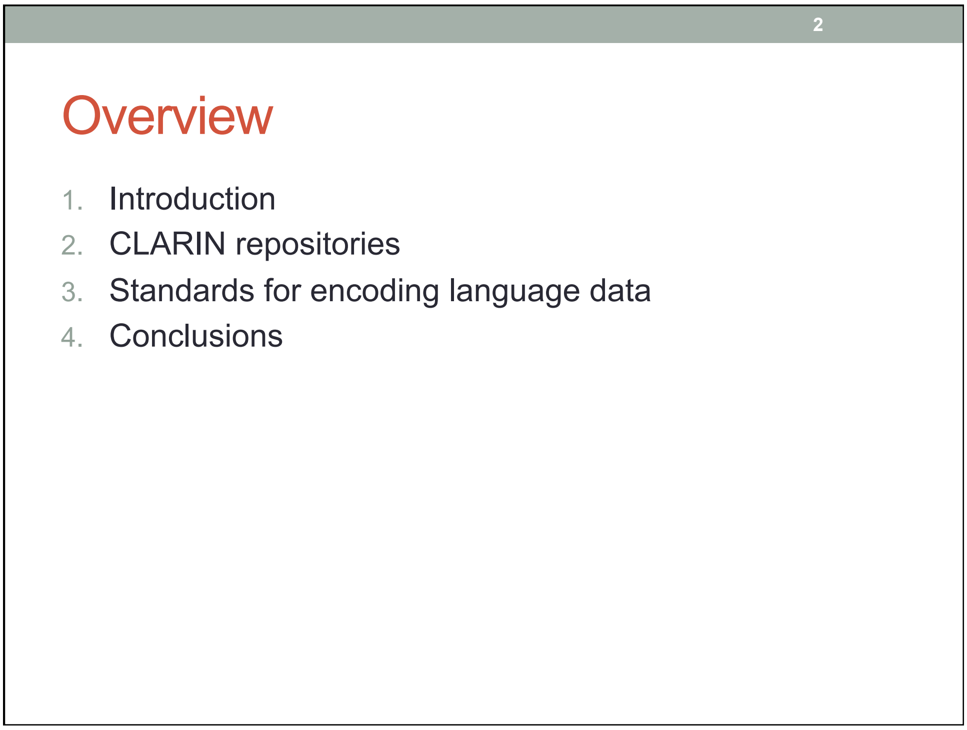
Platforms and standards for data sharing

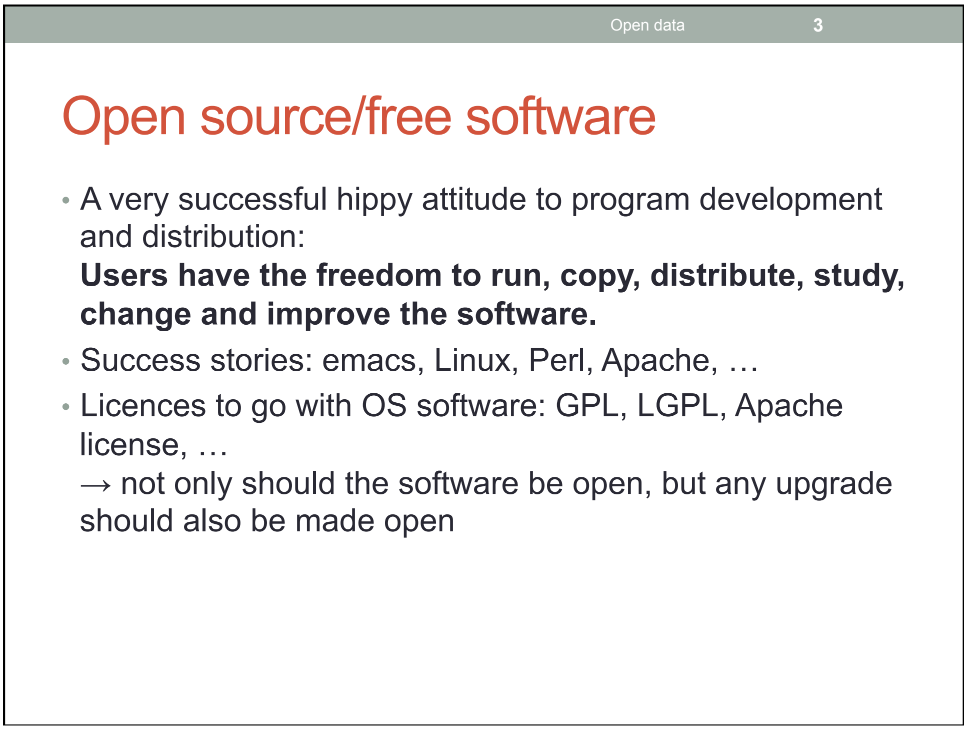
Tomaž Erjavec
Dept. of Knowledge Technologies
Jožef Stefan Institute
Ljubljana

How to make data reusable?
UFSP Sprache und Raum
May 2015



Overview

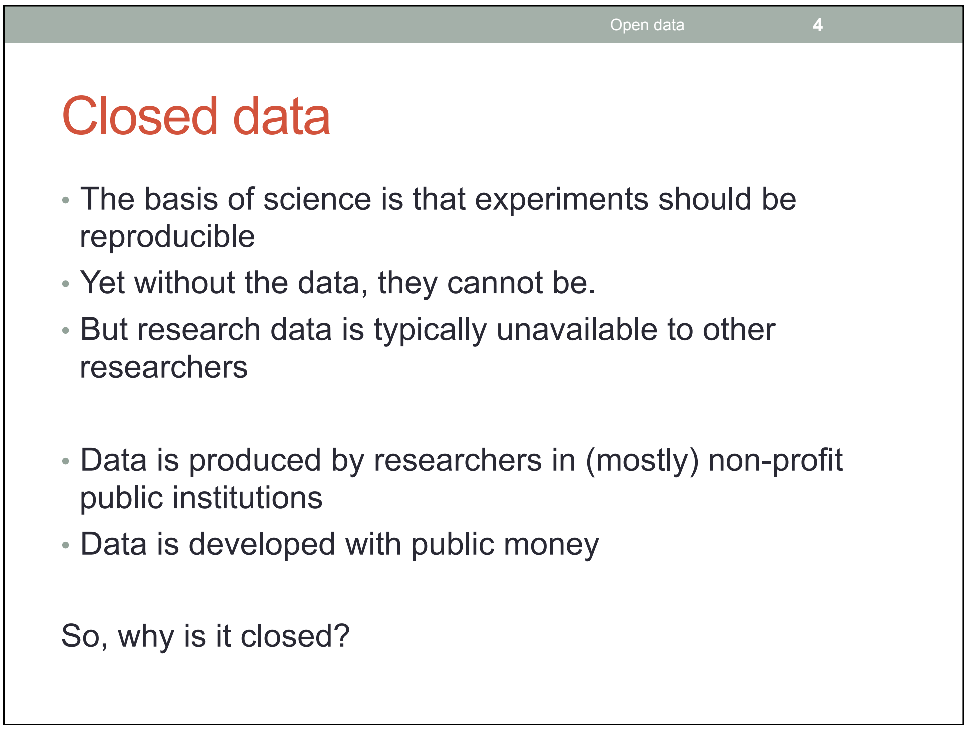
1. Introduction
2. CLARIN repositories
3. Standards for encoding language data
4. Conclusions



Open data 3

Open source/free software

- A very successful hippy attitude to program development and distribution:
  Users have the freedom to run, copy, distribute, study, change and improve the software.
- Success stories: emacs, Linux, Perl, Apache, …
- Licences to go with OS software: GPL, LGPL, Apache license, …
  → not only should the software be open, but any upgrade should also be made open



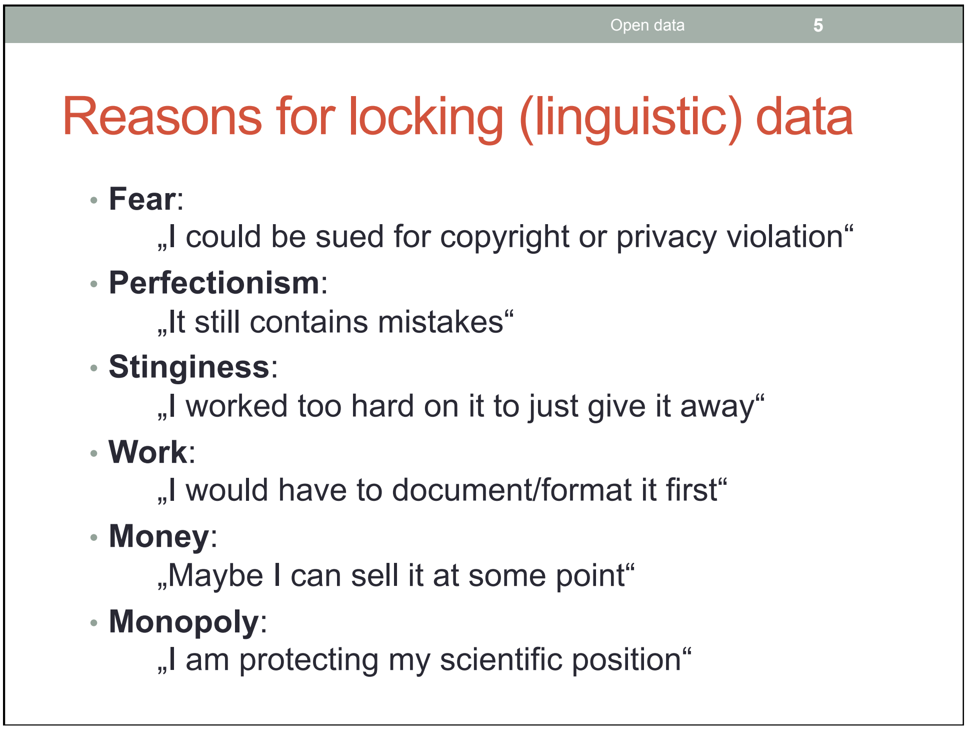
Open data 4

Closed data

- The basis of science is that experiments should be reproducible
- Yet without the data, they cannot be.
- But research data is typically unavailable to other researchers

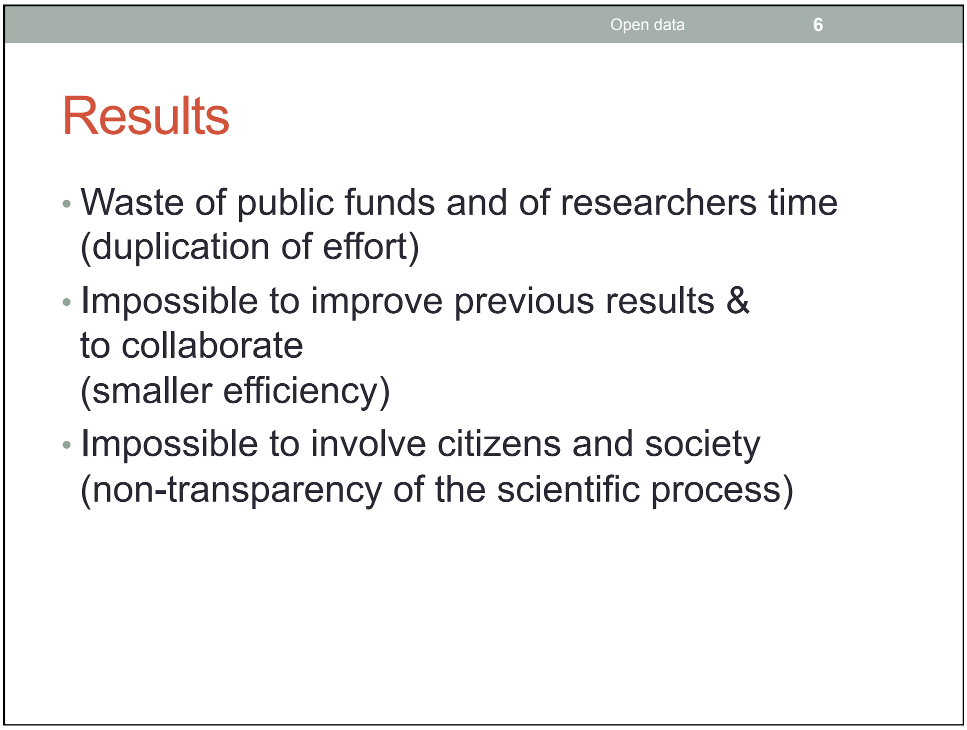
- Data is produced by researchers in (mostly) non-profit public institutions
- Data is developed with public money

So, why is it closed?



Reasons for locking (linguistic) data

- **Fear**:
  „I could be sued for copyright or privacy violation“
- **Perfectionism**:
  „It still contains mistakes“
- **Stinginess**:
  „I worked too hard on it to just give it away“
- **Work**:
  „I would have to document/format it first“
- **Money**:
  „Maybe I can sell it at some point“
- **Monopoly**:
  „I am protecting my scientific position“



Results

- Waste of public funds and of researchers time (duplication of effort)
- Impossible to improve previous results & to collaborate
  (smaller efficiency)
- Impossible to involve citizens and society (non-transparency of the scientific process)

Changing times

Open text repositories:

- MediaWiki, Google Books, OLAC, …

H2020:

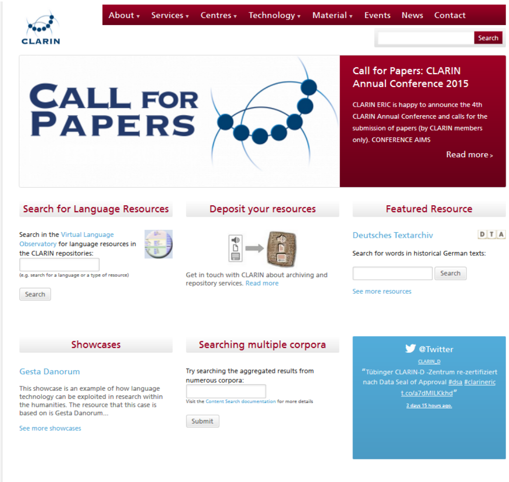
- Open data and publications are a requirement
- This policy is being adopted by EU member states

Research infrastructures:

- EU instrument for establishing long term facilities, resources and related services in order to support research
- Humanities and social sciences: DARIAH, CLARIN

II. Common Language Resource Infrastructure

- CLARIN ERIC: EU legal entity
- 13 national consortia (sites)
- From May 5th, 2015: also Slovenia



Call for Papers: CLARIN Annual Conference 2015

CLARIN ERIC is happy to announce the 4th CLARIN Annual Conference and calls for the submission of papers (by CLARIN members only). CONFERENCE AIMS

Read more ›

Search for Language Resources

Search in the Virtual Language Observatory for language resources in the CLARIN repositories:

(e.g. search for a language or a type of resource)

Deposit your resources

Get in touch with CLARIN about archiving and repository services. Read more

Featured Resource

Deutsches Textarchiv

Search for words in historical German texts:

See more resources

Showcases

Gesta Danorum

This showcase is an example of how language technology can be exploited in research within the humanities. The resource that this case is based on is Gesta Danorum…

See more showcases

Searching multiple corpora

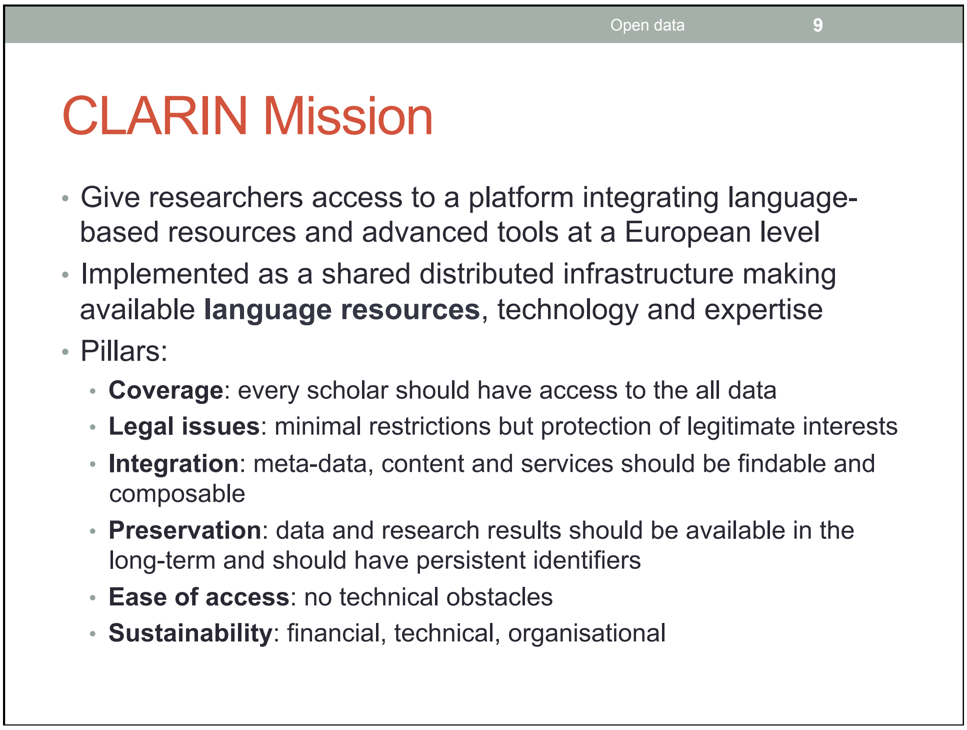
Try searching the aggregated results from numerous corpora:

Visit the Content Search documentation for more details

CLARIN_D

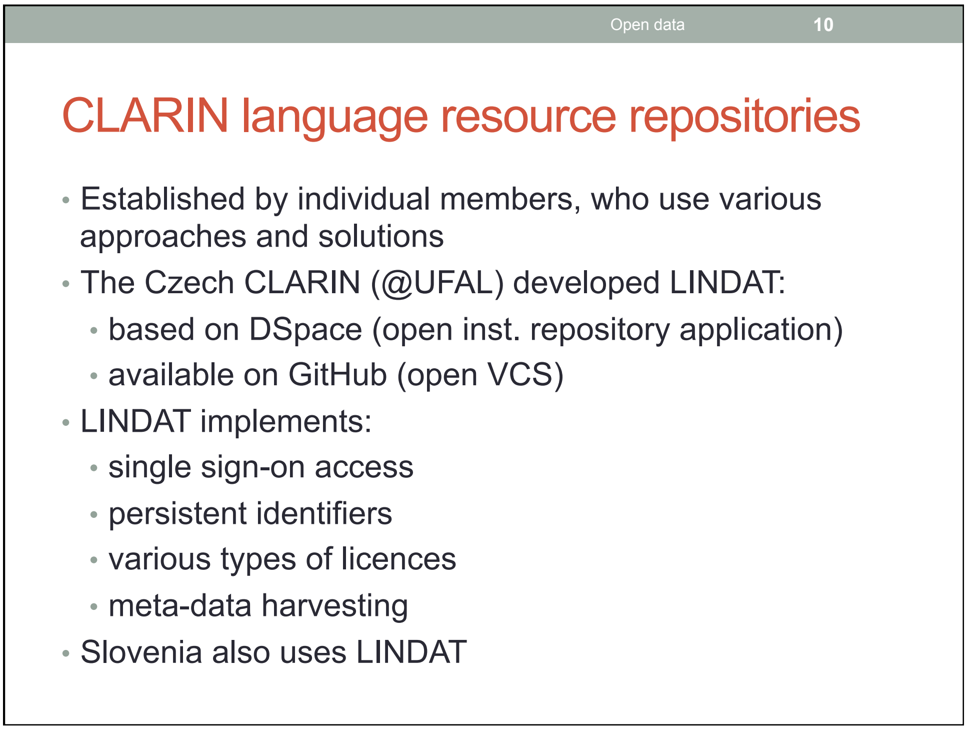
"Tübinger CLARIN-D -Zentrum re-zertifiziert nach Data Seal of Approval #dsa #clarineric t.co/a7dMILKkhd"

3 days 15 hours ago.



CLARIN Mission

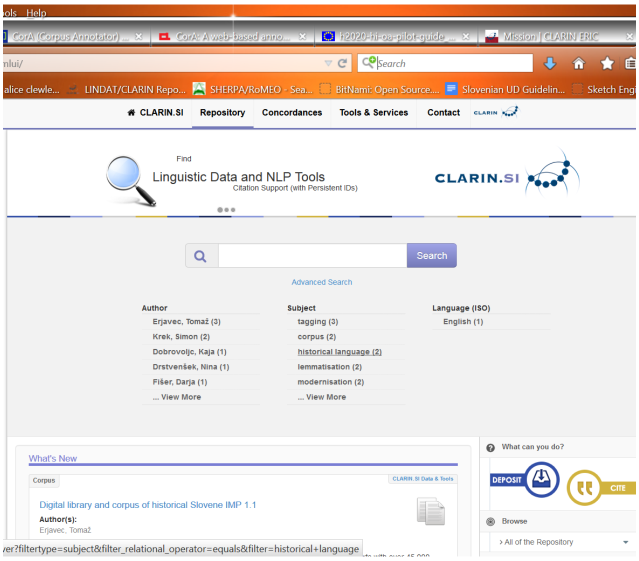
- Give researchers access to a platform integrating language-based resources and advanced tools at a European level
- Implemented as a shared distributed infrastructure making available **language resources**, technology and expertise
- Pillars:
  - **Coverage**: every scholar should have access to the all data
  - **Legal issues**: minimal restrictions but protection of legitimate interests
  - **Integration**: meta-data, content and services should be findable and composable
  - **Preservation**: data and research results should be available in the long-term and should have persistent identifiers
  - **Ease of access**: no technical obstacles
  - **Sustainability**: financial, technical, organisational



CLARIN language resource repositories

- Established by individual members, who use various approaches and solutions
- The Czech CLARIN (@UFAL) developed LINDAT:
  - based on DSpace (open inst. repository application)
  - available on GitHub (open VCS)
- LINDAT implements:
  - single sign-on access
  - persistent identifiers
  - various types of licences
  - meta-data harvesting
- Slovenia also uses LINDAT

CLARIN.SI repository



Find
Linguistic Data and NLP Tools
Citation Support (with Persistent IDs)

Author	Subject	Language (ISO)
Erjavec, Tomaž (3)	tagging (3)	English (1)
Krek, Simon (2)	corpus (2)	
Dobrovoljc, Kaja (1)	historical language (2)	
Drstvenšek, Nina (1)	lemmatisation (2)	
Fišer, Darja (1)	modernisation (2)	
... View More	... View More	

What's New

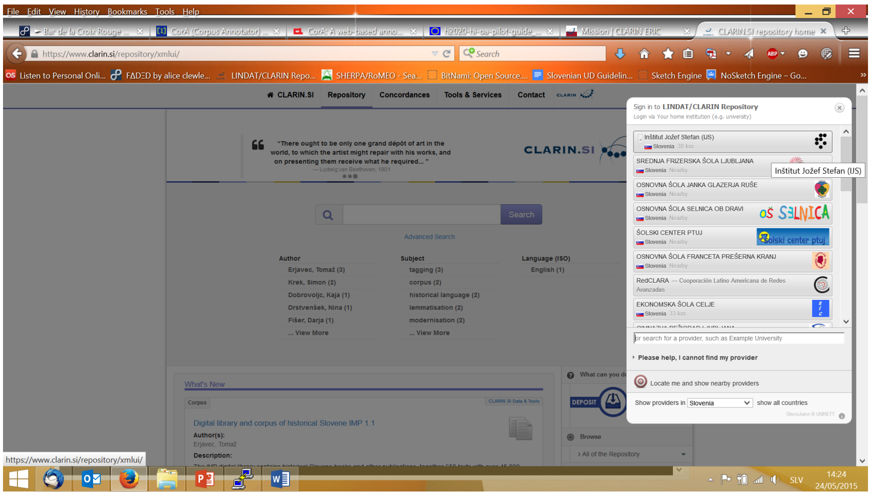
Corpus — CLARIN.SI Data & Tools

Digital library and corpus of historical Slovene IMP 1.1

Author(s):
Erjavec, Tomaž

ver?filtertype=subject&filter_relational_operator=equals&filter=historical+language

Single (AAI) sign in



"There ought to be only one grand dépôt of art in the world, to which the artist might repair with his works, and on presenting them receive what he required..." — Ludwig van Beethoven, 1801

Author	Subject	Language (ISO)
Erjavec, Tomaž (3)	tagging (3)	English (1)
Krek, Simon (2)	corpus (2)	
Dobrovoljc, Kaja (1)	historical language (2)	
Drstvenšek, Nina (1)	lemmatisation (2)	
Fišer, Darja (1)	modernisation (2)	
... View More	... View More	

What's New

Corpus — CLARIN.SI Data & Tools

Digital library and corpus of historical Slovene IMP 1.1

Author(s):
Erjavec, Tomaž

Description:

Sign in to **LINDAT/CLARIN Repository**
Login via Your home institution (e.g. university)

- Inštitut Jožef Stefan (IJS) — Slovenia 38 km
- SREDNJA FRIZERSKA ŠOLA LJUBLJANA — Slovenia Nearby
- OSNOVNA ŠOLA JANKA GLAZERJA RUŠE — Slovenia Nearby
- OSNOVNA ŠOLA SELNICA OB DRAVI — Slovenia Nearby
- ŠOLSKI CENTER PTUJ — Slovenia Nearby
- OSNOVNA ŠOLA FRANCETA PREŠERNA KRANJ — Slovenia Nearby
- RedCLARA — Cooperación Latino Americana de Redes Avanzadas
- EKONOMSKA ŠOLA CELJE — Slovenia 33 km

Inštitut Jožef Stefan (IJS)

or search for a provider, such as Example University

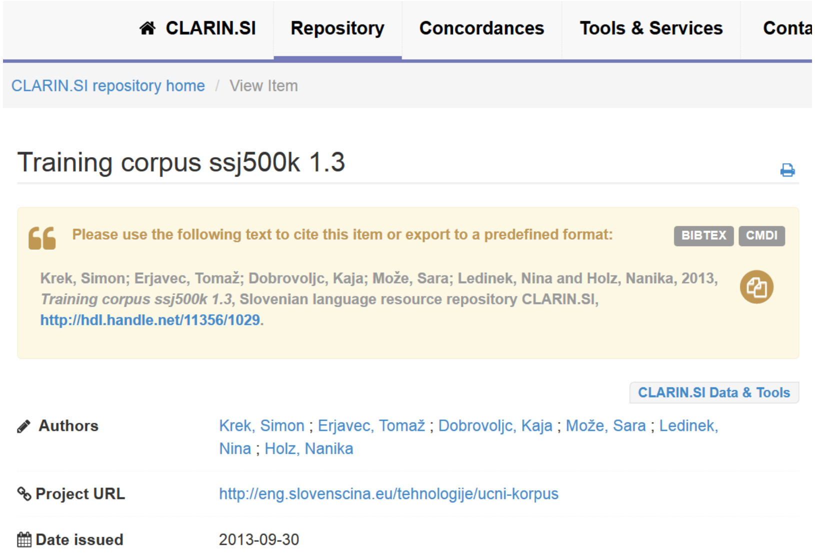
› **Please help, I cannot find my provider**

Locate me and show nearby providers

Show providers in Slovenia show all countries

DiscoJuice © UNINETT

Citation and persistent ID



CLARIN.SI repository home / View Item

Training corpus ssj500k 1.3

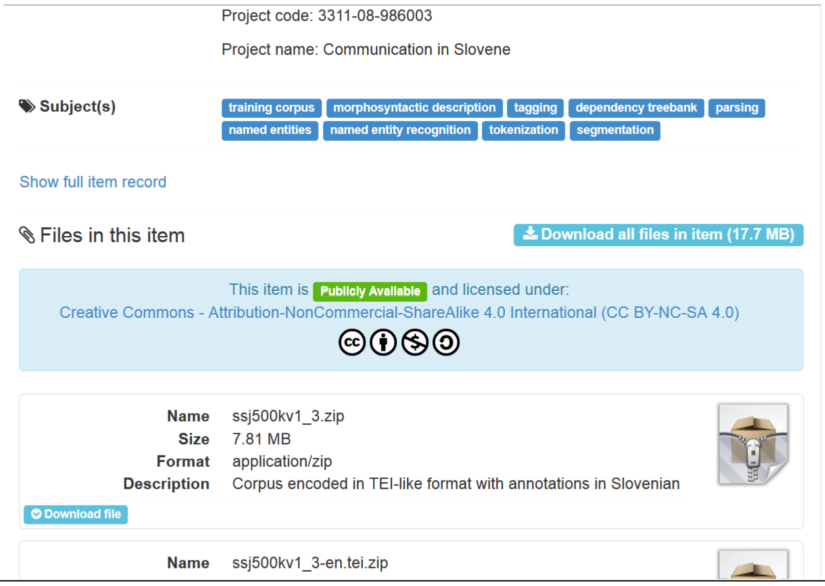
Please use the following text to cite this item or export to a predefined format: BIBTEX CMDI

Krek, Simon; Erjavec, Tomaž; Dobrovoljc, Kaja; Može, Sara; Ledinek, Nina and Holz, Nanika, 2013, *Training corpus ssj500k 1.3*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1029.

CLARIN.SI Data & Tools

Authors	Krek, Simon ; Erjavec, Tomaž ; Dobrovoljc, Kaja ; Može, Sara ; Ledinek, Nina ; Holz, Nanika
Project URL	http://eng.slovenscina.eu/tehnologije/ucni-korpus
Date issued	2013-09-30

Licence and download



Project code: 3311-08-986003

Project name: Communication in Slovene

Subject(s) training corpus | morphosyntactic description | tagging | dependency treebank | parsing | named entities | named entity recognition | tokenization | segmentation

Show full item record

Files in this item

Download all files in item (17.7 MB)

This item is Publicly Available and licensed under:
Creative Commons - Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

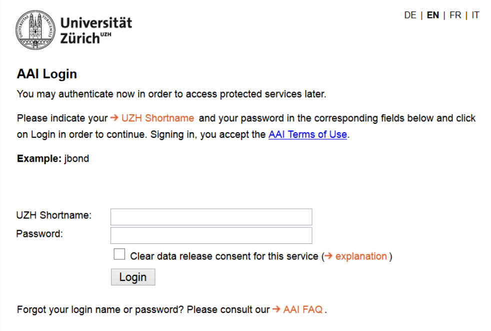
Name	ssj500kv1_3.zip
Size	7.81 MB
Format	application/zip
Description	Corpus encoded in TEI-like format with annotations in Slovenian

Download file

Name	ssj500kv1_3-en.tei.zip

Benefits of depositing

- Safe place for the data
- Maintained platform
- Licenced download
- Citation support
- Discoverable – meta-data harvesting
  - CLARIN ERIC
  - OLAC
  - Thomson Reuters



DE | **EN** | FR | IT

Universität Zürich^UZH

AAI Login

You may authenticate now in order to access protected services later.

Please indicate your → UZH Shortname and your password in the corresponding fields below and click on Login in order to continue. Signing in, you accept the AAI Terms of Use.

Example: jbond

UZH Shortname:

Password:

□ Clear data release consent for this service (→ explanation)

Login

Forgot your login name or password? Please consult our → AAI FAQ .

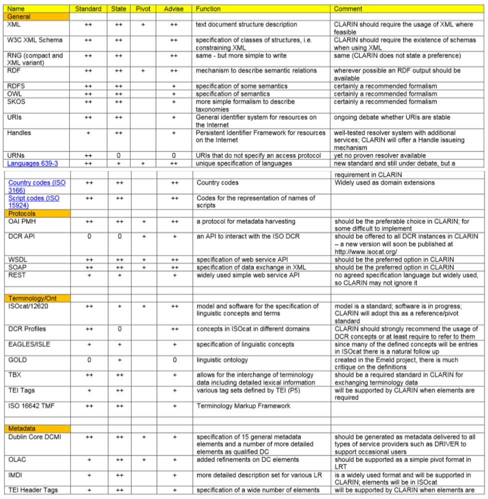
II. Standards for encoding language data

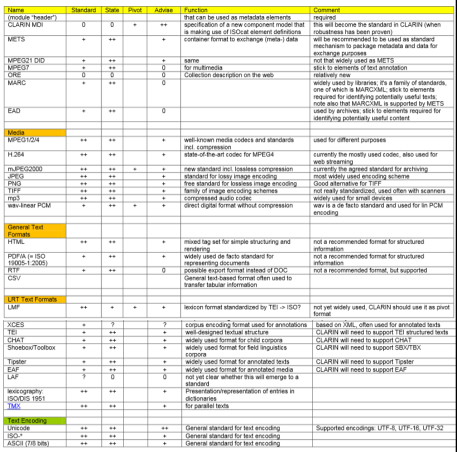
- Bad practice:
  - Data (text and annotations) is in an proprietary and undocumented encoding, tied to a particular piece of software
- Standards exist to make (textual) data
  - Interchangeable: others can use it, on a different platform
  - Reusable: for a different purpose
  - Permanent: for a long time after you made it
- Good practice:
  - Data is stored in an open, documented, maintained and machine-independent format, i.e. it uses standards.

Problem

„The nice thing about standards is that you have so many to choose from.“

List of CLARIN standard recommentdations:





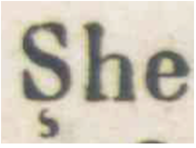
Name	Standard	State	Pivot	Advise	Function	Comment
General						
XML	++	++	+	++	text document structure description	CLARIN should require the usage of XML where feasible
W3C XML Schema	++	++		++	specification of classes of structures, i.e. constraining XML	CLARIN should require the existence of schemas when using XML
RNG (compact and XML variant)	++	++		++	same - but more simple to write	same (CLARIN does not state a preference)
RDF	++	++	+	++	mechanism to describe semantic relations	wherever possible an RDF output should be available
RDFS	++	++		+	specification of some semantics	certainly a recommended formalism
OWL	++	++		+	specification of semantics	certainly a recommended formalism
SKOS	++	++		+	more simple formalism to describe taxonomies	certainly a recommended formalism
URIs	++	++		+	General identifier system for resources on the Internet	ongoing debate whether URIs are stable
Handles	+	++		+	Persistent Identifier Framework for resources on the Internet	well-tested resolver system with additional services; CLARIN will offer a Handle issueing mechanism
URNs	++	0		0	URIs that do not specify an access protocol	yet no proven resolver available
Languages 639-3	++	+	+	++	unique specification of languages	new standard and still under debate, but a requirement in CLARIN
Country codes (ISO 3166)	++	++		++	Country codes	Widely used as domain extensions
Script codes (ISO 15924)	++	++		++	Codes for the representation of names of scripts	
Protocols						
OAI PMH	++	++	+	++	a protocol for metadata harvesting	should be the preferable choice in CLARIN; for some difficult to implement
DCR API	0	0	+	+	an API to interact with the ISO DCR	should be offered to all DCR instances in CLARIN – a new version will soon be published at http://www.isocat.org/
WSDL	++	++	+	++	specification of web service API	should be the preferred option in CLARIN
SOAP	++	++	+	++	specification of data exchange in XML	should be the preferred option in CLARIN
REST	+	+		+	widely used simple web service API	no agreed specification language but widely used, so CLARIN may not ignore it
Terminology/Ont						
ISOcat/12620	++	+	+	++	model and software for the specification of linguistic concepts and terms	model is a standard; software is in progress; CLARIN will adopt this as a reference/pivot standard
DCR Profiles	++	0		++	concepts in ISOcat in different domains	CLARIN should strongly recommend the usage of DCR concepts or at least require to refer to them
EAGLES/ISLE	+	+		+	specification of linguistic concepts	since many of the defined concepts will be entries in ISOcat there is a natural follow up
GOLD	0	+		0	linguistic ontology	created in the Emeld project, there is much critique on the definitions
TBX	++	++		+	allows for the interchange of terminology data including detailed lexical information	should be a required standard in CLARIN for exchanging terminology data
TEI Tags	+	++		+	various tag sets defined by TEI (P5)	will be supported by CLARIN when elements are required
ISO 16642 TMF	++	++		+	Terminology Markup Framework	
Metadata						
Dublin Core DCMI	++	++	+	+	specification of 15 general metadata elements and a number of more detailed elements as qualified DC	should be generated as metadata delivered to all types of service providers such as DRIVER to support occasional users
OLAC	+	++	+	+	added refinements on DC elements	should be supported as a simple pivot format in LRT
IMDI	+	++		+	more detailed description set for various LR	is a widely used format and will be supported in CLARIN; elements will be in ISOcat
TEI Header Tags (module "header")	+	++		+	specification of a wide number of elements that can be used as metadata elements	will be supported by CLARIN when elements are required
CLARIN MDI	0	0	+	++	specification of a new component model that is making use of ISOcat element definitions	this will become the standard in CLARIN (when robustness has been proven)
METS	+	++		+	container format to exchange (meta-) data	will be recommended to be used as standard mechanism to package metadata and data for exchange purposes
MPEG21 DID	+	++		+	same	not that widely used as METS
MPEG7	+	++		0	for multimedia	stick to elements of text annotation
ORE	0	0		0	Collection description on the web	relatively new
MARC	+	++		0		widely used by libraries; it's a family of standards, one of which is MARCXML; stick to elements required for identifying potentially useful texts; note also that MARCXML is supported by METS
EAD	+	++		0		used by archives; stick to elements required for identifying potentially useful content
Media						
MPEG1/2/4	++	++		+	well-known media codecs and standards incl. compression	used for different purposes
H.264	++	++		+	state-of-the-art codec for MPEG4	currently the mostly used codec, also used for web streaming
mJPEG2000	++	++	+	+	new standard incl. lossless compression	currently the agreed standard for archiving
JPEG	++	++		+	standard for lossy image encoding	most widely used encoding scheme
PNG	++	++		+	free standard for lossless image encoding	Good alternative for TIFF
TIFF	++	++		+	family of image encoding schemes	not really standardized, used often with scanners
mp3	++	++		+	compressed audio codec	widely used for small devices
wav-linear PCM	+	++	+	+	direct digital format without compression	wav is a de facto standard and used for lin PCM encoding
General Text Formats						
HTML	++	++		+	mixed tag set for simple structuring and rendering	not a recommended format for structured information
PDF/A (= ISO 19005-1:2005)	+	++		+	widely used de facto standard for representing documents	not a recommended format for structured information
RTF	+	++		0	possible export format instead of DOC	not a recommended format, but supported
CSV					General text-based format often used to transfer tabular information	
LRT Text Formats						
LMF	++	+	+	+	lexicon format standardized by TEI -> ISO?	not yet widely used, CLARIN should use it as pivot format
XCES	+	?		?	corpus encoding format used for annotations	based on XML, often used for annotated texts
TEI	+	++		+	well-designed textual structure	CLARIN will need to support TEI structured texts
CHAT	+	++		+	widely used format for child corpora	CLARIN will need to support CHAT
Shoebox/Toolbox	+	++		+	widely used format for field linguistics corpora	CLARIN will need to support SBX/TBX
Tipster	+	++		+	widely used format for annotated texts	CLARIN will need to support Tipster
EAF	+	++		+	widely used format for annotated media	CLARIN will need to support EAF
LAF	?	0		0	not yet clear whether this will emerge to a standard	
lexicography: ISO/DIS 1951	++	++		+	Presentation/representation of entries in dictionaries	
TMX	++	++		+	for parallel texts	
Text Encoding						
Unicode	++	++		++	General standard for text encoding	Supported encodings: UTF-8, UTF-16, UTF-32
ISO-*	++	++		+	General standard for text encoding	
ASCII (7/8 bits)	++	++		+	General standard for text encoding	

Ladder of standards

- **Character set**: How are characters encoded?
- **Format**: What distinguishes annotations from the text?
- **Schema**: Which annotations does the document use?
- **Metadata**: How is the information about the document encoded?
- **Linguistic categories**: What is the vocabulary of linguistic features?

Character sets

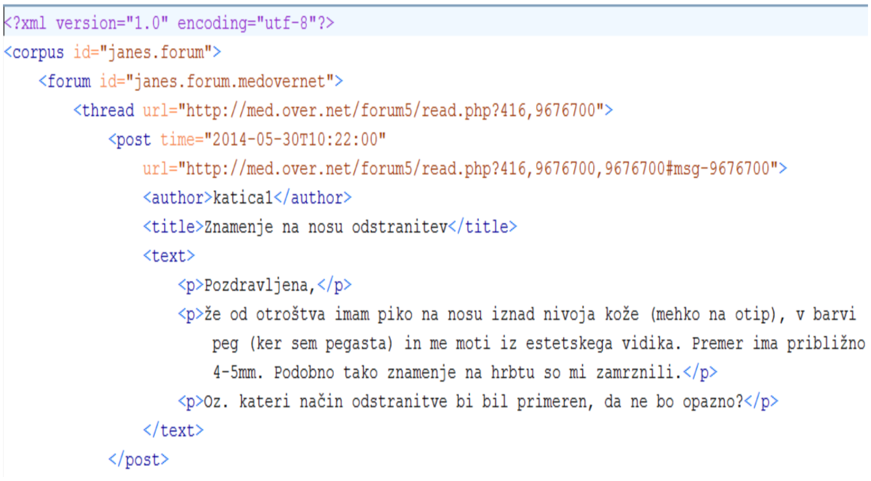
- Do not use Latin-1 (ISO/IEC 8859-1 / Windows-1252) use **Unicode**
- Most characters you will ever need
- Most software now supports it
- Is being extended
- Still, there are always problems:
  - U+0218 LATIN CAPITAL LETTER S WITH COMMA BELOW ?
  - U+015E LATIN CAPITAL LETTER S WITH CEDILLA ?
  - Also: lc(Ş) ≠ ſ (long s)



Encoding format: XML

- If data structure is simple, people still use tabular format
- Otherwise, XML is (almost) mandatory
- Simple syntax
- Formally checking of well-formedness and validity
- A host of associated standards:
  - DTD, XML Schema, RelaxNG
  - XPath, XSLT, XQuery
- Good tool support

Example: Post from UCG corpus in XML

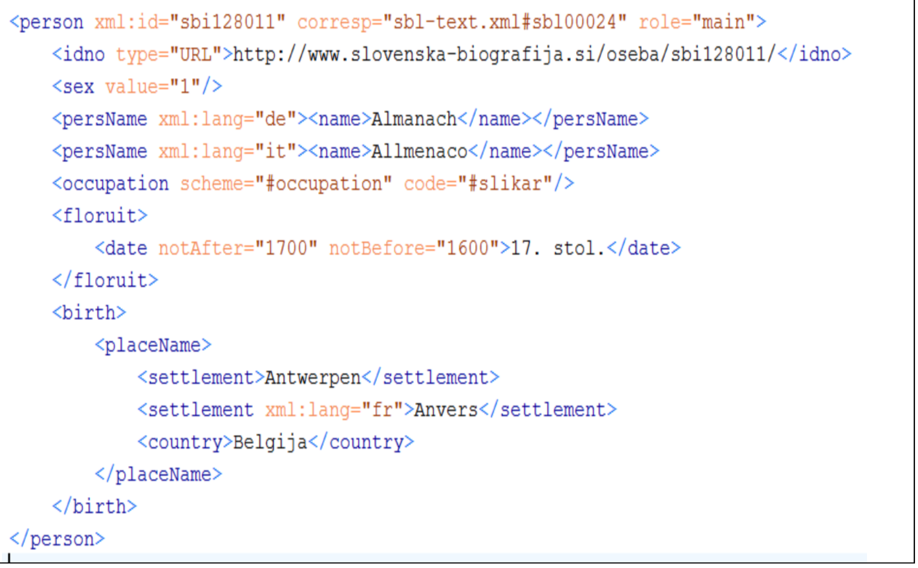


```xml
<?xml version="1.0" encoding="utf-8"?>
<corpus id="janes.forum">
    <forum id="janes.forum.medovernet">
        <thread url="http://med.over.net/forum5/read.php?416,9676700">
            <post time="2014-05-30T10:22:00"
                url="http://med.over.net/forum5/read.php?416,9676700,9676700#msg-9676700">
                <author>katica1</author>
                <title>Znamenje na nosu odstranitev</title>
                <text>
                    <p>Pozdravljena,</p>
                    <p>že od otroštva imam piko na nosu iznad nivoja kože (mehko na otip), v barvi
                        peg (ker sem pegasta) in me moti iz estetskega vidika. Premer ima približno
                        4-5mm. Podobno tako znamenje na hrbtu so mi zamrznili.</p>
                    <p>Oz. kateri način odstranitve bi bil primeren, da ne bo opazno?</p>
                </text>
            </post>
```

Schema: Text Encoding Initiative

- Guidelines and (XML) schemas for encoding scholarly texts: detailed and maintained
- Longest running standardisation effort
- Mostly used for digital humanities, less for HLT
- Good tool support:
  - conversion between formats
  - schema generator
- Active user community:
  - very friendly mailing list
  - annual TEI conferences
  - TEI journal

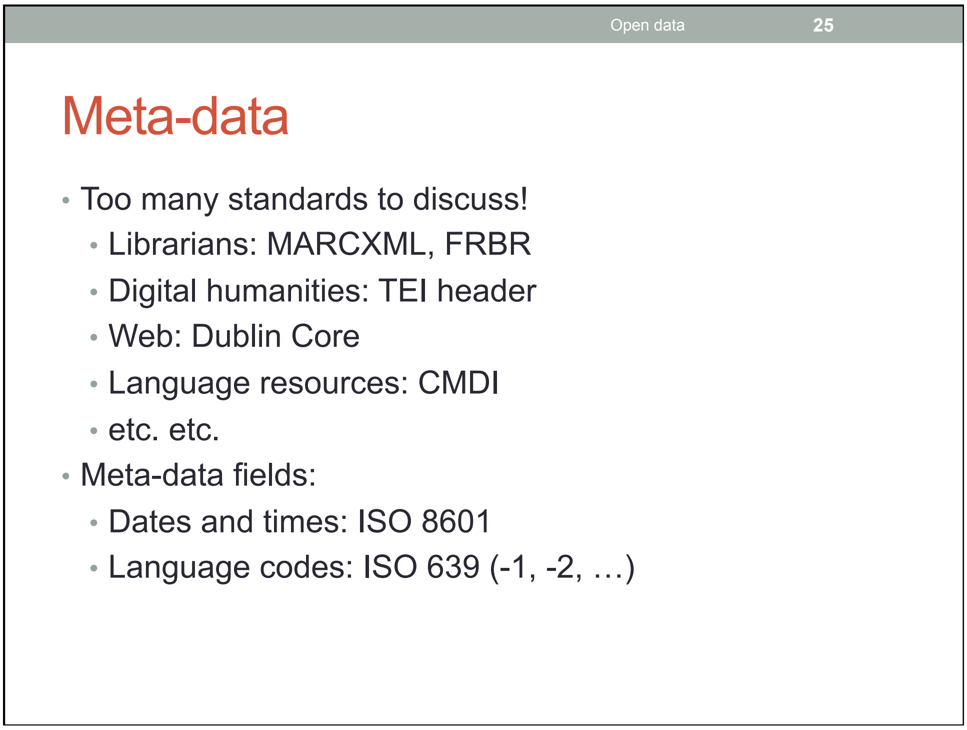
Slovene biographical lexicon in TEI



```
<person xml:id="sbi128011" corresp="sbl-text.xml#sbl00024" role="main">
    <idno type="URL">http://www.slovenska-biografija.si/oseba/sbi128011/</idno>
    <sex value="1"/>
    <persName xml:lang="de"><name>Almanach</name></persName>
    <persName xml:lang="it"><name>Allmenaco</name></persName>
    <occupation scheme="#occupation" code="#slikar"/>
    <floruit>
        <date notAfter="1700" notBefore="1600">17. stol.</date>
    </floruit>
    <birth>
        <placeName>
            <settlement>Antwerpen</settlement>
            <settlement xml:lang="fr">Anvers</settlement>
            <country>Belgija</country>
        </placeName>
    </birth>
</person>
```

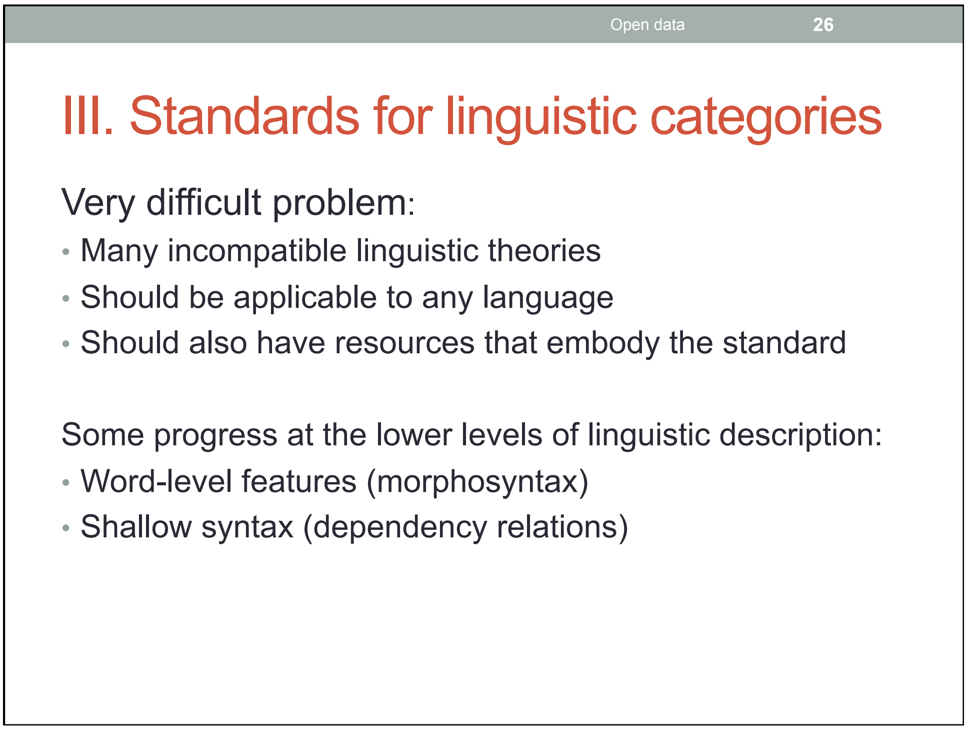
ISO encoding language resources

- ISO TC 37: Technical Committee for Terminology
- In 2004: … **and other language and content resources**
- ISO TC 37 SC4 **Language resource management**:
    - Feature structures: ISO 24610-1:2006
    - Lexical markup framework (LMF): ISO 24613:m2008
    - Morphosyntactic annotation framework: ISO 24611:2012
    - Syntactic annotation framework: ISO 24615-1:2014
    - Semantic annotation framework:
        - Part 1: Time and events: ISO 24617-1:2012
        - Part 2: Dialogue acts: ISO 24617-2:2012
        - Part 4: Semantic roles: ISO 24617-4:2014
        - Part 7: Spatial information: ISO 24617-7:2014
        - etc.
    - etc.



Meta-data

- Too many standards to discuss!
  - Librarians: MARCXML, FRBR
  - Digital humanities: TEI header
  - Web: Dublin Core
  - Language resources: CMDI
  - etc. etc.
- Meta-data fields:
  - Dates and times: ISO 8601
  - Language codes: ISO 639 (-1, -2, …)



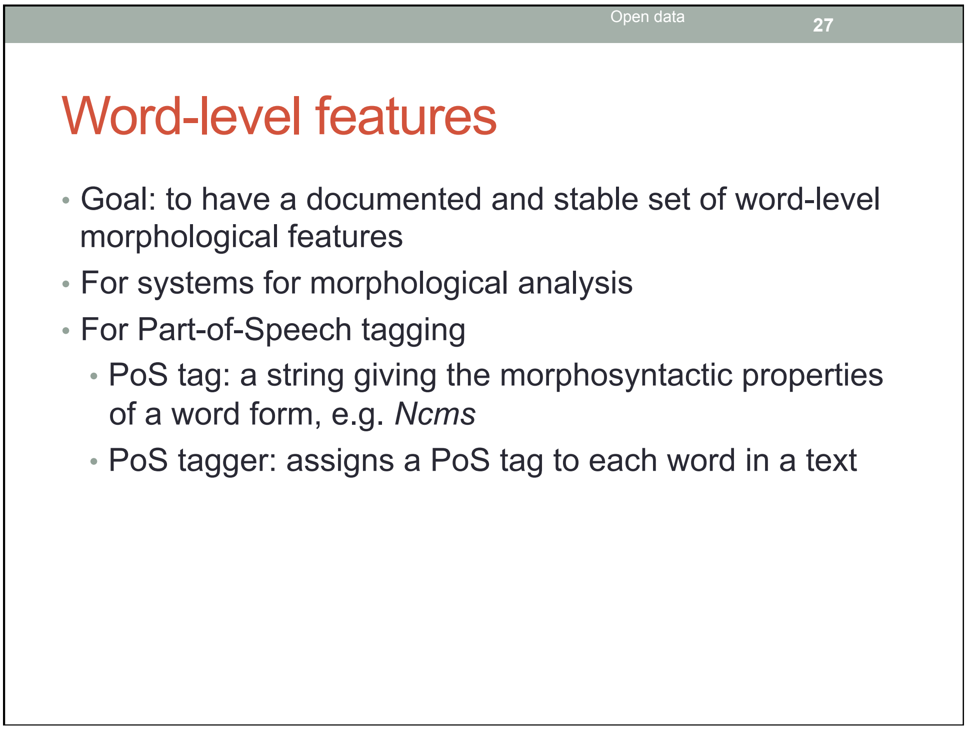
III. Standards for linguistic categories

Very difficult problem:

- Many incompatible linguistic theories
- Should be applicable to any language
- Should also have resources that embody the standard

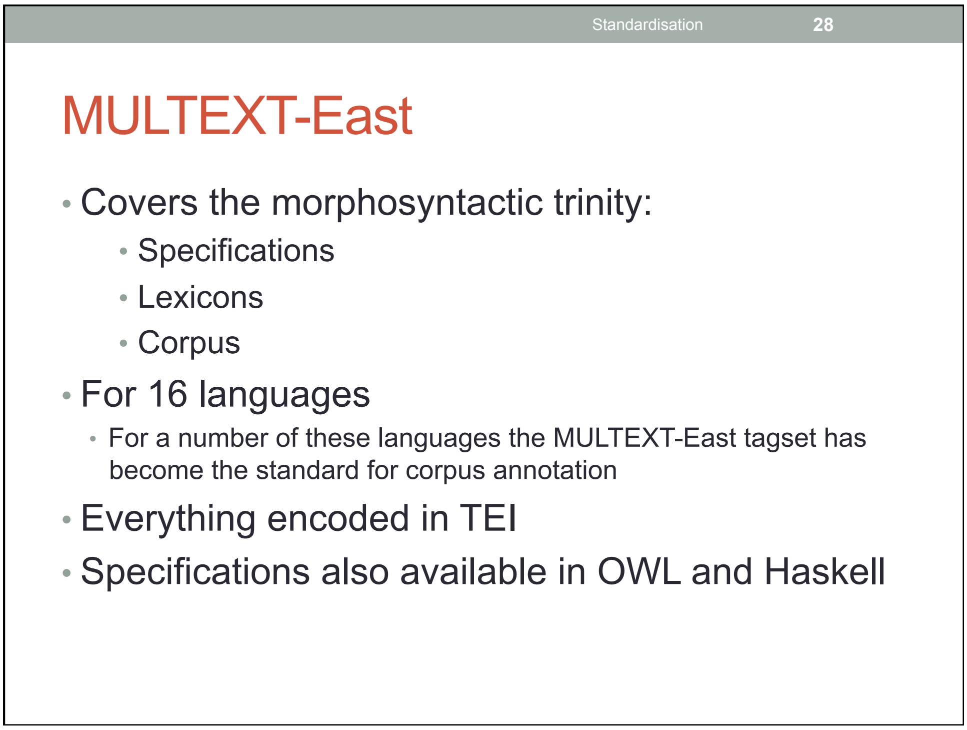
Some progress at the lower levels of linguistic description:

- Word-level features (morphosyntax)
- Shallow syntax (dependency relations)



Word-level features

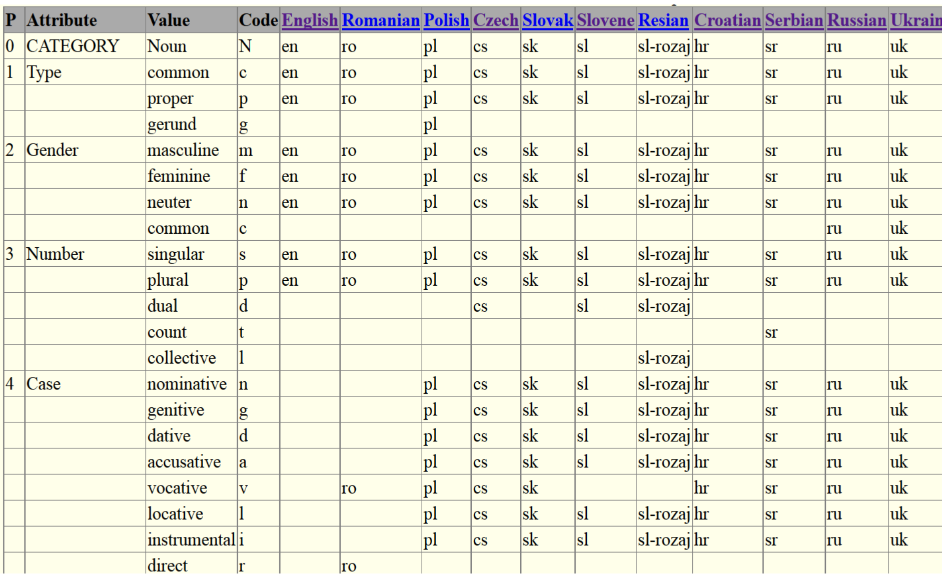
- Goal: to have a documented and stable set of word-level morphological features
- For systems for morphological analysis
- For Part-of-Speech tagging
  - PoS tag: a string giving the morphosyntactic properties of a word form, e.g. *Ncms*
  - PoS tagger: assigns a PoS tag to each word in a text



MULTEXT-East

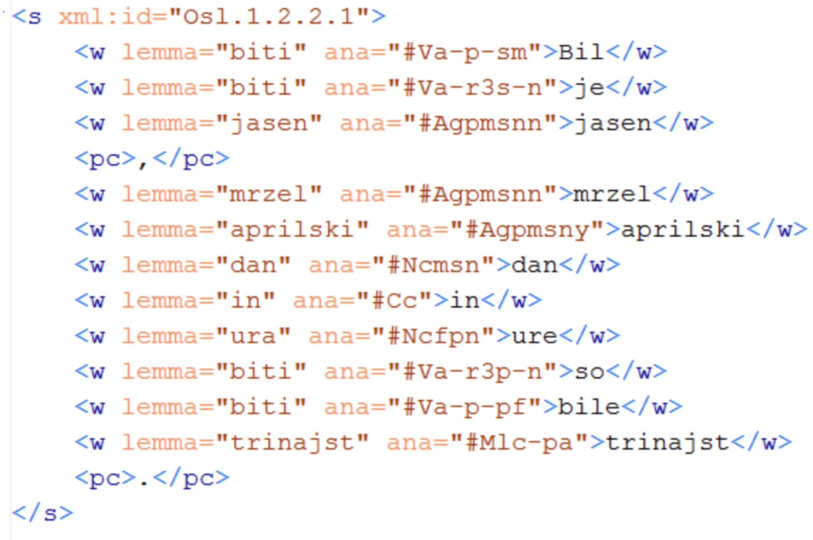
- Covers the morphosyntactic trinity:
  - Specifications
  - Lexicons
  - Corpus
- For 16 languages
  - For a number of these languages the MULTEXT-East tagset has become the standard for corpus annotation
- Everything encoded in TEI
- Specifications also available in OWL and Haskell

MULTEXT-East tables



P	Attribute	Value	Code	English	Romanian	Polish	Czech	Slovak	Slovene	Resian	Croatian	Serbian	Russian	Ukrain
0	CATEGORY	Noun	N	en	ro	pl	cs	sk	sl	sl-rozaj	hr	sr	ru	uk
1	Type	common	c	en	ro	pl	cs	sk	sl	sl-rozaj	hr	sr	ru	uk
		proper	p	en	ro	pl	cs	sk	sl	sl-rozaj	hr	sr	ru	uk
		gerund	g			pl								
2	Gender	masculine	m	en	ro	pl	cs	sk	sl	sl-rozaj	hr	sr	ru	uk
		feminine	f	en	ro	pl	cs	sk	sl	sl-rozaj	hr	sr	ru	uk
		neuter	n	en	ro	pl	cs	sk	sl	sl-rozaj	hr	sr	ru	uk
		common	c										ru	uk
3	Number	singular	s	en	ro	pl	cs	sk	sl	sl-rozaj	hr	sr	ru	uk
		plural	p	en	ro	pl	cs	sk	sl	sl-rozaj	hr	sr	ru	uk
		dual	d				cs		sl	sl-rozaj				
		count	t									sr		
		collective	l							sl-rozaj				
4	Case	nominative	n			pl	cs	sk	sl	sl-rozaj	hr	sr	ru	uk
		genitive	g			pl	cs	sk	sl	sl-rozaj	hr	sr	ru	uk
		dative	d			pl	cs	sk	sl	sl-rozaj	hr	sr	ru	uk
		accusative	a			pl	cs	sk	sl	sl-rozaj	hr	sr	ru	uk
		vocative	v		ro	pl	cs	sk			hr	sr	ru	uk
		locative	l			pl	cs	sk	sl	sl-rozaj	hr	sr	ru	uk
		instrumental	i			pl	cs	sk	sl	sl-rozaj	hr	sr	ru	uk
		direct	r		ro									

Use of MSDs in the corpus

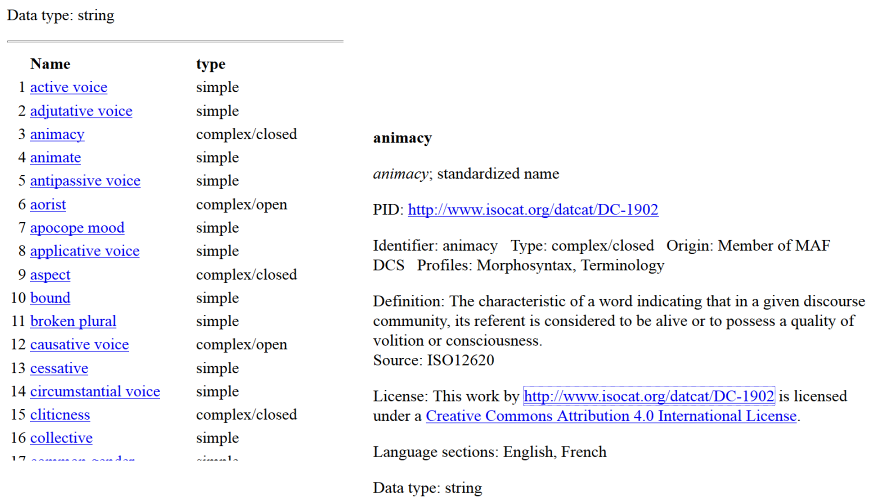


```
<s xml:id="Osl.1.2.2.1">
    <w lemma="biti" ana="#Va-p-sm">Bil</w>
    <w lemma="biti" ana="#Va-r3s-n">je</w>
    <w lemma="jasen" ana="#Agpmsnn">jasen</w>
    <pc>,</pc>
    <w lemma="mrzel" ana="#Agpmsnn">mrzel</w>
    <w lemma="aprilski" ana="#Agpmsny">aprilski</w>
    <w lemma="dan" ana="#Ncmsn">dan</w>
    <w lemma="in" ana="#Cc">in</w>
    <w lemma="ura" ana="#Ncfpn">ure</w>
    <w lemma="biti" ana="#Va-r3p-n">so</w>
    <w lemma="biti" ana="#Va-p-pf">bile</w>
    <w lemma="trinajst" ana="#Mlc-pa">trinajst</w>
    <pc>.</pc>
</s>
```

ISOcat - a Data Category Registry

- Specification of data categories and management of a Data Category Registry for language resources: ISO 12620:2009
- One of the first ISO standards delivered in the form of a database
- ISO DCR used also for defining linguistic terms: ISOcat
- First entries by developers, then a registered interface
- Interface was hosted by MPI but now in the process of migration..

Morphological features



Data type: string

	Name	type
1	active voice	simple
2	adjutative voice	simple
3	animacy	complex/closed
4	animate	simple
5	antipassive voice	simple
6	aorist	complex/open
7	apocope mood	simple
8	applicative voice	simple
9	aspect	complex/closed
10	bound	simple
11	broken plural	simple
12	causative voice	complex/open
13	cessative	simple
14	circumstantial voice	simple
15	cliticness	complex/closed
16	collective	simple
17	common gender	simple

animacy

animacy; standardized name

PID: http://www.isocat.org/datcat/DC-1902

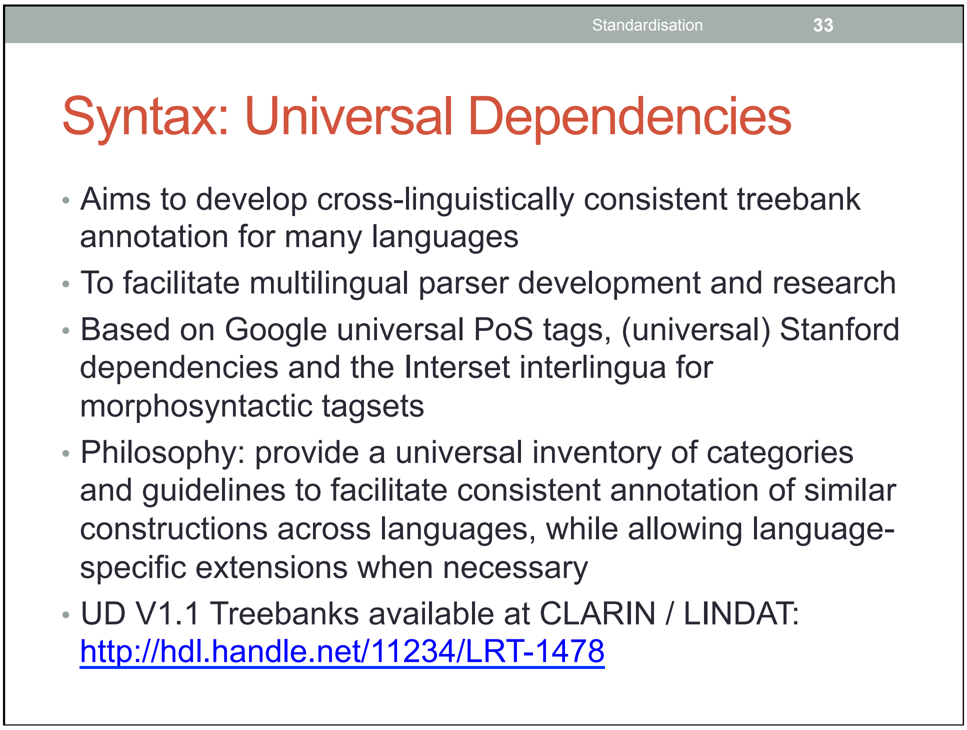
Identifier: animacy Type: complex/closed Origin: Member of MAF DCS Profiles: Morphosyntax, Terminology

Definition: The characteristic of a word indicating that in a given discourse community, its referent is considered to be alive or to possess a quality of volition or consciousness.
Source: ISO12620

License: This work by http://www.isocat.org/datcat/DC-1902 is licensed under a Creative Commons Attribution 4.0 International License.

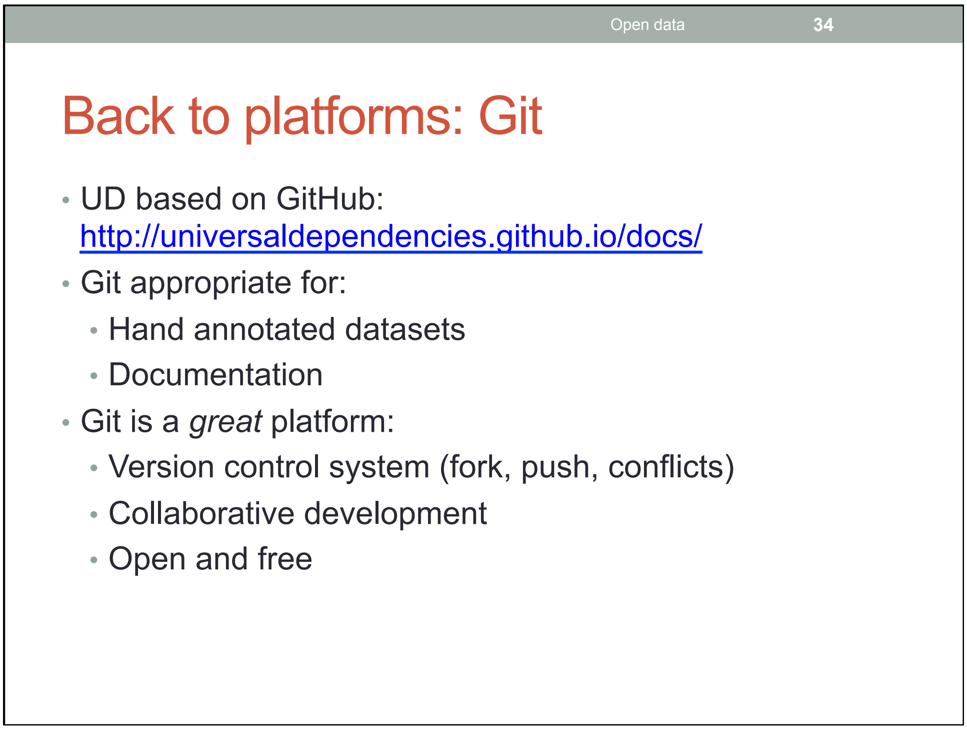
Language sections: English, French

Data type: string



Syntax: Universal Dependencies

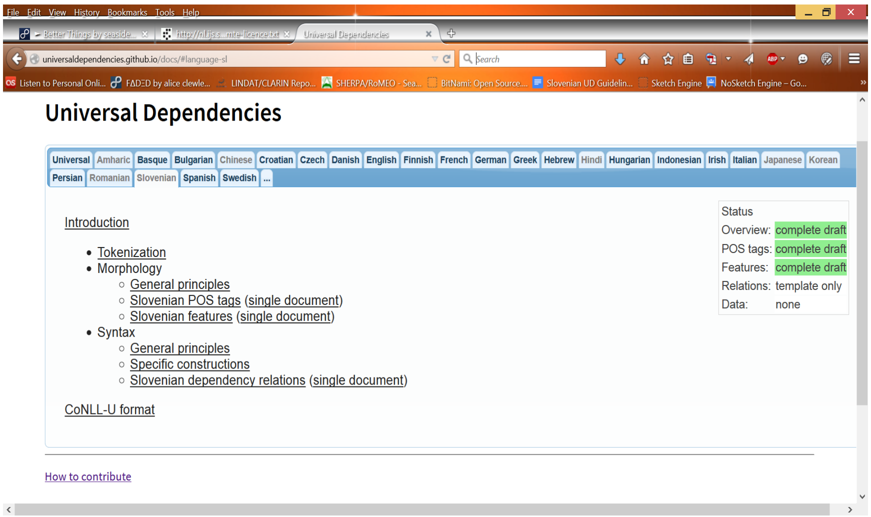
- Aims to develop cross-linguistically consistent treebank annotation for many languages
- To facilitate multilingual parser development and research
- Based on Google universal PoS tags, (universal) Stanford dependencies and the Interset interlingua for morphosyntactic tagsets
- Philosophy: provide a universal inventory of categories and guidelines to facilitate consistent annotation of similar constructions across languages, while allowing language-specific extensions when necessary
- UD V1.1 Treebanks available at CLARIN / LINDAT: http://hdl.handle.net/11234/LRT-1478



Back to platforms: Git

- UD based on GitHub: http://universaldependencies.github.io/docs/
- Git appropriate for:
  - Hand annotated datasets
  - Documentation
- Git is a *great* platform:
  - Version control system (fork, push, conflicts)
  - Collaborative development
  - Open and free

Top level page of Slovenian


File Edit View History Bookmarks Tools Help
► Better Things by seaside...
http://nl.ijs.s...mte-licence.txt
Universal Dependencies
universaldependencies.github.io/docs/#language-sl
Search
Listen to Personal Onli...
FADED by alice clewle...
LINDAT/CLARIN Repo...
SHERPA/RoMEO - Sea...
BitNami: Open Source....
Slovenian UD Guidelin...
Sketch Engine
NoSketch Engine – Go...

Universal Dependencies

Universal | Amharic | Basque | Bulgarian | Chinese | Croatian | Czech | Danish | English | Finnish | French | German | Greek | Hebrew | Hindi | Hungarian | Indonesian | Irish | Italian | Japanese | Korean | Persian | Romanian | Slovenian | Spanish | Swedish | ...

Introduction

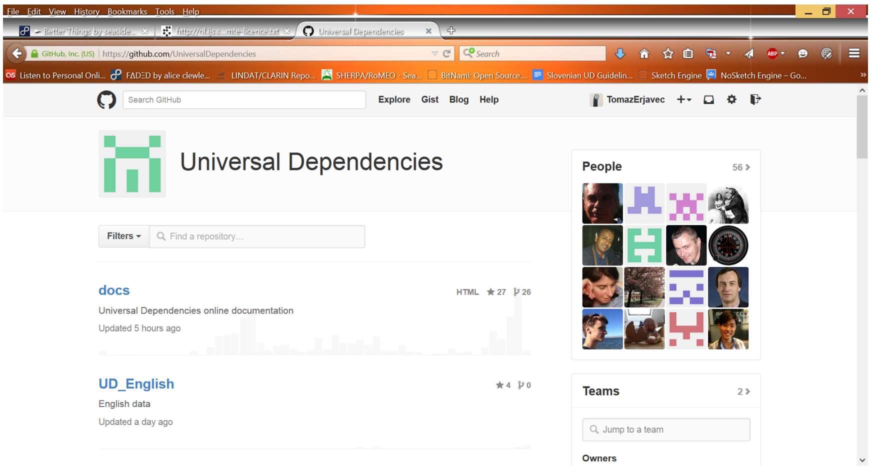
- Tokenization
- Morphology
  - General principles
  - Slovenian POS tags (single document)
  - Slovenian features (single document)
- Syntax
  - General principles
  - Specific constructions
  - Slovenian dependency relations (single document)

CoNLL-U format

Status	
Overview:	complete draft
POS tags:	complete draft
Features:	complete draft
Relations:	template only
Data:	none

How to contribute

Project home page on GitHub



Universal Dependencies

docs HTML ★ 27 26

Universal Dependencies online documentation

Updated 5 hours ago

UD_English ★ 4 0

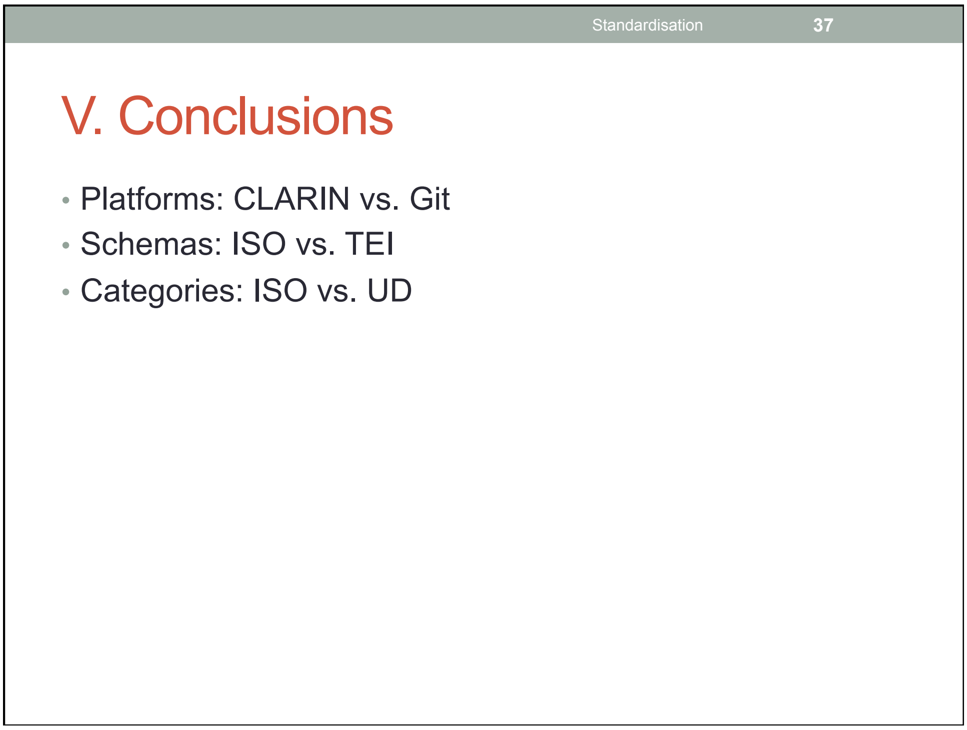
English data

Updated a day ago

People 56 >

Teams 2 >

Owners



V. Conclusions

- Platforms: CLARIN vs. Git
- Schemas: ISO vs. TEI
- Categories: ISO vs. UD