Sharing is Caring: A Text Alignment Approach for Sharing Annotations of Copyrighted Texts

Frederik Arnold[0000−0002−0417−4054] and Robert Jäschke[0000−0003−3271−9653]

Humboldt-Universität zu Berlin
{frederik.arnold,robert.jaeschke}@hu-berlin.de

Abstract. Digital libraries are a crucial infrastructure for researchers to find and access resources. For example, researchers in the digital humanities and computational literary studies make frequent use of digital libraries to access relevant materials, such as text corpora, which they then annotate. These materials are often protected by copyright and are typically released under licensing terms that either prohibit redistribution entirely or impose significant limitations on how they can be shared. We present and evaluate an approach to separate annotations from the underlying text and create a fingerprint, which cannot be used on its own to recreate the original text and can therefore be shared. This fingerprint can be used to merge the separated annotations with another version of the original text. Our framework can easily be adapted to support different file formats and can be integrated into digital libraries to simplify sharing of annotations of copyrighted texts. The code is publicly available under the Apache License 2.0 at `https://hu.berlin/sisc`.

Keywords: annotation · sharing · open science · digital humanities

1 Introduction

Digital libraries are a vital infrastructure and integral part of modern research, providing platforms to find and access digital resources. In fields such as the digital humanities and (computational) literary studies, digital libraries serve as indispensable tools, granting scholars access to extensive corpora. However, these materials are often subject to copyright restrictions that do not allow redistribution, or only under rather restrictive conditions. This results in a number of complications. First, it makes it more difficult to acquire these materials in the first place, even for research purposes. Second, it makes it more difficult or even impossible to share the results of research efforts, for example, a corpus of annotated texts. The laws governing scholarly use of copyrighted material vary by country. In Germany, for instance, sharing of such material is difficult, even for research purposes. This results in situations where it is impossible to publish reproducible research because the underlying data cannot be shared.

Especially for annotations of textual data, one approach to work around these limitations is *stand-off markup* [7]. This describes the separate storage of content and markup, where the markup references the content by positions, for example,

of characters or words. This separation has some advantages over inline markup: The original content is not manipulated, overlapping and multiple annotations are possible, and annotations can be shared without having to provide the underlying content. This is especially useful when copyright prohibits distribution. Although this simplifies sharing annotations of copyrighted content, the original text with the same character positions is needed for meaningful reconstruction. If the receiver of a standoff annotation has access to an identical version of the text, merging is easily possible (see, for example, [17]). Unfortunately, often an identical original is not available and acquisition of the exact same text can be hard or even impossible, since approaches for text extraction (e.g., from PDFs) like optical character recognition (OCR) still produce varying results.

In this work, we present *SisC* (**S**haring **is** **C**aring),[1] an approach to automatically separate annotations from the underlying text. SisC uses a *fingerprint*, that is, a masked version of the text, to merge stand-off annotations with another version of the original text, for example, extracted from a PDF file. The fingerprint cannot be used on its own to recreate (meaningful parts of) the original text and can therefore, to the best of our knowledge, be shared.

Digital libraries have long been battling with the legal and technical constraints that come with handling materials under copyright restrictions [8,3] and addressing these hurdles is essential for fostering collaboration and open science. With this work, we contribute our tool SisC to help alleviate some of the aforementioned issues. Our focus is not on a particular annotation methodology or format but to create a general framework which works with stand-off and inline annotations and can easily be adapted to different (textual) file formats.

This paper is organized as follows: In Section 2, we provide an overview on related work. We then introduce our method in Section 3, followed by an evaluation in Section 4. We conclude with a discussion in Section 5.

2 Related Work

Annotations for textual data can be stored in a variety of file formats (e.g., TXT, CSV, JSON, XML) which are not specific to annotations and only define an abstract structure of the content. To facilitate sharing and collaboration, different annotation formats were developed over time, build on top of the mentioned file formats. For example, CONLL, originally introduced as part of the CONLL-2000 Shared Task [9] grew into a widely used annotation format with a number of variants. TEI XML [15] is another popular format, developed by the Text Encoding Initiative, to provide a set of guidelines for annotating textual data.

Annotations are often realized as inline annotations which comes with the caveat that, without separation of data and annotation, sharing is difficult or impossible in cases where the annotated source is subject to copyright restrictions.

The terms *stand-off markup* and *stand-off annotation* are often used interchangeably and refer to the notion of storing content and markup separately. The

[1] The source code is available under the Apache License 2.0 at `https://hu.berlin/sisc`.

```
<TEI>
  <text>
    <body>
      <p>Some text with <q>an annotated quote</q>.</p>
    </body>
  </text>
</TEI>
```

Listing 1: Example TEI XML with an annotated quote.

idea dates back to the 1990s and an early mention of this concept can be found in [14]. Later, Thompson and McKelvie introduced semantics for hyperlinks for stand-off markup [16]. It is common to use the term stand-off markup to refer to the general concept of storing annotations and text separately, without any restrictions on the form the annotations can have. Sometimes, a stricter definition is used, in which stand-off markup only refers to the case where markup tags are separately stored but still conform to a context-free grammar with a strict hierarchical text structure, see, for example, [10]. Schmidt uses the term *stand-off properties* to refer to a type of stand-off annotation that allows for overlapping annotations. Burghardt and Wolff [4] give an overview over different formats and tools that implement stand-off concepts in various ways.

Derived text formats (DTF) are one approach to handling copyrighted material and sharing information [11,12]. Google N-Grams is an example for a dataset in such a format [5]. However, DTFs only work for cases where the exact text is not relevant, for example, when a term-document-matrix, n-gram counts, or word embeddings, are sufficient. Thus, they are unsuitable for sharing annotated texts.

3 Method

We introduce the overall idea of SisC using the example of the TEI XML file format which is widely used for annotating texts. Our approach works by creating an *exchange TEI XML file* that includes the annotations plus metadata for later alignment with (another version of) the text. Someone (e.g., a researcher or a digital library) who receives that file and has access to the original text can then reconstruct the annotated TEI/XML file.

3.1 Creating the Exchange File

From a TEI XML file with annotations (Listing 1), we create an intermediate TEI XML file (Listing 2) which consists of the original XML tags, but lacks the annotated text. Instead, it contains a *fingerprint* which *masks* the annotated text such that it cannot be reconstructed without additional information. During this process, the attributes `sisc_start` and `sisc_end` are added to the XML tags and refer to the start and end character positions of the text in the fingerprint,

```
<TEI>
  <text>
    <body sisc_start="0" sisc_end="36">
      <p sisc_start="1" sisc_end="35">
        <q sisc_start="16" sisc_end="34" />
      </p>
    </body>
  </text>
  <standoff>
  S___ _ex_ ___h __ __no_____ q____.
  </standoff>
</TEI>
```

Listing 2: The annotated text from Listing 1 represented in our proposed TEI XML exchange format with *uniform* fingerprinting (for $n = 2$ and $d = 5$).

respectively, which is stored in the `standoff` tag. Prior to the fingerprint creation all XML tags are removed such that the resulting fingerprint only consists of text. This implementation decision was made to have a clear separation of the data handled in the different processing steps. For example, the masking step only works on strings without the need for additional information such as XML tags. This simplifies the addition of support for new file formats.

Properly handling whitespace in XML files is not trivial.[2] Currently, whitespace is preserved in the fingerprint, except around newlines. We plan to support further options in the future.

SisC implements two variants for handling page layout information during fingerprint creation. The first variant, *in-place*, is sufficient for works that consist of running text only. However, the page layout is often more complex and some elements do not follow a linear order. In particular, footnotes appear at the end of pages in a PDF file but might be moved to different positions in the TEI XML file, for example, their anchor positions. To handle such cases, SisC implements a second variant, *move-fn*, which moves footnotes to the end of the pages and makes the order of elements of the fingerprint match the order in the PDF file. This includes handling of footnotes that run over multiple pages. This variant requires that footnotes and page breaks are annotated in the TEI XML file. During processing, we introduce three additional attributes: `sisc_text_start` and `sisc_text_end` are added to the XML tags for footnotes and refer to the start and end character positions of the text in the fingerprint, respectively, after the text is moved to the new position. The attribute `sisc_skip` is added to the XML tags for page breaks to store the length of text that needs to be ignored during reconstruction. Specifically, for a footnote which starts on one page and ends on the next page, during fingerprinting that footnote is split and the length of the text between to the two parts needs to be known during reconstruction.

[2] `https://www.w3.org/TR/xml/#sec-white-space`

```
Running text at the end of a page some header text text on the next page
Ru_____ t___ _t ___ _nd __ _ p___ ----------------_ex_ __ th_ ___t ____
```

Listing 3: Alignment example with superfluous text (highlighted and aligned with a gap ('-') in the fingerprint). The first line shows the text that was extracted from the PDF file and the second line shows the fingerprint.

We propose uniform masking, that is to keep n characters every d-th character of the original text and replace characters at other positions. We only replace letter characters with '_' but keep all numbers and punctuation (see Listing 2).

Keeping the placeholder '_' instead of removing characters makes the process more comprehensible and can simplify debugging. When optimizing for (space) efficiency, the placeholder could be compressed (e.g., using run-length encoding).

3.2 Reconstruction of the Annotated Text

From this exchange format, we can reconstruct the original TEI XML if a version of the original text is available, for example, in a PDF file. To automatically extract the text from a PDF, we use pdf2image[3] and tesseract[4] and then align the fingerprint and the text with BioPython.[5] We then use the start and end positions (`sisc_start` and `sisc_end`) to merge the text and the TEI XML file.[6]

Text documents often contain segments, such as headers and footers, that are not necessarily included in their annotated versions. For our use case, such additional text segments need to be removed after alignment, in order to reconstruct the original TEI XML file as accurately as possible. Thus, we remove segments of text which are longer than 10 characters and are only present in the PDF file. An example is shown in Listing 3. The highlighted text is only present in the PDF file and aligned with a gap ('-') in the fingerprint.

Table 1. Statistics for the two corpora of annotated scholarly articles.

Literary work	Die Judenbuche	Michael Kohlhaas
Scholarly articles	44	49
Scholarly articles' characters	2 614 061	2 748 559
Footnotes	2 025	2 331
Footnotes' characters	471 027	482 099

[3] `https://github.com/Belval/pdf2image`
[4] `https://github.com/tesseract-ocr/tesseract`
[5] `https://biopython.org`
[6] As described in Sec 3.1, this is the simple case and more XML attributes are used for moved footnotes. More details are in the documentation and source code of SisC.

4 Evaluation

4.1 Setup

We evaluate SisC on a dataset of 44 scholarly articles which interpret the novella *Die Judenbuche* by Annette von Droste Hülshoff, and 49 scholarly texts which interpret the novella *Michael Kohlhaas* by Heinrich von Kleist. The texts are in TEI XML format and all direct quotations were manually annotated (i.e., enclosed by `<q>...</q>`). Footnotes and page breaks are also annotated and the texts of footnotes are moved from the end of the page to their anchor positions. The texts were also manually corrected for OCR errors. The texts in both corpora contain many footnotes which make up around 18 % of the text (see Table 1).[7]

These documents represent a typical research corpus for literary studies, with rich annotation of document structure, like paragraphs, section headings, page breaks, and footnotes. We use the corpus in our research project,[8] where we face the challenge of sharing copyrighted literary interpretations, which have been annotated, specifically, with citations to the two above-mentioned novellas.

We assume that the first page of PDF files contains the start of the main text. That is, author and title information may be present but the table of contents and other unrelated content has been removed.

We evaluate how faithfully SisC can reconstruct the original annotated document by measuring the average normalized Levenshtein similarity between the original and the reconstructed text. We set the number of kept characters to $n = 2$ and vary the distance d between 5 and 100 characters in steps of 5. Additionally, we evaluate two variants of our approach. In the first variant (*punct*), which resembles the setting $n = 0$ and $d = \infty$, only punctuation and whitespace characters are kept in the fingerprint but no additional characters. The second variant (*space*) further restricts this to only whitespace characters. All mentioned scenarios are evaluated with moved footnotes (*Move-fn*) and without moved footnotes (*In-place*). For the calculation of the Levenshtein similarity, we only take letter and number characters into account.

4.2 Results

The results in Figure 1 show that the reconstruction of the TEI XML files works best when the positions of the footnotes in the fingerprint match the positions in the text extracted from a PDF (*Move-fn*). We get consistently high Levenshtein similarity between 0.949 and 0.963, independent of the distance between unmasked characters. Even keeping only information about the positions of whitespace and punctuation (*punct*) or only whitespace (*space*) in the fingerprint is sufficient to achieve such a high similarity (between 0.938 and 0.956). That is, on average, only 5 characters every 100 characters are misaligned, equivalent to roughly 75 characters per page (assuming 1500 characters per page).

[7] One of the texts from the Judenbuche corpus is excluded from the evaluation, as it is over 200 000 characters long and thus took too long to process using BioPython.

[8] `https://hu.berlin/keypassages`

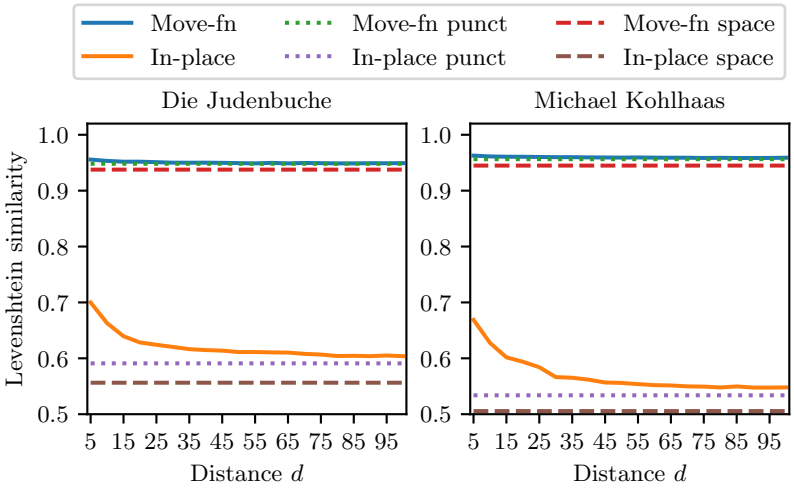

Fig. 1. Normalized Levenshtein similarity for *Die Judenbuche* and *Michael Kohlhaas* with (*Move-fn*) and without (*In-place*) moving footnotes for distances between 5 and 100 and two variants *punct* and *space*.

We identified three types of issues that remain. Firstly, general, smaller OCR errors. Secondly, variations between the fingerprint and OCR text due to tables and figures with descriptions. And thirdly, the first page of PDF files often contains a title, authors, date, affiliations, and so on. This can sometimes lead to alignment issues in the beginning of texts.

The results are worse when footnotes are not moved (*In-place*). With $d = 5$, we get a Levenshtein similarity of 0.70 for *Die Judenbuche*. With an increase in distance, the similarity gradually decreases to about 0.62 at $d = 35$, where it levels out and stays the same for larger d. We observe similar behavior for *Michael Kohlhaas*. This shows that there is a clear impact of the distance on the quality of the alignment. This observation is further supported by the performance of *punct* and *space* where we notice that the latter performs the worst.

We conclude that punctuation and whitespace alone already contain a lot of information which can be sufficient for good alignment, specifically in the case of moved footnotes. For the more difficult *In-place* case, we find that punctuation and whitespace characters only contain sufficient information for a certain level of quality of the alignment and that more information improves the alignment.

We also evaluated a second masking approach, where we keep text before and after annotations in a window of 10 characters on both sides and replace the remaining characters. The text of the annotations themselves is also masked. Our original hypothesis was that context masking might help in the *In-place* footnote case and that, even if the whole text cannot be reconstructed, reconstruction of certain annotations which are of special interest might be improved by this masking approach. This turned out not to be the case as this approach generally performed worse.

5 Discussion

We presented SisC, an approach for sharing of annotations of copyrighted texts. Our tool offers parameters to adjust the masking of text to fit personal preferences and specific legal requirements, which can vary by country.

Our evaluation shows that the approach works very well as long as footnotes appear at the same position in the fingerprint as in the PDF file. In our use-case of TEI XML files, which support rearranging the order of running text and footnotes, this is not an issue. But this could be a problem in other scenarios where rearrangement is not possible. The logical solution is to rearrange footnotes during alignment, but this needs information on the page layout of the PDF file.

Variations between the fingerprint and the OCR text due to figures and tables with captions can lead to alignment issues. This did not much impact our results as literary interpretations rarely contain tables or figures. We consider SisC most suitable for fields (like law or literary studies) that frequently annotate texts with infrequent use of tables and figures and whose licenses do not allow redistribution. Texts from other fields might need further development and testing.

Automatic layout detection of PDF files is a hard task in general [1] but recent approaches based on large language models show impressive results in OCR and layout detection tasks, even handling complex layouts [6,2,18,13]. We conducted tests with SpaCy Layout[9] and Mistral OCR[10] but found that both tools cannot reliably handle footnotes. For further details, see the appendix.

Our approach also assumes specific concepts for characters and words. Languages with variations of these concepts (e.g., non-Latin languages such as Chinese, Arabic, or Hindi) might require more sophisticated approaches.

To foster open science, we envision the integration of SisC into digital libraries and research data repositories. Scholars could then upload their annotations to repositories that create and publish exchange files. Repositories could then indicate to researchers that access an annotation, whether their local library has licensed the original work and forward them to its digital library. The digital library then retrieves the annotation from the repository and, using the licensed original text, reconstructs the annotated text. Digital libraries could also subscribe to repositories and show the availability of annotations for the licensed documents they provide and then allow their users to download annotated versions of the documents. Such integrations would take the burden from researchers to handle the processing of files and could foster the sharing of annotations.

Acknowledgments. Parts of this research were funded by the German Research Foundation (DFG) priority programme (SPP) 2207 *Computational Literary Studies* project *Is Expert Knowledge Key? Scholarly Interpretations as Resource for the Analysis of Literary Texts in Computational Literary Studies* (grant no. 424207720).[11]

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

[9] `https://github.com/explosion/spacy-layout`
[10] `https://docs.mistral.ai/capabilities/document/`
[11] `https://www.projekte.hu-berlin.de/en/schluesselstellen`

Appendix: Layout Detection

SpaCy Layout can extract layout including headers, running text, tables, and footnotes but does not link footnotes to their anchor position in the main text. Mistral OCR has similar functionality but does not reliably link footnotes and their anchor positions. The resulting text format often contains an indication of the start of the footer section but no separation between individual footnotes or a link between a footnote's text and its anchor in the running text. Footnote numbers are often marked as superscript but are otherwise not distinguishable from other numbering.

We tested passing the resulting markdown from Mistral OCR to the large language model (LLM) GPT-4.1-mini to extract text from the markdown with the following prompt:

> This is pdf content in markdown:
> <BEGIN_IMAGE_OCR>
> {pdf_ocr_markdown}
> <END_IMAGE_OCR>.
> Convert this into plain text. Include the text of footnotes in the running text surrounded by triple brackets, for example, [[[Text of a footnote]]].

We found that this approach would sometimes work quite well and sometimes the LLM would combine multiple footnotes into one, include that combined footnote in triple brackets in the running text, and then later reference that earlier footnote instead of using the actual footnote text. Considering the recent developments, we conclude that it is only a matter of time until these tools can handle those cases, but at the current time this is beyond the scope of this work.

References

1. Binmakhashen, G.M., Mahmoud, S.A.: Document layout analysis: A comprehensive survey. ACM Computing Surveys **52**(6) (2019). https://doi.org/10.1145/3355610
2. Blecher, L., Cucurull, G., Scialom, T., Stojnic, R.: Nougat: Neural optical understanding for academic documents (2023), arXiv:2308.13418
3. Breemen, V.: Digital libraries under EU copyright law: A relationship set in stone? European Papers – A Journal on Law and Integration **8**(2), 689–712 (2023)
4. Burghardt, M., Wolff, C.: Stand off-Annotation für Textdokumente: Vom Konzept zur Implementierung (zur Standardisierung?). In: Chiarcos, C., de Castillo, R.E., Stede, M. (eds.) Von der Form zur Bedeutung: Texte automatisch verarbeiten = from form to meaning: processing texts automatically: proceedings of the Biennial GSCL Conference 2009, pp. 53–59. Narr, Tübingen (2009), https://epub.uni-regensburg.de/14223/
5. Goldberg, Y., Orwant, J.: A dataset of syntactic-ngrams over time from a very large corpus of English books. In: Diab, M., Baldwin, T., Baroni, M. (eds.) Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity. pp. 241–247. Association for Computational Linguistics, Atlanta, Georgia, USA (2013), https://aclanthology.org/S13-1035/

6. Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., Park, S.: OCR-free document understanding transformer. In: European Conference on Computer Vision (ECCV). Lecture Notes in Computer Science, vol. 13688, pp. 498–517. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19815-1_29
7. Klug, H.W.: Stand-off-markup. In: Helmut W. Klug in collaboration with Selina Galka and Elisabeth Steiner in the HRSM project (ed.) KONDE Weißbuch, pp. 453–455 (2021), https://hdl.handle.net/11471/562.50.171
8. Samuelson, P.: Copyright and digital libraries. Communications of the ACM **38**(3) (1995)
9. Sang, E.F.T.K., Buchholz, S.: Introduction to the CoNLL-2000 shared task chunking. In: Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop (2000), https://aclanthology.org/W00-0726/
10. Schmidt, D.A.: Using standoff properties for marking-up historical documents in the humanities. it – Information Technology **58**(2), 63–69 (2016). https://doi.org/10.1515/itit-2015-0030
11. Schöch, C., Döhl, F., Rettinger, A., Gius, E., Trilcke, P., Leinen, P., Jannidis, F., Hinzmann, M., Röpke, J.: Abgeleitete Textformate: Prinzip und Beispiele. Recht und Zugang **1**(2) (2020). https://doi.org/10.5771/2699-1284-2020-2-160
12. Schöch, C., Döhl, F., Rettinger, A., Gius, E., Trilcke, P., Leinen, P., Jannidis, F., Hinzmann, M., Röpke, J.: Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen. Zeitschrift für digitale Geisteswissenschaften (2020). https://doi.org/10.17175/2020_006
13. Shehzadi, T., Stricker, D., Afzal, M.Z.: A hybrid approach for document layout analysis in document images. In: Barney Smith, E.H., Liwicki, M., Peng, L. (eds.) Document Analysis and Recognition - ICDAR 2024. pp. 21–39. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-70546-5_2
14. Souter, C.: Towards a standard format for parsed corpora. Tech. Rep. 93.5, University of Leeds, School of Computer Studies (Jan 1993)
15. TEI Consortium, eds.: TEI P5: Guidelines for electronic text encoding and interchange, version 4.4.0 (2022), https://www.tei-c.org/Guidelines/P5/
16. Thompson, H., McKelvie, D.: Hyperlink semantics for standoff markup of read-only documents. In: SGML Europe. Graphical Communications Association (1997)
17. Zehe, A., Konle, L., Guhr, S., Dümpelmann, L., Gius, E., Hotho, A., Jannidis, F., Kaufmann, L., Krug, M., Puppe, F., Reiter, N., Schreiber, A.: Shared Task on Scene Segmentation@KONVENS 2021. In: Proceedings of the Shared Task on Scene Segmentation (Sep 2021)
18. Zhong, Z., Wang, J., Sun, H., Hu, K., Zhang, E., Sun, L., Huo, Q.: A hybrid approach to document layout analysis for heterogeneous document images. In: Fink, G.A., Jain, R., Kise, K., Zanibbi, R. (eds.) Document Analysis and Recognition - ICDAR 2023. pp. 189–206. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-41734-4_12