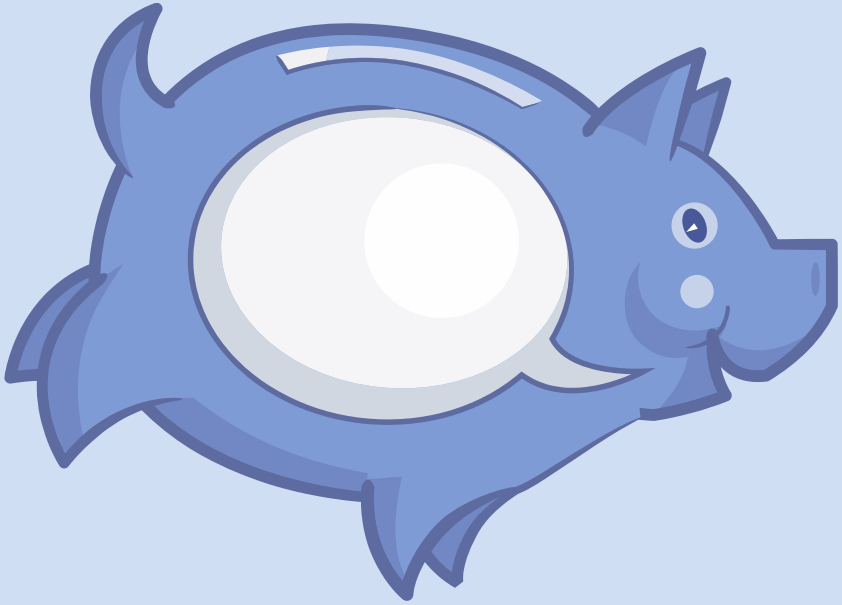
CLARIN Annual Conference 2018

Pisa, 8-10 October 2018



KIELIPANKKI
The Language Bank of Finland

Versioning with PIDs

Martin Matthiesen, CSC – IT Center for Science
Ute Dieckmann, University of Helsinki
CLARIN annual conference, Pisa, Italy 10.10.2018

FIN-CLARIN

Contents

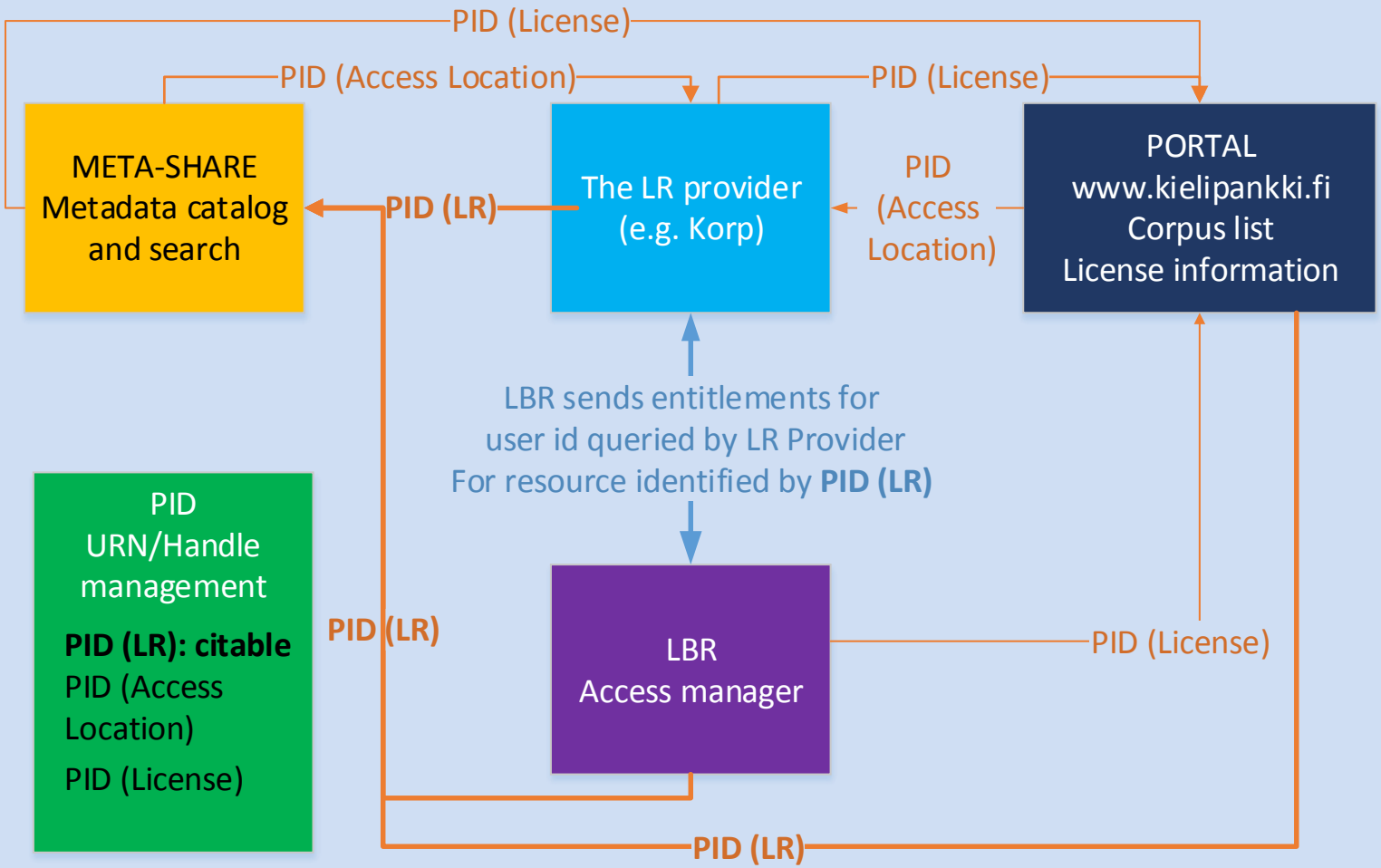
- What we do: The repository in a nutshell
  - The Language Bank
  - Publication process
  - Using PIDs in the components
- The dataset update
  - The original plan
  - Problems
  - Versioning with PIDs
- Reflections on PIDs
  - A PID is a promise


KIELIPANKKI
The Language Bank of Finland
LANGUAGE BANK ACCESS CORPORA TOOLS ORGANIZATION SUPPORT SUOMEKSI PÅ SVENSKA
Access
Apply for rights to use our language resources.
Corpora
Browse our corpora.
Tools
Try our tools.
Search the Language Bank Portal:
Researcher of the Month: Annika Lilli
Organization
Who are the Language Bank?
Support
Help and instructions.
News
What is the Language Bank of Finland?
CLARIN CENTRE B
CORE TRUST SEAL

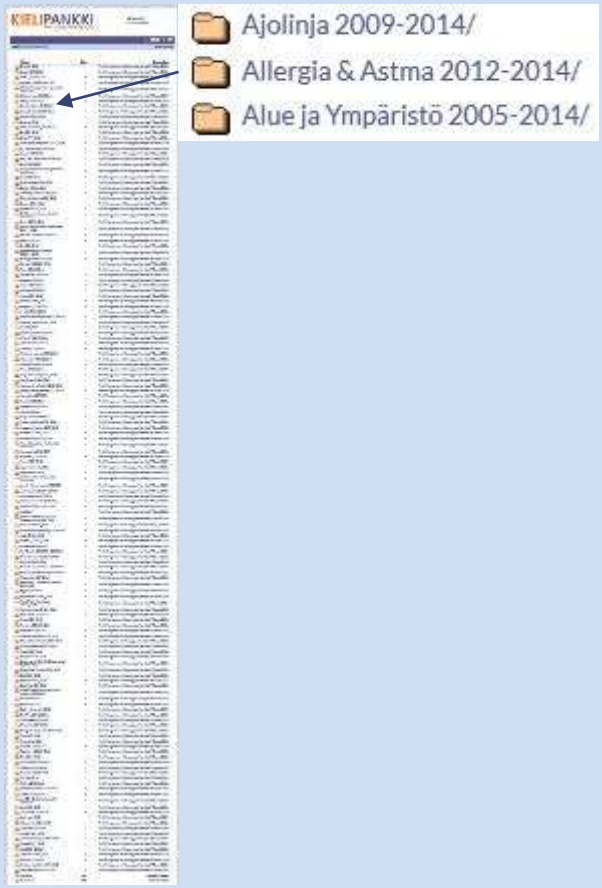
Publication of a language resource at the Language Bank of Finland

- Descriptive metadata in META-SHARE
- Packaging (e.g. Download, Korp)
- PID assignment
- Access control
  - ACA: Login (eduPersonAffiliation=member)
  - RES: Personal application needed via Language Bank Rights

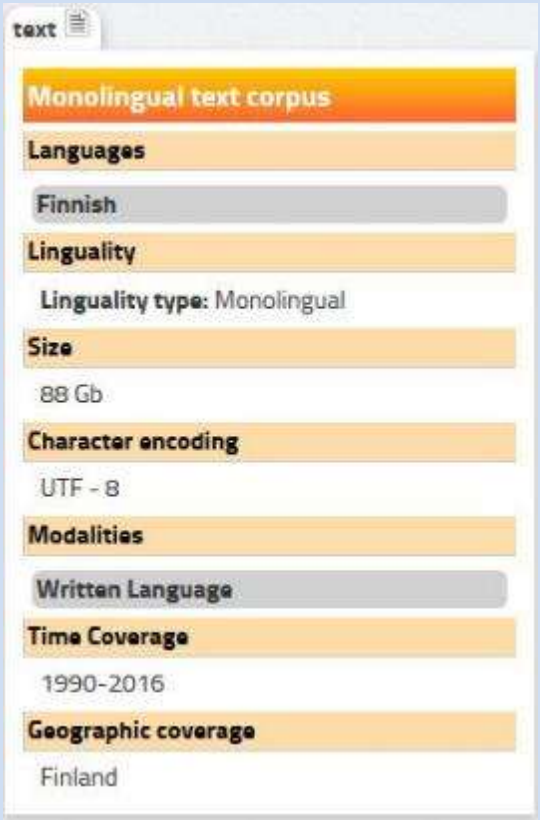
The components



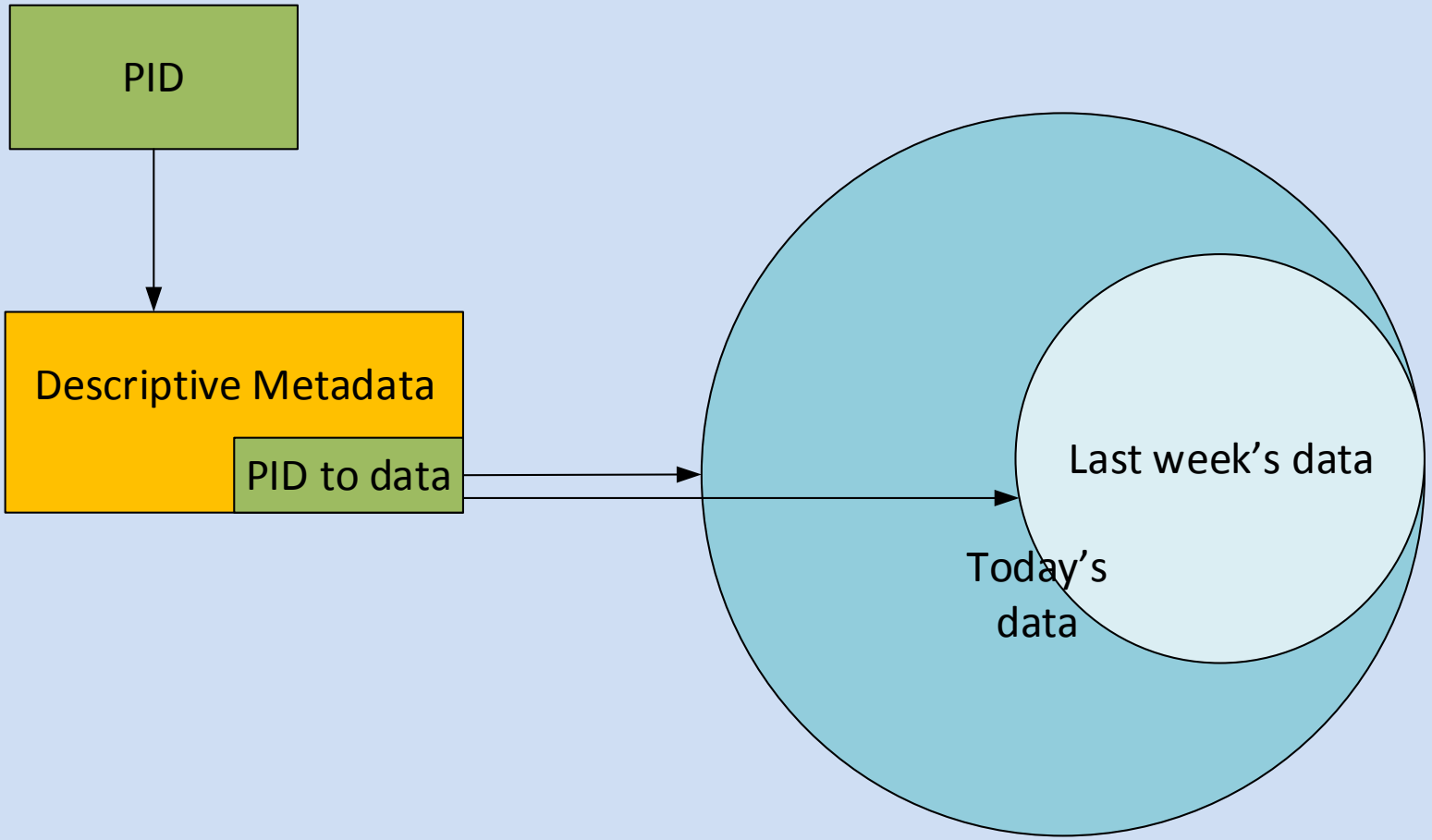
Before the update: “Corpus of Finnish Magazines and Newspapers from the 1990s and 2000s” in 2 variants: Download and Korp



Before the update: “Corpus of Finnish Magazines and Newspapers from the 1990s and 2000s”, Download and Korp



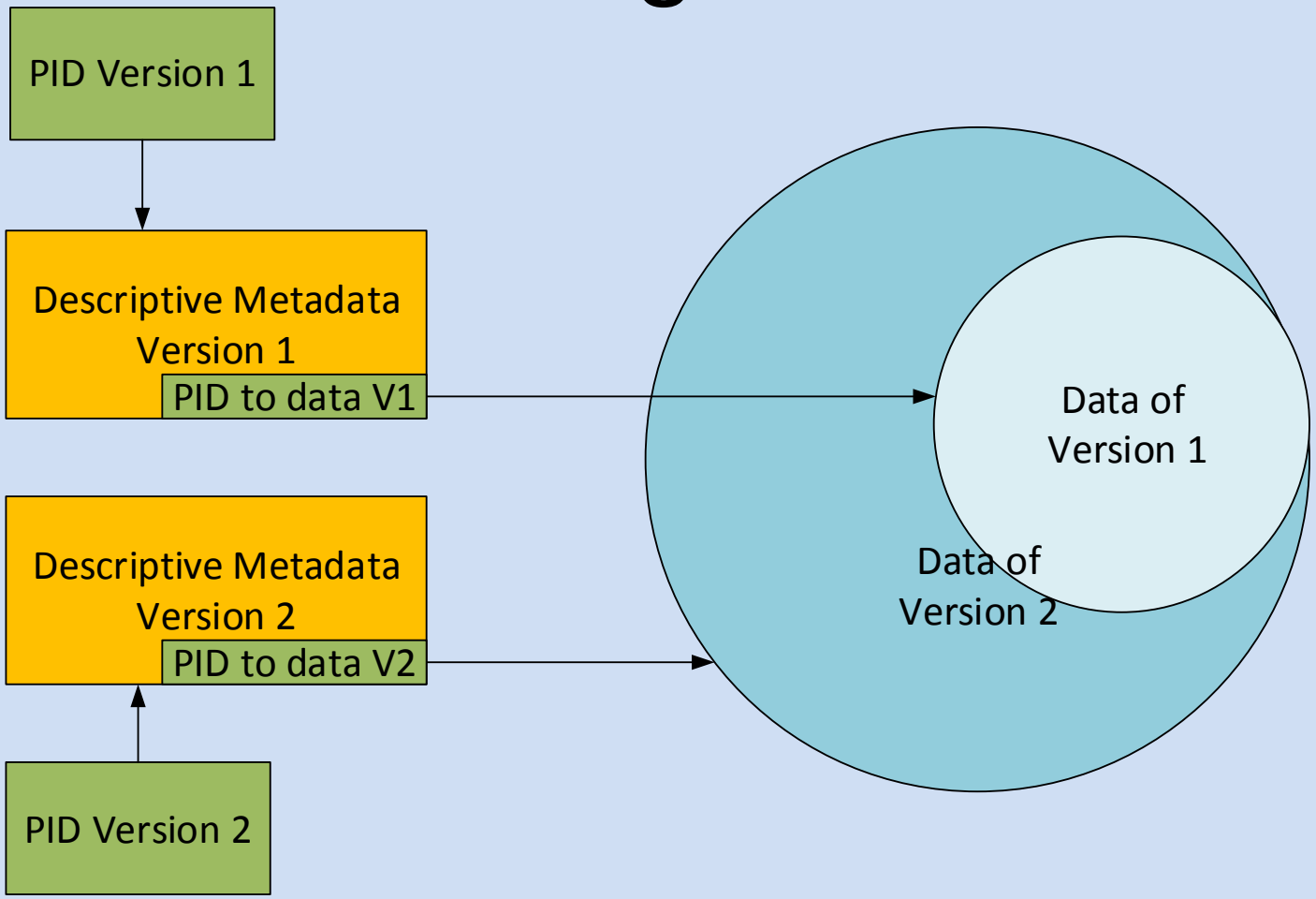
Unversioned



The original plan

- Synchronize the Download and Korp variants of the accruing dataset
- Add new subcorpora
- Update the PIDs and metadata

Introducing versions



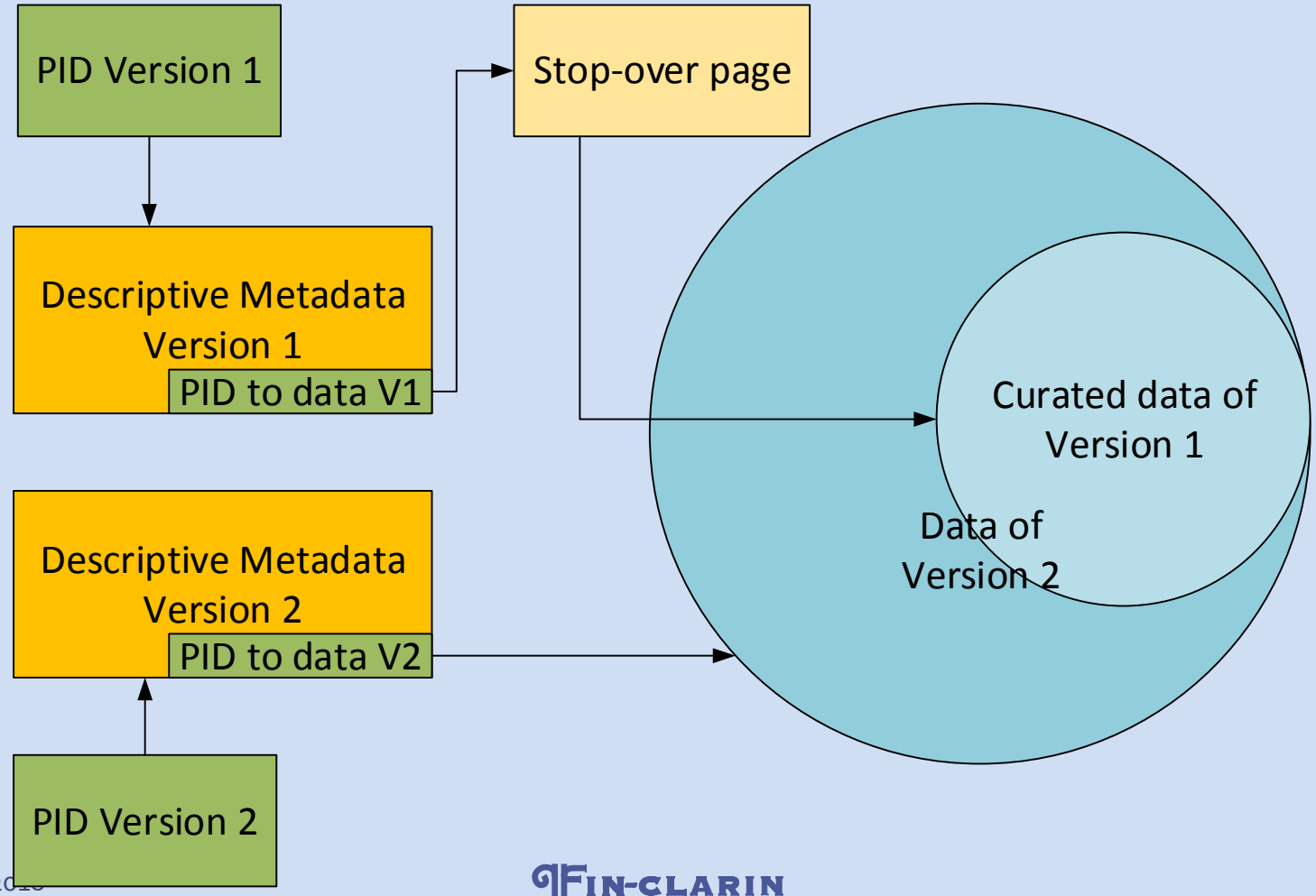
Problems

- Missing data in Korp and Download
- Inconsistencies, typos in zip files
- Irrelevant temporary files/directories (“.tmp”)
- How to define version 1
  - Variants not in sync
- What to do with the old, uncorrected data?

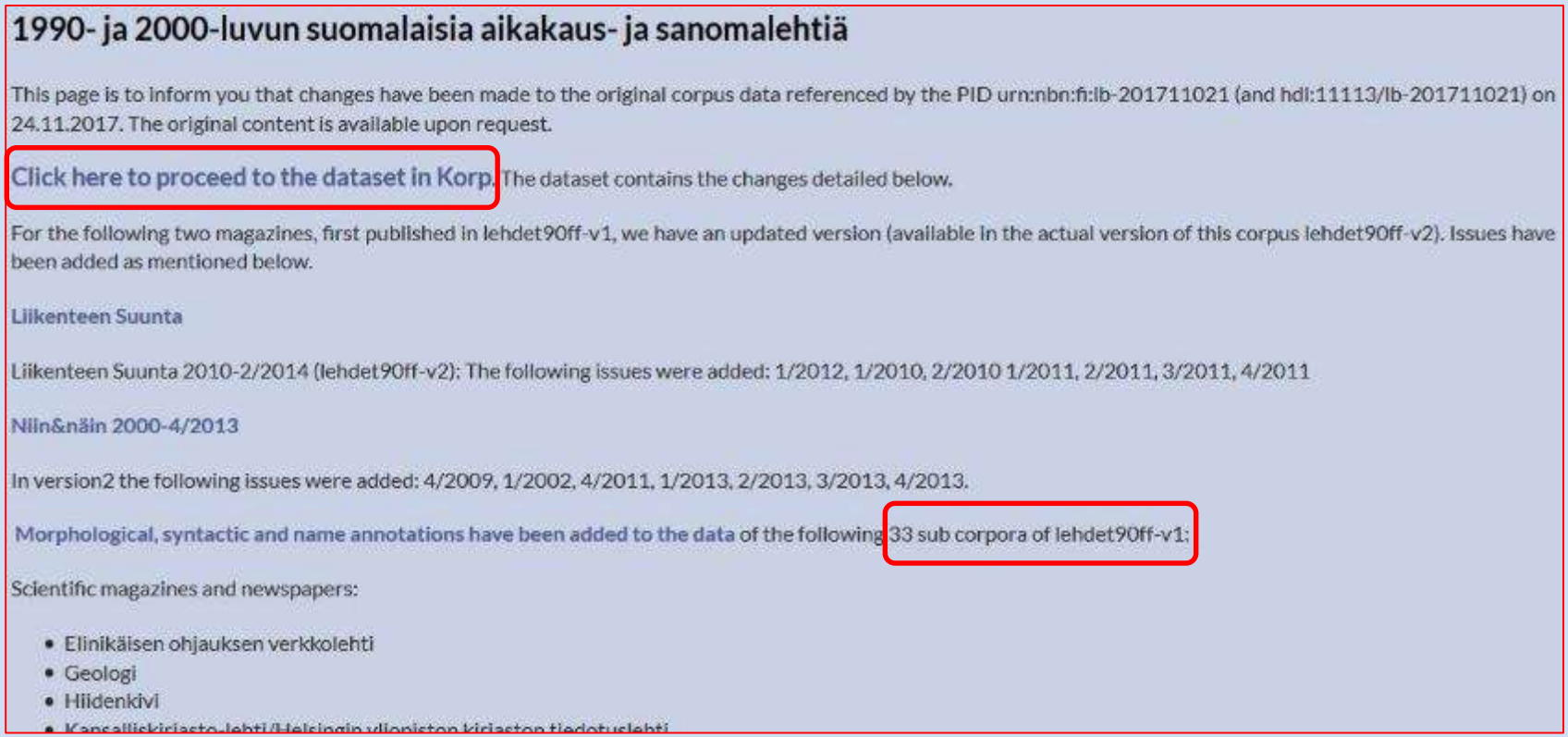
Decisions

- Documented not in sync version 1
- Minor changes: Change Log
- Major changes: The new version.
- The new subset: Stop-over page

The stop-over page



The stop-over page



1990- ja 2000-luvun suomalaisia aikakaus- ja sanomalehtiä

This page is to inform you that changes have been made to the original corpus data referenced by the PID urn:nbn:fi:lb-201711021 (and hdl:11113/lb-201711021) on 24.11.2017. The original content is available upon request.

Click here to proceed to the dataset in Korp. The dataset contains the changes detailed below.

For the following two magazines, first published in lehdet90ff-v1, we have an updated version (available in the actual version of this corpus lehdet90ff-v2). Issues have been added as mentioned below.

Liikenteen Suunta

Liikenteen Suunta 2010-2/2014 (lehdet90ff-v2): The following issues were added: 1/2012, 1/2010, 2/2010 1/2011, 2/2011, 3/2011, 4/2011

Niin&näin 2000-4/2013

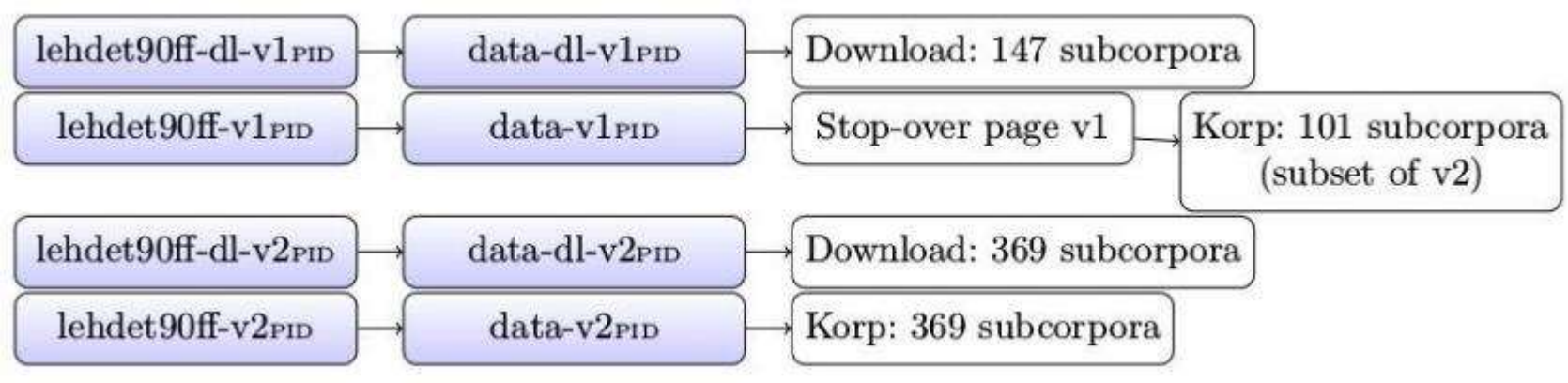
In version2 the following issues were added: 4/2009, 1/2002, 4/2011, 1/2013, 2/2013, 3/2013, 4/2013.

Morphological, syntactic and name annotations have been added to the data of the following 33 sub corpora of lehdet90ff-v1:

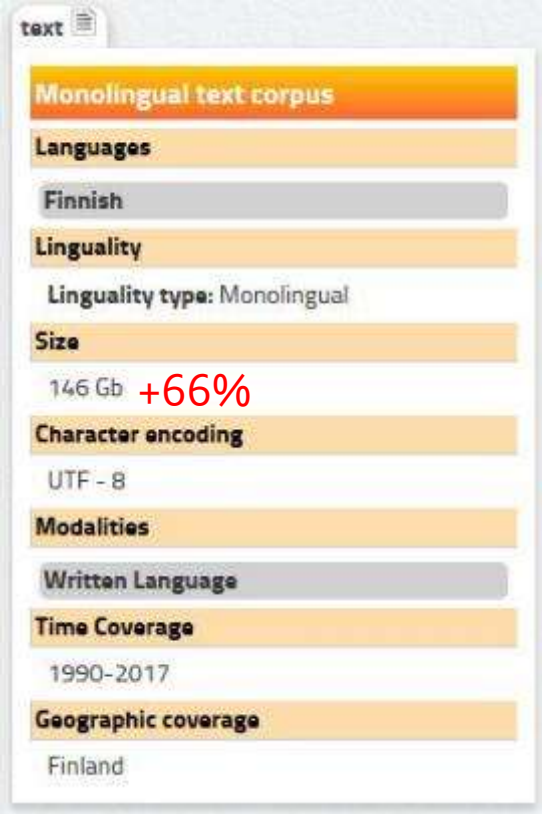
Scientific magazines and newspapers:

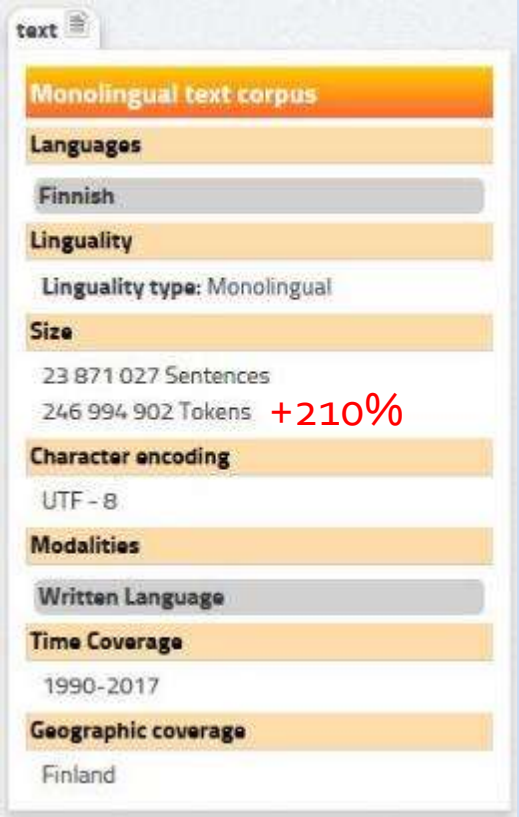
- Elinikäisen ohjauksen verkkolehti
- Geologi
- Hiidenkivi
- Kansalliskirjasto-lehti/Helsingin yliopiston kirjaston tiedotuslehti

After the update

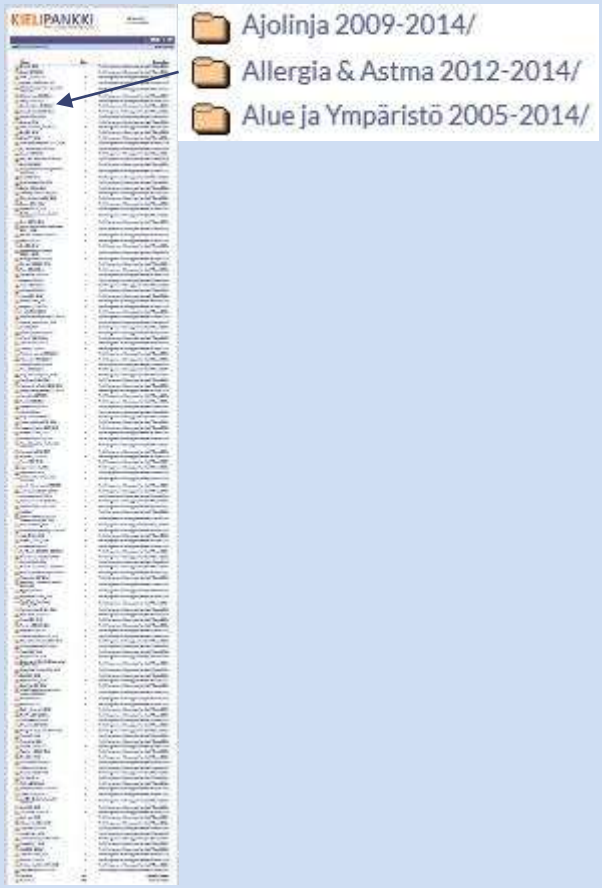


"Corpus of Finnish Magazines and Newspapers from the 1990s and 2000s"
Version 2, in 2 variants: Download and Korp



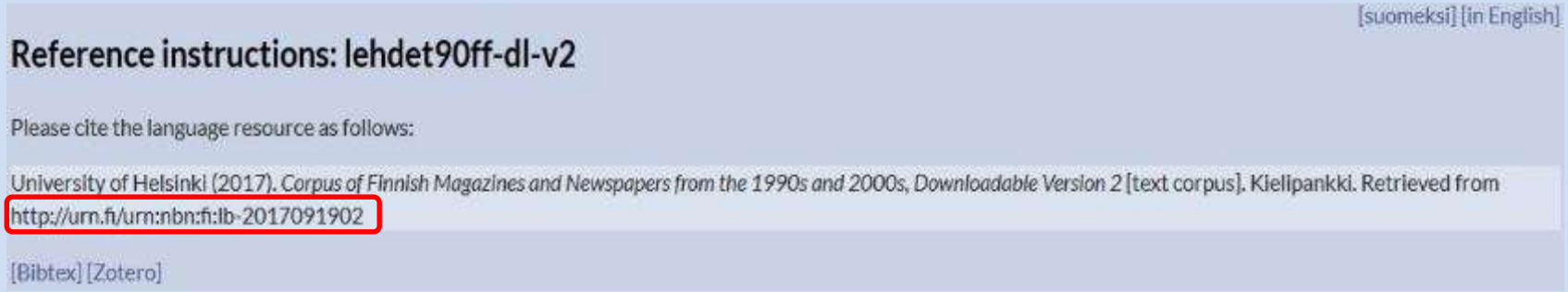


“Corpus of Finnish Magazines and Newspapers from the 1990s and 2000s” in 2 variants: Download and Korp



What are PIDs *really* for?

- Not for data objects.
- Not for data access locations.
- Not for license pages.
- Only for **citable** authoritative metadata:


[suomeksi] [in English]

Reference instructions: lehdet90ff-dl-v2

Please cite the language resource as follows:

University of Helsinki (2017). *Corpus of Finnish Magazines and Newspapers from the 1990s and 2000s, Downloadable Version 2* [text corpus]. Kielipankki. Retrieved from http://urn.fi/urn:nbn:fi:lb-2017091902

[Bibtex] [Zotero]

PIDs and reproducibility

- Reproducibility will become more relevant for our field.
- Only **citable** PIDs together with transparent data update policies support reproducibility.

A PID is a promise

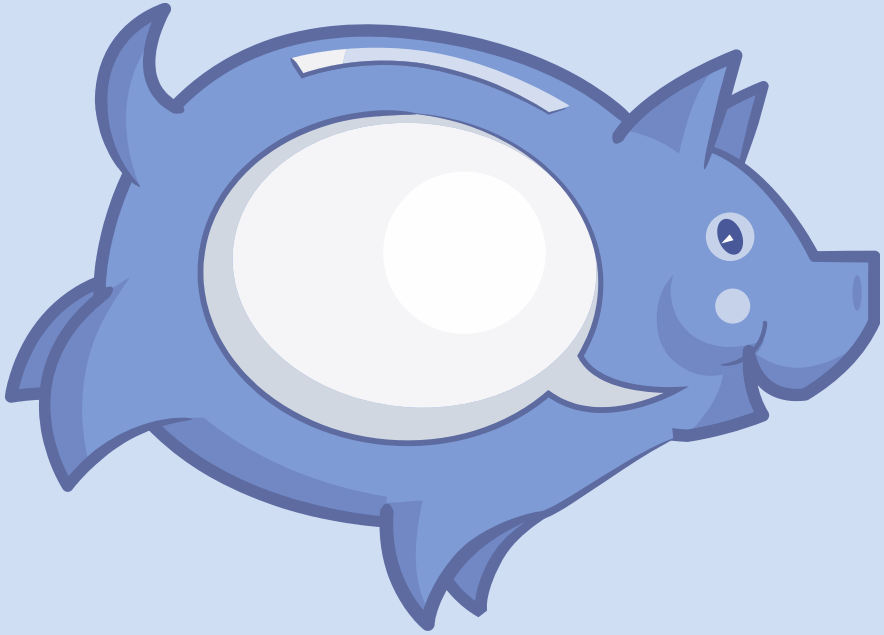
Less is more

- From a paper about PIDs: hdl:*11314.2/d5396a97c316a0eaca055846ba4233ac*
- ►http://**38.100.130.13**:8002/registrar/?view=11314.2/07841c3f84cbe0d4ff8687d0028c2622
- IP belongs to: **CNRI**, the Handle Registry

Conclusions

- Don’t issue PIDs you cannot curate.
  - In our case: 4 PIDs
- Assume PIDs need curation over time.
- Keep reproducibility in mind.

KIELIPANKKI
The Language Bank of Finland



Thank you!