



národní  
úložiště  
šedé  
literatury

### **CLARIN-DSpace repozitář v LINDAT/CLARIN**

Straňák, Pavel; Košarko, Ondřej; Mišutka, Jozef  
2019

Dostupný z <http://www.nusl.cz/ntk/nusl-407834>

Dílo je chráněno podle autorského zákona č. 121/2000 Sb.

Licence Creative Commons Uveďte původ-Nezpracovávejte 4.0

Tento dokument byl stažen z Národního úložiště šedé literatury (NUŠL).

Datum stažení: 11.05.2024

Další dokumenty můžete najít prostřednictvím vyhledávacího rozhraní [nusl.cz](http://nusl.cz) .

Konference o šedé literatuře a repozitářích, Praha, 17. 10. 2019

# LINDAT/CLARIN

FAIR REPOSITORY FOR LANGUAGE DATA

PAVEL STRAŇÁK, ONDŘEJ KOŠARKO, JOZEF MIŠUTKA



via <http://www.nusl.cz/ntk/nusl-407834>

LINDAT/CLARIN

LINDAT/CLARIN

LINDAT/CLARIN

LINgUistic

LINDAT/CLARIN

LINGuistic

DATa

LINDAT/CLARIN

LINGuistic

DATA ... very broadly

LINDAT/CLARIN

LINgustic

DATa

/

CLARIN



LINDAT/CLARIN

LINguistic

DATa

/

Common

Language

Research and technology

INfrastructure

# LINDAT/CLARIN

- Czech national project; node of CLARIN ERIC
- Operational since 2014
- Users:
  - Researchers in SSH and Computational Linguistics
- Technology:
  - Repository (resources), Services, Applications
- Knowledge, Support and Training

# LANGUAGE TECHNOLOGY

- Natural Language Processing - NLP
  - Analysis, synthesis of spoken and written language
  - Machine Translation, Information Extraction, ...
  - Search in texts, audio, video, images
- State-of-the-art technology in NLP
  - “Statistical” methods:
    - Machine learning incl. neural networks
    - Need for (large) Language Resources – Texts, multimodal
      - Repositories, identification, replication of experiments, standards

# USERS

- Everyone
  - communicates in and works with natural language!
- ... immediate users of the infrastructure:
  - Language Technology researchers
    - Universities, Research organisations
      - Need lots of data, easy to get, clean open licensing
  - "Content" users:
    - Linguists, historians, teachers, psychologists, sociologists, ...
      - Need identifiable data, preprocessed, searchable, easy-to-use services and applications

# Data Repository

PRESERVE AND FIND LANGUAGE DATA AND NLP TOOLS




Search

[Advanced Search](#)

## Author

Hajič, Jan (47)

Žabokrtský, Zdeněk (32)

Straka, Milan (29)

Zeman, Daniel (29)

Bojar, Ondřej (28)

[... View More](#)

## Subject

Germanistik (47)

machine translation (39)

corpus (34)

treebank (30)

morphology (26)

[... View More](#)

## Language (ISO)

English (222)

Czech (192)

German (159)

Dutch (92)

Spanish (83)

[... View More](#)

## What's New

ToolService

LINDAT / CLARIN

### CorpusExplorer

#### Author(s):

Rüdiger, Jan Oliver

#### Description:

Software for corpus linguists and text/data mining enthusiasts. The CorpusExplorer combines over 45 interactive



? What can you do?

DEPOSIT




CITE

🎯 Browse

> All of the Repository

# DATA REPOSITORY

- **OPEN**  **ACCESS** (whatever can be – Public License Selector)
- > 500 registered users
  - submitters & users signing licenses (not everything can be OA)
- 200+ Data Records
  - > 1000 Metadata Records
  - 80 languages
- 100 TB+ Data in Repository (+ 1PB of UCS Shoah Foundation Archive)

# DATA REPOSITORY

- > 500 registered users
  - submitters & users signing licenses (not everything can be Open Access)
- 200+ Data Records
  - > 1000 Metadata Records
  - 80 languages
- 100 TB+ Data in Repository (+ 1PB of UCS Shoah Foundation Archive)

The screenshot shows the LINDAT/CLARIN Repository search results page. At the top, there are navigation links: LINDAT CLARIN, Repository, TreeQuery, Treex, and More Apps. Below the navigation is a search bar with a magnifying glass icon and a 'Search' button. Underneath the search bar is an 'Advanced Search' link. The main content area is divided into two columns. The left column, titled 'Limit your search', contains several dropdown menus for filtering results: Author, Subject, Rights, Language (ISO), Type, Contain Files, and Community. The right column, titled 'Showing 1 through 10 out of 1038 results', displays a list of search results. The first result is 'AKCES 2 ver. 2' (Charles University in Prague, ÚČJTK / 2013-12-18) by Šebesta, Karel ; Goláňová, Hana, with a file size of 3.85 MB and a 'Publicly Available' status. The second result is 'A Gold Standard Word Alignment for English-Swedish (2015-10-12)' (Linköping University / 2015-10-12) by Ahrenberg, Lars ; Holmqvist, Maria, with a file size of 590 KB and a 'Publicly Available' status. The third result is 'MorphoDiTa: Morphological Dictionary and Tagger' (Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics (UFAL) / 2014-02-14) by Straka, Milan ; Straková, Jana, with no files and a 'Publicly Available' status. The page also features a pagination control showing '1 2 3 > 104'.

# DATA REPOSITORY

- Safe preservation  
(upload and don't worry)
- Discovery & Reuse
- Direct data citation  
(works in Google Scholar)
- Licensing  
(Open Access, but also more options)
- Versioning
- Language data and tools
- Worldwide (for everyone), easy to use

The screenshot shows the LINDAT/CLARIN Repository search results page. At the top, there are navigation links for "LINDAT/CLARIN Repository Home" and "Search". Below this is a search bar with a magnifying glass icon and a "Search" button. Underneath the search bar is a link for "Advanced Search".

The main content area is divided into two columns. The left column, titled "Limit your search", contains several dropdown menus for filtering results: "Author", "Subject", "Rights", "Language (ISO)", "Type", "Contain Files", and "Community".

The right column, titled "Showing 1 through 10 out of 1038 results", displays a list of search results. Each result is shown in a card format with a category label (e.g., "Corpus", "LexicalConceptualResource", "ToolService") and a title. The first result is "AKCES 2 ver. 2" (Charles University in Prague, ÚČJTK / 2013-12-18) by Šebesta, Karel ; Goláňová, Hana, containing 1 file (3.85 MB). The second result is "A Gold Standard Word Alignment for English-Swedish (2015-10-12)" (Linköping University / 2015-10-12) by Ahrenberg, Lars ; Holmqvist, Maria, containing 1 file (590 KB). The third result is "MorphoDiTa: Morphological Dictionary and Tagger" (Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics (UFAL) / 2014-02-14) by Straka, Milan ; Straková, Jana, containing no files. Each result card also includes a "Publicly Available" badge and a Creative Commons license icon.



# DATA REPOSITORY

- Safe preservation (upload and don't worry)
- Discovery & Reuse
- Direct data citation (works in Google Scholar)
- Licensing (Open Access, but also more options)
- Versioning
- Language data and tools
- Worldwide (for everyone), easy to use

The screenshot shows the LINDAT/CLARIN Repository search results page. At the top, there are navigation links: LINDAT/CLARIN, Repository, TreeQuery, Treex, More Apps, and an 'All' link. Below the navigation is a search bar with a magnifying glass icon and a 'Search' button. Underneath the search bar is a link for 'Advanced Search'. The main content area is divided into two columns. The left column, titled 'Limit your search', contains several dropdown menus: Author, Subject, Rights, Language (ISO), Type, Contain Files, and Community. The right column, titled 'Showing 1 through 10 out of 1038 results', displays a list of search results. The first result is 'AKCES 2 ver. 2' (Charles University in Prague, ÚČJTK / 2013-12-18) by Šebesta, Karel ; Goláňová, Hana, with a file size of 3.85 MB and a 'Publicly Available' status. The second result is 'A Gold Standard Word Alignment for English-Swedish (2015-10-12)' (Linköping University / 2015-10-12) by Ahrenberg, Lars ; Holmqvist, Maria, with a file size of 590 KB and a 'Publicly Available' status. The third result is 'MorphoDiTa: Morphological Dictionary and Tagger' (Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics (UFAL) / 2014-02-14) by Straka, Milan ; Straková, Jana, with no files and a 'Publicly Available' status. The page also features a pagination control showing '1 2 3 > 104'.

# UPLOAD AND DON'T WORRY

## How to Deposit

Only authenticated users can deposit items. If you cannot find your home organisation in the Login dialog list of organisations then register at [clarin.eu](http://clarin.eu) and authenticate using "clarin.eu website account". In case you cannot use any authentication method above or if you encounter a problem, do not hesitate to contact our [Help Desk](#) and we can create a local account for you.

### Step 1: Login

To start a new submission you have to login first. Click Login under My Account in the right menu panel.

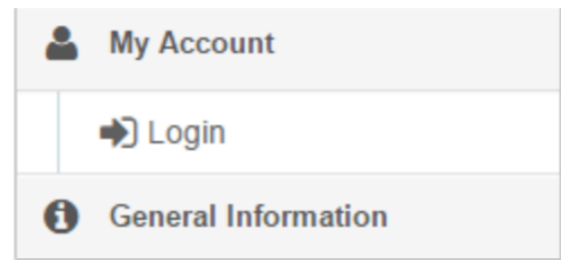
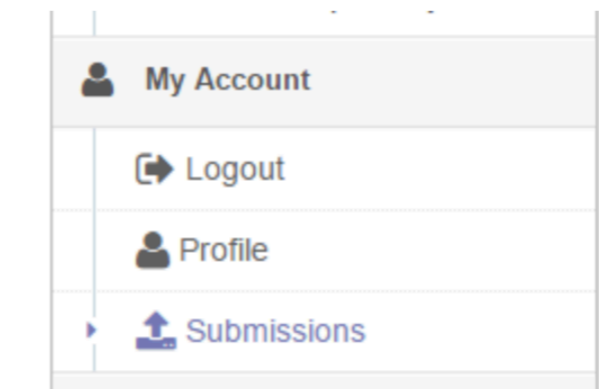


Fig1. Menu Login

### Step 2: Starting a new submission

Now you have a new menu item 'Submissions' under My Account. Click on Submissions to go to the Submissions screen.



The screenshot shows the LINDAT CLARIN user interface. At the top right, there are logos for LINDAT CLARIN and CLARIN. Below the logos is a navigation menu with the following items:

- What can you do?** (with a question mark icon)
  - DEPOSIT** (with a download icon)
  - CITE** (with a quote icon)
- Browse** (with a magnifying glass icon)
  - > All of the Repository (with a dropdown arrow)
- My Account** (with a user icon)
  - Login (with a right arrow icon)
- General Information** (with an information icon)
  - Deposit (with an upload icon)
  - Cite (with a quote icon)
  - Submission Lifecycle (with a refresh icon)
  - FAQ (with a question mark icon)
  - About (with an exclamation mark icon)
  - Help Desk (with an envelope icon)

# UPLOAD AND DON'T WORRY

## How to Deposit

Only authenticated users can deposit items. If you cannot find your home organisation in the Login dialog list of organisations then register at [clarin.eu](http://clarin.eu) and authenticate using "clarin.eu website account". In case you cannot use any authentication method above or if you encounter a problem, do not hesitate to contact our [Help Desk](#) and we can create a local account for you.

### Step 1: Login

To start a new submission you have to login first. Click Login under My Account in the right menu panel.

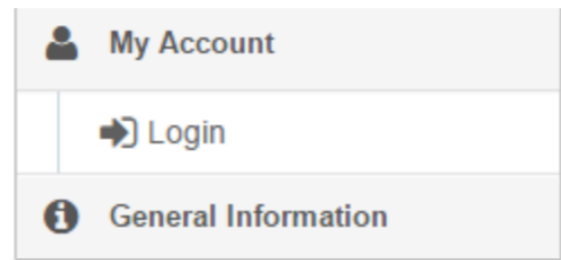
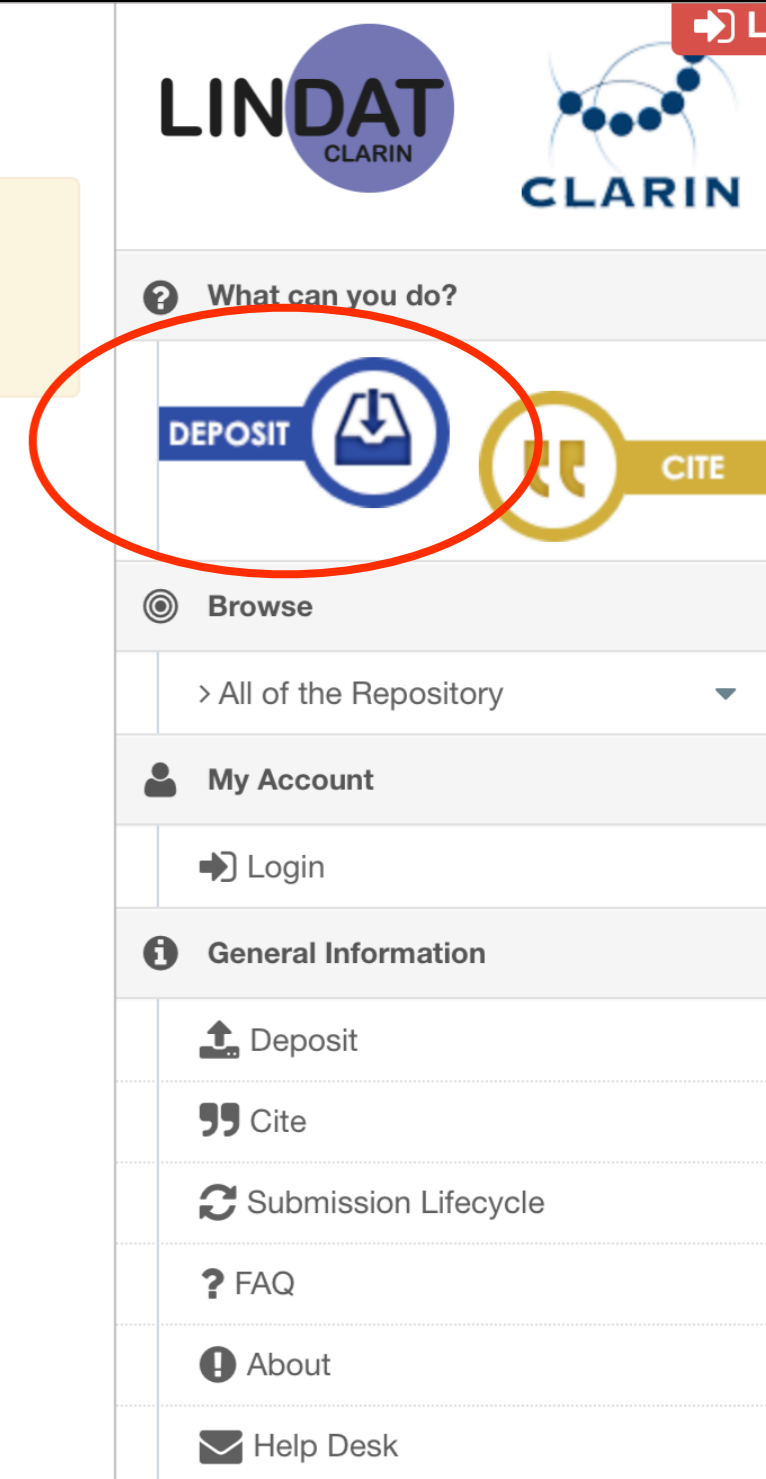
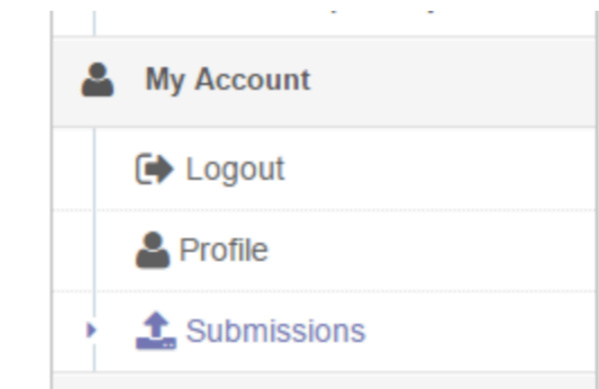


Fig1. Menu Login

### Step 2: Starting a new submission

Now you have a new menu item 'Submissions' under My Account. Click on Submissions to go to the Submissions screen.



# DEPOSITION GUIDE

- step-by-step description

1. login

2. fill-in metadata

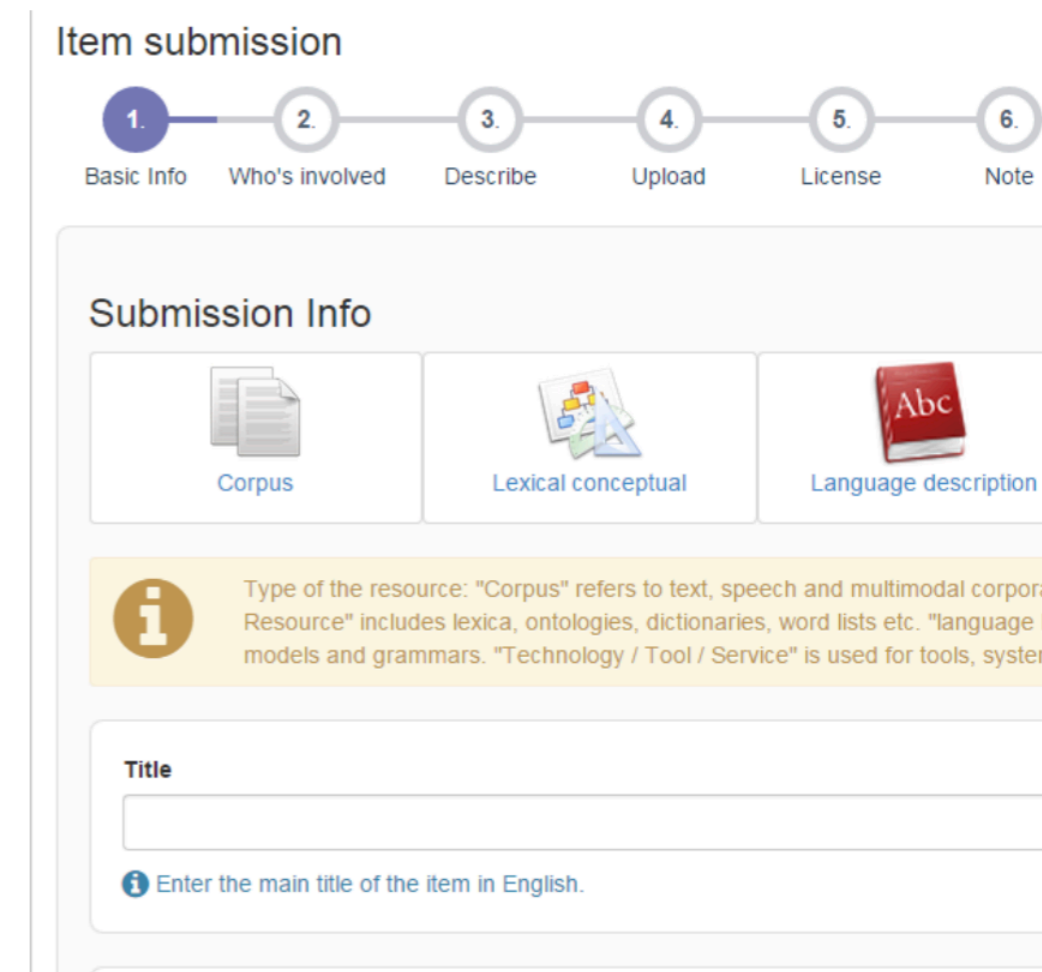
3. drag&drop data

4. select a license

5. submit

## Step 3: Select type of your submission

You have initiated a new workflow item. In the next few steps you will provide the details, select the type of the resource you are about to submit.



Item submission

1. Basic Info 2. Who's involved 3. Describe 4. Upload 5. License 6. Note

Submission Info

Corpus Lexical conceptual Language description

**i** Type of the resource: "Corpus" refers to text, speech and multimodal corpora. "Resource" includes lexica, ontologies, dictionaries, word lists etc. "language description" includes language models and grammars. "Technology / Tool / Service" is used for tools, systems and services.

Title

**i** Enter the main title of the item in English.

Fig4. Submission info

Click on one of the type buttons e.g. Corpus. Proceed with filling the basic information in the following step.

## Step 4: Describe your item

In the following two steps you will provide more details for your item. First describe the item.



Item submission

Basic Info 2. Who's involved Describe Upload License Note

# LOGIN TO DEPOSIT


- institutional logins (EduID-cz, EduGAIN)
- CLARIN account for the “homeless researchers”
- minimal personal info


## Sign in to LINDAT/CLARIN Repository


Login via Your home institution (e.g. university)

- Univerzita Karlova v Praze  
Czech Republic 6 km
- Institute of Biotechnology CAS, v.v.i.  
Czech Republic 4 km
- Czech University of Life Sciences Prague  
Czech Republic 4 km
- Institute of Art History of the Academy of Sciences of the Czech Republic  
Czech Republic 4 km
- College of Polytechnics Jihlava  
Czech Republic 4 km
- Global Change Research Institute CAS  
Czech Republic 4 km
- Czech Language Institute of the Czech Academy of Sciences  
Czech Republic 4 km
- Institute of Chemical Process Fundamentals of the AS CR  
Czech Republic 4 km
- Institute of Theoretical and Applied Mechanics AS CR  
Czech Republic 4 km

► **Please help, I cannot find my provider**

 Locate me and show nearby providers

Show providers in Czech Republic  show all countries

DiscoJuice © UNINETT 

## FACETED SEARCH

[Advanced Search](#)

## Limit your search

Author

Subject

Rights

Language (ISO)

Type

Contain Files

Community

Showing 1 through 10 out of 1038 results

[1](#)
[2](#)
[3](#)
[>](#)
[104](#)


Corpus

LINDAT / CLARIN

AKCES 2 ver. 2

(Charles University in Prague, ÚČJTK / 2013-12-18)

Author(s):

Šebesta, Karel ; Goláňová, Hana

This item contains 1 file (3.85 MB).

Publicly Available

LexicalConceptualResource

LRT + Open Submissions

A Gold Standard Word Alignment for English-Swedish (2015-10-12)

(Linköping University / 2015-10-12)

Author(s):

Ahrenberg, Lars ; Holmqvist, Maria

This item contains 1 file (590 KB).



What can you do?

DEPOSIT



CITE



Browse

&gt; All of the Repository

My Account

Login

General Information

Deposit

Cite

Submission Lifecycle

FAQ

About

Help Desk

# FACETED SEARCH

[Advanced Search](#)


### Limit your search

- Author ▾
- Subject ▾
- Rights ▾
- Language (ISO) ▾
- Type ▾
- Contain Files ▾
- Community ▾

Showing 1 through 10 out of 1038 results

1
2
3
>
104
⚙️ ▾

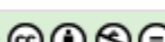
Corpus
LINDAT / CLARIN

[AKCES 2 ver. 2](#) 


(Charles University in Prague, ÚČJTK / 2013-12-18)

**Author(s):**  
Šebesta, Karel ; Goláňová, Hana

This item contains 1 file (3.85 MB).

Publicly Available 



LexicalConceptualResource
LRT + Open Submissions

[A Gold Standard Word Alignment for English-Swedish \(2015-10-12\)](#) 


(Linköping University / 2015-10-12)


**Author(s):**  
Ahrenberg, Lars ; Holmqvist, Maria

This item contains 1 file (590 KB).

**?** What can you do?

DEPOSIT



CITE





**🎯** Browse

> All of the Repository ▾

**👤** My Account

➔ Login

**ℹ️** General Information

-  Deposit
-  Cite
-  Submission Lifecycle
- ?** FAQ
- !** About
-  Help Desk

## FACETED SEARCH

[Advanced Search](#)



? What can you do?



🎯 Browse

> All of the Repository

👤 My Account

➔ Login

📘 General Information

📄 Deposit

🗉 Cite

🔄 Submission Lifecycle

? FAQ

📄 About

📧 Help Desk

### Limit your search

Author

Subject

Rights

Language (ISO)

Type

Contain Files

Community

Showing 1 through 10 out of 1038 results

1 2 3 > 104



Corpus

LINDAT / CLARIN

AKCES 2 ver. 2

(Charles University in Prague, ÚČJTK / 2013-12-18)



**Author(s):**

Šebesta, Karel ; Goláňová, Hana

This item contains 1 file (3.85 MB).

Publicly Available

LexicalConceptualResource

LRT + Open Submissions

A Gold Standard Word Alignment for English-Swedish (2015-10-12)

(Linköping University / 2015-10-12)



**Author(s):**

Ahrenberg, Lars ; Holmqvist, Maria

This item contains 1 file (590 KB).



DISCOVERY

GOOGLE



prague dependency treebank 3.0



Vše

Obrázky

Nákupy

Mapy

Videa

Více

Nastavení

Nástroje

Přibližný počet výsledků: 13 900 (0,44 s)

## Vědecké články o prague dependency treebank 3.0

**Prague dependency treebank 3.0** - [Bejček](#) - Počet citací tohoto článku: 47

**Prague Dependency Treebank** - [Hajič](#) - Počet citací tohoto článku: 385

**The Prague dependency treebank** - [Böhmová](#) - Počet citací tohoto článku: 423

## Prague Dependency Treebank 3.0 | ÚFAL

<https://ufal.mff.cuni.cz/pdt3.0> ▼ Přeložit tuto stránku

Introduction. The **Prague Dependency Treebank 3.0** (PDT 3.0) annotates the same texts as the PDT 2.0 (Hajič et al. 2006), PDT 2.5 (Bejček et al. 2011), and the Prague Discourse Treebank 1.0 (PDiT 1.0, Poláková et al. 2012). The annotation on the four layers was further fixed and improved in various aspects. Moreover ...

## The Prague Dependency Treebank 2.0.

<https://ufal.mff.cuni.cz/pdt2.0/> ▼ Přeložit tuto stránku

The **Prague Dependency Treebank 2.0** (PDT 2.0) contains a large amount of Czech texts with complex and interlinked morphological (2 million words), syntactic (1.5 MW) and complex semantic annotation ... Please note that new versions of this corpus have been published: PDT 3.0 (2013), PDiT 1.0 (2012), PDT 2.5 (2012).

## Prague Dependency Treebank 3.0 (PDT 3.0)

[https://lindat.mff.cuni.cz/repository/xmlui/bitstream/.../PDT30\\_index\\_lindat.html?...](https://lindat.mff.cuni.cz/repository/xmlui/bitstream/.../PDT30_index_lindat.html?...)

**Prague Dependency Treebank 3.0** (PDT 3.0). Overview. The **Prague Dependency Treebank 3.0** (PDT 3.0) annotates the same texts as the PDT 2.0 (Hajič et al. 2006), PDT 2.5 (Bejček et al. 2011), and the Prague Discourse Treebank 1.0 (PDiT 1.0, Poláková et al. 2012). The annotation on the four layers was further fixed ...

DISCOVERY

GOOGLE



prague dependency treebank 3.0



Vše

Obrázky

Nákupy

Mapy

Videa

Více

Nastavení

Nástroje

Přibližný počet výsledků: 13 900 (0,44 s)

## Vědecké články o prague dependency treebank 3.0

**Prague dependency treebank 3.0** - [Bejček](#) - Počet citací tohoto článku: 47

**Prague Dependency Treebank** - [Hajič](#) - Počet citací tohoto článku: 385

**The Prague dependency treebank** - [Böhmová](#) - Počet citací tohoto článku: 423

## Prague Dependency Treebank 3.0 | ÚFAL

<https://ufal.mff.cuni.cz/pdt3.0> ▼ Přeložit tuto stránku

Introduction. The **Prague Dependency Treebank 3.0** (PDT 3.0) annotates the same texts as the PDT 2.0 (Hajič et al. 2006), PDT 2.5 (Bejček et al. 2011), and the Prague Discourse Treebank 1.0 (PDiT 1.0, Poláková et al. 2012). The annotation on the four layers was further fixed and improved in various aspects. Moreover ...

## The Prague Dependency Treebank 2.0.

<https://ufal.mff.cuni.cz/pdt2.0/> ▼ Přeložit tuto stránku

The **Prague Dependency Treebank 2.0** (PDT 2.0) contains a large amount of Czech texts with complex and interlinked morphological (2 million words), syntactic (1.5 MW) and complex semantic annotation ... Please note that new versions of this corpus have been published: PDT 3.0 (2013), PDiT 1.0 (2012), PDT 2.5 (2012).

## Prague Dependency Treebank 3.0 (PDT 3.0)

[https://lindat.mff.cuni.cz/repository/xmlui/bitstream/.../PDT30\\_index\\_lindat.html?...](https://lindat.mff.cuni.cz/repository/xmlui/bitstream/.../PDT30_index_lindat.html?...)

**Prague Dependency Treebank 3.0** (PDT 3.0). Overview. The **Prague Dependency Treebank 3.0** (PDT 3.0) annotates the same texts as the PDT 2.0 (Hajič et al. 2006), PDT 2.5 (Bejček et al. 2011), and the Prague Discourse Treebank 1.0 (PDiT 1.0, Poláková et al. 2012). The annotation on the four layers was further fixed ...

## CREDIT FOR DATA



prague dependency treebank 3.0



Vše

Obrázky

Nákupy

Mapy

Videa

Více

Nastavení

Nástroje

Přibližný počet výsledků: 13 900 (0,44 s)

**Vědecké články o prague dependency treebank 3.0****Prague dependency treebank 3.0** - Bejček - Počet citací tohoto článku: 47**Prague Dependency Treebank** - Hajič - Počet citací tohoto článku: 385**The Prague dependency treebank** - Böhmová - Počet citací tohoto článku: 423**Prague Dependency Treebank 3.0 | ÚFAL**<https://ufal.mff.cuni.cz/pdt3.0> ▼ Přeložit tuto stránku

Introduction. The **Prague Dependency Treebank 3.0** (PDT 3.0) annotates the same texts as the PDT 2.0 (Hajič et al. 2006), PDT 2.5 (Bejček et al. 2011), and the Prague Discourse Treebank 1.0 (PDiT 1.0, Poláková et al. 2012). The annotation on the four layers was further fixed and improved in various aspects. Moreover ...

**The Prague Dependency Treebank 2.0.**<https://ufal.mff.cuni.cz/pdt2.0/> ▼ Přeložit tuto stránku

The **Prague Dependency Treebank 2.0** (PDT 2.0) contains a large amount of Czech texts with complex and interlinked morphological (2 million words), syntactic (1.5 MW) and complex semantic annotation ... Please note that new versions of this corpus have been published: PDT 3.0 (2013), PDiT 1.0 (2012), PDT 2.5 (2012).

**Prague Dependency Treebank 3.0 (PDT 3.0)**[https://lindat.mff.cuni.cz/repository/xmlui/bitstream/.../PDT30\\_index\\_lindat.html?...](https://lindat.mff.cuni.cz/repository/xmlui/bitstream/.../PDT30_index_lindat.html?...)

**Prague Dependency Treebank 3.0** (PDT 3.0). Overview. The **Prague Dependency Treebank 3.0** (PDT 3.0) annotates the same texts as the PDT 2.0 (Hajič et al. 2006), PDT 2.5 (Bejček et al. 2011), and the Prague Discourse Treebank 1.0 (PDiT 1.0, Poláková et al. 2012). The annotation on the four layers was further fixed ...

# CREDIT FOR DATA



enTenTen

“ Please use the following text to cite this item or export to a predefined format: ”

BIBTEX CMDI

Masaryk University, NLP Centre, 2011, *enTenTen*, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, <http://hdl.handle.net/11858/00-097C-0000-0001-CCDF-8>.



This resource is also integrated in following services:

Share:

KonText

Item identifier	<a href="http://hdl.handle.net/11858/00-097C-0000-0001-CCDF-8">http://hdl.handle.net/11858/00-097C-0000-0001-CCDF-8</a>
Date issued	2011-12-16
Type	corpus
Language(s)	English
Description	Very large English web corpus enTenTen, comprising 3,268,798,627 tokens.
Publisher	Masaryk University, NLP Centre
Acknowledgement	Lexical Computing Ltd.
Subject(s)	English large corpus
Collection(s)	LINDAT / CLARIN Data & Tools

[Show full item record](#)

What can you do?

DEPOSIT CITE

Browse

> All of the Repository

My Account

Logout

Profile

Submissions

Context

> Edit this item

> Export Item

> Export Metadata

Administrative

Control Panel

Access Control

## AS OPEN AS POSSIBLE

## Choose a License

Answer the questions or use the search to find the license you want

↻ Start again



What do you want to deposit?

Software

Data

Search for a license...

### Public Domain Mark (PD)

The work identified as being free of known restrictions under copyright law, including all related and neighboring rights.

Publicly Available 

### Public Domain Dedication (CC Zero)

CC Zero enables scientists, educators, artists and other creators and owners of copyright- or database-protected content to waive those interests in their works and thereby place them as completely as possible in the public domain, so that others may freely build upon, enhance and reuse the works for any purposes without restriction under copyright or database law.

Publicly Available   OPEN DATA

# AS OPEN AS POSSIBLE (NOT MORE)

**Publisher** [Faculty of Arts, Institute of the Czech National Corpus, Charles University in Prague](#)

**Acknowledgement** Ministerstvo školství, mládeže a tělovýchovy  
 Project code: LM2011023  
 Project name: Český národní korpus

**Subject(s)** representative corpus written language

**Collection(s)** [LINDAT / CLARIN Data & Tools](#)

[Show full item record](#)

**Files in this item**



Download instructions for command line

This item is Academic Use and licensed under:  
 Czech National Corpus (Shuffled Corpus Data)



<b>Name</b>	syn2015.gz
<b>Size</b>	1.48 GB
<b>Format</b>	application/x-gzip
<b>Description</b>	corpus
<b>MD5</b>	e0242cc77e999794af6cfaf57f843c12



[Download file](#)

# AS OPEN AS POSSIBLE (NOT MORE)

**Publisher** Faculty of Arts, Institute of the Czech National Corpus, Charles University in Prague


**Acknowledgement** Ministerstvo školství, mládeže a tělovýchovy  
 Project code: LM2011023  
 Project name: Český národní korpus

**Subject(s)** representative corpus written language


**Collection(s)** LINDAT / CLARIN Data & Tools


[Show full item record](#)

**Files in this item**

  
 Download instructions for command line

This item is Academic Use and licensed under:  
 Czech National Corpus (Shuffled Corpus Data)



<b>Name</b>	syn2015.gz	
<b>Size</b>	1.48 GB	
<b>Format</b>	application/x-gzip	
<b>Description</b>	corpus	
<b>MD5</b>	e0242cc77e999794af6cfaf57f843c12	

[Download file](#)

CLEAR RULES  
 CUSTOM LICENSES  
 LICENSE SIGNING

ANY LICENSE (OPEN SOURCE / OPEN DATA PREFERRED)

# LICENSING FRAMEWORK

## Manage Licenses

All Licenses

Define License

Define License Label

	License Name	Definition (URL)	Confirmation	Required user info	License Label	Extended Labels	Used by Bitstreams
<input type="checkbox"/>	Universal Derivations v0.5 License Agreement	<a href="https://lindat.mff.cuni.cz/repository/xmlui/page/licence-UDer-0.5">https://lindat.mff.cuni.cz/repository/xmlui/page/licence-UDer-0.5</a>	Not required		PUB	CC	1
<input type="checkbox"/>	Licence Universal Dependencies v2.4	<a href="https://lindat.mff.cuni.cz/repository/xmlui/page/licence-UD-2.4">https://lindat.mff.cuni.cz/repository/xmlui/page/licence-UD-2.4</a>	Not required		PUB	CC GPLv3	4
<input type="checkbox"/>	License agreement for The Multilingual corpus of literal occurrences of multiword expressions	<a href="https://lindat.mff.cuni.cz/repository/xmlui/page/licence-mwe-literal">https://lindat.mff.cuni.cz/repository/xmlui/page/licence-mwe-literal</a>	Not required		PUB	CC GPLv3	5
<input type="checkbox"/>	Licence Universal Dependencies v2.3	<a href="https://lindat.mff.cuni.cz/repository/xmlui/page/licence-UD-2.3">https://lindat.mff.cuni.cz/repository/xmlui/page/licence-UD-2.3</a>	Not required		PUB	CC GPLv3 GPLv2	4
<input type="checkbox"/>	PARSEME Shared Task Data (v. 1.1) Agreement	<a href="https://lindat.mff.cuni.cz/repository/xmlui/page/licence-mwe-1.1">https://lindat.mff.cuni.cz/repository/xmlui/page/licence-mwe-1.1</a>	Not required		PUB	CC GPLv3	22
<input type="checkbox"/>	Licence Universal Dependencies v2.2	<a href="https://lindat.mff.cuni.cz/repository/xmlui/page/licence-UD-2.2">https://lindat.mff.cuni.cz/repository/xmlui/page/licence-UD-2.2</a>	Not required		PUB	CC GPLv3 GPLv2	7
<input type="checkbox"/>	Licence Universal Dependencies v2.1	<a href="https://lindat.mff.cuni.cz/repository/xmlui/page/licence-UD-2.1">https://lindat.mff.cuni.cz/repository/xmlui/page/licence-UD-2.1</a>	Not required		PUB	CC GPLv3 GPLv2	4
<input type="checkbox"/>	PARSEME Shared Task Data (v. 1.0) Agreement	<a href="https://lindat.mff.cuni.cz/repository/xmlui/page/licence-mwe-1.0">https://lindat.mff.cuni.cz/repository/xmlui/page/licence-mwe-1.0</a>	Not required		PUB	CC GPLv3	21
<input type="checkbox"/>	Licence Universal Dependencies v2.0	<a href="https://lindat.mff.cuni.cz/repository/xmlui/page/licence-UD-2.0">https://lindat.mff.cuni.cz/repository/xmlui/page/licence-UD-2.0</a>	Not required		PUB	CC GPLv3 GPLv2	14
<input type="checkbox"/>	Licence Universal Dependencies v1.4	<a href="https://lindat.mff.cuni.cz/re">https://lindat.mff.cuni.cz/re</a>	Not		PUB	CC	4

LINDAT  
CLARIN



Pavel Straňák | Logout

What can you do?



Browse

> All of the Repository

My Account

Logout

Profile

Submissions

Administrative

Control Panel

Access Control

> Content Administration

Registries

Collections & Communities

Statistics

Curation Tasks

Licenses



# DEFINING A NEW LICENSE

# LICENSING FRAMEWORK



## Manage Licenses

All Licenses

Define License

Define License Label

### Define new licence

License name	<input type="text"/>
License definition URL	<input type="text"/>
License requires confirmation	<input type="text" value="Not required"/>
License Labels	<input checked="" type="radio"/> Publicly Available (PUB) <input type="radio"/> Academic Use (ACA) <input type="radio"/> Restricted Use (RES)
License Labels	<input type="checkbox"/> Attribution Required (BY) <input type="checkbox"/> Share Alike (SA) <input type="checkbox"/> Noncommercial (NC) <input type="checkbox"/> Redeposit Modified (ReD) <input type="checkbox"/> No Derivative Works (ND) <input type="checkbox"/> Inform Before Use (Inf) <input type="checkbox"/> Distributed under Creative Commons (CC) <input type="checkbox"/> No Copyright (ZERO) <input type="checkbox"/> GNU General Public License, version 3.0 (GPLv3) <input type="checkbox"/> GNU General Public License, version 2.0 (GPLv2) <input type="checkbox"/> BSD (BSD) <input type="checkbox"/> The MIT License (MIT) <input type="checkbox"/> The Open Source Initiative (OSI)
Additional required user info	<input type="checkbox"/> The user will receive an email with download instructions. <input type="checkbox"/> User Name <input type="checkbox"/> Date of Birth <input type="checkbox"/> Address <input type="checkbox"/> Country



Pavel Straňák | Logout

What can you do?



Browse

> All of the Repository

My Account

Logout

Profile

Submissions

Administrative

Control Panel

Access Control

> Content Administration

Registries

Collections & Communities

Statistics

Curation Tasks

Licenses

# DEFINING A LICENSING LABEL / ATTRIBUTE

# LICENSING FRAMEWORK

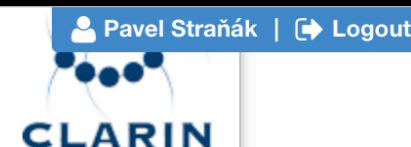


## Manage Licenses

[All Licenses](#)[Define License](#)[Define License Label](#)

### Define New Label

Short Label	<input type="text"/>
Label Title	<input type="text"/>
Is extended	<input type="text" value="Yes"/>
Icon image	<input type="button" value="Choose File"/> no file selected
	<input type="button" value="Save"/> <input type="button" value="Return"/>



What can you do?



Browse

> All of the Repository

My Account

Logout

Profile

Submissions

Administrative

Control Panel

Access Control

> Content Administration

Registries

Collections & Communities

Statistics

Curation Tasks

Licenses

## PREFER LATEST, PRESERVE ALL

Project name: Internet jako jazykový korpus

Ministerstvo školství, mládeže a tělovýchovy České republiky

Project code: LN00A063

Project name: Centrum počítační lingvistiky

Ministerstvo školství, mládeže a tělovýchovy České republiky

Project code: MSM 0021620838

Project name: Moderní metody, struktury a systémy informatiky

**Subject(s)**

MorphoDiTa

Czech

morphological analysis

morphological generation

PoS tagging

**Collection(s)**

LINDAT / CLARIN Data &amp; Tools



This item is replaced by a newer submission:

<http://hdl.handle.net/11234/1-1836>

Please refer to the submission above for the latest available data. If you nevertheless need the original data, please click [here](#).

List all versions ▼

## PREFER LATEST, PRESERVE ALL

Project name: Internet jako jazykový korpus

Ministerstvo školství, mládeže a tělovýchovy České republiky

Project code: LN00A063

Project name: Centrum počítační lingvistiky

Ministerstvo školství, mládeže a tělovýchovy České republiky

Project code: MSM 0021620838

Project name: Moderní metody, struktury a systémy informatiky

 Subject(s)

MorphoDiTa

Czech

morphological analysis

morphological generation

PoS tagging

 Collection(s)

LINDAT / CLARIN Data & Tools



This item is replaced by a newer submission:

<http://hdl.handle.net/11234/1-1836>

Please refer to the submission above for the latest available data. If you nevertheless need the original data, please click [here](#).

List all versions ▼

# VERSIONING

# PREFER LATEST, PRESERVE ALL

Collection(s) [LINDAT / CLARIN Data & Tools](#)

## Other versions

List all versions ▾

[Show full item record](#)

## Files in this item



Download instructions for command line

This item is **Publicly Available** and licensed under:

Creative Commons - Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)



<b>Name</b>	czech-morfflex-pdt-161115.zip
<b>Size</b>	69.18 MB
<b>Format</b>	application/zip
<b>Description</b>	Czech Models (MorFlex CZ 161115 + PDT 3.0) for MorphoDiTa 161115
<b>MD5</b>	adde38cd363219759e19165b06baa4ce



[Download file](#)

[Preview](#)

# VERSIONING

# PREFER LATEST, PRESERVE ALL

Collection(s) LINDAT / CLARIN Data & Tools

Other versions

List all versions ▾

[Show full item record](#)

Files in this item



Download instructions for command line

This item is **Publicly Available** and licensed under:

Creative Commons - Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)



<b>Name</b>	czech-morfflex-pdt-161115.zip
<b>Size</b>	69.18 MB
<b>Format</b>	application/zip
<b>Description</b>	Czech Models (MorFlex CZ 161115 + PDT 3.0) for MorphoDiTa 161115
<b>MD5</b>	adde38cd363219759e19165b06baa4ce



Download file

Preview

# VERSIONING

# PREFER LATEST, PRESERVE ALL

Collection(s)

LINDAT / CLARIN Data & Tools

Other versions

List all versions ▾

- ▶ Czech Models (Morfflex CZ 161115 + PDT 3.0) for MorphoDiTa 161115
- Czech Models (Morfflex CZ 160310 + PDT 3.0) for MorphoDiTa 160310
- Czech Models (Morfflex CZ + PDT) for MorphoDiTa

[Show full item record](#)

Files in this item



Download instructions for command line

This item is **Publicly Available** and licensed under:

Creative Commons - Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)



<b>Name</b>	czech-morfflex-pdt-161115.zip
<b>Size</b>	69.18 MB
<b>Format</b>	application/zip
<b>Description</b>	Czech Models (Morfflex CZ 161115 + PDT 3.0) for MorphoDiTa 161115
<b>MD5</b>	adde38cd363219759e19165b06baa4ce



[Download file](#)

[Preview](#)

# VERSIONING – TECHNICAL

- Clone and modify previous version
  - auto-fill linking metadata:  
dc.relation.replaces  
dc.relation.isreplacedby
- hide files of older versions and point to the newest
- Promote the latest in search
- No “all-versions PID”



# WHY CLARIN-DSPACE

- Modifications for external assignment of PIDs
- Licensing framework (CC is not everything)
- User Experience (search-centric)
- Admin experience: better control panel
- Citations (Force11 – direct data citations)
- Statistics (global – Matomo; item: graphs, reports)
- Integrations: CLARIN VLO, Clarivate DCI, OpenAIRE, EUDAT, ...

# MUCH MORE

- content negotiation for XML metadata or HTML page
- EUDAT B2SAFE replication
  - external tool linked to DSpace
- summary statistics
- bibtex format for bibliography (for typesetting in LaTeX)
- optional request for user information  
(custom form before download)

The screenshot shows a web browser window displaying the 'Open SDP' item page. The page includes a citation instruction, social sharing options, a metadata table, and a detailed description. A sidebar on the right contains navigation and utility links.

**Open SDP**

Please use the following text to cite this item or export to a predefined format: [BIBTEX](#) [CMDI](#)

Flickinger, Dan; Hajič, Jan; Ivanova, Angelina; et al., 2016, *Open SDP*, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University in Prague, <http://hdl.handle.net/11234/1-1742>.

Share: [f](#) [t](#) [g+](#)

**Authors** Flickinger, Dan ; Hajič, Jan ; Ivanova, Angelina ; Kuhlmann, Marco ; Miyao, Yusuke ; Oepen, Stephan ; Zeman, Daniel

**Project URL** <http://sdp.delph-in.net/>

**Date issued** 2016-06-25

**Type** corpus

**Size** 391062932 bytes

**Language(s)** Czech English ,

**Description**

The original SDP 2014 and 2015 data collections were made available under task-specific 'evaluation' licenses to registered SemEval participants. In mid-2016, all original data has been bundled with system submissions, supporting software, an additional SDP-style collection of semantic dependency graphs, and additional background material (from which some of the SDP target representations were derived) for release through the Linguistic Data Consortium (with LDC catalogue number LDC2016T10).

One of the four English target representations (viz. DM) and the entire Czech data (in the PSD target representation) are not derivative of LDC-licensed annotations and, thus, can be made available for direct download (Open SDP; version 1.1; April 2016) under a more permissive licensing scheme, viz. the Creative Common Attribution-NonCommercial-ShareAlike license. This package also includes some 'richer' meaning representations from which the English bi-lexical DM graphs derive, viz. scope-underspecified logical forms and more abstract, non-lexicalized 'semantic networks'. The latter of these are formally (if not linguistically) similar to Abstract Meaning Representation (AMR) and are available in a range of serializations, including in AMR-like syntax.

**Right Sidebar:**

- LINDAT CLARIN
- What can you do?
- DEPOSIT
- CITE
- Browse
  - > All of the Repository
- My Account
  - Login
- Statistics
  - Statistics **BETA**
- General Information
  - Deposit
  - Cite
  - Submission Lifecycle
  - FAQ
  - About
  - Help Desk

# Item Statistics

- View item visits and downloads in time
- Subscribe to monthly statistics of the item

The screenshot shows a web page for an item titled "Open SDP". The page includes a citation instruction, social sharing options, and a metadata table. On the right, a sidebar menu contains various navigation options, with "Statistics" highlighted by a red circle. The "Statistics" option is accompanied by a bar chart icon and a "BETA" label.

**Open SDP**

Please use the following text to cite this item or export to a predefined format: [BIBTEX](#) [CMDI](#)

Flickinger, Dan; Hajič, Jan; Ivanova, Angelina; et al., 2016, *Open SDP*, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University in Prague, <http://hdl.handle.net/11234/1-1742>.

Share: [f](#) [t](#) [g+](#)

**Authors** Flickinger, Dan ; Hajič, Jan ; Ivanova, Angelina ; Kuhlmann, Marco ; Miyao, Yusuke ; Oepen, Stephan ; Zeman, Daniel

**Project URL** <http://sdp.delph-in.net/>

**Date issued** 2016-06-25

**Type** corpus

**Size** 391062932 bytes

**Language(s)** Czech English ,

**Description**  
The original SDP 2014 and 2015 data collections were made available under task-specific 'evaluation' licenses to registered SemEval participants. In mid-2016, all original data has been bundled with system submissions, supporting software, an additional SDP-style collection of semantic dependency graphs, and additional background material (from which some of the SDP target representations were derived) for release through the Linguistic Data Consortium (with LDC catalogue number LDC2016T10).  
  
One of the four English target representations (viz. DM) and the entire Czech data (in the PSD target representation) are not derivative of LDC-licensed annotations and, thus, can be made available for direct download (Open SDP; version 1.1; April 2016) under a more permissive licensing scheme, viz. the Creative Common Attribution-NonCommercial-ShareAlike license. This package also includes some 'richer' meaning representations from which the English bi-lexical DM graphs derive, viz. scope-underspecified logical forms and more abstract, non-lexicalized 'semantic networks'. The latter of these are formally (if not linguistically) similar to Abstract Meaning Representation (AMR) and are available in a range of serializations, including in AMR-like syntax.

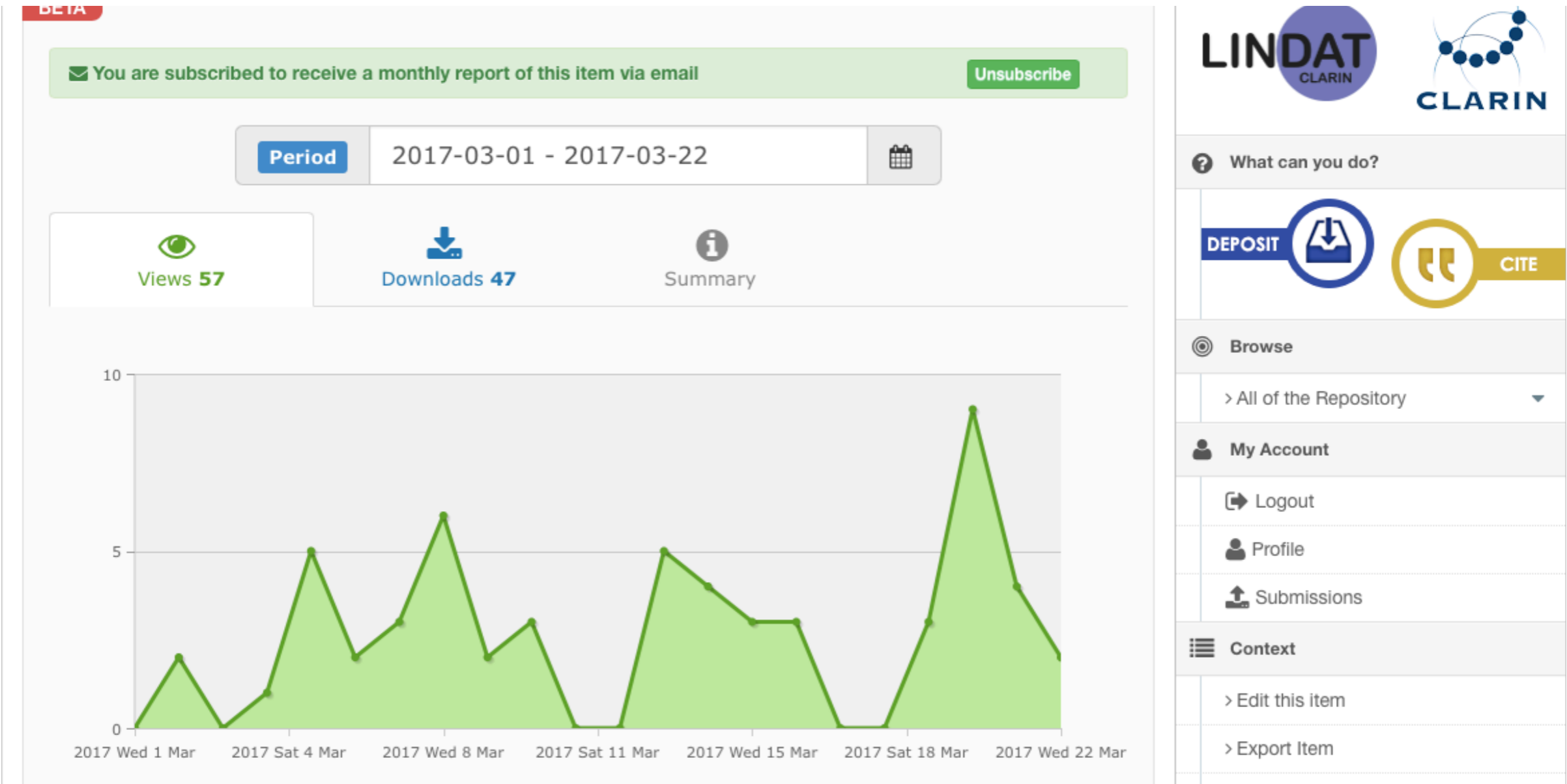
**Statistics** **BETA**

**General Information**

- Deposit
- Cite
- Submission Lifecycle
- FAQ
- About
- Help Desk

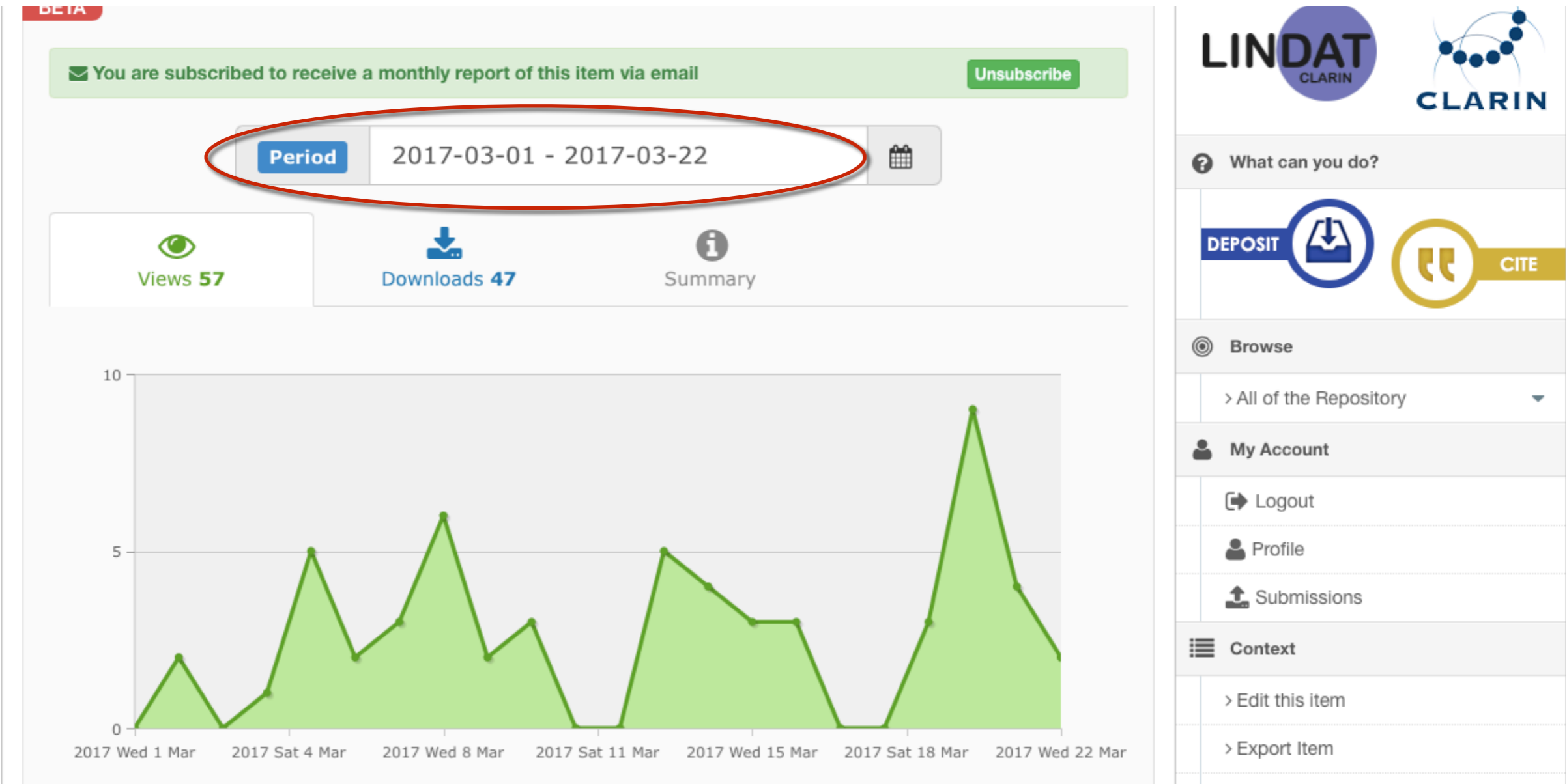
# Item Statistics

- View item visits and downloads in time
- Subscribe to monthly statistics of the item



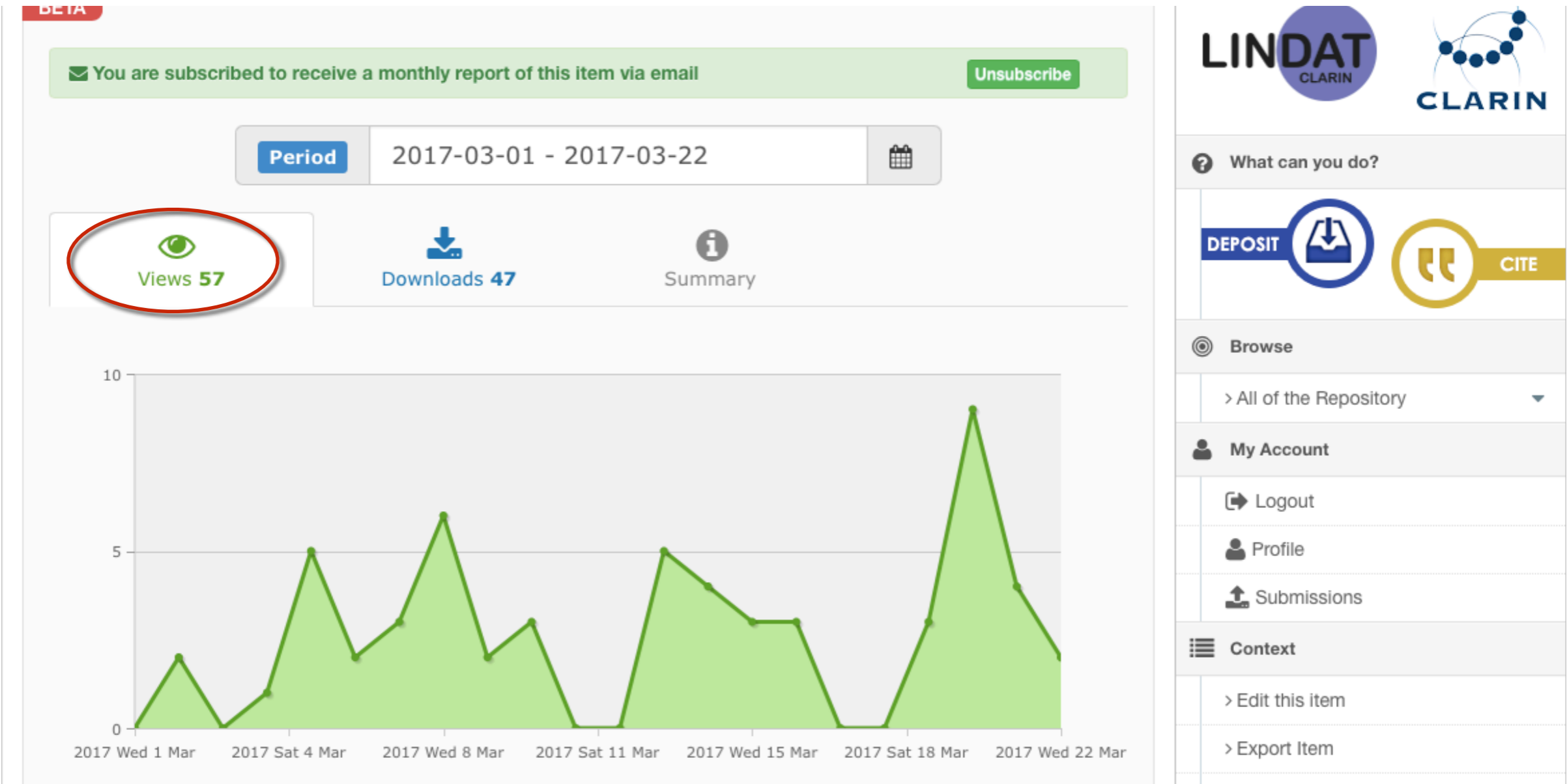
# Item Statistics

- View item visits and downloads in time
- Subscribe to monthly statistics of the item



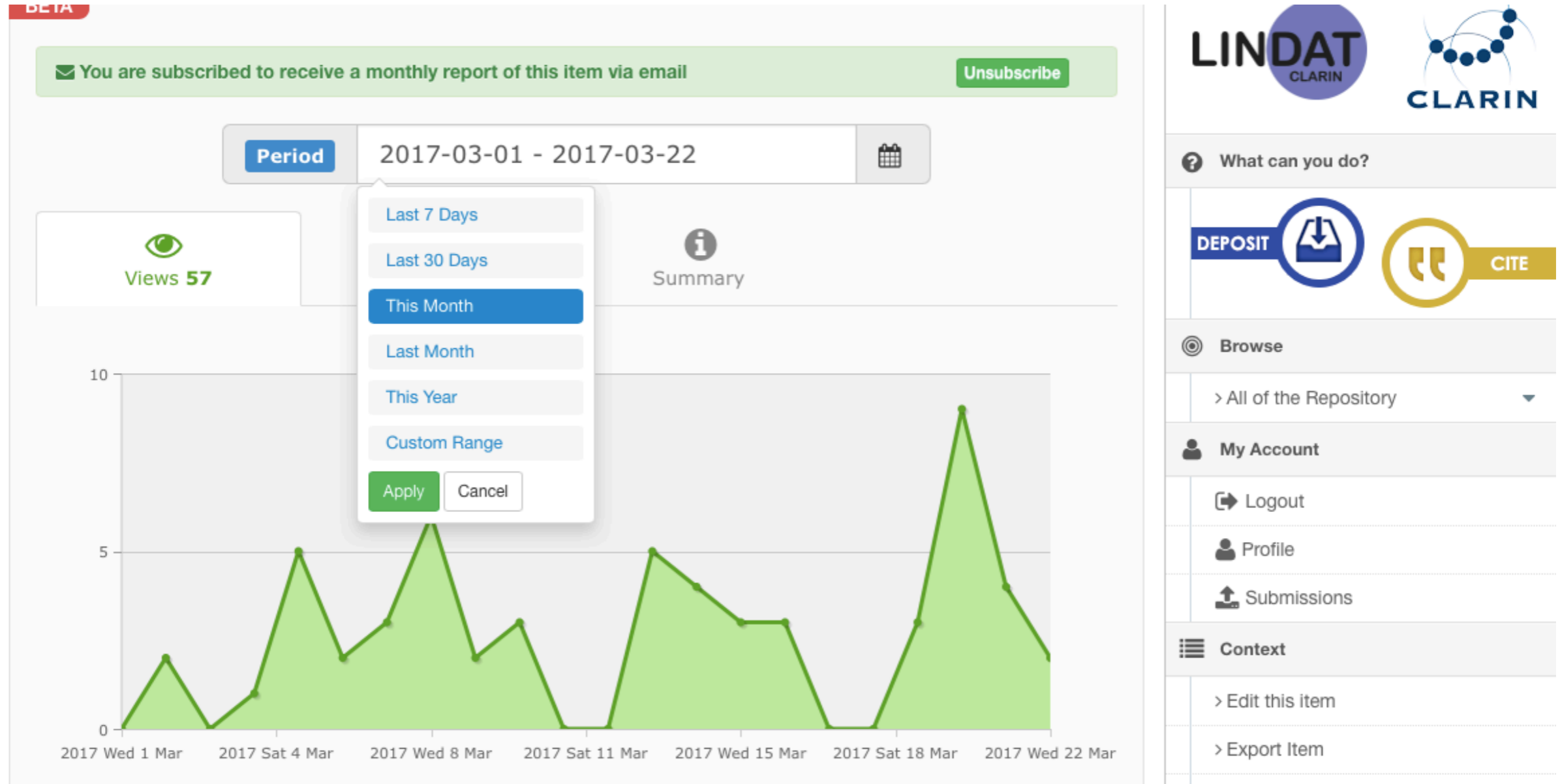
# Item Statistics

- View item visits and downloads in time
- Subscribe to monthly statistics of the item



# Item Statistics

- View item visits and downloads in time
- Subscribe to monthly statistics of the item



# Item Statistics

- View item visits and downloads in time
- Subscribe to monthly statistics of the item



**BETA**

✉ Get a monthly report for this item via email Subscribe

Period 2016-01-01 - 2017-01-01

Views **2036**
 Downloads **1540**
 Summary

**2036** pageviews, **1493** unique pageviews.  
**1438** visits, **1321** unique visitors.  
**1540** downloads, **1237** unique downloads.

What can you do?

DEPOSIT
 CITE

Browse

> All of the Repository

My Account

Logout  
 Profile  
 Submissions

Context

> Edit this item  
 > Export Item

# Item Statistics

- View item visits and downloads in time
- Subscribe to monthly statistics of the item

**BETA**

✉ Get a monthly report for this item via email [Subscribe](#)

Period 2016-01-01 - 2017-01-01

Views **2036** Downloads **1540** Summary

**2036** pageviews, **1493** unique pageviews.  
**1438** visits, **1321** unique visitors.  
**1540** downloads, **1237** unique downloads.

**LINDAT** CLARIN

What can you do?

**DEPOSIT** **CITE**

Browse

> All of the Repository

My Account

Logout

Profile

Submissions

Context

> Edit this item

> Export Item

# Item Statistics

- View item visits and downloads in time
- Subscribe to monthly statistics of the item

BETA

Get a monthly report for this item via email [Subscribe](#)

Period 2016-01-01 - 2017-01-01

Views **2036** Downloads **1540** Summary

2036 pageviews, 1493 unique pageviews.  
1438 visits, 1321 unique visitors.  
1540 downloads, 1237 unique downloads.

LINDAT CLARIN

What can you do?

DEPOSIT CITE

Browse

> All of the Repository

My Account

Logout Profile Submissions

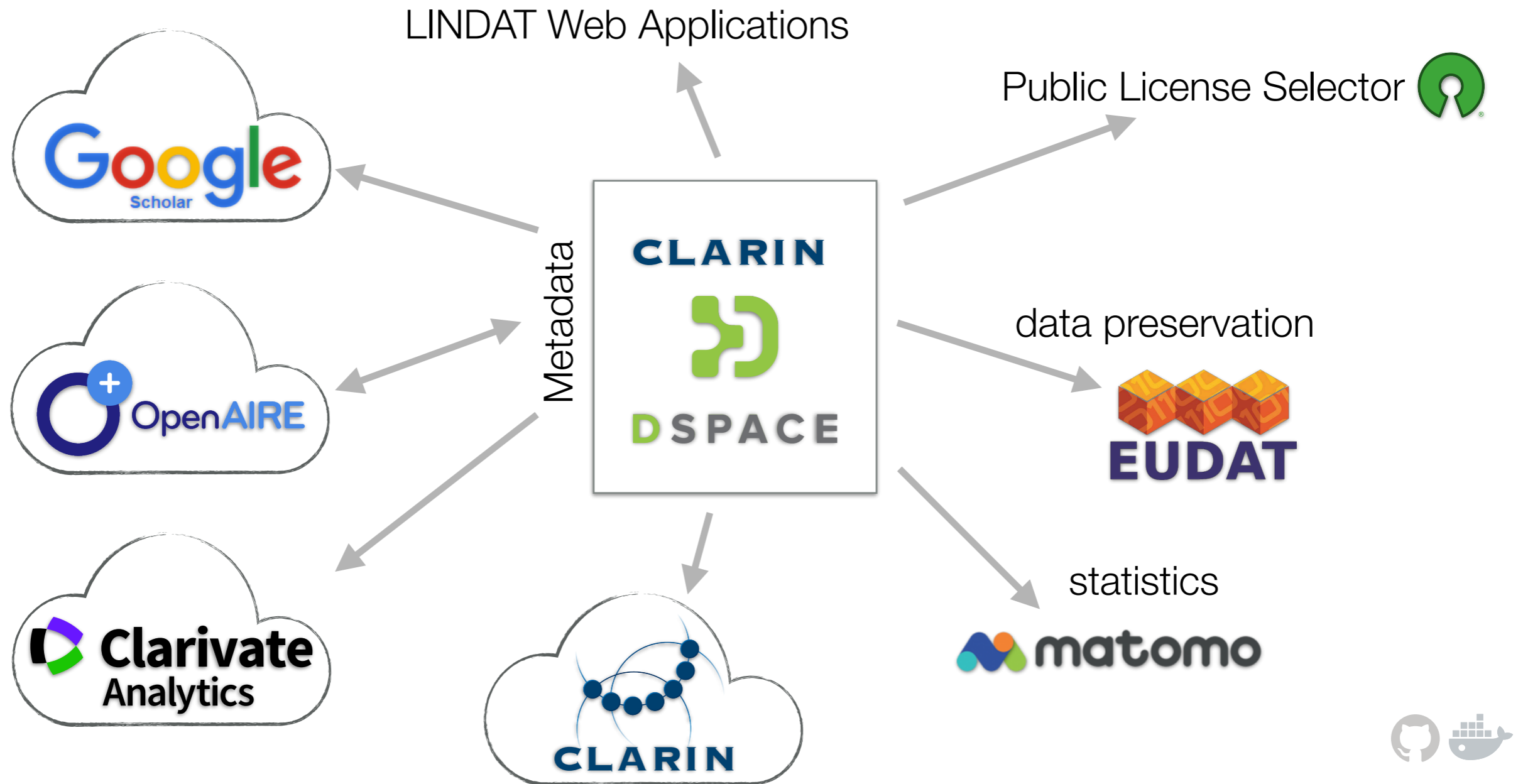
Context

> Edit this item  
> Export Item

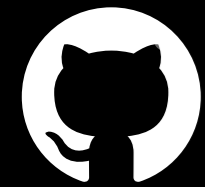
# Item Statistics

- View item visits and downloads in time
- Subscribe to monthly statistics of the item

# Integrations



# PUBLIC LICENSE SELECTOR



<https://github.com/ufal/public-license-selector>

- public license: no signatures, public distribution
- data / software
- explanations provided
- choose as open as possible
- open source / open data
  - best licenses chosen



Pull requests Issues Marketplace Explore

ufal / public-license-selector

<> Code Issues 6 Pull requests 0 Projects 0 Wiki

Tool that will help you select the right open license for your data or software

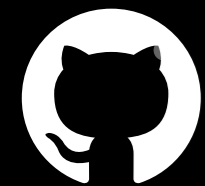
95 commits 3 branches 5 releases

Branch: master New pull request

stranak Merge pull request #9 from Aldarone/Aldarone-patch-1

ci	Removed remnants of Clarin version and changed b
src	Change software copyleft question wording to be o
.editorconfig	Removed remnants of Clarin version and changed b
.gitignore	Removed remnants of Clarin version and changed b
.mversionrc	Removed remnants of Clarin version and changed b
.travis.yml	Fix year in license file and set node version to stabl
LICENSE.md	Fix year in license file and set node version to stabl
Makefile	Removed remnants of Clarin version and changed b
README.md	Update README.md
bower.json	Releasing 0.0.6
index.html	Removed remnants of Clarin version and changed b
package.json	Releasing 0.0.6
webpack.config.js	Removed remnants of Clarin version and changed b

# CLARIN-DSPACE



<https://github.com/ufal/clarin-dspace>

- MIT license
- DSpace + licensing, versioning and more
- LINDAT's project converting to community
- Issues, Documentation
- 15+ deployments  
10+ countries
- since 24 October 2009



This screenshot shows the GitHub repository page for 'ufal / clarin-dspace', which is a fork of 'DSpace/DSpace'. The repository has 8,658 commits and 43 branches. The current branch is 'clarin', which is 1368 commits ahead and 2394 commits behind the 'DSpace:master' branch. A recent merge by user 'kosarko' is shown, merging the 'release-2018.01' branch into 'clarin'. Below the merge, a list of sub-projects is displayed, each with a link to its respective issue or pull request.

Sub-project	Issue/Pull Request
<a href="#">dspace-api</a>	<a href="#">Resolves #412 - Access other versions</a>
<a href="#">dspace-jspui</a>	<a href="#">Resolves #816 - Merge DSpace-5.8 (#</a>
<a href="#">dspace-lni</a>	<a href="#">Resolves #816 - Merge DSpace-5.8 (#</a>
<a href="#">dspace-oai</a>	<a href="#">Resolves #816 - Merge DSpace-5.8 (#</a>
<a href="#">dspace-rdf</a>	<a href="#">Resolves #816 - Merge DSpace-5.8 (#</a>
<a href="#">dspace-rest</a>	<a href="#">Resolves #412 - Access other versions</a>
<a href="#">dspace-services</a>	<a href="#">Resolves #816 - Merge DSpace-5.8 (#</a>
<a href="#">dspace-solr</a>	<a href="#">Resolves #816 - Merge DSpace-5.8 (#</a>
<a href="#">dspace-sword</a>	<a href="#">Resolves #816 - Merge DSpace-5.8 (#</a>
<a href="#">dspace-swordv2</a>	<a href="#">Resolves #816 - Merge DSpace-5.8 (#</a>

# FAIR SUMMARY

**F***indable*: Google, Google Scholar, Data Citation Index, CLARIN VLO, OLAC... and the repository itself

**A***ccessible*: records with data (even when restricted), complete licensing, Open Access (Public License Selector), login only when needed, CESNET, EUDAT

**I***nteroperable*: common data formats, full documentation (enhanced metadata, documentation bitstreams)

**R***eusable*: records with data, complete licensing, full versioning, direct data citations, maximal OA

Thank you!

<http://lindat.cz>

