AKADEMIE DER WISSENSCHAFTEN IN HAMBURG

Towards Continuous Quality Control for Spoken Language Corpora

Anne Ferger and Hanna Hedeland

University of Hamburg

INEL

Grammatical Descriptions, Corpora and Language Technology for **I**ndigenous **N**orthern **E**urasian **L**anguages

HZSK

Hamburg Center for Language Corpora

Akademie der Wissenschaften in Hamburg

Union of the German Academies of Sciences and Humanities in Hamburg

CLARIN

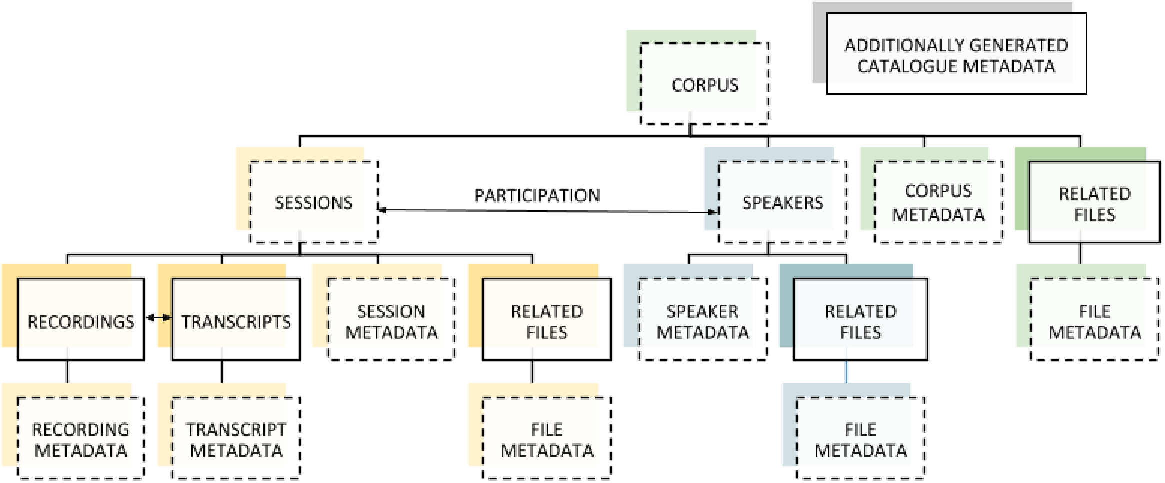
Common Language Resources and Technology Infrastructure

Aim of the Presentation

- Our approach on optimizing a **linguistic data creation and curation workflow** aiming towards **continuous integration** of speech corpora
- The **data** we work with
- Our **framework** and **practical issues** we overcame
- Our **perspective**

Language Corpora

Structure of a spoken language corpus



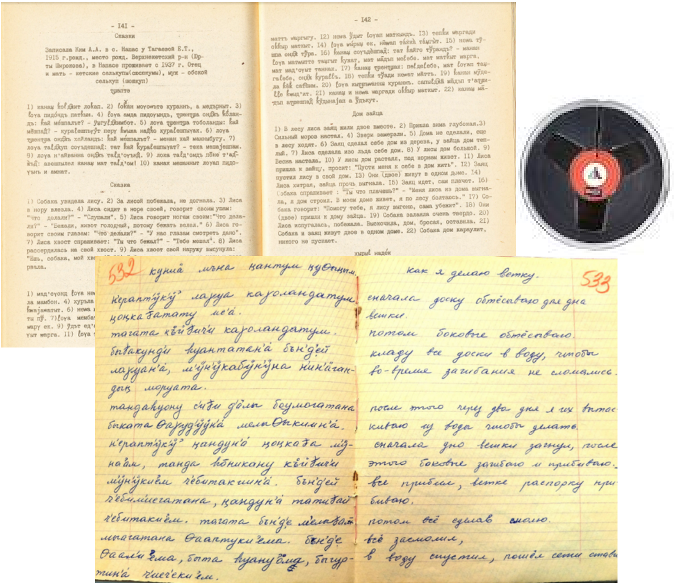
Exemplary Data in the INEL Project

Language Data

- with and without audio and video
- transcriptions in XML format

Metadata

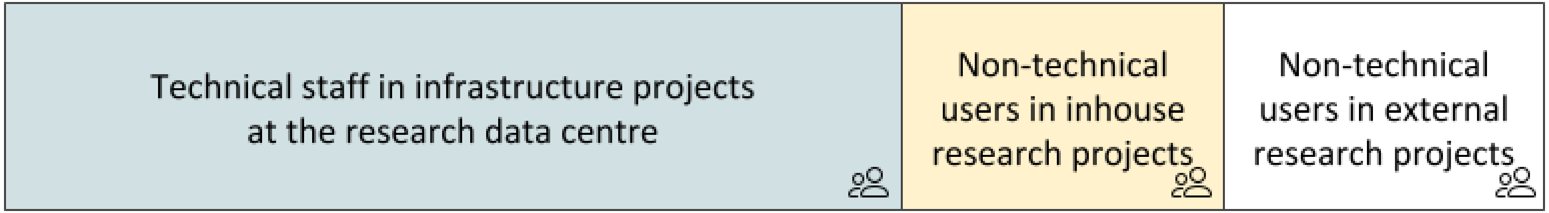
- diverse corpus-wide metadata
- in XML and CMDI format



Quality Management for Spoken Language Corpora

- Existing linguistic **data creation and curation** workflow: mostly manually and non-reproducible
- In our case: Creating **searchable, consistent language corpora** that can be used for quantitive or qualitative analysis
- Publishing that corpora in a Fedora **repository**
- **Completely automated curation** is **not possible** because it would require unacceptable constraints on the creation of the data

Users and Maintainers of Our Approach



Technical staff in infrastructure projects at the research data centre

Non-technical users in inhouse research projects

Non-technical users in external research projects

Our Approach

Why do we need continuous quality control for spoken language corpora?

- Limiting (expensive) **manual work**
- **Avoid** unnecessary **data curation**
- Increasing the amount of **automatic enhancements** of the data
- Creating **high quality resources** suitable for different research needs
- Making the **publishing** of the resources as fast, spontaneous and easy as possible
- Enabling the **conversion** of the data into various different formats

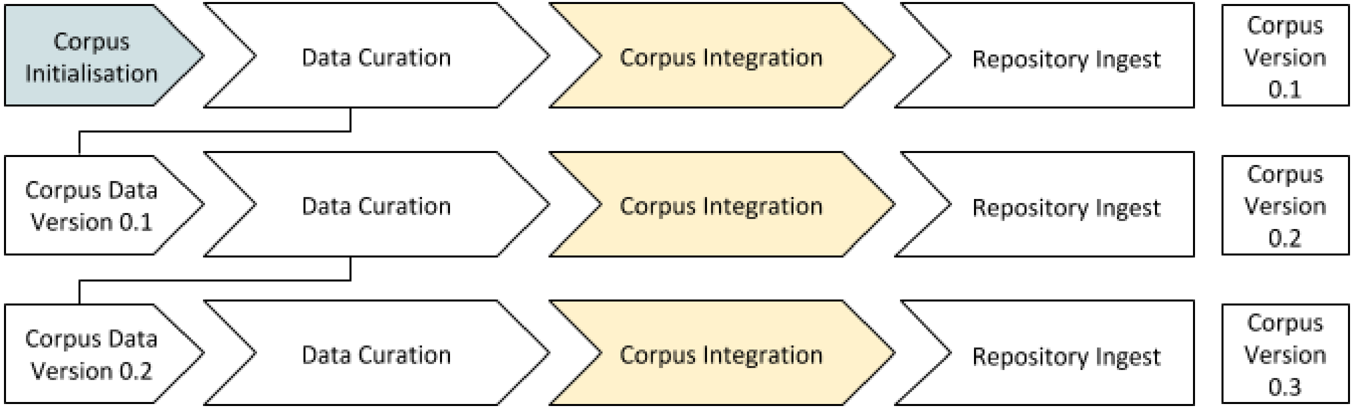
Formalizing the Workflows

Corpus Initialisation

- Prepare files and folders
- Set up Git repository
- Validate and assess Data
- Create error reports
- Create curation tasks

Corpus Integration

- Create derived corpus file formats
- Clean up file references within corpus
- Complete metadata and documentation
- Generate generic and specific metadata formats
- Finalise and tag publication version



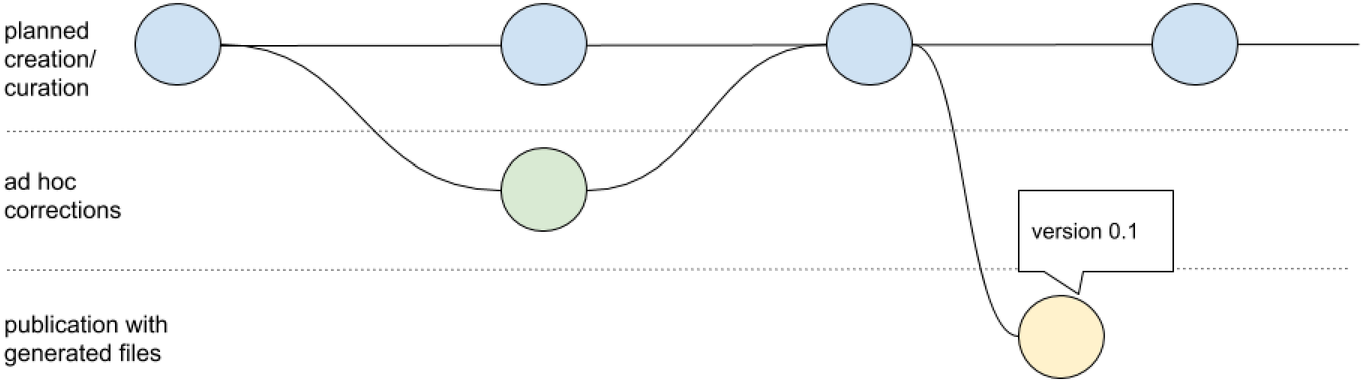
Corpus Initialisation → Data Curation → Corpus Integration → Repository Ingest → Corpus Version 0.1

Corpus Data Version 0.1 → Data Curation → Corpus Integration → Repository Ingest → Corpus Version 0.2

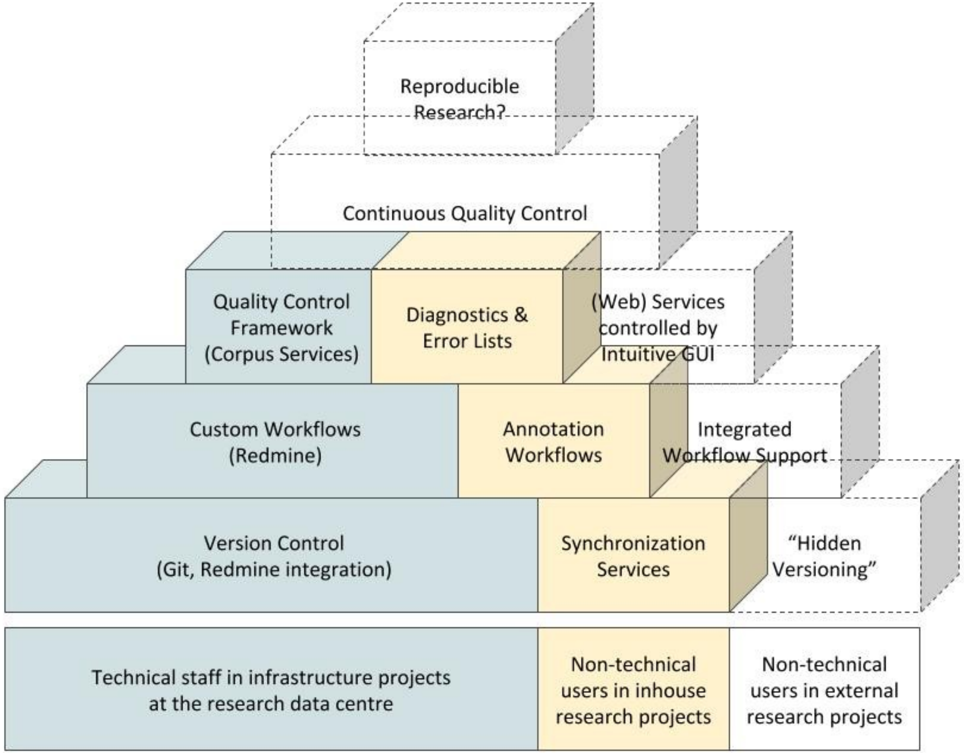
Corpus Data Version 0.2 → Data Curation → Corpus Integration → Repository Ingest → Corpus Version 0.3

Using Git for Version Control

Using a git workflow



The Modules Used for Our Approach



Reproducible Research?

Continuous Quality Control

Quality Control Framework (Corpus Services)

Diagnostics & Error Lists

(Web) Services controlled by Intuitive GUI

Custom Workflows (Redmine)

Annotation Workflows

Integrated Workflow Support

Version Control (Git, Redmine integration)

Synchronization Services

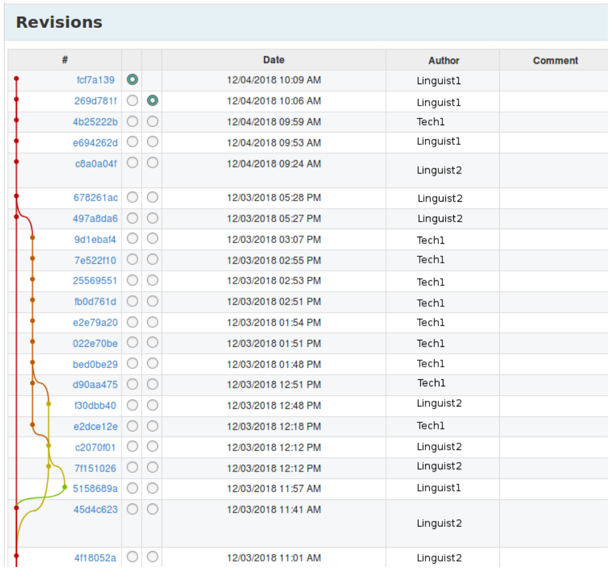
“Hidden Versioning”

Technical staff in infrastructure projects at the research data centre

Non-technical users in inhouse research projects

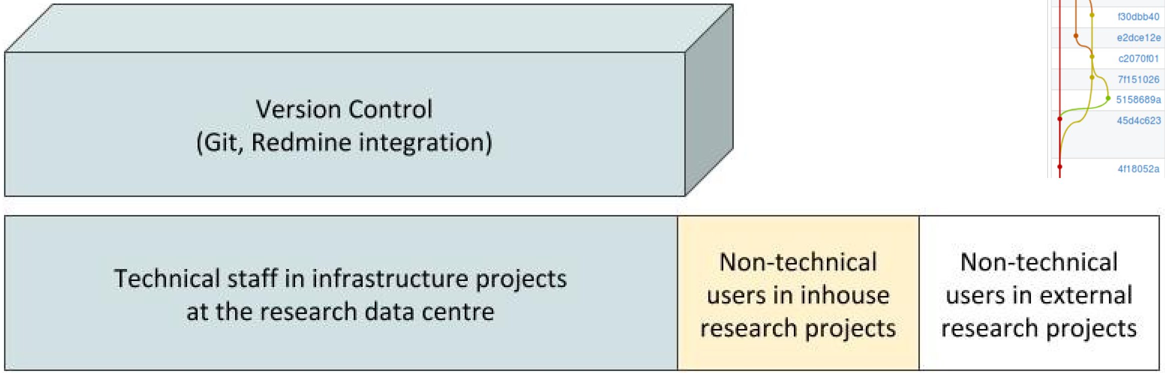
Non-technical users in external research projects

Specific Git Solutions I



Revisions

#			Date	Author	Comment
fcf7a139	●		12/04/2018 10:09 AM	Linguist1	
269d781f	○	●	12/04/2018 10:06 AM	Linguist1	
4b25222b	○	○	12/04/2018 09:59 AM	Tech1	
e694262d	○	○	12/04/2018 09:53 AM	Linguist1	
c8a0a04f	○	○	12/04/2018 09:24 AM	Linguist2	
678261ac	○	○	12/03/2018 05:28 PM	Linguist2	
497a8da6	○	○	12/03/2018 05:27 PM	Linguist2	
9d1ebaf4	○	○	12/03/2018 03:07 PM	Tech1	
7e522f10	○	○	12/03/2018 02:55 PM	Tech1	
25569551	○	○	12/03/2018 02:53 PM	Tech1	
fb0d761d	○	○	12/03/2018 02:51 PM	Tech1	
e2e79a20	○	○	12/03/2018 01:54 PM	Tech1	
022e70be	○	○	12/03/2018 01:51 PM	Tech1	
bed0be29	○	○	12/03/2018 01:48 PM	Tech1	
d90aa475	○	○	12/03/2018 12:51 PM	Tech1	
f30dbb40	○	○	12/03/2018 12:48 PM	Linguist2	
e2dce12e	○	○	12/03/2018 12:18 PM	Tech1	
c2070f01	○	○	12/03/2018 12:12 PM	Linguist2	
7f151026	○	○	12/03/2018 12:12 PM	Linguist2	
5158689a	○	○	12/03/2018 11:57 AM	Linguist1	
45d4c623	○	○	12/03/2018 11:41 AM	Linguist2	
4f18052a	○	○	12/03/2018 11:01 AM	Linguist2	



Version Control
(Git, Redmine integration)

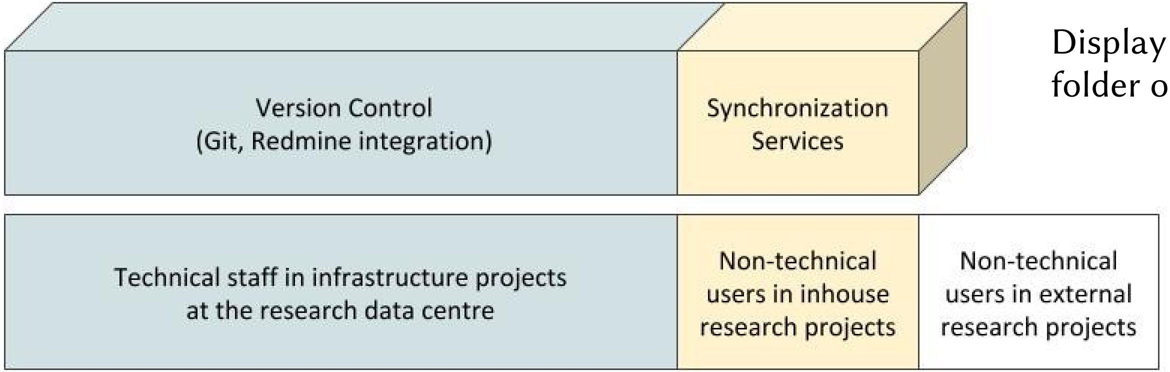
Technical staff in infrastructure projects at the research data centre

Non-technical users in inhouse research projects

Non-technical users in external research projects

Using different branches for publication

Specific Git Solutions II



Version Control
(Git, Redmine integration)

Synchronization Services

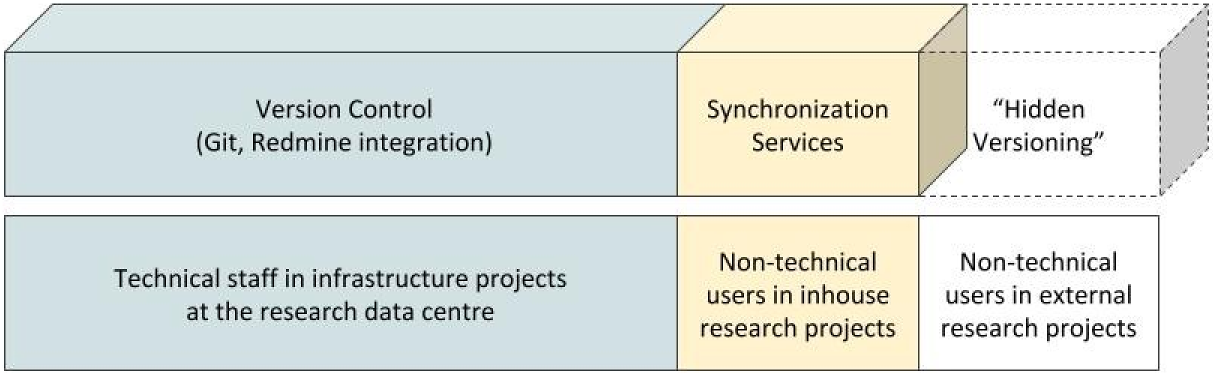
Technical staff in infrastructure projects at the research data centre

Non-technical users in inhouse research projects

Non-technical users in external research projects

Displaying the git repository as a folder on a shared drive

Specific Git Solutions III



Version Control
(Git, Redmine integration)

Synchronization
Services

“Hidden
Versioning”

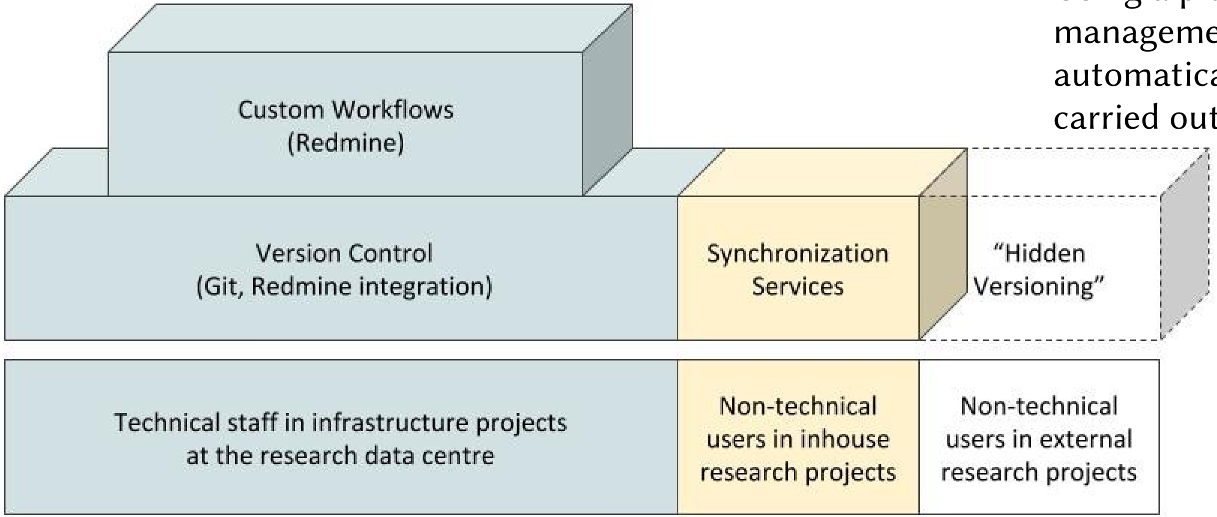
Technical staff in infrastructure projects
at the research data centre

Non-technical
users in inhouse
research projects

Non-technical
users in external
research projects

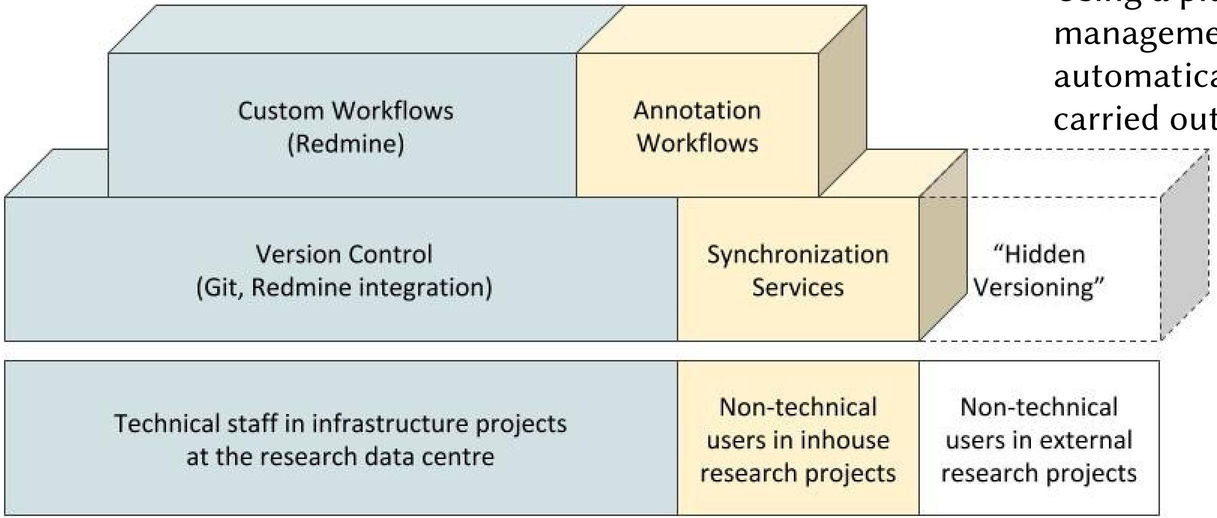
Scripts to let users work with git without noticing it

Automatically Supported Workflows I



Using a plugin in the project management software git to automatically create issues to be carried out

Automatically Supported Workflows II



Custom Workflows (Redmine)

Annotation Workflows

Version Control (Git, Redmine integration)

Synchronization Services

“Hidden Versioning”

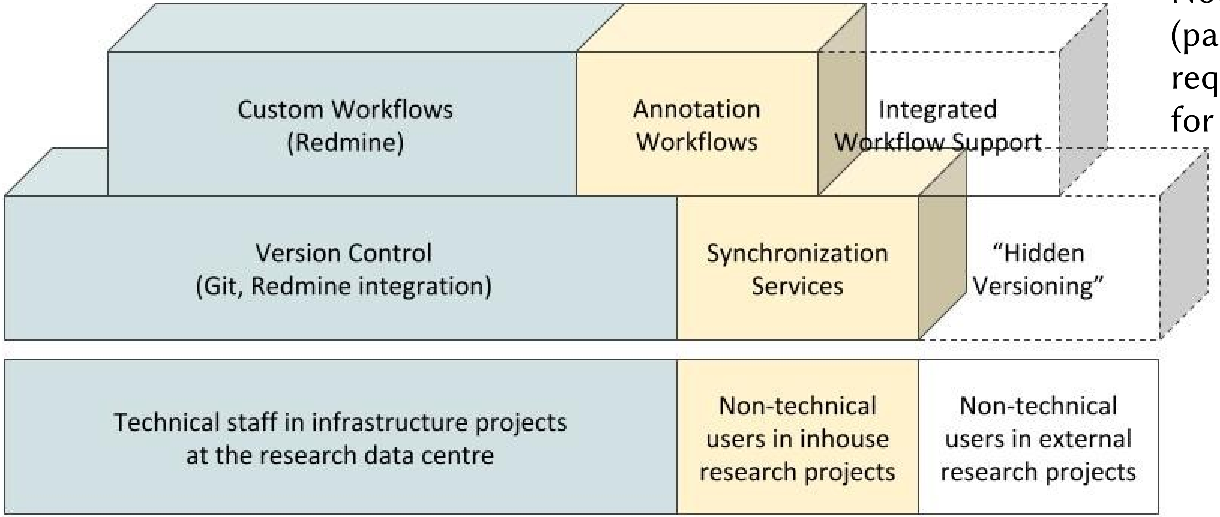
Technical staff in infrastructure projects at the research data centre

Non-technical users in inhouse research projects

Non-technical users in external research projects

Using a plugin in the project management software git to automatically create issues to be carried out

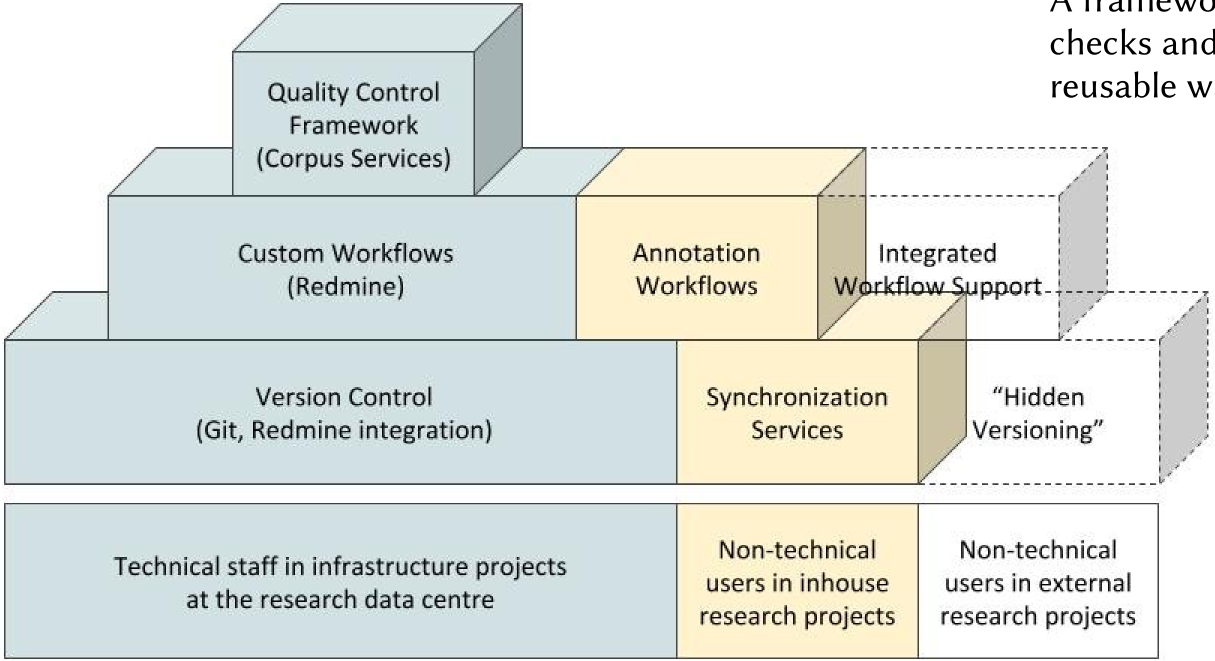
Automatically Supported Workflows III



Custom Workflows (Redmine)
Annotation Workflows
Integrated Workflow Support
Version Control (Git, Redmine integration)
Synchronization Services
“Hidden Versioning”
Technical staff in infrastructure projects at the research data centre
Non-technical users in inhouse research projects
Non-technical users in external research projects

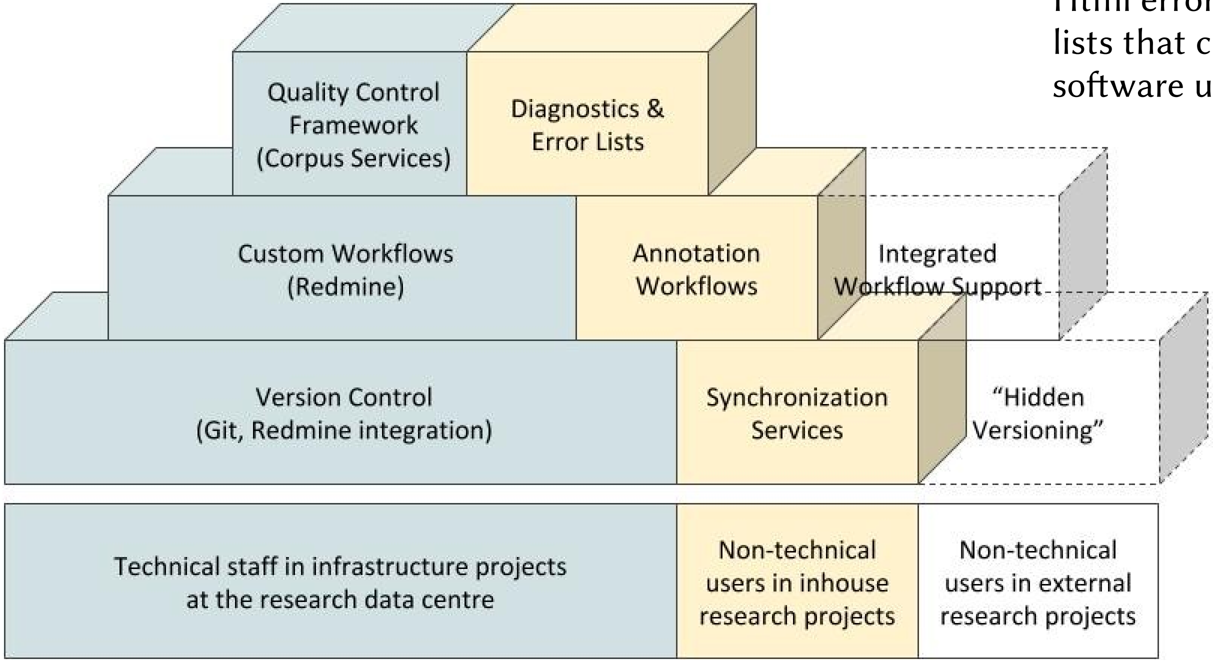
Not only create, but also (partly) carry out the required issues automatically for users in another infrastructure

Quality Control I


Quality Control Framework (Corpus Services)
Custom Workflows (Redmine)
Annotation Workflows
Integrated Workflow Support
Version Control (Git, Redmine integration)
Synchronization Services
"Hidden Versioning"
Technical staff in infrastructure projects at the research data centre
Non-technical users in inhouse research projects
Non-technical users in external research projects

A framework to gather existing checks and fixes in a consistent and reusable way

Quality Control II



Quality Control Framework (Corpus Services)

Diagnostics & Error Lists

Custom Workflows (Redmine)

Annotation Workflows

Integrated Workflow Support

Version Control (Git, Redmine integration)

Synchronization Services

“Hidden Versioning”

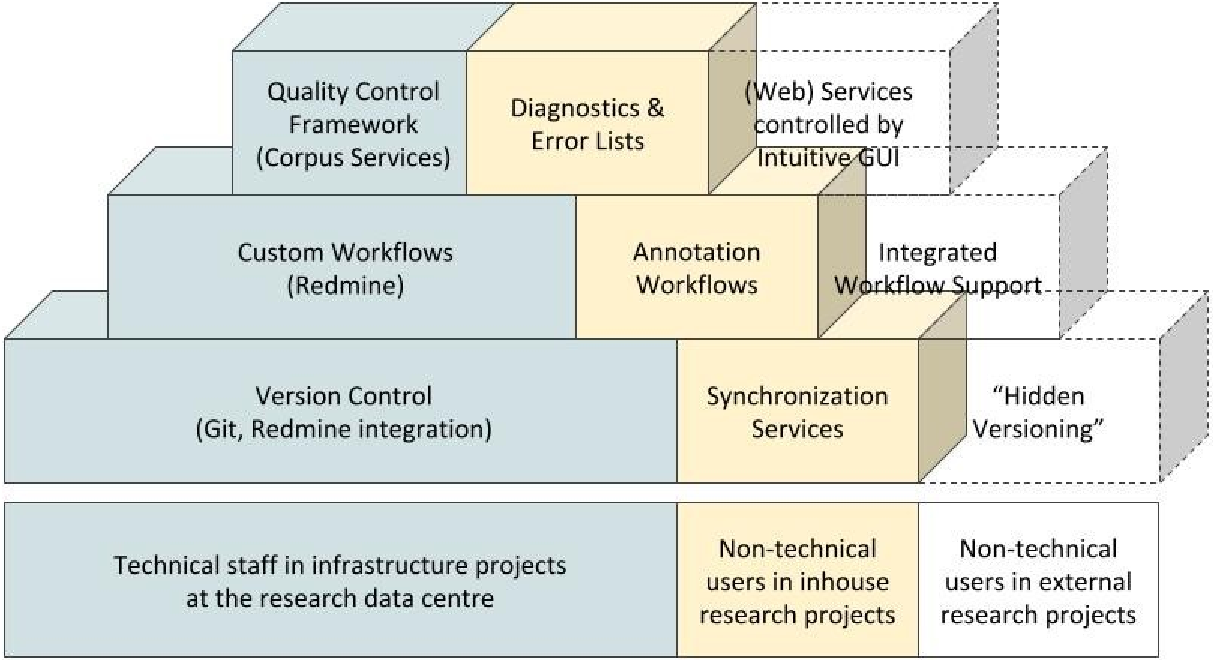
Technical staff in infrastructure projects at the research data centre

Non-technical users in inhouse research projects

Non-technical users in external research projects

Html error list along with XML error lists that can be opened in the software used to produce the data

Quality Control III



Quality Control Framework (Corpus Services)

Diagnostics & Error Lists

(Web) Services controlled by Intuitive GUI

Custom Workflows (Redmine)

Annotation Workflows

Integrated Workflow Support

Version Control (Git, Redmine integration)

Synchronization Services

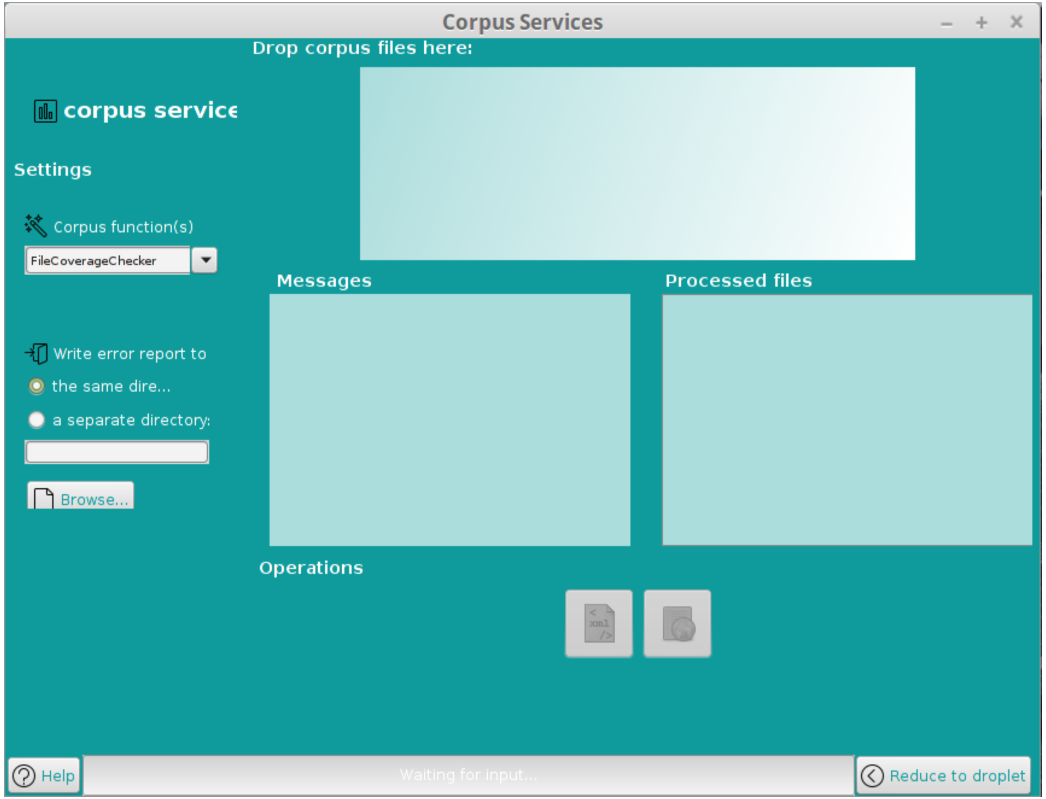
“Hidden Versioning”

Technical staff in infrastructure projects at the research data centre

Non-technical users in inhouse research projects

Non-technical users in external research projects

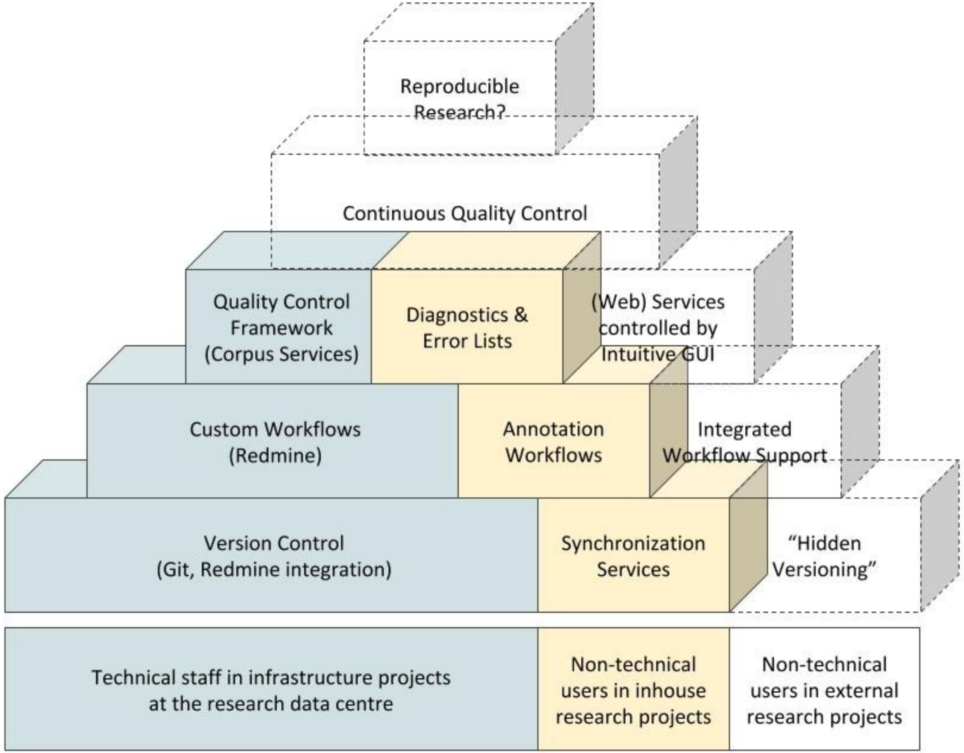
Quality Control III - Example



Corpus function(s)
FileCoverageChecker

Write error report to
● the same dire...
○ a separate directory:

Summary


Reproducible Research?
Continuous Quality Control
Quality Control Framework (Corpus Services)
Diagnostics & Error Lists
(Web) Services controlled by Intuitive GUI
Custom Workflows (Redmine)
Annotation Workflows
Integrated Workflow Support
Version Control (Git, Redmine integration)
Synchronization Services
“Hidden Versioning”
Technical staff in infrastructure projects at the research data centre
Non-technical users in inhouse research projects
Non-technical users in external research projects

Conclusions

- Technical solutions for non-technical users are needed
- Git for Humanities
- Technical support will still be needed for Humanities projects

Adaptability to other data

- Technical support will be needed for the project
- Resources should be versionable using Git

Perspective

- Enhance the hidden versioning
- Make the workflows more open to external projects/users
- Enhance the GUI options
- Adapt Framework to be even more user-friendly and robust

Acknowledgements

INEL

https://inel.corpora.uni-hamburg.de/

Project leader	Prof. Dr. Beáta Wagner-Nagy (IFUU, Universität Hamburg)
Applicants	Prof. Dr. Beáta Wagner-Nagy, Dr. Michael Rießler, Hanna Hedeland, Timm Lehmberg
Researchers/Developers	Dr. Alexandre Arkhipov (Research Coordinator), Timm Lehmberg (Technical Coordinator), Dr. Maria Brykina, Chris Lasse Däbritz, Anne Ferger, Dr. Valentin Gusev, Daniel Jettka, Dr. Svetlana Orlova

HZSK/CLARIN

https://corpora.uni-hamburg.de

Spokesperson/Project leader	Prof. Dr. Kristin Bührig
Managing director	Hanna Hedeland, M.A.
Deputy Managing Director	Daniel Jettka, M.A.
Researchers/Developers	Tommi Pirinen, PhD Hanna Hedeland, M.A.

Thank you!

Contact:
anne.ferger@uni-hamburg.de
hanna.hedeland@uni-hamburg.de
inel@uni-hamburg.de
corpora@uni-hamburg.de

Optional additional Information

Links:

- https://corpora.uni-hamburg.de
- https://inel.corpora.uni-hamburg.de
- https://exmaralda.org/en/