

CLARIN-D

Requirements, Examples & Experiences

Dirk Goldhahn

Natural Language Processing Group

Department of Computer Science, University of Leipzig

dgoldhahn@informatik.uni-leipzig.de

UNIVERSITÄT LEIPZIG

Institut für Informatik

GEFÖRDERT VOM



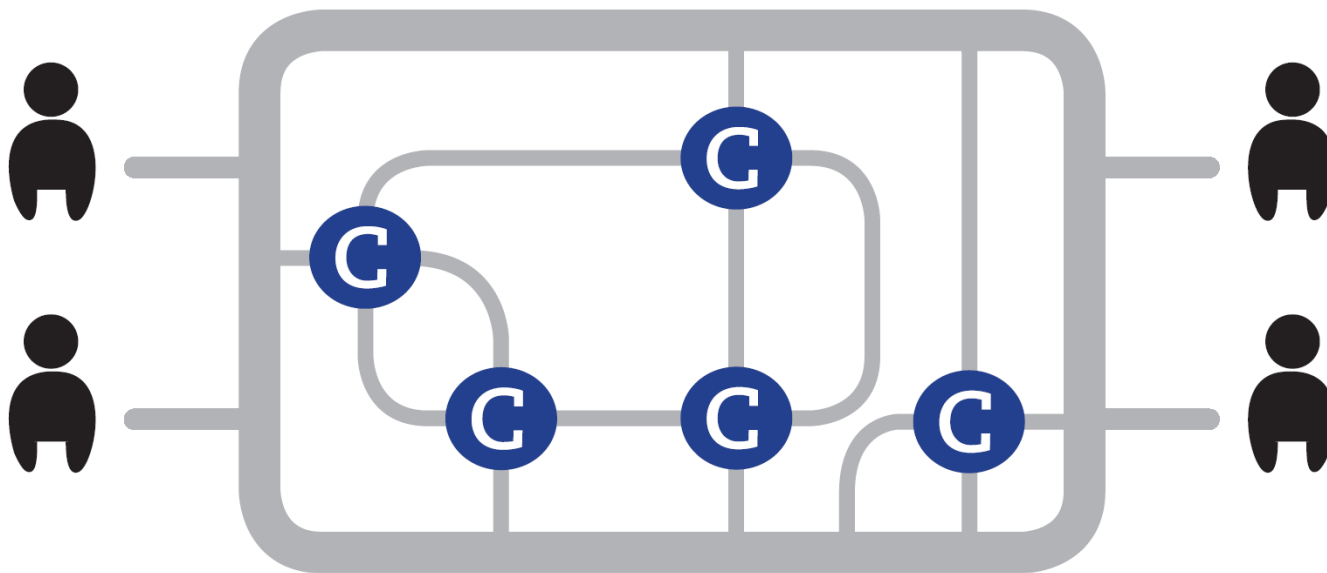
Bundesministerium
für Bildung
und Forschung

- CLARIN(-D):
 - A web and centres-based research infrastructure for the social sciences and humanities
 - Aims to provide relevant, useful data and tools in an integrated, interoperable and scalable way
 - Infrastructure is built in close collaboration with expert scholars in the humanities and social sciences, to ensure that it meets the needs of users in a systematic and easily accessible way

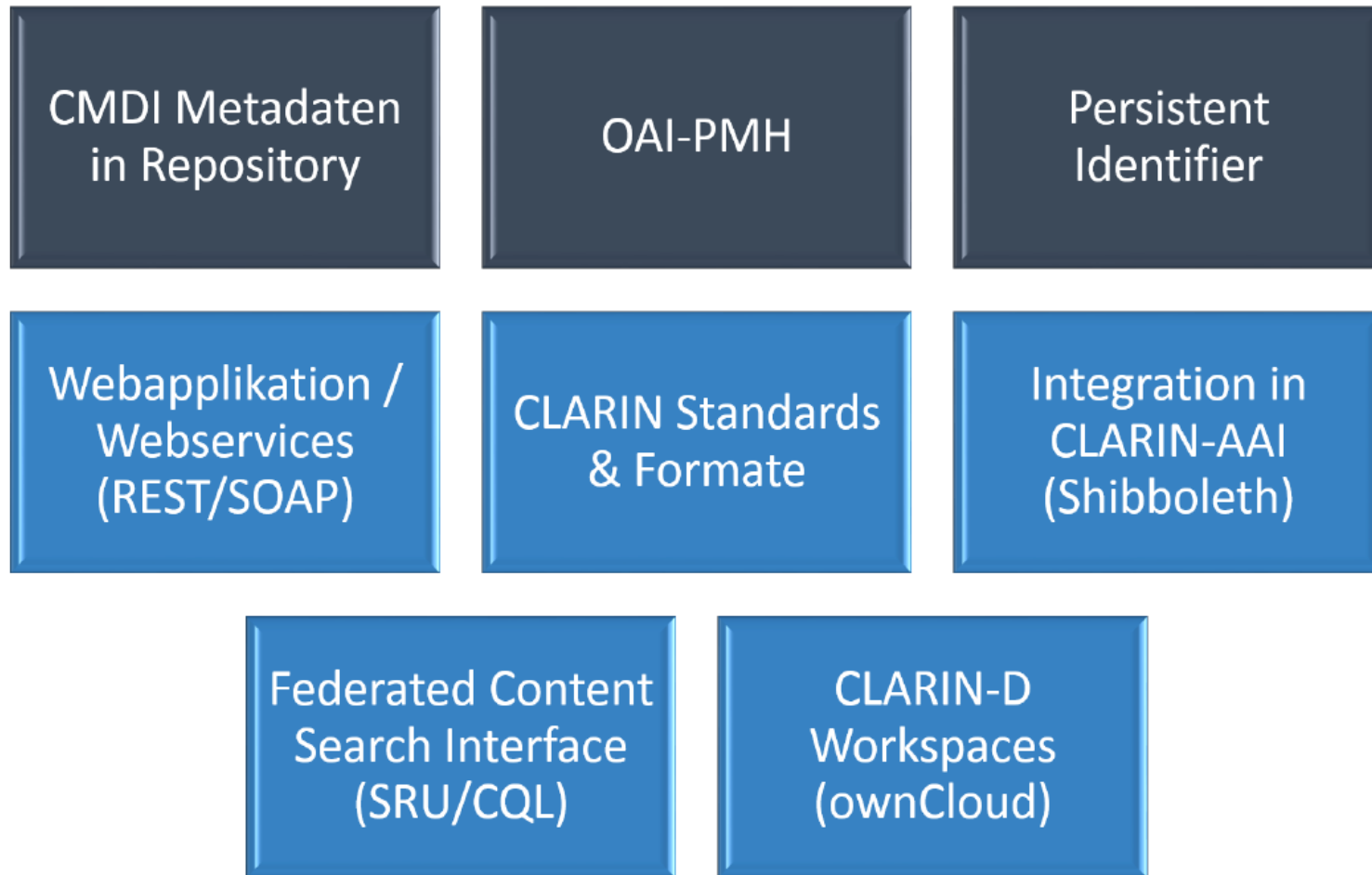


Requirements

- Not one monolithic setup
 - But using a well-working paradigm (think of the Internet): distributed architecture
- services to researcher



Requirements or „means to get there“

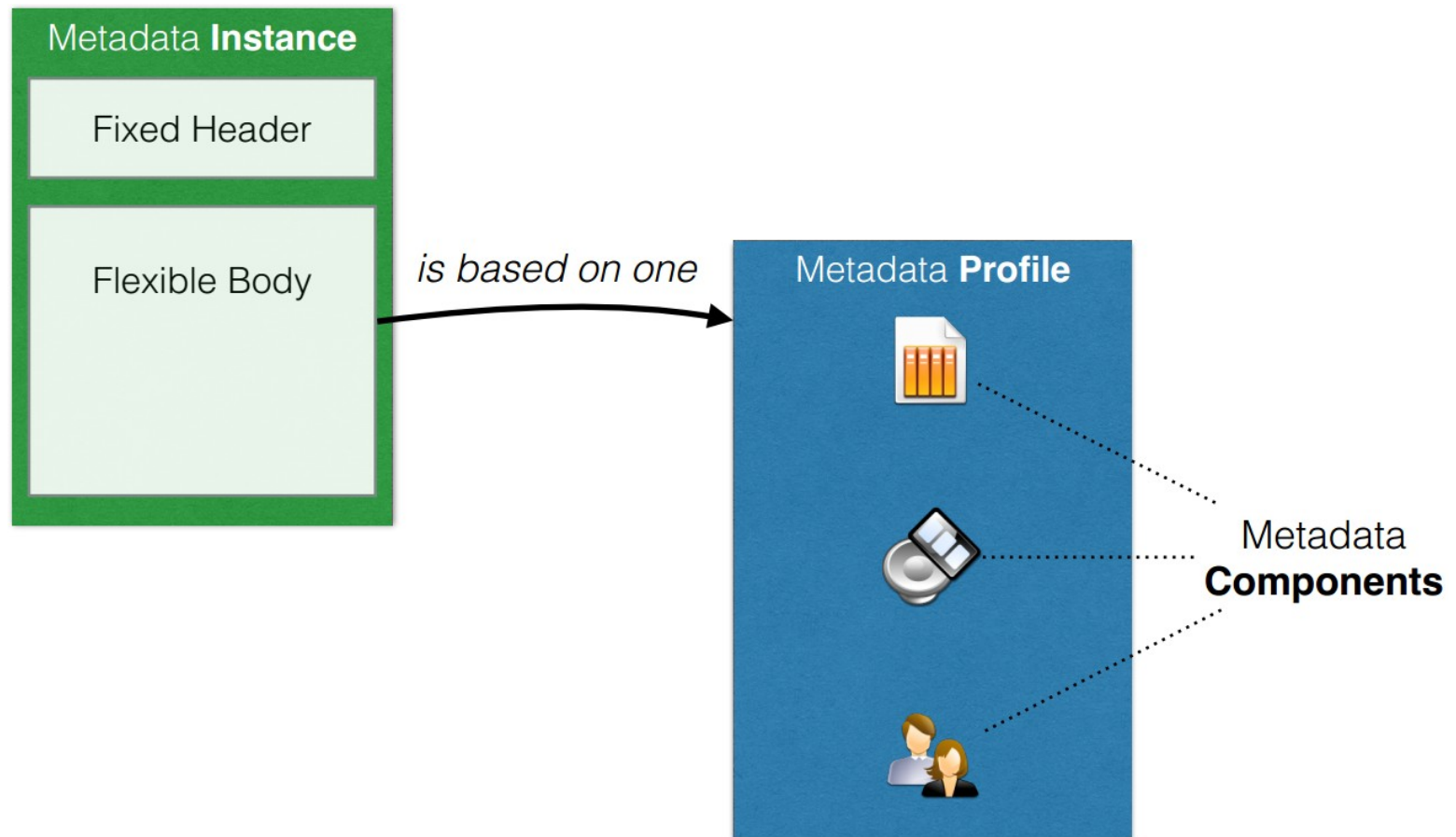


CLARIN chose for a component approach: CMDI

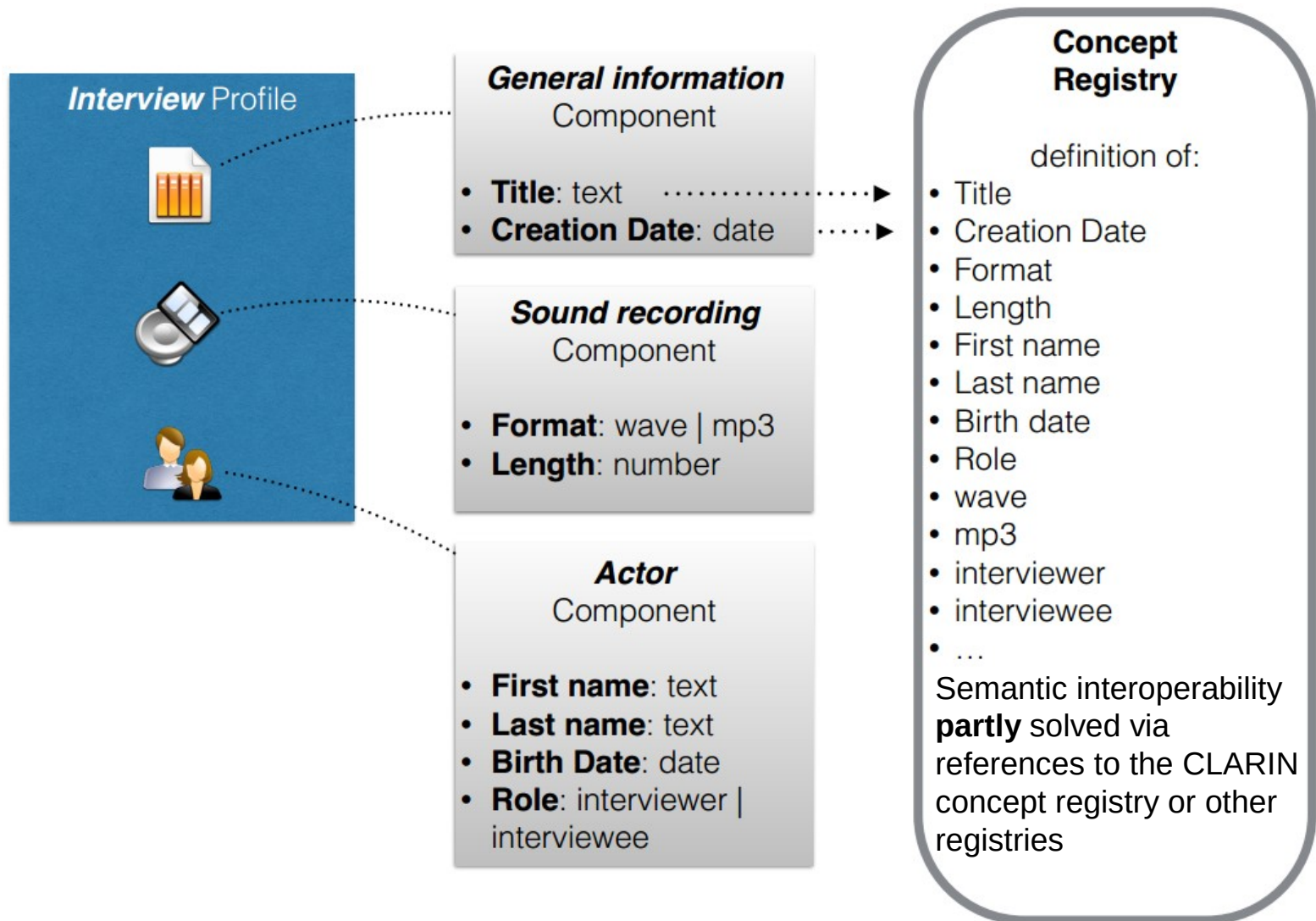
- NOT a single new metadata schema
- ...but rather allow coexistence of many controlled schemas
- ...with explicit semantics for interoperability
- Profiles for describing resources of the same type
- Reuse via components

How does it work?

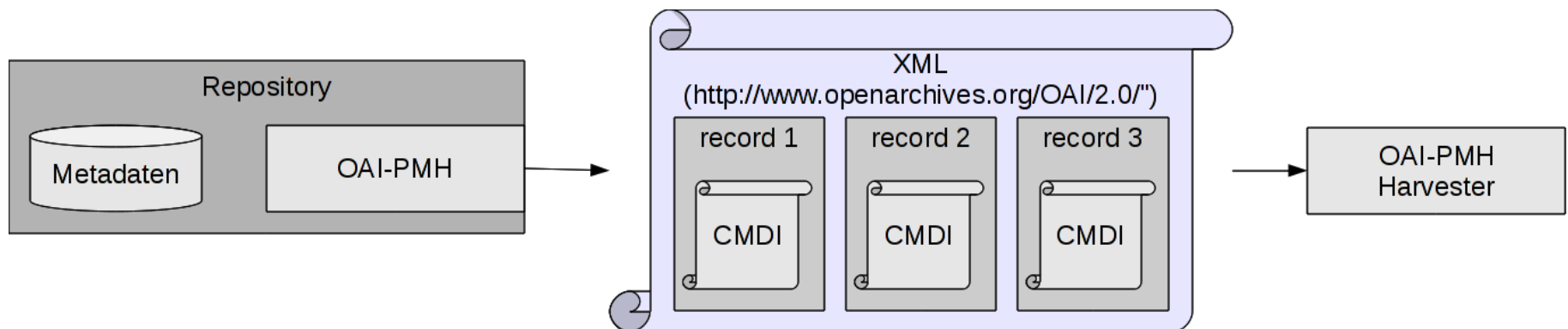
- Components are bundles of related metadata elements that describe an aspect of the resource
- A complete description of a resource may require several components
- Components may contain other components
- Components should be designed for reusability



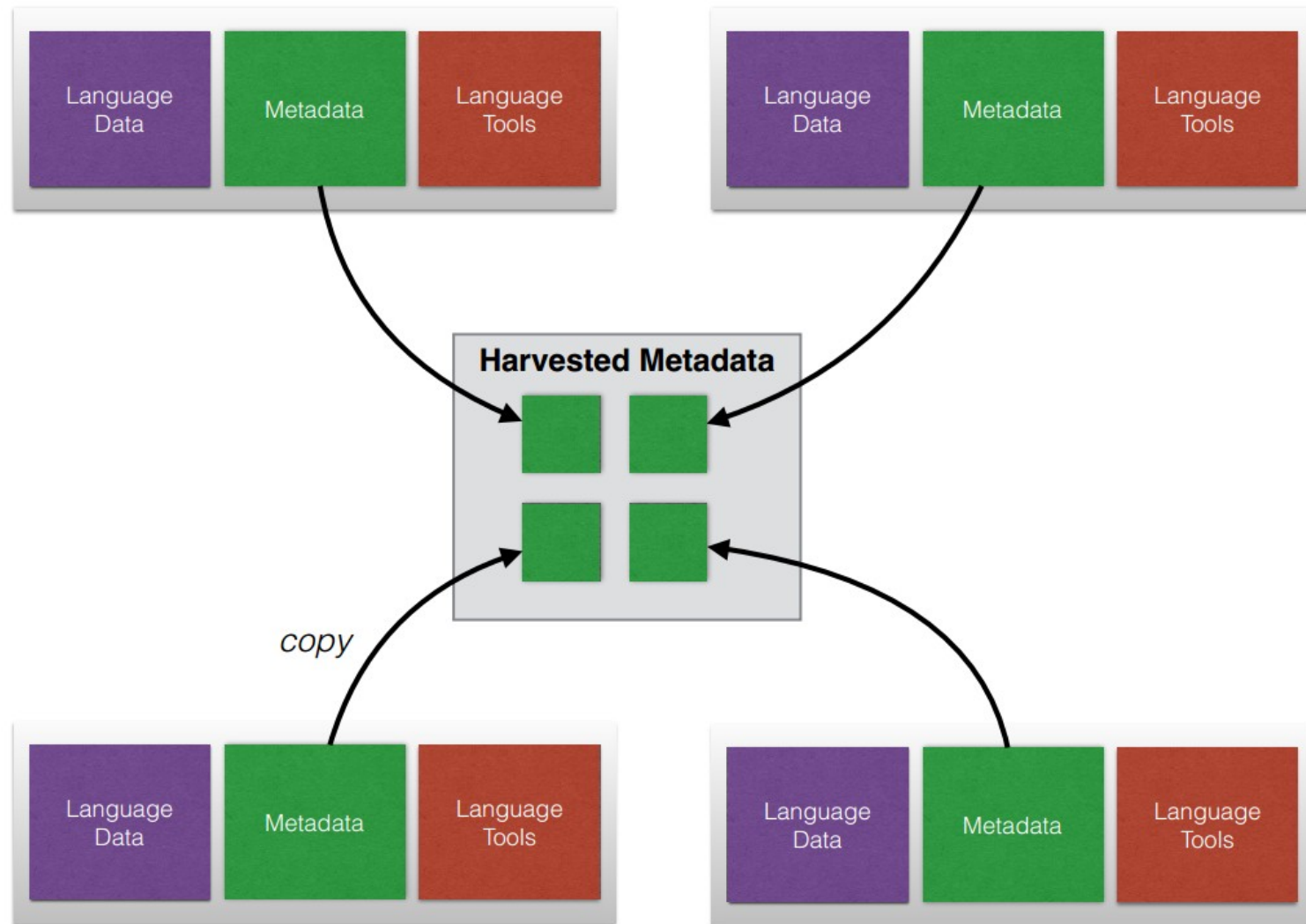
Reuse of existing profiles & components is recommended



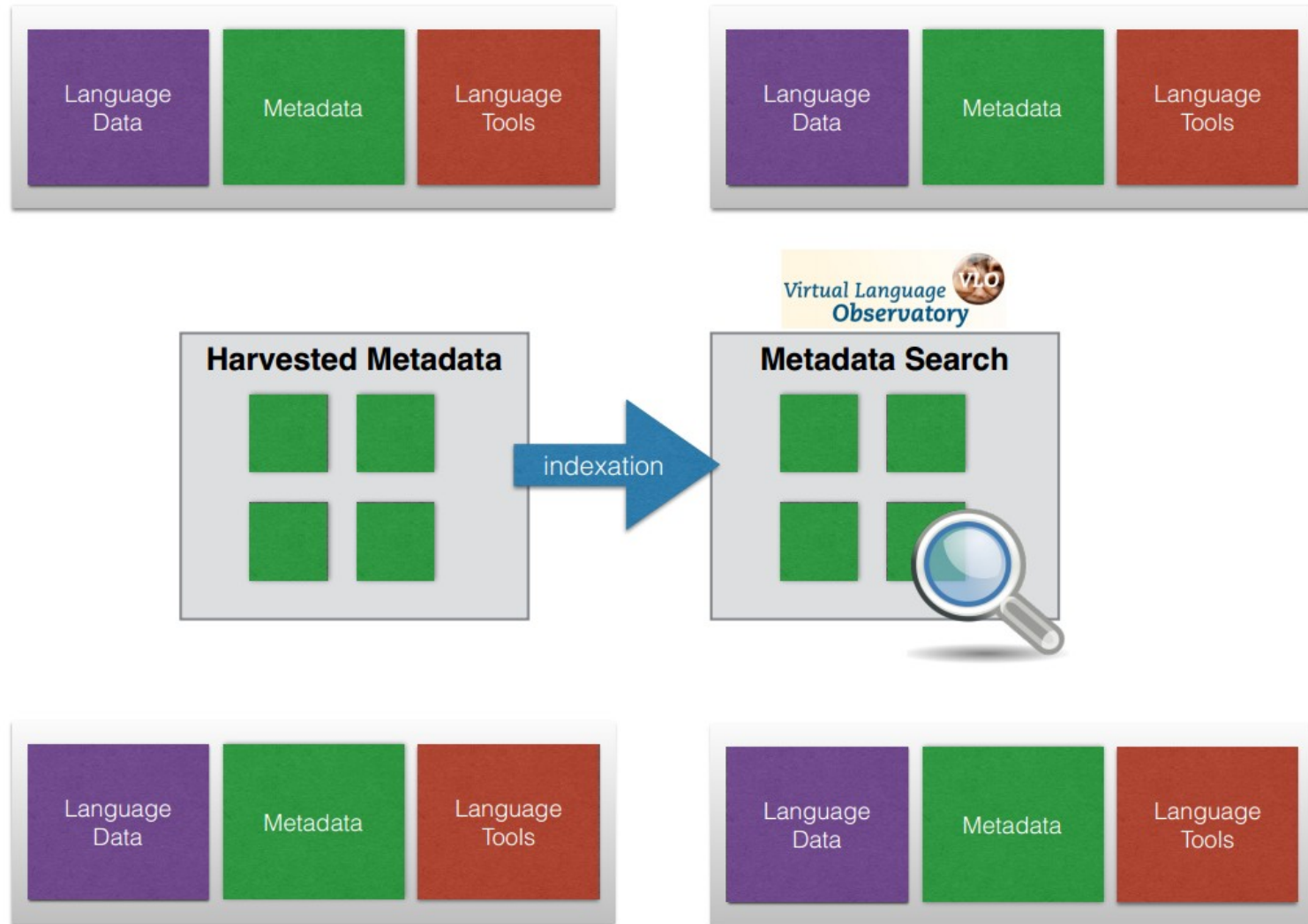
- OAI-PMH – **O**pen **A**rchives **I**nitiative **P**rotocol for **M**etadata **H**arvesting
 - typically: Access datasets at data provider (e.g.: metadata for books, ...)
 - CLARIN: simple access to all metadata stored at the repositories (CMDI-format)
 - Based on HTTP/REST and XML
 - OAI-PMH Harvester: Collects Metadata of all CLARIN-centres
 - Example: Which datasets in CMDI-format have been added since 01.03.2014 (last visit)?



Harvesting (1)



Harvesting (2)

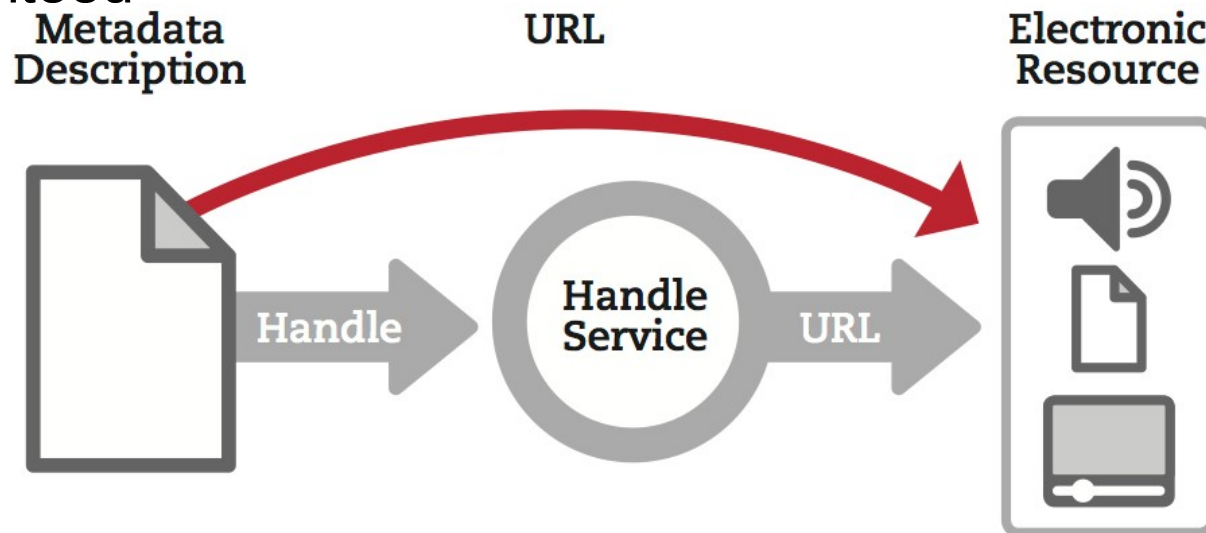


Why?

- To ensure the stability of scientific citations of language resources and the associated metadata descriptions

How?

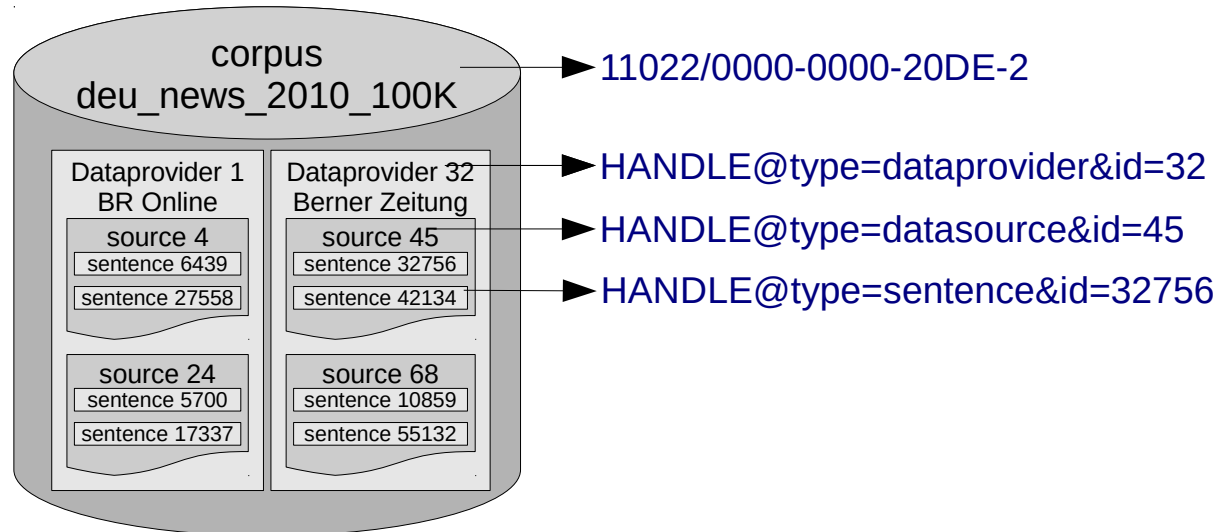
- By adding a level of indirection when resolving an identifier towards a URL, the long-term stability of the references can be guaranteed



Prüfsumme



11858/00	229C	0000-0001-B06F-3	@type=source&id=6921
Handle-Inhaber (GWDG)	Institution (ASV)	Objekt (Korpus „deu_news_2008_10K“)	PartIdentifier (Quelle mit id=6921; „Berner Zeitung“)

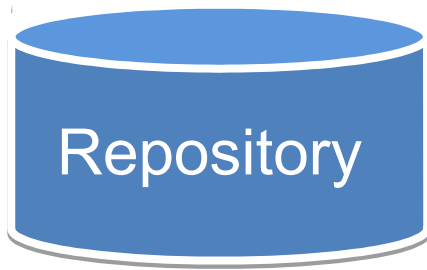


W O R T S C H A T Z
U N I V E R S I T Ä T L E I P Z I G

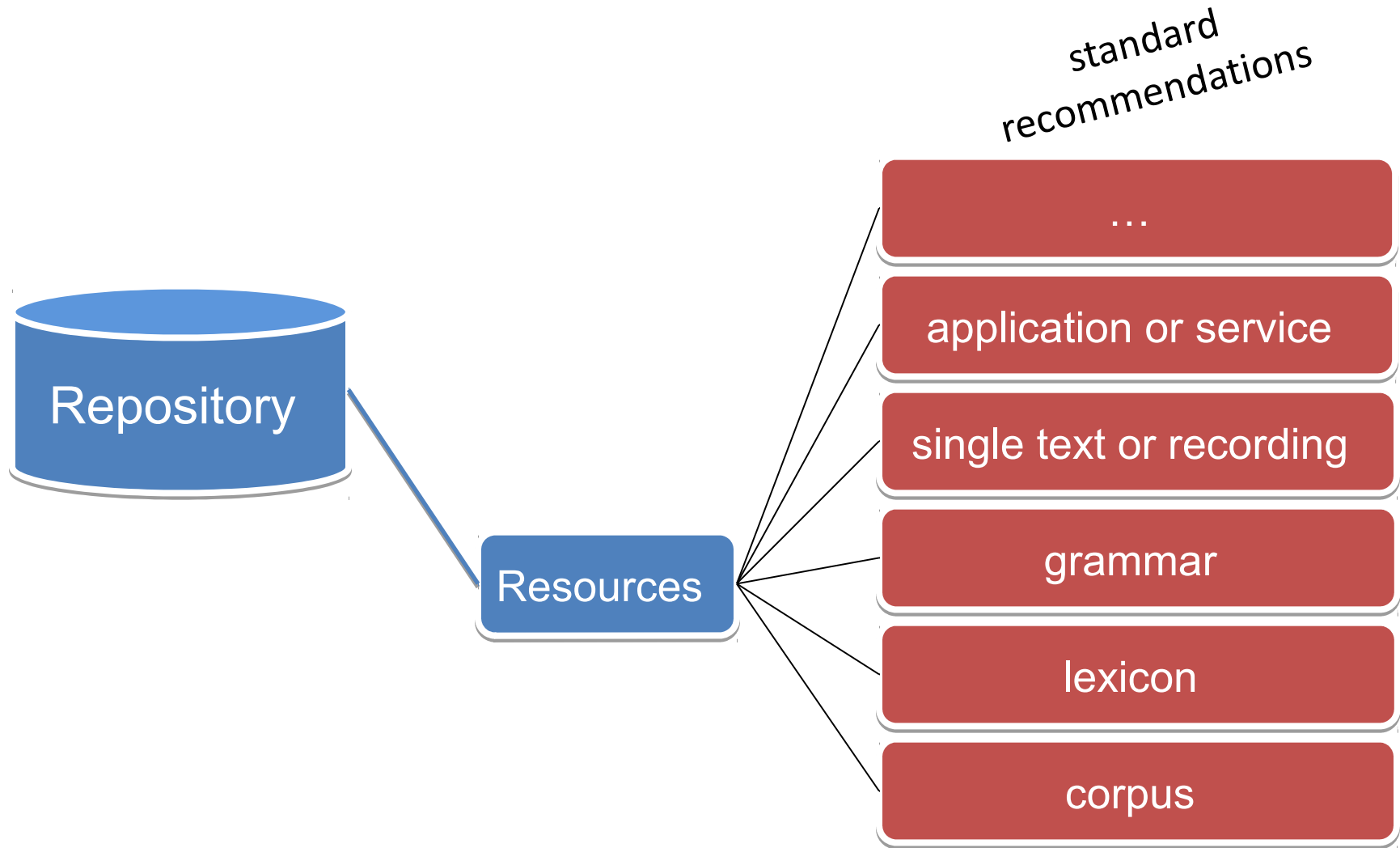
- Access via Webapplications and/or Webservices
- REST-webservices (also SOAP)
- Retrieval of data & access to tools
- optional: Accessibility via WebLicht (webapp)
 - REST
 - TCF-format (XML-TEI in preparation)
 - CMDI metadata (specification of input & output)
 - chaining of webservices (primarily annotation)

- Recommendation: Usage of well established standards
- Conversion, semi-automatic integration, ...

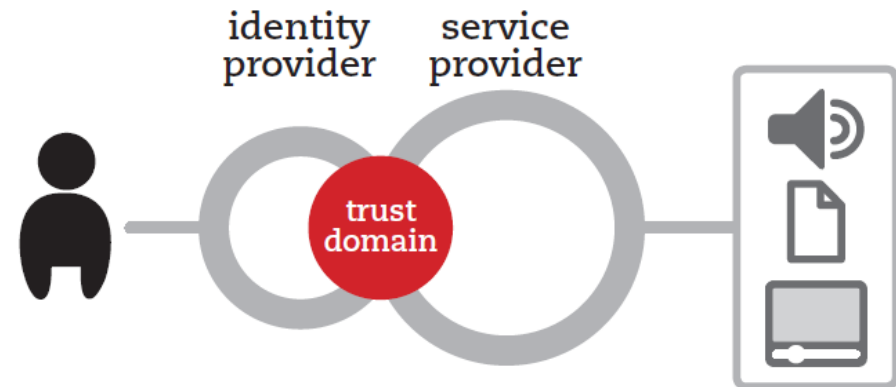
Abbreviation/Name	Topic(s)	Standard body	CLARIN Center(s)
CES	Generic Corpus Annotation	EAGLES	
CHAT	File Formats Transcription	Other	
CMDI	Metadata	ISO	BAS, CLARIN-PL, IDS, TLA, UC
Controlled Vocabulary	Controlled Vocabulary Knowledge Representation Thesaurus	NISO	CLARIN-PL, TLA
CQLF	Query	ISO	IDS
DCAM	Metadata	DCMI	
DCMES	Metadata	DCMI	TLA
DCR	Data Categorization	ISO	CLARIN-PL, IDS, TLA
DiAML	Markup Language Semantic Annotation	ISO	

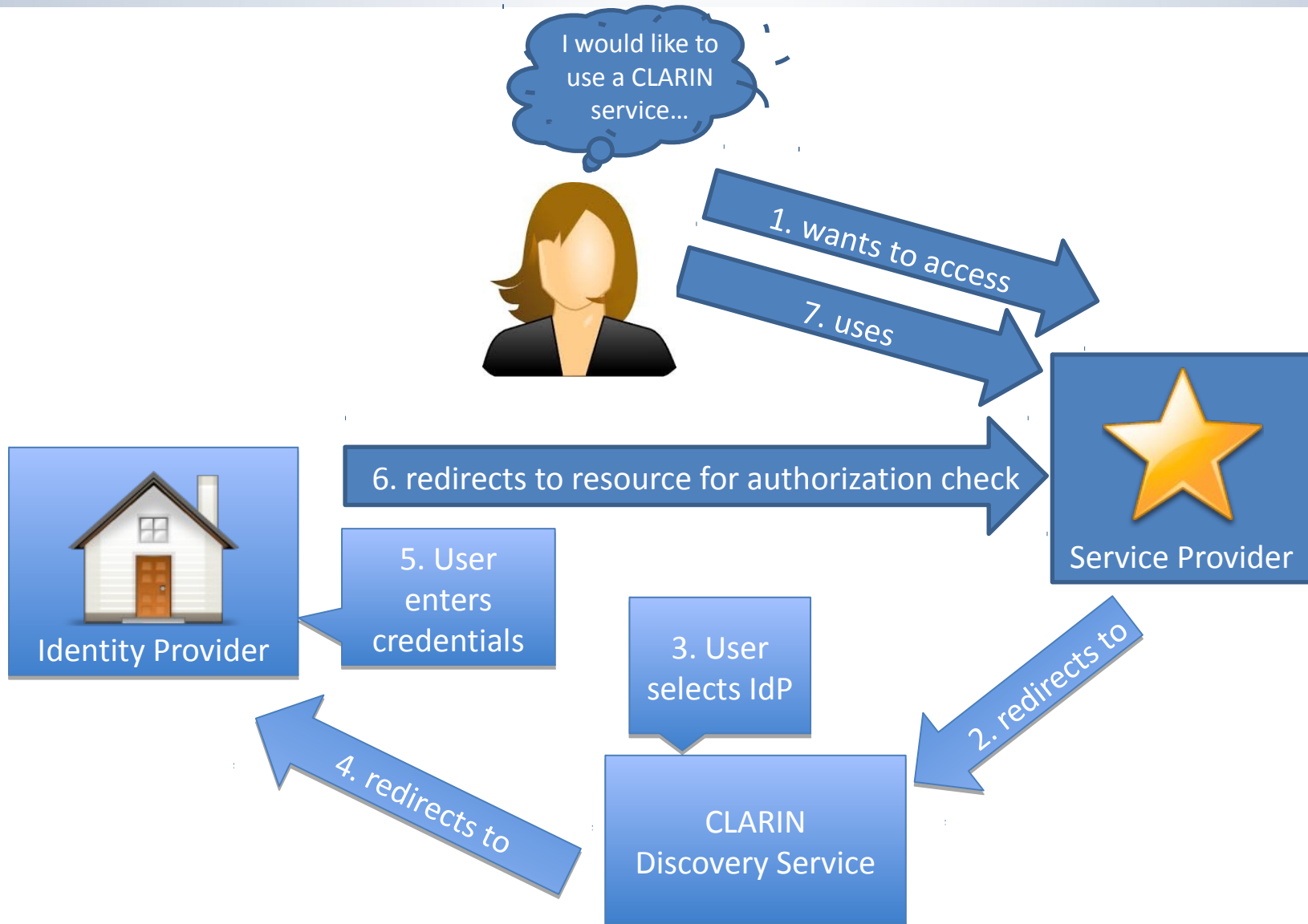


Repository = sustainable
store at a CLARIN center
that can be accessed via
the internet



- Use your own (institutional) login to access password-protected resources → Single Sign On
- Trust domain (service provider, identity provider)
- Decentralized user management
- Requirement for B centers
- Available Services:
 - Service Provider Federation
 - CLARIN Identity Provider
 - Easy-to-use discovery service





User

[Getting Started](#)

[Use Cases and Examples](#)

[User Manual](#)

[FAQ](#)

Developer

[Developer Manual](#)

Associated Applications

[WebLicht](#)

[Annotation Viewer](#)

[TViewer](#)

[TüNDRA - the Tübingen aNnotated Data Retrieval Application](#)

[FCS Aggregator](#)

[CiNaViz - Visualization of European City Names](#)

Personal tools

[Wiki login](#)

Main Page

Welcome to WebLicht!

WebLicht is an execution environment for automatic annotation of text corpora. Linguistic tools such as tokenizers, can be combined by the user into custom processing chains. The resulting annotations can then be visualized in an appropriate way.

Many of the tools incorporated into WebLicht have existed as command-line or desktop tools for many years and are now available for use with WebLicht. By making these tools available on the web and by use of a common data format for storing the results, they can be used in a more flexible way.

It is often the case that a tool relies on one or more annotation layers to exist in the input data before it can begin processing. WebLicht provides a framework in which tools are able to operate on the data at any given point and offering only those tools as options to add to the chain.

Development of WebLicht started in October 2008 as part of the [D-SPIN](#) project, the predecessor project of [CLARIN-D](#), aiming to make WebLicht a fully-functional virtual research environment.

Tools

Tools and processing tasks are described in more detail in the [User Manual](#) under [Tools in Detail](#).

Try It Out!

If you have an account in the CLARIN Service Provider Federation (for example a DFN-AAI account, provided by many universities), you can log in to WebLicht. The user interface has been upgraded to include more help and documentation, and also includes a new "Easy Mode". Currently "Easy Mode" is available for German text only, but other languages will be added soon.

[Start WebLicht](#)



Einloggen bei **Clarin EU Service Provider**

Select your Identity Provider

If you cannot find your institution in the list above please select the "Clarin.eu website account" and use your credentials of the **CLARIN website**. For questions please contact spf@clarin.eu.


 clarin.eu website account
 European Union

The University Hospital Brno
 Czech Republic




Universiteit Utrecht
 Netherlands




Hochschule Kempten
 Germany



Fachhochschule Dortmund
 Germany



Universiteit Hasselt
 Belgium

Universität Potsdam
 Germany



Meine Position bestimmen und Anbieter in der Nähe anzeigen

Zeige Anbieter in Based on DiscooJuice © UNINETT 



CLARIN EU website Identity Provider.

You are now trying to access <https://sp.catalog.clarin.eu>

Please login using **your e-mail address** by which you are subscribed to the [CLARIN website](#) and the corresponding password.
In case of problems please contact webmaster@clarin.eu.

Email:

Password:

Login

Main Page

+ New Chain

WebLicht



Welcome to WebLicht

WebLicht consists of a collection of web-based linguistic annotation tools, distributed repositories for storing and retrieving information about the tools, and this web application, which allows you to easily create and execute tool chains without downloading or installing any software on your local computer.

This application and its associated tools are continually being updated and improved.

For more information, visit our websites at [WebLicht](#), [CLARIN-D](#), and [CLARIN](#).

What's New

There are 2 modes for building tool chains:

- Easy Mode lets you choose pre-defined processing chains
- Advanced Mode allows you to build customized tool chains.

In this version, the input selection/upload process was made more intuitive.

Getting Started

Click on the "+ New Chain" tab at the top left of this page or click on the "Start" button below:

Start >>

Available Annotations for: German Plain Text

- ☒ POS Tags/Lemmas
- ☐ Morphology
- ☐ Constituent Parses
- ☐ Dependency Parses
- ☐ Named Entities

TreeTagger

Annotation Layers:

- Simple view
 - ☐ text
 - ☐ sentences
- Table view
 - ☒ tokens
 - ☒ POSTags
 - ☒ lemmas

language = de

 1 / 5

token ID	tokens	POSTags
t1	CLARIN	NN
t2	is	ADJA
t3	the	FM
t4	Common	NE
t5	Language	NN
t6	Resources	NN
t7	and	FM
t8	Technology	NE
t9	Infrastructure	NN

[Download TCF](#)

Input and Chain Selection

CLARIN is the C [Plain

CLARIN is the Common Language Resources and Technology Infrastructure, which aims to provide easy and sustainable access for

SFS - To TCF Converter

Document Type
 Language: German
 Document Type: TCF

IMS - Tokenizer

Sentences
 Tokens

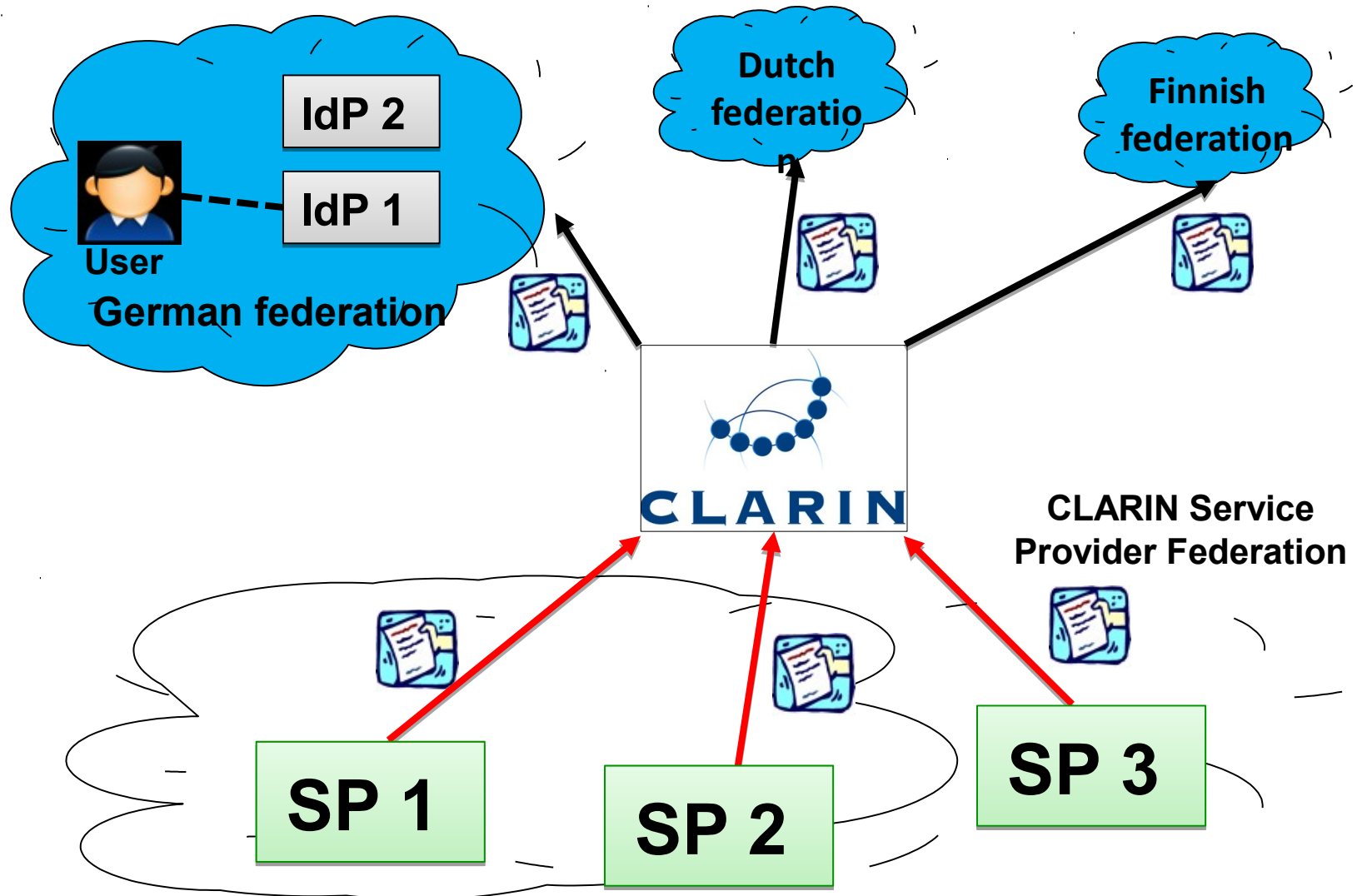
IMS - TreeTagger

Part of Speech: STTS Tagset
 Lemmas

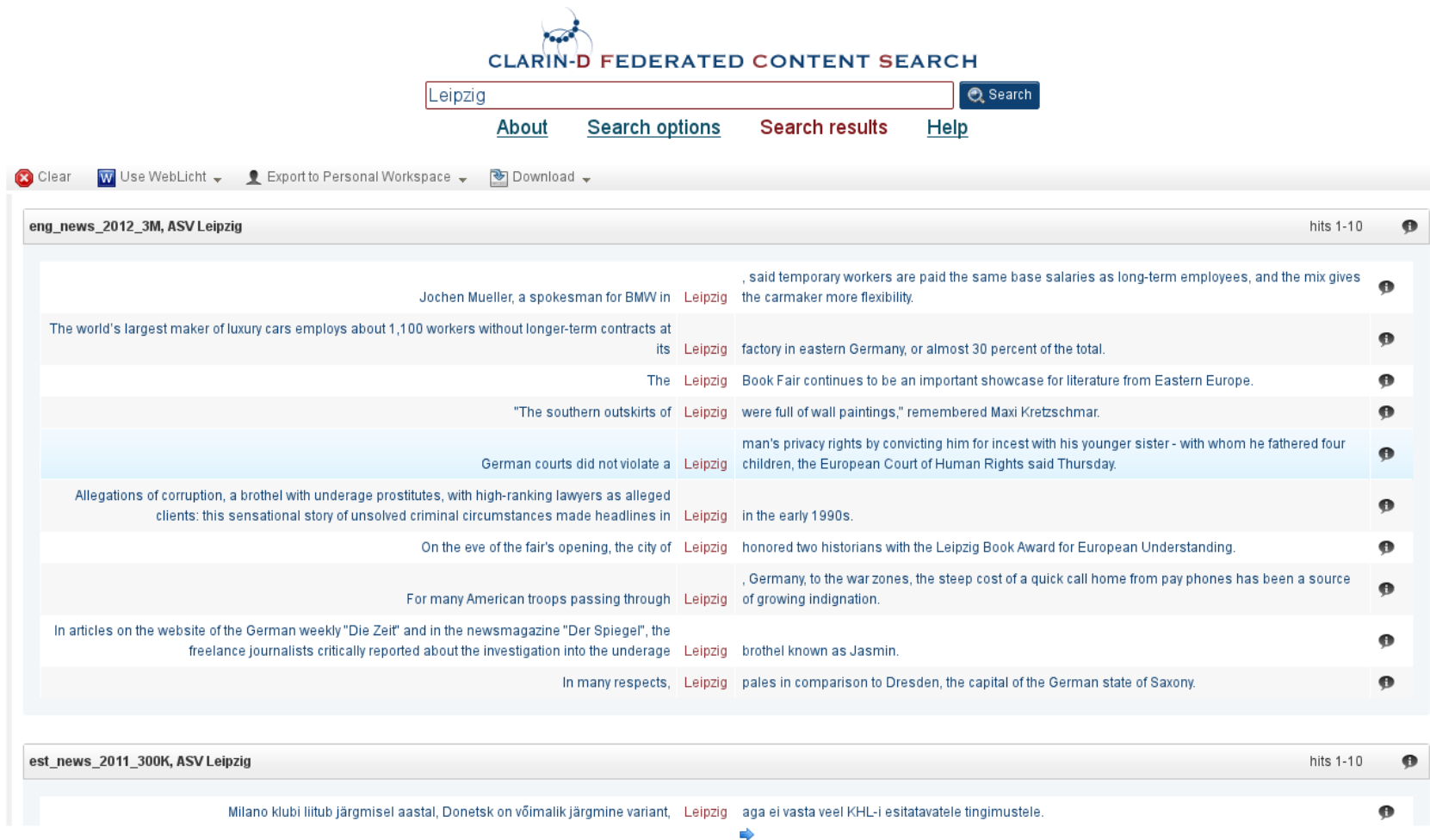
- Member countries without identity federation (e.g. Bulgaria, Lithuania)
- Institutions without identity provider (e.g. Hamburg university)
- Lacking attribute release

→ CLARIN IdP: fallback-system to log in to CLARIN SPs

Service Provider Federation



- FCS – Federated Content Search



The screenshot displays the CLARIN-D Federated Content Search web application. At the top, the title "CLARIN-D FEDERATED CONTENT SEARCH" is centered. Below it is a search bar containing the text "Leipzig" and a "Search" button. Navigation links for "About", "Search options", "Search results", and "Help" are positioned below the search bar. A toolbar at the top of the results area includes a "Clear" button, a "Use WebLight" dropdown, an "Export to Personal Workspace" dropdown, and a "Download" button. The main results section is titled "eng_news_2012_3M, ASV Leipzig" and shows "hits 1-10". It contains a list of search results, each with a snippet of text and a "Leipzig" label. The results are displayed in a table-like format with alternating light and dark blue rows. At the bottom, another section titled "est_news_2011_300K, ASV Leipzig" shows "hits 1-10" and a single result snippet in Estonian.

CLARIN-D FEDERATED CONTENT SEARCH

Leipzig Search

About Search options Search results Help

Clear Use WebLight Export to Personal Workspace Download

eng_news_2012_3M, ASV Leipzig hits 1-10

Jochen Mueller, a spokesman for BMW in Leipzig, said temporary workers are paid the same base salaries as long-term employees, and the mix gives the carmaker more flexibility.

The world's largest maker of luxury cars employs about 1,100 workers without longer-term contracts at its Leipzig factory in eastern Germany, or almost 30 percent of the total.

The Leipzig Book Fair continues to be an important showcase for literature from Eastern Europe.

"The southern outskirts of Leipzig were full of wall paintings," remembered Maxi Kretzschmar.

German courts did not violate a Leipzig man's privacy rights by convicting him for incest with his younger sister - with whom he fathered four children, the European Court of Human Rights said Thursday.

Allegations of corruption, a brothel with underage prostitutes, with high-ranking lawyers as alleged clients: this sensational story of unsolved criminal circumstances made headlines in Leipzig in the early 1990s.

On the eve of the fair's opening, the city of Leipzig honored two historians with the Leipzig Book Award for European Understanding.

For many American troops passing through Leipzig, Germany, to the war zones, the steep cost of a quick call home from pay phones has been a source of growing indignation.

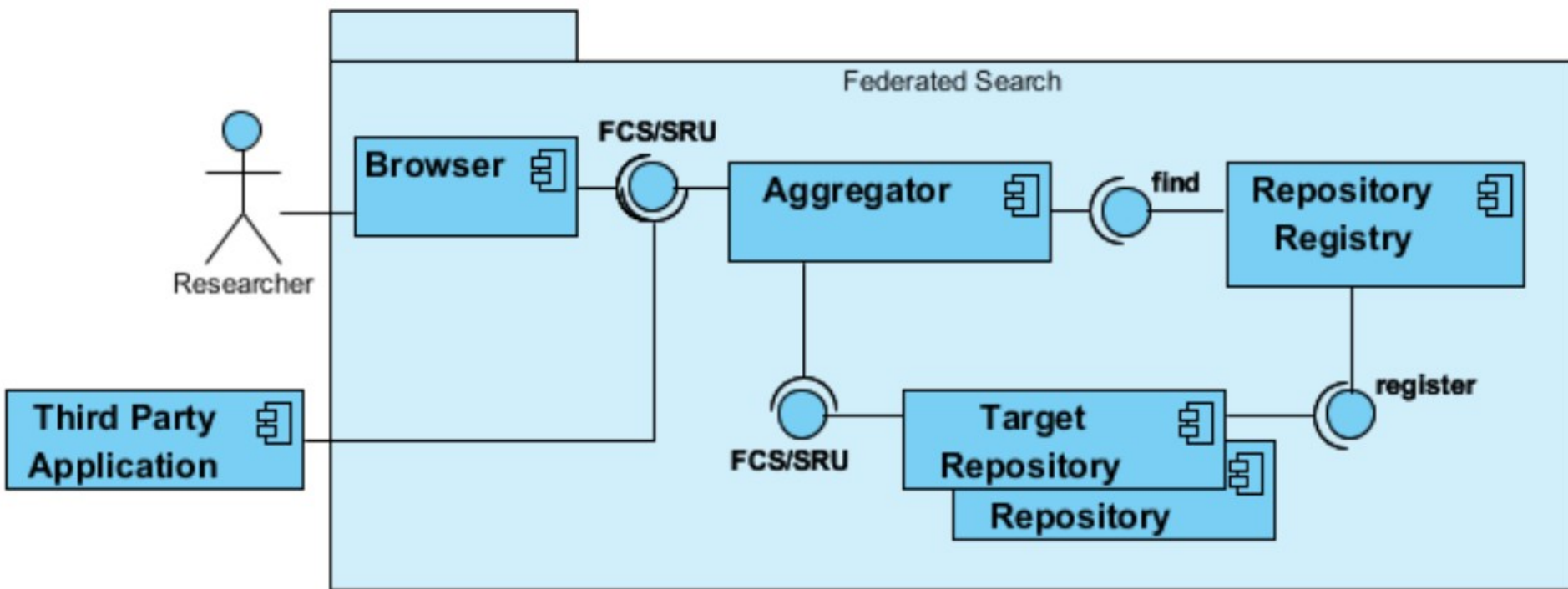
In articles on the website of the German weekly "Die Zeit" and in the newsmagazine "Der Spiegel", the freelance journalists critically reported about the investigation into the underage Leipzig brothel known as Jasmin.

In many respects, Leipzig pales in comparison to Dresden, the capital of the German state of Saxony.

est_news_2011_300K, ASV Leipzig hits 1-10

Milano klubi liitub järgmisel aastal, Donetsk on võimalik järgmine variant, Leipzig aga ei vasta veel KHL-i esitatavatele tingimustele.

- Based on SRU/CQL (Search/Retrieve via URL + Contextual Query Language)
- Querying of content from multiple sources using a standardized interface



A *Personal Workspace* in CLARIN-D consists of online storage for individual users, which can be accessed via a programming API by CLARIN-D applications.

So you can e.g.:

- Save search results of the Federated Content Search
- Import data into WebLicht web services and save results of your processing

OwnCloud was deemed best suited for use as the basis software for implementing personal workspaces in CLARIN-D.

- Data exchange between different applications

