rHd Revista de Humanidades Digitales

RHD 8 (2023)

ISSN

2531-1786

Between Corpora, Tools, and Authority Files: TextGrid Repository for Hispanic Studies

Entre corpus, herramientas y archivos de autoridad: TextGrid Repository para los Estudios de Literatura Española

José CALVO TELLO
State and University Library Göttingen
calvotello@sub.uni-goettingen.de
https://orcid.org/0000-0002-1129-5604

Mathias GÖBEL
State and University Library Göttingen
goebel@sub.uni-goettingen.de
https://orcid.org/0000-0002-1102-5284

Nanette RißLER-PIPKA
Max Weber Stiftung
Rissler-Pipka@MaxWeberStiftung.de
https://orcid.org/0000-0002-0719-9003

Lukas WEIMER
State and University Library Göttingen
weimer@sub.uni-goettingen.de
https://orcid.org/0000-0001-6919-3646

Stefan E. FUNK
State and University Library Göttingen
funk@sub.uni-goettingen.de
https://orcid.org/0000-0003-1259-2288

Daniel KURZAWE
State and University Library Göttingen
kurzawe@sub.uni-goettingen.de
https://orcid.org/0000-0001-5027-7313

Ubbo VEENTJER
State and University Library Göttingen
veentjer@sub.uni-goettingen.de
https://orcid.org/0000-0002-9726-3135

RESUMEN

En este artículo abordamos las formas en que los proyectos de Humanidades Digitales organizan sus recursos y argumentamos a favor de colaborar con servicios especializados de otras instituciones para tareas específicas, al igual que hacen los editores de investigación. Más concretamente, defendemos que las bibliotecas tienen un nuevo papel central que desempeñar en la publicación de datos de investigación a través de repositorios. Tomando como ejemplo el TextGrid Repository (TGR), presentamos sus principales componentes, su historia, sus principales características y discutimos sus ventajas y desventajas en comparación con repositorios y tecnologías similares. A continuación, presentamos su desarrollo actual dentro del consorcio Text+, tanto en relación a nuevas funcionalidades, así como a la integración de corpus ya existentes. Estos nuevos corpus hacen que este repositorio no solo sea interesante para la literatura alemana (como hasta ahora), sino también para otras tradiciones, especialmente en español.

PALABRAS CLAVE

Texto, infraestructura, metadatos, almacenamiento, codificación.

ABSTRACT

In this article, we discuss the ways in which Digital Humanities projects organize their resources and argue for the need to collaborate with specialized services from other institutions for specific tasks, just as research publishers do. More specifically, we argue that libraries have a new central role to play in the publication of research data in repositories. Using the TextGrid Repository (TGR) as an example, we present its main components, its history, its main features, and discuss its advantages and disadvantages in comparison with similar repositories and technologies. We then present its current development within the Text+ consortium, both in the form of new features but also in the integration of existing corpora, making it interesting not only for German literature (as has been the case until now), but also for other traditions, especially Spanish literature.

KEYWORDS

Text, Infrastructure, Metadata, Storage, Encoding.

Revista de Humanidades Digitales http://revistas.uned.es/index.php/RHD/index
Recibido 04/08/2023 – Aceptado 20/10/2023
Editado por R. De León

1. CURRENT CHALLENGES FOR DH PROJECTS WORKING WITH SPANISH LITERATURE

Digital Humanities (DH) researchers working with Spanish literature face several difficulties in comparison when working with languages such as English or German. First, few corpora with Spanish literature are openly available in a reusable format, contain textual structure (for example in TEI XML), the editions are of philological quality and the files are annotated with metadata linked to authority files (such as the ones curated by National libraries) or knowledge basis (such as Wikidata).

In a recent survey about the Spanish-speaking community of users of TEI XML, the researchers have found that the majority of the respondents do not share their data (del Rio Riande and Allés-Torrent, 2023). As these authors have explained in another article, pioneer projects such as Biblioteca Virtual Miguel de Cervantes (BVMC) and Corpus Diacrónico del Español (CORDE) did not make their data available for the rest of the community in TEI XML (Allés-Torrent and del Rio Riande, 2019). The survey also highlights the lack of infrastructures, which can be argued that it impacts the Spanish-community stronger than other Romance languages because of the late participation of Spain in infrastructure consortia like DARIAH or CLARIN (Spain only joined in summer 2023) in comparison to countries like France or Italy which were founding members of DARIAH.

Second, projects tend to focus on short-term outcomes, such as online interactive visualizations and publications for reports at the end of specific projects. As we will discuss later, the previously mentioned survey points out that almost all projects rely on their own portals or instances of technologies (such as eXist-DB, TEI Publisher, or TEITOK) (del Rio Riande and Allés-Torrent, 2023). However, it is rarely asked whether the data that many DH projects spend a long time collecting will be available in a few years. We take it for granted that libraries should preserve books and journals for decades and even centuries. But what will happen to our portals and instances of technology containing digital editions, corpora, and databases in 20 years or 50 years (when the project leader is likely to have retired)? Who is going to take care that they will still be available?

Third, the increase in technological and ethical requirements around DH projects raises whether a single study or research project can respond to all these aspects. To some extent, DH projects are expected to be FAIR, CARE, use Linked Open Data technology, use complex deep algorithms, be open, smart, and user-friendly. All these are or can be positive aspects, but they add new layers of complexity to any project, layers that need to be considered when managing the resources of the projects.

All projects are affected by limited resources, and this is especially urgent in situations where resources are severely constrained. The previously mentioned survey names the insufficient funding that affects many projects of the Spanish-speaking community, especially in Latin America (del Rio Riande and Allés-Torrent, 2023). The greater the economic restrictions are, the more urgent it is to manage them efficiently, both in the short and in the long-term.

Many projects working with Spanish literature still relies on their own for the different tasks, such as digitization, data curation, transformation, and publication, software development, user identification and the maintenance of all these aspects. This centralized research data management is questionable for at least two reasons: first, it is unlikely that a single project has sound expertise for all these tasks, therefore, the solutions for many of these areas would be mediocre. Second, the long-term development and especially the accessibility of the tools and data is not given. Actually, the survey highlights the isolation of the researchers as one of the main problems (del Rio Riande and Allés-Torrent, 2023).

For these reasons, we invite projects to change their perspective of structuring projects from what we call a single-project perspective to a distributed-project perspective. While distributed project management in the context of agile software programming is well known and also applied in DH projects (Ermolaev et al. 2018; Tabak 2017), our perspective is to include infrastructure outside of the project. In a single-project perspective, each project tends to cover all the previously mentioned areas (digitization, curation, transformation, publication, analysis, visualization, maintenance, etc.). We would argue that the majority of projects working with Spanish literature manage their resources in this way.

The alternative is to adopt a distributed project perspective. In this perspective, each project would be responsible for only some of the areas mentioned above and would be supported (either partially or fully) by other projects, especially infrastructure projects (such as DARIAH or CLARIN), or other research and cultural heritage institutions (such as libraries, museums, computing centers, etc.). Each project would embrace the advantages and challenges of being connected to a distributed resource environment driven by the community. The collaboration between this plurality of actors is mainly determined by the use of technical standards and the openness of the resources. These two aspects are at the core of the implementation of the FAIR principles (Wilkinson et al., 2016, 2018), one of the most important guidelines for research data management. The main advantage of this perspective is that each project does not have to take care of all the tasks but can share the responsibilities and benefit from the expertise of others. Although we are convinced that this perspective is advantageous, we are also aware that it brings some challenges, for example that very specialized features of projects are not treated as they would be in a single-project perspective or that the difference in speeds and goals in the different projects can create tensions. In the next section, we describe how this is currently developing in Germany.

2. TEXT+

The national research data infrastructure (NFDI) in Germany is built to tackle the above mentioned challenges –not only for humanities or computational literary studies but in cooperation of all disciplines (Kraft et al., 2021). The expertise and services regarding research data management should no longer be part of every single project and in some generic aspects not even of every single discipline but will be treated in distributed teams, consortia and in the end in one distributed infrastructure.

Text+ is a consortium of the NFDI for language and text-based research data (Hinrichs et al., 2022; Kett et al., 2022). The main objective of the consortium is to make discipline specific datasets sustainably accessible, linked and qualitatively (re)usable for the language –and text– based communities, but also for the entire German science system, in accordance with the FAIR principles (Findable, Accessible, Interoperable, Reusable) (Wilkinson et al., 2016). Text+ is organized in three data domains Collections, Lexical Resources and Editions, although further data domains are to be expected. These are framed by the task area Infrastructure/Operations for the overall research data management services.

As a locally distributed infrastructure, Text+ is structured along data and competence centres that focus on the three data domains. The resources already available at these centres are successively being integrated into the overall Text+ infrastructure; further resources are being added by the wider community via Cooperation Projects[1] and external data providers. Due to the composition of the consortium the data has a clear focus on German, and English, language data. Other languages, like Spanish on the other hand, are underrepresented and particularly addressed in order to create a balance.

The Göttingen State and University Library (SUB) leads together with the German National Library (DNB) the cluster "unstructured text" which is part of the data domain Collections. Both libraries aim to foster the transformation of unstructured text which is mostly the result of mass digitization into structured text. For this purpose, the SUB provides the TextGrid Repository. The repository focuses on TEI XML-encoded texts and images and has no restrictions regarding the languages. Text+ pushes the expansion of its portfolio to include more different language resources, e.g., Spanish.

3. TEXTGRID

3.1. Components and History

TextGrid has been developed in several projects over the years, always funded as part of research programmes in Germany. It was first funded as part of the TextGrid project between 2006 and 2015. Since then, it has been further developed within DARIAH-DE (2011-2019) and the joint project between CLARIN-D and DARIAH-DE in the form of CLARIAH-DE (2019-2021). Today, TextGrid is part of the services offered by the Association for Research Infrastructures in the Humanities and Cultural Studies (GKFI e.V.) and in-kind contribution for DARIAH ERIC (Weimer, 2022). It is operated and further developed within the NFDI consortium Text+ by the Göttingen State and University Library in cooperation with computing center GWDG.

During the original TextGrid project, three components were conceived as the core: the laboratory (TGL), the repository (TGR) and the community. The community was thought to consist of the people and institutions associated with the funded project, but also of the users of the

[1] See: https://www.text-plus.org/en/research-data/cooperation-projects/.

laboratory and the repository. Due to the active community and one of the initial collections (the Digital Library, Betz, 2015)[2] the TextGrid project had a clear focus towards German literature and texts, with several leading researchers coming from this field (Wegstein et al., 2015, 27). In the initial phase, 2006, the DH landscape in Germany (and also internationally) was not yet visible. Only slowly the community formed and organized itself in an association (DHd[3] since 2012) and in the above-mentioned research infrastructure consortia. This aspect has changed considerably in the German-speaking world since then, with new associations, consortia, COST actions and other funded projects being set up and in which many of the original initiators of TG are still active.

The TextGrid Laboratory (TGL) is the client component of the infrastructure, offering a graphical user interface not only to interact with the server components but also for editing documents, organizing projects and collections of documents (Aschenbrenner, 2015). This software runs on the local computer and works on TextGrid resources stored at the remote servers. This client is feature-complete covering the workflow starting at file import (e.g., images, documents such as transcriptions and also technical files for processing and visualizing the data), creating new ones and editing and –finally– publishing at the TextGrid Repository.

The third component is the TextGrid Repository (TGR), which is the main topic of this article. This repository is not intended as a repository for research publications such as arXiv or those based on Digital Commons, but as a repository for research data. More specifically, it was designed as a repository for XML (usually TEI) editions and the images from the digitization. The main workflow was for researchers to create their editions in the TGL and then publish them in the TGR.

3.2. Characteristics of TGR

An important aspect of the TextGrid environment is that a relatively complex model of bibliographic records has been used to structure the metadata associated with the text. We refer to the Functional Requirements for Bibliographic Records (or FRBR) model, a very widespread theoretical model in libraries (Hickey et al., 2002; Patton, 2013; Chambers, 2013). This model distinguishes between following instances:

- works (an abstract creation, for example *Pride and Prejudice* by Jane Austen),
- expressions (for example the original English version of this work, or a specific translation into, say, French),
- manifestations (the Penguin edition of *Pride and Prejudice* in English)
- and items (a specific exemplar that I can hold in my hand of the Penguin edition of the English version of *Pride and Prejudice*).

Following the FRBR model, TGR foresees works, editions (which correspond to the manifestation in the FRBR terminology) and items (Funk and Pempe, 2015). These items would be in TGR the TEI files with the texts encoded. Besides, TGR also offers the collections object, in which

[2] See: https://textgrid.de/en/digitale-bibliothek.

[3] See: https://dig-hum.de/.

different works, editions and items can be associated for different reasons. For example, while metadata about the year of publication of a particular edition would be stored in the manifestation metadata file, metadata about the year of creation of the work would be stored in the work metadata file. Although this is a much more accurate way of modelling the information, it also adds complexity when using the already published text in the TGR, and especially when publishing new texts, which we will discuss later.

An important milestone of TextGrid was the publication of the Digital Library, a collection of more than 100,000 literary texts in German, many of them translated from other languages (Betz, 2015). This collection was purchased from the company Zeno and was part of a larger collection of documents that also included other non-literary texts, such as dictionaries, which had not yet been published. It can be argued that this Digital Library has become one of the most widely used textual resources in the German-speaking countries. Many other corpora containing German literature, such as KOLIMO (Herrmann and Lauer, 2018) and the German subcorpora within DraCor (Börner and Trilcke, 2023; Fischer et al., 2019) or ELTeC (Burnard et al., 2021; Schöch et al., 2021), rely on texts from the Digital Library.

A further characteristic of the repository is that it has obtained the CoreTrustSeal (CTS) from 2020 to 2023. The CTS is an international, community-driven NGO that provides a professional baseline certification for trustworthy repositories (L'Hours et al., 2019). The requirements for obtaining the CTS are updated in a three-year cycle; regular reapplication is considered necessary. Until March 2023, the TGR was certified with the CTS; it is currently in the process of reapplying.

3.3. Alternatives to TGR

What alternatives to TGR do researchers have for publishing and archiving their TEI XML-coded documents? We should distinguish between two types of repositories: generalist and XML-specific repositories. In both of them, we highlight some solutions that exist or existed within the Spanish-speaking community.

The generalist repositories provide a place for researchers to make their research data available. These repositories tend to offer a range of services such as the ability to tag and metadata, to assign persistent identifiers, to send references to the data to other harvesting services, and so on. Currently, the most widely used platform for publishing research data is Git repositories, either in GitHub (hosted by Microsoft) or in institutional instances of GitLab. These Git repositories are not a long-term solution, as any change in the institutions behind them could jeopardize the accessibility of the data.

Many researchers choose to publish their Git repositories in Zenodo, an open generalist repository supported by CERN. Another generalist repository oriented on research data is RADAR, operated by FIZ Karlsruhe. It is a fee-based service for researchers at German universities and non -university research institutions. However, there are already cost-free disciplinary instances, such as RADAR4Culture, which is focused on cultural studies research data. Other examples of generalist repositories can be found in re3data[4]. These repositories can have a geographical or institutional focus at universities, countries or states, or consortia and are usually open to the wider community. Many of these repositories are instances of software such as DSpace, Fedora Commons, or Dataverse. An example of the last-mentioned software is the data repository e-cienciaDatos[5] of the consortium Madroño[6] from several universities in Madrid. We describe these repositories as generalist because users can publish data in any digital format; the repositories receive research data as black boxes and give it back to other users as black boxes. The repositories do not normally attempt to analyze or extract information from the research data published in them. For many cases, the flexibility of these repositories is an advantage.

However, if a community uses a particular format extensively, it can be argued that it would be beneficial to use repositories that can parse the research data to some degree. We agree with Burnard that the "Text Encoding Initiative (TEI) is one of the longest-lived and most influential projects in the field now known as the Digital Humanities" (2014), so there is an advantage in having repositories specific to this format, such as TGR.

There are several solutions and repositories for TEI XML encoded texts. The repository most similar to TGR is Gams[7], which is a repository hosted by the University of Graz in Austria (Steiner et al., 2022). An important resource published here are the corpora of Spectators (or Moral Weeklies)[8] in English, French, German and Italian (Ertler, 2014). Gams has also obtained the CTS. However, for projects to be allowed to publish their resources in Gams, they must first sign an agreement with the Austrian Centre for Digital Humanities. In general, the aim of this centre is to support projects from Austria on a national level only.

The second option most similar to TGR is the TAPAS repository, hosted by the Northeastern University Library (in Boston, USA). Since its launch in 2014, TAPAS has been a benefit of membership in the TEI Consortium, but this changed in 2021, when it became available to all users with a TAPAS account, which can be obtained free of charge. The repository is not currently recruiting new users, as the main efforts are directed towards updating various components and technologies of the repository.

Another solution is the Deutsches Textarchiv (DTA), a project that collects and publishes TEI XML-coded texts, but also allows other projects to publish their own corpora (Kampkaspar, 2017; Wiegand et al., 2018). However, as the name of the project implies, these texts must be in German.

Another option available as of 2021 is the new not-for-profit association Sources Online[9], an endeavor of various edition partners and projects, many of them based in Switzerland (Kränzle et al., 2023). This platform offers several services to projects, including hosting of the TEI Publisher and IIIF servers. The cost of using these services (which are available online)[10] depends on the

[4] See: https://www.re3data.org/.
[5] See: https://edatos.consorciomadrono.es/.
[6] See: http://www.consorciomadrono.es/ .
[7] See: https://gams.uni-graz.at/.
[8] See: https://gams.uni-graz.at/archive/objects/context:mws/methods/sdef:Context/get?locale=en.
[9] See: https://sources-online.org.

complexity of the edition, with one-time fees at the beginning and annual costs, both in the thousands of Swiss francs. Considering the problems of funding for many projects dealing with Spanish literature, these costs will exclude many of them.

A final alternative is for a project to install and host an instance of software such as eXist-DB (Retter, 2014)[11], TEI publisher[12], or TEITOK[13]. The survey about the users of the TEI shows that using instances of these software is one of the most frequently used manners of publishing TEI files within the Spanish-speaking community (del Rio Riande and Allés-Torrent, 2023). The DraCor project uses an eXist-db[14], which has published various drama corpora in European languages, but without a medium- or long-term plan (Börner and Trilcke, 2023; Fischer et al., 2019). While we believe that using these software solutions is a more sustainable approach than using ad hoc portals, in both cases the responsibility for maintaining and upgrading the software remains with each project.

In the Spanish-speaking community, for a time there was a tool for editing and publishing texts in TEI that combined several of the technologies mentioned in the previous paragraph, such as TEI Publisher and eXist-DB. It was called EVI-LINHD[15] and was conceived as a virtual research environment that also functioned as a repository (González-Blanco et al., 2017). The website is no longer available, and the project leaders confirmed to us in personal communication that the tool is not being continued. Thus, to our knowledge, the Spanish-speaking community has no current tool specifically for publishing TEI files.

3.4. Advantages and Disadvantages of TGR

In comparison with other projects, it is possible to identify some advantages of TGR over similar solutions. TGR is free of charge, open to groups from any country and multilingual. TGR is supported by a library and the GWDG computing center, institutions whose primary funding is not linked to specific projects, and which provide various services on a long-term basis. As mentioned before, TGR has obtained the CoreTrust Seal (CTS) from 2020 to 2023 and currently is in the process of reapplying. In addition, the data published in TGR is then harvested by other platforms such as OpenAire, EOSC, or the CLARIN Virtual Language Observatory and Virtual Collection Registry via OAI-PMH[16]. As we will explain in the section on the current state of the repository, the fact that TGR is an XML-specific repository allows the data in the published files to be used in a variety of ways, and so its XML-specificity can be seen as an advantage.

Of course, this specificity could also be seen as a problem, especially for those projects that use other formats. Although TEI XML is the de facto standard format for scholarly digital editions, there is no accepted format for the corpus community (both working with linguistic and literary data, Burnard et al., 2021). While many projects use TEI XML, many others prefer plain text, often

[10] See: https://archives-online.org/sources_online_flyer.pdf.
[11] See: http://exist-db.org.
[12] See: https://teipublisher.com.
[13] See: http://www.teitok.org/.
[14] See: https://dracor.org/doc/what-is-dracor.
[15] See: https://linhd.uned.es/entorno-virtual-de-investigacion.
[16] See: https://textgridlab.org/doc/services/submodules/oai-pmh/docs_tgrep/index.html.

extracted from HTML, or model their data in other formats such as JSON, tables, or other XML specifications.

Another disadvantage of TGR is the complexity of the metadata model, which is partly due to the instances (work, edition, and item) based on the FRBR model explained before. Although the structure of the metadata at these levels allows for a very specific description, the majority of projects tend to encode only one item per edition and per work. The three different levels are therefore redundant for many projects. In addition, in recent years the identification of works has been mainly done through authority file identifiers (such as those assigned by the Spanish National Library or the GND for the German-speaking world) and knowledge bases (such as Wikidata and VIAF). The complexity of the metadata models makes it difficult to navigate in TGR, but it is particularly challenging when users want to publish the data in TGR, because they have to generate a data file and a metadata file for each level. In other words, in order to publish a single TEI XML file, users should correctly generate at least six files. In addition, the current import process requires users to use different technologies and techniques that may be new to them. We are working with other Text+ partners to develop a new import process that will guide users through the process. As with any software development, the development of TGR is constrained by the technologies and resources available.

A final drawback of TGR is that it is a generic solution for many projects. Many edition projects still create their own website to publish the encoded files and present them exactly as they wish. Although TGR offers different options to projects, many of the structural elements are common to all projects. This can lead to a situation where a project cannot show a particular feature or characteristic in TGR in a specific way. The TGR development team is aware of this and, as we will explain later, some of the latest developments are trying to improve this by giving more flexibility in how the projects can present themselves, how they transform the text for reading, and how they use the metadata behind the facets for filtering and searching.

4. INTEGRATION OF NEW RESOURCES

In this section, we would like to present two resources that have recently been added to the TGR and are relevant to Spanish literature. In addition to these corpora, we are currently considering and in contact with several other projects to support their publication in the repository. As described in the section "Current Developments", we are designing new and simpler import methods for future corpora.

4.1. Corpus of Novels of the Spanish Silver Age

The first corpus recently added to TGR relevant for this publication is the Corpus of Novels of the Spanish Silver Age (CoNSSA). This corpus contains prose works (mostly novels) by Spanish authors published between 1880 and 1939. The original corpus contains 358 texts, but since a number of authors died in the second half of the 20th century, their texts are still protected by copyright and cannot be published openly. For this reason, currently it is possible to publish 219 texts. The corpus is encoded in TEI XML, with numerous metadata in the header and a superficial codification of the text (with specific TEI elements for paragraphs, verses and poems, headings, cursive passages, etc.). The corpus has also been annotated with several lexicographic resources and natural language processing tools, has been previously published on GitHub and Zenodo, and has been described in various publications (Calvo Tello, 2021a, 2021b).

Since the beginning of 2023, the corpus is also available in TGR[17] and benefits from features of an XML-specific repository. Users can send texts on the fly to different tools, which can be used in the classroom. For example, the prose texts by Valle-Inclán *La media noche* obtains now a persistent identifier[18], and users can send the text to different tools without downloading or installing locally neither texts nor tools. Actually, the import of the texts into different tools obtains a URL that can be shared with the group for an even more comfortable experience. For example, the text can be sent to Voyant Tools[19] or to Named Entities Recognition tools connected to the CLARIN Switchboard[20].

4.2. European Literary Text Collection (ELTeC)

The second project that has been incorporated into TGR recently and which is also relevant for Spanish literature but also for many other literatures is the European Literary Text Collection (ELTeC) corpora. These have been compiled in the frame of the COST Action "Distant Reading for European Literary History", containing novels in different European languages, published for the first time between 1840-1920. The goal of the project was to obtain 100 novels per language following a series of composition criteria and offer them encoded in TEI XML. The corpora and the details about the composition of several of the corpora have been already documented extensively in several publications (Burnard et al., 2021; Schöch et al., 2021).

In 2023, 15 of the corpora of the ELTeC with at least 50 texts were integrated into TGR[21]. This represents a total of 1.365 new texts in the repository in 14 European languages, including Spanish, many of them without any exemplar of those languages until now (Rißler-Pipka et al., 2023). For this import, several new features were developed for the TGR, options that are now available for any other project and which we will describe in the next sections.

5. CURRENT FEATURES OF TGR

As a repository, the main purpose of TGR is to make data available to other researchers for the future. Data can be retrieved, for example, through the Persistent Identifiers (PID), more

[17] See: https://textgridrep.org/project/TGPR-8b44ca41-6fa1-9b49-67b7-6374d97e29eb.

[18] See: https://hdl.handle.net/21.11113/0000-000F-774F-4.

[19] See: https://voyant-tools.org/?corpus=bf906c7495bbdf86e9454475d9f9a61a&input=https://textgridlab.org/1.0/tgcrud-public/rest/textgrid:40wc3.0/data.

[20] See: https://lindat.mff.cuni.cz/services/nametag/?data=https%3A%2F%2Fswitchboard.clarin.eu%2Fapi%2Fstorage%2F63e3f7a8-2f24-46e5-8b12-6b432e5b9475%3Fmediatype%3Dtext%252Fplain&model=spa.

[21] See: https://textgridrep.org/project/TGPR-99d098e9-b60f-98fd-cda3-6448e07e619d.

specifically ePIC Handles[22], that each object in TGR receives. Not only is each corpus or collection assigning a PID (as in the case of Zenodo), but also each individual object, such as the data and metadata files for the different levels (item, edition and work) of TGR. If a project publishes images of the digitisations, each image file would also receive a PID. The portal offers citation suggestions that include the object's PID, making it easier to cite the entire corpus or specific objects within it[23].

As mentioned above, one of the advantages of TGR is the fact that the repository is able to parse and transform specific data formats, in particular TEI XML. Probably the most important feature in this regard is that TGR converts the original TEI XML to HTML, allowing users to read the encoded text as they would any other content on the web. In other words, users can not only deposit their data in TGR, but also create a publication of the texts at the same time. TGR also offers other transformations of the files, such as eBook (ePUB, although this transformation is currently being developed for the Digital Library and may not be acceptable for other projects) or plain text (specifically for further processing in other programs).

Another recent feature related to parsing and processing the internal data of the files is that TGR generates a Table of Contents of the TEI XML files and displays it in the left menu. This gives users an overview of the text and allows them to navigate to its sections. This, and the transformation into other formats, are clear advantages of TGR over generic repositories.

Another new feature of TGR is the ability to search for different types of objects. Firstly, it is possible to search for words both in the whole repository and in specific projects. For example, we can search for the word *Austria* in the ELTeC corpora[24]. We can also use special characters (similar to regular expressions), e.g., searching for the string "españ*" would give any word starting with that root. TGR allows you to search not only in the text in the TEI XML file, but also in the metadata of the TGR files. For example, we can search for works that were written in a certain period of time (e.g. between 1900 and 1910: work.dateOfCreation.value:>1900 work.dateOfCreation.value:<1910)[25] or the gender of the author. We can also combine these queries and search within ELTeC for texts written by female authors, written between 1900 and 1910 and containing any word with the root "españ*" (work.vocab.eltec.authorGender:female AND work.dateOfCreation.value:>1890 AND work.dateOfCreation.value:<1900 AND Españ*)[26]. In other words, we can combine criteria for the metadata of the work, the author, and aspects of the full text in a single query.

TGR is also integrated with other tools in the DH environment. Users can send any text (or even corpora, as we will explain later with the shelf function) to different tools simply by clicking on the tool on the left. The advantage of this integration is that users can use specific tools without having to download (and then upload) the texts to their local computer or install any software. We

[22] See: https://www.pidconsortium.net.

[23] See: https://textgridrep.org/browse/40wc3.0#citation.

[24] See: https://textgridrep.org/search?query=Austria%20&order=relevance&limit=20&mode=list&filter=project.id:TGPR-99d098e9-b60f-98fd-cda3-6448e07e619d.

[25] See: https://textgridrep.org/search?query=work.dateOfCreation.value%3A%3E1900+work.dateOfCreation.value%3A%3C1910+&order=relevance&limit=20&filter=project.id%3ATGPR-99d098e9-b60f-98fd-cda3-6448e07e619d&filter=format%3Atext%2Fxml.

[26] See: https://textgridrep.org/search?query=work.vocab.eltec.authorGender%3Afemale+AND+work.dateOfCreation.value%3A%3E1890+AND+work.dateOfCreation.value%3A%3C1900+AND+Espa%C3%B1*&order=relevance&limit=20&filter=project.id%

think this is particularly useful for integration in classrooms or workshops, but also for researchers to explore texts and tools without spending time on installations. There are currently three tools integrated in this way: Voyant Tools, CLARIN Switchboard and an annotation tool.

Voyant Tools is one of the standard tools for corpus exploration and analysis in DH (Sinclair and Rockwell, 2016). It offers many possibilities for quantitative analysis of the texts and visualization of this analysis, such as key-word-in-context analysis, generation of n-grams and collocations, distribution of frequency of terms, etc. Since Voyant can also import TEI XML, users can use the textual structure for these analyses. Recent developments in Voyant have seen it evolve into an environment called Spyro, in which the former Voyant modules are now functions that can be used in an environment similar to Jupyter Notebooks. As we will see in the next sections, TGR is also evolving in this direction.

The second tool currently connected to TGR is CLARIN Language Resource Switchboard (LRS) (Zinn, 2018). The LRS automatically recognizes the data format and language of the texts or text corpora to be analyzed and directs users only to those analysis tools that are suitable for their case. In this way, it allows users to test and apply NLP tools to both texts and corpora without the need to install any software.

Finally, users can also use an annotation tool developed in the context of DARIAH-DE. The use of this tool requires access to a DARIAH account (which can be created for free and does not require academic affiliation)[27], and can also be accessed via eduGAIN. Once logged in, users can make annotations (both individually and in groups) in the form of free text or categories. These annotations are structured in relation to specific TEI sections using xPaths. The annotations can then be exported. This module can be used as part of a research project, for example to quantitatively analyze these annotations later, or to train algorithms to predict the annotated categories in other texts. But it can also be used for annotating texts in a classroom context, in different settings: each student annotates a different text, all students annotate the same texts but with different phenomena, all students annotate the phenomenon in the same text, etc.

Another feature of TGR is the Shelf function, which allows any user to select texts and other TGR objects, create a collection of them, and then perform various functions on that collection[28]. For example, the user can download these objects in a single ZIP file (both plain text and TEI), or send the collections to some tools such as Voyant (the CLARIN Switchboard does not yet accept collections of texts).

6. CURRENT DEVELOPMENTS

In this section, we present some features that have been developed in recent months, some of them motivated by the release of the ELTeC corpora. In the previous section, we presented

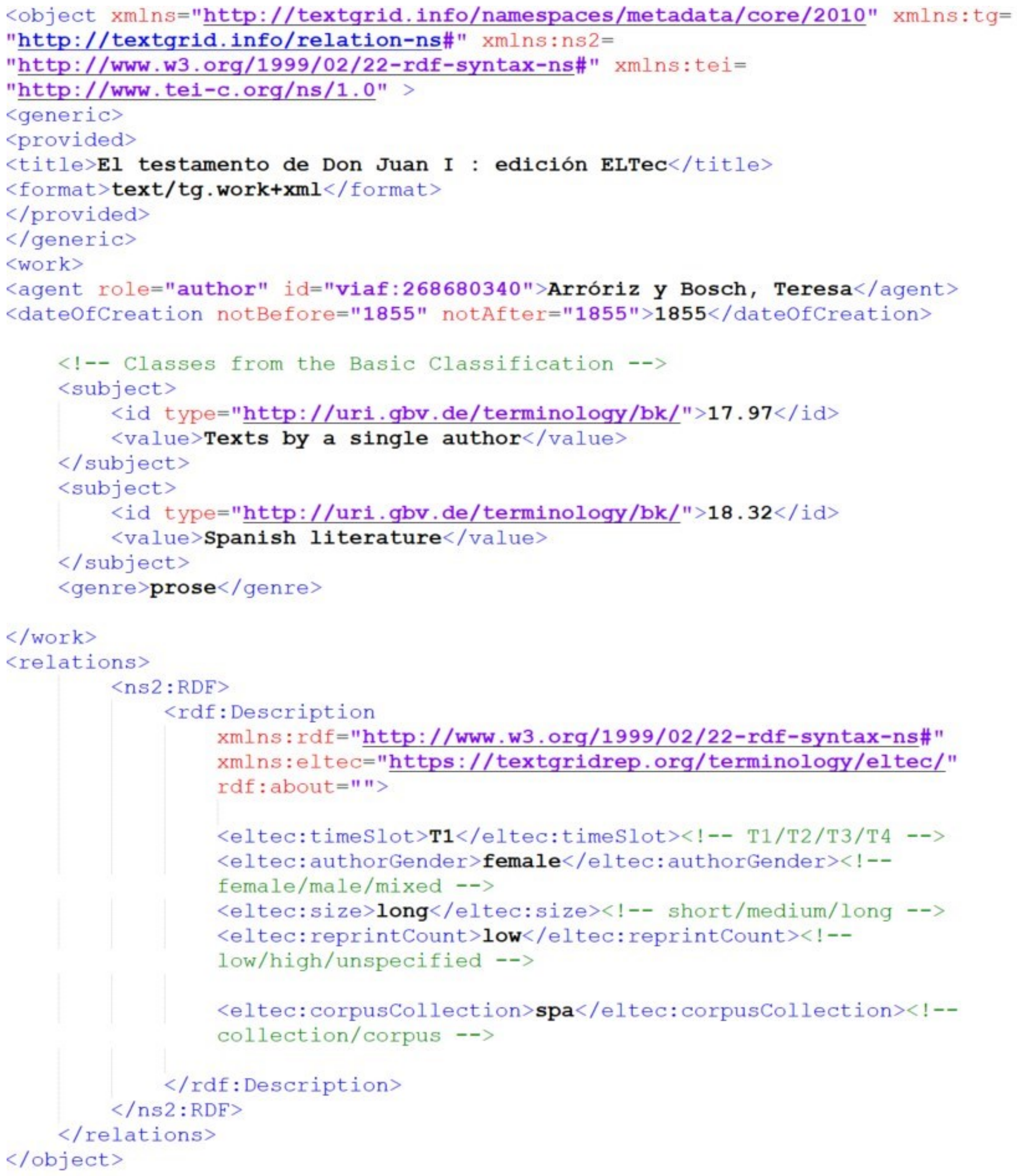
3ATGPR-99d098e9-b60f-98fd-cda3-6448e07e619d.

[27] See DARIAH-DE Terms of use: http://dx.doi.org/10.20375/0000-000B-CB44-4.

[28] See: https://textgridrep.org/docs/shelf?lang=en.

[29] See: https://dev.textgridlab.org/doc/services/submodules/kolibri/kolibri-addon-textgrid-import/docs/

several options for querying TGR using metadata. For the ELTeC corpora, these options have been further developed so that each project can now define project-specific metadata[29]. This metadata must be included in the TGR metadata file for the work. In this file, the element *relations* can contain triples expressed in RDF which is serialised in XML. The following example corresponds to the complete metadata file of the work *El testamento de Don Juan I* in the Spanish ELTeC corpus:



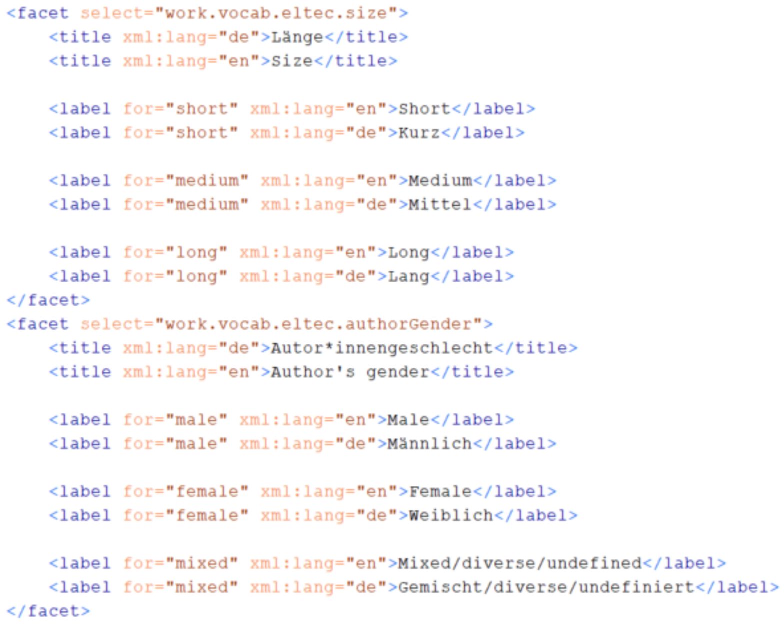
```
<object xmlns="http://textgrid.info/namespaces/metadata/core/2010" xmlns:tg=
"http://textgrid.info/relation-ns#" xmlns:ns2=
"http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:tei=
"http://www.tei-c.org/ns/1.0" >
<generic>
<provided>
<title>El testamento de Don Juan I : edición ELTec</title>
<format>text/tg.work+xml</format>
</provided>
</generic>
<work>
<agent role="author" id="viaf:268680340">Arróriz y Bosch, Teresa</agent>
<dateOfCreation notBefore="1855" notAfter="1855">1855</dateOfCreation>

    <!-- Classes from the Basic Classification -->
    <subject>
        <id type="http://uri.gbv.de/terminology/bk/">17.97</id>
        <value>Texts by a single author</value>
    </subject>
    <subject>
        <id type="http://uri.gbv.de/terminology/bk/">18.32</id>
        <value>Spanish literature</value>
    </subject>
    <genre>prose</genre>

</work>
<relations>
        <ns2:RDF>
            <rdf:Description
                xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
                xmlns:eltec="https://textgridrep.org/terminology/eltec/"
                rdf:about="">

                <eltec:timeSlot>T1</eltec:timeSlot><!-- T1/T2/T3/T4 -->
                <eltec:authorGender>female</eltec:authorGender><!--
                female/male/mixed -->
                <eltec:size>long</eltec:size><!-- short/medium/long -->
                <eltec:reprintCount>low</eltec:reprintCount><!--
                low/high/unspecified -->

                <eltec:corpusCollection>spa</eltec:corpusCollection><!--
                collection/corpus -->

            </rdf:Description>
        </ns2:RDF>
    </relations>
</object>
```

Figure 1. Metadata file of the work *El testamento de Don Juan* I. Source: Teresa Arróriz y Bosch, in the Spanish ELTeC.

These work-level metadata values need to be configured at the project level. This is done in an XML file, in which each project can define the project-specific categories, their possible values and labels to be displayed in both languages of the TGR portal: English and German.

import_and_configuration.html#project-specific-landing-page-and-project-metadata.

[30] See: https://textgridrep.org/projects.



```
<facet select="work.vocab.eltec.size">
    <title xml:lang="de">Länge</title>
    <title xml:lang="en">Size</title>

    <label for="short" xml:lang="en">Short</label>
    <label for="short" xml:lang="de">Kurz</label>

    <label for="medium" xml:lang="en">Medium</label>
    <label for="medium" xml:lang="de">Mittel</label>

    <label for="long" xml:lang="en">Long</label>
    <label for="long" xml:lang="de">Lang</label>
</facet>
<facet select="work.vocab.eltec.authorGender">
    <title xml:lang="de">Autor*innengeschlecht</title>
    <title xml:lang="en">Author's gender</title>

    <label for="male" xml:lang="en">Male</label>
    <label for="male" xml:lang="de">Männlich</label>

    <label for="female" xml:lang="en">Female</label>
    <label for="female" xml:lang="de">Weiblich</label>

    <label for="mixed" xml:lang="en">Mixed/diverse/undefined</label>
    <label for="mixed" xml:lang="de">Gemischt/diverse/undefiniert</label>
</facet>
```

Figure 2. Section of the XML file for the configuration of the project. Source: Teresa Arróriz y Bosch, in the Spanish ELTeC.

This file for the configuration of the project offers further options, such as the possibility of showing a logo of the project and a short description of the project. Both the logo and the short description are used in the list of projects of TGR[30].

One of the categories we have incorporated in various ways into CoNSSA and ELTeC is subject classification through the so-called Basic Classification. Subject classification systems are taxonomies of classes that libraries assign to publications (journals, books). There are several of them, such as the Dewey Decimal Classification (DDC), and others derived from it, such as the Universal Classification System or the Library of Congress Subject System (Gantert, 2016). In the German-speaking world, one of the most widely used classification systems is called Basic Classification (Balakrishnan and Voß, 2022; Schulz, 1991)[31]. With about 2,000 classes, it is a relatively small classification system compared to others with more than half a million classes or even with a potentially infinite number of classes (such as the DDC). One of the advantages of this classification system is that it is openly published, also expressed as a Linked Open Data resource. Assigning Basic Classification metadata to each work allows the user to use the hierarchical structure of the Basic Classification to query, for example, not only a language, but a group of languages. For example, users can now use a query to select all ELTeC corpora of Romance languages, or the corpora of the Ibero-Romance group[32].

Another option now available is to create a project page, using a simple Markdown formatted file[33]. Several projects have already made use of this option, including CoNSSA and ELTeC. This option can be used, for example, to link to the main portal of the project or to mention publications that present or analyze the texts published in TGR.

[31] See: https://wiki.k10plus.de/pages/viewpage.action?pageId=437452809.

[32] See: https://textgridrep.org/search?query=work.subject.id.value%3A%5B18.30+TO+18.38%5D&order=relevance&limit=20&filter=project.id%3ATGPR-99d098e9-b60f-98fd-cda3-6448e07e619d.

[33] See: https://dev.textgridlab.org/doc/services/submodules/kolibri/kolibri-addon-textgrid-import/docs/import_and_configuration.html#project-specific-landing-page-and-project-metadata.

A final new feature relating to the flexibility for projects for customizing their content is the ability to publish an XSLT transformation stylesheet for the transformation from TEI XML to HTML. In this way, projects can decide that the standard transformation of TGR would not be used to their data and apply a specific-project one, giving the projects the option to highlight or present differently specific elements of their textual codification.

These are all features that are already implemented in TGR, and projects are already using them. We would now like to present three more features that are currently under development. The first is the integration of data from Linked Open Data resources such as Wikidata or the GND into TGR. The aim is to enrich existing files with further identifiers for these databases in order to create links between resources that are not currently linked. The next step is to allow users to query TGR with metadata that was not originally included in the metadata files but is available in the authority files and the knowledge base. In this way, users who publish their corpora in TGR will automatically benefit from any further information in these other resources.

The second is the development of a Python library called TextGrid Python Clients or *tgclients*. This library is currently designed for querying TGR and accessing the data (metadata, text in plain text and TEI format) directly from Python. The library uses the already existing APIs to access the data and makes it simpler to use it in a Python environment. A GitLab repository[34], documentation and several Jupyter Notebooks as examples for its use are already available, one of them specific for the CoNSSA project[35].

Finally, we want to mention a new development currently being designed and potentially integrated in the *tgclients*[36]. Its goal is to create new import workflows for future edition and corpus projects that want to publish their texts in TGR. This process should help users to extract and validate their import metadata, evaluate it, support the completion of missing metadata, and create the several TGR metadata files (at the work, editions, and item files). This is currently under discussion with other project partners within Text+.

7. OUTREACH

In this article, we have not only introduced the TextGrid Repository and other components. Our main aim is to reflect on the way in which projects manage their resources and collaborate with other projects. For many projects, digital technologies are no longer secondary components. They have become core components of the research process. For this reason, we need to think about the long-term preservation of data and tools.

Academia relies on professional publishers (both corporate and associated with public institutions such as universities) to design and print its traditional publications (such as monographs and journals). An individual research project is not expected to print its own monographs, but to rely on professional publishers. Similarly, projects should not be expected to publish their data on

[34] See: https://gitlab.gwdg.de/dariah-de/textgridrep/textgrid-python-clients.
[35] See: https://gitlab.gwdg.de/jose.calvotello/hispanistentag_2023_tgr.
[36] See: https://dariah-de.pages.gwdg.de/textgridrep/textgrid-python-clents/docs/index.html.

the web in their own portal, but to rely on repositories with a long-term plan for their data. These repositories should be supported by institutions whose primary role is to ensure long-term access to the data, such as libraries and computing centres.

For these reasons, we invite projects to change their digital strategy from a single-project perspective to a distributed project perspective, relying on external services and institutions for specific services, especially the publication of data. The way in which the components of the different agents are connected should be through standard technologies within the community (such as TEI XML or RDF) and controlled vocabularies (such as authority files, knowledge bases such as Wikidata, or library classification systems). The resources of this distributed project perspective will be better integrated if they are open and oriented towards the FAIR criteria.

Digital tools and methods come and go very quickly. Very few of the tools that were used in DH twenty years ago are still in use today. Digital editions and corpora can last much longer and can be reused by other projects several decades after their initial digital publication. Because it is multilingual and for free, the TextGrid repository can be the place for many projects to publish their corpora and editions in TEI XML. While the TextGrid repository has, and probably will continue to have, an understandable focus on German literature, with the publication of CoNSSA and ELTeC, Spanish is currently overrepresented compared to other European languages. This can be seen as an opportunity for the publication of further corpora, their use in teaching and workshops, or their integration into other tools. They will benefit from its current features, but also from future developments.

BIBLIOGRAPHIC REFERENCES

Allés-Torrent, S., & Riande, G. del R. (2019). The Switchover: Teaching and Learning the Text Encoding Initiative in Spanish. *Journal of the Text Encoding Initiative, Issue 12*. https://doi.org/10.4000/jtei.2994

Aschenbrenner, A. (2015). Share It-Kollaboratives Arbeiten in TextGrid. In Heike Neuroth, Andrea Rapp, & Sibylle Söring (Eds.), *TextGrid: Von der Community-Für die Community* (pp. 201-209). Verlag Werner Hülsbusch. https://publications.goettingen-research-online.de/handle/2/14388

Balakrishnan, U., & Voß, J. (2022, May 31). Automatische Anreicherung der Sacherschließung des Verbundkatalogs K10plus mittels coli-rich. *#FreiräumeSchaffen*. 8. Bibliothekskongress. https://bid2022.abstractserver.com/program/#/details/presentations/27

Betz, K. (2015). Ein virtuelles Bücherregal: Die Digitale Bibliothek im TextGrid Repository. In Heike Neuroth, Andrea Rapp, & Sibylle Söring (Eds.), *TextGrid: Von der Community-Für die Community* (pp. 229-239). Verlag Werner Hülsbusch. https://publications.goettingen-research-online.de/handle/2/14388

Börner, I., & Trilcke, P. (2023). *CLS INFRA D7.1 On Programmable Corpora*. https://zenodo.org/record/7664964

Burnard, L. (2014). *What is the Text Encoding Initiative?: How to add intelligent markup to digital resources*. OpenEdition Press. http://books.openedition.org/oep/426

Burnard, L., Schöch, C., & Carolin Odebrecht. (2021). In search of comity: TEI for distant reading. *Journal of the Text Encoding Initiative, Issue 14*. https://doi.org/10.4000/jtei.3500

Calvo Tello, J. (2021a). Corpus de novelas de la Edad de Plata, en XML-TEI. *Signa: Revista de la Asociación Española de Semiótica*, *30*(0). https://doi.org/10.5944/signa.vol30.2021.29299

Calvo Tello, J. (2021b). *The Novel in the Spanish Silver Age: A Digital Analysis of Genre Using Machine Learning*. transcript.

Chambers, S. (Ed.). (2013). *Catalogue 2.0: The future of the library catalogue*. Facet Publishing.

CoreTrustSeal Standards and Certification Board. (2019). *CoreTrustSeal Trustworthy Data Repositories Requirements: Extended Guidance 2020-2022*. https://doi.org/10.5281/zenodo.3632533

del Rio Riande, G., & Allés-Torrent, S. (2023). ¿Quién conforma la comunidad de la TEI en español? Análisis de los datos de una encuesta. *Journal of the Text Encoding Initiative*, *16*. https://doi.org/10.4000/jtei.4927

Ermolaev, N., Munson, R., Li, X., Siemens, L., Siemens, R., Kaufman, M., & Boyd, J. (2018). Project Management for The Digital Humanities. *DH2018*. https://dh2018.adho.org/en/project-management-for-the-digital-humanities/

Ertler, K.-D. (2014). Die Gattung der frankophonen ‘Spectators’ im Spiegel der zeitgenössischen Medienrevolution. In *Literaturwissenschaft im digitalen Medienwandel* (pp. 18-35). web.fu-berlin.de/phin/beiheft7/b7t01.pdf

Fischer, F., Börner, I., Göbel, M., Hechtl, A., Kittel, C., Milling, C., & Trilcke, P. (2019). Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. *Proceedings of DH2019: ‘Complexities’, Utrecht, July 9-12, 2019*. https://doi.org/10.5281/zenodo.4284002

Funk, S. E., & Pempe, W. (2015). Vom Konzept zur Umsetzung—Einblicke in die Entstehung des TextGrid Repository. In Heike Neuroth, Andrea Rapp, & Sibylle Söring (Eds.), *TextGrid: Von der Community—Für die Community* (pp. 191-200). Verlag Werner Hülsbusch. https://publications.goettingen-research-online.de/handle/2/14388

Gantert, K. (2016). Bibliothekarisches Grundwissen. In *Bibliothekarisches Grundwissen*. De Gruyter Saur. https://www.degruyter.com/view/title/302969

González-Blanco, E., Cantón, C. M., del Rio Riande, G., Ros, S., Pastor, R., Robles-Gómez, A., Caminero, A., Díez Platas, M. L., del Olmo, Á., & Urízar, M. (2017). EVI-LINHD, a virtual research environment for the Spanish-speaking community. *Digital Scholarship in the Humanities*, *32* (suppl. 2), ii171-ii178. https://doi.org/10.1093/llc/fqx025

Herrmann, J. B., & Lauer, G. (2018). Korpusliteraturwissenschaft. Zur Konzeption und Praxis am Beispiel eines Korpus zur literarischen Moderne. *Osnabrücker Beiträge zur Sprachtheorie*, *92*, 127-156.

Hinrichs, E., Leinen, P., Geyken, A., Speer, A., & Stein, R. (2022). *Text+: Language- and text-based Research Data Infrastructure*. https://doi.org/10.5281/zenodo.6452002

Kampkaspar, D. (2017). Deutsches Textarchiv. *RIDE*, *6*. https://doi.org/10.18716/ride.a.6.8

Kett, J., Kudella, C., Rapp, A., Stein, R., & Trippel, T. (2022). Text+ und die GND - Community-Hub und Wissensgraph. *Zeitschrift Für Bibliothekswesen Und Bibliographie*, *69*(1-2), 37-47. https://doi.org/10.3196/1864295020691262

Kraft, S., Schmalen, A., Seitz-Moskaliuk, H., Sure-Vetter, Y., Knebes, J., Lübke, E., & Wössner, E. (2021). Nationale Forschungsdateninfrastruktur (NFDI) e. V.: Aufbau und Ziele. *Bausteine Forschungsdatenmanagement*, *2*. https://doi.org/10.17192/bfdm.2021.2.8332

Kränzle, A., Ritter, G., & Sieber, C. (2023). Sources Online: Eine nachhaltige Infrastruktur für digitale wissenschaftliche Texteditionen auf der Grundlage von TEI Publisher und IIIF. *ABI Technik*, *43*(3), 158-167. https://doi.org/10.1515/abitech-2023-0030

Patton, G. E. (Ed.). (2013). *Functional requirements for authority data: A conceptual model*. Saur. http://www.ifla.org/files/assets/cataloguing/frad/frad_2013.pdf

Retter, A. (2014). *eXist: A NoSQL document database and application platform* (First edition). O'Reilly.

Rißler-Pipka, N., Calvo Tello, J., Funk, S. E., Odebrecht, C., Schöch, C., & Veentjer, U. (2023). The European Literary Text Collection in TextGrid Repository. In W. Scholger, G. Vogeler, T. Tasovac, A. Baillot, E. Raunig, M. Scholger, E. Steiner, & P. Helling (Eds.), *Collaboration as Opportunity*. ADHO. https://doi.org/10.5281/ZENODO.8107707

Schöch, C., Erjavec, T., Patras, R., & Santos, D. (2021). Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives. *Modern Languages Open*, *1*. https://doi.org/10.3828/mlo.v0i0.364

Schulz, U. (1991). Die niederländische Basisklassifikation: Eine Alternative für die 'Sachgruppen' im Fremddatenangebot der Deutschen Bibliothek. *Bibliotheksdienst*, *25*, 1196-1219.

Sinclair, S., & Rockwell, G. (2016). *Voyant Tools* [Computer software]. http://voyant-tools.org/

Tabak, E. (2017). A hybrid model for managing dh projects. *Digital Humanities Quarterly*, *011*(1). http://www.digitalhumanities.org/dhq/vol/11/1/000284/000284.html

Weimer, L. (2022, July 15). Verein „Geistes- und kulturwissenschaftliche Forschungsinfrastrukturen" geht nächsten Schritt im Zuge der Vereinstransformation. *DHd-Blog*. https://dhd-blog.org/?p=18116

Wiegand, F., Thomas, C., Haaf, S., Geyken, A., Jurish, B., & Boenig, M. (2018). Recherchieren, Arbeiten und Publizieren im Deutschen Textarchiv: Ein Praxisbericht. *Zeitschrift Für*.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. 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., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3* (*1*), 160018. https://doi.org/10.1038/sdata.2016.18

Wilkinson, M. D., Sansone, S.-A., Schultes, E., Doorn, P., Bonino da Silva Santos, L. O., & Dumontier, M. (2018). A design framework and exemplar metrics for FAIRness. *Scientific Data*, *5*(1). https://doi.org/10.1038/sdata.2018.118

Zinn, C. (2018). The Language Resource Switchboard. *Computational Linguistics*, *44*(4), 631-639. https://doi.org/10.1162/coli_a_00329