# FAIR Derived Data in TEI and its Publication in the TextGrid Repository

## Authors

* José Calvo Tello, Göttingen State and University Library, [calvotello@sub.uni-goettingen.de](mailto:calvotello@sub.uni-goettingen.de), <https://orcid.org/0000-0002-1129-5604>
  + José Calvo Tello works as a researcher and subject librarian at the Göttingen State and University Library. He obtained his doctorate in Humanities from the University of Würzburg (Germany) with the title: The Novel in the Spanish Silver Age: A Digital Analysis of Genre Using Machine Learning (transcript, 2021). His research focuses on the application and development of statistical and computational methods applied to Romance Literatures and library records.
* Mathias Göbel, Göttingen State and University Library, [goebel@sub.uni-goettingen.de](mailto:goebel@sub.uni-goettingen.de), <https://orcid.org/0000-0002-1102-5284>
  + Mathias Göbel is a digital humanist and software engineer by training. His work is focused on processing text data, not exclusively TEI encoded, from small-scale digital editions to larger corpora. Struggling with invalid TEI is part of his job at Göttingen State and University Library.
* Ubbo Veentjer, Göttingen State and University Library, [veentjer@sub.uni-goettingen.de](mailto:veentjer@sub.uni-goettingen.de), <https://orcid.org/0000-0002-9726-3135>
  + Ubbo Veentjer is a software engineer and information scientist working at the Göttingen State and University Library. He has a master's degree in library and information science and has been involved in several research projects mainly in the area of the Digital Humanities.
* Stefan E. Funk, Göttingen State and University Library, [funk@sub.uni-goettingen.de](mailto:funk@sub.uni-goettingen.de), <https://orcid.org/0000-0003-1259-2288>
  + Stefan E. Funk is a library and information scientist with deep roots in computer sciences. He has been long working in the field of long-term archiving and is heavily involved in research data repositories for the humanities at the moment. He is working as a process manager for technology and software engineer and developer for the TextGrid and DARIAH-DE Repository in the Göttingen State and University Library.
* Nanette Rißler-Pipka, Max Weber Stiftung, [Rissler-Pipka@MaxWeberStiftung.de](mailto:Rissler-Pipka@MaxWeberStiftung.de), <https://orcid.org/0000-0002-0719-9003>
  + Nanette Rißler-Pipka is a digital humanist, literary scholar, specialist in French and Spanish literature. She is co-managing director of DARIAH-DE and National Coordinator of Germany for DARIAH-ERIC. As such she is also part of the coordination committee of the Association for Research Infrastructure in the Humanities and Cultural Studies (GKFI). She has a master’s degree in comparative literature, Romance languages and economics, Dr. phil. and habilitation in Romance literature (University of Siegen). At the headquarter of the Max Weber Foundation (German Humanities Institutes Abroad) she is currently working as Digital Humanities and Research Data Management Officer.
* Keli Du, Trier Center for Digital Humanities,, [duk@uni-trier.de](mailto:duk@uni-trier.de), <https://orcid.org/0000-0001-7800-0682>
  + Keli Du is a researcher at the Trier Center for Digital Humanities with a background in digital humanities who seeks collaborations with computer scientists, computational linguists, statisticians, and humanities scholars. His goal is to translate humanities research questions into data analytics problems, develop solutions to address those problems, as well as interpret the outcomes of the analysis.

## Abstract

Many projects face legal restrictions on the publication of texts. In recent decades, several projects have circumvented these restrictions by deleting some parts of the data and publishing derived data from the original files. We discuss the limitations of the commonly used ad hoc solutions and the deprecation of the FAIR status that they cause. In contrast, we propose to model derived data in TEI and present several variants with five corpora from different languages, genres and periods. We also present the implementation of several features for publishing such data in the TextGrid Repository and the publication of derived data from a corpus of Spanish novels and a corpus of American plays.

## Keywords

Derived Data, FAIR Principles, Repository, Literature, Publication, Legal Issues, Copyright

# 1. Derived Data

In recent years, many research projects and institutions have faced the problem that some of the data (among others, text in the form of collections or corpora) they collect cannot be published openly for a variety of reasons. One of the most important reasons is copyright restrictions, which in many countries expire 70 years after the death of the author.

In an attempt to get around this problem, many projects have already found their own solutions by releasing not all the data, but only some of it. One of the most famous cases is the HathiTrust Research Center Extracted Features Dataset (Jett et al. 2020). In this case, the data published in a JSON file can be traced back to each individual publication, with specific metadata and identifiers for each. Each JSON file contains elements for each page of the original publication, with metadata about each page and the frequency of the tokens that appear in each page. The specific order of the words on the page is not preserved, so the text is no longer readable. However, many methods, especially quantitative methods, can be applied to this type of data. Two other examples of the publication of data derived from copyrighted texts are the Google Books Ngrams (in which the information has been aggregated by language and year; see Lin et al. 2012)[[1]](#footnote-2) or the files of the CORDE corpus with metadata and frequencies that the Spanish *Real Academia de la Lengua* has made available in recent years (Sánchez Sánchez and Domínguez Cintas 2007). Many other projects have undertaken similar goals in the past years (Rehm et al. 2007; Goldhahn, Eckart, and Quasthoff 2012; Ortiz Suárez, Sagot, and Romary 2019).

Although with different motivations and backgrounds, library catalogs also provide some data about texts that cannot be published in digital form. For example, the number of pages in a book is a quantitative measure of the content of the publication. In other words, metadata in a library catalog can be already considered as derived data from copyrighted documents, and its publication is not seen as problematic by anyone.

Within the German NFDI (National Research Data Infrastructure) consortium Text+ (Kett et al. 2022; Hinrichs et al. 2022), providing solutions for derived data (often referred to as derived text formats, in German *abgeleitete Textformate*, see Schöch et al. 2020) is an important task. In the consortium, research teams from academies, universities and libraries (such as the German National Library and the Göttingen State and University Library) work together with the main goal of integrating already existing resources under consideration of the FAIR criteria for long-term preservation. The solutions and implementations in the TextGrid Repository (TextGridRep) presented here are part of the activities within this task (see Calvo Tello et al. 2023 for further new developments in TextGridRep).[[2]](#footnote-3)

# 2. FAIR Principles and TEI

The above mentioned examples, and others (see Schöch et al. 2020 for more examples), have used ad hoc solutions for modelling the data and the specific format. In many cases, projects use plain text files or spreadsheets to quantify the text (e.g. token frequency tables). In many cases, the metadata is either reduced to the name of the plain text file or stored in a separated spreadsheet.

Although easy to produce and easy to use for closely related projects, it is arguable that the quality of both data and metadata published in this form is problematic when considering the FAIR principles (Wilkinson et al. 2016). In many of the cases mentioned, neither the data nor the metadata can be automatically validated. The perspective that these derived data are products of secondary importance compared to the original files leads to decisions that affect the interoperability and reusability of the data. For example, the metadata tends to be sparse and do not use identifiers from other resources (such as Q-ids from Wikidata or authority files) or controlled vocabularies. Most of all, the metadata is not linked to the data via a persistent identifier or included in the structure of the data.

In comparison to the solutions used by other projects, we argue for TEI as a more appropriate format for modelling and publishing derived data, even though to our knowledge no previous project has done so. The FAIR criteria (Wilkinson et al. 2016) can be followed much more closely in TEI than in plain text files and spreadsheets, which is the most commonly used solution (Creamer et al. 2021; Gengnagel, Neuber, and Schulz 2023). For example, TEI provides specific elements for the various sub-criteria of the FAIR principles, in particular for findability, interoperability and reusability. In terms of findability, TEI offers specific elements for identifiers (F1), the ability to use rich metadata (F2), and to relate data and metadata in the same file (F4). In TEI-specific repositories such as TextGridRep, specific metadata fields can be indexed to make them findable (F3). In terms of interoperability, TEI is an accessible, shared and widely applicable language (I1), it can be argued that it itself follows the FAIR principles (I2), and it has specific elements and attributes for relating the data to other data (I3). Regarding reusability, TEI provides options for assigning a variety of relevant attributes (R1), such as licenses (R1.1), provenance (R1.2), and is, and allows for the integration of other community standards (R1.3).

Although metadata could also be integrated into spreadsheets, this would happen without a common framework (I1) having a negative impact on the interoperability. Some projects would argue that they prefer JSON (such as JSON-LD) to XML-TEI. We agree that in some current environments, such as Python, it is more convenient to manage data in JSON than in XML. However, ease of data handling is not a consideration in the FAIR principles. Moreover, the XML-TEI is better known in many of the communities working with humanities data than any specification of JSON (I1 and R1.3). A recent survey on metadata and the creation of literary corpora also asked about the preferred format for digital literary corpora (Calvo Tello et al. 2023). The respondents show that TEI is much more widely accepted than JSON.

# 3. Corpora

For the exploration of possibilities presented in the section, it is important for us to consider solutions that fit heterogeneous real data. Therefore, we use five existing corpora which cover different periods of literature, languages and genres, collected by different projects. Here we present the main characteristics of each corpus:

* **Gutenberg.de**: the German clone of “Project Gutenberg”, with mostly literary texts after copyright expires.
* **American Drama 1714-1915 (CD-ROM)**: A total of 1558 plays written by North American dramatists,provided on CD-ROM. The corpus covers a broad range of subgenres from drawing room plays, one-acts and monologues to extensive tragedies. Since the texts are delivered as a pre-compiled corpus on CD-ROM, intellectual property rights might be claimed by the curator or publisher. A minor part of the collection is still covered by copyright terms based on authors' death, but the database itself might form some intellectual property.
* **Spanish CoNSSA corpus:** The original CoNSSA corpus (Corpus of Novels of the Spanish Silver Age) contained 358 prose texts by Spanish authors published for the first time between 1880 and 1939 (Calvo Tello 2021), with many texts still under copyright. Part of the corpus has been made available in the TextGridRep and in Zenodo (Calvo Tello et al. 2023), which includes derived data from the original files, containing the frequency of tokens as bag-of-words models and metadata in spreadsheets.
* **French corpus**: The French corpus is a balanced collection of 320 French novels from 1980s and 1990s. The corpus consists of novels of four different subgenres: crime fiction, science fiction, romance novel and high-brow literature, with 40 novels from each decade in each genre. Since the texts are all contemporary novels, they are copyright protected. However, the texts were transformed in a derived text format and were made available in GitHub with detailed metadata[[3]](#footnote-4). In addition to the tokens in texts, the corresponding POS-tags and lemma of each token word are also included in the corpus.
* **Chinese corpus**: The Chinese corpus contains 158 texts, written between 1918 and 1936 by Lu Xun, who was a leading figure of modern Chinese literature. The corpus consists of 33 novels and 125 essays with a total of more than 430,000 words. Because the author died in 1936, his works are in the public domain. The original plain texts are available online, for example in Wikisource (in traditional Chinese)[[4]](#footnote-5) or in GitHub (in simplified Chinese)[[5]](#footnote-6) with detailed metadata.

Because almost all the corpora face legal restrictions about its publication, we only use a single text of each corpus for the possible solutions of the next section.

# 4. Variants for Modelling Derived Data in TEI

## 4.1. Quantification of Elements

A first step in publishing some of the content of a document might be to quantify some encoded aspects. This information could be placed in TEI in <measure> elements within a <measureGrp> element, which would make explicit through an @corresp attribute the textual unit being quantified, such as the entire document or smaller units, such as chapters or acts, paragraph-like elements, sentences, etc.

In the following excerpts, we show each time a section of the data of the five corpora to address differences in languages, periods and genres.[[6]](#footnote-7) In Example 1 we have selected some frequent TEI elements. In Section 5b we will demonstrate this approach with the American Drama corpus.

<extent><!--German corpus-->

<measureGrp type="elements">

<measureGrp corresp="#gb14397">

<measure commodity="div" quantity="16"/>

<measure commodity="sp" quantity="0"/>

<measure commodity="p" quantity="0"/>

<measure commodity="l" quantity="0"/>

<measure commodity="w" quantity="54491"/>

<measure commodity="pc" quantity="10814"/>

</measureGrp>

</measureGrp>

</extent>

<extent><!--English corpus-->

<measureGrp type="elements">

<measureGrp corresp="#amdram1">

<measure commodity="div" quantity="7"/>

<measure commodity="sp" quantity="334"/>

<measure commodity="p" quantity="338"/>

<measure commodity="l" quantity="0"/>

<measure commodity="w" quantity="8582"/>

<measure commodity="pc" quantity="2098"/>

</measureGrp>

</measureGrp>

</extent>

<extent><!--Spanish corpus-->

<measureGrp type="elements">

<measureGrp corresp="#ne0120">

<measure commodity="div" quantity="8"/>

<measure commodity="sp" quantity="0"/>

<measure commodity="p" quantity="354"/>

<measure commodity="l" quantity="4"/>

<measure commodity="w" quantity="30477"/>

<measure commodity="pc" quantity="4672"/>

</measureGrp>

</measureGrp>

</extent>

<extent><!--French corpus-->

<measureGrp type="elements">

<measureGrp corresp="#Aarons\_Attila">

<measure commodity="div" quantity="13"/>

<measure commodity="sp" quantity="0"/>

<measure commodity="p" quantity="683"/>

<measure commodity="l" quantity="0"/>

<measure commodity="w" quantity="17823"/>

<measure commodity="pc" quantity="3166"/>

</measureGrp>

</measureGrp>

</extent>

<extent><!--Chinese corpus-->

<measureGrp type="elements">

<measureGrp corresp="#Luxun\_Kuangren">

<measure commodity="div" quantity="14"/>

<measure commodity="sp" quantity="0"/>

<measure commodity="p" quantity="76"/>

<measure commodity="l" quantity="0"/>

<measure commodity="w" quantity="2780"/>

<measure commodity="pc" quantity="710"/>

</measureGrp>

</measureGrp>

</extent>

Example 1: Quantification through frequency of certain aspects of the text (such as sections, speeches, paragraphs, lines, words, and punctuation) encoded in TEI

In Example 1, the data is reduced to a level that the majority of actors would not even consider this derived data, or that it could be legally problematic. In fact, we are used to finding similar measures in repositories or library catalogs, even when data is not available, such as the number of pages in a book. In this way, derived data can be seen on a continuum with metadata, whose publication is clearly seen as unproblematic and used as research data (Fischer and Jäschke 2022; Calvo Tello 2022).

## 4.2. Bag-of-words Model at Document Level

A frequently seen method by many projects is to split the entire text into tokens and count each different token (i.e. type) in the whole document, generating a bag-of-word model at document level. The frequencies of the types can then be placed in the <teiHeader>, e.g. each type in a different <measure> element within a <measureGrp> element with a @corresp attribute with the identifier of the document.

<extent><!--German corpus-->

<measureGrp type="form">

<measureGrp corresp="#gb14397">

<measure commodity="," quantity="4511"/>

<measure commodity="." quantity="2473"/>

<measure commodity="und" quantity="1331"/>

<measure commodity="die" quantity="1294"/>

<!-- … -->

<measureGrp>

<measureGrp>

</extent>

<extent><!--English corpus-->

<measureGrp type="form">

<measureGrp corresp="#amdram1">

<measure commodity="." quantity="688"/>

<measure commodity="," quantity="660"/>

<measure commodity="I" quantity="324"/>

<measure commodity="!" quantity="231"/>

<!-- … -->

<measureGrp>

<measureGrp>

</extent>

<extent><!--Spanish corpus-->

<measureGrp type="form">

<measureGrp corresp="#ne0120">

<measure commodity="," quantity="2414"/>

<measure commodity="de" quantity="1503"/>

<measure commodity="que" quantity="1312"/>

<measure commodity="y" quantity="1086"/>

<!-- … -->

<measureGrp>

<measureGrp>

</extent>

<extent><!--French corpus-->

<measureGrp type="form">

<measureGrp corresp="#Aarons\_Attila">

<measure commodity="." quantity="1441"/>

<measure commodity="," quantity="1028"/>

<measure commodity="de" quantity="581"/>

<measure commodity="—" quantity="476"/>

<!-- … -->

<measureGrp>

<measureGrp>

</extent>

<extent><!--Chinese corpus-->

<measureGrp type="form">

<measureGrp corresp="#Luxun\_Kuangren">

<measure commodity="，" quantity="366"/>

<measure commodity="。" quantity="155"/>

<measure commodity="的" quantity="137"/>

<measure commodity="我" quantity="119"/>

<!-- … -->

<measureGrp>

<measureGrp>

</extent>

Example 2: Bag-of-words models at the document level

TEI files containing metadata and token frequencies could provide all the necessary data for many research questions in the humanities, especially in the digital humanities, as we will demonstrate in Section 5a with Voyant Tools and stylometric methods. To a certain degree, this would also solve the problem of poor OCR results in quantitative analysis.

This variant could be combined with the one presented in 4.1 through multiple <measureGrp> elements. In addition, frequencies about linguistically annotated data (such as lemmata, part-of-speech, or named-entity recognition information) can be added.

## 4.3. Bag-of-Words Model at Chapter or Paragraph Level

Instead of creating the bag-of-words models at the document level, they can be created at smaller textual units. For example, if a file is structured in pages, the frequencies of tokens per page could be counted. In fact, this is exactly the information that can be found in the HathiTrust Research Center Extracted Features Dataset (Jett et al. 2020), with the only difference that in this case it is expressed in JSON-LD.

One of the most important decisions in this variant is the textual unit selected for its quantification. For example, selecting the <body> element would make almost no difference compared to the variant presented in 4.2. Selecting text units much deeper in the text structure, such as sentences or even phrases, would in many cases create such small bags of words that the reconstruction of the original text would be trivial. We think a reasonable compromise is to consider <div> or paragraph-like elements (<p>, <ab>, <l>, <stage>, etc.). These elements would require an @xml:id attribute and use this identifier in the @corresp attribute in the corresponding <measureGrp> element. This step is a significant increase in information compared to the bag of words at the document level presented in 4.2.

<measureGrp type="form"><!--German corpus-->

<measureGrp corresp="#gb14397\_chap1\_ab1">

<measure commodity="die" quantity="2"/>

<measure commodity="," quantity="2"/>

<measure commodity="Als" quantity="1"/>

<measure commodity="ihm" quantity="1"/>

<!-- … -->

<measureGrp>

<measureGrp>

<measureGrp type="form"><!--English corpus-->

<measureGrp corresp="#emdram1\_act1\_stage3">

<measure commodity="." quantity="2"/>

<measure commodity="SCENE" quantity="1"/>

<measure commodity="—" quantity="1"/>

<measure commodity="Drawing" quantity="1"/>

<!-- … -->

<measureGrp>

<measureGrp>

<measureGrp type="form"><!--Spanish corpus-->

<measureGrp corresp="#ne0120\_chap1\_p1">

<measure commodity="," quantity="6"/>

<measure commodity="en" quantity="4"/>

<measure commodity="de" quantity="3"/>

<measure commodity="puedo" quantity="2"/>

<!-- … -->

<measureGrp>

<measureGrp>

<measureGrp type="form"><!--French corpus-->

<measureGrp corresp="#Aarons\_Attila\_ch01\_p1">

<measure commodity="," quantity="6"/>

<measure commodity="." quantity="4"/>

<measure commodity="son" quantity="3"/>

<measure commodity="de" quantity="2"/>

<!-- … -->

<measureGrp>

<measureGrp>

<measureGrp type="form"><!--Chinese corpus-->

<measureGrp corresp="#Luxun\_Kuangren\_ch00\_p0">

<measure commodity="，" quantity="23"/>

<measure commodity="。" quantity="10"/>

<measure commodity="；" quantity="4"/>

<measure commodity="一" quantity="4"/>

<measure commodity="所" quantity="4"/>

<!-- … -->

<measureGrp>

<measureGrp>

Example 3: Bag-of-words models at paragraph-like elements level

## 4.4. Empty Textual Structure

Even when certain TEI elements in the 4.1 variant are counted, the actual tree structure would be erased. For example, it wouldn't be possible to answer whether the first and last scenes of a play tend to contain more stages than the scenes in the middle.

A simple way to preserve this information without raising legal concerns would be to publish the entire TEI structure hanging from the <text> element, but as empty elements, with no actual text in them, as shown in Example 4.

<body xml:lang="de"><!--German corpus-->

<div type="chapter" xml:id="gb14397\_chap1">

<head xml:id="gb14397\_chap1\_head0">I. Hauptmann Bitterlin</head>

<ab xml:id="gb14397\_chap1\_ab1"/>

<ab xml:id="gb14397\_chap1\_ab2"/>

<ab xml:id="gb14397\_chap1\_ab3"/>

<ab xml:id="gb14397\_chap1\_ab4"/>

<!-- … -->

</div>

</body>

<body xml:lang="en"><!--English corpus-->

<div type="act" n="1" xml:id="emdram1\_act1">

<head xml:id="emdram1\_act1\_head0">ACT I.</head>

<div type="scene" xml:id="emdram1\_act1\_div1">

<head xml:id="emdram1\_act1\_head2">Scene</head>

<stage xml:id="emdram1\_act1\_stage3"/>

<stage xml:id="emdram1\_act1\_stage4"/>

<sp who="#mrclover" xml:id="emdram1\_act1\_sp5">

<speaker xml:id="emdram1\_act1\_speaker6">Clo.</speaker>

<stage xml:id="emdram1\_act1\_stage7"/>

<p xml:id="emdram1\_act1\_p8"/>

</sp>

<sp who="#agnesclover" xml:id="emdram1\_act1\_sp9">

<speaker xml:id="emdram1\_act1\_speaker10">Agnes.</speaker>

<p xml:id="emdram1\_act1\_p11"/>

</sp>

<!-- … -->

</div>

</div>

</body>

<body xml:lang="es"><!--Spanish corpus-->

<div n="1" type="chapter" xml:id="ne0120\_chap1">

<head xml:id="ne0120\_chap1\_head0">Capítulo I</head>

<p xml:id="ne0120\_chap1\_p1"/>

<p xml:id="ne0120\_chap1\_p2"/>

<p xml:id="ne0120\_chap1\_p3"/>

<p xml:id="ne0120\_chap1\_p4"/>

<p xml:id="ne0120\_chap1\_p5"><seg rend="italic" xml:id="ne0120\_chap1\_seg6"></seg></p>

<!-- … -->

</div>

</body>

<body xml:lang="fr"><!--French corpus-->

<div type="chapter" xml:id="Aarons\_Attila\_ch01">

<head xml:id="Aarons\_Attila\_ch01\_head0">I</head>

<p xml:id="Aarons\_Attila\_ch01\_p1"/>

<p xml:id="Aarons\_Attila\_ch01\_p2"/>

<p xml:id="Aarons\_Attila\_ch01\_p3"/>

<p xml:id="Aarons\_Attila\_ch01\_p4"/>

<p xml:id="Aarons\_Attila\_ch01\_p5"/>

<!-- … -->

</div>

</body>

<body xml:lang="zh"><!--Chinese corpus-->

<div type="chapter" xml:id="Luxun\_Kuangren\_ch00">

<p xml:id="Luxun\_Kuangren\_ch00\_p0"/>

</div>

<div type="chapter" xml:id="Luxun\_Kuangren\_ch01">

<head xml:id="Luxun\_Kuangren\_ch01\_head0">一</head>

<p xml:id="Luxun\_Kuangren\_ch01\_p1"/>

<p xml:id="Luxun\_Kuangren\_ch01\_p2"/>

<p xml:id="Luxun\_Kuangren\_ch01\_p3"/>

</div>

<!-- … -->

</body>

Example 4: Descendant elements of <body> without any text

In Example 4, we have only retained the actual text of the elements in the <head> elements. The texts in these elements make up the file's table of contents, which is often published (as metadata or digitized as a PDF) by libraries and publishers. In some genres, such as poetry or short stories, the text in these elements could be used to identify specific literary works. In Section 5, it will be shown how TextGridRep uses this information.

In some specific genres, it could be argued that the text of certain elements should be retained, especially in cases where the text is very short. This could be the case for plays, for example, with the <speaker> elements (which can be seen in Example 4) or the dramatis personae.

This variant of derived data could be combined with some of the previously seen variants. For example, the empty TEI elements could be enriched with statistical values in measure elements (as seen in Section 4.1) that quantify the number of words or characters in each element, as we will show in Section 5b.

## 4.5. Randomization of Tokens within Element

The combination of variants 4.3 and 4.4 in a single file, containing both the TEI elements and the frequencies of all the tokens but not in their original order, would preserve much of the original information. However, this combination would make these files much more difficult to work with. Tools that accept TEI as an input format (such as stylo or Voyant Tools) would not be able to work with these files without ad-hoc processing. Although our goal is to increase the reusability of the data, with this specific combination we are also degrading the reusability of the tools and thus the data.

An improvement to this combination would be to preserve both the out-of-order tokens and the textual structure expressed in the TEI elements, and still be able to work with the files in a user-friendly way. To achieve this, we propose to randomize the tokens (including punctuation) within the paragraph-like elements. Before the randomization takes place, the text contained in each of these elements could be annotated, and this additional data could also be placed in the TEI document. In Example 5, we show one way of adding this information as inline annotation via attributes; more annotation could be retained using other strategies, such as with the <standOff> element.

<div type="chapter" xml:id="gb14397\_chap1"><!--German corpus-->

<head xml:id="gb14397\_chap1\_head0">I. Hauptmann Bitterlin</head>

<ab xml:id="gb14397\_chap1\_ab1">

<w lemma="wie" pos="ADP">wie</w>

<w lemma="eckig" pos="ADJ">eckigen</w>

<w lemma="ich" pos="PRON">er</w>

<w lemma="der" pos="PRON">die</w>

<w lemma="Schriftzügen" pos="NOUN">Schriftzügen</w>

<pc lemma="," pos="PUNCT">,</pc>

<!-- ... -->

</ab>

</div>

<div type="act" n="1" xml:id="emdram1\_act1"><!--English corpus-->

<head xml:id="emdram1\_act1\_head0">ACT I.</head>

<div type="scene" xml:id="emdram1\_act1\_div1">

<head xml:id="emdram1\_act1\_head2">Scene</head>

<stage xml:id="emdram1\_act1\_stage3">

<w lemma="in" pos="ADP">in</w>

<w lemma="'s" pos="PART">'s</w>

<w lemma="London" pos="PROPN">London</w>

<w lemma="at" pos="ADP">at</w>

<!-- ... -->

</stage>

</div>

</div>

<div n="1" type="chapter" xml:id="ne0120\_chap1"><!--Spanish corpus-->

<head xml:id="ne0120\_chap1\_head0">Capítulo I</head>

<p xml:id="ne0120\_chap1\_p1">

<w lemma="misterio" pos="NOUN">misterio</w>

<w lemma="alma" pos="NOUN">alma</w>

<w lemma="verdadero" pos="ADJ">verdadera</w>

<w lemma="abacial" pos="ADJ">abacial</w>

<w lemma="calmar" pos="VERB">calmar</w>

<w lemma="yo" pos="PRON">me</w>

<!-- ... -->

</p>

</div>

<div type="chapter" xml:id="Aarons\_Attila\_ch01"><!--French corpus-->

<head xml:id="Aarons\_Attila\_ch01\_head0">I</head>

<p xml:id="Aarons\_Attila\_ch01\_p1">

<w lemma="médecin" pos="NOUN">médecin</w>

<pc lemma="," pos="PUNCT">,</pc>

<w lemma="luisant" pos="ADJ">luisantes</w>

<w lemma="possible" pos="ADJ">possible</w>

<w lemma="à" pos="ADP">à</w>

<w lemma="Stein" pos="PROPN">Stein</w>

<!-- ... -->

</p>

</div>

<div type="chapter" xml:id="Luxun\_Kuangren\_ch01"><!--Chinese corpus-->

<head xml:id="Luxun\_Kuangren\_ch01\_head0">一</head>

<p xml:id="Luxun\_Kuangren\_ch01\_p1">

<w pos="NOUN">晚上</w>

<pc pos="PUNCT">。</pc>

<w pos="VERB">狠</w>

<w pos="ADJ">好的</w>

<w pos="NOUN">月光</w>

<w pos="NOUN">今天</w>

<pc pos="PUNCT">，</pc>

<!-- ... -->

</p>

</div>

Example 5: Randomization of tokens within paragraph-like elements

The amount of information in this variant is almost identical to the information in the combination of 4.3 and 4.4. In both cases, the presence of the tokens at specific TEI elements and the entire textual structure are preserved, without the original order of the tokens. The advantage of this variant is that the files can be analyzed by many tools. However, not all the analyses offered by these tools would be appropriate with these files, even though they are possible. Any attempt to account for the actual order of tokens within paragraph-like elements would only produce noise, such as n-gram extraction. Nevertheless, other analyses that consider a flexible distance within the paragraph-like elements could be appropriate.

This randomization process could be applied in other ways. In this variant, we keep the paragraph-like element as a container and randomize the tokens as descendant elements. Ortiz Suárez, Sagot, and Romary decided to randomize the tokens in each file (2019), which could be expressed in TEI taking the <text> as a container, deleting the TEI-elements hanging from it and then randomize the tokens. Another strategy would be to choose the chapters as containers and randomize the order of other descendant elements, such as paragraphs, or even go a step further with sentences. Although each paragraph or sentence could be read and understood, the chapters, and in general the work as a whole, could not be understood, or at least not in the same way. Maintaining sentences as a whole could support other use-cases. However, unlike the rest of the case, retaining entire sentences would preserve much of the original order of the tokens and give readers the sense that they can read and understand most of the text, and therefore might raise more legal concerns.

## 4.6. Randomization of Fixed N-Grams

Many current computational methods applied to text use the distributional information of the original order of the tokens (Blei 2012; Devlin et al. 2019; Jawanpuria et al. 2019). In this way, we would like to obtain derived data in which a part of the distributional information is preserved without legal concerns.

In this variant, after linguistic annotation of the original text, a preprocessing step would generate n-grams with the tokens of certain container elements, for example <p>. The order of the tokens within the n-grams is kept fixed in a new TEI element, for example <seg>. The order of the <seg> elements within the container element (<p>) would then be randomized. These steps will still spoil the majority of the distributional information and give the impression that the text is unintelligible. However, some distributional methods could still use the original order within the n-grams in the <seg> elements.

<div type="chapter" xml:id="gb14397\_chap1"><!--German corpus-->

<head xml:id="gb14397\_chap1\_head0">I. Hauptmann Bitterlin</head>

<ab xml:id="gb14397\_chap1\_ab1">

<seg><w lemma="Quartierliste" pos="NOUN">Quartierliste</w><w lemma="der" pos="DET">der</w><w lemma="Armee" pos="NOUN">Armee</w></seg>

<seg><w lemma="schreiben" pos="VERB">schrieb</w><w lemma="ich" pos="PRON">er</w><w lemma="eigenhändig" pos="ADV">eigenhändig</w></seg>

<seg><w lemma="als" pos="SCONJ">Als</w><w lemma="ich" pos="PRON">ihm</w><w lemma="der" pos="DET">die</w></seg>

<seg><w lemma="vorlegen" pos="VERB">vorgelegt</w><w lemma="werden" pos="AUX">wurde</w><pc lemma="," pos="PUNCT">,</pc></seg>

<!-- ... -->

</ab>

</div>

<div type="act" n="1" xml:id="emdram1\_act1"><!--English corpus-->

<head xml:id="emdram1\_act1\_head0">ACT I.</head>

<div type="scene" xml:id="emdram1\_act1\_div1">

<head xml:id="emdram1\_act1\_head2">Scene</head>

<stage xml:id="emdram1\_act1\_stage3">

<seg><w lemma="at" pos="ADP">at</w><w lemma="London" pos="PROPN">London</w><pc lemma="." pos="PUNCT">.</pc></seg>

<seg><w lemma="'s" pos="PART">'s</w><w lemma="House" pos="PROPN">House</w><pc lemma="," pos="PUNCT">,</pc></seg>

<seg><w lemma="in" pos="ADP">in</w><w lemma="Mr." pos="PROPN">Mr.</w><w lemma="Clover" pos="PROPN">Clover</w></seg>

<seg><w lemma="SCENE" pos="PROPN">SCENE</w><pc lemma="." pos="PUNCT">.</pc><pc lemma="—" pos="PUNCT">—</pc></seg>

<!-- ... -->

</div>

</div>

<div n="1" type="chapter" xml:id="ne0120\_chap1"><!--Spanish corpus-->

<head xml:id="ne0120\_chap1\_head0">Capítulo I</head>

<p xml:id="ne0120\_chap1\_p1">

<seg><w lemma="de" pos="ADP">de</w><w lemma="misterio" pos="NOUN">misterio</w><pc lemma="," pos="PUNCT">,</pc></seg>

<seg><w lemma="libremente" pos="ADV">libremente</w><pc lemma="..." pos="PUNCT">...</pc><w lemma="ya" pos="ADV">ya</w></seg>

<seg><w lemma="poder" pos="AUX">puedo</w><w lemma="calmar" pos="VERB">calmar</w><w lemma="en" pos="ADP">en</w></seg>

<seg><pc lemma="," pos="PUNCT">,</pc><w lemma="tanto" pos="ADV">tan</w><w lemma="lleno" pos="ADJ">llena</w></seg>

<seg><w lemma="Dios" pos="INTJ">Dios</w><w lemma="mío" pos="PRON">mío</w><pc lemma="!" pos="PUNCT">!</pc></seg>

<!-- ... -->

</p>

</div>

<div type="chapter" xml:id="Aarons\_Attila\_ch01"><!--French corpus-->

<head xml:id="Aarons\_Attila\_ch01\_head0">I</head>

<p xml:id="Aarons\_Attila\_ch01\_p1">

<seg><pc lemma="," pos="PUNCT">,</pc><w lemma="Stein" pos="PROPN">Stein</w><w lemma="être" pos="AUX">était</w></seg>

<seg><pc lemma="," pos="PUNCT">,</pc><w lemma="ébouriffé" pos="NOUN">ébouriffé</w><pc lemma="," pos="PUNCT">,</pc></seg>

<seg><w lemma="luisant" pos="ADJ">luisantes</w><w lemma="de" pos="ADP">de</w><w lemma="sueur" pos="NOUN">sueur</w></seg>

<seg><w lemma="que" pos="SCONJ">que</w><w lemma="possible" pos="ADJ">possible</w><w lemma="dans" pos="ADP">dans</w></seg>

<!-- ... -->

</p>

</div>

<div type="chapter" xml:id="Luxun\_Kuangren\_ch00"><!--Chinese corpus-->

<p xml:id="Luxun\_Kuangren\_ch00\_p0">

<seg><pc pos="PUNCT">。</pc><w pos="VERB">勞君</w><w pos="ADJ">遠道</w></seg>

<seg><w pos="ADV">多</w><w pos="VERB">荒唐之言</w><pc pos="PUNCT">；</pc></seg>

<seg><w pos="ADV">亦</w><w pos="ADV">不</w><w pos="VERB">著</w></seg>

<seg><w pos="VERB">良友</w><pc pos="PUNCT">；</pc><w pos="VERB">分隔</w></seg>

<!-- ... -->

</p>

</div>

Example 6: Randomization of n-grams tokens within paragraph-like elements

In the English corpus, the first paragraph-like element would contain the following text: “at London . ' s House , in Mr. Clover SCENE . — Drawing - Room”. As readers, we recognize some of the n-grams as intelligible, such as “Mr. Clover”. The same sentence with the order of the tokens completely randomized (as presented in 4.5) could be “in 's London at Mr. Clover Drawing . — , Room SCENE House . -”. Although the order in this case is completely randomized, some of the tokens could resemble a possible original order, such as "at Mr. Clover". Thus, a similar sense of unintelligible text and lucky guess arises with both variants. However, we believe that this variant preserving n-grams can raise more legal concerns.

# 5. Implementation in TextGrid

In this section, we want to present several new developments and the publication of two corpora as derived data in TextGridRep. This repository is an open access, for-free, multilingual repository mainly conceived for TEI documents. Originally, it was part of the German Grid initiative, and since then it has being maintained and developed under further programs such as DARIAH-DE, CLARIAH-DE, and more recently Text+ (Neuroth, Rapp, and Söring 2015; Schmunk and Funk 2016; Calvo Tello et al. 2023; Rißler-Pipka and Stein 2023). TextGridRep contains one of the largest open access corpora in German, the Digital Library (Betz 2015). Together with the publication of the CoNSSA and the ELTeC corpora in the past years (Rißler-Pipka et al. 2023), two features in TextGridRep have been developedwhich are now available to every TextGridRep user: The *portalconfig* and *readme* files[[7]](#footnote-8) provide a project with several options about their configuration and presentation in TextGridRep. Users can describe their TextGrid project through the *readme* file, which generates a project specific landing page for your TextGrid project. The *readme* file is written in Markdown and its content is shown after a short description and the list of the objects at the landing page of the project. The *portalconfig* file can be used to configure a project specific presentation page that is used in exchange to the general landing page for TextGrid projects. The project’s title can be configured here as well as a project logo can be added. Furthermore a project description can be added and also project specific search facets.

Another feature that is provided by the TextGridRep is the project specific XSL transformation to present content of published TEI files as HTML. It is possible to provide:

1. a project specific XSLT file for every project’s TEI file altogether (through the *portalconfig* file),
2. a reference to an XSLT file in each TEI to be transformed, or
3. an XSLT file for all files of one mimetype (through the *portalconfig* file).

The first two are already implemented, and the latter will be developed in the near future, so that a fine-grained configuration of XSL transformations is made possible to visualize project specific TEI files. This can be specially interesting for derived data for which the standard transformation to HTML can lead to poor results.

When publishing data in TextGridRep, researchers can now use derived data as a possible mimetype, which can be then selected by users as a further facet on the left menu. All derived data files are then shown with a warning that the presented document does not represent the original text.

In the following sections, we present the publication of two corpora in TextGridRep following two different strategies. These corpora have been already briefly described in Section 3 and have been part of the examples of Section 4.

## 5a. CoNSSA

As mentioned before, the majority of this corpus has been already published in several platforms (Calvo Tello et al. 2023; Calvo Tello 2021). The texts now published in TextGridRep as TEI-derived data are 59 texts by 23 authors (such as Gómez de la Serna, Fernández Flórez, Jarnés, Concha Espina, Poncela, Azorín, Pérez de Ayala, Francisco Ayala and others).

The data of the CoNSSA files as derived data are transformed according to the variant presented here in 4.5: randomization of tokens in paragraph-like elements. Before the randomization took place, the paragraphs are tokenized and linguistically annotated (lemma and POS) using attributes in the <w> and <pc> elements.[[8]](#footnote-9)

This derived data is published in TextGrid using a special XSLT stylesheet that displays some metadata (authority file identifiers, abstract, measures), as can be seen in Figure 1. The text in the <head> elements is kept in its original form for the correct generation of the table of contents.

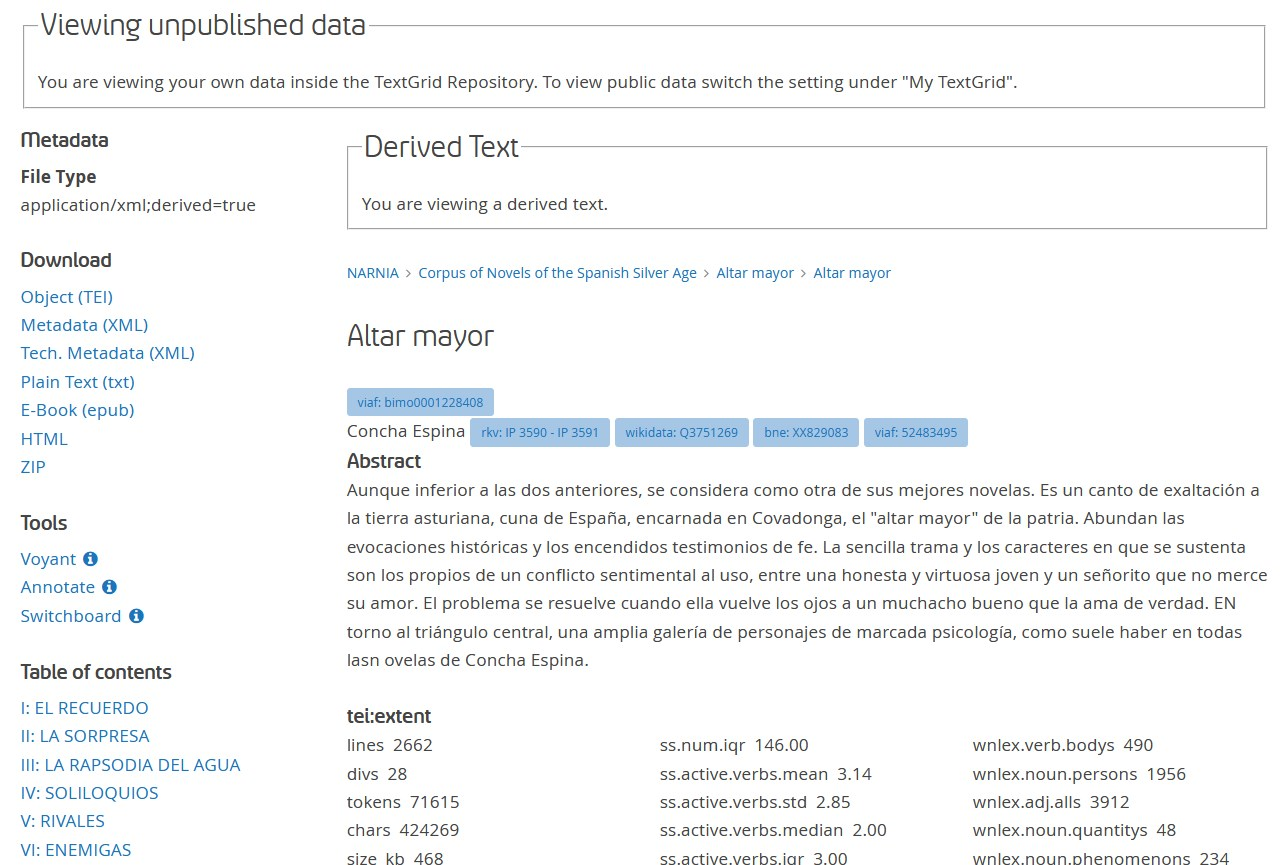


Figure 1: Presentation of derived data (metadata, table of content) in TextGridRep for a text in CoNSSA

Figure 2 shows the beginning of the text transformed into HTML in the TextGridRep portal. As a native speaker, one can see that the text resembles paragraphs, but the lack of a word in uppercase at the beginning of the paragraph and the lack of punctuation at the end suggest that this is not ordinary text. Even with short one-sentence paragraphs, such as the last two paragraphs in Figure 2, the task of reconstruction seems far from trivial.

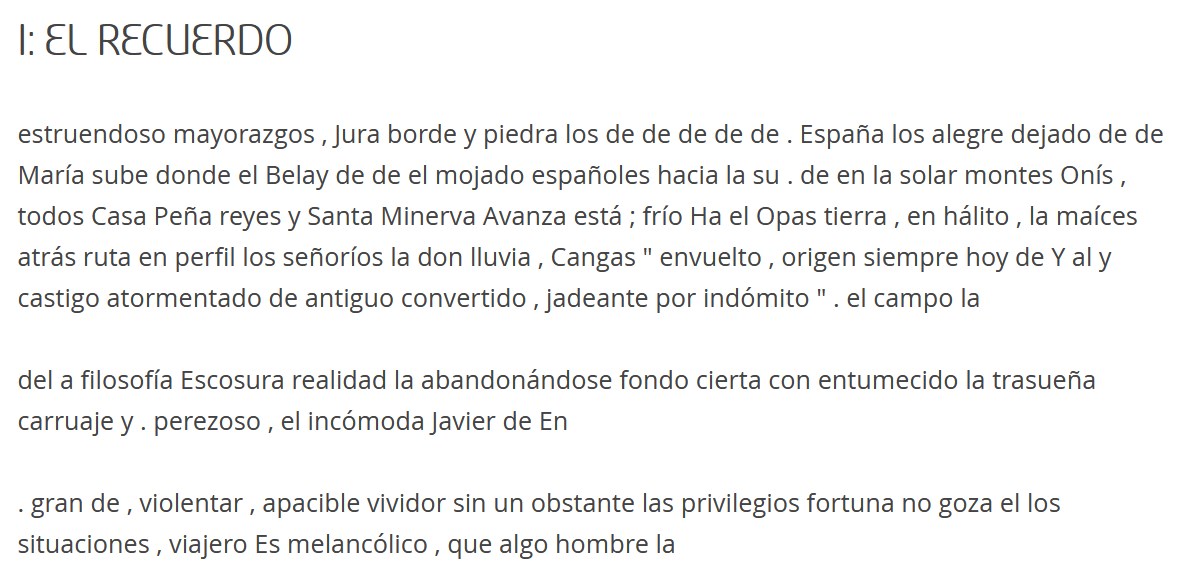


Figure 2: Presentation of derived data (randomized tokens within paragraphs) in TextGridRep for a text in CoNSSA

What can research do with these nonsensical paragraphs in TEI? As explained in 4.5, tools that accept TEI as an input format can analyze these files without further modification. We show two examples here, the first with a closer look at a single text and the second with a more distant reading approach. In both cases, the tools would produce the same results if the original texts had been used instead of a derived data document.

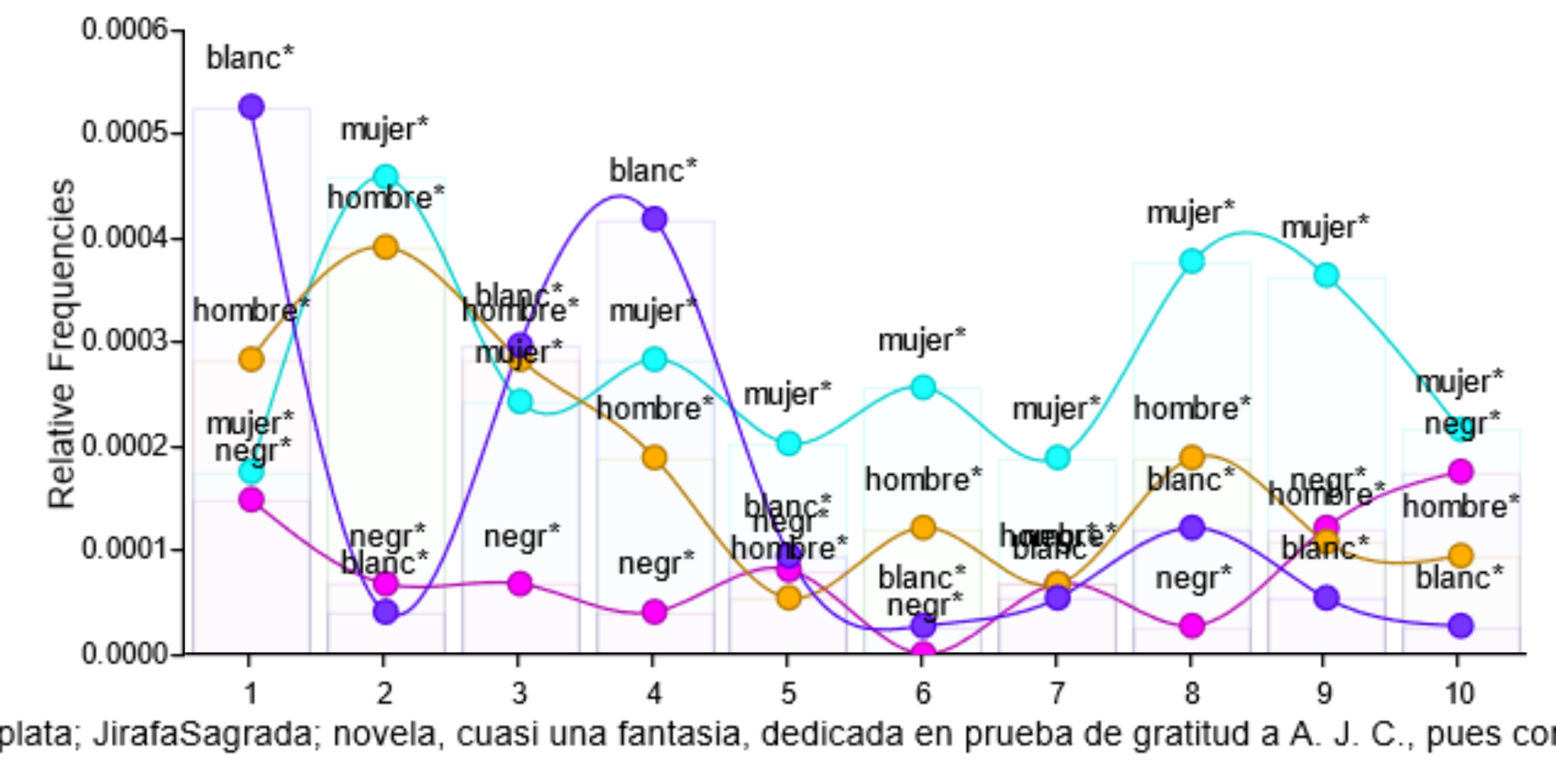


Figure 3: Distribution of frequencies over the length of the novel in a single novel using randomized order of tokens as derived data

In Figure 3, use Voyant Tools (Sinclair and Rockwell 2016) to explore a single novel: *La jirafa sagrada* by Salvador Madariaga. This novel presents an alternative universe in which black women are the highest group in society, providing an original reflection on discrimination. In Voyant Tools, researchers can explore the development of frequencies along the text of the roots in Spanish for the words *women*, *men*, *black* and *white*. While the development of *women* and *men* is similar, the frequencies of *black* and *white* are completely dissimilar.

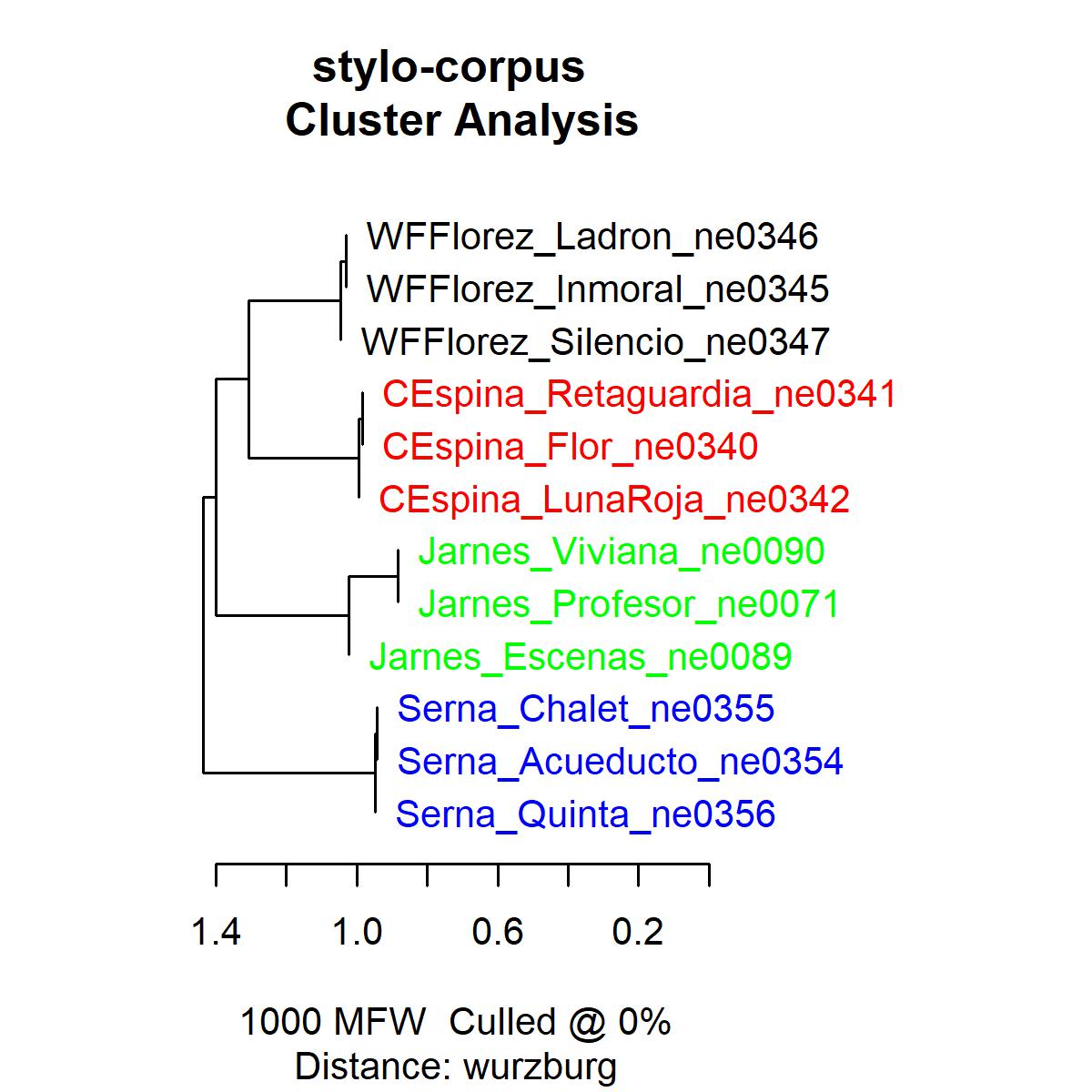


Figure 4: Clusters of a corpus of derived data using randomized order of tokens as derived data

Another use case for these data could be stylometric methods such as Delta (Burrows 2002), for example for authorship attribution. Figure 4 shows a dendrogram generated by the R library stylo (Eder, Kestemont, and Rybicki 2016) using Cosine Delta (Evert et al. 2017; Smith and Aldridge 2011) on a smaller subcorpus of derived data from CoNSSA. As can be seen, all texts are correctly clustered by authorship.

We want to stress that the analyses shown in Figure 3 and 4 would produce identical results if they would be conducted with the original files with the correct order of tokens.

## 5b. American Drama Corpus

For the corpus “American Drama 1714-1915”, most of its total of 1558 documents are out of copyright as of today. But since the texts are delivered as a pre-compiled corpus on CD-ROM, intellectual property rights might be claimed by the curator or publisher. While the database and the access-providing software offers a research tool including full-text and metadata search functionality to the reader as well as further information like related bibliographic and biographic data, a data extraction is required to make data-driven research reproducible. This mainly covers structural markup as it is one key for analyzing dramatic texts on a larger scale and also requires algorithms applied to the content that are not part of the software shipped with the CD-ROM. The corpus can be also found in a platform which offers more features.[[9]](#footnote-10)

The software offers functionality to read and export the texts. A SGML export function puts structural information to the local clipboard. The structural information in the TEI encoding could be then used for the creation of character networks of co-occurrence. For this purpose, we derived the data of these files as proposed in Section 4.4, i.e. empty TEI elements without the actual text. To these empty elements, we add a <note> element with several <measure> elements quantifying different aspects of the text, similarly as presented in Section 4.1.

Only a section of the entire corpus is intended to be published in TextGridRep.[[10]](#footnote-11) Currently, there are 772 TEI documents by 330 authors, plus a few texts of anonymous origin. The collection contains a huge diversity in terms of genre from short plays like drawing room plays to extensively long tragedies, where to the most extent one “The Sons of Usna: A Tragi-Apotheosis, in Five Acts” by T. H. Chivers. Regarding the number of acting persons, we find monologues and on the other end of this scale a play featuring 90 persons (“The Garden of Paradise” by Edward Sheldon).

One goal is to allow researchers to use this derived data at the DraCor platform. The platform offers programmatic access as well as a web application to explore the content, not focusing on the text in the first place, but on network data extracted from TEI documents.

<div type="act">

<head> ACT I. </head>

<div type="scene">

<head>SCENE I.</head>

<stage> — Room in a Welsh Inn .</stage>

<stage>Enter DAVY and JANET, from R. D., in flat .</stage>

<sp who="#davy">

<speaker>Davy .</speaker>

<stage>Off at door as he enters</stage>

<p>

<note>Contains <measureGrp>

<measure commodity="words" quantity="5">5 words</measure> in

<measure commodity="sentences" quantity="1">1 sentence</measure> using

<measure commodity="chars" quantity="30">30 chars</measure>

</measureGrp>.</note>

</p>

</sp>

<sp who="#janet">

<speaker>Janet .</speaker>

<p>

<note>Contains <measureGrp>

<measure commodity="words" quantity="5">5 words</measure> in

<measure commodity="sentences" quantity="2">2 sentences</measure> using

<measure commodity="chars" quantity="26">26 chars</measure>

</measureGrp>.</note>

</p>

</sp>

<!--...-->

</div>

Example 7: XML sample snippet of a speech acts in a TEI-encoded drama using a derived format

The data presented in Example 6 allows for the creation of network data, corresponding metrics and visualization.

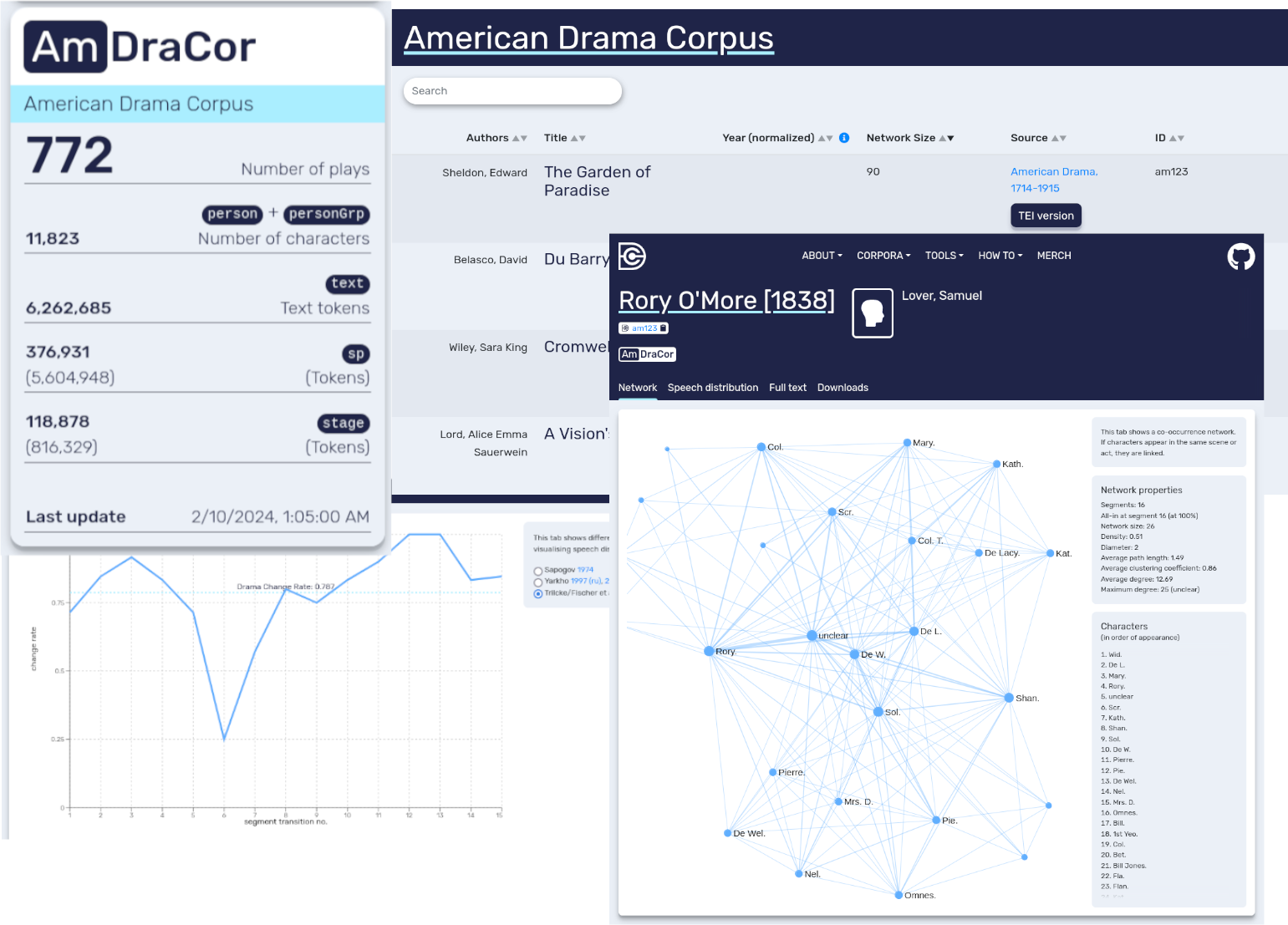


Figure 5: Rendering of derived data within a local instance of the DraCor software

The DraCor backend software is not ready for the derived data yet, as some metrics are calculated from text nodes within the TEI files, e.g. number of tokens. The derived format provides this information, but it has to be parsed differently. Also the TEI application profile (the dracor-schema) is not aware of this format.

# 6. Outreach

There is already a long tradition of projects that have challenged legal restrictions on the publication of research data. So far, these projects have used ad hoc solutions, models and formats, often spreadsheets or plain text. In this article we have criticized these solutions for degrading the FAIR status of the data, particularly in terms of findability, interoperability and usability. Publishing derived data in TEI, as presented here, can improve their FAIR status. Compared to other formats, such as JSON-LD, TEI has a much wider acceptance in the humanities.

Moreover, a TEI file can accommodate the different types of data (structure, metadata, text, annotation) better than other formats. In recent years some of the discussions about derived data have focused more on the text part (see the German term *derived text format* or *abgeleitete Textformate*), to the detriment of other types of data. In a TEI file, it is more obvious than in other formats that the text is only one component of the document's data. Metadata, textual structure and annotations can be just as interesting for research as text, and are subject to fewer legal restrictions.

With this article we want to move the conversation and practice of derived data forward with real corpora from different genres, time periods and languages. As seen in the prose and drama examples, the characteristics of the genre can determine which derived data variants might be more interesting for researchers. In addition to these two specific corpora, we have presented how derived data can now be published in TextGridRep. The conversation and practice need to continue developing, and for this reason we invite the community to give us feedback on the derived data already published, but also to bring their corpora of derived data to TextGridRep. Together we can hope for a future with more and better data, even with legal restrictions.

# 7. References

Betz, Katrin. 2015. ‘Ein virtuelles Bücherregal: Die Digitale Bibliothek im TextGrid Repository’. In *TextGrid: Von der Community - für die Community*, edited by Heike Neuroth, Andrea Rapp, and Sibylle Söring, 229–39. Glückstadt: Verlag Werner Hülsbusch.<https://publications.goettingen-research-online.de/handle/2/14388>.

Blei, David M. 2012. ‘Probabilistic Topic Models’. *Communications of the ACM* 55 (4): 77–84.

Burrows, John. 2002. ‘“Delta”: A Measure of Stylistic Difference and a Guide to Likely Authorship’. *Literary and Linguistic Computing* 17 (3): 267–87.<https://doi.org/10.1093/llc/17.3.267>.

Calvo Tello, José. 2021. *The Novel in the Spanish Silver Age: A Digital Analysis of Genre Using Machine Learning*. Digital Humanities Research 4. Bielefeld: transcript.<https://www.transcript-verlag.de/978-3-8376-5925-2/the-novel-in-the-spanish-silver-age/?c=331025282>.

———. 2022. ‘Where Are Romance Studies Heading? A Bibliographic Data Science Analysis Using Regression’. Edited by José Calvo Tello, Nanette Rißler-Pipika, and Jan Rohden. *Apropos [Perspektiven Auf Die Romania]* 9 (December): 119–52.<https://doi.org/10.15460/apropos.9.1894>.

Calvo Tello, José, Stefan E. Funk, Mathias Göbel, Daniel Kurzawe, Nanette Rißler-Pipika, and Ubbo Veentjer. 2023. ‘Between Corpora, Tools, and Authority Files: TextGrid Repository for Hispanic Studies’, Revista de Humanidades Digitales, 8: 90–108.<https://doi.org/10.5944/rhd.vol.8.2023.37994>.

Calvo Tello, José, Nanette Rißler-Pipika, Florian Barth, Jung Kerstin, and Christof Schöch. 2023. ‘Questionnaire of the Survey: How Do You Compose Your Literary Corpus or Literary Collection?’ Questionnaire. Zenodo.<https://doi.org/10.5281/zenodo.10203624>.

Calvo Tello, José, and Nanette Rißler-Pipka. 2023. ‘¿Qué hacer con textos que no se pueden publicar? Datos derivados, criterios FAIR y TEI’. *Journal of the Text Encoding Initiative*, no. 16 (May).<https://doi.org/10.4000/jtei.4720>.

Creamer, Andrew, Gaia Lembi, Elli Mylonas, and Michael Satlow. 2021. ‘Archiving a TEI Project FAIRly’. *Journal of the Text Encoding Initiative*, no. Issue 14 (March).<https://doi.org/10.4000/jtei.4324>.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. ‘BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding’. *arXiv:1810.04805 [Cs]*, May.<http://arxiv.org/abs/1810.04805>.

Eder, Maciej, Mike Kestemont, and Jan Rybicki. 2016. ‘Stylometry with R: A Package for Computational Text Analysis’. *The R Journal* 16 (1): 1–15.

Evert, Stefan, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch, and Thorsten Vitt. 2017. ‘Understanding and Explaining Delta Measures for Authorship Attribution’. *Digital Scholarship in the Humanities*.<https://doi.org/10.1093/llc/fqx023>.

Fischer, Frank, and Robert Jäschke. 2022. ‘Ein Quantum Literatur. Empirische Daten zu einer Theorie des literarischen Textumfangs’. In *Digitale Literaturwissenschaft: DFG-Symposion 2017*, edited by Fotis Jannidis, 777–812. Germanistische Symposien. Stuttgart: J.B. Metzler.<https://doi.org/10.1007/978-3-476-05886-7_32>.

Gengnagel, Tessa, Frederike Neuber, and Daniela Schulz. 2023. ‘FAIR Enough? Evaluating Digital Scholarly Editions and the Application of the FAIR Data Principles’. Application/tei+xml,application/pdf,text/html. *RIDE*, no. 16.<https://doi.org/10.18716/RIDE.A.16.0>.

Goldhahn, Dirk, Thomas Eckart, and Uwe Quasthoff. 2012. ‘Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages’. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, 759–65. Istanbul, Turkey: European Language Resources Association (ELRA).<http://www.lrec-conf.org/proceedings/lrec2012/pdf/327_Paper.pdf>.

Hinrichs, Erhard, Peter Leinen, Alexander Geyken, Andreas Speer, and Regine Stein. 2022. ‘Text+: Language- and Text-Based Research Data Infrastructure’. Zenodo. https://doi.org/10.5281/zenodo.6452002.<https://zenodo.org/record/6452002>.

Jawanpuria, Pratik, Arjun Balgovind, Anoop Kunchukuttan, and Bamdev Mishra. 2019. ‘Learning Multilingual Word Embeddings in Latent Metric Space: A Geometric Approach’. *Transactions of the Association for Computational Linguistics* 7 (April): 107–20.<https://doi.org/10.1162/tacl_a_00257>.

Jett, Jacob, Boris Capitanu, Deren Kudeki, Timothy Cole, Yuerong Hu, Peter Organisciak, Ted Underwood, Eleanor Dickson Koehl, Ryan Dubnicek, and J. Stephen Downie. 2020. ‘The HathiTrust Research Center Extracted Features Dataset’. HathiTrust Research Center.<https://doi.org/10.13012/R2TE-C227>.

Kett, Jürgen, Christoph Kudella, Andrea Rapp, Regine Stein, and Thorsten Trippel. 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>.

Lin, Yuri, Jean-Baptiste Michel, Erez Aiden Lieberman, Jon Orwant, Will Brockman, and Slav Petrov. 2012. ‘Syntactic Annotations for the Google Books NGram Corpus’. In *Proceedings of the ACL 2012 System Demonstrations*, 169–74. Jeju Island, Korea: Association for Computational Linguistics.<https://aclanthology.org/P12-3029>.

Neuroth, Heike, Andrea Rapp, and Sibylle Söring, eds. 2015. *TextGrid: Von der Community - für die Community: eine virtuelle Forschungsumgebung für die Geisteswissenschaften*. Glückstadt: Verlag Werner Hülsbusch.<https://publications.goettingen-research-online.de/handle/2/14388>.

Ortiz Suárez, Pedro Javier, Benoît Sagot, and Laurent Romary. 2019. ‘Asynchronous Pipelines for Processing Huge Corpora on Medium to Low Resource Infrastructures’. Application/pdf, 337 KB.<https://doi.org/10.14618/IDS-PUB-9021>.

Rehm, Georg, Andreas Witt, Heike Zinsmeister, and Johannes Dellert. 2007. ‘Masking Treebanks for the Free Distribution of Linguistic Resources and Other Applications’. In , edited by Koenraad De Smedt, Jan Hajič, and Sandra Kübler, 127–38. Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories. Tartu University Library (Estonia): Northern European Association for Language Technology (NEALT).<http://hdl.handle.net/10062/4814>.

Rißler-Pipka, Nanette, José Calvo Tello, Stefan E. Funk, Carolin Odebrecht, Christof Schöch, and Ubbo Veentjer. 2023. ‘The European Literary Text Collection in TextGrid Repository’. In *Collaboration as Opportunity*, edited by Walter Scholger, Georg Vogeler, Toma Tasovac, Anne Baillot, Elisabeth Raunig, Martina Scholger, Elisabeth Steiner, and Patrick Helling. Graz: ADHO.<https://doi.org/10.5281/ZENODO.8107707>.

Rißler-Pipka, Nanette, and Regine Stein. 2023. ‘Die Wege des virtuellen Skriptoriums St. Matthias. Von TextGrid über die NFDI zu EOSC’. In *Computer im Musenhain. Von träumenden Büchern und der Aura des Digitalen*, edited by Sabine Bartsch, Luise Borek, and Philipp Hegel, 01|2023:173–82. Darmstadt: Universitäts- und Landesbibliothek Darmstadt.<https://doi.org/10.26083/tuprints-00024712>.

Sánchez Sánchez, Mercedes, and Carlos Domínguez Cintas. 2007. ‘El banco de datos de la RAE: CREA y CORDE’. *Per Abbat: boletín filológico de actualización académica y didáctica*, no. 2: 137–48.

Schmunk, Stefan, and Stefan E. Funk. 2016. ‘Das DARIAH-DE- Und Das TextGrid-Repositorium: Geistes- Und Kulturwissenschaftliche Forschungsdaten Persistent Und Referenzierbar Langzeitspeichern’. *Bibliothek Forschung Und Praxis* 40 (2): 213–21.<https://doi.org/10.1515/bfp-2016-0020>.

Schöch, Christof, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann, and Jörg Röpke. 2020. ‘Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen’. Html,PDF,Xml. *Zeitschrift für digitale Geisteswissenschaften*.<https://doi.org/10.17175/2020_006>.

Smith, Peter W. H., and W. Aldridge. 2011. ‘Improving Authorship Attribution: Optimizing Burrows’ Delta Method’. *Journal of Quantitative Linguistics* 18 (1): 63–88.<https://doi.org/10.1080/09296174.2011.533591>.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. ‘The FAIR Guiding Principles for Scientific Data Management and Stewardship’. *Scientific Data* 3 (March).<https://doi.org/10.1038/sdata.2016.18>.

1. <https://books.google.com/ngrams/info> [↑](#footnote-ref-2)
2. In 2023, two of the authors of this paper published an article in Spanish discussing some solutions for the specific case of Spanish literature (Calvo Tello and Rißler-Pipka 2023). Although the current paper is based on a similar starting point, it addresses a broader audience, explores further possibilities, uses examples in five languages, and shows the implementation and publication of two corpora in the TextGridRep. [↑](#footnote-ref-3)
3. <https://github.com/Zeta-and-Company/derived-formats>. [↑](#footnote-ref-4)
4. <https://zh.wikisource.org/wiki/Author:魯迅>. [↑](#footnote-ref-5)
5. <https://github.com/dkltimon/LuXun_Works>. [↑](#footnote-ref-6)
6. The order of the corpora follows an alphabetical order considering the language code expressed following the ISO 639 as two characters: German, English, Spanish, French, and Chinese. [↑](#footnote-ref-7)
7. [https://textgridlab.org/doc/services/submodules/kolibri/kolibri-addon-textgrid-import/docs/import\_and\_configuration.html#project-specific-landing-page-and-project-metadata](https://textgridlab.org/doc/services/submodules/kolibri/kolibri-addon-textgrid-import/docs/import_and_configuration.html" \l "project-specific-landing-page-and-project-metadata). [↑](#footnote-ref-8)
8. Currently, the data can be visualized in this test instance of TextGridRep: <https://test.textgridrep.org/browse/8wtr.0?lang=en>. [↑](#footnote-ref-9)
9. https://www.nationallizenzen.de/angebote/nlproduct.2006-03-10.9045864083, (22.01.2024) [↑](#footnote-ref-10)
10. <https://test.textgridrep.org/project/TGPR-6285e2a3-63eb-f3c1-1445-65d857a140ab>. See example: . <https://test.textgridrep.org/browse/8x7r.1>. [↑](#footnote-ref-11)