How GermaParl Evolves

Improving Data Quality by Reproducible Corpus Preparation and User Involvement

Andreas Blätte, Julia Rakers and Christoph Leonhardt

ParlaCLARIN III Workshop, Marseille, 2022-06-20

KonsortSWD
Consortium for the Social, Behavioural, Educational and Economic Sciences

Text+
Collections - Lexical Resources - Editions
Infrastructure / Operations

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

The Need for a Reproducible Workflow for Corpus Creation

- large parliamentary text corpora are widely used
- being FAIR (Wilkinson et al. 2016) is increasingly intended
- data quality is crucial
- but not all flaws can be ruled out in large data collections

Requirement for large high-quality data: A process for evolving data quality with a reproducible data preparation workflow and community feedback as central building blocks

Outline

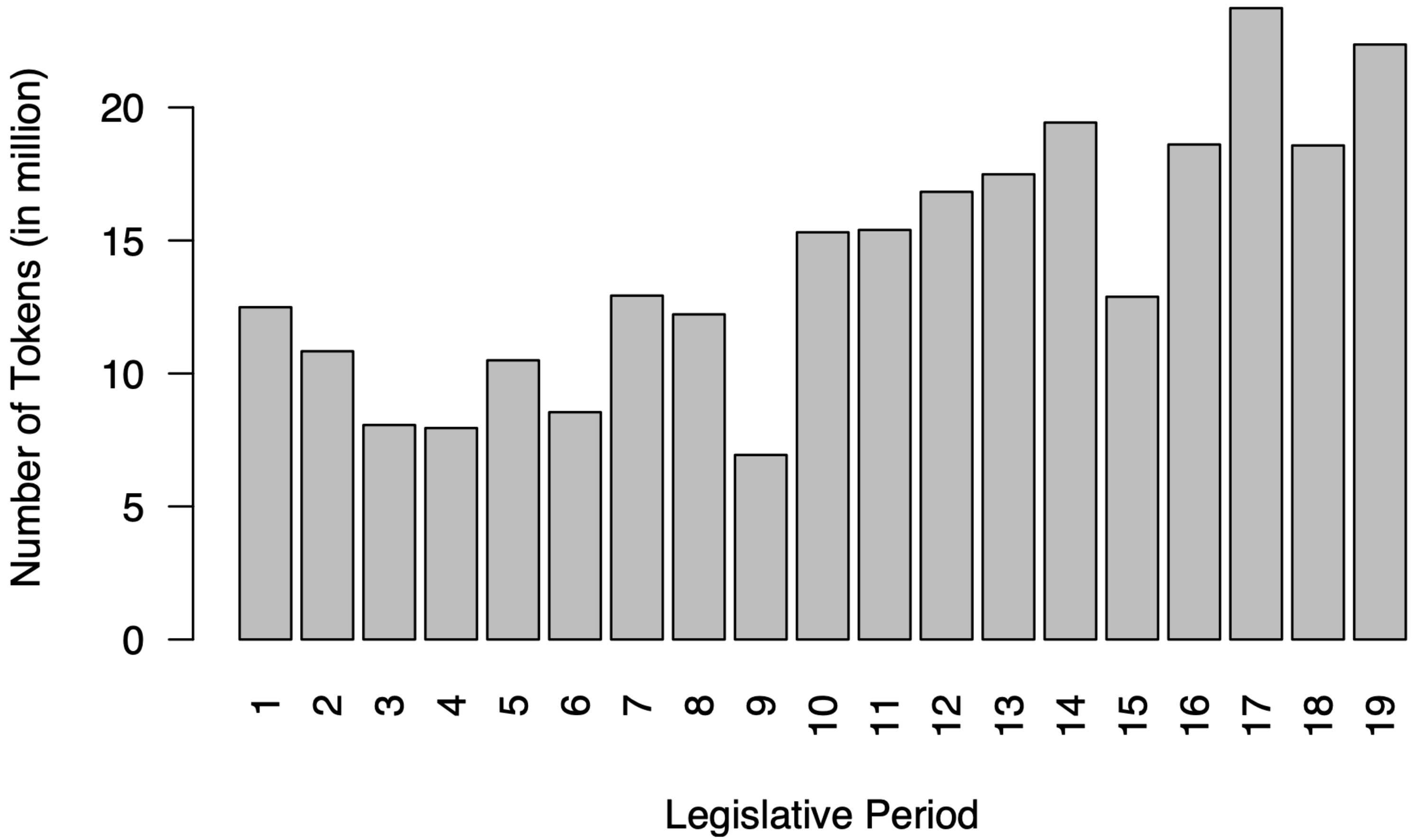
1. The GermaParl Corpus of Parliamentary Debates
2. A Reproducible Corpus Preparation Pipeline
3. Data Quality by Reproducibility and User Involvement

GermaParl: An Overview

- corpus of parliamentary debates in the German Bundestag
- released version: 1996 – 2016 (Blätte and Blessing 2018)
- **GermaParl v2: 1949 – 2021**
- Project Contexts
  - PolMine project (see also R package polmineR, Blätte 2020)
  - KonsortSWD and Text+
- not the only corpus of parliamentary debates in the German Bundestag
  - ParlSpeech (Rauh and Schwalbach 2020), DeuParl (Kirschner et al. 2021) or Open Discourse (Richter et al. 2020) (among others) as meaningful contributions
- but: GermaParl as a comprehensive, universally applicable and evolving resource

GermaParl v2: Comprehensive in Volume

- 19 legislative periods
- 72 years (1949 - 2021)
- 271 million tokens



Source: GermaParl

GermaParl v2: Comprehensively Annotated

- Structurally and linguistically annotated
- Structural Annotation (“structural attributes”):
  - mostly metadata
  - nested structures like named entities

Structural Attribute	Description
lp	legislative period
protocol_no	session number
date	date
year	year
speaker	speaker name
parliamentary group	parliamentary group of speaker
party	party of speaker
role	role of speaker
…	

Source: GermaParl, see Blätte et al. 2022, Table 2

KonsortSWD

Collections - Lexical Resources - Editions
Text+
Infrastructure / Operations

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

GermaParl v2: Comprehensively Annotated (cont.)

- Linguistic Annotation ("positional attributes"):
  - Word
  - POS-Tags
    - Universal Dependencies tag set
    - language specific STTS
  - Lemmata

cpos	word	upos	xpos	lemma
0	Meine	PRON	PPOSAT	Mein
1	Damen	NOUN	NN	Dame
2	und	CCONJ	KON	und
3	Herren	NOUN	NN	Herr
4	!	PUNCT	$.	!
		…		

Source: GermaParl, see Blätte et al. 2022, Table 3

A Reproducible Corpus Preparation Pipeline

- Size of data as a **challenge**
- **Reproducibility** an important aspect
  = the technical basis for a **feedback loop for quality control** during creation as well as after the initial release of a corpus

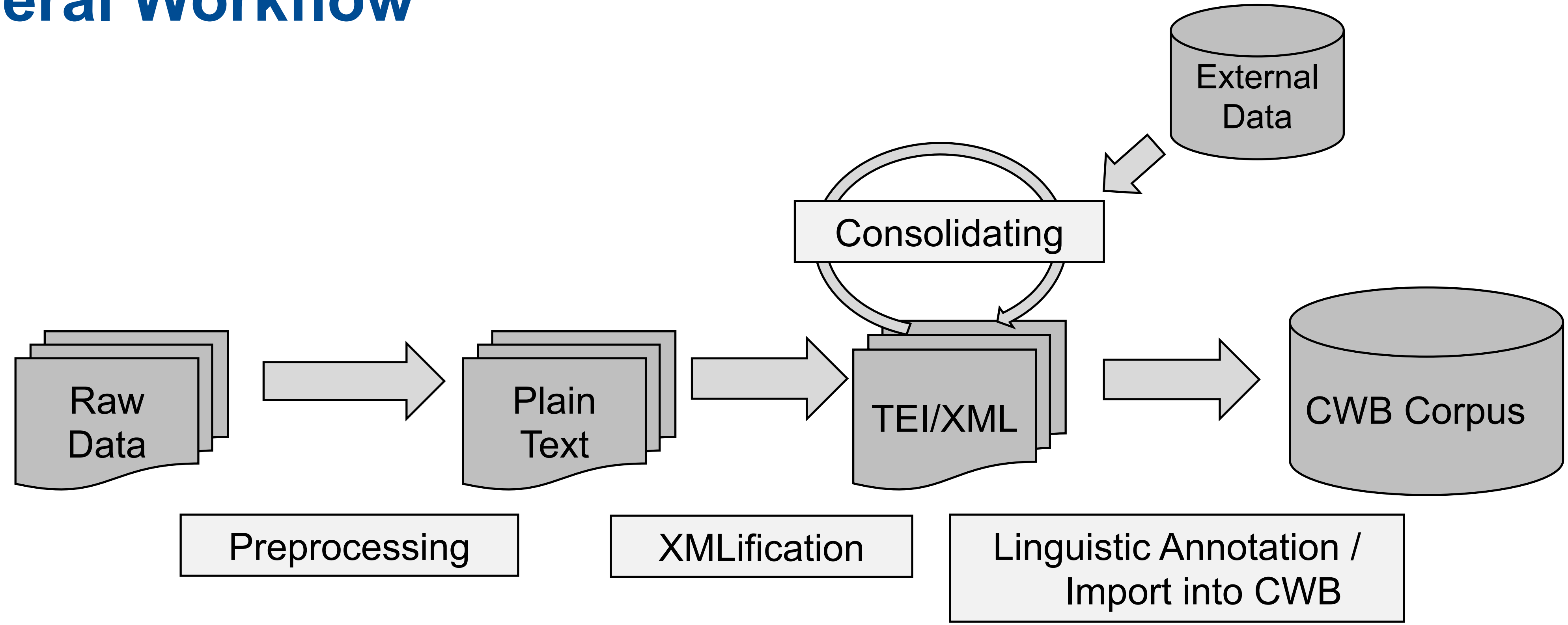
  ➢ Evolving Data Quality

General Workflow

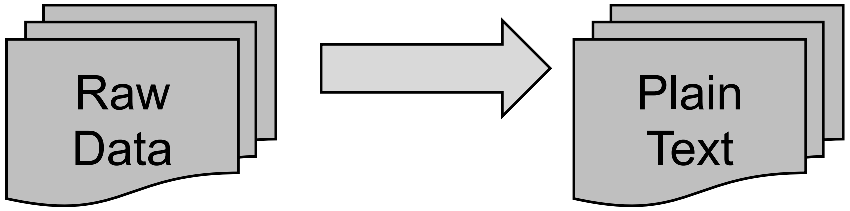
- **Goal:** An iterative workflow for corpus creation
- Successfully used in other corpus preparation processes, especially GermaParl (Blätte and Blessing 2018)
- Flexible, portable, local

- **Output formats:**
  - TEI-XML (sustainable, interoperable)
    - perspectively: ParlaMint format (Erjavec et al. 2022)
  - CWB Corpus (powerful corpus management and query tool, Evert and Hardie 2011)

General Workflow



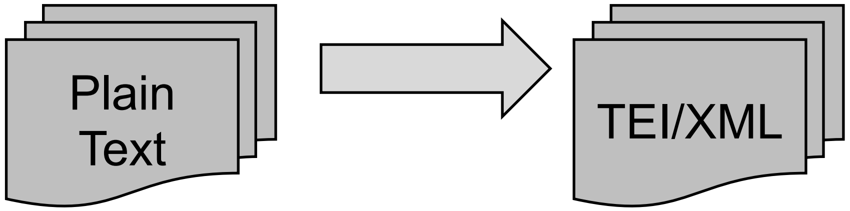
Preprocessing



- **Input / Raw Data:**
  - new parliamentary protocols retrieved from https://www.bundestag.de/services/opendata
    - (mostly unstructured) XML
    - PDF
- **Notes:**
  - existing GermaParl TEI added in consolidation step
  - Special Case 19th Legislative Period

- **Output:** Clean Plain Text
- **Steps:**
  - Removal of header and footer lines, removal of table of contents and appendices...

XMLification

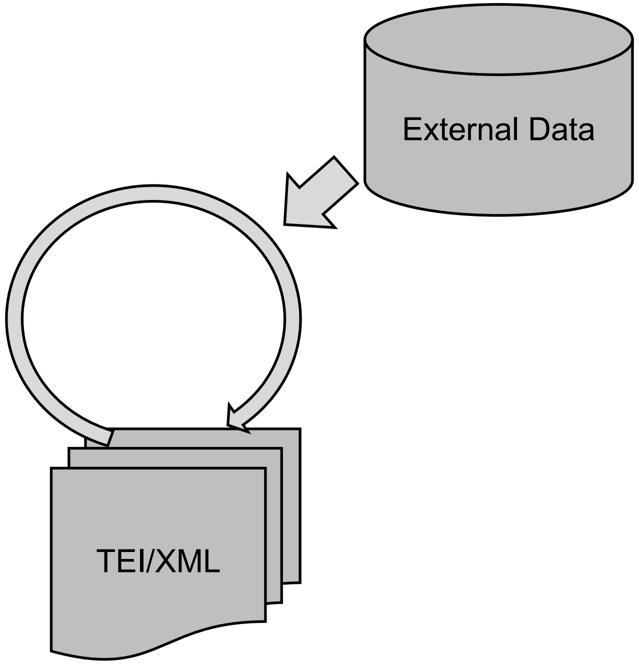


- Process to **reconstruct debate structure from unstructured text**
- Essential idea: a battery of regular expressions for speakers, interjections, etc.
- See Blätte and Blessing 2018

- Using the **Framework for Parsing Plenary Protocols** (frappp)
  - generic tool set, integrating standardized steps
  - further information and code examples: https://polmine.github.io/frappp_slides

Consolidating

- Initial correction of missing speakers and other flaws in the data
- Enrichment of speaker attributes not included in the protocols via external data sources (Wikipedia and the *Stammdaten* file of the German Bundestag (Deutscher Bundestag 2021))



KonsortSWD Text+ Universität Duisburg Essen Open-Minded

Consolidating

- *Wikipedia* for speakers' **party affiliations** and **full speaker names of speakers which are not members of parliament**
- *Stammdaten* file of the German Bundestag for **full speaker names of members of parliament**

```
<sp who="Adenauer" parliamentary_group="CDU/CSU" role="mp"
position="NA" party="CDU" name="Konrad Adenauer">
   <speaker>Dr. Adenauer (CDU/CSU) :</speaker>
   …
</sp>
```

(04/002.xml)

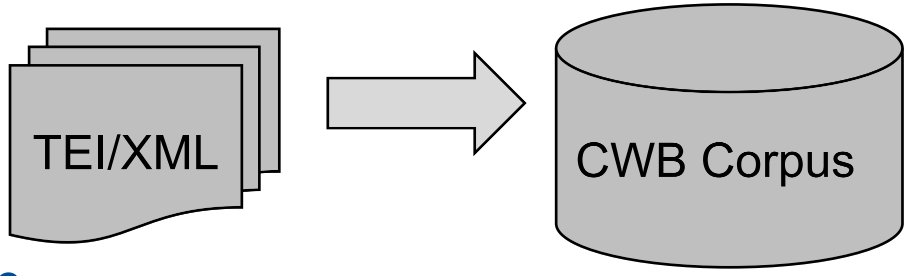
TEI-XML Output

- Available on **GitHub** for released data

(https://github.com/PolMine/GermaParlTEI/blob/master/13/BT_13_087.xml)


```xml
<?xml version="1.0" encoding="UTF-8"?>
<TEI>
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Plenarprotokoll</title>
        <legislativePeriod>13</legislativePeriod>
        <sessionNo>87</sessionNo>
      </titleStmt>
      <editionStmt>
        <edition>
          <package>ctk.plpr.bt.txt</package>
          <version>0.1.1</version>
          <birthday>2016-01-06</birthday>
        </edition>
      </editionStmt>
      <publicationStmt>
        <publisher>Deutscher Bundestag</publisher>
        <date when="">1996-02-09</date>
        <page/>
      </publicationStmt>
      <sourceDesc>
        <filetype>txt</filetype>
        <url>http://webarchiv.bundestag.de/archive/2005/1205/bic/plenarprotokolle/pp/1996/13087a.zip</url>
        <date>2015-10-20 19:19:06</date>
      </sourceDesc>
    </fileDesc>
    <encodingDesc>
      <projectDesc>PolMine-Project (www.polmine.de)</projectDesc>
      <samplingDecl/>
      <editorialDecl/>
    </encodingDesc>
    <profileDesc/>
    <revisionDesc/>
  </teiHeader>
  <text>
    <body>
      <div type="agenda_item" n="1" what="NA" desc="NA">
        <sp who="Dr. Rita Süssmuth" parliamentary_group="NA" role="presidency" position="Präsidentin" party="CDU" name="Rita Süssmuth">
          <speaker>Präsidentin Dr. Rita Süssmuth:</speaker>
          <p>Liebe Kolleginnen und Kollegen! Die Sitzung ist eröffnet.</p>
```

Linguistic Annotation and Import into the Corpus Workbench



- **Linguistic Annotation**
  - Segmentation into Tokens and Sentences (Stanford CoreNLP, Manning et al. 2014)
  - POS-Tagging (Universal Dependencies) (Stanford CoreNLP)
  - Named Entity Recognition (Stanford CoreNLP)
  - POS-Tagging (STTS) (TreeTagger, Schmid 1995)
  - Lemmata (TreeTagger)
- Used from within R (e.g., bignlp R package)

Linguistic Annotation and Import into the Corpus Workbench

- **Import into the Corpus Workbench (CWB)**
  - CWB as a fast corpus management and querying tool (https://cwb.sourceforge.io/)
  - Import via the cwbtools R package (on CRAN)

- **Dissemination** of linguistically annotated, CWB-indexed corpora
  - GermaParl stored on Zenodo
  - Beta Version of GermaParl v2 (1949-2021) currently as restricted access on Zenodo

Data Quality by Reproducibility and User Involvement

- **User Feedback** important to detect and remedy remaining flaws and bugs in the data
- **User Involvement** via
  - GitHub Issues
  - User Workshops
- **Reproducibility**
  - Feedback most useful when it can be incorporated into the data efficiently
  - Entire process fully reproducible
  - Transparency through Versioning/DOIs via Zenodo (for data) and GitHub

Release Plan of GermaParl v2.0

- Restricted Beta Release for interested persons (request access via Zenodo) in May 2022
- Issue-Only Repository on GitHub accompanies the release for feedback during beta phase
- User Workshop collects feedback on bugs, but also usability, convenience and further feature requests
- Full, public release in October 2022

https://zenodo.org/record/6539967

Conclusion

- **Contribution**
  - GermaParl as a large, comprehensively annotated resource
  - Thoroughly checked but flaws will remain, given its size
  - Reproducible, replicable data preparation as a precondition for high-quality data, especially in large datasets and corpora
  - Workflow facilitates this for GermaParl, allowing efficient user feedback
- **Next Steps**
  - Increasing interoperability by providing XML as ParlaMint corpus (Erjavec et al. 2022)
  - GermaParl as a basis for linkage to other types of data (“Linking Textual Data” within KonsortSWD)

References

Blätte, A. and Blessing, A. (2018). The GermaParl Corpus of Parliamentary Protocols. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 810–816, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Blätte, A. (2020). polmineR: Verbs and Nouns for Corpus Analysis. https://doi.org/10.5281/zenodo.4042093.

Deutscher Bundestag. (2021). Stammdaten aller Abgeordneten seit 1949 im XML-Format. https://www.bundestag.de/resource/blob/472878/d5743e6ffabe14af60d0c9ddd9a3a516/MdB-Stammdaten-data.zip.

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrímsson, S., Çöltekin, Ç., de Does, J., Depuydt, K., Agnoloni, T., Venturi, G., Pérez, M. C., de Macedo, L. D., Navarretta, C., Luxardo, G., Coole, M., Rayson, P., Morkevičius, V., Krilavičius, T., Darģis, R., Ring, O., van Heusden, R., Marx, M., and Fišer, D. (2022). The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*.

Evert, S. and Hardie, A. (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 conference*, University of Birmingham, UK.

References

Kirschner, C., Walter, T., Eger, S., Glavas, G., Lauscher, A., and Ponzetto, S. P. (2021). DeuParl. https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2889.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland, June. Association for Computational Linguistics.

Rauh, C. and Schwalbach, J. (2020). The ParlSpeech V2 data set: Fulltext corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies. Harvard Dataverse, V1. https://doi.org/10.7910/ DVN/L4OAKN.

Richter, F., Koch, P., Franke, O., Kraus, J., Kuruc, F., Thiem, A., Högerl, J., Heine, S., and Schöps, K. (2020). Open Discourse. Harvard Dataverse, V3. https://doi.org/10.7910/DVN/FIKIBO.

Schmid, H. (1995). Improvements in Part-of-Speech Tagging With an Application To German. Revised version of a paper originally presented at the EACL SIGDAT workshop in Dublin in 1995. https://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/data/tree-tagger2.pdf.

References

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., ’t Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018.

Text+ Logo: https://www.nfdi.de/wp-content/uploads/2021/12/Textplus-Logo-en.png

KonsortSWD Logo: https://www.konsortswd.de/wp-content/uploads/Logo-KonsortSWD_en.svg
https://www.konsortswd.de/wp-content/uploads/Logo-KonsortSWD_kurz.svg

Thank you for your attention!

Christoph Leonhardt | christoph.leonhardt@uni-due.de

KonsortSWD
Consortium for the Social, Behavioural, Educational and Economic Sciences

Text+
Collections - Lexical Resources - Editions
Infrastructure / Operations

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

ParlaCLARIN III Workshop, Marseille, 2022-06-20