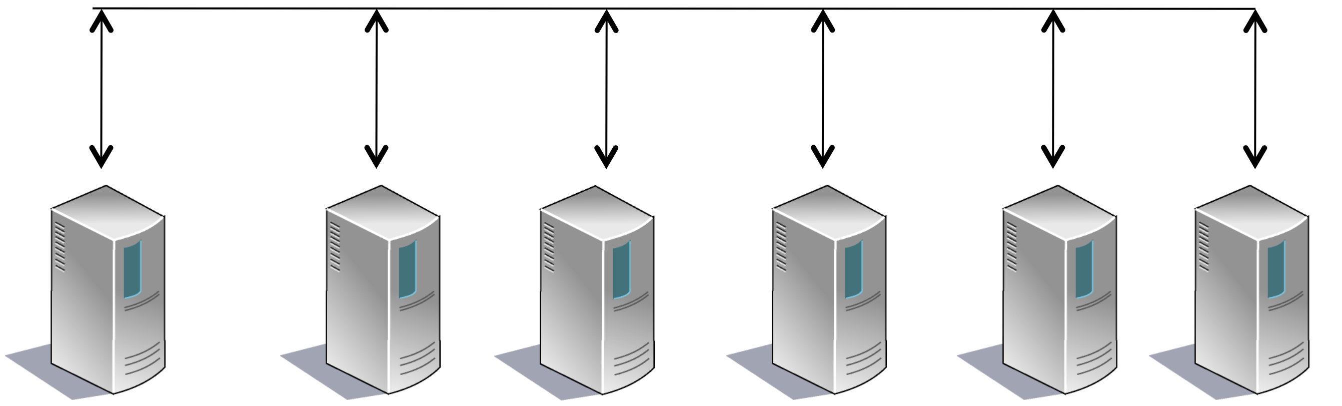

WebLicht: Web-Based Linguistic Chaining Tool
Tools
Constituent Parser (IMS,TCF0.3,deutsch)
TEST: POS Tagger - TübaDZ (SfS,TCF0.3,deutsch)
PDF Converter (SfS,TCF0.3,english)
RTF Converter (SfS,TCF0.3,deutsch)
Semantic Annotator (SfS,TCF0.3,deutsch)
BBAW Tokenizer (TCF 0.3)
Microsoft Word Converter (SfS,TCF0.3,english)
Tokenizer (IMS,TCF0.3,deutsch)
Microsoft Word Converter (SfS,TCF0.3,deutsch)
PDF Converter (SfS,TCF0.3,deutsch)
BBAW Person Name Recognizer (TCF 0.3)
Negra Converter (SfS,TCF0.3,deutsch)
Plaintext Converter (SfS,TCF0.3,deutsch)
POS Tagger (IMS,TCF0.3,deutsch)
Plaintext Converter (SfS,TCF0.3,english)
RTF Converter (SfS,TCF0.3,english)
POS Tagger - TübaDZ (SfS,TCF0.3,deutsch)
BBAW Tagger (TCF 0.3)
File
Schema Version
Next Tool Choices
Constituent Parser (IMS,TCF0.3,deutsch)
POS Tagger (IMS,TCF0.3,deutsch)
BBAW Tagger (TCF 0.3)
POS Tagger - TübaDZ (SfS,TCF0.3,deutsch)
BBAW Tokenizer (TCF 0.3)
BBAW Person Name Recognizer (TCF 0.3)
Add >
Clear
Run
Selected Tool Chain
Plaintext Converter (SfS,TCF0.3,deutsch)
Tokenizer (IMS,TCF0.3,deutsch)
Input
Plaintext Converter (SfS,TCF0.3,deutsch)
Tokenizer (IMS,TCF0.3,deutsch)
View as Table
Download...

```
<?xml version="1.0" encoding="UTF-8"?>
<D-Spin xmlns="http://www.dspin.de/data" version="0.3">
  <tns:MetaData xmlns:tns="http://www.dspin.de/data/metadata">
    <tns:source/>
  </tns:MetaData>
  <tns:TextCorpus xmlns:tns="http://www.dspin.de/data/textcorpus" lang="de">
    <tns:text>Charles Perrault,Das Rotkäppchen.
Es war einmal ein kleines Mädchen, ein herziges Ding, das alle Welt liebhatte.
Am liebsten hatte es die Großmutter, die kaufte ihm ein Mäntelchen mit einer
roten Kapuze daran, und danach hieß es Rotkäppchen. Eines Tages, da die Mutter
Kuchen gebacken, sagte sie zu Rotkäppchen: »Rotkäppchen, die Großmama ist krank,
 geh hin und erkundige dich, wie es ihr geht, und bringe ihr hier von den
schönen Kuchen, solange sie noch frisch sind, und etwas Wein und Butter dazu
und allerlei gute Sachen, die ich in das Körbchen packe.«
Rotkäppchen ging immer gerne zur Großmutter, obwohl es ein langer Weg war,
denn man weiß es ja, wie die Großmütter die Enkelchen lieben, und das Enkelchen
möcht ich sehen, das nicht auch die Großmutter liebhätte.
Ehe es ging, sagte noch die Mutter: »Kind, Kind, gehe immer geradeaus, sieh
nicht rechts, nicht links, und lasse dich durch niemanden vom geraden Weg
ablocken!«
So ungefähr sagen ja die Mütter immer, wenn die Töchter hinausgehen. Manchmal
nützt es, öfter auch nicht, denn die besten Reden sind ja aus Luft gemacht.
```

Web 2.0 Application for
Tool Chaining
and Execution

Leipzig

Repository



Stuttgart Tübingen Berlin Leipzig Finland Romania

WebLicht – The Services

- Services are implemented as REST style webservices
- HTTPs POST method is used to send data from the UI to the services
- As client, anything which is able to use the HTTP protocol, can be used:
  - Browser
  - Commandline tools (wget, curl)
  - Programming Languages
- Anyone can implement his/her own interface to WebLicht

WebLicht – The Repository

- Implemented at the ASV Leipzig
- It offers information and a query engine for the services:
  - Which services are available?
  - How can I combine them?
  - Which input/output format does a service accept/produce?

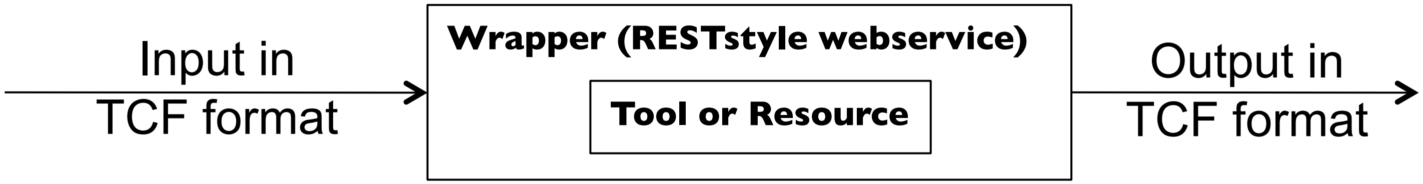
- Example: a tokenizer is already applied to a plain text, which services can be used next?

WebLicht – The User Interface

- Web 2.0 Application for Tool Chaining and Execution
- Implemented at the SfS Tübingen
- Java application, deployed in Apache Tomcat
- Allows the user to:
  - upload a text (plain text, MS Word, RTF or PDF files)
  - construct a text from corpora in Leipzig
  - use some hardwired example texts
  - Build a chain of linguistic tools
  - Executes the tool chain with the uploaded text and presents the results
- During the chaining process, it queries the repository for available services

WebLicht – Integrating new Services

- Building a webservice for WebLicht consists of the following steps:
  - Create a RESTstyle webservice around the tool as wrapper
  - Make in- and output compatible with WebLicht's TCF format
  - Register the service in the repository



Input in TCF format

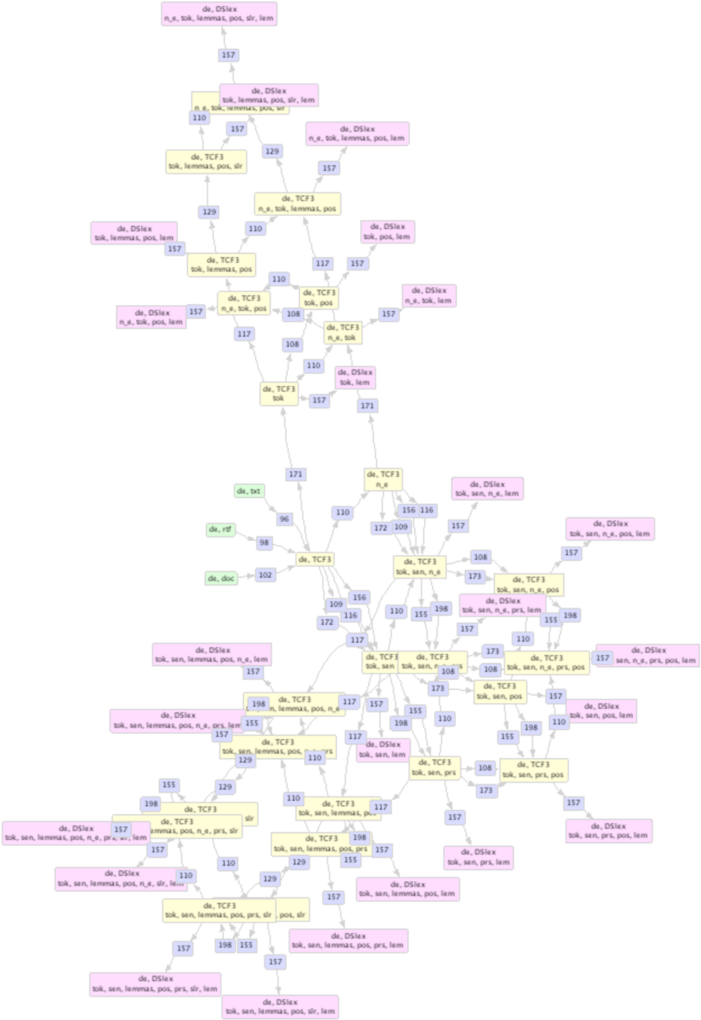
Wrapper (RESTstyle webservice)

Tool or Resource

Output in TCF format

- *You can find further information and a tutorial online:*
  - *http://weblicht.sfs.uni-tuebingen.de/englisch/weblicht.shtml*

WebLicht - Combinations

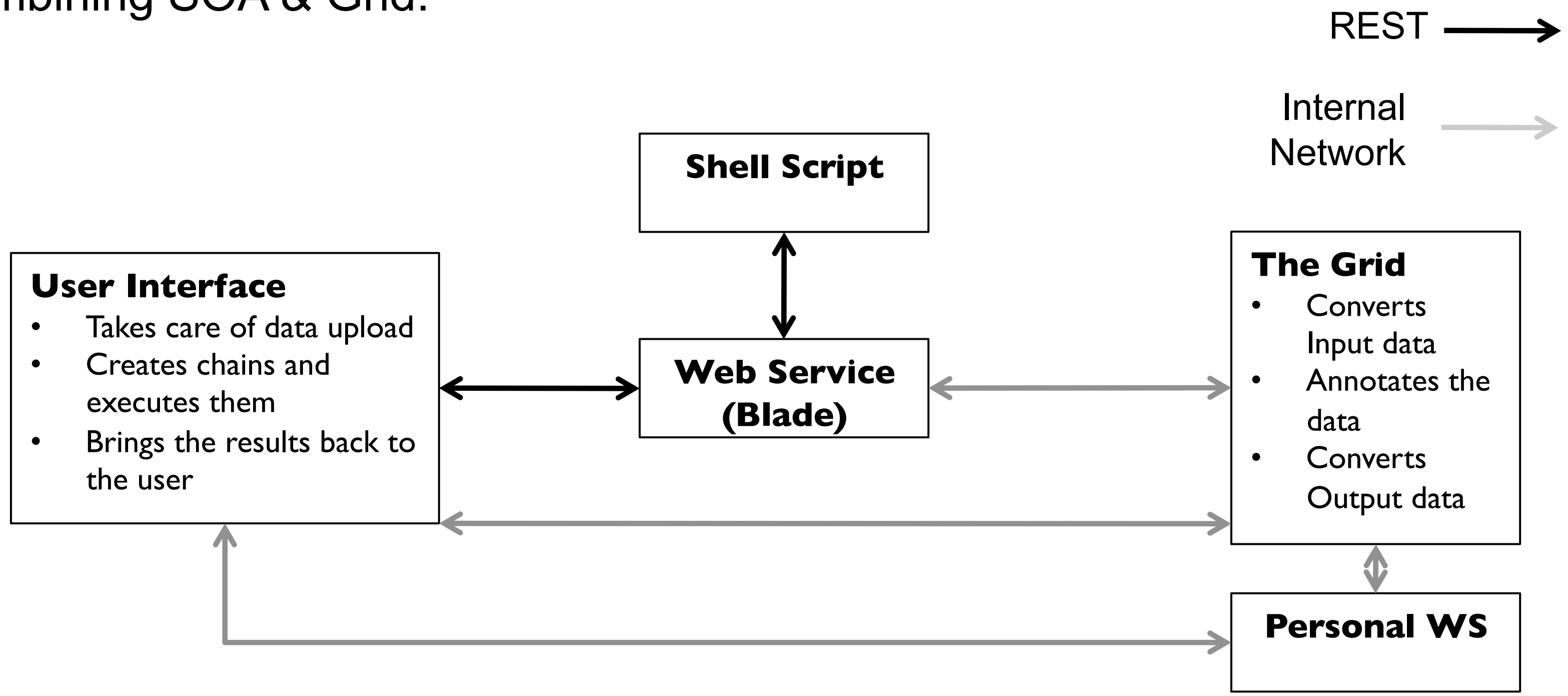


WebLicht – Further Work

- Add more services and functionality
- Usability tests
- Make WebLicht more scaleable
- Creating personal workspaces
- Implement an asynchronous workflow, based on Grid technology

WebLicht goes Grid

Combining SOA & Grid:



Links etc.

- The D-Spin homepage: http://www.d-spin.org
- DFN AAI-Federation based login to WebLicht and some other webapplications: https://weblicht.sfs.uni-tuebingen.de/

Thomas Zastrow
Seminar für Sprachwissenschaft
Universität Tübingen

Wilhelmstr. 19
D-72074 Tübingen

thomas.zastrow@uni-tuebingen.de
http://www.thomas-zastrow.de

Tel.: 07071/29-73968
Fax: 07071/29-5214