CLARIN
Common Language Resources and Technology Infrastructure

DOI & DataCite workshop
Welcome and introduction

Dieter Van Uytvanck
CLARIN ERIC
dieter@clarin.eu
2014-10-23

Agenda

- 10:00 - 10:15: Welcome and introduction (Dieter Van Uytvanck, CLARIN ERIC)
- 10:15 - 11:15: Introduction to DataCite (Madeleine de Smaele, TU Delft Library)
- 11:15 - 11:30: coffee break
- 11:30 - 12:30: Some CLARIN centres present their plans/ questions/issues with DOIs and DataCite:
  - CELR (EE, Neeme Kahusk)
  - DANS (NL, Marnix van Berchum)
  - Oxford Text Archive (UK, Martin Wynne)
  - LINDAT (CZ, Pavel Straňák)
- 12:30 - 13:00: Question and discussion slot
- 13:00: wrap-up and lunch

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)

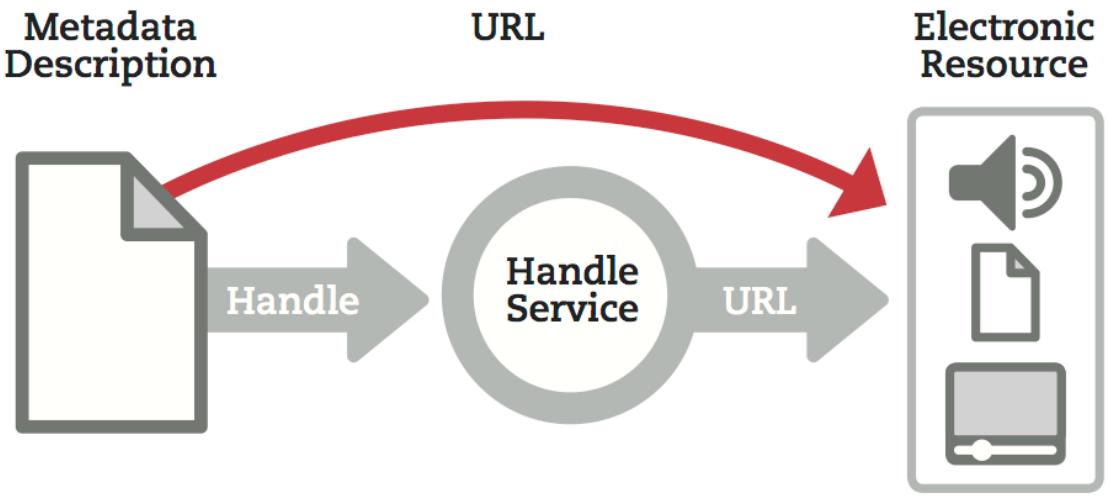
Source: Koehler, W. (2004) A longitudinal study of Web pages continued: a report after six years. *Information Research*, 9(2) paper 174 [Available at http://InformationR.net/ir/9-2/paper174.html]

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

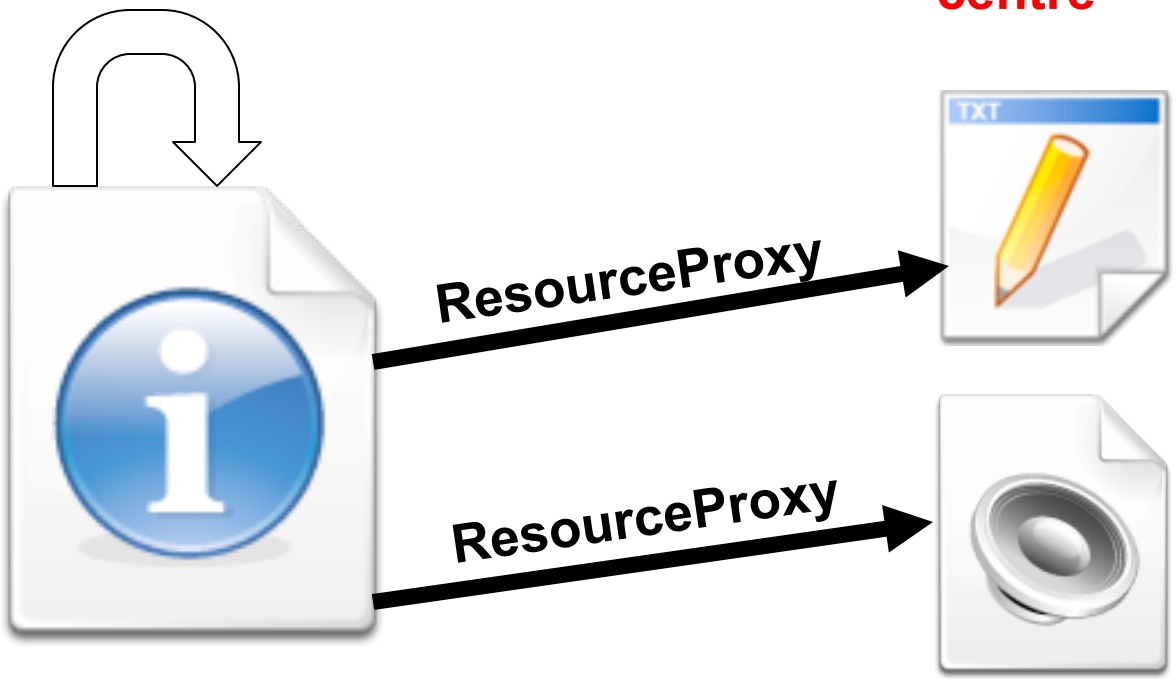
- B-centres need to associate (handle) PIDs with their **metadata records**. These PIDs should be suitable for both human and machine interpretation, taking into account the HTTP- accept header.

- **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

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
- 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
- EPIC is also a DataCite registration authority
- So at first sight it looks positive, but let's listen to some experts…

Thank you for your attention!