CLARIN
Common Language Resources and Technology Infrastructure

The Role of Persistent Identifiers in CLARIN

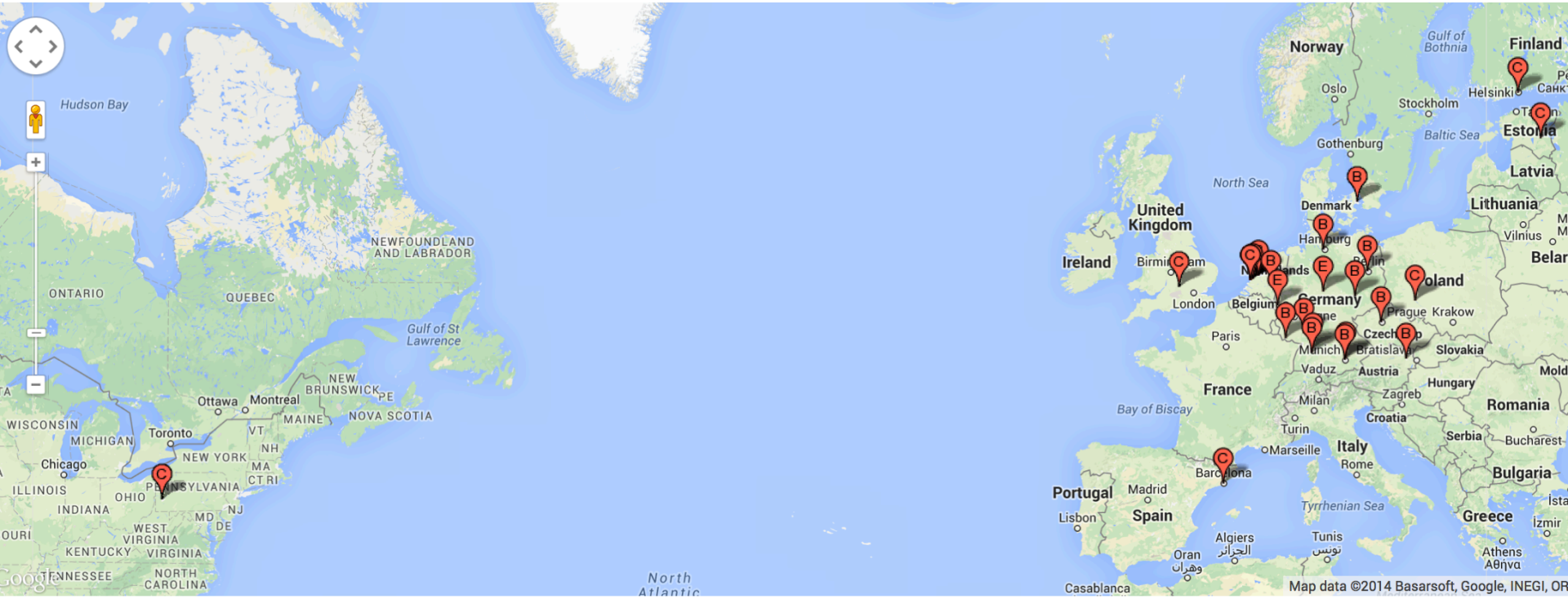
Dieter Van Uytvanck
CLARIN ERIC
dieter@clarin.eu

DASISH PID workshop
2014-12-08

CLARIN?

- **C**ommon **La**nguage **R**esources and Technology **In**frastructure
- European (ESFRI) Research Infrastructure – ERIC since February 2012
- aims at providing easy and sustainable access for scholars in the **humanities and social sciences**
  - to **digital language data** (in written, spoken, video or multimodal form)
  - to **advanced tools** to discover, explore, exploit, annotate, analyse or combine them

CLARIN centres



Persistent Identifiers: why?

Study	Resource type	Resource half-life
Koehler (1999 and 2002)	Random Web pages	about 2.0 years
Nelson and Allen (2002)	Digital Library Object	about 24.5 years
Harter and Kim (1996)	Scholarly Article Citations	about 1.5 years
Rumsey (2002)	Legal Citations	about 1.4 years
Markwell and Brooks (2002)	Biological Science Education Resources	about 4.6 years
Spinellis (2003)	Computer Science Citations	about 4.0 years (p. 74)

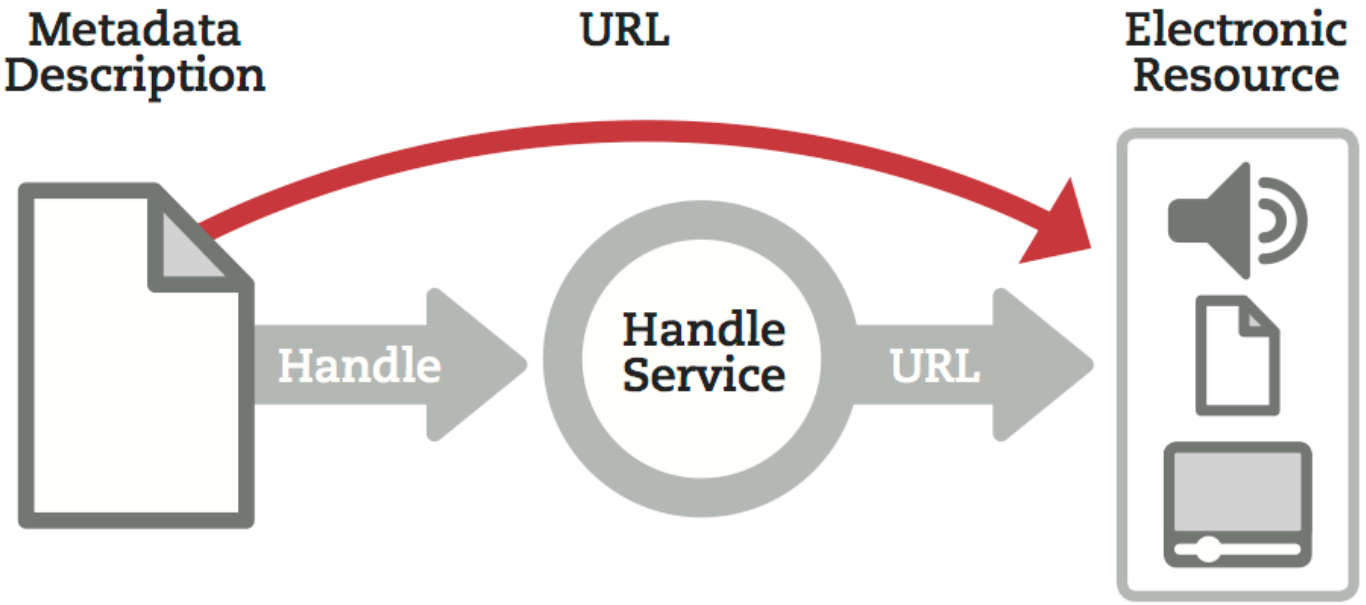
Source: Koehler, W. (2004) A longitudinal study of Web pages continued: a report after six years. *Information Research*, 9(2) paper 174 [Available at http://InformationR.net/ir/9-2/paper174.html]

How to prevent decaying links?

- Mentality: creating awareness about link rot

How to prevent decaying links?

- Technically: adding a level of indirection



PIDs in CLARIN

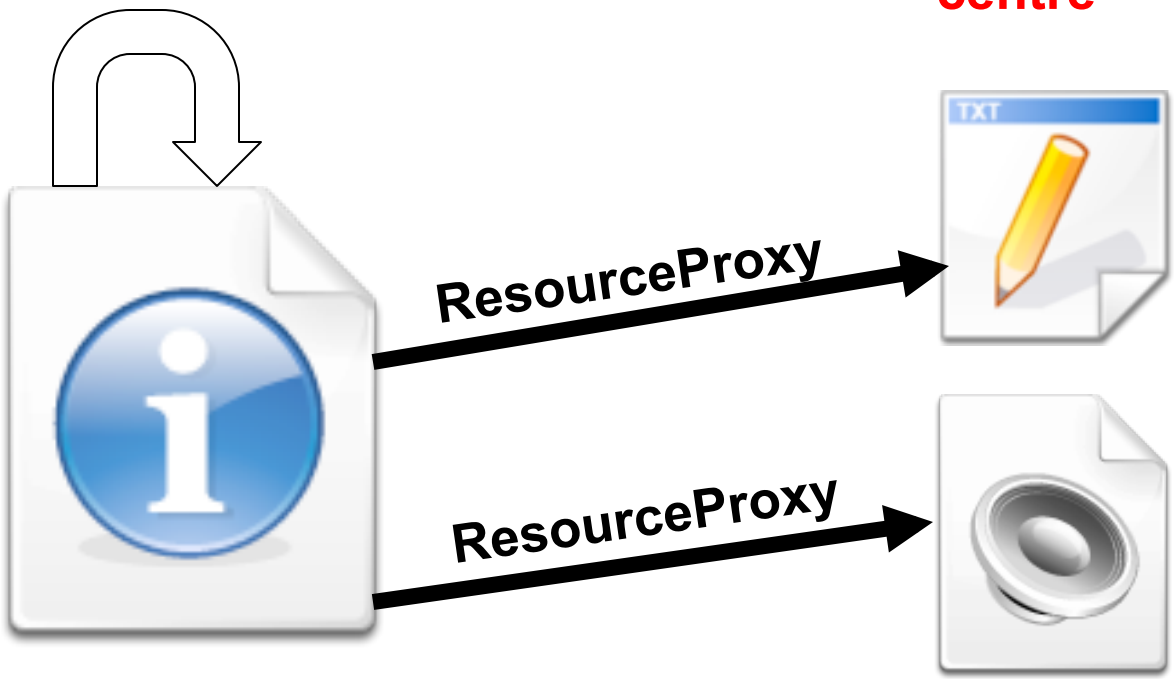
- B-centres need to associate **handles** with their **metadata records**. These PIDs should be suitable for both human and machine interpretation, taking into account the HTTP-accept header.

- **Non-metadata files** should receive a PID or a PID in combination with a part identifier, **if** these files:
  - are **accessible** via internet
  - are considered to be **stable** by the data provider
  - are considered to be **worth to be accessed directly** (not via metadata records) by the data provider

Object model

PID required

Handle + content negotiation



PID probably good idea, but depends on centre

Metadata (CMDI) XML file: PID in MdSelfLink

Language resources: PID or URL in metadata description

Why PIDs for metadata?

- Metadata is standardized:
  - After harvesting, clear point to start workflows
  - Self-reference available (MdSelfLink)
  - References to files and websites available with additional information:
    - Mime type
    - Service type (landing page, search service, search page)
- … so it is the ideal starting point for further processing:
  - Web service chains
  - Web applictions
  - “Add to virtual collection"

Why content negotiation?

- Requirement: a metadata PID should support content negotiation for:
  - CMDI (application/x-cmdi+xml) > **machine-processing**
  - HTML (text/html) > **human consumption**

- Ensures **standardized access to the digital objects**. After harvesting the metadata, one can always:
  - **Process** the described language resources **automatically**, based on the machine-readable XML description
  - Use a **browser** to access a **cited metadata record**

Why handles?

- **Scalable, proven technology with a universal resolution protocol**
- Decision taken during CLARIN's **preparatory phase**, supported by experiences from earlier projects (DAM-LR, starting in 2005)
- Service offer to CLARIN centres via agreement with **EPIC** consortium

Requirements

- Already in the preparatory phase (**2009**), CLARIN put forward some clear recommendations:
  - http://hdl.handle.net/1839/00-DOCS.CLARIN.EU-30
- In **2012**, the centre committee adopted the official centre criteria document:
  - http://hdl.handle.net/1839/00-DOCS.CLARIN.EU-77
- In **2013**, after several meetings of the PID taskforce, the criteria for the usage of PIDs were made more detailed and explicit (in relation to metadata records):
  - https://www.clarin.eu/node/3757

Requirements **2009**

- centres should get acquainted with PIDs
- and with repositories that support PIDs
- keep PIDs in mind during software development
- establish CLARIN-wide PID service
- one system which is performant, scalable and robust enough and that offers enough flexibility: **handle**
- talk to CNRI about requirements that are not met yet:
  - Global Handle Registry mirror
  - part identifiers
- make versioning compulsory for digital objects (with PIDs – comply or explain)
- establishing PID service independent of any commercial business model (DOI ok as individual choice but should not be default option)
- investigate various options of sharing a registration and resolution service with other disciplines

Requirements **2012**

- ***Centres need to associate PIDs records according to the CLARIN agreements with their objects and add them to the metadata record***.
- This should be indicated by:
  - An indication of the **Handle** assignment policies and procedures and ways to check that they are systematically applied.
  - An indication where to find **Handle** information in the metadata records and whether clicking on them will bring you to the data.

Requirements **2013**

- Centres need to associate (handle) PIDs with their metadata records. These PIDs should be suitable for both human and machine interpretation, taking into account the HTTP- accept header.

- Non-metadata files should receive a PID or a PID in combination with a part identifier, if these files:
  - are accessible via internet
  - are considered to be stable by the data provider
  - are considered to be worth to be accessed directly (not via metadata records) by the data provider

Recommendations

- PID taskforce, november 2013:
  - **get your own prefix** (costs: 50$ registration + 50$/year, see http://www.handle.net/service_agreement.html)
  - it is **not obligatory to use EPIC**, you can also run your own handle server
  - if you use **EPIC**, make sure to use **API version 2**

CLARIN services with PIDs: Virtual Collection Registry

Absolute spatial deixis and proto-toponyms in Kata Kolok

General

Name: Absolute spatial deixis and proto-toponyms in Kata Kolok

Type: extensional

Creation Date: 2014-09-26

Description: Digital references for De Vos, C. (2014). Absolute spatial deixis and proto-toponyms in Kata Kolok. NUSA: Linguistic studies of languages in and around Indonesia, 56, 3-26.

Purpose: research

Reproducibility: intended

Persistent identifier: hdl:11372/VC-1001

Keywords:
- sign language
- Kata Kolok

Creators

Resources

Reference	Type
Journal Article (fulltext) This paper presents an overview of spatial deictic structures in Kata Kolok, a sign language which is indigenous to a Balinese village community.	Resource
Footnote 3 - video Absolute versus absolute transpositional pointing signs	Resource
Footnote 4 - video COME-HERE-FROM-A and GO-FROM-HERE-TO-B	Resource

CLARIN services with PIDs: Language Resource Inventory

LINDAT/CLARIN Repository Home / Search

Search

Selected filters

Community: LRT + Open Submissions | Clear All

Advanced Search

Limit your search

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

Showing 1 through 10 out of 987 results

1 2 3 › 99

Corpus | LRT + Open Submissions

jos100k 2.0

(Jožef Stefan Institute, Dept. of Knowledge Technologies / 2010-03-18)

Author(s):

Erjavec, Tomaž ; Fišer, Darja ; Krek, Simon ; Ledinek, Nina

This item contains 1 file (4.39 MB).

Publicly Available

LINDAT CLARIN

CLARIN

What can you do?

DEPOSIT

CITE

Browse

› All of the Repository

My Account

Login

General Information

Deposit

Cite

Submission Lifecycle

FAQ

About

What about DOIs?

- After all, it is based on the handle protocol as well
- At the time of the choice for handles, DOIs were still limited to the commercial publishing world: issues with costs and business model (especially costs for high amounts of PIDs)
- New kid on the block: DataCite – more directed to research data repositories

Are DataCite DOIs CLARIN-compliant?

- They are handles
- Technically, some first experiments seem to show that the content negotiation for CMDI files works
  - wget --header "Accept: application/x-cmdi+xml" http://test.datacite.org/handle/10.5072/11148/0000-0003-203F-3 ➔ CMDI XML
  - wget --header "Accept: text/html" http://test.datacite.org/handle/10.5072/11148/0000-0003-203F-3 ➔ HTML
- Business and cost models should be evaluated case-by-case

Handles vs DataCite DOIs

	Handles	DataCite DOIs
Prefix/PID ownership (transferability)	**yes**	**Only at level of registration agency**
Digital Object referencing (e.g. single data file)	**yes**	**no**
Integrated metadata catalogue	no	yes
Resolution statistics	no	yes
Impact statistics (e.g. Thompson Reuters)	not automatic but possible	automatic
Resolution to multiple URLs	yes	no
Part Identifiers	yes	no
Content Negotiation	yes	yes

Broader context

- Joint Declaration of Data Citation Principles - https://www.force11.org/datacitation
- RDA dynamic data citation working group: https://rd-alliance.org/groups/data-citation-wg.html
- RDA PID information types working group: https://rd-alliance.org/groups/pid-information-types-wg.html

Conclusion

- CLARIN has made the choice to use **handles**
- **Clear requirements**:
  - Minimally a PID for each metadata record
  - Support for content negotiation
- Strong preference to acquire **an own prefix** (= no lock-in)
- Within this context, centres make a well-informed **choice between providers**:
  - Host-it-yourself
  - EPIC
  - DataCite DOIs (can fulfill minimal requirements)

Thank you for your attention!

http://clarin.eu/node/4005