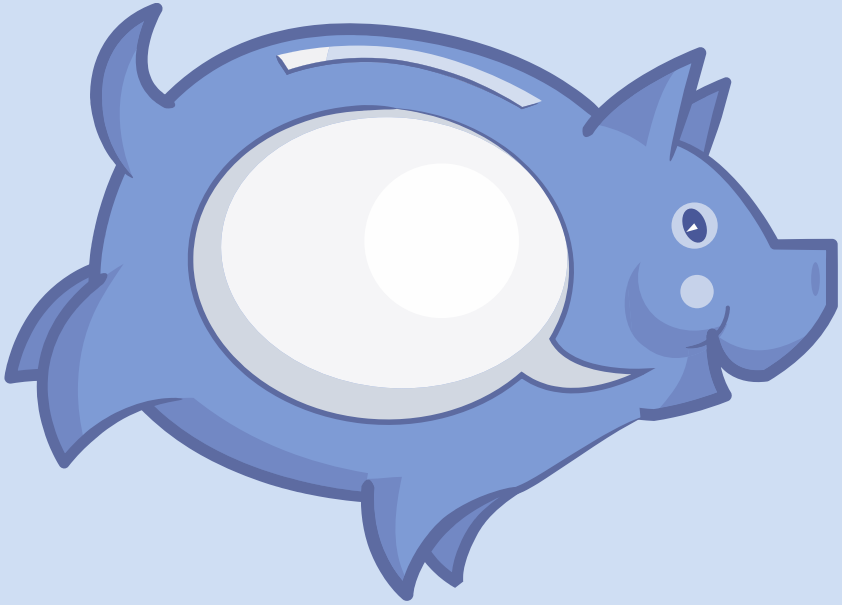


KIELIPANKKI
The Language Bank of Finland

# Updating a dataset using PIDs and Metadata

Martin Matthiesen, Ute Dieckmann

CLARIN Centre Meeting, Utrecht 5.6.2018

FIN-CLARIN

# Contents

- Background
- The dataset
- Before the update
- Decisions
- After the update
- Conclusions
- References

# Background

- Experience with accruing datasets (Suomi24)
- This one (”lehdet90ff”) considered different.
- PID on a moving target

# The dataset

- "Corpus of Finnish Magazines and Newspapers from the 1990s and 2000s"
- To be updated in irregular intervals
- 2 (P)IDs, 2 related corpora:
  - Zip files for dowload (partly with PDFs)
  - Text only version in Korp
- 2+ PIDs for access locations

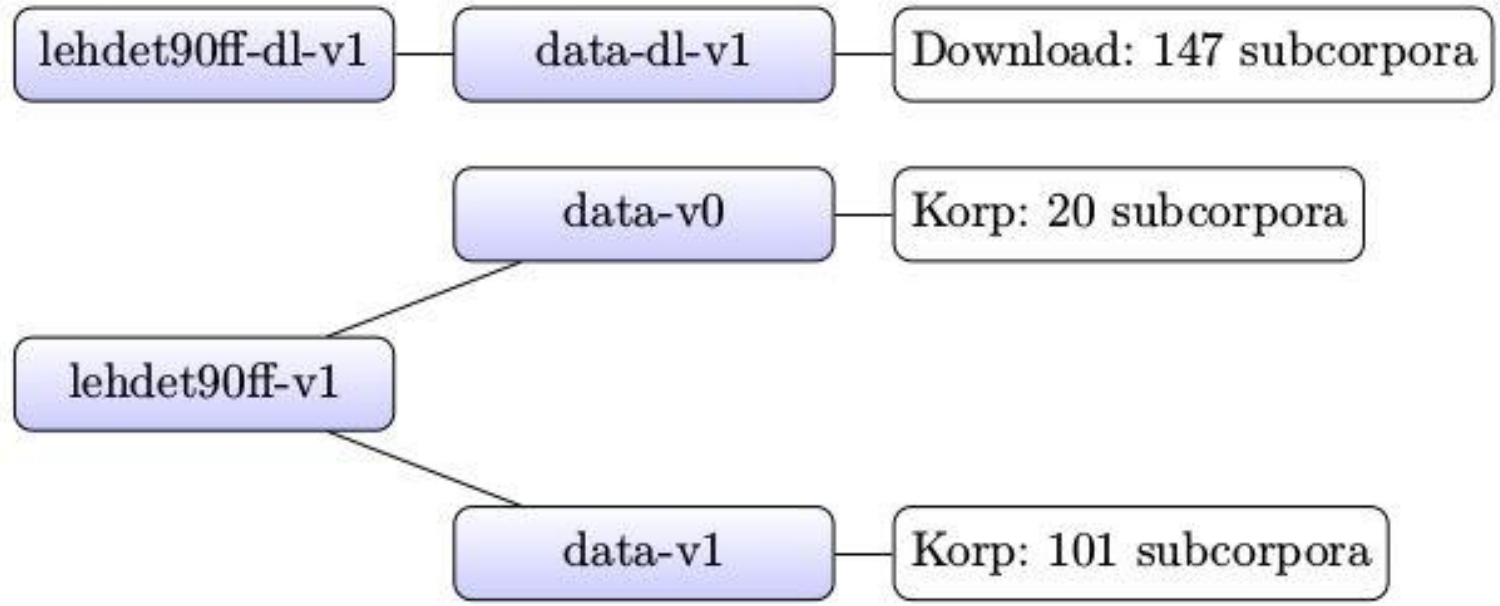
# Before the update 1/2

- PIDs to Metadata not updated when data changed. New subcorpora listed in Metadata.
- Korp: Access PIDs updated inconsistently:
  - PID1: 2 subcorpora
  - PID2: 20 subcorpora
- Last PID did not reference all 101 Korp subcorpora
- Download: PID to top level directory references new data automatically.
- No version number

# Before the update 2/2

- Some Korp subcorpora were not properly annotated.
- Some Korp subcorpora were missing data due to problems with the conversion.
- Missing license and README information in zip files
- Missing directories in zip files.
- Directory structure of zips was generally not consistent.
- Files zipped on a Mac had filename encoding problems on Linux.
- Some zip files contained thumbnails and other irrelevant temporary files/directories.

# PIDs before update



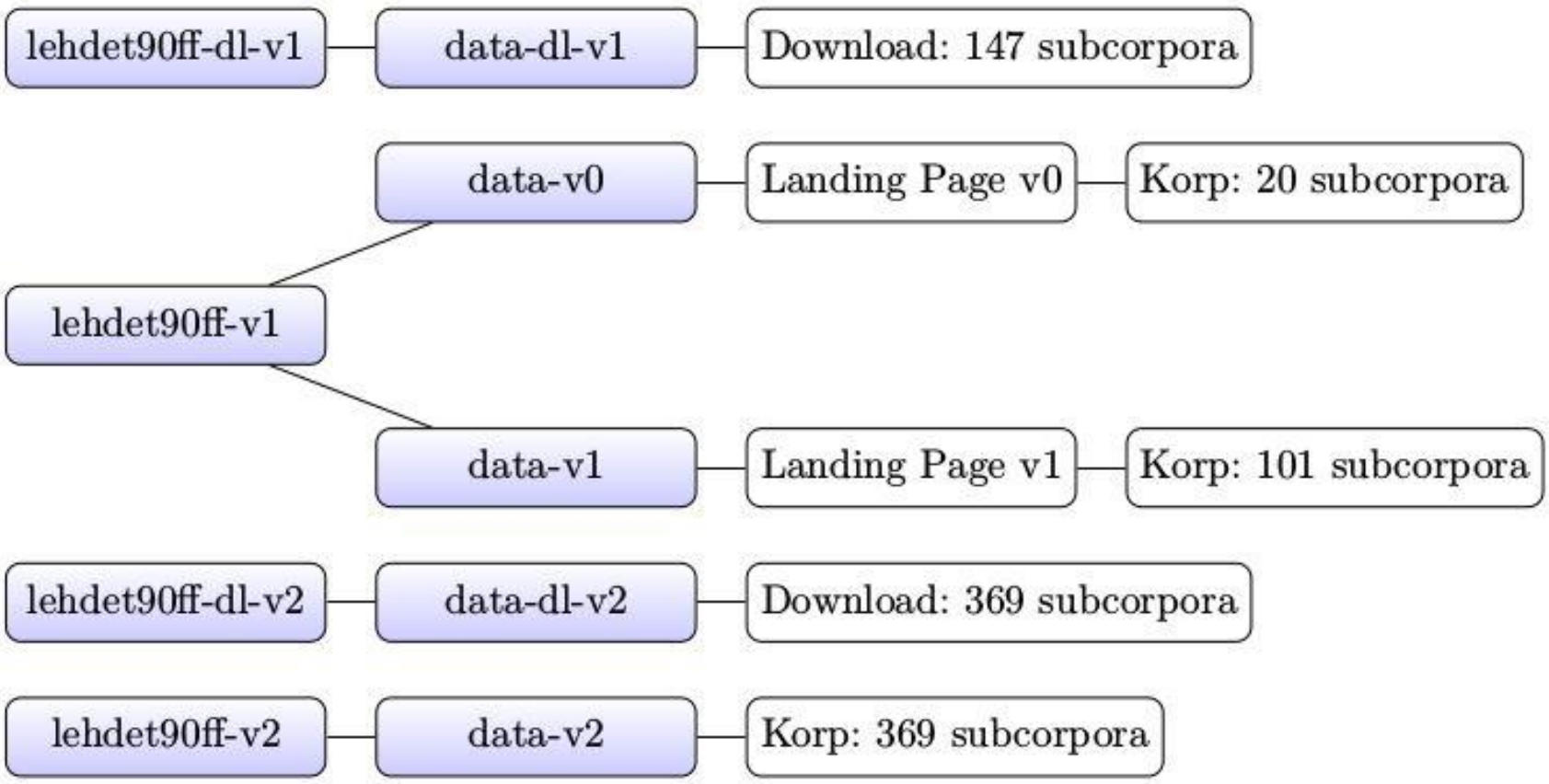
# Decisions

- Abandon the idea of an accruing dataset variant behind a single PID.
- Begin versioning:
  - Create a version 1 of the current Korp variant of the dataset.
  - Obsolete access PIDs to former Korp variants hidden, changes marked in landing page
  - Create a version 1 of the current downloadable dataset.
  - Make explicit, that the variants of version 1 are not in sync.
- Keep the idea of having one PID per variant and version.
- Introduce “intermediate landing pages” for PIDs pointing to corrected data in version 1.
- Non-significant changes are marked in a new Change Log section in the Metadata.

# After the update

- PIDs partly redirected to ”Stop Over Landing Pages”
- Change Log for ”minor changes”
- Detailed list of changes in sub corpora

# PIDs after update



# No PIDs to individual files

- Version 2 of the downloadable corpus
  - 369 subcorpora
  - 574 zip files
  - 88718 individual files.
- Update would have broken most old direct PIDs

# Conclusions

- Accruing datasets need versioning.
- PIDs should be assigned sparingly.
- Not every change warrants a new PID.

# Example PIDs

- Access location ”v0”(20 subcorpora, (leads to landing page): http://urn.fi/urn:nbn:fi:lb-2016021202
- lehdet90ff-v1: http://urn.fi/urn:nbn:fi:lb-2016011101 (Access PID in Metadata leads also to landing page, Relation leads to version 2)

# Thank you!