

# TTNWW

TST tools voor het Nederlands als Webservices in een Workflow

Marc Kemps-Snijders

Computational Linguistics in the Netherlands  
June 25<sup>th</sup>, Nijmegen

*Marc.kemps.snijders@meertens.knaw.nl*



## CLARIN-NL in a nutshell



The CLARIN-NL project is a large national project in the Netherlands (2009-2015) which aims to make a significant contribution to the European CLARIN infrastructure.

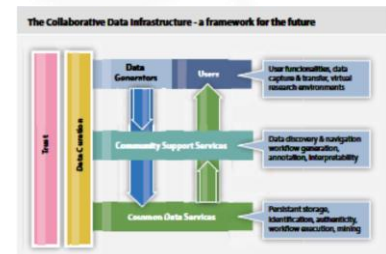
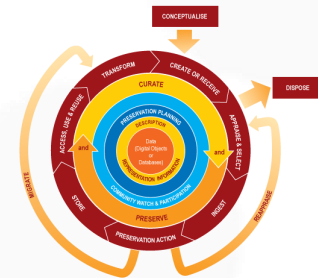
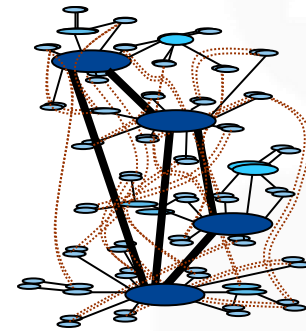
The CLARIN infrastructure is a **research infrastructure** intended for humanities researchers that work with language data and tools. The researcher will be able to find data and tools relevant for his/her research via the CLARIN infrastructure and get access to them. We also aim to make it possible to apply tools to data in such a way that no technical background is needed or ad-hoc adaptations to the tools or the data are necessary.

The researcher's research project may result in new **data and tools**, which he/she can make available to other researchers via the CLARIN infrastructure. The CLARIN research infrastructure will thus make humanities research related to language easier, faster, better, and in some cases even possible for the first time.

<http://www.clarin.nl>

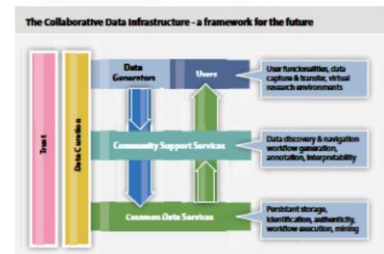
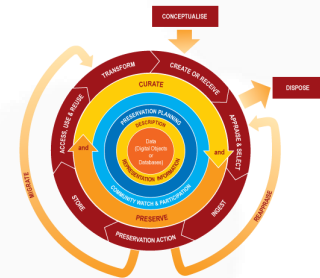
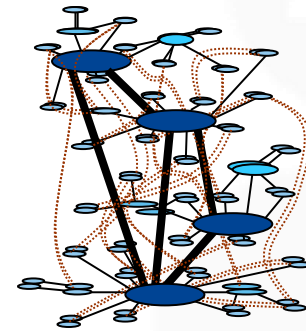
# VIEWS ON INFRASTRUCTURES

- Centers view
  - How do organizations fit in?
- Data lifecycle view
  - What about the data?
- Architectural view
  - Tools and services



# VIEWS ON INFRASTRUCTURES

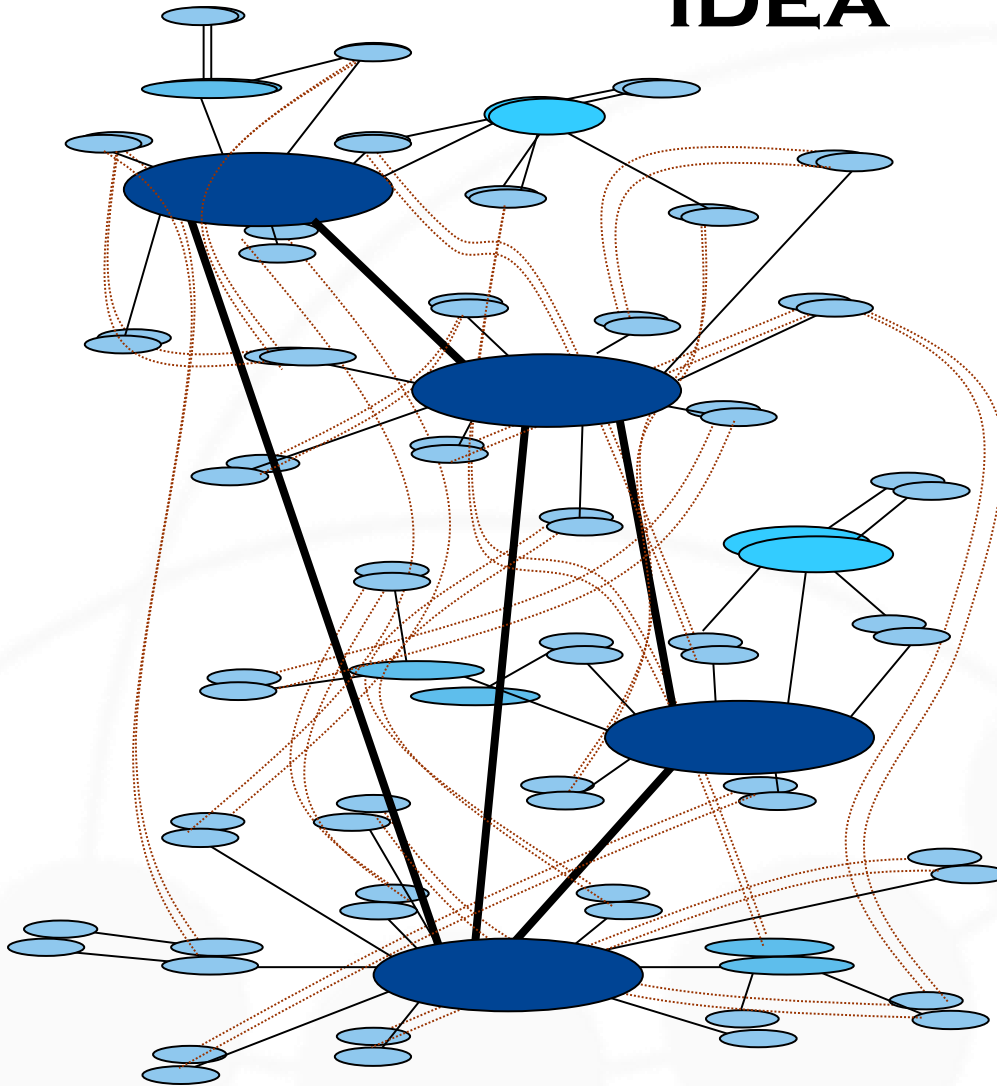
- **Centers view**
  - How do organizations fit in?
- Data lifecycle view
  - What about the data?
- Architectural view
  - Tools and services





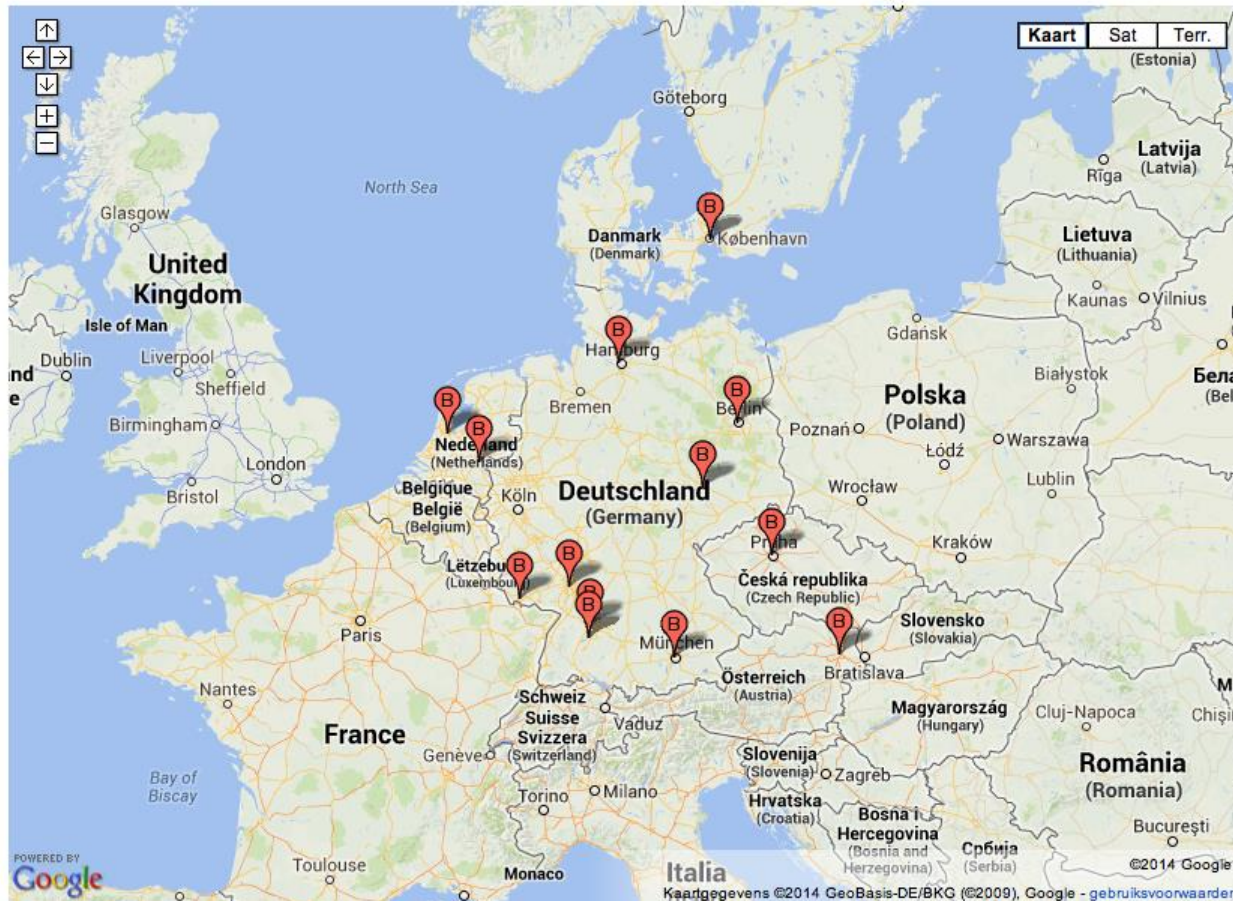
# ORIGINAL CLARIN CENTER

## IDEA



Scenario where dedicated services centres of new type interact in a stable way and give persistent and easy-to-use services to the community.  
Scenario characterized mainly by accidental and temporary interactions  
Researchers must be able to rely on the services offered

# CLARIN CENTERS 2014



The following centres have successfully passed the CLARIN centre assessment procedure and now can officially carry the CLARIN B centre label:

- ASV Leipzig
- Bayerisches Archiv für Sprachsignale
- Berlin-Brandenburg Academy of Sciences and Humanities
- CLARIN Center Vienna
- Eberhard Karls Universität Tübingen
- Hamburger Zentrum für Sprachkorpora (HZSK)
- IMS, Universität Stuttgart
- Institut für Deutsche Sprache
- LINDAT-Clarín
- Meertens Instituut
- MPI for Psycholinguistics
- The Clarin center at University of Copenhagen
- Universität des Saarlandes

# MAKING DATA AND TOOLS AVAILABLE

Virtual I

From here y



## Virtual Language Observatory

Explore the world of language resources and technology from different perspectives

[Link](#) | [Report](#) | [Help](#)

[VLO](#) > [Faceted search](#) > [Search: "dutch"](#)

### SEARCH

### SEARCH RESULTS

32431 results Showing 1 to 10

<< < 1 2 3 4 5 6 7 8 9 10 >> >>

**"Al zijn we katholiek, we zijn Nederlanders" : opkomst en verval van de katholieke kerk in Nederland sinds 1795 vanuit rational choice perspectief** Expand

The Dutch Catholic Church has experienced an immense growth and an equal measure of decline in the last two centuries. Four phases can be discerned: slow growth 1795-1870; fast growth 1870-1960; early decline 1920-1960; decline 1960-. Existing explanations for these developments seem not fully adequate: they either exp...

Resources: | 1 image | 22 text documents | 1 other |

**"De stormen van natuur en vrindschap tegenstaan!" J. Nomsz, Anthonius Hambroek, of de Belegering van Formoza. Treurspel. (1775)** Expand

This article shows - starting from A. de Haas' boek on the Dutch theatre in de 18th century - how Joannes Nomsz in his drama Anthonius Hambroek, of de Belegering van Formoza (1775) tries to combine the conventions of the French classicist dramatical theory with new ideas, especially on the national matter and the poeti...

Resources: | 1 image |

**"He is the best soccer player of this world" An exploration of the language use of** Expand

**Note: Information is dynamically harvested from centers**

#### NARROW DOWN

Use the categories below to limit the search results to those matching the selected value(s).

- + LANGUAGE
- + COLLECTION
- + RESOURCE TYPE
- + CONTINENT
- + COUNTRY
- + MODALITY
- + GENRE
- + SUBJECT
- + FORMAT
- + ORGANISATION
- + NATIONAL PROJECT
- + KEYWORD
- + DATA PROVIDER

[Expand all +](#)  
[Collapse all -](#)

Virtual I

From here y



## Virtual Language Observatory

Explore the world of language resources and technology from different perspectives

[Link](#) | [Report](#) | [Help](#)

[VLO](#) > [Faceted search](#) > [Search: "dutch"](#)

### SEARCH

### SEARCH RESULTS

32431 results Showing 1 to 10

<< < 1 2 3 4 5 6 7 8 9 10 >> >>

**"Al zijn we katholiek, we zijn Nederlanders" : opkomst en verval van de katholieke kerk in Nederland sinds 1795 vanuit rational choice perspectief** Expand

The Dutch Catholic Church has experienced an immense growth and an equal measure of decline in the last two centuries. Four phases can be discerned: slow growth 1795-1870; fast growth 1870-1960; early decline 1920-1960; decline 1960-. Existing explanations for these developments seem not fully adequate: they either exp...

Resources: | 1 image | 22 text documents | 1 other |

**"De stormen van natuur en vrindschap tegenstaan!" J. Nomsz, Anthonius Hambroek, of de Belegering van Formoza. Treurspel. (1775)** Expand

This article shows - starting from A. de Haas' boek on the Dutch theatre in de 18th century - how Joannes Nomsz in his drama Anthonius Hambroek, of de Belegering van Formoza (1775) tries to combine the conventions of the French classicist dramatical theory with new ideas, especially on the national matter and the poeti...

Resources: | 1 image |

**"He is the best soccer player of this world" An exploration of the language use of** Expand

**Note: Information is dynamically harvested from centers**

Expand all +  
Collapse all -

#### NARROW DOWN

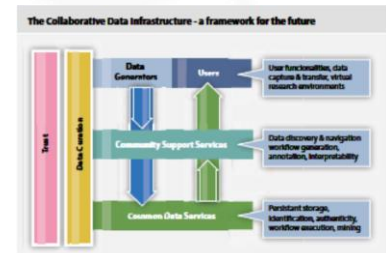
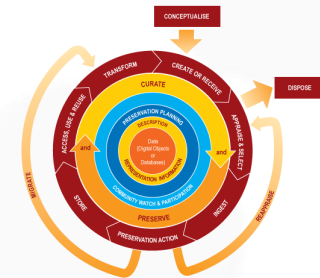
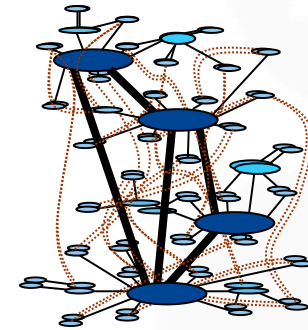
Use the categories below to limit the search results to those matching the selected value(s).

- + LANGUAGE
- + COLLECTION
- + RESOURCE TYPE
- + CONTINENT
- + COUNTRY
- + MODALITY
- + GENRE
- + SUBJECT
- + FORMAT
- + ORGANISATION
- + NATIONAL PROJECT
- + KEYWORD
- + DATA PROVIDER



# VIEWS ON INFRASTRUCTURES

- Centers view
  - How do organizations fit in?
- **Data lifecycle view**
  - What about the data?
- Architectural view
  - Tools and services



# DATA MANAGEMENT LIFE CYCLE

nodorlab

Commonly used Metadata Standards for encoding Archive Collections

Sum # metadata records		
standard	Content Provider	Totaal
? (no answer)	VGA	14621
<b>Total ? (no answer)</b>		<b>14621</b>
Dublin Core	KEE/OSA	289
	UIPIP(BDIC)	872
<b>Total Dublin Core</b>		<b>1161</b>
EAD 2002	CGIL	30696
	Génériques	5324
	KNAW-IISG	4292
<b>Total EAD 2002</b>		<b>40312</b>
idiosyncratic	Amsab-ISG	6500
	FES Archive	183312
	FMS	31394
	KEE/OSA	125783
	TA	10727
<b>Total idiosyncratic</b>		<b>357716</b>
<b>Total:</b>		<b>413810</b>

3,53 %

0,28 %

9,74 %

86,44 %  
archive' metadata records



You are here



You are here



**CMDI**  
Component Metadata Infrastructure

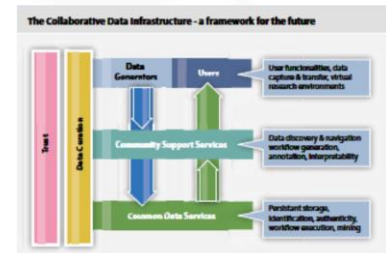
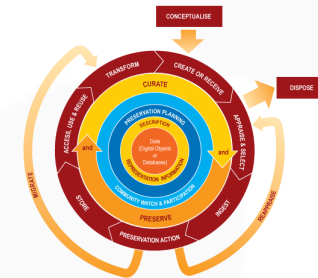
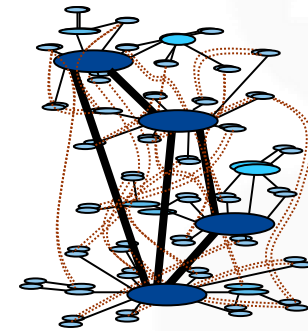
10744/mi\_a19eb1ae-8e90-4cd6

Naming authority

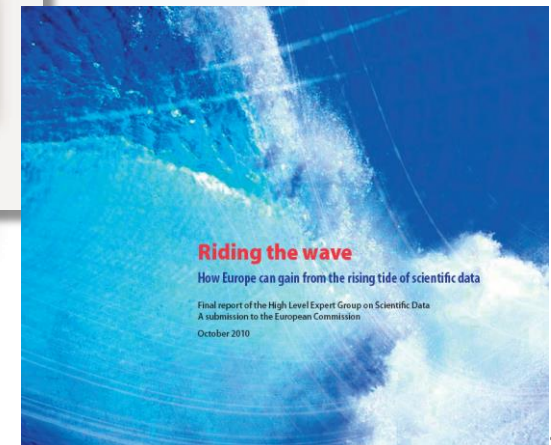
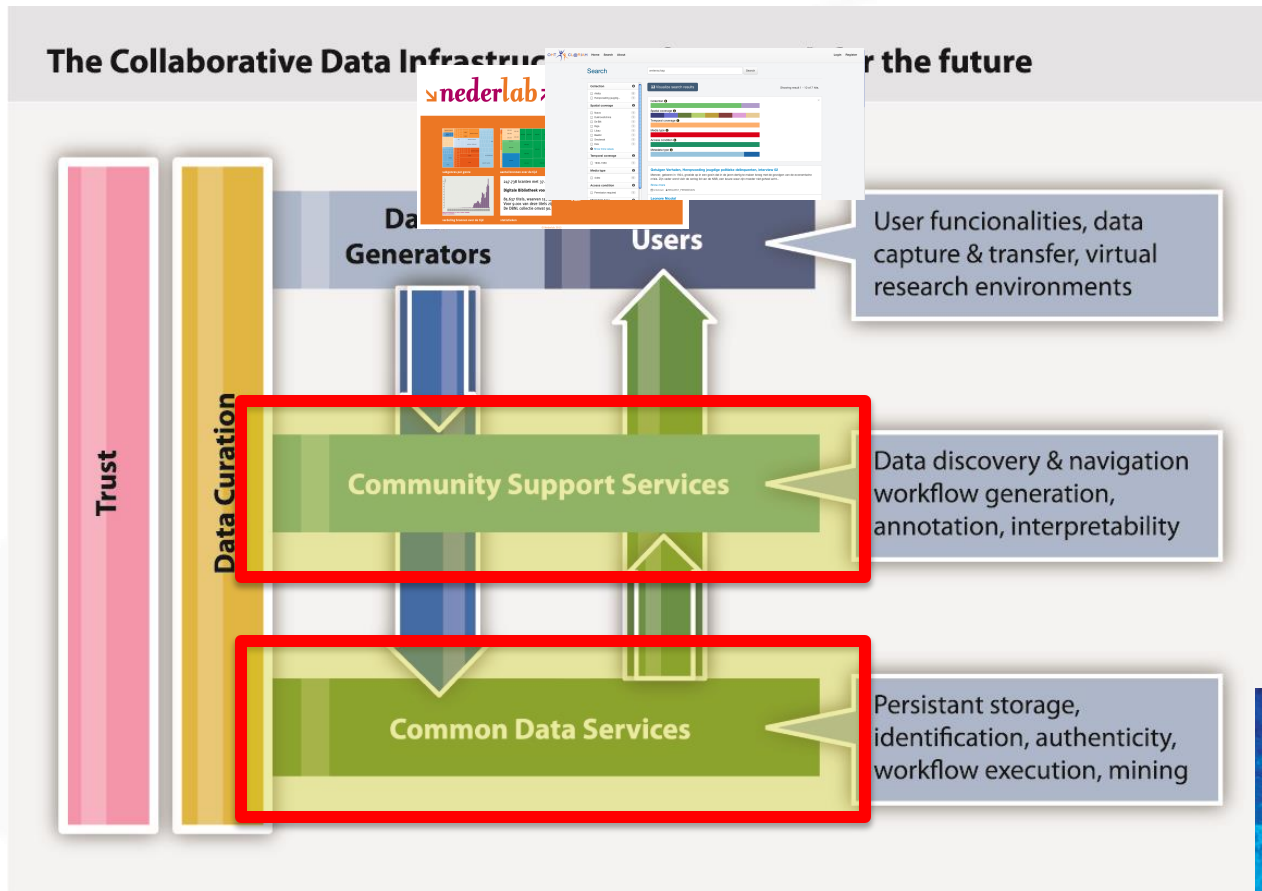
Suffix

# VIEWS ON INFRASTRUCTURES

- Centers view
  - How do organizations fit in?
- Data lifecycle view
  - What about the data?
- **Architectural view**
  - **Tools and services**



# TOOLS AND SERVICES

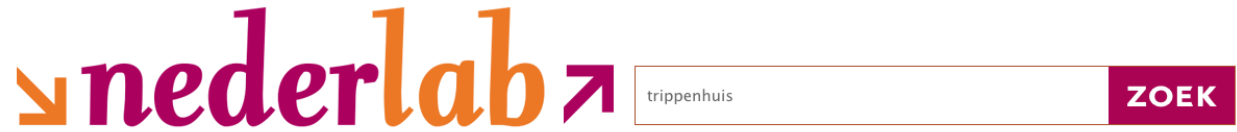


- A two-year program with the intention to make **existing components** from previous project (a.o. CGN and STEVIN) available **as web services** and combine them into **workflows** for use by **HSS researchers** with little or no technical background. This enable them to:
  - More easily answer research questoins
  - Provide a basis to formulate new research questions

# WHAT DO WE PROVIDE?

- **Text**
  - Make several basic text analysis components available:
    - Corpus cleanup
    - Part of Speech tagging
    - Named Entity Recognition
    - Semantic role labelling
    - Coreference
- **Speech**
  - Provide components for speech recognition

# WHAT'S THE USE?



- For texts m... techniques
  - E.g. extra...
  - Analysis c... possible

Direct naar: [Uitgebreid zoeken](#) [Zoeken op woordsoort en -kenmerk](#) [Collectie kiezen](#)

statistieken

Momenteel omvat Nederlab 37.509.003 documenten met in totaal 1.277.188.758 woorden.

De volgende collecties zijn beschikbaar gemaakt:

**KB kranten (1900 - 1940)**  
247.738 kranten met 37.009.183 artikelen.

**Digitale Bibliotheek voor de Nederlandse Letteren**  
85.637 titels, waarvan 157.444 onzelfstandige titels zijn afgesplitst. Voor 9.001 van deze titels zijn teksten beschikbaar. De DBNL collectie omvat 90.706 auteurs-beschrijvingen.

subgenres per genre

aantal bronnen

247.738

OHT CL@RIAH Home Search About

Search wetenschap Search

Visualize search results Showing result 1 - 10 of 7 hits.

Collection

- Aletta
- Heropvoeding jeugd...

Spatial coverage

- Narva
- Dikhovshchina
- De Bit
- Riga
- Libau
- Beelitz
- Smolensk
- Kiev
- [Show more values](#)

Temporal coverage

- 1930-1950

Media type

- video

Access condition

- Permission required

Metadata type

Getuigen Verhalen, Heropvoeding jeugdige politieke delinquenten, interview 02

Meneer, geboren in 1924, groeide op in een gezin dat in de jaren dertig te maken kreeg met de gevolgen van de economische crisis. Zijn vader werd vóór de oorlog lid van de NSB, een keuze waar zijn moeder niet geheel acht...

[Show more](#)

Unknown REQUEST\_PERMISSION

Leonore Nicolai

- For speech it... transcribe sp... available for



# WEB SERVICE DEFINITION



There are many things that might be called "Web services" in the world at large. *However, for the purpose of this Working Group and this architecture, and* without prejudice toward other definitions, we will use the following definition:

A Web service is a software system designed to support interoperable machine-to-machine interaction over a network. *It has an interface described in a machine-processable format (specifically WSDL).* Other systems interact with the Web service in a manner prescribed by its description *using SOAP-messages, typically conveyed using HTTP with an XML serialization* in conjunction with other Web-related standards.



**Frog**  
*mijnProject*

---

**Status**

Accepting new input files and selection of parameters Abort and delete project

---

**Input**

**Input files**

Show  entries Search:

Input File	Template	Format	Actions
No data available in table			

Showing 0 to 0 of 0 entries

---

**Upload a file from disk**

Use this to upload files from your computer to the system.

**Step 1** First select what type of file you want to add:

**Step 2** Set the parameters for this type of file:  
Select a type first

**Step 3**

---

**Grab a file from the web**

Retrieves an input file from another location on the web.

**Step 1** First select the desired input type:

**Step 2** Set the parameters for this type of file:  
Select a type first

**Step 3** Enter the URL where to retrieve the file

**Step 4**

---

**Add input from browser**

You can create and add new files from within your browser:

---

**Parameter Selection**

**Modules**

**Skip modules**  
*Are there any components you want to skip? Skipping the parser and chunker speeds up the process*



# THE CLOUD

SARA HPC-Cloud Documentation | Support | Community Welcome marc.kemps.snijders%40meertens.knaw.nl | Sign out

Dashboard  
Virtual Machines  
Templates  
Images  
Create VM  
Network filter  
Users  
Upload images

Show 10 entries Search:

All	ID	Owner	Group	Name	Status	Network	Hostname	Start Time	VNC Access
<input type="checkbox"/>	2934	marc.kemps.snijders@meertens.knaw.nl	ctnww-meertens	alpino-erikt	RUNNING	IP = 145.100.57.148	d-node02	14:46:50 04/04/2012	
<input type="checkbox"/>	3617	marc.kemps.snijders@meertens.knaw.nl	ctnww-meertens	alpino-erikt-ghent	RUNNING	IP = 145.100.57.29	d-node12	13:07:06 04/25/2012	
<input type="checkbox"/>	5806	marc.kemps.snijders@meertens.knaw.nl	ctnww-meertens	FROG_Martin	RUNNING	IP = 145.100.57.13	d-node13	13:33:31 06/27/2012	
<input checked="" type="checkbox"/>	5851	marc.kemps.snijders@meertens.knaw.nl	ctnww-meertens	InPolder-martin	RUNNING	IP = [ 10.0.129.10, 145.100.57.78 ]	d-node16	12:20:11 06/28/2012	
<input type="checkbox"/>	8128	marc.kemps.snijders@meertens.knaw.nl	ctnww-meertens	mks-test1	RUNNING	IP = [ 10.0.129.16, 145.100.58.189 ]	d-node03	16:02:44 08/20/2012	

VM information | VM Template | VM log | Monitoring information

Virtual Machine information - TTNWW-test5

ID	8186
Name	TTNWW-test5
Owner	marc.kemps.snijders@meertens.knaw.nl
Group	ctnww-meertens
State	ACTIVE
LCM State	RUNNING
Hostname	d-node04
Start time	15:15:22 08/21/2012
Deploy ID	one-8186
Permissions	

Monitoring information

Net_TX	2368316
Net_RX	5210628
Used Memory	8G
Used CPU	1
VNC Session (port 43962)	<a href="#">Reset VNC</a>

Project Quota Usage: 36%

sara

Copyright 2002-2012 © OpenNebula Project Leads (OpenNebula.org). All Rights Reserved. OpenNebula 3.2.1

# TRY IT YOURSELF

- Pos tagging

– <http://145.100.57.19/frog/>

Frog

**Start a new Project**

Project ID:

– Select your newly created project

**Add input from browser**

You can create and add new files from within your browser:

**Parameter Selection**

**Modules**

**Skip modules**  
*Are there any components you want to skip? Skipping the parser and chunker speeds up the process considerably.*

Tokeniser	<input type="checkbox"/>
Multi-Word Detector	<input type="checkbox"/>
Parser	<input type="checkbox"/>
Chunker / Shallow parser	<input type="checkbox"/>

# TRY IT YOURSELF

- Named Entity Recognition

- <http://145.100.57.35/>

The Nerd System

**Start a new Project**

Project ID:

- Select your newly created project

**Add input from browser**

You can create and add new files from within your browser:

- Copy output from POS tagging into the window

**Start**

- Workflow is a term used to describe the tasks, procedural steps, organizations or people involved, required input and output information, and tools needed for each step in a business process.

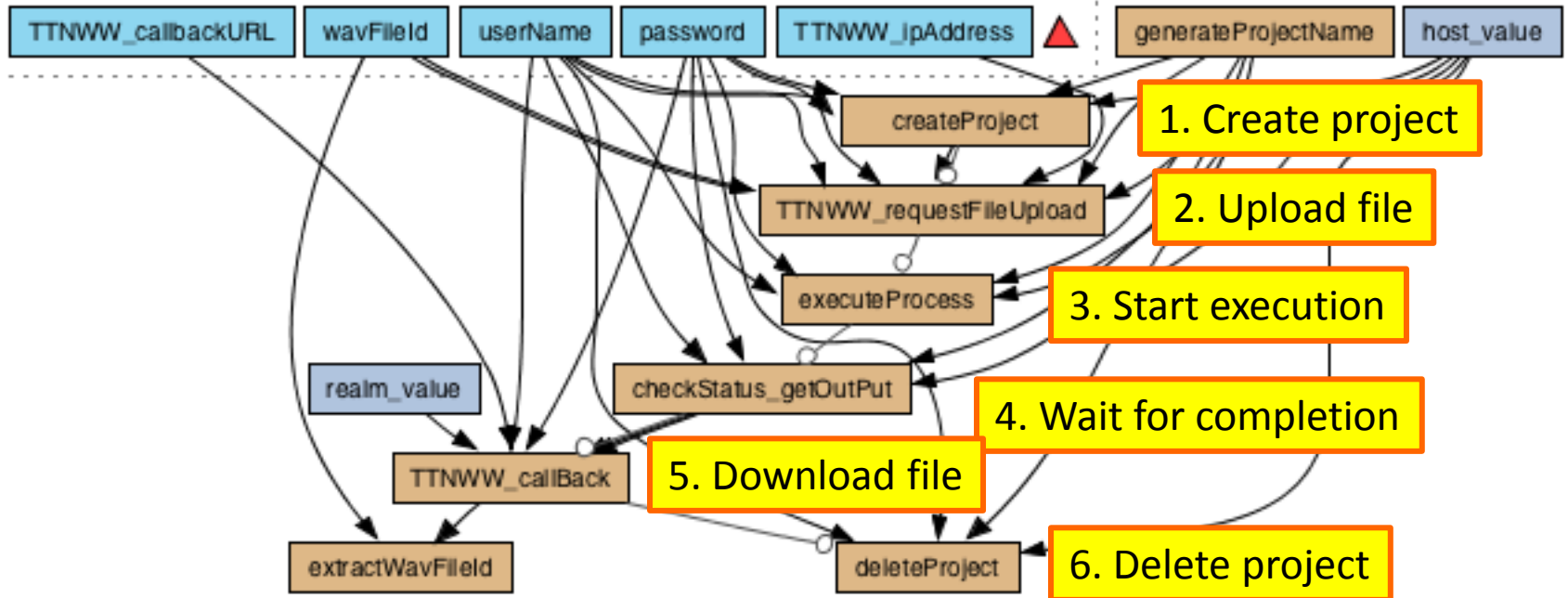
*<http://searchcio.techtarget.com/definition/workflow>*

Workflows have been commonly used in commercial and scientific processes for decades.

# WORKFLOWS

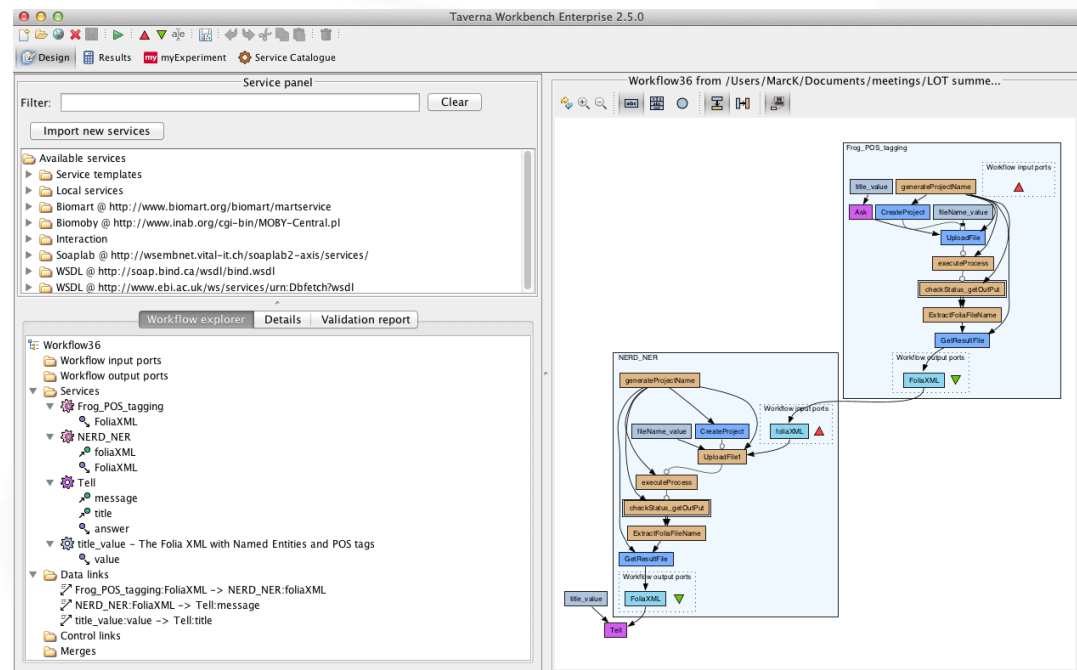
TYING SERVICES TOGETHER

Workflow input ports



# TRY IT YOURSELF

- Download Taverna workbench from <http://www.taverna.org.uk/>
- Use the workflow connecting Frog and Nerd.

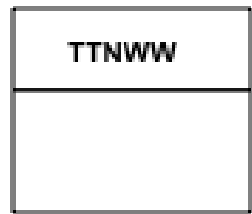
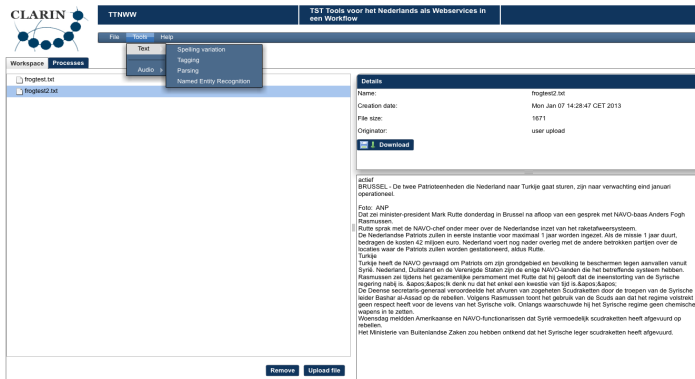


# **TTNWW APPLICATION**





# ARCHITECTUUR



1. Submit workflow



2. Evaluate workflow



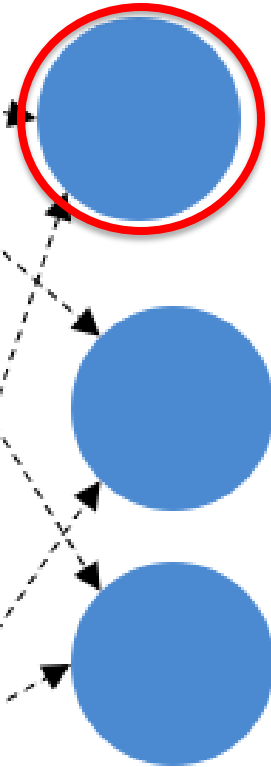
4. Execute workflow

5. Remove VM

3. Deploy image



**Currently disabled**



In SARA HPC cloud gedeployed



# TTNWW APPLICATIE



TTNWW TST Tools voor het Nederlands als Webservices in een Workflow

File Tools Help

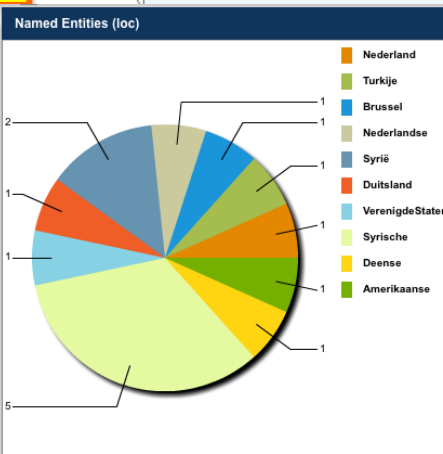
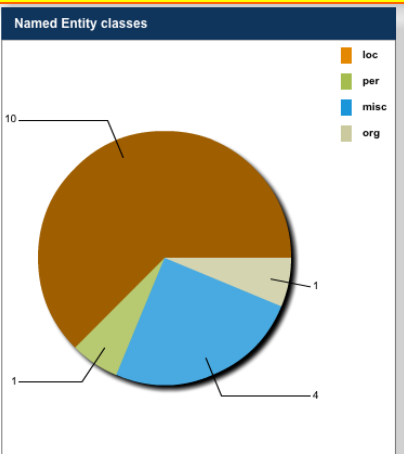
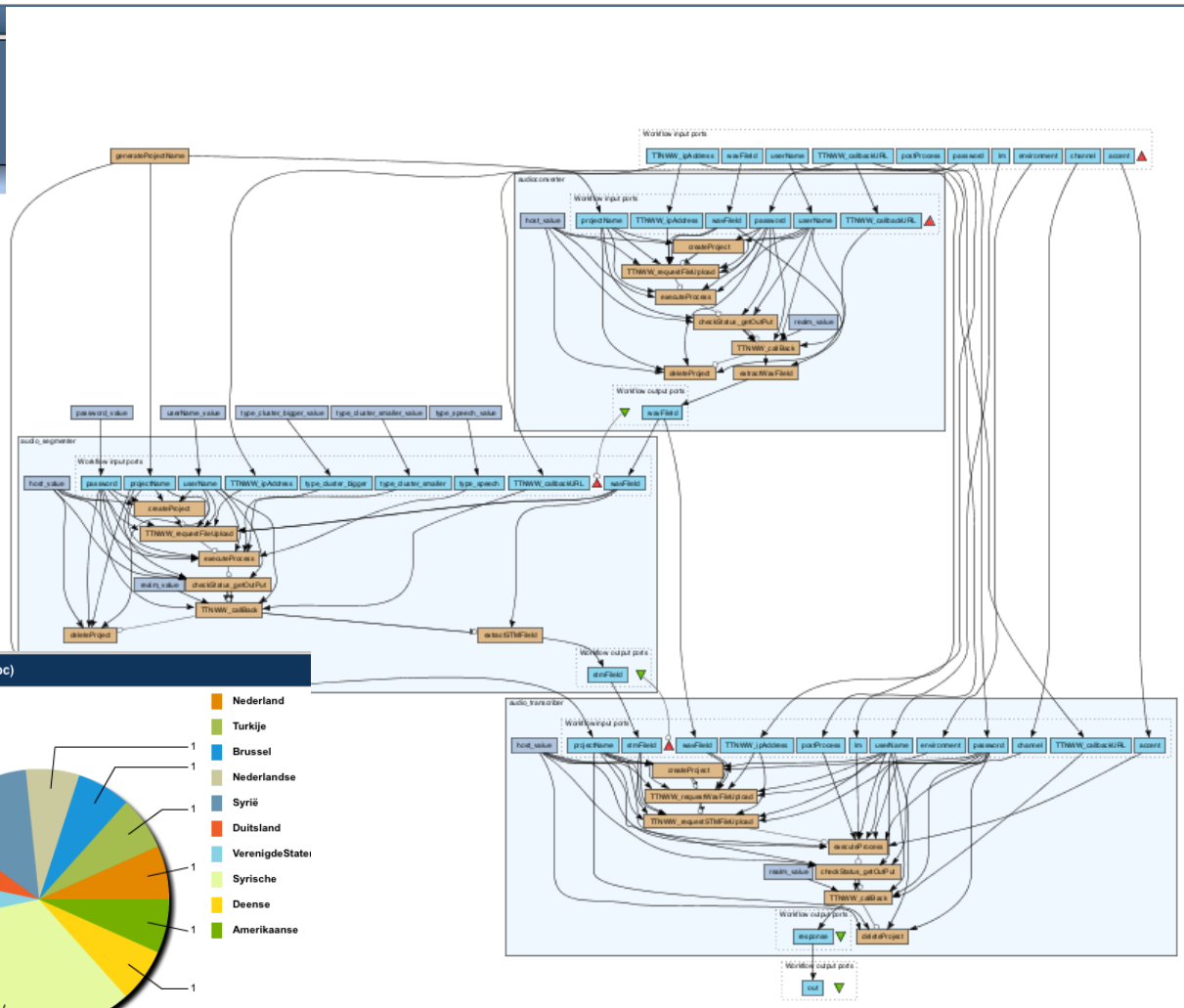
Text Spelling variation Tagging Parsing Named Entity Recognition

Audio

Workspace Processes

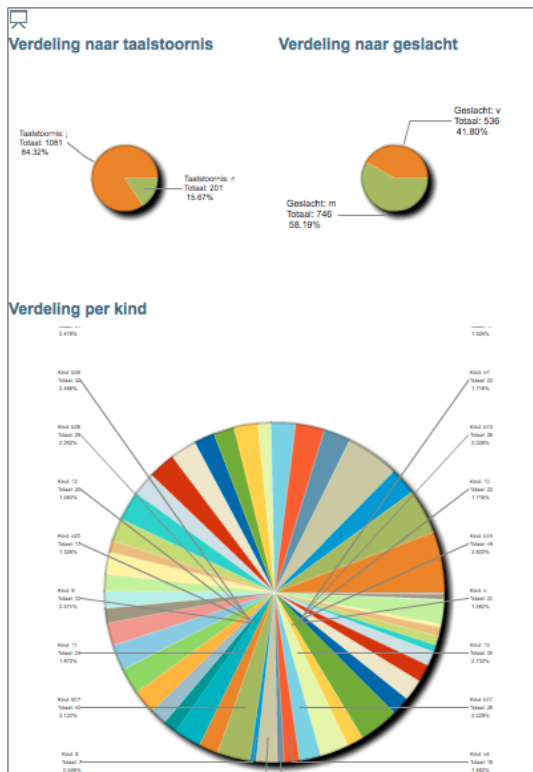
frogtest.txt frogtest2.txt

- Functionaliteiten**
- Data selectie (upload/download)
  - Selectie van processen
  - Aansturing van processen
  - Terugkoppeling resultaten in tijdelijke werkruimte
  - Only generic viewers



# TRY IT YOURSELF

- Use TTNWW at <http://yago.meertens.knaw.nl/apache/TTNWW/>



'de'[pos='adj.\*']{0,3}[pos='.\*onz.\*'] Zoek Found 1282 hits

Left	Hit	Right
de hondje gaat in	de bakketje	zit in haar hoofd in
en die hondje slaapt op	de bed	.
hij springt naar	de bije	. weg . hij roepe
roepen bij de boom en	de bije	gaan weg weer .
de hond gaat vluchten .	de bije	gaan achter haar aan .
en de hond pakt	de bijerenhuisje	. en de jongen kijkt
Ik wijs naar	de blauw vliegtuig	
Ik wijs naar	de blauw vliegtuig	
Ik wijs naar	de blauw vliegtuig	
Ik wijs naar	de blauwe vliegtuig	
Ik wijs naar	de blauwe vliegtuig	
Ik wijs naar	de blauwe vliegtuig	
Ik wijs naar	de blauwe vliegtuig	
Ik wijs naar	de blauwe vliegtuig	
Ik wijs naar naar	de blauwe vliegtuig	
Ik wijs naar	de blauwe vliegtuig	
Ik wijs naar	de blauwe vliegtuig	
Ik wijs naar wijs naar	de blauwe vliegtuig	
Ik wijs naar naar	de blauwe vliegtuig	
Ik wijs naar	de blauwe vliegtuig	
Kroko staat naast	de boek	
Smurf staat op	de boek	
Kroko staat naast	de boek	
Ik wil naar	de boek	

Sectie: A.3\_Blauwe\_map  
 Taak: elicitation  
 Zinsnummer: 65  
 Doelvorm: Ik wijs naar het blauwe witte vliegtuig

Kind: b16  
 Leeftijd: 6.9  
 Geslacht: v  
 Taal: tu  
 data

ik	wijs	naar	de	blauw	vliegtuig
VNW(pers,pron,	N(soort,ev,basis,	VZ(init)	LID(bep,stan,res	ADJ(prenom,bas	N(soort,ev,basis,
0.999688	1	0.999754	0.981886	0.973684	0.996804



# OTHER APPROACHES IN THE '*CLARIN*' DOMAIN

- Weblicht
  - Custom designed workflow system
  - Webbased
- Panacea
  - Taverna based workflow system
  - Workbench oriented

## Sign in to Clarin EU Service Provider

Select your Identity Provider

[View Tool List](#) [HELPDESK](#)


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 [webmaster@clarin.eu](mailto:webmaster@clarin.eu).



← Clarin.eu website account  
 European Union

RUG  
 Netherlands

Universiteit Leiden  
 Netherlands

Graafschap College  
 Netherlands

TERENA Secretariat  
 Netherlands

Universiteit van Amsterdam  
 Netherlands

Tilburg University  
 Netherlands

Vrije Universiteit  
 Netherlands

NIOO (KNAW)  
 Netherlands

SURFnet bv  
 Netherlands



**Shibboleth.**

<https://weblicht.sfs.uni-tuebingen.de/WebLicht-4/>

or search for a provider, such as Example University.

Main Page **Chain 2** x +

Show tools with status:  dev  development  production  withdrawn

Next Choices (Double-click on an icon to add it to the chain)

<b>IMS</b> Morphology morphology i	<b>SfS</b> Convert to Negra Document Type: NEGRA Form: i	<b>Berlin-Brandenburg</b> Person Nar Named Entities: person i	<b>SfS</b> Berkeley Parser - Berke Parsing: tuebadztb i	<b>Berlin-Brandenburg</b> Lemmas2L Language: German Document Type: Lexicon Form: TCF Version: 0.4 entries.type: lemmas i	<b>Berlin-Brandenburg</b> Tokens2Le: Language: German Document Type: Lexicon Form: TCF Version: 0.4 entries.type: types i	<b>IMS</b> Constituent Parser Parsing: Tiger Treebank Tagset i	<b>Berlin-Brandenburg</b> CAB orthog orthography i	<b>SfS</b> German Named Entity F Model conll2003 i
--	--	---	---	---	--	--	--	---

Input and Chain Selection

[Run Tools](#) [Clear Results](#) [Download chain](#)

<b>My Input</b> Plain Text Es war einmal ein Mädchen so schön das die Himmel weinte. ↕	<b>SfS</b> To TCF Converter Language: German Document Type: TCF TCF Version: 0.4 Text i X	<b>Berlin-Brandenburg</b> Tokenizer a Sentences Tokens i X	<b>IMS</b> Stuttgart Dependency F Part of Speech: STTS Tagset Parsing (Dep): No Empty Token Lemmas Parsing (Dep): tiger Parsing (Dep): false i X
--	--	---	--

### Available Annotations for: German Plain Text

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

#### German Named Entity Recognizer

Annotation Layers:

- Simple view
  - text
  - sentences
  - Table view
    - tokens
    - POSTags
    - lemmas
    - namedEntities
  - Highlighted view
    - namedEntities

language = de

[Download as Excel sheet](#) | [Download as CSV](#)

token ID	tokens	POSTags	lemmas	namedEntities
t1	Es	PPER	es	
t2	war	VAFIN	sein	
t3	einmal	ADV	einmal	
t4	ein	ART	ein	
t5	Mädchen	NN	Mädchen	
t6	so	ADV	so	
t7	schön	ADJD	schön	

[Download TCF](#)

#### Input and Chain Selection

[Run Tools](#) | [Clear Results](#) | [Download chain](#)

**My Input** Plain Text

Es war einmal ein Mädchen so schön das die Himmel weinte.

[↔](#)

**SfS** To TCF Converter

Language: German  
Document Type: TCF  
TCF Version: 0.4  
Text

[↓](#) [↕](#) [i](#)

**IMS** Tokenizer

Sentences  
Tokens

[↓](#) [↕](#) [i](#)

**IMS** TreeTagger

Part of Speech: STTS Tagset  
Lemmas

[↓](#) [↕](#) [i](#)

**SfS** German Named Entity F

Named Entities: tuebadz8

[↓](#) [↕](#) [i](#)

| Done running tools.



Available Annotations for: German Plain Text

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

German Named Entity Recognizer

Annotation Layers:

- Simple view
  - text
  - sentences
  - Table view
    - tokens
    - POSTags
    - lemmas
    - namedEntities
  - Highlighted view
    - namedEntities

language = de

[Download as Excel sheet](#) [Download as CSV](#)

1 / 2

[ 1 - 30 / 46 ]

token ID	tokens	POSTags	lemmas	namedEntities
t1	Israelische	ADJA	israelisch	
t2	Kampfflugzeuge	NN	Kampfflugzeug	
t3	haben	VAFIN	haben	
t4	in	APPR	in	
t5	der	ART	d	
t6	Nacht	NN	Nacht	
t7	neun	CARD	neun	

[Download TCF](#)

Input and Chain Selection

[Run Tools](#) [Clear Results](#) [Download chain](#)

<p><b>My Input</b> Plain Text</p> <p>Israelische Kampfflugzeuge haben in der Nacht neun Stellungen der syrischen Armee beschossen. Die Armee teilte mit, es handele sich um eine Antwort auf den von</p>	<p><b>SFS</b> To TCF Converter</p> <p>Language: German Document Type: TCF TCF-Version: 0.4 Text</p>	<p><b>IMS</b> Tokenizer</p> <p>Sentences Tokens</p>	<p><b>IMS</b> TreeTagger</p> <p>Part of Speech: STTS Tagset Lemmas</p>	<p><b>SFS</b> German Named Entity F</p> <p>Named Entities: tuebadz8</p>
--	---	---	--	---

Done running tools.

### Welcome to PANACEA

PANACEA, a STREP Project under EU-FP7, has developed a factory of Language Resources (LRs) in the form of a production line that automates all steps involved in the acquisition, production, maintenance and updating of the LRs required by Machine Translation and other Language Technologies.

The factory is a Web Service-based platform that integrates advanced technological components for:


- Monolingual and Parallel Text Acquisition and Pre-Processing
- Parallel corpora Alignment
- Bilingual Dictionary Production
- Monolingual Rich Information Lexica Production


Click [HERE](#) to learn how to begin using PANACEA.


Click [HERE](#) to get the acquired resources, i.e. the datasets produced during the project:


- Monolingual Corpora (raw text, n-grams, parsed, for EL, EN, ES, FR and IT)
- Parallel Corpora (sentence aligned, EN-EL and EN-FR)
- Monolingual Lexica (Verbal Subcategorization for EN, ES and IT, Noun Lexical-Semantic Classes, for EN and ES; MultiWords for IT)
- Bilingual Glossaries (EL-EN, FR-EN, DE-EN)

### PANACEA Web Portals

  
 Registry

  
 MyExperiment

  
 Tutorials

  
 Documentation

### News

**MT Summit 2013 PANACEA tutorial now available**

18  
Sep  
2013

[Read more](#)

**MT SUMMIT 2013 PANACEA TUTORIAL**

04  
Mar  
2013

PANACEA will hold a tutorial at the [Machine Translation Summit XIV](#) in Nice, France on 2 - 6 September 2013.

[Read more](#)

**8 PANACEA Papers accepted at COLING 2012**

23  
Nov  
2012

List of Accepted papers at the forthcoming 24th International Conference on Computational Linguistics COLING 2012, December Mumbai, India.

[Read more](#)

### Latest Services Available


TaaS API 2014-06-06, 07:15 am

- Test service 2013-10-09, 03:34 am
- corpus\_analysis 2013-04-16, 05:57 am
- tpc\_subcat\_inductive 2013-04-15, 04:27 am
- graf\_resource\_header\_generator 2013-04-03, 06:56 am

Home » Workflows

## Workflows

Search filter terms

Sort by: Rank

Filter by type

Taverna 2 2

Filter by tag

- directory 2
- download 2
- freeing 2
- upload 2
- dependency 1
- graf 1
- tagging 1

Filter by user

- Marcpoch 33
- Muntsa Padró 13
- Valeria Quochi 8
- atoral 7
- Laura Rimell 3
- Thurmair 3
- Francesco Ru... 2
- Marta Villegas 2
- Prokopis 2
- Panacea Admin 1

Filter by licence

by-sa 2

Showing 2 results. Use the filters on the left and the search box below to refine the results.

User: Marta Villegas

[Remove all filters](#)

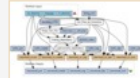
**Taverna 2** **Freeling dependency for plain text data with input upload and output download (v1)** View Download (v1)

**Original Uploader**  
 Created: 14/02/12 @ 15:44:15 | Last updated: 14/03/12 @ 09:00:03  
 Credits: Marta Villegas

License: Creative Commons Attribution-Share Alike 3.0 Unported License



Marta Villegas



This is the dependency workflow for plain text files in local directory using Freeling Graf. Freeling is run using the "keeptags" option to remove boilerplate and to keep paragraph tags info from the input data. The output is converted to graf format using the grafconverter\_postagging service. This workflow uses the download\_url processor for users who want to download the output files during the workflow execution and want the output files to follow the name of the input file. Exam...

Rating: 0.0 / 5 (0 ratings) | Versions: 1 | Reviews: 0 | Comments: 0 | Citations: 0

Viewed: 13 times | Downloaded: 7 times

Tags (6):

[dependency](#) | [directory](#) | [download](#) | [freeling](#) | [graf](#) | [upload](#)

### New/Upload

Workflow

### Log in / Register

Username or Email:

Password:

Remember me:

Need an account?  
[Click here to register](#)

[Forgot Password?](#)

### Popular Tags

25 tags

[\[All Tags\]](#)

[basicxcex](#) | [bilingual](#) | [cleaner](#) | [cqp](#) | [crawled](#) | [dependency](#) | [dear](#) | [directory](#) | [download](#) | [english](#) | [example](#) | [freeling](#) | [graf](#) | [hunalign](#) | [ilsp](#) | [lexical acquisition](#) | [lists](#) | [merge](#) | [noun classification](#) | [panacea](#) | [parser](#) | [pos tagging](#) | [sentence alignment](#) | [spanish](#) | [tagging](#)

# FURTHER READING

- Martin W. C. Reynaert. Character confusion versus focus word-based correction and OCR variants in corpora  
<http://link.springer.com/article/10.1007%2Fs10032-010-0133-5>
- Van den Bosch, A., Busser, G.J., Daelemans, W., and Canisius, S. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch, In F. van Eynde, P. Dirix, I. Schuurman, and V. Vandeghinste (Eds.), Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting, Leuven, Belgium, pp. 99-114.  
<http://ilk.uvt.nl/downloads/pub/papers/tadpole-final.pdf>
- Daelemans, Walter, Sabine Buchholz, and Jorn Veenstra. "Memory-based shallow parsing." arXiv preprint cs/9906005 (1999).  
<http://arxiv.org/pdf/cs.CL/9906005.pdf>
- Gertjan van Noord. At Last Parsing Is Now Operational. In: Piet Mertens, Cedrick Fairon, Anne Dister, Patrick Watrin (editors): TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles. Page 20--42.  
<http://odur.let.rug.nl/vannoord/papers/taln.pdf>
- Iris Hendrickx, Gosse Bouma, Walter Daelemans, and Veronique Hoste. "COREA: Coreference Resolution for Extracting Answers for Dutch." Essential Speech and Language Technology for Dutch. Springer Berlin Heidelberg, 2013. 115-128.  
[http://download.springer.com/static/pdf/777/chp%253A10.1007%252F978-3-642-30910-6\\_7.pdf?auth66=1403690524\\_2af68c8e695d8395e42b6465edc837d1&ext=.pdf](http://download.springer.com/static/pdf/777/chp%253A10.1007%252F978-3-642-30910-6_7.pdf?auth66=1403690524_2af68c8e695d8395e42b6465edc837d1&ext=.pdf)
- Monachesi, Paola, Gerwert Stevens, and Jantine Trapman. "Adding semantic role annotation to a corpus of written Dutch." Proceedings of the Linguistic Annotation Workshop. Association for Computational Linguistics, 2007.  
[http://www.researchgate.net/publication/220029776\\_Associating\\_facial\\_displays\\_with\\_syntactic\\_constituents\\_for\\_generation/file/9fcfd50c5f3003762c.pdf#page=91](http://www.researchgate.net/publication/220029776_Associating_facial_displays_with_syntactic_constituents_for_generation/file/9fcfd50c5f3003762c.pdf#page=91)
- Schuurman, Ineke. "Spatiotemporal annotation on top of an existing treebank." Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories. 2007.  
[http://www.lrec-conf.org/proceedings/lrec2010/pdf/180\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2010/pdf/180_Paper.pdf)

Thank you for your attention.

