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

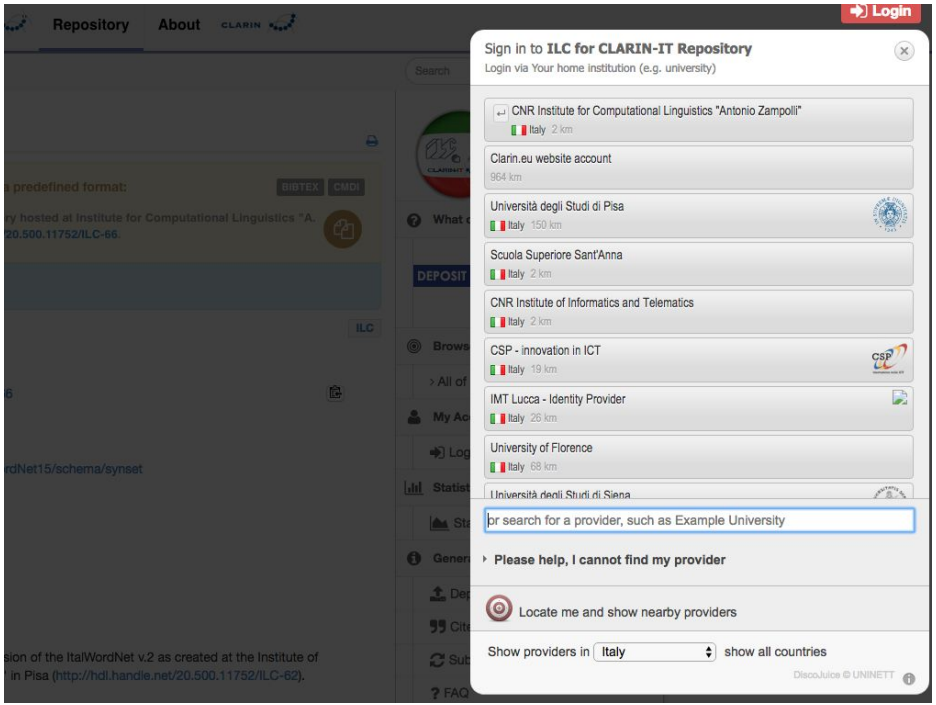
- The **pillars** of the Common Language Resources and Technology Infrastructure (CLARIN)*
- Some **resources**, **tools** and **services**
- What are **FAIR** Data?
- Why do we **need** a Research Infrastructure?

*http://bit.ly/2NYWkhV

CLARIN-IT [ve]dph

Federated identity

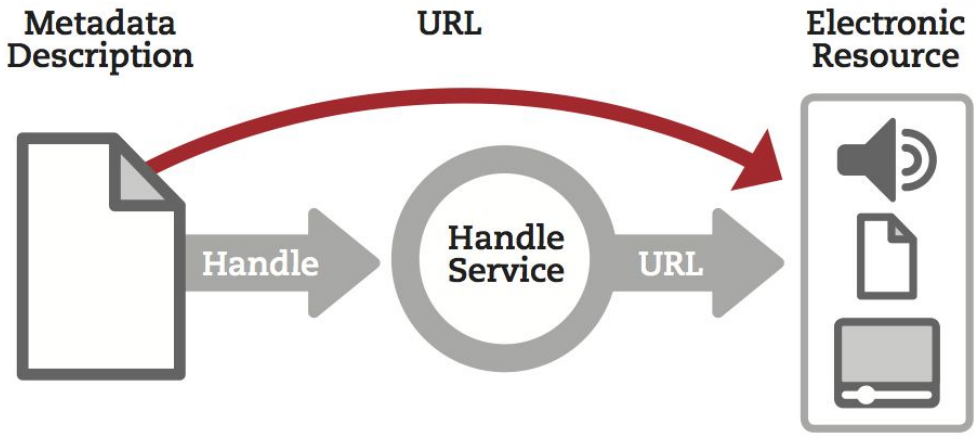
Using institutional credentials to access **all** the resources provided by the infrastructure



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

Persistent identifiers

- CLARIN Persistent Identifiers (PIDs) are similar to DOI
- PIDs enable s**ustainable 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

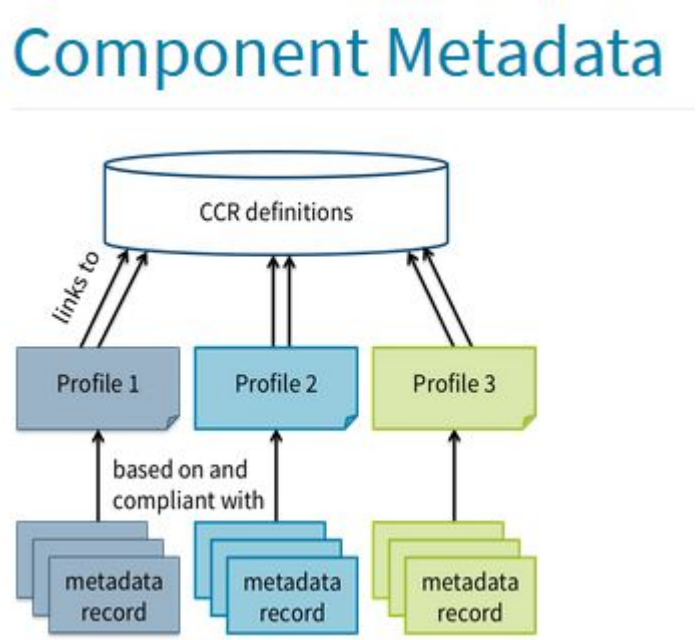
“Storing language resources and related datasets is something that requires a sound organization and attention for digital sustainability. After all, one of the important aims of CLARIN is to ensure that digital language resources are made available to a broad community on a **long-term basis**. This 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**)

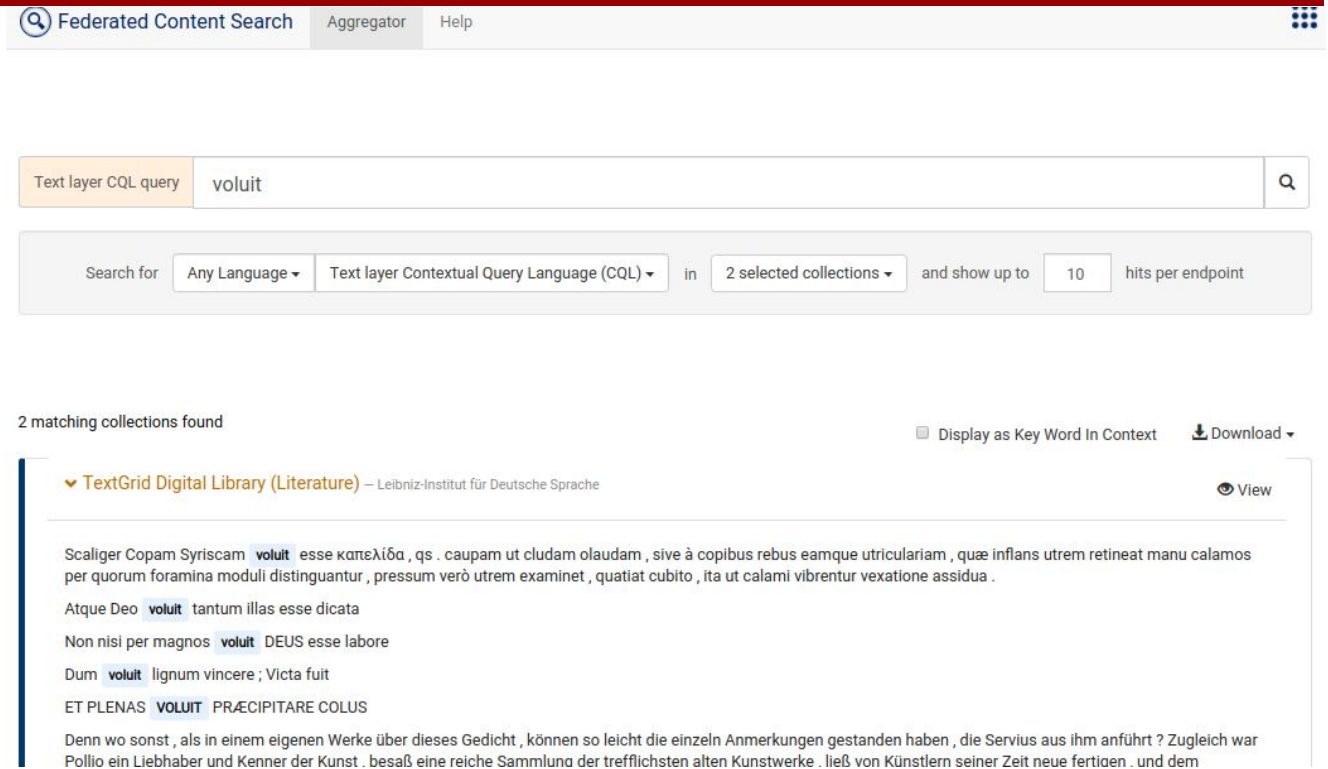


https://www.clarin.eu/content/component-metadata

CLARIN-IT [ve]dph

Federated content search

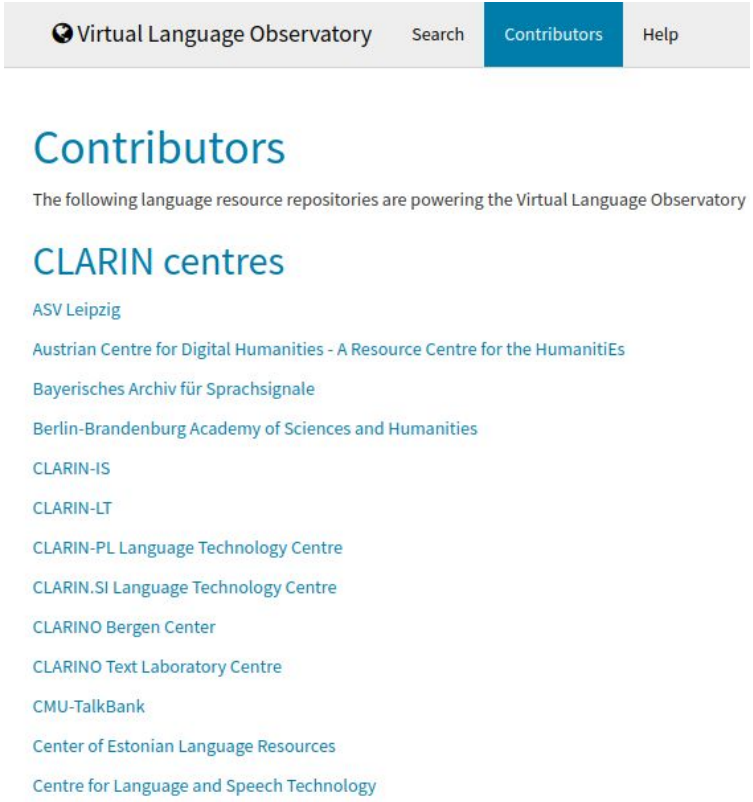
Concordances (**keywords in context**) through multiple corpora

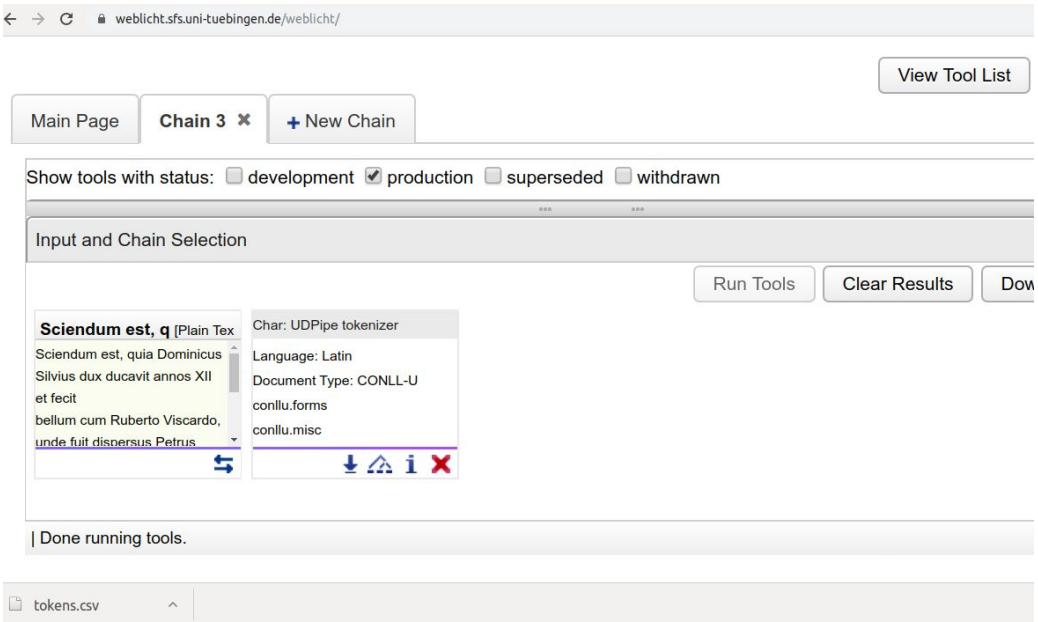


Web service chaining

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





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 Finnic 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

"PolDiLemma" Middle Polish Diachrone Lemmatised Corpus

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

The PolDiLemma 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…

CLARIN-IT [ve]dph

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

CLARIN-IT [ve]dph

Findable

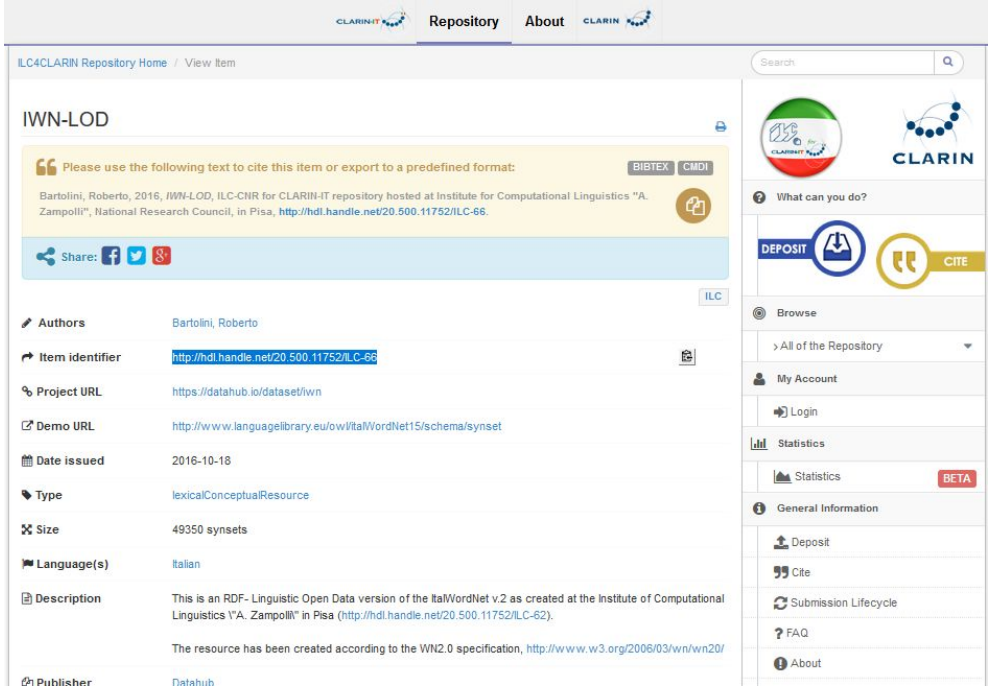
“F1. (meta)data are assigned a **globally unique and persistent identifier**

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)



Accessible

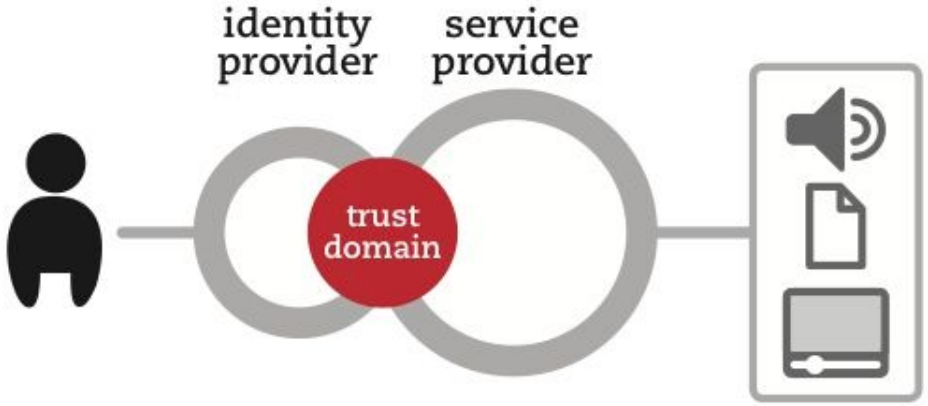
“A1. (meta)data are retrievable by their identifier using a standardized communications protocol

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)



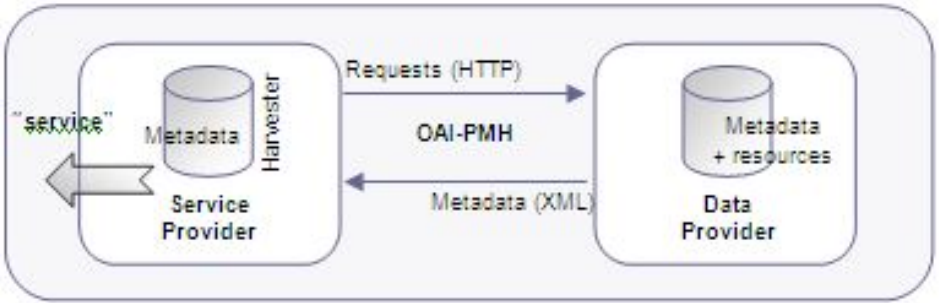
Interoperable

“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

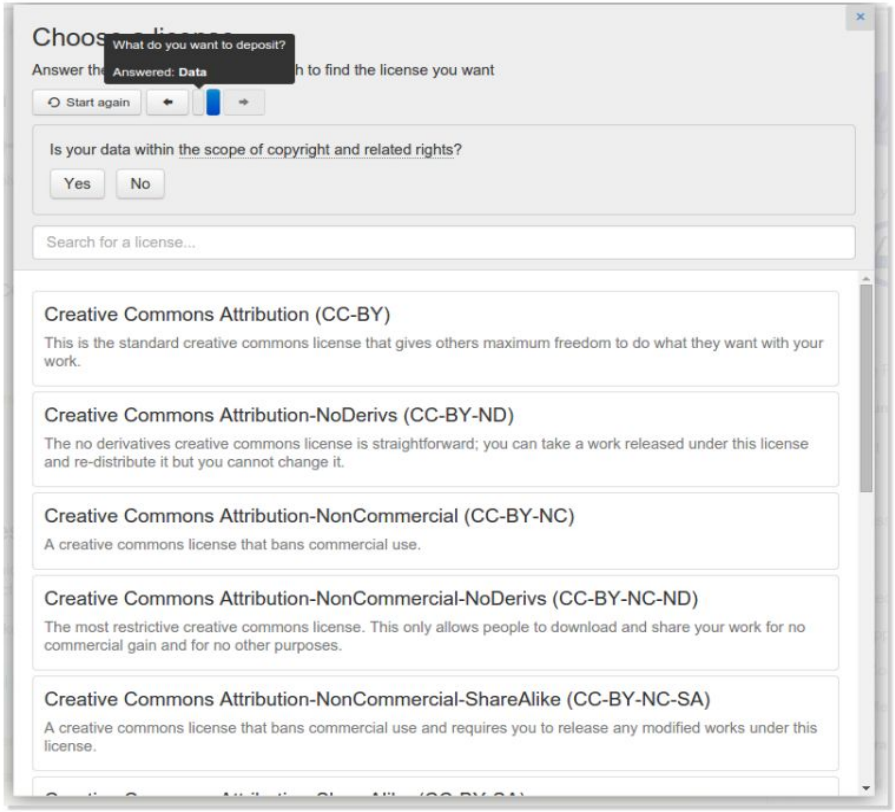
“R1. meta(data) are **richly** described with a **plurality** of **accurate** and relevant **attributes**

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)



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

H. Plesser, **Replicability vs. Reproducibility**, 2018, http://bit.ly/36w2dte

CLARIN-IT [ve]dph

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