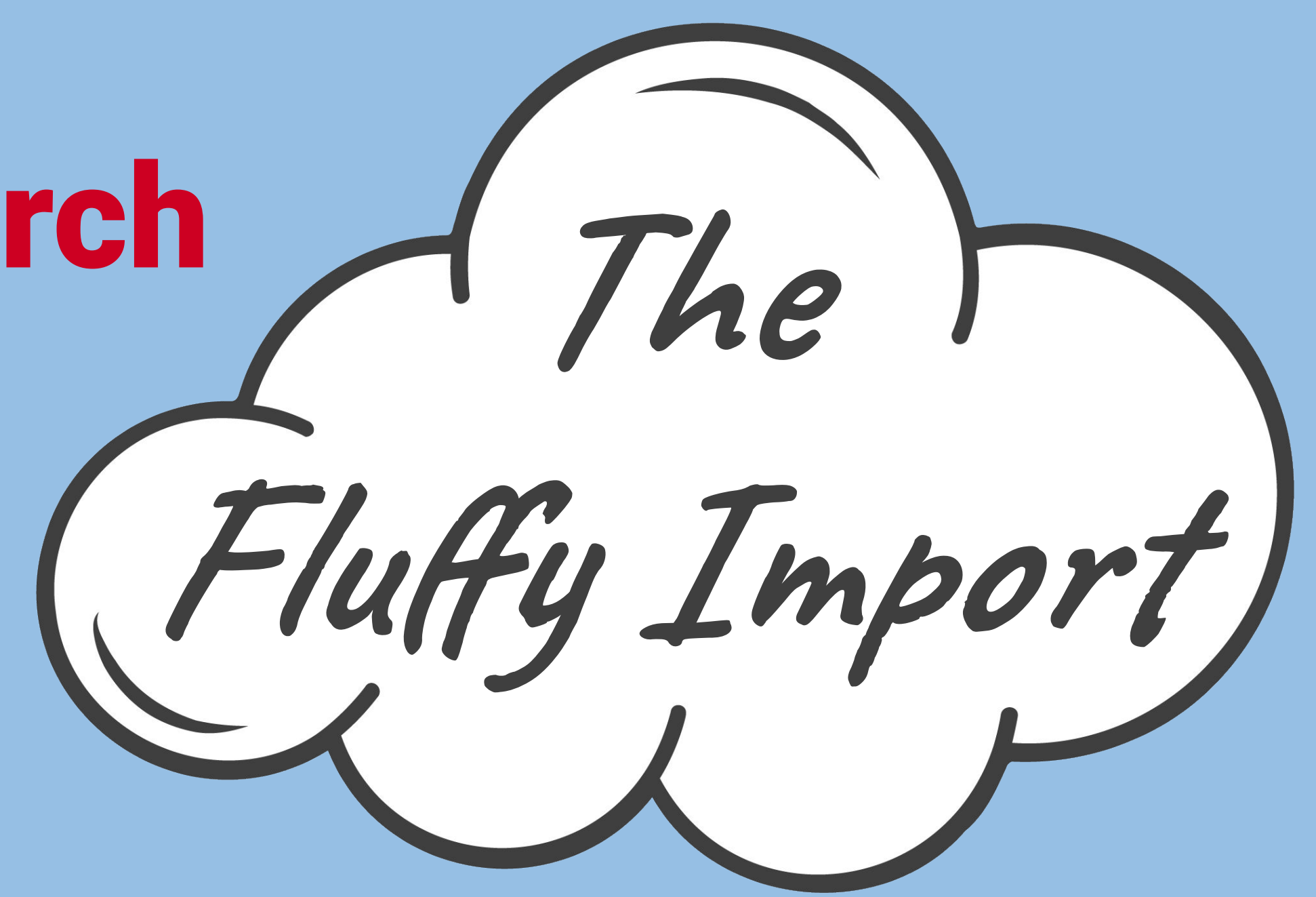




Preserving Humanities Research

Data: Data Depositing in the TextGrid Repository



If "**Preservation is use**" (John Cotton Dana), research data infrastructures have to emphasize the potentials and re-usability of data. **How to get there in three steps?**



Step 1 Living Infrastructure

TGRep is a pioneer of the Digital Humanities across the German-speaking area. Today, TGRep is part of the Text+ portfolio, the NFDI consortium for language and text-based research data in Germany.



Step 2 Project Sensitive Presentation

ELTeC in TGRep, the European Literary Text Collection, is a good example for the identification, consultation, ingest, transformation, enrichment, publication and integration in the portfolio of Text+, **spelling out re-usability and interoperability**. ELTeC is a state-of-the-art, open access multilingual collection of corpora. More than 2000 full-text novels in XML-TEI in 21 languages, distributed via multiple platforms (such as GitHub and Zenodo). 1365 full-texts in 15 languages are also published in the TGRep.

Step 3 The Data Depositing Workflow in TGRep

The solution implemented in Text+ is a **workflow that automates creating the technical files required** for repository import and assists with missing metadata.

Users interact with a **web-based interface in a Jupyter notebook**, specifying the location of TEI files to be imported. **The tool analyzes the files, finds common XPath with the required metadata**, and in a subsequent manual step, users can accept or modify the suggested XPath or specify fixed values for missing metadata. Users then generate and upload these metadata files to a beta area in the TGRep, available only to project researchers. After final evaluation, the files can be permanently published.

This **new workflow improves data import and serves as a blueprint for easy-to-build applications combining libraries and notebooks**, relying on the versatile Jupyterlab environment, deployable both locally and in the cloud.

Portalconfig

Researchers can now personalize various aspects of their TGRep projects presentation using two documents:

Portalconfig: an XML file containing a brief description, the project logo, an XSLT document for project specific transformation into HTML, and the metadata categories to be displayed as facets.

Readme: a Markdown file where projects can describe their project and link to publications and resources.

How does it look? → <https://textgridrep.org/projects>

TG-model and TG-clients

The Jupyter import notebook uses the TG-model and TG-clients (both library & command line interface) for **extracting and generating TextGrid metadata files from TEI data files** and managing the import and update of data files, as well as TextGrid project management.



Would you like to publish a collection of language- and text-based research data in the TextGrid Repository? **Get in touch with us: anfragen@textgrid.de** We are happy to advise and assist you with the ingest process. We offer to organise hands-on workshops and code sprints. Together, we learn about and test services, ingest your data, and last but not least have fun together.