



[ve]dph

Latin digital archives and research infrastructures: just a trendy option or a substantive need?

F. Boschetti^{*}, R. Del Gratta[#], M. Monachini[#]

Institute for Computational Linguistics “A. Zampolli”, CNR

^{*}CNR-ILC-VeURT%UniVE-DSU-VeDPH, [#]CNR-ILC of Pisa

Introduction

- The **pillars** of the Common Language Resources and Technology Infrastructure (CLARIN)*
- Some **resources, tools** and **services**
- What are **FAIR** Data?
- Why do we **need** a Research Infrastructure?

*<http://bit.ly/2NYWkhV>

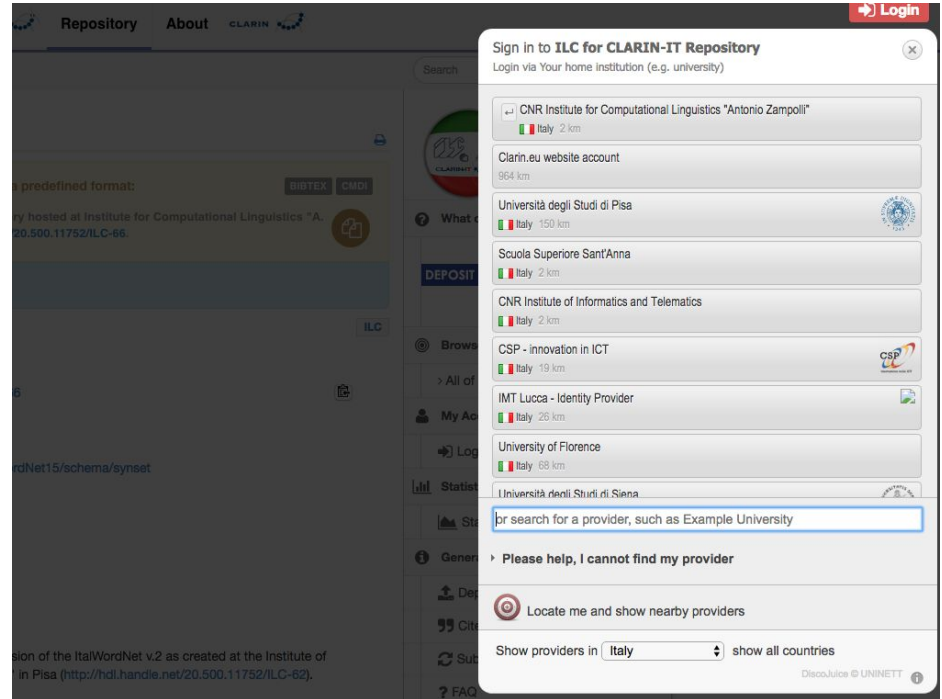


[ve]dph

Federated identity

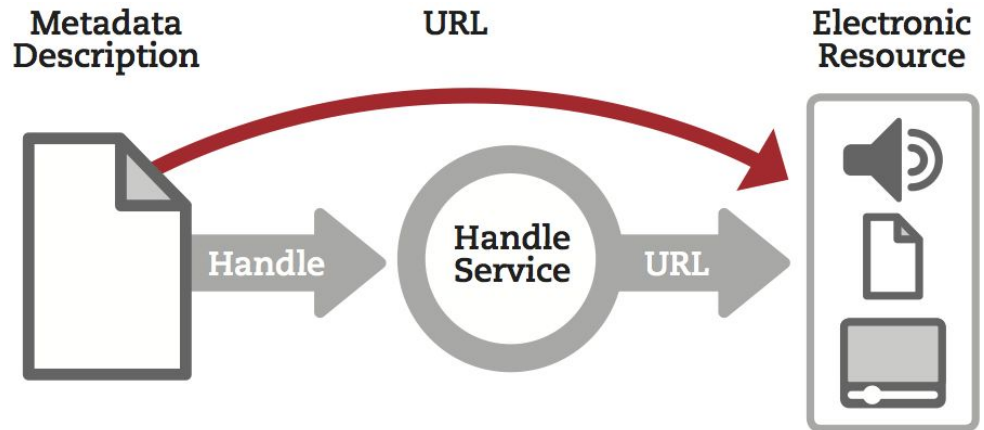
Using institutional credentials to access **all** the resources provided by the infrastructure

Both open and protected resources can be used in a similar way



Persistent identifiers

- CLARIN Persistent Identifiers (PIDs) are similar to DOI
- PIDs enable **sustainable citations** of digital resources
- They are disjoined from the actual location of a resource
- The **reference** to the actual location of the resource can be changed



Sustainable repositories

“Storing language resources and related datasets is something that requires a sound organization and attention for digital sustainability. After all, one of the important aims of CLARIN is to ensure that digital language resources are made available to a broad community on a **long-term basis**. This is achieved by establishing **data repositories** at the **centres**, which host **digital files** and the **associated metadata**. For reference purposes, these repositories also assign **persistent identifiers** to the resources, so that e.g. a specific dataset can be easily **cited** in a paper.” (<http://bit.ly/2U3wLjI>)



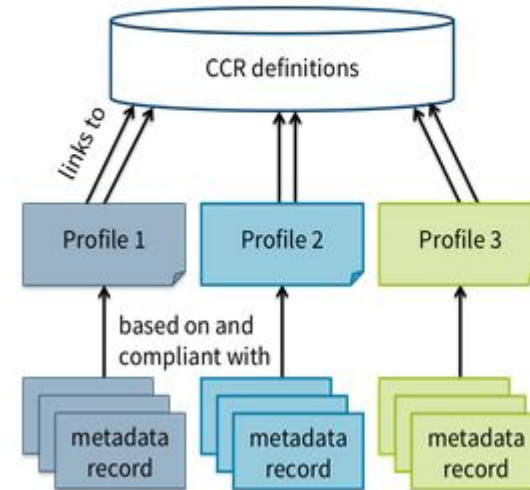
Flexible metadata and concept definitions

- Textual and linguistic resources are produced and consumed by multiple communities with different perspectives on data and metadata
- These communities along the years have elaborated different guidelines for metadata, more or less detailed, according to the typical use of the resources
- For instance, **Dublin Core** metadata are more suitable for librarians, **teiHeader** for philologists

Component MetaData Infrastructure (CMDI)

A component metadata infrastructure is a **flexible** and **community-oriented** solution, because it is **modular** and based on **multiple** schemes (one of them is the **teiHeader**)

Component Metadata



<https://www.clarin.eu/content/component-metadata>

Federated content search

Concordances
(keywords in
context)
through multiple
corpora

Federated Content Search Aggregator Help

Text layer CQL query

Search for in and show up to hits per endpoint

2 matching collections found Display as Key Word In Context Download

TextGrid Digital Library (Literature) – Leibniz-Institut für Deutsche Sprache View

Scaliger Copam Syriscam **voluit** esse κατελίδα , q̄s . caupam ut cludam olaudam , sive à copibus rebus eamque utriculariam , quæ inflans utrem retineat manu calamos per quorum foramina moduli distinguantur , pressum verò utrem examinet , quatiat cubito , ita ut calami vibrentur vexatione assidua .

Atque Deo **voluit** tantum illas esse dicata

Non nisi per magnos **voluit** DEUS esse labore

Dum **voluit** lignum vincere ; Victa fuit

ET PLENAS **VOLUIT** PRÆCIPITARE COLUS

Denn wo sonst , als in einem eigenen Werke über dieses Gedicht , können so leicht die einzeln Anmerkungen gestanden haben , die Servius aus ihm anführt ? Zugleich war Pollio ein Liebhaber und Kenner der Kunst . besaß eine reiche Sammluna der trefflichsten alten Kunstwerke . ließ von Künstlern seiner Zeit neue fertiaen . und dem

Web service chaining

In the CMDI file we can describe which **web services** (e.g. which linguistic analyzers) can be applied to the textual or linguistic resources. We can describe also how web services can be **piped**, in order to create **chains of analysis**: e.g. tokenization, morphological analysis, syntactic analysis (syntactic parsing).



Virtual Language Observatory & WebLicht

Virtual Language Observatory Search Contributors Help

Contributors

The following language resource repositories are powering the Virtual Language Observatory

CLARIN centres

[ASV Leipzig](#)

[Austrian Centre for Digital Humanities - A Resource Centre for the Humanities](#)

[Bayerisches Archiv für Sprachsignale](#)

[Berlin-Brandenburg Academy of Sciences and Humanities](#)

[CLARIN-IS](#)

[CLARIN-LT](#)

[CLARIN-PL Language Technology Centre](#)

[CLARIN.SI Language Technology Centre](#)

[CLARINO Bergen Center](#)

[CLARINO Text Laboratory Centre](#)

[CMU-TalkBank](#)

[Center of Estonian Language Resources](#)

[Centre for Language and Speech Technology](#)

weblicht.sfs.uni-tuebingen.de/weblicht/

View Tool List

Main Page Chain 3 x + New Chain

Show tools with status: development production superseded withdrawn

Input and Chain Selection

Run Tools

Clear Results

Down

Scindum est, q (Plain Tex

Scindum est, quia Dominicus
Silvius dux ducavit annos XII
et fecit
bellum cum Ruberto Viscardo,
unde fuit dispersus Petrus

Char: UDPipe tokenizer

Language: Latin

Document Type: CONLL-U

conllu.forms

conllu.misc

Done running tools.

tokens.csv



[ve]dph

Latin Resources

Language

Latin ✕

Collection

Resource type

Modality

Format

Keyword

▼

Availability

Search options

PROIEL collection

(Part of [Clarino Bergen Centre - INESS](#))

📖 A collection of dependency treebanks for early attestations of Indo-European, including a set of parallel treebanks of the New Testament.

🏠 [Landing page for this record](#)



Finnish Folk Poetry

📖 The corpus is available in Kielipankki - the Language Bank of Finland (korp.csc.fi), <http://urn.fi/urn:nbn:fi:lb-2014052711>. A 34-volume collection of Finnish oral poetry, lyric, short rhymes, incantations etc., collected and recorded from the 16th century to the 1930s and published mostly between 1908 and 1948, with a...



Index Thomisticus Treebank

(Part of [Tübingen Curated Resource](#))

📖 13th century Church Latin; Dependency-based annotation of excerpts from three works of Thomas Aquinas: (1) *Scriptum super Sententiis Magistri Petri Lombardi*, (2) *Summa contra Gentiles*, (3) *Summa Theologiae*; The Index Thomisticus Treebank is the syntactically annotated subset of the Index Thomisticus corpus. The Index T...

🏠 [Landing page for this record](#)



"PoDiLemma" Middle Polish Diachrone Lemmatised Corpus

(Part of [Universität des Saarlandes CLARIN-D-Zentrum, Saarbrücken](#))

📖 The PoDiLemma corpus is a diachronic corpus made of political, religious, scientific and historical texts from different authors of the Middle Polish period (16th-18th century). It contains in total ca. 7 million tokens. Characteristic for this period is the slow development of a supra-regional standard language, a pr...



Scientific method applied to historical data

The scientific method and the historical method are irreducible one to another; they are complementary: there is no science of individuals as such (Aristotle) and specularly there is no (hi)story of universals as such

BUT ...

the **scientific method** can be applied to **historical data** (if they are treated as **scientific data**) and symmetrically the **historical method** can be applied to **scientific data** (if they are treated as **historical data**)



Data must be FAIR

- **F**indable
- **A**ccessible
- **I**nteroperable
- **R**eusable

Wilkinson, M. D., et al. (2016).
The FAIR Guiding Principles
for scientific data management
and stewardship. *Scientific
data*, 3, 160018.
doi:10.1038/sdata.2016.18



Findable

- “ F1. (meta)data are assigned a **globally unique and persistent identifier**
- F2. data are described with **rich metadata** (defined by R1 below)
- F3. metadata clearly and explicitly include the identifier of the data it describes
- F4. (meta)data are registered or indexed in a searchable resource ”
- (Wilkinson et al., 2016)

The screenshot shows the ILC4CLARIN Repository Home page for the IWN-LOD dataset. The page features a navigation bar with 'Repository' and 'About' links, and a search bar. The main content area displays the dataset title 'IWN-LOD' and a citation instruction: 'Please use the following text to cite this item or export to a predefined format: Bartolini, Roberto, 2016, /IWN-LOD, ILC-CHR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa, http://hdl.handle.net/20.500.11752/ILC-66'. Below this, there are social media share buttons for Facebook, Twitter, and Google+, and a 'Share' button. The metadata section lists: Authors: Bartolini, Roberto; Item identifier: http://hdl.handle.net/20.500.11752/ILC-66; Project URL: https://datahub.io/dataset/wn; Demo URL: http://www.languagelibrary.eu/ow/ItaWordNet15/schema/synset; Date issued: 2016-10-18; Type: lexicalConceptualResource; Size: 49350 synsets; Language(s): Italian; Description: This is an RDF- Linguistic Open Data version of the ItaWordNet v.2 as created at the Institute of Computational Linguistics "A. Zampolli" in Pisa (http://hdl.handle.net/20.500.11752/ILC-62). The resource has been created according to the WN2.0 specification, http://www.w3.org/2006/03/wn/wn20/; Publisher: Datahub. The right sidebar contains navigation links for 'What can you do?' (DEPOSIT, CITE), 'Browse' (All of the Repository), 'My Account' (Login), 'Statistics' (BETA), and 'General Information' (Deposit, Cite, Submission Lifecycle, FAQ, About).



Accessible

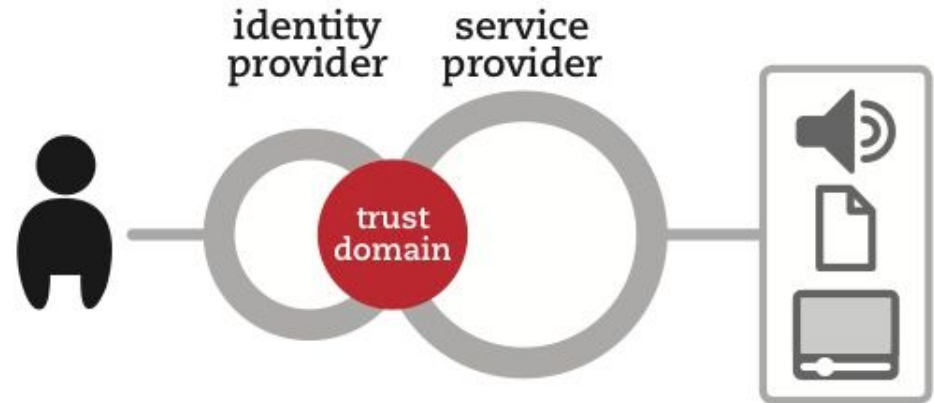
“ A1. (meta)data are retrievable by their identifier using a standardized communications protocol

A1.1 the protocol is open, free, and universally implementable

A1.2 the protocol allows for an authentication and authorization procedure, where necessary

A2. **metadata are accessible, even when the data are no longer available** ”

(Wilkinson et al., 2016)



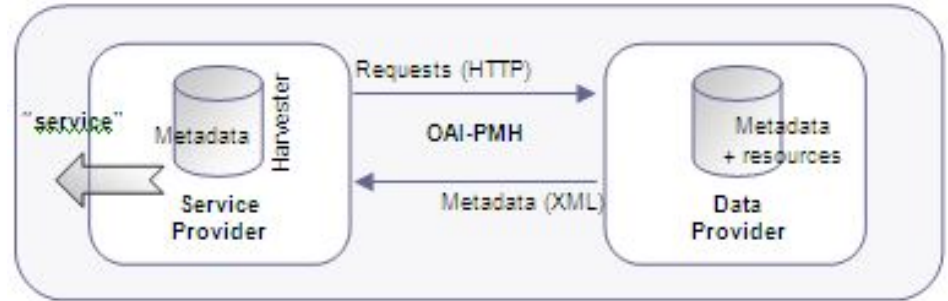
Interoperable

“ I1. (meta)data use a **formal**, accessible, shared, and broadly applicable language for knowledge representation

I2. (meta)data use **vocabularies that follow FAIR principles**

I3. (meta)data include **qualified references to other (meta)data**”

(Wilkinson et al., 2016)



Reusable

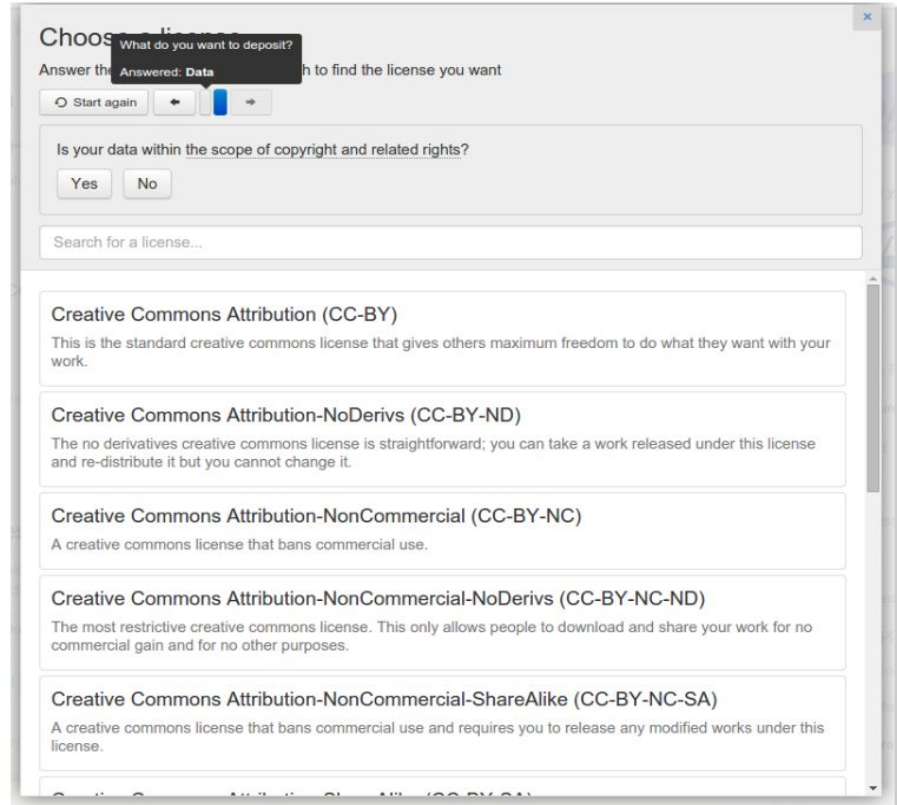
“ R1. meta(data) are **richly** described with a **plurality** of **accurate** and relevant **attributes**

R1.1. (meta)data are released with a clear and accessible **data usage license**

R1.2. (meta)data are associated with detailed **provenance**

R1.3. (meta)data meet **domain-relevant community standards** ”

(Wilkinson et al., 2016)



The screenshot shows a web interface for selecting a Creative Commons license. At the top, there is a question: "What do you want to deposit?" with a dropdown menu currently showing "Data". Below this, there is a question: "Is your data within the scope of copyright and related rights?" with "Yes" and "No" buttons. A search bar labeled "Search for a license..." is present. Below the search bar, several license options are listed with their descriptions:

- Creative Commons Attribution (CC-BY)**: This is the standard creative commons license that gives others maximum freedom to do what they want with your work.
- Creative Commons Attribution-NoDerivs (CC-BY-ND)**: The no derivatives creative commons license is straightforward; you can take a work released under this license and re-distribute it but you cannot change it.
- Creative Commons Attribution-NonCommercial (CC-BY-NC)**: A creative commons license that bans commercial use.
- Creative Commons Attribution-NonCommercial-NoDerivs (CC-BY-NC-ND)**: The most restrictive creative commons license. This only allows people to download and share your work for no commercial gain and for no other purposes.
- Creative Commons Attribution-NonCommercial-ShareAlike (CC-BY-NC-SA)**: A creative commons license that bans commercial use and requires you to release any modified works under this license.

Versioning

- How to ensure that we are talking about the **same thing** under the **same name**?

PROBLEM

- How to migrate (=reuse) annotations from an older to a newer version?

Obsolescence management

LONG-TERM PRESERVATION ISSUES

“one no longer preserves **tangible physical objects** per se, but views **abstract representations** of such objects that can be reconstructed in an **unpredictable** technological future.”

J.P. Chanod, *Will Your Data Still Be Around Tomorrow?*, 2013, <http://bit.ly/2tMRP3c>



Repeatability, Replicability and Reproducibility

ACM DEFINITIONS

- **Repeatability (Same team, same experimental setup)**
- **Replicability (Different team, same experimental setup)**
- **Reproducibility (Different team, different experimental setup)**

GOODMAN'S DEFINITIONS

- **Methods reproducibility (=ACM replicability):** provide **sufficient detail** about **procedures** and **data** so that the same procedures could be exactly repeated
- **Results reproducibility (=ACM reproducibility):** obtain the same **results** from an **independent** study with procedures as closely matched to the original study as possible
- **Inferential reproducibility:** draw the same **conclusions** from either an **independent replication** of a study or a **reanalysis** of the original study

Conclusion

If printed critical editions need authoritative publishers and the complex infrastructure constituted by national and academic libraries, digital scholarly editions **need** digital research infrastructures