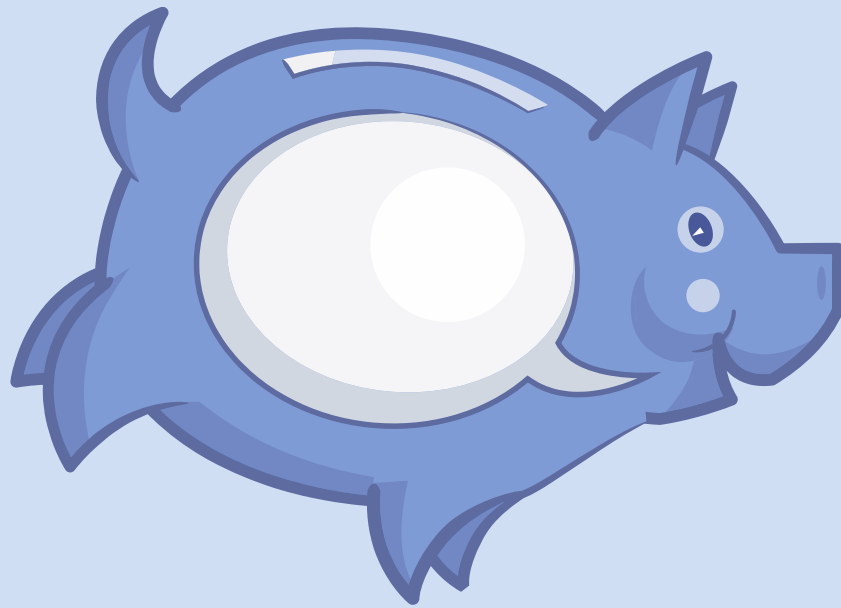




# **CLARIN Annual Conference 2018**

**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



# 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.

**Organization**  
Who are the Language Bank?

**Support**  
Help and instructions.

Search the Language Bank Portal:  
[Search bar] [Like]

Researcher of the Month: Anniika Lilli

**News**

- Researcher of the Month: Anniika Lilli (13.9.2018)
- Researcher of the Month: Marja Hukka-Pillanen (6.8.2018)
- Researcher of the Month: Miirta-Liina Laitinen (9.7.2018)
- Researcher of the Month: Maximilian Murrmann (11.4.2018)
- Researcher of the Month: Jarna Kivikukka (7.5.2018)

What is the Language Bank of Finland?  
The Language Bank of Finland is a service for researchers using language resources. The Language Bank has a wide va-

**CLARIN CENTRE B**

**CORE TRUST SEAL**

© 2015-2018 The Language Bank of Finland, FIN-CLARIN and CSC - IT Center for Science  
Privacy practices for the Language Bank of Finland

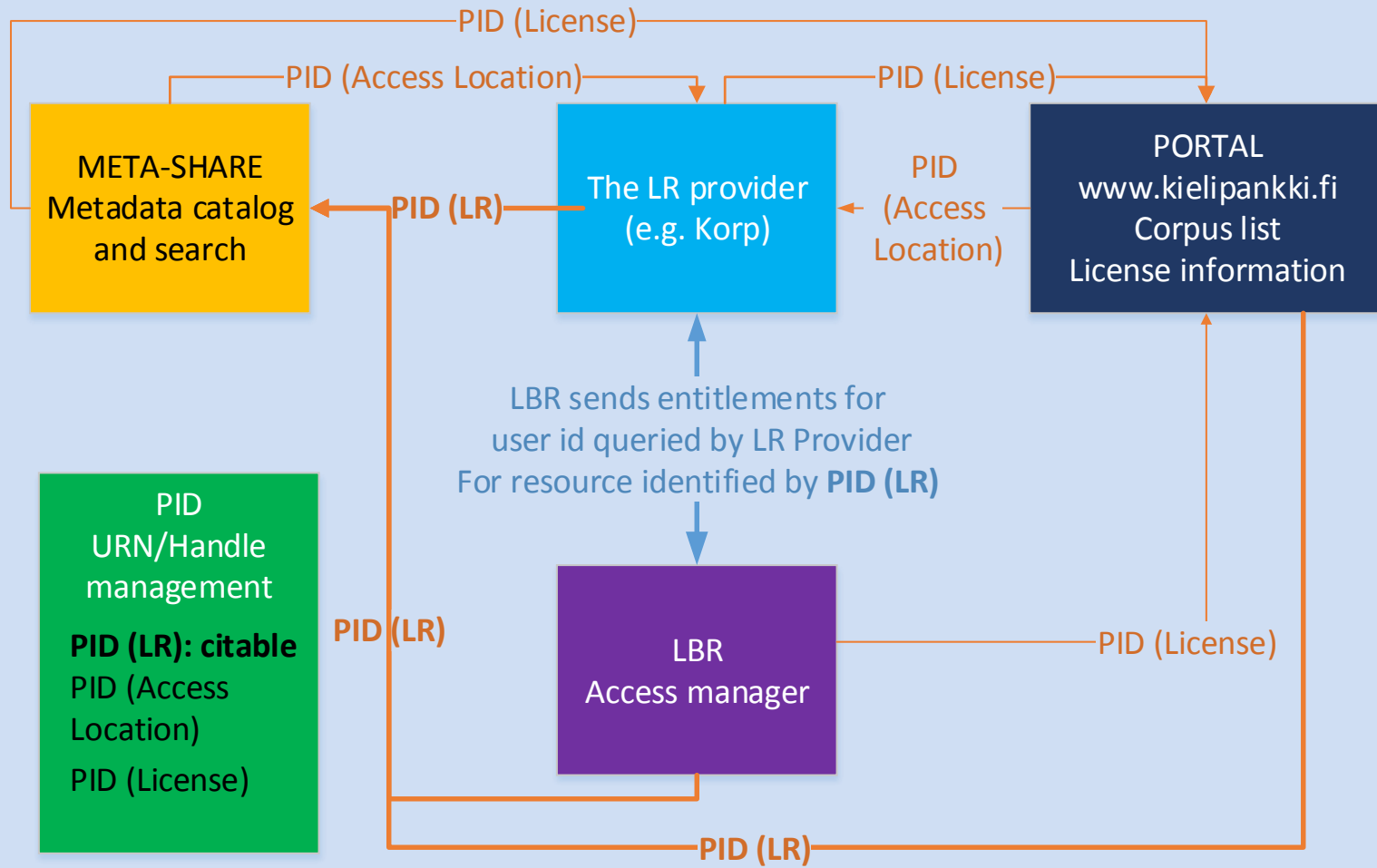


# Publication of a language resource at the Language Bank of Finland

- 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

Ajolinja 2009-2014/  
Allergia & Astma 2012-2014/  
Alue ja Ympäristö 2005-2014/

101 of 932 corpora selected – 80,49M of 8,74G tokens

1990- ja 2000-luvun suomalaisia aikakaus- ja sanomalehtiä (349)

Tiedelehtiä (99)

A-G (21)

- 30 Päiväs
- Aakusti
- Agricolan Tietosanommat
- Aikakauskirja Äidinkielen opetustiede
- Aikuskasvatus
- Alue ja ympäristö
- Ammattikasvatuksen aikakauskirja
- Apollon
- Arelopagi
- ATS-Ydintekniikka
- Auraica
- Automaatioväylä
- Aysin
- Baptria
- Bryobrotherella
- Diakonlan tutkimus - aikakauskirja
- Elinikäisen ohjauksen verkkolehti
- Ennen ja nyt
- GeoForummi
- Geologi
- Glossae

H-K (20)



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

text

**Monolingual text corpus**

**Languages**

Finnish

**Linguality**

Linguality type: Monolingual

**Size**

88 Gb

**Character encoding**

UTF - 8

**Modalities**

Written Language

**Time Coverage**

1990-2016

**Geographic coverage**

Finland

text

**Monolingual text corpus**

**Languages**

Finnish

**Linguality**

Linguality type: Monolingual

**Size**

6 632 115 Sentences  
79 756 383 Tokens

**Character encoding**

UTF - 8

**Modalities**

Written Language

**Time Coverage**

1990-2016

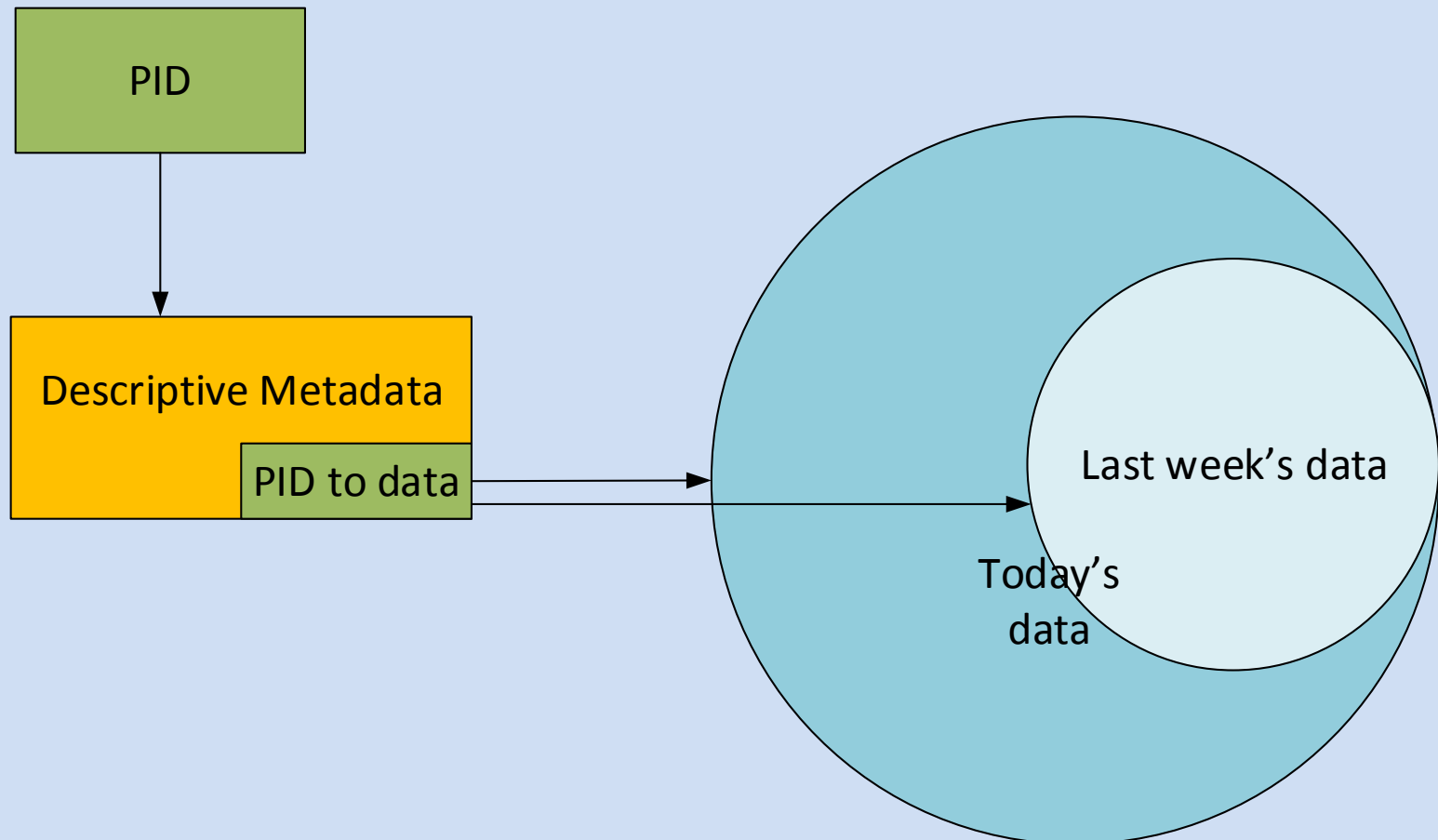
**Geographic coverage**

Finland





# Unversioned



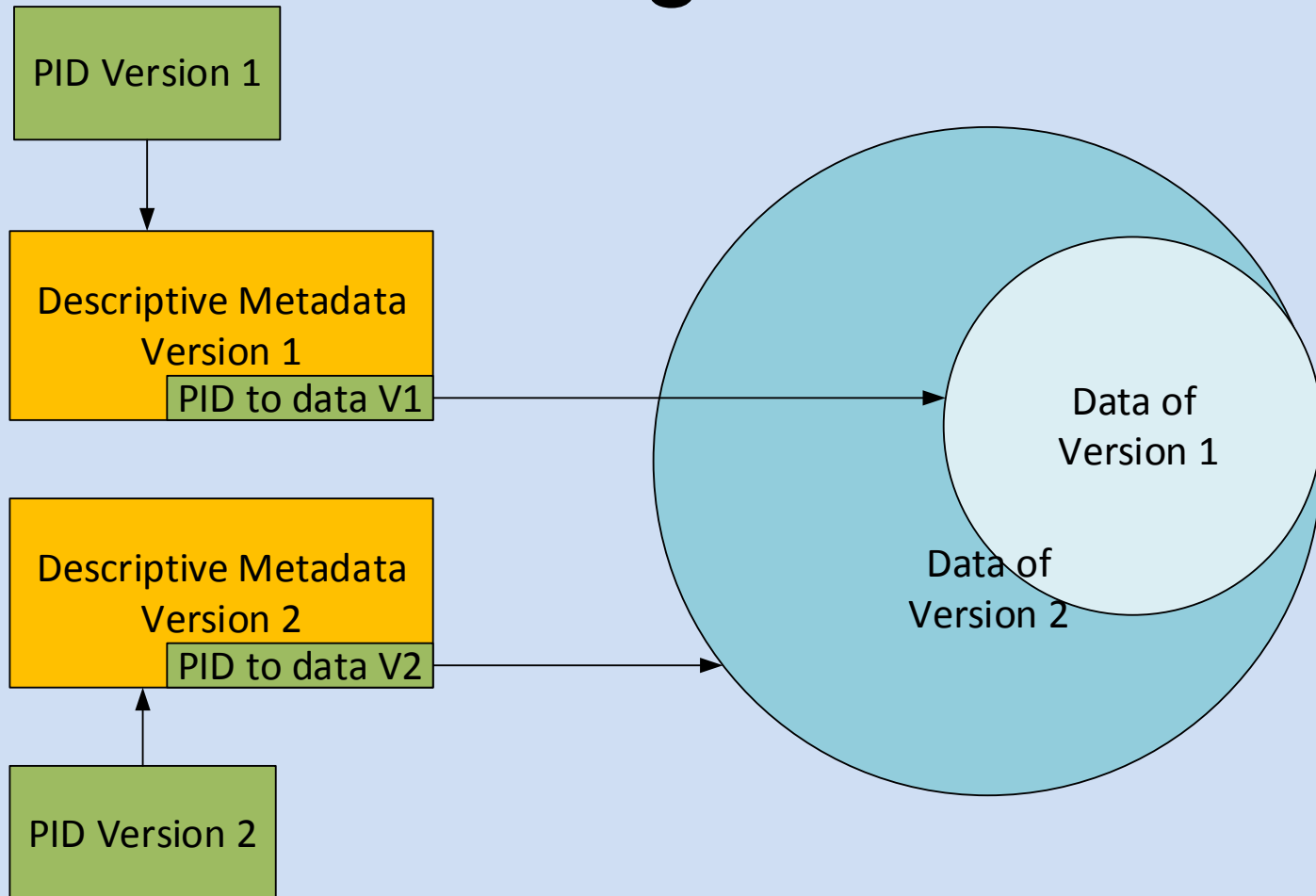


# The original plan

- 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?

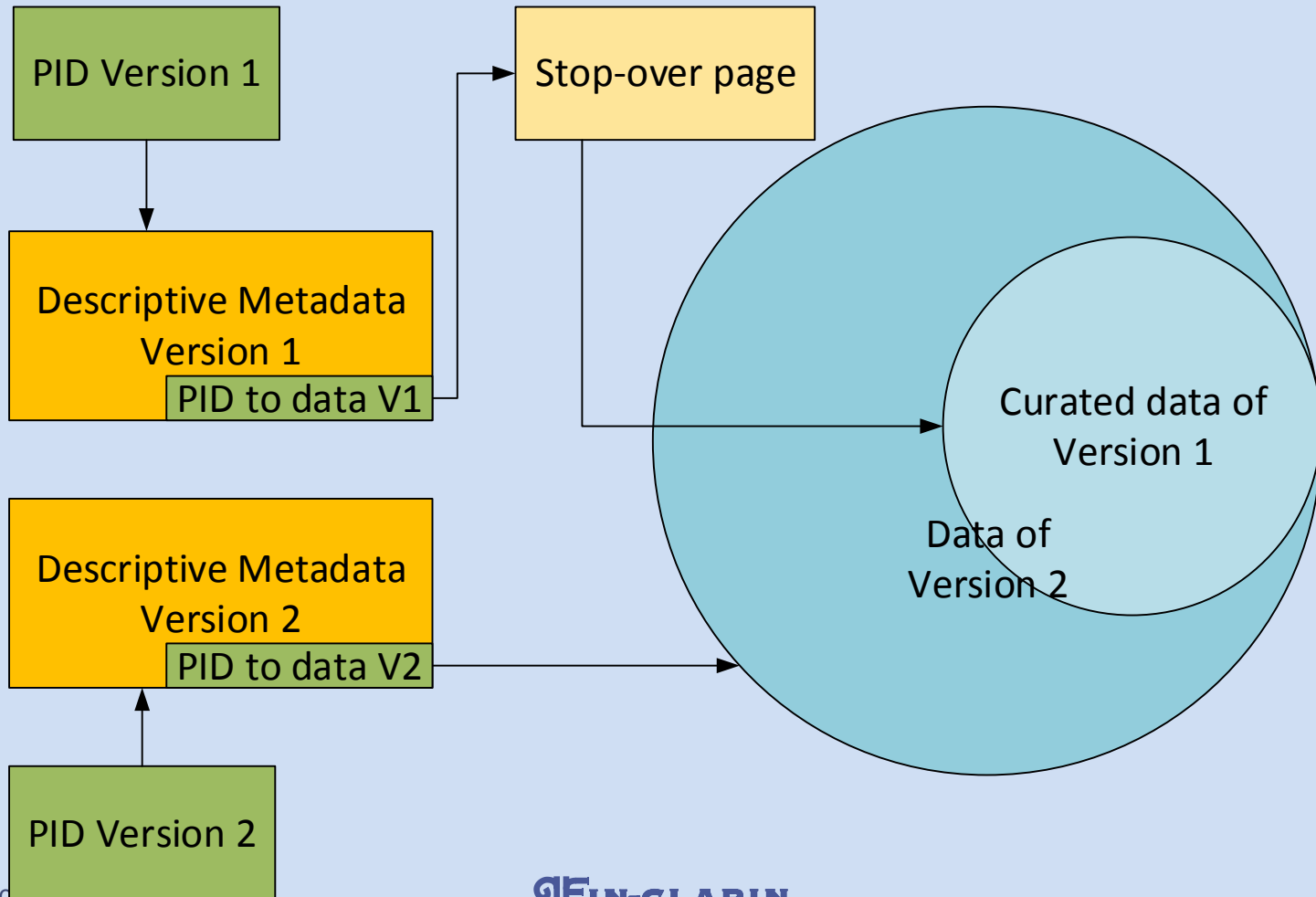


# Decisions

- 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.

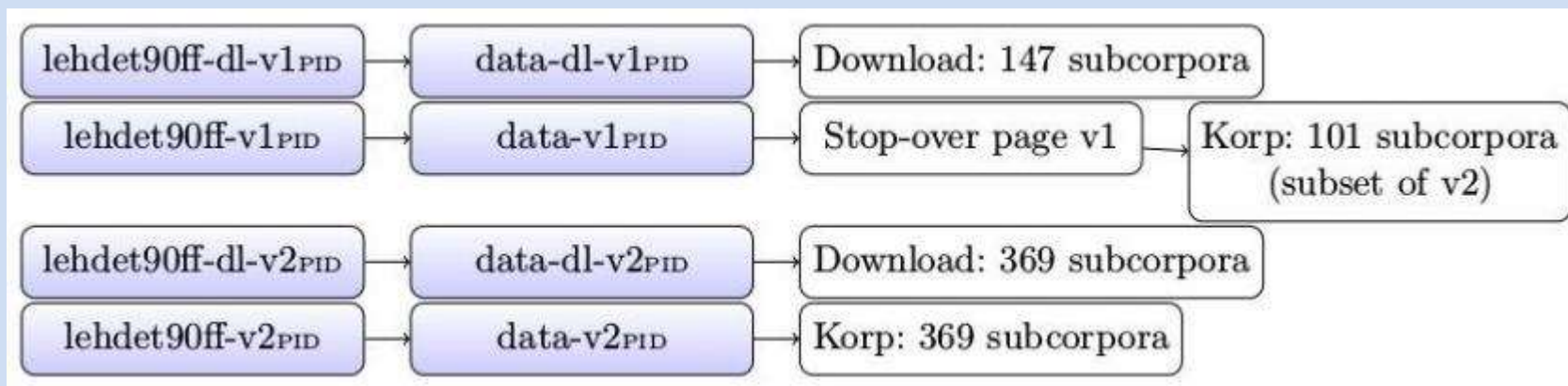
Morphological, syntactic and name annotations have been added to the data of the following [33 sub corpora of lehdet90ff-v1](#):

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

text

**Monolingual text corpus**

**Languages**

Finnish

**Linguality**

Linguality type: Monolingual

**Size**

146 Gb **+66%**

**Character encoding**

UTF - 8

**Modalities**

Written Language

**Time Coverage**

1990-2017

**Geographic coverage**

Finland

text

**Monolingual text corpus**

**Languages**

Finnish

**Linguality**

Linguality type: Monolingual

**Size**

23 871 027 Sentences  
246 994 902 Tokens **+210%**

**Character encoding**

UTF - 8

**Modalities**

Written Language

**Time Coverage**

1990-2017

**Geographic coverage**

Finland



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

Ajolinja 2009-2014/  
Allergia & Astma 2012-2014/  
Alue ja Ympäristö 2005-2014/

101 of 932 corpora selected – 80,49M of 8,74G tokens

1990- ja 2000-luvun suomalaisia aikakaus- ja sanomalehtiä (349)

Tiedelehtiä (99)

A-G (21)

- 30 Päivää
- Aakusti
- Agricolan Tietosanomat
- Aikakauskirja Äidinkielen opetustiede
- Aikuskasvatus
- Alue ja ympäristö
- Ammattikasvatuksen aikakauskirja
- Apollon
- Arelopagi
- ATS-Ydintekniikka
- Auraica
- Automaatioväylä
- Aysin
- Baptria
- Bryobrotherella
- Diakonlan tutkimus -aikakauskirja
- Elinikäisen ohjauksen verkkolehti
- Ennen ja nyt
- GeoForummi
- Geologi
- Glossae

H-K (20)



# 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



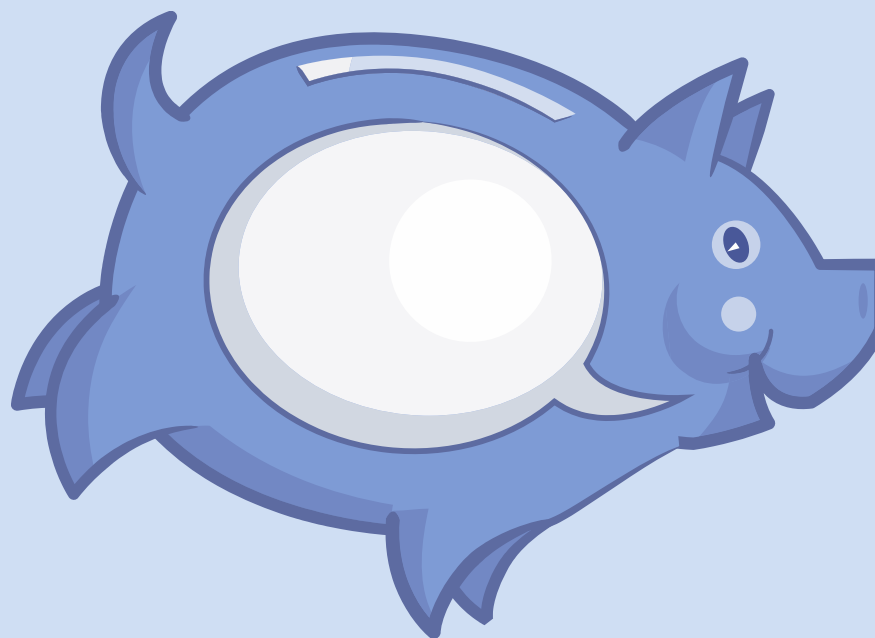


# Conclusions

- 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!