# Use Case Analysis

## PID Adoption in the Text+ Consortium using the Example of the SUB Göttingen

Jana Böhm

20 December 2024

# Imprint

| | |
|---|---|
| **Authored by** | Jana Böhm (https://orcid.org/0009-0004-9802-113X) |
| **Contributions by** | Stefan Buddenbohm (https://orcid.org/0000-0002-3469-6101)<br>Stefan Funk (https://orcid.org/0000-0003-1259-2288) |
| **Reviewed by** | Stephanie Hagemann-Wilholdt (https://orcid.org/0000-0002-0474-2410)<br>Torsten Kahlert (https://orcid.org/0009-0003-3264-5006)<br>Marc Lange (https://orcid.org/0000-0002-7742-3867)<br>Jana Tatscheck (https://orcid.org/0009-0002-1285-2027)<br>Lukas Weimer (https://orcid.org/0000-0001-6919-3646) |
| **Date of publication** | 20 December 2024 |
| **DOI** | https://doi.org/10.5281/zenodo.14327691 |
| **License** | This work is licensed under https://creativecommons.org/licenses/by/4.0/ |
| **Deliverable** | Part of D1.2 and D2.1 |

**About**

PID4NFDI (https://base4nfdi.de/projects/pid4nfdi) is the basic service for persistent identifiers in development for the German National Research Data Infrastructure (NFDI). PID4NFDI is part of and funded through Base4NFDI.

Funded by DFG as part of NFDI. DFG Grant Number: 521466146

base4nfdi Basic Services for NFDI

Funded by
DFG Deutsche Forschungsgemeinschaft
German Research Foundation

# Contents

**Background** **1**

Criteria for Use Case Selection 1

Knowledge Base for Use Case Analysis 2

**PID Adoption in the Text+ Consortium using the Example of the SUB** **3**

Context and Overview 3

Resource Types 4

Infrastructures 5

The TextGrid Repository 6

Publication Workflows 7

(Persistent) Identifier Types 8

Metadata 8

The DARIAH-DE Repository 9

Publication Workflow 9

PID Types 9

Metadata 10

Schema 10

Metadata Quality Report 10

Discoverability 11

PID Provider Considerations 11

Metadata 11

Technical Interoperability 12

Training and Support 12

For End-Users 12

For Staff 13

PID Policy & Governance 13

Outreach 14

**Summary** **15**

**References** **16**

# Background

The initial phase of the PID4NFDI project is dedicated to establishing a comprehensive understanding of the persistent identifier (PID) landscape within the German National Research Data Infrastructure (NFDI). This foundational phase encompasses use case analysis, requirements engineering, and concept development. These activities are not only designed to address immediate project requirements but also to ensure that the insights gained will facilitate seamless integration and ongoing development in subsequent project phases.

As part of our deliverables for the initialization phase work packages 1 and 2, this use case analysis serves several key purposes:

- **Work Package 1 (WP1):** This encompasses **D1.1**, which focuses on exploring the landscape of PID practices across different NFDI services, and **D1.2**, which involves a requirements analysis of selected use cases to derive practical and relevant insights. These analyses set the stage for a broader understanding of how PIDs are implemented and managed across diverse research contexts within NFDI.
- **Work Package 2 (WP2):** Within WP2, deliverable **D2.1** aims to develop a conceptual framework for mapping selected use cases to existing PID services. The goal here is to ensure that our understanding of use cases translates into actionable strategies for integrating PID services effectively across NFDI consortia.

## Criteria for Use Case Selection

The selection of use cases was a crucial step in ensuring a diverse and representative analysis. We established the following criteria for choosing which use cases to examine:

1. **Diversity in duration of operation and PID service providers:** We aimed to reflect a range of maturities, encompassing both well-established and newer initiatives. This variety helps us understand PID usage at different stages of project development and their engagement with different stakeholders (i.e. PID providers).
2. **Disciplinary breadth:** The use cases needed to span multiple disciplines to assess the adaptability and versatility of PID adoption across different scientific fields and sectors.

3. **Active engagement:** The use cases had to demonstrate active contributions and engagement.

## Knowledge Base for Use Case Analysis

This use case analysis is primarily the result of a guideline-based interview of Göttingen State and University Library (Staats- und Universitätsbibliothek Göttingen – SUB)[1] representatives that was conducted in November 2024 and an extensive survey [1] designed by the PID4NFDI project team, which was distributed in March 2024 and completed by a Text+ representative with responsibilities related to PID management. Further details were added to the text based on comprehensive online documentations of the services.

The survey yielded valuable insights into the experiences, challenges, and potential areas for improvement related to PID management within the Text+ consortium. This allows for a comparison of the responses from Text+ with those from other NFDI consortia, thus enabling an assessment of whether experiences and challenges are commonly shared between consortia or if they are specific to particular use cases.

The interview provided the opportunity to gain a deeper understanding of how infrastructure components work together within the consortium and which role PIDs take on in this framework. Additionally, it shed light on different approaches for metadata granularity, quality and completeness for different repository solutions, workflows for PID registration, and usage of PID types for a broad range of resources.

[1] Webpage: https://www.sub.uni-goettingen.de/en/news/

# PID Adoption in the Text+ Consortium using the Example of the SUB

## Context and Overview

The Text+[2] consortium is one of four NFDI consortia from the humanities, including also NFDI4Culture[3], NFDI4Memory[4], and NFDI4Objects[5]. Within the humanities sector, Text+ is the consortium for text- and language-based research data and therefore addresses linguistics, literary studies and philology in the broadest sense.

Text+ relies on a distributed infrastructure that consists of so-called Text+ centers[6]. Each center might either be a data center (providing technical infrastructures for the collection, storage, management or provision of research data), a competence center (taking on tasks in research data management, e.g. counseling, training, networking), or a combination of both. Data centers provide their own data, but also accept data from third parties. The data stored in these infrastructures together form the Text+ data space.

The SUB together with the Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen (GWDG)[7] act as competence and data centers in the context of Text+, providing several repository solutions such as the DARIAH-DE Repository[8] and the TextGrid Repository[9]. This use case analysis will focus on PID practices within the SUB as a center in the Text+ consortium, and more specifically, on the previously named repository solutions. The aim is to describe the usage of PIDs, identify challenges, gaps, and opportunities for improvement.

The SUB was selected as a use case partner based on the following reasons:

1. **Broad perspective:** This use case gives the opportunity to inspect the PID adoption in several services of one partner (SUB) within a large NFDI consortium. By giving an overview of the most important services related to PID management, we can take on a comprehensive perspective.

[2] Webpage: https://text-plus.org/.
[3] NFDI consortium for material and immaterial cultural heritage, https://nfdi4culture.de/index.html.
[4] NFDI consortium for historical data, https://4memory.de/.
[5] NFDI consortium for the material remains of human history, https://www.nfdi4objects.net/.
[6] Overview of Text+ centers: https://text-plus.org/en/ueber-uns/textpluszentren/.
[7] Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen, https://gwdg.de/.
[8] Webpage: https://repository.de.dariah.eu/.
[9] Webpage: https://textgridrep.org/.

2. **Mature and well-integrated infrastructures:** The Text+ repositories provided by the SUB are embedded into DARIAH-DE[10], which is a mature research data infrastructure for the humanities and cultural studies. This infrastructure acts as the German partner of DARIAH[11], a research infrastructure in the European context. Generally, Text+ integrates DARIAH-DE, CLARIN-D[12] and CLARIAH-DE[13]. Both the DARIAH-DE Repository and the TextGrid Repository have been operated by the SUB since 2017 and 2011, respectively. Hence they act as mature infrastructures, also even before they have been integrated into Text+. Both repositories integrate ePIC[14] PIDs. Additionally, DOIs[15] (provided by DataCite[16]) are used in the DARIAH-DE Repository.

3. **Commitment:** Representatives of the SUB/Text+[17] have taken part in the initialisation phase of PID4NFDI for this use case, providing valuable information in a guideline-based interview. Further information was provided via the survey by another consortium representative.

A variety of stakeholder groups are involved in Text+: researchers (end-users), infrastructure, service and repository operators and developers, and stakeholders on NFDI level, for example, the Knowledge Graph Working Group from NFDI Section (Meta)Data, Terminologies, Provenance[18]. In the future, the large group of Text+ data centers with their specific repositories will be a target group for a fully ramped-up PID4NFDI basic service.

## Resource Types

PIDs play an important role for the SUB/Text+, involving the following resource types:

[10] DARIAH-DE (Digital Research Infrastructure for the Arts and Humanities) is a digital research infrastructure for the humanities and cultural studies in Germany, see https://de.dariah.eu/.

[11] Webpage: https://www.dariah.eu/. DARIAH has 23 European partners and 19 cooperating partners in 10 Non-member countries, see https://www.dariah.eu/network/members-and-partners/.

[12] CLARIN-D (Common Language Resources and Technology Infrastructure) is a digital research infrastructure for language resources in the humanities and social sciences, https://www.clarin-d.net/de/ https://www.clarin-d.net/de/ueber/partner-dariah-de..

[13] CLARIAH-DE is the merger of the two research infrastructure networks CLARIN-D and DARIAH-DE. CLARIN-D and DARIAH-DE closely work together to complement their offers for users, https://dig-hum.de/forschung/projekt/clariah-de.

[14] European Persistent Identifier consortium, https://www.pidconsortium.net/.

[15] Digital Object Identifiers, https://www.doi.org/.

[16] Webpage: https://datacite.org/.

[17] Stefan Buddenbohm, https://orcid.org/0000-0002-3469-6101, and Stefan Funk, https://orcid.org/0000-0003-1259-2288.

[18] The section facilitates the collaborative development and implementation of common data and metadata standards by the NFDI consortia, see https://www.nfdi.de/section-meta/?lang=en.

- **Editions, lexical resources, collections and objects within collections:** PIDs are vital for referencing not only editions and lexical resources, but also more general research data. For research data in Text+, PIDs are applied at the collection level, where a collection refers to a set of objects that belong together in some way, and at the object level, whereby standardised object types (e.g., XML files, images in selected formats, WAV files) are used.

Furthermore, PIDs are or were under discussion for the following entities:

- **Layers in digital editions** (see also [2]): Digital editions typically consist of individual layers. As an example, a digital edition of the Theodor Fontane Notebooks[19] may consist of scans of the notebooks as bottom layer, topped with an XML/TEI transcript layer and another layer with annotations. PIDs could be used to reference the different layers.

- **Search strings and intermediate statuses** (see also [2]): Researchers may want to reference intermediate statuses within complex software environments, for example statuses during the editing process. In this case, a reference would need to point to a fixed state of the object, for example a query term or a search string which might refer to a database state (if the database remains stable).

- **Services:** Service descriptions are metadata describing a service and are increasingly often referenced in practice. As an example, reference [3] is an URL pointing to a service description of the TextGrid Repository within the SSH Open Marketplace[20]. Currently, the service would be cited via the given URL, which is not persistent by definition, but expected not to change frequently. However, a PID that resolves to the given URL as a landing page would be reasonable. PID4NFDI could evaluate if this would be a (small) use case or whether this concept is already applied by other consortia.

## Infrastructures

In the context of PIDs, repositories and search and retrieval services (such as the Text+ Registry[21] and the Federated Content Search[22]) are particularly relevant in Text+.

[19] https://fontane-nb.dariah.eu/index.html.

[20] The SSH (Social Sciences and Humanities) Open Marketplace is a discovery portal that aggregates and contextualises resources for social sciences and humanities, https://marketplace.sshopencloud.eu/.

[21] The Text+ registry is available at https://registry.text-plus.org/.

[22] The Federated Content Search is available at https://www.clarin.eu/content/content-search.

- **Repositories:** PID integration is primarily of interest in the context of research data publication, for example when integrated into a data deposit workflow. Some repositories with integrated PID assignment are already operational, and more are planned for the future of Text+. Many partner institutions run their own repositories or plan to do so. This is due, among other reasons, to differences in expertise, or different collection/research focuses. In repositories like TextGrid and DARIAH-DE, the repository acts as the PID-issuing instance and the users initiate the allocation. Both repositories are based on the same technological framework, but, depending on the data format and the repository, offer different data presentation and curation models, as well as interfaces to enable reuse (see also the joint documentation of both repositories [4]). Future developments are largely focused on the TextGrid Repository. Text+ aims to achieve certification for its repositories, for example the CoreTrustSeal[23] or nestor Seal[24], not only to be seen as a trustworthy provider, but also to comply with current standards, such as in the area of PIDs. Both repositories are already CoreTrustSeal certified[25], although the certification for the TextGrid Repository was easier to achieve than for the DARIAH-DE Repository, because the former provides more comprehensive metadata of higher quality.
- **Search and Retrieval:** The Text+ Registry is a central, cross-domain research and information tool that enables searching resources such as lexical resources, collections, editions, repositories and services of the various partners within Text+. Ideally, the resources have PIDs and enriched metadata descriptions before they enter the system. The registry is currently being filled with data.

## The TextGrid Repository

The TextGrid Repository is a long-term archive and repository for text-based research data and digital editions from projects based on collections. In general, data in the TextGrid Repository are carefully and comprehensively catalogued. The PID assignment is integrated into the modularized upload-workflows. Detailed

[23] The CoreTrustSeal (https://www.coretrustseal.org/) offers for any trustworthy research data repository a certification according to the CoreTrustSeal requirements, see [5].

[24] The nestor Seal (https://www.langzeitarchivierung.de/Webs/nestor/EN/Zertifizierung/nestor_Siegel/nestor_siegel_node.html) offers any trustworthy digital archive certification based on the DIN 31644 standard.

[25] The TextGrid and DARIAH-DE Repository requirements for the CoreTrustSeal can be found at https://doi.org/10.34894/L1F7BS respectively https://doi.org/10.34894/SGNHLL.

descriptions below are based on the interview, additionally enriched with information from the repository's documentation [6].

### Publication Workflows

The challenge of publishing data in the TextGrid Repository is the provision of comprehensive TextGrid metadata. There are several workflows that can be used to create this metadata and then publish the data in the TextGrid Repository.

- The TextGrid Repository is optimised for XML/TEI formats and the editorial work and publications from the TextGrid Laboratory[26]. Researchers can jointly edit, label and enrich XML/TEI files as TextGrid objects in the TextGrid Laboratory. Finally, the data is published using the TextGrid-publish GUI. An ePIC PID is assigned to the files at this stage.
- It is also possible to publish data directly in the TextGrid Repository without using the TextGrid Laboratory. Data can be imported by copying files and folders manually using the TextGrid Import Tool[27], plus the required TextGrid metadata files must be created manually by the user.
- A new user-friendly import workflow, the Fluffy Import[28], has recently been introduced. This workflow is intended to replace the other workflows in the long run. It is advantageous compared to the first two workflows because it does not require the user to know the necessary metadata files of TextGrid, nor does it require manual processing of TextGrid metadata [7]. The tool is available as a web-application[29] based on a Jupyter notebook and supports the automated mapping of TEI metadata to the corresponding TextGrid metadata, including manual adaptations by the user [8]. The deployment of the software for the import workflow was also realised on the Text+ JupyterHub, such that Text+ users can use the software directly in the browser without having to install it first.

[26] The TextGrid Laboratory is available to download under this URL: https://textgrid.de/en/download.
[27] TextGrid Import Tool: https://textgrid.de/en/datenimport.
[28] A comprehensive documentation of the Fluffy Import is currently being worked on.
[29] Gitlab: https://gitlab.gwdg.de/textplus/textgrid-import-ui.

## (Persistent) Identifier Types

Resources in the TextGrid Repository are organized in projects. While a project[30] itself would be referenced via an URL, elements within the projects[31], such as corpora, editions and texts, all receive an ePIC PID (under the prefix 21.11113) when being published. It is also planned to additionally assign DOIs in the future. ROR ID[32], GND[33], VIAF[34], ORCID[35], and other IDs referring to persons can be added in the TextGrid metadata.

TextGrid URIs are stable URIs used to identify every TextGrid object internally and externally (especially also non-public objects), where every TextGrid object consists of TextGrid metadata and a content file. Technical metadata are extracted automatically in the publishing process and can be used later for long-term preservation actions[36].

## Metadata

The TextGrid Repository emphasises a granular description of the archived research data and relies on its own TextGrid metadata schema[37]. The metadata schema was developed as a chained schema that serves editions, collections, aggregations, works, and items with different mandatory metadata for different user experience levels. By adopting this approach, researchers can effectively enrich their metadata, whether they are new to IT and metadata or have considerable experience in these areas. [9]. A comprehensive documentation of the metadata schema is available in [10].

Data stored in the TextGrid Repository are well discoverable as multiple portals harvest the information from the repository. For example, data are findable via the

[30] Example for a project: The “European Literary Text Collection (ELTeC)” would be cited via URL https://textgridrep.org/project/TGPR-99d098e9-b60f-98fd-cda3-6448e07e619d?lang=en.

[31] Example for a corpus and a text in the corpus: “The German ELTeC Novel Corpus (ELTeC-deu)” (https://hdl.handle.net/21.11113/0000-000F-F165-F and https://hdl.handle.net/21.11113/0000-000F-F165-F?noredirect) and “Weisse Sclaven oder die Leiden des Volkes: ELTeC ausgabe” (https://hdl.handle.net/21.11113/0000-000F-F4F8-6 and https://hdl.handle.net/21.11113/0000-000F-F4F8-6?noredirect), both would be cited using an ePIC PID.

[32] Research Organization Registry, https://ror.org/.

[33] Gemeinsame Normdatei, https://gnd.network/Webs/gnd/DE/Home/home_node.html.

[34] Virtual International Authority File, https://viaf.org/.

[35] Open Researcher and Contributor ID, https://orcid.org/.

[36] Example for metadata: The URLs https://textgridlab.org/1.0/tgcrud-public/rest/textgrid:41fwf.0/metadata and https://textgridlab.org/1.0/tgcrud-public/rest/textgrid:41fwf.0/tech (TextGri URI: textgrid:41fwf.0) reference the TextGrid metadata and the technical metadata extracted for long-term preservation of the text from the previous footnote.

[37] The schema can be downloaded here: https://textgrid.info/namespaces/metadata/core/2010.

TextGrid-search API[38], the DARIAH-DE generic search[39], OpenAIRE[40], and VLO[41]. Information can be harvested via the TextGrid OAI-PMH service[42].

## The DARIAH-DE Repository

The DARIAH-DE Repository is a long-term archive and data format agnostic repository for research data from the arts, humanities, and cultural studies. As in the TextGrid Repository, groups of objects can be combined into a collection. Archived research data are subject to less stringent requirements than in the TextGrid Repository, for example regarding metadata requirements. The PID assignment is integrated into the upload-workflow.

### Publication Workflow

The DARIAH-DE Publikator[43], a graphical interface located within the DARIAH-DE portal, is a tool used by researchers to publish data. The platform allows users to define and describe collections of objects and enrich them with metadata, along with their associated files. Once the collection has been prepared, it can then be imported into the DARIAH-DE Repository.

Technically, data and metadata are stored at the OwnStorage (i.e., non-public) of the repository as long as the researchers are working within the Publikator. When the publication process is initiated in the Publikator, data and metadata are passed on to the DARIAH-DE Publish Service (performing metadata validation and generating technical metadata) and from there to the DARIAH-DE CRUD Service (writing all objects to the PublicStorage and assigning PIDs) [11].

### PID Types

Currently, ePIC PIDs[44] (under the prefix 21.11113, same as TextGrid Repository) are used to identify objects in the DARIAH-DE Repository both before and after publication, they are used for the more administrative tasks. DOIs[45] (provided by

[38] Documentation of TextGrid-Search API: https://textgridlab.org/doc/services/submodules/tg-search/docs/index.html.
[39] The DARIAH-DE generic search is available at https://search.de.dariah.eu/search/.
[40] OpenAIRE: https://www.openaire.eu/.
[41] The CLARIN Virtual Language Observatory is available at https://vlo.clarin.eu/?2.
[42] Documentation of TextGrid OAI-PMH: https://textgridlab.org/doc/services/submodules/oai-pmh/docs_tgrep/.
[43] The Publikator is available at https://repository.de.dariah.eu/publikator/.
[44] Example for a collection: “DARIAH-DE Repository – Terms of Use”: https://hdl.handle.net/21.11113/0000-000B-C8EF-7 and https://hdl.handle.net/21.11113/0000-000B-C8EF-7?noredirect.
[45] The same collection as in the previous footnote is identified via https://doi.org/10.20375/0000-000B-C8EF-7 and https://doi.org/10.20375/0000-000B-C8EF-7?noredirect.

DataCite, under the prefix 10.20375) are assigned in the publication process and are recommended to cite the object. Hence, each object has both an ePIC PID and a DOI after publication. The DataCite metadata analysis report (see below) revealed that there are also some URNs identifying older data in the repository.

## Metadata

### Schema

Access to the publication mechanism is intended to be kept as low-threshold as possible in the DARIAH-DE Repository. Consequently, Dublin Core Simple[46] was selected as the metadata standard, comprising a minimal set of only 15 elements, three of which (title, creator, and rights) are mandatory. Furthermore, any additional metadata can be stored and published as objects, but will not be taken into account during indexing.

An extension of the metadata (with DCMI Metadata Terms[47] and the Datacite Metadata Schema[48] as possible schemas) has long been considered as a feature, but will not be implemented in the near future. DCMI Metadata Terms are currently being used to describe administrative metadata.

### Metadata Quality Report

As part of the PID4NFDI project, a metadata quality analysis[49] was conducted for those metadata that end up at DataCite for resources registered at the DARIAH-DE Repository. The findings indicated that there is room for improvement of metadata quality by

- adding ROR IDs for organizations and ORCIDs for creators and contributors,
- adding project funding information,
- adding contributor information,
- standardizing subject descriptions (for example in licensing/rights information).
- Collections in the repository are currently mapped to the DataCite resource type “dataset”, where “collection” might be the more reasonable choice.

[46] Specification: https://www.dublincore.org/specifications/dublin-core/
[47] Specification: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
[48] Specification: https://schema.datacite.org/
[49] The report is available at: https://docs.google.com/document/d/1eOa87NmhbbL5920XxUYZm_Za3KnCkloIbzsDVh6N2E8/edit?usp=sharing

By addressing these gaps, discoverability, reusability and the overall quality of the research outputs could be improved. Furthermore, comprehensive metadata of resources reduces the efforts required for obtaining certifications (such as CoreTrustSeal).

**Discoverability**

Collections and data published in the DARIAH-DE Repository via the Publikator are well discoverable. The collections are added to the Repository Collection Registry[50]. Collections and data are searchable via the DARIAH-DE Repository Search[51], the DARIAH-DE Generic Search, and DataCite Commons[52]. Metadata can be harvested via the DARIAH-DE Repository OAI-PMH service[53].

## PID Provider Considerations

Text+ does not decide as a consortium which PID providers to work with, but rather the data centers operating the repositories. Furthermore, the popularity of a particular PID integration is not considered to be decisive for the acceptance of a repository.

- The SUB has a PID-competent local partner in the GWDG (as member of the ePIC consortium). One advantage is the direct line to colleagues in GWDG, because both the SUB and GWDG are based in Göttingen. GWDG has been a reliable partner of SUB for many years, not only for providing PIDs, but also for many other collaborations.

- The SUB is a DataCite member (consortium lead organisation). The advantage of assigning DOIs is that they are more widespread than handles and also well known in the humanities.

## Metadata

The DARIAH-DE Repository and the TextGrid Repository take different approaches to make research data upload attractive to users: While the DARIAH-DE Repository enables users to archive research data with a minimal set of indexing metadata, the TextGrid Repository places great emphasis on the quality and granularity of the indexing metadata. The creation of collections of objects is possible in both

[50] The DARIAH-DE repository collection is available at https://repository.de.dariah.eu/colreg-ui/.
[51] The DARIAH-DE repository search is available at https://repository.de.dariah.eu/search/.
[52] DataCite Commons is available at https://commons.datacite.org/. The DARIAH-DE Repository can be found at https://commons.datacite.org/repositories/q98cjo within DataCite Commons.
[53] Documentation of DARIAH-DE OAI-PMH: https://repository.de.dariah.eu/doc/services/submodules/dhoaipmh/docs_dhrep/index.html.

repositories, and particularly useful in the TextGrid Repository for presenting a project or research context of the research data.

With the DARIAH-DE data federation architecture and, in the future, the Text+ Registry, two key infrastructure components will address the interoperability of research data, e.g. through mappings between different data models to enable standardised search and retrieval. Potential use of services such as TS4NFDI[54] or the NFDI Knowledge Graph are included in the work plan, although there is currently no focus on these. Machine-readability of metadata can be regarded as a basic requirement, also for the other services in the consortium.

In the TextGrid Repository, a scientific quality check takes place during the upload (Fluffy Import), including automatic extraction of metadata and subsequent curation by the user. Additionally, each published object (i.e., editions and collections) is reviewed again by SUB staff for quality and scope. A general metadata quality check is occasionally carried out for specific reasons, such as when metadata is newly indexed in another provider's portfolio (e.g. OpenAIRE) or prior to certification (e.g. CoreTrustSeal). The curation of the DARIAH-DE Repository involves a brief checking of basic metadata, e.g. addition of basic metadata or documentation.

## Technical Interoperability

Both the DARIAH-DE Repository and TextGrid Repository use the ePIC API[55] in their backend to register ePIC PIDs. To register DOIs, the DARIAH-DE repository uses the DataCite API[56]. So far, there were no issues, difficulties or restrictions reported regarding the use of the APIs. Both repositories also offer APIs themselves for uploading or publishing data. The repository systems are updated as needed. New versions and updates for bugs/features are available on Gitlab[57] via a CI/CD workflow.

## Training and Support

### For End-Users

PIDs have not been an issue in the helpdesk inquiries yet, but they do come up in discussions about data and repositories. Documentation and accurate description

[54] TS4NFDI: https://base4nfdi.de/projects/ts4nfdi
[55] Documentation: https://doc.pidconsortium.eu/docs/
[56] Documentation: https://support.datacite.org/docs/api
[57] GitLab: https://gitlab.gwdg.de/dariah-de/dariah-de-crud-services/-/releases

of metadata is generally difficult because users come with very diverse prior knowledge. Some users would like to only use their specific own metadata. Others wonder why they need to enter metadata at all, just wanting to publish their research results quickly. An attempt by the SUB to support researchers is to provide the translation of the DC Simple metadata from the Kompetenzzentrum Interoperable Metadaten (KIM)[58] with further additions and examples[59] in the Publikator of the DARIAH-DE Repository. If the support can illustrate the implications of rich metadata, researchers are usually receptive, unless they don't have time for the enrichment process. It is also regarded as useful to offer advice during the project and development phase in order to achieve the best possible documentation.

Many different training formats have been offered during the DARIAH-DE phases from 2012 to 2021. Currently, within Text+, there are still occasional formats that introduce the use of the services, e.g. specific training for the TextGrid Fluffy Import. However, the focus here is on the general use of the services/repositories, not on PIDs or metadata specifically.

### For Staff

So far, there has been no need for staff training on PID integration, which can certainly be explained by the fact that the GWDG and the SUB, as well as other institutions, already have extensive expertise in this area. Another issue concerns the division of labour with regard to PIDs: the employees or areas of an institution that are involved in the topic are often not congruent with the working structures of the consortium. Prospective training for repository operators/developers in the context of PID4NFDI is considered helpful.

## PID Policy & Governance

In the current phase of Text+, the focus is generally on infrastructure-related tasks that do not concern PID governance. Also, Text+ does not have a policy or set of guidelines regarding the management or use of PIDs. However, the topic of PIDs might present itself differently as soon as a fully ramped-up PID4NFDI basic service would be available.

[58] Webpage: https://dini.de/standards

[59] Gitlab: https://gitlab.gwdg.de/dariah-de/dariah-de-repository/dariah-de-publikator/-/blob/develop/src/main/webapp/schema/dcelements_de.ttl?ref_type=heads

For the DARIAH-DE Repository and the TextGrid Repository, two infrastructure partners with extensive PID expertise, namely GWDG and the SUB, are responsible for the services, so there are no current challenges for Text+ in terms of organisation or governance.

## Outreach

Apart from the ideally seamless integration of PID allocation into the services, in particular repositories, PIDs are currently not a topic in the consortium's outreach activities.

# Summary

This use case analysis presents the use of PIDs by the SUB as a data and competence center in the Text+ consortium, with a particular focus on the TextGrid Repository and the DARIAH-DE Repository.

PIDs in the context of Text+ are generally applied to a wide range of resource types from the digital humanities, such as editions, lexical resources, collections and objects within collections. In addition to these resources that are typically stored in repositories, PIDs have been discussed in the context of service descriptions, search strings and intermediate statuses, but have not been implemented for these purposes yet.

Both the TextGrid Repository and the DARIAH-DE Repository have been operational for many years. As a result, PIDs (DOIs and ePIC PIDs) are already integrated seamlessly into the systems. Both repositories have different policies on metadata granularity and quality to ensure that every researcher has the opportunity to describe their research data according to their level of metadata affinity. Improvements are currently being made to the upload workflow of the TextGrid Repository to make it more user-friendly. The focus is also on increasing the discoverability and trustworthiness of research data by integrating metadata into search and retrieval tools at the German and European level and by certifying repositories (e.g. CoreTrustSeal). However, in Text+, issues such as governance, policies, outreach and training are being addressed without a specific focus on PIDs.

The analysed use case with its repositories provides a very mature service which existed prior to NFDI. The repositories do not only cover NFDI requirements, but also have long standing experiences in managing PIDs and are therefore facing less challenges than NFDI services in development. Currently, the SUB has no imminent need for support from PID4NFDI, but can serve as a best practice model for emerging infrastructure services within NFDI. However, a fully ramped-up PID4NFDI basic service in the future would provide an opportunity for the SUB to further discuss and develop issues such as PID-related training for the large group of Text+ data centers and issues around PID policy and governance.

# References

[1] El-Gebali, S., Hagemann-Wilholt, S., Böhm, J., & Kahlert, T. (2024). D1.1 Landscape of PID practices within NFDI services. Zenodo. https://doi.org/10.5281/zenodo.14277479.

[2] Bingert, S., Buddenbohm, S., Loizides, F. (2016). Research data center services for complex software environments in the humanities. Information Services and Use. 36(3-4):189-202. https://doi.org/10.3233/ISU-160817.

[3] “TextGrid Repository & Laboratory”, https://marketplace.sshopencloud.eu/tool-or-service/oKFMi6 [Online; accessed 15-Dec-2024]

[4] “Das DARIAH-DE Repository und das TextGrid Repository”, https://doc.de.dariah.eu/Das-DARIAH-DE-Repository-und-das-TextGrid-Repository/ [Online; accessed 15-Dec-2024]

[5] CoreTrustSeal Standards and Certification Board. (2022). CoreTrustSeal Requirements 2023-2025 (V01.00). Zenodo. https://doi.org/10.5281/zenodo.7051012.

[6] “The TextGrid Repository Documentation”, https://textgridlab.org/doc/services/index.html [Online; accessed 15-Dec-2024]

[7] Buddenbohm, S., Calvo Tello, J., Funk, S. E., Klammer, R., Rißler-Pipka, N., Steckel, A., Veentjer, U., Weimer, L., & Dogaru, G. (2024). Fluffy Import: Preserving Humanities Research Data with the TextGrid Repository. In TRANSFORMATIONS - A DARIAH Journal (Number Workflows: Digital Methods for Reproducible Research Practices in the Arts and Humanities). Zenodo. https://doi.org/10.5281/zenodo.14025199.

[8] “Fluffy Workflow: Neue Tools für den Datenimport ins TextGridRep”, https://textplus.hypotheses.org/10328 [Online; accessed 15-Dec-2024]

[9] “Data Policies”, https://textgridlab.org/doc/services/data-policies.html [Online; accessed 15-Dec-2024]

[10] “Metadata”, https://doc.textgrid.de/Metadata/ [Online; accessed 15-Dec-2024]

[11] “Publikator”, https://repository.de.dariah.eu/doc/services/submodules/publikator/docs/index.html [Online; accessed 15-Dec-2024]