



A Generic Chaining Algorithm for NLP Webservice

Volker Boehlke

University of Leipzig

BoehlkeV@informatik.uni-leipzig.de

UNIVERSITÄT LEIPZIG

Institut für Informatik



Automatische Sprachverarbeitung

WSPP2010

Valetta

2010-05-18

- situation:
 - scientist that visited a conference
 - heard a talk about POS-tagging based on a tagset “STTS“
 - wants to generate POS-tags on a given text
 - unexperienced user, no deeper knowledge about:
 - webservices
 - chaining/pipelines
 - POS-tagging
 - tokenization
 - ...

use case I

A screenshot of a KWrite text editor window. The title bar reads 'langer text.txt - KWrite'. The menu bar includes 'Datei', 'Bearbeiten', 'Ansicht', 'Lesezeichen', 'Extras', 'Einstellungen', and 'Hilfe'. The toolbar contains icons for file operations (open, save, print, delete, undo, redo, cut, copy, paste, find, find and replace) and search functions. The text area contains three paragraphs of German text. The first paragraph discusses the historical context of the unified contribution rate in the health insurance fund. The second paragraph, with the phrase 'Jahresanfang' highlighted, describes the current plan for the health insurance fund. The third paragraph mentions the CDU financial expert Otto Bernhardt's views on the contribution rate.

zufrieden. "Der staatlich festgelegte Einheitsbeitragssatz im Gesundheitsfonds gehört der Vergangenheit an", sagte Bahr der dpa. Die Krankenkassen würden künftig wieder selbst über die Höhe ihrer Beiträge entscheiden.

Am Donnerstag hatten die künftigen Koalitionspartner den Plan verworfen, das durch die Krise bedingte Defizit der Krankenversicherung über einen Schattenhaushalt zu decken. Derzeit gilt für die rund 180 Kassen ein einheitlicher Beitragssatz von 14,9 Prozent. Das Geld fließt zusammen mit Steuermitteln in den am Jahresanfang gestarteten Gesundheitsfonds. Von dort wird es an die Kassen verteilt, wobei Versicherungen mit mehr Kranken höhere Zuweisungen bekommen.

Der CDU-Finanzexperte Otto Bernhardt sagte am frühen Morgen im rbb-Inforadio, die Begrenzung der zusätzlichen Kassenbeiträge auf ein Prozent des Bruttoeinkommens werde vermutlich gekippt. Auch das bedeutet: Die Kassen haben mehr Freiheit, den von den Arbeitnehmern und Rentnern zu zahlenden Beitrag je nach Finanzbedarf selbst zu erhöhen. Bernhardt sagte, Erhöhungen seien möglich. Einige Kassen würden ohne Zusatzbeiträge auskommen, anderen werde auch mehr als ein Prozent nicht ausreichen.

use case II

ASV DSpin Registry Management Tool v0.02a

add service
browse registry
manage formats
chain services
auto chain
direct call
log

96 - Plaintext Converter (SfS,TCF0.3,deutsch)

query

Id	Name	Id	Name

execute chain << >>

service	result size	time

save

use case III

ASV DSpin Registry Management Tool v0.02a

add service
browse registry
manage formats
chain services
auto chain
direct call
log

96 - Plaintext Converter (SfS,TCF0.3,deutsch)

stts

Id	Name	Id	Name
117	POS Tagger (IMS,TCF0.3,deutsch)		
112	BBAW Tagger (TCF 0.2)		
108	BBAW Tagger (TCF 0.3)		
61	POS Tagger (IMS,TCF0.2,deutsch)		
167	Morph analyzer (TCF0.3, Finnish)		
173	POS Tagger - OpenNLP Project		

service	result size	time
---------	-------------	------

use case IV

ASV DSpin Registry Management Tool v0.02a

add service
browse registry
manage formats
chain services
auto chain
direct call
log

96 - Plaintext Converter (SfS,TCF0.3,deutsch)

stts

Id	Name	Id	Name
117	POS Tagger (IMS,TCF0.3,deutsch)	96	Plaintext Converter (SfS,TCF0.3,d
112	BBAW Tagger (TCF 0.2)	116	Tokenizer (IMS,TCF0.3,deutsch)
108	BBAW Tagger (TCF 0.3)	117	POS Tagger (IMS,TCF0.3,deutsch)
61	POS Tagger (IMS,TCF0.2,deutsch)		
167	Morph analyzer (TCF0.3, Finnish)		
173	POS Tagger - OpenNLP Project		

<< 1/10 >>

service	result size	time
---------	-------------	------

use case V

ASV DSpin Registry Management Tool v0.02a

add service
browse registry
manage formats
chain services
auto chain
direct call
log

96 - Plaintext Converter (SfS,TCF0.3,deutsch)

stts query

Id	Name	Id	Name
117	POS Tagger (IMS,TCF0.3,deutsch)	96	Plaintext Converter (SfS,TCF0.3,d
112	BBAW Tagger (TCF 0.2)	156	ULei - Tokenizer - deutsch
108	BBAW Tagger (TCF 0.3)	117	POS Tagger (IMS,TCF0.3,deutsch)
61	POS Tagger (IMS,TCF0.2,deutsch)		
167	Morph analyzer (TCF0.3, Finnish)		
173	POS Tagger - OpenNLP Project		

execute chain << 3/10 >>

service	result size	time
---------	-------------	------

save

use case VI

ASV DSpin Registry Management Tool v0.02a

add service
browse registry
manage formats
chain services
auto chain
direct call
log

96 - Plaintext Converter (SfS,TCF0.3,deutsch)

stts

Id	Name	Id	Name
117	POS Tagger (IMS,TCF0.3,deutsch)	96	Plaintext Converter (SfS,TCF0.3,d
112	BBAW Tagger (TCF 0.2)	156	ULei - Tokenizer - deutsch
108	BBAW Tagger (TCF 0.3)	117	POS Tagger (IMS,TCF0.3,deutsch)
61	POS Tagger (IMS,TCF0.2,deutsch)		
167	Morph analyzer (TCF0.3, Finnish)		
173	POS Tagger - OpenNLP Project		

<< 3/10 >>

service	result size	time
Plaintext Converter (SfS,TCF0.3,deutsch)	5 KB	230
ULei - Tokenizer - deutsch	21 KB	312
POS Tagger (IMS,TCF0.3,deutsch)	59 KB	6110

use case VII



```
tokID="136">180</tns:lemma>      <tns:lemma tokID="137">ein</tns:lemma>      <tns:lemma
tokID="138">einheitlich</tns:lemma>      <tns:lemma tokID="139">Beitragssatz</tns:lemma>
<tns:lemma tokID="140">@card@</tns:lemma>      <tns:lemma tokID="141">d</tns:lemma>      <tns:lemma
tokID="142">Geld</tns:lemma>      <tns:lemma tokID="143">fließen</tns:lemma>      <tns:lemma
tokID="144">zusammen</tns:lemma>      <tns:lemma tokID="145">mit</tns:lemma>      <tns:lemma
tokID="146">Steuermittel</tns:lemma>      <tns:lemma tokID="147">Jahresanfang</tns:lemma>
<tns:lemma tokID="148">gestartet</tns:lemma>      <tns:lemma tokID="149">von</tns:lemma>
<tns:lemma tokID="150">dort</tns:lemma>      <tns:lemma tokID="151">werden</tns:lemma>
<tns:lemma tokID="152">es</tns:lemma>      <tns:lemma tokID="153">an</tns:lemma>      <tns:lemma
```

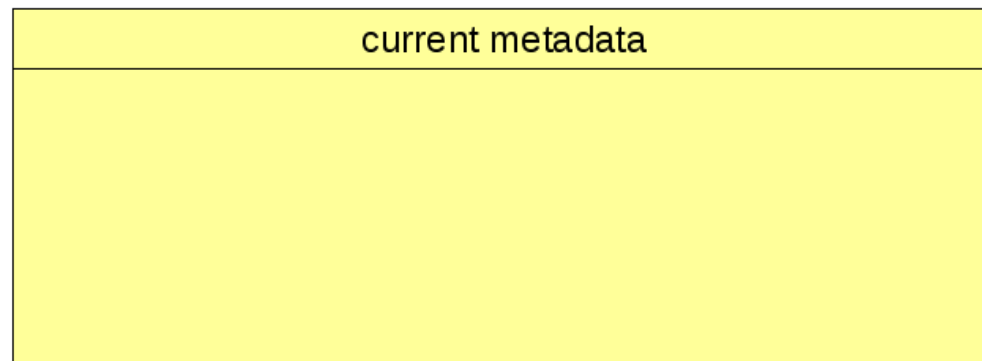
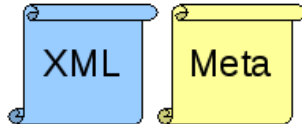
```
tokID="140">CARD</tns:tag>      <tns:tag tokID="141">ART</tns:tag>      <tns:tag
tokID="142">NN</tns:tag>      <tns:tag tokID="143">VVFIN</tns:tag>      <tns:tag
tokID="144">ADV</tns:tag>      <tns:tag tokID="145">APPR</tns:tag>      <tns:tag
tokID="146">NN</tns:tag>      <tns:tag tokID="147">NN</tns:tag>      <tns:tag
tokID="148">ADJA</tns:tag>      <tns:tag tokID="149">APPR</tns:tag>      <tns:tag
tokID="150">ADV</tns:tag>      <tns:tag tokID="151">VAFIN</tns:tag>      <tns:tag
tokID="152">PPER</tns:tag>      <tns:tag tokID="153">APZR</tns:tag>      <tns:tag
```

- viewer/editor needed, but the problem (text → POS) is solved

- How does the chaining work?
 - reduce to: Which services are executable after service A has run?
 - better: Is service B executable after service A was invoked?
- we are searching for „perfect matches“
- mathematics/programming: $c(b(a()))$
- basically the same problem to be solved like on the type checking level of a compile run: Check if all input needs of a certain function are satisfied => correct number and compatible type of parameters
- example:
 - Question: “Which data is needed?”
 - Answer: “A german (iso-6183) text split into tokens encoded in utf8 given in the format DSpin-TextCorpus.”
 - Question: “Which data is produced?”
 - Answer: “POS-Tags according to the STTS standard for each token.”

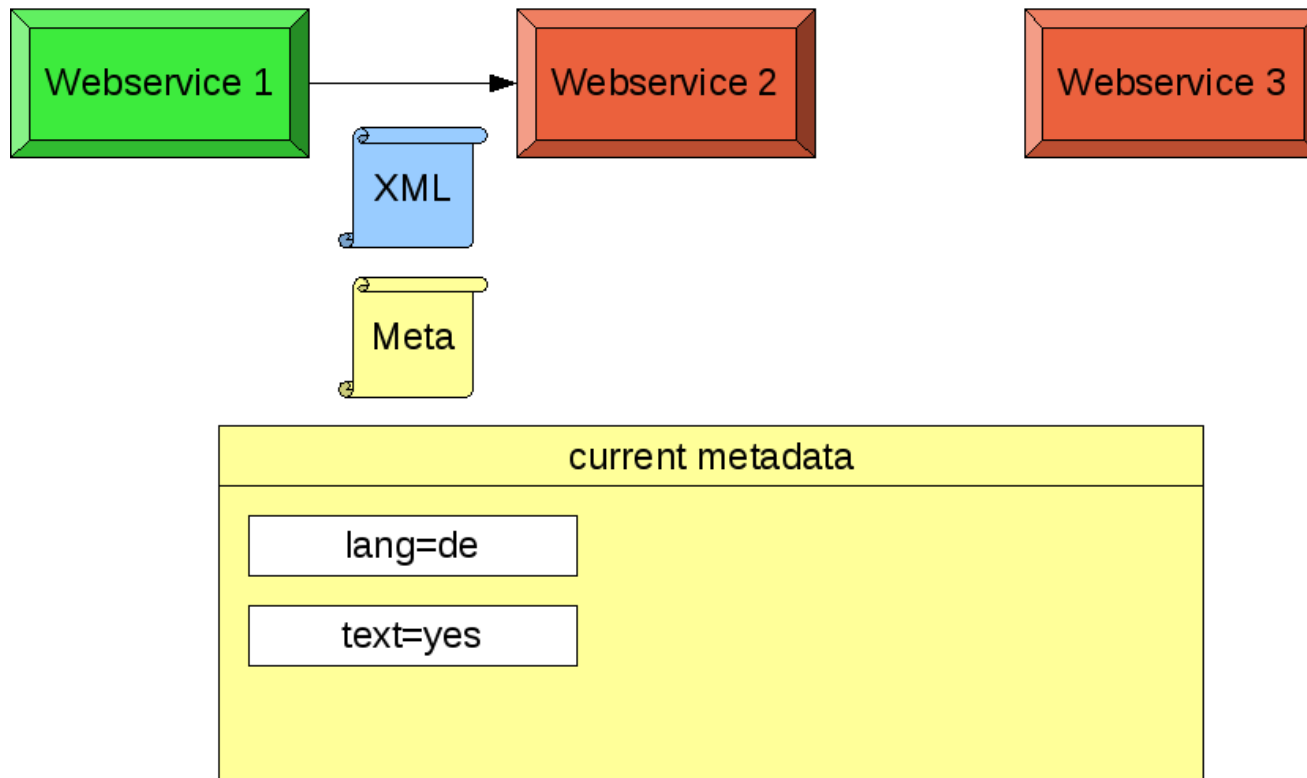
chaining - example I

input	ouput
format=binary	format=tcf0.3
<div style="border: 1px solid black; padding: 5px; width: fit-content;">format=txt</div>	<div style="border: 1px solid black; padding: 5px; width: fit-content;">lang=de</div> <div style="border: 1px solid black; padding: 5px; width: fit-content; margin-top: 10px;">text=yes</div>



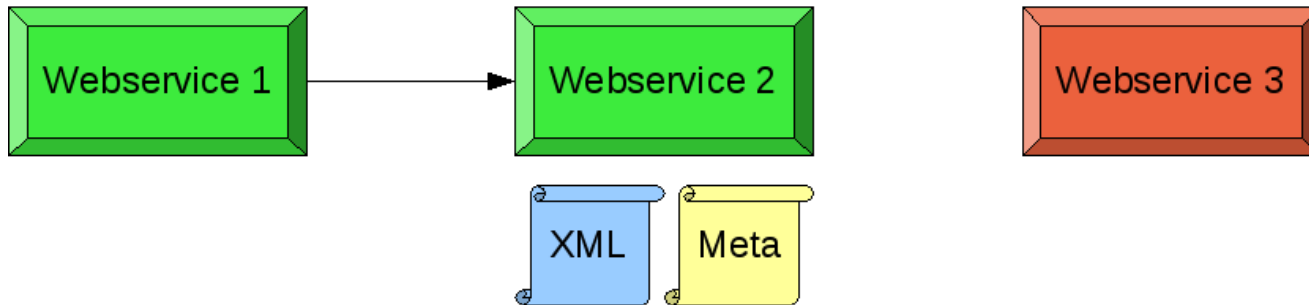
chaining - example II

input	ouput
format=tcf0.3	format=tcf0.3
lang=de	sentences=yes
text=yes	tokens=yes



chaining - example III

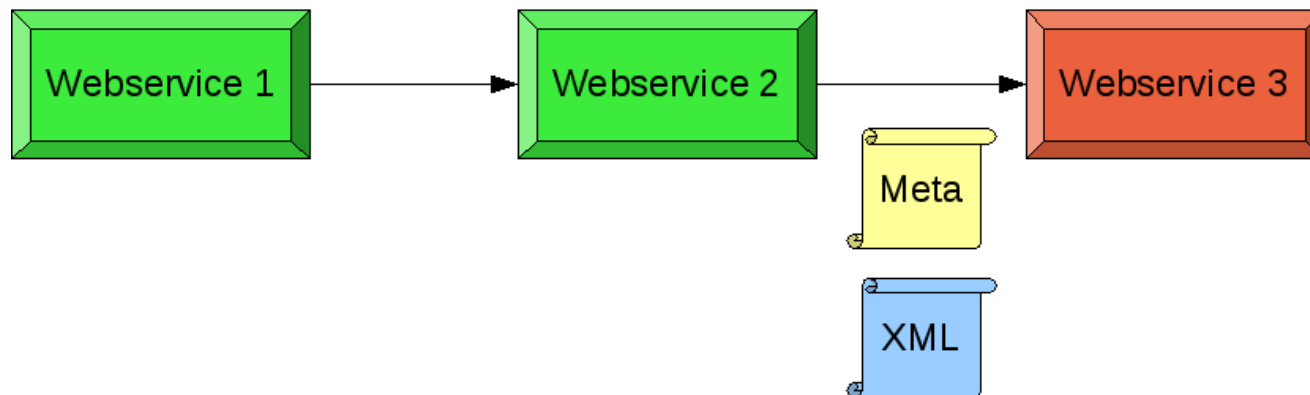
input	ouput
format=tcf0.3	format=tcf0.3
lang=de	sentences=yes
text=yes	tokens=yes



current metadata	
lang=de	tokens=yes
text=yes	
sentences=yes	

chaining - example IV

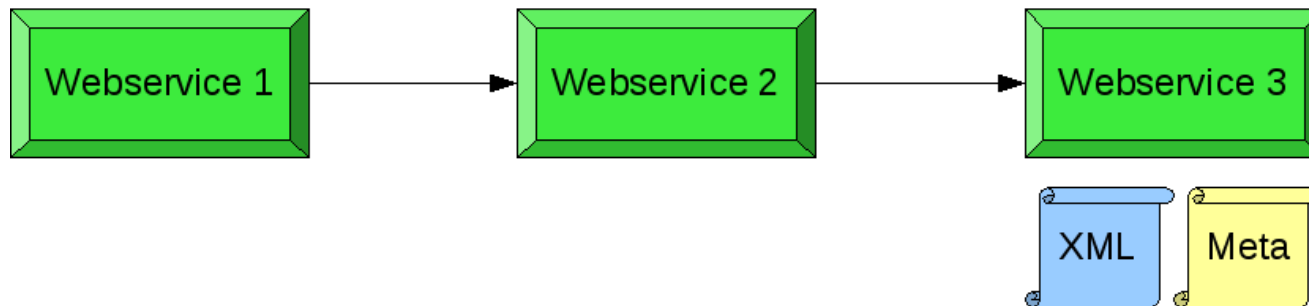
input	output
format=tcf0.3	format=tcf0.3
lang=de	pos=STTS
sentences=yes	
tokens=yes	



current metadata	
lang=de	tokens=yes
text=yes	
sentences=yes	

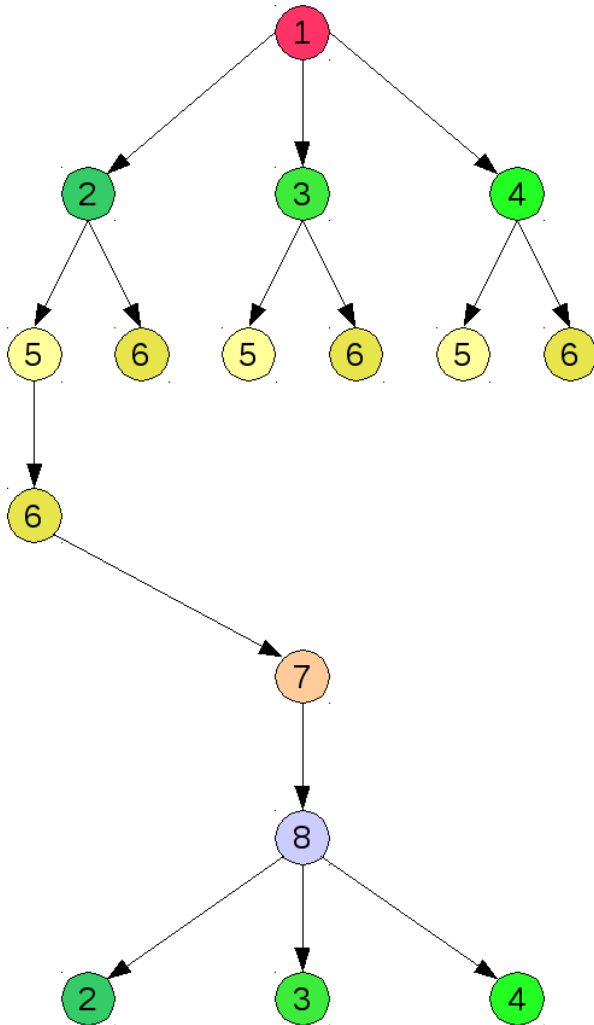
chaining - example V

input	output
format=tcf0.3	format=tcf0.3
lang=de	pos=STTS
sentences=yes	
tokens=yes	



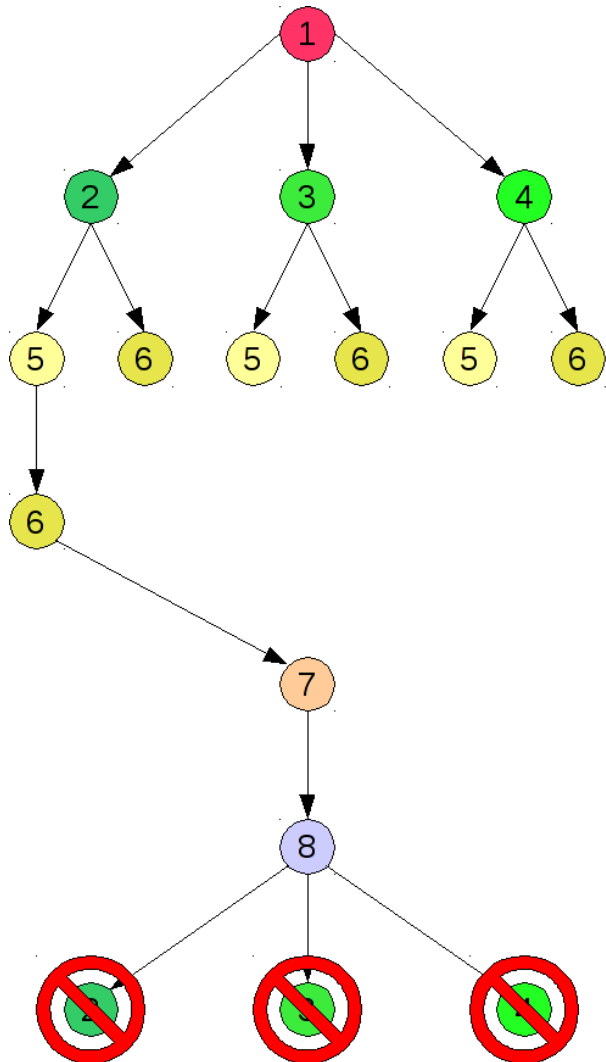
current metadata	
lang=de	tokens=yes
text=yes	pos=STTS
sentences=yes	

autochain – circles I



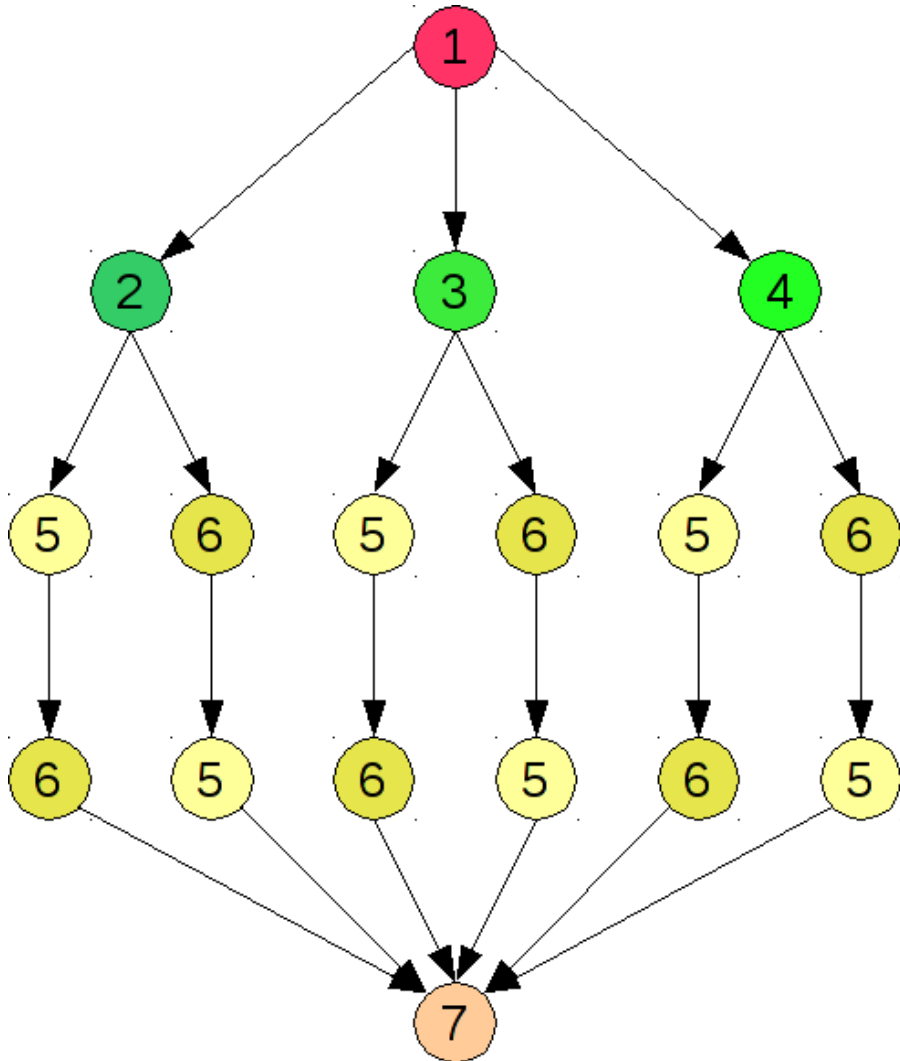
- circle: a service occurs twice in a single chain
- „circles“ can be correct, but:
 - how many iterations are right?
 - performance problem

autochain – circles II



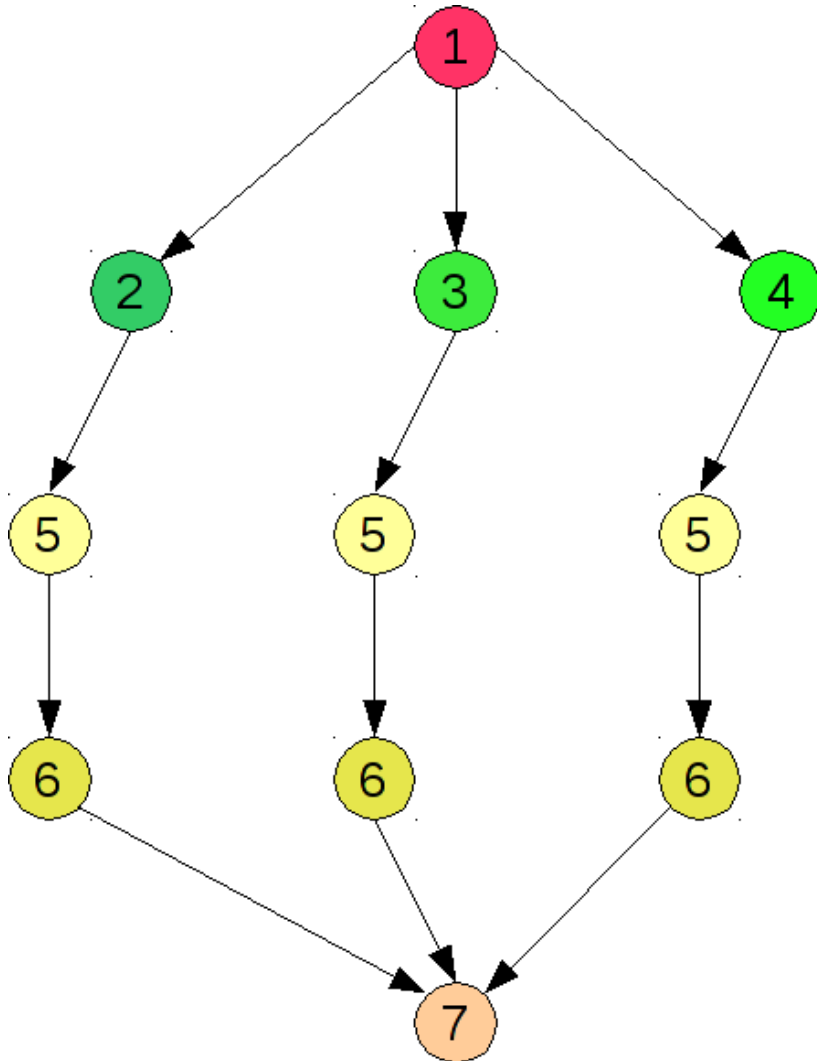
- circle: a service occurs twice in a single chain
- „circles“ can be correct, but:
 - how many iterations are right?
 - performance problem
- pragmatic solution: forbid circles

autochain – doublets I



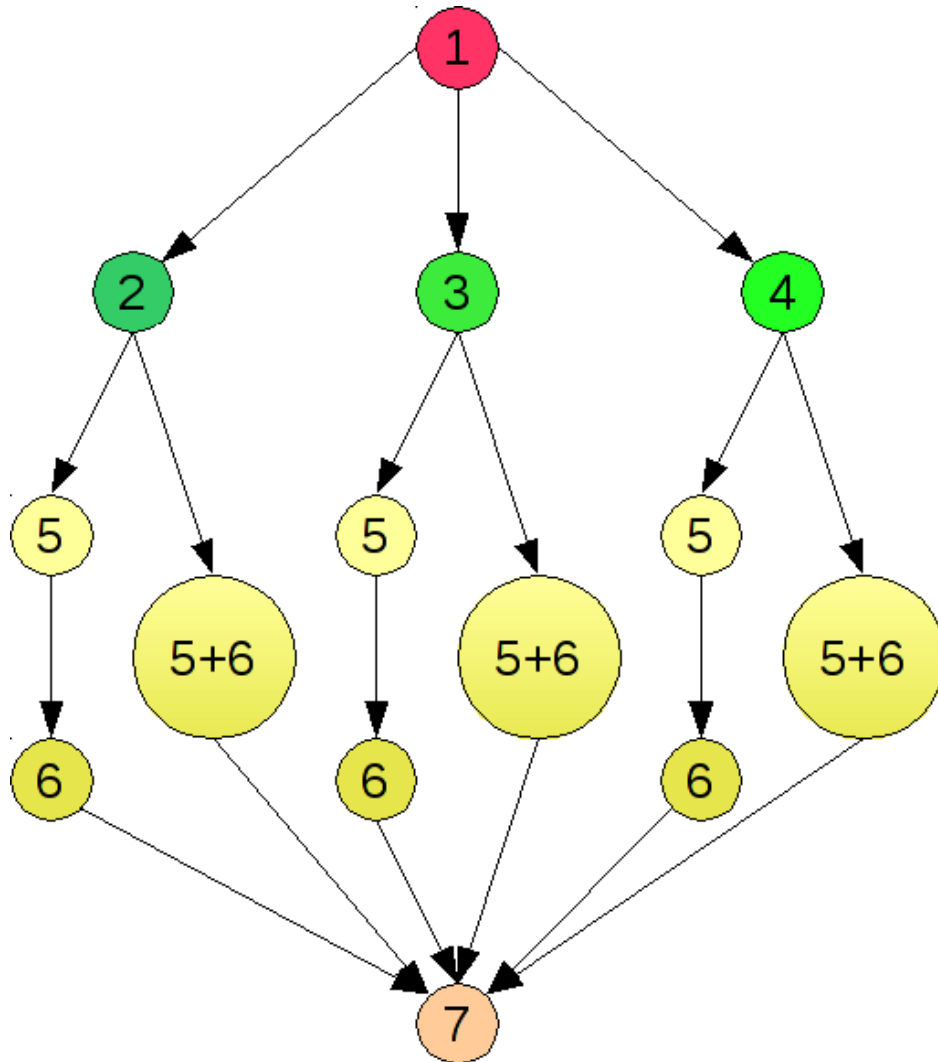
- doublet: a chain that is a reordering of another chain
- problem: too many combinations
 - performance issue
 - What is the difference?
 - How to explore?

autochain – doublets III



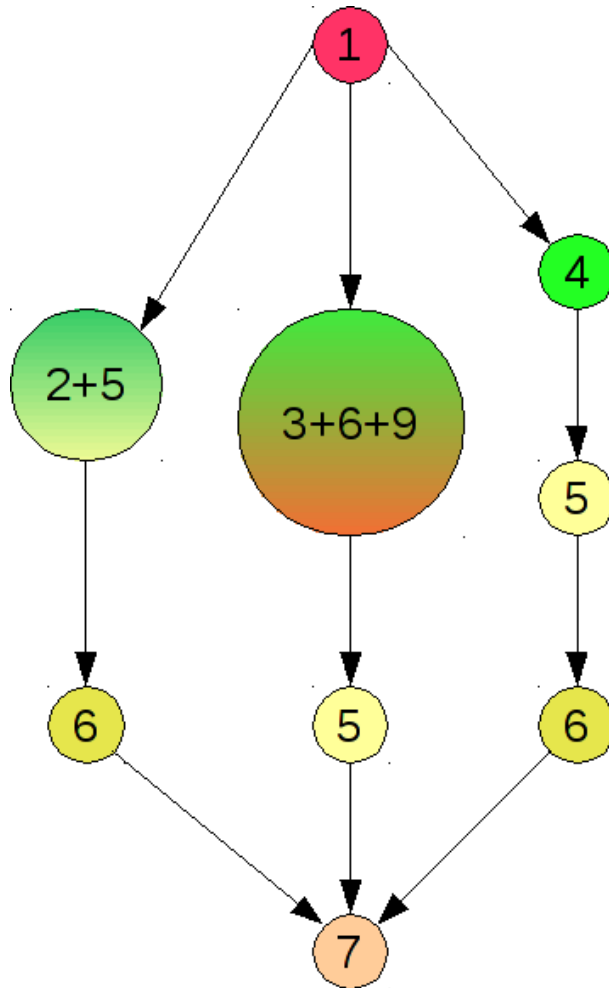
- doublet: a chain that is a reordering of another chain
- problem: too many combinations
 - performance issue
 - What is the difference?
 - How to explore?
- pragmatic solution: forbid doublets
- filter out as early as possible

autochain – aggregates I



- aggregate: a service that is composed out of number of other services
- shorter chain → better chain?

autochain – aggregates II



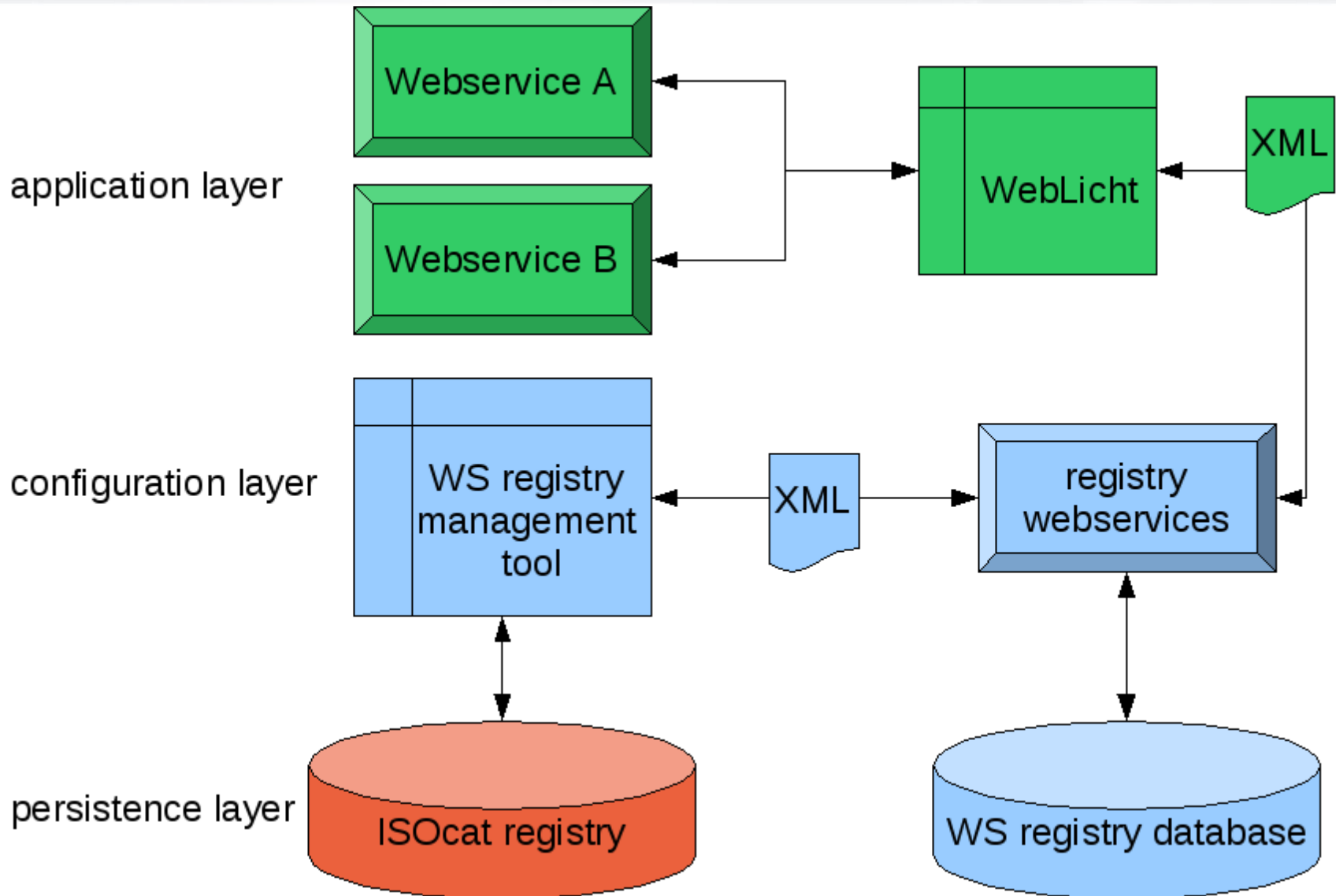
- aggregate: a service that is composed out of number of other services
- shorter chain → better chain?
- using the length of a chain as ranking criteria leads to:
 - aggr. chains are ranked higher than other ones
 - unneeded information and effort (3+6+9)
- find a better ranking criteria

DSpin prototype - chaining



- the chaining algorithm is:
 - independent from the data being computed
 - independent from format/encoding of the data (currently DSpin TextCorpus and -Lexicon services available)
- the DSpin prototype is based on
 - REST-Webservices
 - usage of HTTP-POST and HTTP-GET
- main goals:
 - solve/demonstrate the orchestration problem
 - gather experience in building a SOA-infrastructure for NLP

DSpin prototype - architecture



DSpin prototype - services



- resources:
 - wortschatz-webservices (baseform, frequencies, cooccurrences, sentences, ...) for german
 - GermanNet (Tübingen)
 - TüBa/DZ (german treebank)
- tools:
 - different tokenizers and POS-taggers (Leipzig, Tübingen, Stuttgart, BBAW)
 - semantic annotator, constituent parser, morphological analyzer, sentence segmentation, lemmatization, cooccurrence annotation, named entity- and person name recognition
- mainly german DSpin partners, but also from Finland (Helsinki), Romania (RACAI), ...

DSpin prototype - WebLicht



WebLicht: Web-Based Linguistic Chaining Tool

Tool Filters Language: TCF Version:

Name	Creator	Lang	Version
Semantic Annotator	SfS: Uni-Tuebingen	de	0.2
BBAW Person Name Rec...	BBAW	de	0.2
BBAW Tagger	BBAW	de	0.3
ULei - Frequency	ASV Universiaet Leip...	de	
ULei - Baseform	ASV Universiaet Leip...	de	
POS Tagger - TübaDZ	SfS: Uni-Tuebingen	de	0.3
Microsoft Word Conve...	SfS: Uni-Tuebingen	de	0.3
ULEI - TextCorpus2Le...	ASV Universiaet Leip...	de	0.3
TreeTagger	IMS: Uni-Stuttgart	it	0.3
Plaintext Converter	SfS: Uni-Tuebingen	en	0.3
Plaintext Converter	BBAW: Berlin	de	0.2
Tokenizer	IMS: Uni-Stuttgart	it	0.3
Tokenizer	GL: Uni-Helsinki	fi	0.3
TreeTagger	IMS: Uni-Stuttgart	en	0.3
POS Tagger	IMS: Uni-Stuttgart	de	0.3
Morph analyzer	GL: Uni-Helsinki	fi	0.3
POS Tagger - TübaDZ	SfS: Uni-Tuebingen	de	0.2
ULei - Tokenizer - d...	ASV Universiaet Leip...	de	0.3
Tokenizer	IMS: Uni-Stuttgart	de	0.3
ULei - Cooccurrences...	ASV Universiaet Leip...	de	
RTF Converter	SfS: Uni-Tuebingen	fr	0.3
RTF Converter	SfS: Uni-Tuebingen	en	0.3
TreeTagger	IMS: Uni-Stuttgart	fr	0.3
Constituent Parser	IMS: Uni-Stuttgart	en	0.3
Frequency - deutsch	ASV Universitaet Lei...	de	0.2
RTF Converter	SfS: Uni-Tuebingen	de	0.3
Microsoft Word Conve...	SfS: Uni-Tuebingen	it	0.3

Build Chain

Next Tool Choices:

Name	Creator	Lang	Versio
Constituent Parser	IMS: Uni-Stuttgart	de	0.3
Semantic Annotator	SfS: Uni-Tuebingen	de	0.3
ULEI - TextCorpus2Le...	ASV Universiaet Leip...	de	0.3

Add »
Clear
Run

Selected Tools:

Name	Creator	Lang	Versio
Plaintext Converter	SfS: Uni-Tuebingen	de	0.3
ULei - Tokenizer - d...	ASV Universiaet Leip...	de	0.3
POS Tagger	IMS: Uni-Stuttgart	de	0.3
BBAW Person Name Rec...	BBAW	de	0.3

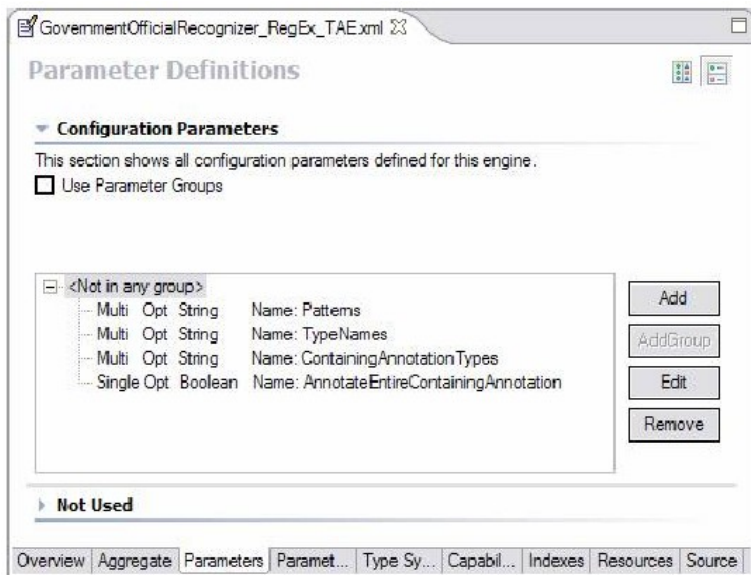
Results

Input

Karin fliegt nach New York. Sie will dort Urlaub machen.

Edit the text and click the OK button. Uploaded files may not be edited. To edit text after a file selection, choose 'Enter Plain Text' from the Input menu.

CA9E90165CC4F46DE1332FC347791F08



- current WS in DSpin-prototype similar to UIMA-annotators
- similar input/output specifications
- UIMA defines a container (CAS) and provides a framework
 - read annotations of previously executed annotators
 - add annotations, build/specify aggregates, specify flows, ...
- DSpin chaining algorithm:
 - service description: format + input/output
 - WS can be based on established formats
- automatic chain builder for UIMA?
 - mandatory/optional parameters
 - service repository + namespace for parameters

CLARIN

Common Language Resources and Technology Infrastructure



Thank you for your attention

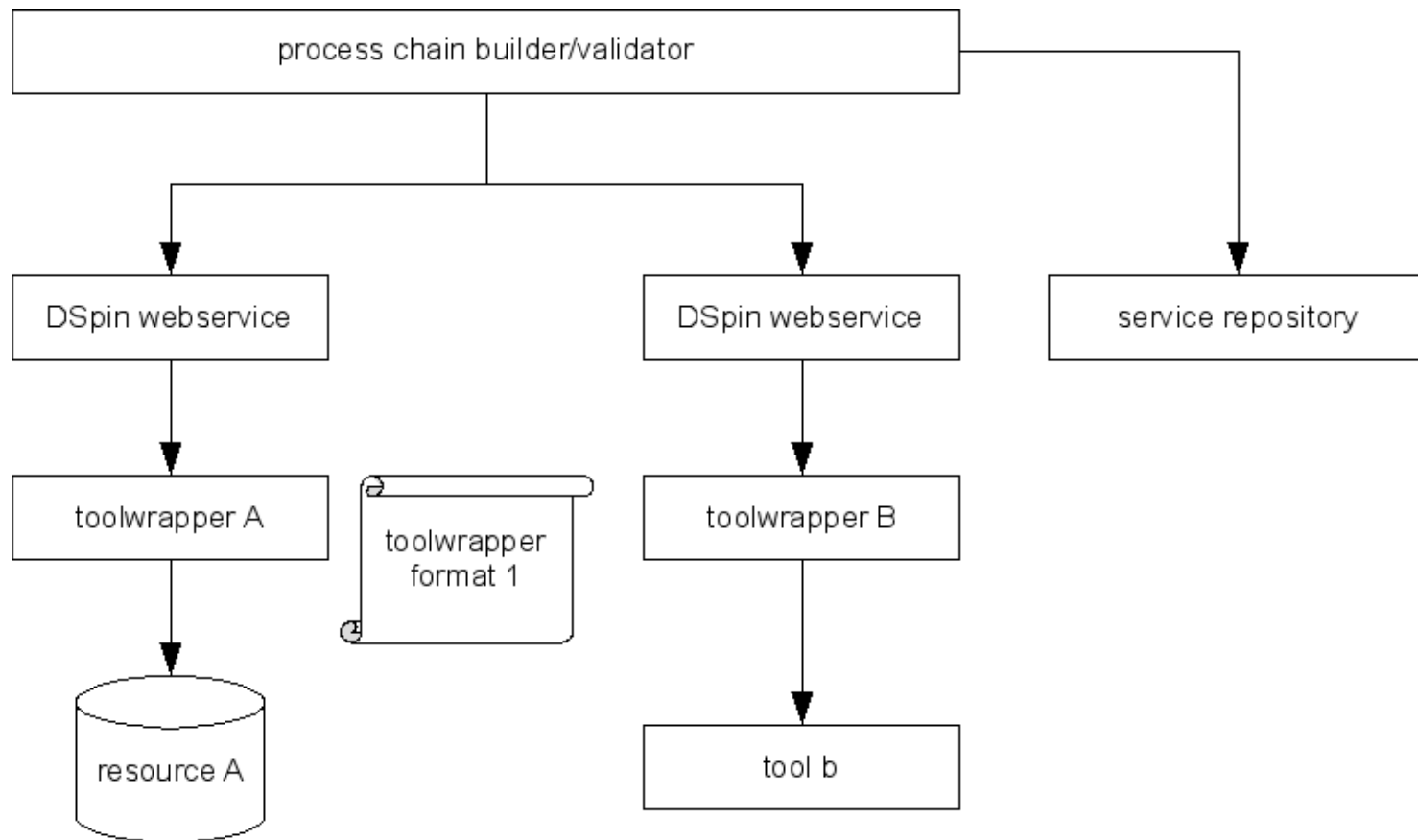
CLARIN has received funding from
the European Community's Seventh Framework Programme
under grant agreement n° 212230

Dspin prototype - milestones



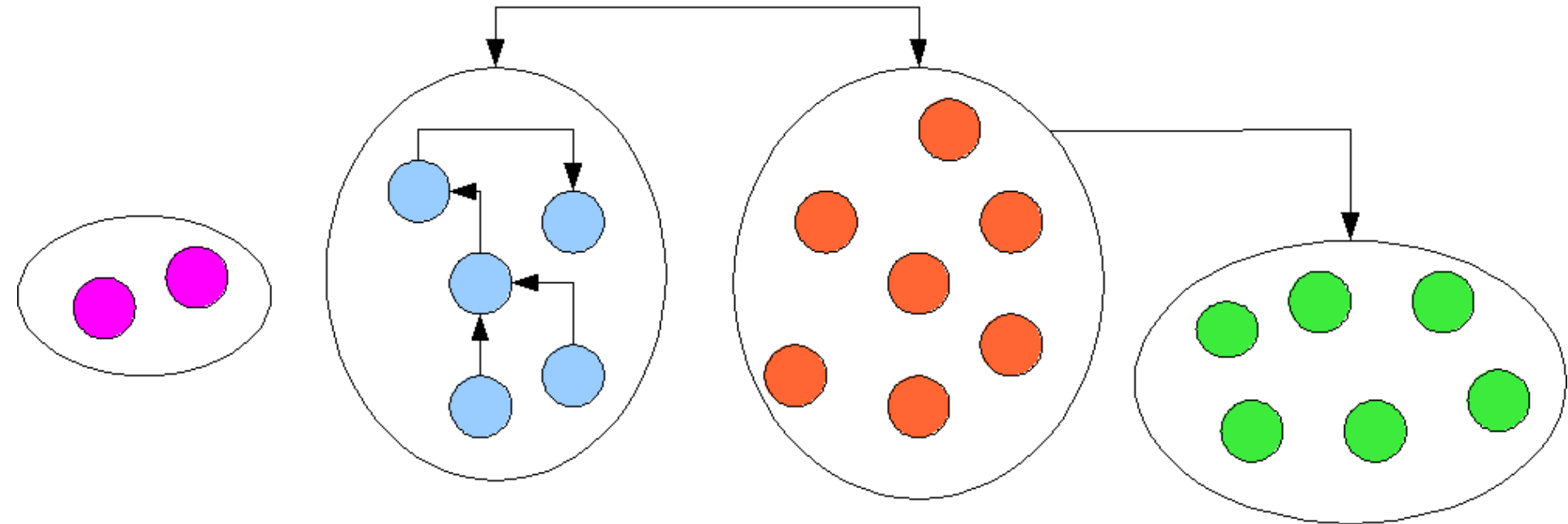
- done:
 - simple registry for webservices: url description, contact, + service description (format, input/output)
 - implementation of a registry management tool (easy registration of services)
 - implementation and registration of various services from different partners
 - implementation of the chaining algorithm
 - first successfull tests in two „workflow“ tools: Tübingen, Leipzig
 - implemenation of the automatic chain building algorithm
- work in progress:
 - impl. of harvesting-interface (OAI-PMH) for Clarin (WS metadata component)
 - test/integration of a QoS component for WS
 - save/load service chains
 - Maybe:
 - implementation of the registry management tool in GWT
 - usage of the ISOcat datacategory registry
 - contribution to the Clarin European Demonstrator

DSpin prototype - concept



chaining & service bubbles

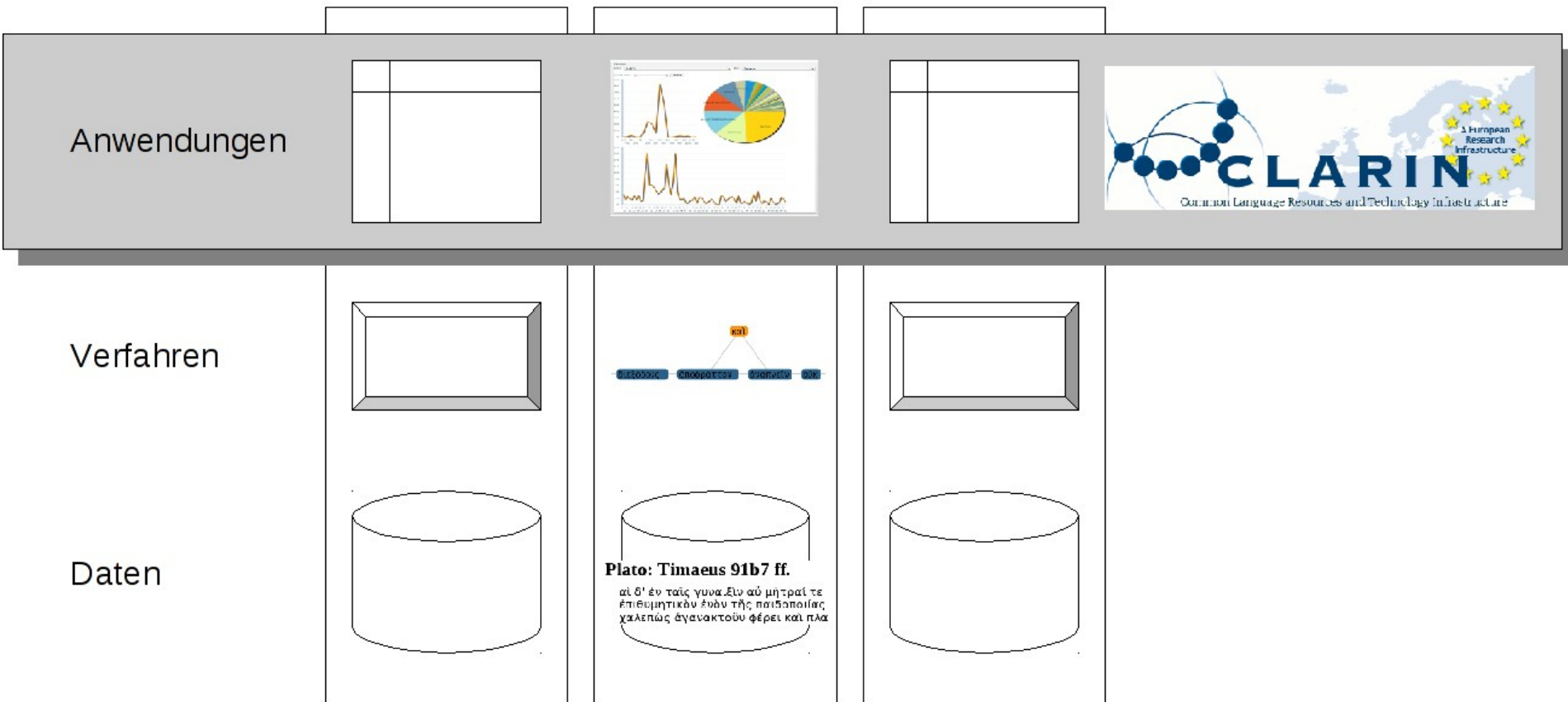
converter/transformer repository (web 2.0; automatic inference, ...)



eHumanities & integration I

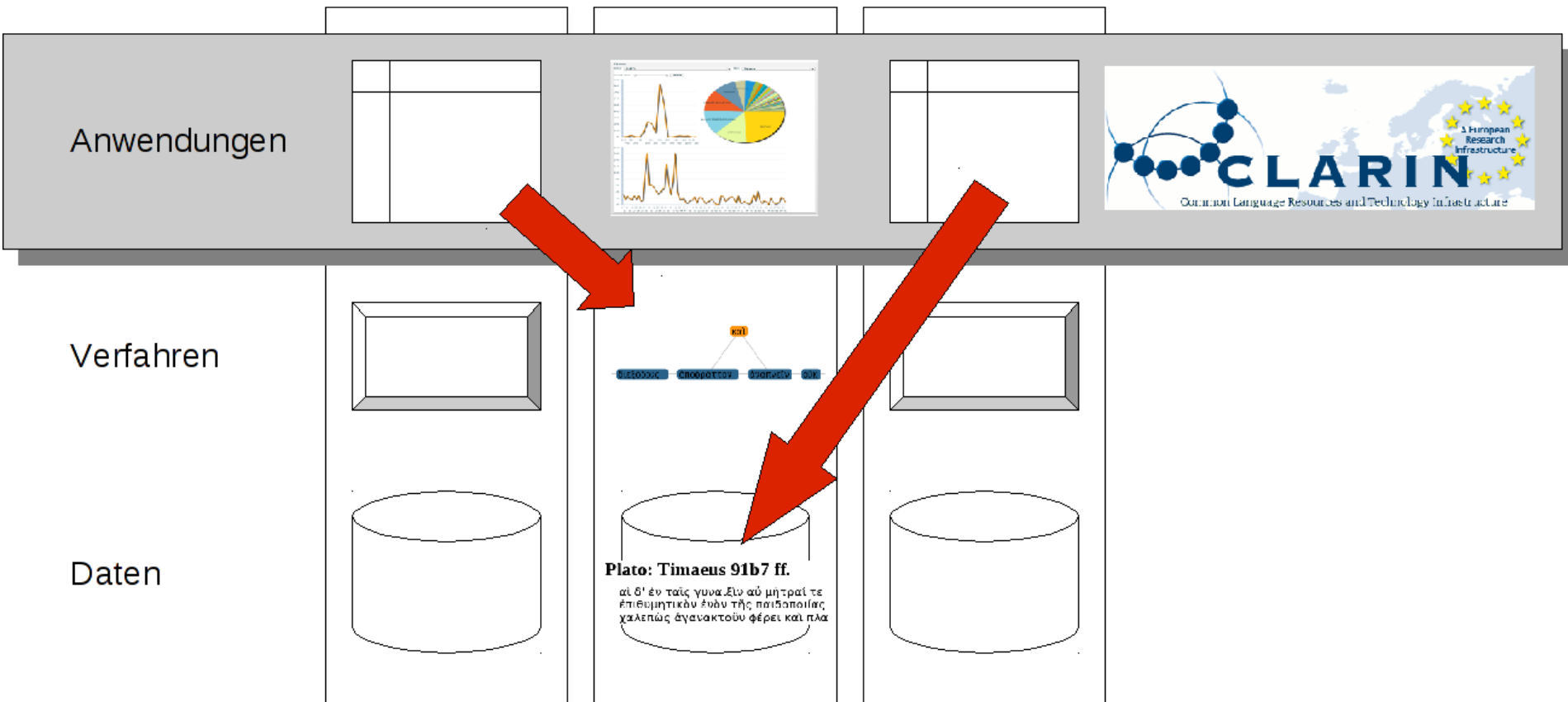
- integration of eHumanities projects

eAQUA



eHumanities & integration II

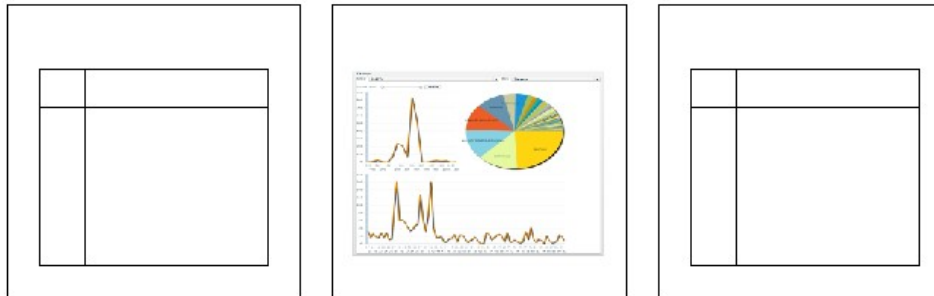
eAQUA



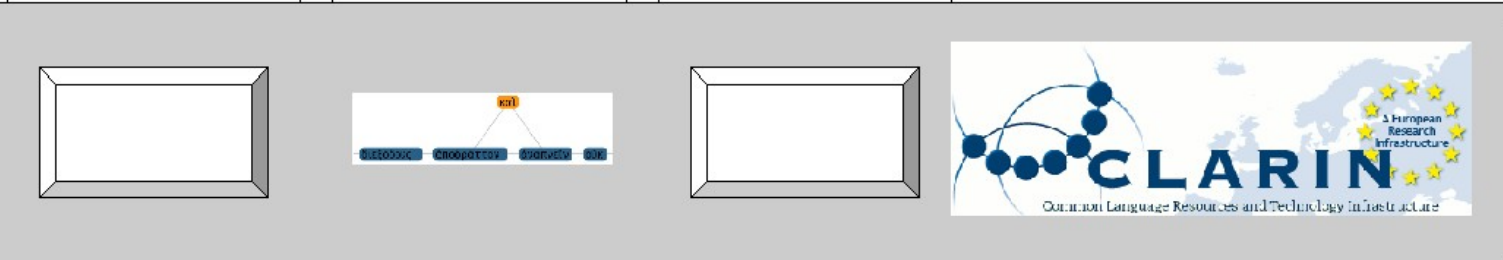
eHumanities & integration III

eAQUA

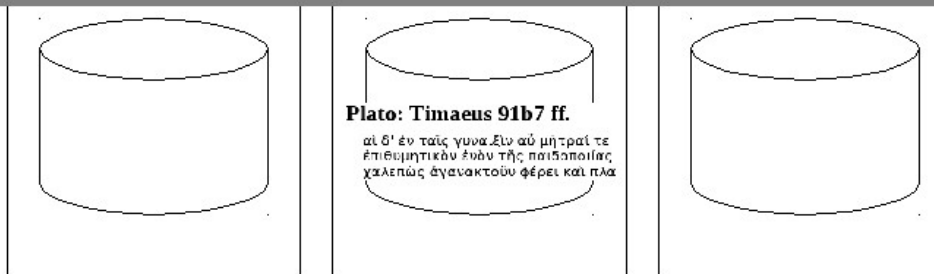
Anwendungen



Verfahren



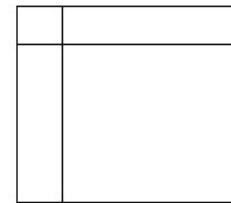
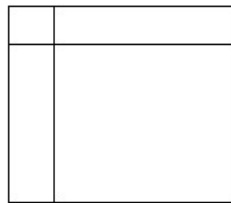
Daten



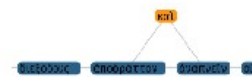
eHumanities & integration IV

eAQUA

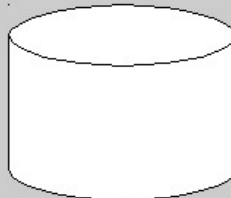
Anwendungen



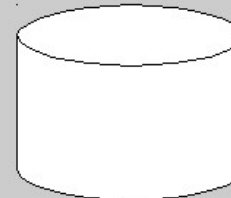
Verfahren



Daten



Plato: Timaeus 91b7 ff.
αἱ δ' ἐν ταῖς γυναῖξιν αὐτῆσιν τε
ἐπιθυμητικόν ἐκόν τῆς παιδαγωγίας
χαλεπῶς ἀναγκαῖον φέροι καὶ πλάττειν



registry management tool I



ASV DSpin Registry Management Tool v0.01a

- add service
- browse registry**
- manage formats
- chain services
- direct call
- log

Id	Name	Description
47	TEST TCF0.2 converter (deutsch)	
48	TEST Tokenizer-Stuttgart 0.2 (deutsch)	test
49	TEST Tagger-Stuttgart 0.2 (deutsch)	test
50	TEST Tokenizer Leipzig (deutsch)	test
53	TEST Query Wortschatz	test
58	Tokenizer (IMS,TCF0.2,deutsch)	Tokenizer for German text.

Id: Name:

ShortDescription:

Url:

Description:

Creator:

Contact:

InputWrapper: OutputWrapper:

Name	Standard
Text	Utf8
Sentence	Utf8
Language	German
Token	Utf8

Name	Standard
POS	STTS
Lemma	Utf8

Password:

registry management tool II



ASV DSpin Registry Management Tool v0.01a

- add service
- browse registry
- manage formats
- chain services**
- direct call
- log

Id	Name	Id	Name
		47	TEST TCF0.2 converter (deutsch)
		53	TEST Query Wortschatz
		94	Plaintext Converter (Sfs,TCF0.2,deutsch)
		95	Plaintext Converter (BBAW,TCF0.2,deutsch)
		96	Plaintext Converter (Sfs,TCF0.3,deutsch)
		97	Plaintext Converter (Sfs,TCF0.3,english)
		98	RTF Converter (Sfs,TCF0.3,deutsch)
		99	RTF Converter (Sfs,TCF0.3,english)
		100	PDF Converter (Sfs,TCF0.3,deutsch)
		101	PDF Converter (Sfs,TCF0.3,english)
		102	Microsoft Word Converter (Sfs,TCF0.3,deutsch)
		103	Microsoft Word Converter (Sfs,TCF0.3,english)
		104	Negra Converter (Sfs,TCF0.3,deutsch)

clear chain execute chain

service	result size	time
---------	-------------	------

save

registry management tool III



WSP2010

Valetta

2010-05-18

www.clarin.eu

ASV DSpin Registry Management Tool v0.01a

- add service
- browse registry
- manage formats
- chain services**
- direct call
- log

Id	Name
47	TEST TCF0.2 converter (deutsch)

Id	Name
48	TEST Tokenizer-Stuttgart 0.2 (deutsch)
50	TEST Tokenizer Leipzig (deutsch)

clear chain

execute chain| service | result size | time |
| --- | --- | --- |

save

registry management tool IV



ASV DSpin Registry Management Tool v0.01a

- add service
- browse registry
- manage formats
- chain services**
- direct call
- log

Id	Name
47	TEST TCF0.2 converter (deutsch)
49	TEST Tagger-Stuttgart 0.2 (deutsch)
50	TEST Tokenizer Leipzig (deutsch)

Id	Name
48	TEST Tokenizer-Stuttgart 0.2 (deutsch)

clear chain

execute chain

service	result size	time
TEST TCF0.2 converter (deutsch)	6 KB	1293
TEST Tokenizer Leipzig (deutsch)	58 KB	331
TEST Tagger-Stuttgart 0.2 (deutsch)	122 KB	1202

save

Clarín – resources & tools



http://www.clarin.eu/view_resources

Name▲	Languages	Type	Description	Country	Institute (not a CLARIN member)	Finalization year	Distribution Type
Aligner 2.0.6.7	-- language not in list --	Application / Tool	A language-independent tag-oriented semi-automatic paragraph and sentence aligner. Works on MS Windows. Produces XML valid documents. Allows recording detailed bibliographical information. It has been used for creating English-Lithuanian Parallel corpus.	Lithuania		2007	
alinea	Catalan English Spanish	Application / Tool	A tool for parallelizing translated texts, which has been specially designed for specialized corpora and also	Spain			