



Processing Large-Scale Archival Records: The Case of the Swiss Parliamentary Records

Journal Article

Author(s):

Salamanca Miño, Luis ; Brandenberger, Laurence ; Gasser, Lilian; Schlosser, Sophia; Balode, Marta; Jung, Vincent; Perez-Cruz, Fernando; Schweitzer, Frank

Publication date:

2024

Permanent link:

<https://doi.org/10.3929/ethz-b-000666852>

Rights / license:

[Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International](#)

Originally published in:

Swiss Political Science Review, <https://doi.org/10.1111/spsr.12590>







Funding acknowledgement:

184963 - Analyzing Co-Sponsorship Networks from 127 Years of the Swiss Federal Assembly (SNF)

RESEARCH NOTE



Processing Large-Scale Archival Records: The Case of the Swiss Parliamentary Records

Luis Salamanca¹  | Laurence Brandenberger²  | Lilian Gasser¹ |
 Sophia Schlosser²  | Marta Balode² | Vincent Jung²  |
 Fernando Perez-Cruz¹  | Frank Schweitzer² 

¹SDSC, Switzerland

²ETH Zurich, Switzerland

Correspondence

Laurence Brandenberger, ETH Zurich,
 Weinbergstr. 56, 8092 Zürich.
 Email: lbrandenberger@ethz.ch

Funding information

Schweizerischer Nationalfonds zur
 Förderung der Wissenschaftlichen
 Forschung, Grant/Award Number: 184963;
 Swiss Data Science Center, ETH Zürich;
 A Research Platform for Data-Driven
 Democracy Studies

Abstract

Legislative bodies generally keep records of their activities. While the digitization wave spurred the availability of archival documents, their processing remains a challenge. The Swiss parliamentary records are no exception.

In this paper we present a supervised pipeline for extracting and structuring of content of archival records. Our pipeline consists of five steps, starting with an assessment of which elements need extraction and how they relate to each other. Step two involves general pre-processing to prepare the PDF documents and is followed by an element classification step. Step four involves post-processing and the final step is a validation of the extracted information. With our supervised approach, we are able to process over 200,000 pages of Swiss parliamentary records (spanning the years 1891–1995), a feat that would exceed the budget of most projects using manual curation. We discuss validation of individual steps and offer guidance to researchers engaged in similar data processing efforts.

KEYWORDS

archival records, parliamentary proceedings, Swiss parliament, text processing, text-to-data

This article is supported by the SDSC Grant entitled ‘A Research Platform for Data-Driven Democracy Studies in Switzerland’ as well as the SNF-Grant entitled ‘Analyzing Co-Sponsorship Networks from 127 Years of the Swiss Federal Assembly’ (grant nr. 184963). We would like to thank the Swiss Federal Archives (Dr. Stefan Nellen) as well as the Parliamentary Services for their efforts in digitizing the records of the Swiss parliament. We also thank our student assistant—Clemens Hutter—for his engagement and effort in helping us parse the documents. We further thank the Editor and the reviewers for helpful criticism and guidance. The data and script that support the findings of this paper and contain hands-on guidance for researchers working on processing archival records are openly available in our online repository <https://renkulab.io/projects/luis.salamanca/processing-large-records>.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Authors. *Swiss Political Science Review* published by John Wiley & Sons Ltd on behalf of Swiss Political Science Association.

Zusammenfassung

Die gesetzgebenden Organe führen im Allgemeinen Aufzeichnungen über ihre Tätigkeit. Obwohl die Digitalisierungswelle die Verfügbarkeit von Archivadokumenten gefördert hat, bleibt ihre Bearbeitung eine Herausforderung. Die Schweizer Parlamentsakten sind keine Ausnahme. In diesem Beitrag stellen wir eine supervised Pipeline für die Extraktion und Strukturierung von Inhalten aus solchen Archivadokumenten vor. Unsere Pipeline besteht aus fünf Schritten, beginnend mit einem Assessment, welche Elemente extrahiert werden müssen und wie sie zueinander in Beziehung stehen. Der zweite Schritt umfasst eine allgemeines pre-processing zur Vorbereitung der PDF-Dokumente, gefolgt von einem Schritt zur Elementklassifizierung. Der vierte Schritt umfasst das post-processing und der letzte Schritt ist eine Validierung der extrahierten Informationen. Mit unserem supervised Ansatz sind wir in der Lage, über 200.000 Seiten Schweizer Parlamentsakten (aus den Jahren 1891–1995) zu verarbeiten, eine Leistung, die das Budget der meisten Projekte mit manueller Kuratation übersteigen würde. Wir erörtern die Validierung der einzelnen Schritte und bieten Forschenden, die sich mit ähnlichen Datenverarbeitungsprozessen beschäftigen, eine Anleitung.

Résumé

Les organes législatifs conservent généralement des archives de leurs activités. Si la vague de numérisation a stimulé la disponibilité des documents d'archives, leur traitement reste un défi. Les archives parlementaires suisses ne font pas exception. Dans cet article, nous présentons un pipeline supervisé pour l'extraction et la structuration du contenu de ces documents d'archives. Notre pipeline se compose de cinq étapes, commençant par une évaluation des éléments à extraire et de leurs relations entre eux. La deuxième étape consiste en un prétraitement général pour préparer les documents PDF et est suivie d'une étape de classification des éléments. La quatrième étape concerne le post-traitement et la dernière étape est une validation des informations extraites. Grâce à notre approche supervisée, nous sommes en mesure de traiter plus de 200 000 pages de documents parlementaires suisses (couvrant les années 1891–1995), un exploit qui dépasserait le budget de la plupart des projets utilisant la curation manuelle. Nous discutons de la validation des étapes individuelles et offrons des conseils aux chercheurs engagés dans des efforts similaires de traitement des données.

Riassunto

Gli organi legislativi generalmente conservano i documenti delle loro attività. Sebbene la digitalizzazione abbia favorito la disponibilità di documenti d'archivio, il loro trattamento rimane una sfida. I documenti parlamentari svizzeri non fanno eccezione. In questo lavoro presentiamo un canale sorvegliato (“supervised pipeline”) per l'estrazione e la strutturazione del contenuto di tali documenti d'archivio. Il nostro canale consiste in cinque fasi, di cui la prima comporta una valutazione degli elementi da estrarre e delle loro relazioni reciproche. La seconda fase prevede una pre-elaborazione generale per la preparazione di documenti PDF ed è seguita da una fase di classificazione degli elementi. La quarta fase riguarda la post-elaborazione, e la fase finale è la validazione delle informazioni estratte. Con questo approccio siamo in grado di elaborare oltre 200.000 pagine di documenti parlamentari svizzeri (che coprono gli anni 1891-1995), un'impresa che supererebbe il budget della maggior parte dei progetti che utilizzano la gestione manuale. Discutiamo la validazione delle singole fasi e offriamo una guida ai ricercatori e alle ricercatrici impegnate in questo tipo di elaborazione dei dati.

INTRODUCTION

The digitization wave has changed the way we view and handle archival data. Records and proceedings are scanned at large volumes and made public via online portals (Jensen et al., 2012; Michel et al., 2011; Owens & Padilla, 2021; Solberg, 2012). These documents are a promising data source to address a magnitude of research questions, in particular in the social sciences. However, their efficient and reliable processing has proven to be a challenge. In this paper we present a processing chain for archival documents. The goal of the paper is to offer guidance to researchers tackling large-scale data extraction projects. We detail our pipeline on the extraction of data from over 200,000 pages of Swiss parliamentary records spanning the years 1891–1995. Our extraction efforts result in a large-scale database of parliamentary activity, including information on proposed bills, how they were debated, votes and committee activities spanning over 100 years of archived documents.

Our proposed pipeline for extracting data from scanned documents consists of five steps, illustrated in Figure 1. In step 1, we propose an assessment of the content of the PDF documents. This provides a roadmap as to which information is lifted from the PDF pages and which information discarded. In step 2, PDF documents are pre-processed in order to obtain clean XML-files. This step entails the detection of text margins, column splits for the text, the detection of lines, cleaning of smudges and correcting the text flow. In step 3, text boxes are classified. We show how machine learning tools can help the annotation process of text boxes. A learning model can assist researchers by offering suggestions as to which category a text box belongs to as the annotation process unfolds. In the case of large-scale projects, it even allows researchers to only annotate a small percentage of the records—in our case 1 percent—and obtain predictions for all the non-annotated records. In step 4, post-processing tasks are performed to improve the resultant database. Step 5 consists of a thorough validation of the

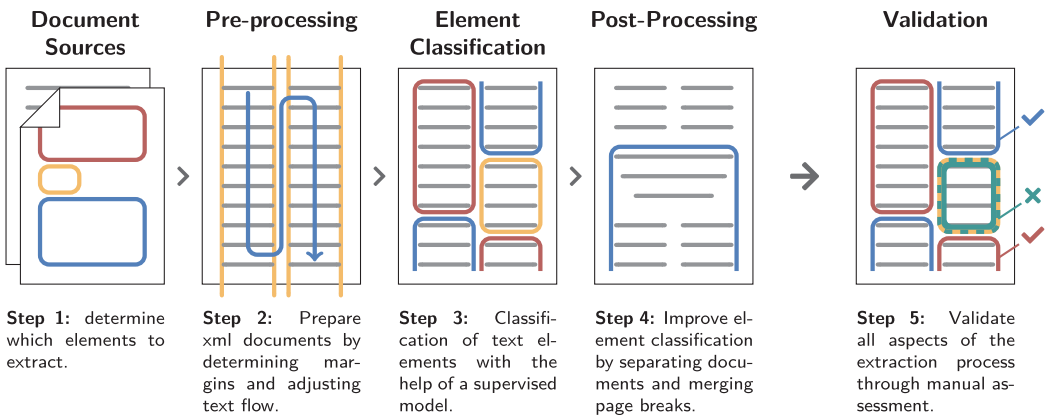


FIGURE 1 Pipeline steps: Step 1 entails a detailed assessment of the content and structure of the document sources, and the definition of the main information to extract. Step 2 tackles the pre-processing of PDF documents to extract clean and workable XML-files. Step 3 approaches the problem of element classification. Step 4 entails post-processing steps. Step 5 involves a validation procedure.

extraction efforts. This validation is performed by comparing computer-based outcomes with hand-annotated outcomes, allowing for a quantification of extraction errors.

Our illustration of the pipeline is built upon the parliamentary records of the Swiss national parliament. The Swiss Parliament has been recording its activities since 1891 in the form of a stenographic record of proceedings. This stenographic record (called *Amtliches Bulletin*) contains debates, verbatim, as well as the accompanying legislative proposition texts and votes. Recent efforts in digitizing these proceedings resulted in over 35,000 scanned PDF documents,¹ adding up to more than 200,000 pages. Our efforts result in a queryable database of Swiss parliamentary activities that has never been available to researchers for quantitative and computational analyses before.

By expanding our data requisition efforts we gain new insights in legislative politics. A more cost-effective extraction pipeline allows researchers to process more archival records, allowing for (i) longer time-spans in the analyses or (ii) a comparative perspective. For instance, we can gauge historical levels of polarization to better understand under which circumstances parliamentary polarization in- and decreases (e.g., Bornschier, 2019; Goet, 2019; Schlosser et al., 2023); we can study professionalization dynamics of members of parliament to better understand what difference the level of professional engagement of MPs makes on legislative outcomes (e.g., Schlosser et al., 2023); or we can examine how institutional reforms have affected legislative outcomes and improved representation (e.g., Childs, 2023; Sieberer et al., 2011). By employing a supervised machine-learning approach to the extraction of data from archival records, we are able to cut down on time and resources and thus expand the time horizon on data projects.² Our paper presents a supervised pipeline and offers guidance for such large-scale data acquisition efforts.

¹By PDF documents we refer to hard copies that have been scanned, and then the text extracted by optical character recognition software, without further formatting. Therefore, these documents are searchable, but they have no structure beyond that extracted by the OCR software.

²For instance, the *Amtliches Bulletin* consists of 207,417 pages. If we were to manually extract the content of these pages into a spreadsheet (by manually selecting the text, copying it to a spreadsheet and labeling the elements), we would expect to invest 12,167 hours (1,521 full working days (8 hr)). Two of the authors conducted a trial and manually extracted random pages from the AB. On average, it took them 211 seconds (3.5 minutes) per page. This includes identifying text elements and labeling them. The problem with manual extraction is that the extracted text entities are not linked to each other and thus not usable for research without additional processing.

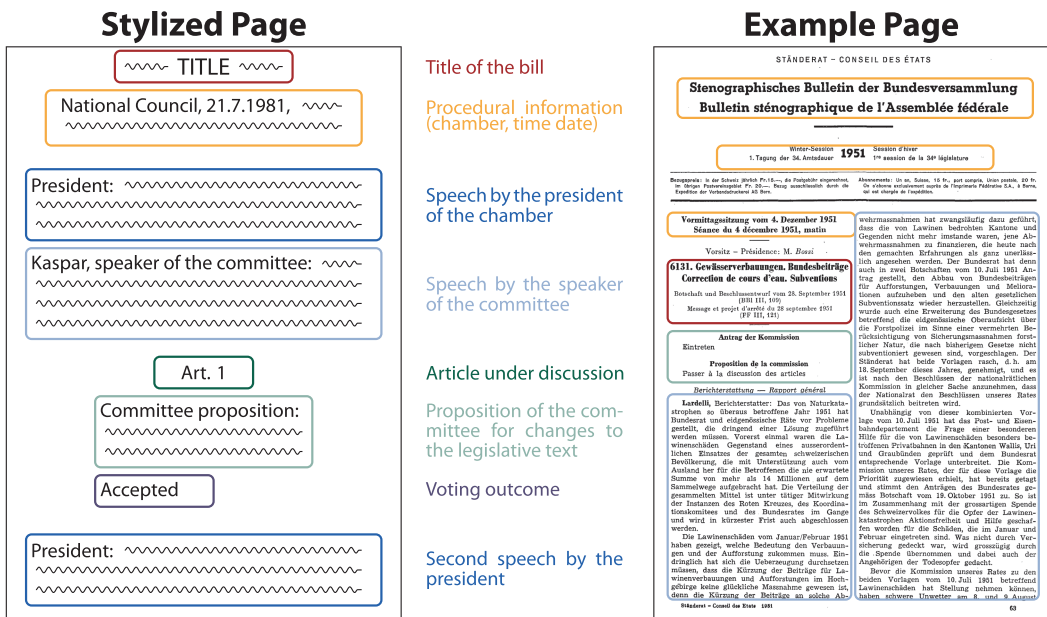


FIGURE 2 Schema of the information contained in the AB. Apart from speeches (blue), the AB contains detailed procedural information (yellow) as well as legislative propositions (green) and votes (violet). In reality, every page contains a set of these elements but in different arrangements (see right panel).

STEP 1: ASSESSING THE STRUCTURE OF ARCHIVAL RECORDS

When tackling a project processing archival records, we propose to start with an assessment of the elements contained on the pages. **Figure 2** depicts the elements contained on the pages of the AB: (i) speeches of MPs (verbatim), (ii) information on bill texts and wordings of legislative propositions and articles (e.g., in detailed discussions on federal enactment drafts), and (iii) voting outcomes. These three elements are complemented with section headers, procedural information, footnotes and separating lines.

Incidentally, most of these elements are present in parliamentary proceedings around the world. France and Germany, for instance, both report proceedings in PDF form in chronological order, i.e., listing speeches alongside parliamentary proposals and voting outcomes in their PDF records. Moreover, state legislative bodies often adopt similar styles in reporting the minutes of their plenary sessions, with the most prominent elements being the speeches, followed by legislative propositions and voting outcomes (see for instance the proceedings of the cantonal parliament of Basel-Stadt); and even the UN General Assembly presents their session minutes in a similar form.

While the elements are similar, the structure of the pages can look very different. **Figure 3** shows an example of how a discussion of an interpellation proposed in 1944 is reported in the AB: procedural information (yellow) is reported next to bill identifiers (title, sponsor, submission date; dark blue) and the submitted bill (light green), followed by a speech (red). **Figure 3**, on the other hand, shows a discussion of a federal decree draft and has a more complicated structure including propositions (dark green article header, titles in dark blue and texts in light green) and speeches (red). Some proposed changes are accompanied by speeches (red) and are often directly followed by a report of the voting outcome (magenta).

The structure and relation of the elements differ greatly, not only across time, but also across different forms of parliamentary discussions. Essentially, every page in the AB is unique

<p>Vormittags-sitzung vom 15. Juni 1944. Séance du 15 juin 1944, matin.</p> <p>Vorsitz — Présidence: Hr. Gysler.</p>	<p>Art. 47.</p> <p>Antrag der Kommission.</p> <p>Die Vorschriften des Zivilgesetzbuches über die Pfandstelle beim Grundpfand (Art. 813—815) finden auf die Schiffsverschreibung entsprechende Anwendung.</p>
<p>4515. Interpellation Duttweiler vom 24. März 1944.</p>	<p>Proposition de la commission.</p> <p>Les dispositions du code civil sur la case hypothécaire en matière de gage immobilier (art. 813 à 815) sont applicables à l'hypothèque sur bateau.</p>
<p>Verfassungsmässige Rechte der Bürger. Rétablissement des libertés constitutionnelles.</p>	<p>Geel, Berichterstatter der Kommission: Zustimmung zum Bundesrat und den gedruckten redaktionellen Ergänzungen.</p>
<p>Ist der Bundesrat nicht auch der Auffassung, es seien die Einschränkungen der verfassungsmässig gewährleisteten Freiheitsrechte der Bürger (Versammlungsverbot, Verbot neuer Presseorgane, Verbot von politischen Parteien) sowie die Handhabung der Zensur für Erzeugnisse der Druckerpresse, sofort zu lockern mit dem Ziel, auf diesem Gebiete die Verfassung wieder vollumfänglich in Kraft zu setzen?</p>	<p>Angenommen. — Adopté.</p> <p>Art. 48.</p> <p>Antrag der Kommission.</p>
<p>Le Conseil fédéral n'est-il pas d'avis qu'il y aurait lieu d'assouplir sans délai la restriction des libertés constitutionnelles (interdiction d'assemblées, interdiction de nouveaux organes de presse, interdiction de partis politiques), comme aussi la censure de la presse, afin de rétablir dans ce domaine l'application intégrale de la constitution.</p>	<p>Zustimmung zum Entwurf des Bundesrates.</p> <p>Proposition de la commission.</p> <p>Adhésion au projet du Conseil fédéral.</p> <p>Geel, Berichterstatter der Kommission: Zustimmung zum Bundesrat.</p>
<p>Die Interpellation wird unterstützt von den Herren — La demande d'interpellation est appuyée par MM.:</p>	<p>Angenommen. — Adopté.</p> <p>Art. 49.</p> <p>Antrag der Kommission.</p>
<p>Barben, Eggenberger, Gadiet, Lanicca, Leupin, Maag, Munz, Sappeur, Spindler, Sprecher, Trüb, Zigerli. (12)</p>	<p>Zustimmung zum Entwurf des Bundesrates.</p> <p>Proposition de la commission.</p>
<p>Duttweiler: Gestern hat Herr Zellweger seine Interpellation begründet. Sie war abgestellt auf einen einseitigen parteipolitischen Zweck, während unsere Auffassung dahingeht, es solle die Gleichberechtigung auf politischen Gebiet wieder hergestellt werden. Wir lehnen eine Politik des „je nach dem und des je nach wem“ in dieser Hauptfrage ausdrücklich ab. Man wird kaum behaupten wollen, dass die frontistischen und nationalsozialistischen Bestrebungen in der Schweiz heute noch irgendwie aktuell wären. Und wenn sie es auch wären, so würde das nicht verhindern, dass wir diesem Grundsatz der Gleichberechtigung, diesem fundamentalen Grundsatz dennoch treu bleiben würden.</p>	<p>Adhésion au projet du Conseil fédéral.</p> <p>Geel, Berichterstatter der Kommission: Art. 49 entspricht dem Art. 822, Abs. 1 und 2, des Zivilgesetzbuches. Zustimmung zum Bundesrat.</p>
<p>Von dieser Auffassung ausgehend, muss ich etwas weiter ausholen. «La liberté, c'est un bloc» hat ein französischer Parlamentarier mit Recht gesagt. Die Wirtschaftsfreiheit ist ein Teil dieses Blockes. Deshalb beginne ich mit einer Übersicht über den Zerfall auf dem Gebiet der wirtschaftlichen Freiheiten.</p>	<p>Angenommen. — Adopté.</p> <p>Art. 50.</p> <p>Antrag der Kommission.</p>
<p>Die Geschichte lehrt uns, dass Krieg und Krise von jeher freiheitszerstörend gewirkt haben. Im</p>	<p>Zustimmung zum Beschluss des Bundesrates.</p> <p>Proposition de la commission.</p> <p>Adhésion au projet du Conseil fédéral.</p>
<p>(a) Bill #4515, 1944</p>	<p>Geel, Berichterstatter der Kommission: Hier finden wir zusammengefasst die Bestimmungen 832 und 834 des Zivilgesetzbuches, deren Grundsätze über Eigentum und Schuldnerschaft bei Veräusserung eines verschriebenen Schiffes sowie über die Anzeige der Schuldübernahme wichtig genug sind, um hier</p> <p>(b) Bill #1673, 1923</p>

FIGURE 3 Example columns from the AB. Three different elements can be extracted: information on bill texts and written propositions (light green); speeches (red) and voting outcomes (magenta). Additionally, procedural information (dates = yellow; president of the session = dark red) and bill titles (dark blue) can be distinguished. Every page has a different layout with a combination of these elements.

in how the different elements are presented. This makes extraction of data via hard-coded rules impossible. We, therefore, propose a more flexible way of extracting the data in Step 3, after pre-processing the documents.

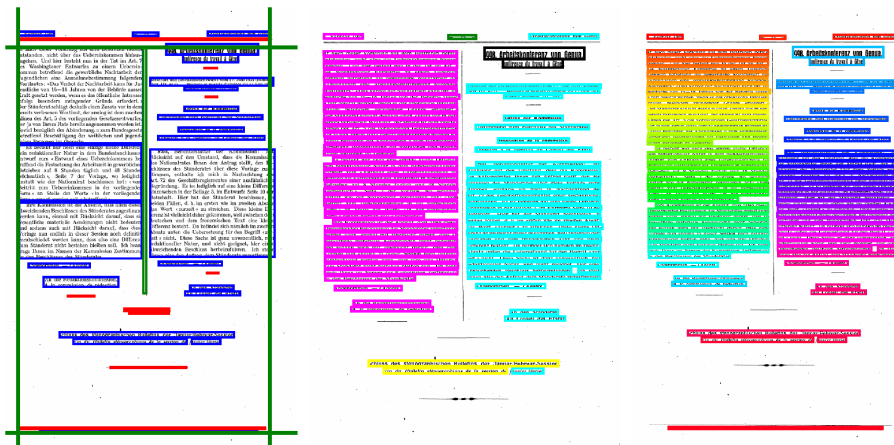


FIGURE 4 Visual inspection of pre-processing step: Left panel: Dividing pages into text boxes (dark blue) and margins (green lines). Middle panel: Separating text into one column (dark blue) and two columns (pink and light blue). Right panel: Assessing text ordering with the help of a rainbow scale.

STEP 2: PRE-PROCESSING PDF DOCUMENTS

Before extracting information from records, the PDF documents need to be pre-processed. The scanned documents need to run through an OCR routine. This entails that the text and layout information from each PDF page can be extracted into an XML-file or equivalent. In our case, the OCR was performed on all scanned documents by the Swiss Federal Archives. Alternatively, Tesseract is an easy-to-use OCR tool (Smith, 2007).

While most XML-files have some inherent structure, the ordering of text and grouping into text boxes is not always present (or coherent). Furthermore, archival documents often contain stains and smudged text. These stains from the scanning process were also commonly identified as valid characters, causing problems with the existing text on the pages.

We recommend to pre-process XML-files. First, we ensure that margins are accurately assigned and black margins from open-book scanning are removed. Second, horizontal, central and vertical lines are detected. In Figure 4, left panel, we can observe the main elements identified, green lines for margins and central column separator, and red boxes for horizontal lines. Third, we ensure the correct separation of text lines (see middle panel in Figure 4). Fourth, it is essential that textlines are ordered correctly (see right panel in Figure 4). As archival records often contain stains that OCR scans cannot distinguish from actual text, they can cause the text flow from breaking.³

Performance Assessment of Pre-processing Steps

An important aspect in processing archival data is the continued validation. Therefore, we assess whether our pre-processing steps result in clean XML-files. For each year, we draw 100 random pages, resulting in a sample of 10,500 pages ready for manual assessment. We assess five different performance aspects: (i) whether or not text lines are correctly classified as

³These four steps are quite universal and do not depend on the structure of the PDFs. We recommend a careful pre-processing of XML-files to everyone working with archival records. By ensuring the text lifted off the PDF documents is clean and ordered, the classification of text into actual data is less error prone. In the SI we provide methodological details on all pre-processing tasks. In our online repository (<https://renkulab.io/projects/luis.salamanca/processing-large-records>) we provide code for these four steps.

spanning the full page (one-column lines) or only one column (two-column lines); (ii) whether the text flow has been adequately captured; (iii) whether there is text missing; (iv) whether text boxes are correctly split between different elements, and (v) whether horizontal lines on the page have been identified correctly (see detailed description and assessment in the SI). Out of the 10,500 random documents, 48 pages report at least one error. Since the pages are randomly sampled, this results in an error of 0.5%, i.e., we project that 99.5% of documents are correctly pre-processed.

STEP 3: ELEMENT CLASSIFICATION

A crucial step processing archival records involves separating information contained on pages. In the AB we want to separate speeches, votes and legislative propositions from one another.

We face the problem that every PDF page is unique in terms of the information they contain and how text boxes are arranged. So, how can we label each text box accurately? The simplest solution is to annotate the text boxes by hand, i.e., assign a ‘speech’-label to every text box that is part of a speech. However, hand-annotation is labor-intensive and with growing corpus size not always feasible. A second solution is to rely on machine learning.

Machine learning techniques span algorithmic and statistical tools and are often used for making predictions (for a primer, see Grimmer et al., 2021). One class of tools are supervised learning models (Bishop, 2006). They are used to make label-predictions for every segment, in our case a text box. To make the predictions, the model relies on similarities between text boxes. Thus, for every text box, we compute a set of features to characterize the text box. We then feed the model with hand-annotated text boxes. Out of these hand-annotated samples, we let the model learn commonalities between text boxes and elements, e.g., learning if speeches always start with a specific word or if vote text boxes contain a specific set of font types. The more hand-annotated text boxes are fed into the model, the better the label-predictions of text boxes becomes. We determine how good the model is at predicting the labels of text boxes by assessing the models' performance. For the performance assessment, we train the model on part of the hand-annotated data, and use the rest to evaluate the quality of the training.

Feature extraction. For all text boxes in the training and test set, we create a set of features with the following characteristics:

- **Visual information:** size of the bounding box, the position of the text boxes and its level of indentation.
- **General text:** count of character types (numbers, letters and punctuation), length of the text boxes in characters, and formatting features related with the size and font type.
- **Vocabulary features:** occurrence of specific terms.

These features are general and can be applied to any PDF document. We have explored more specific features designed just for our use-case, but have found the more general features to be more powerful in sorting elements.

Learning model. In machine learning language our model is posed as a supervised classification problem where the model, a classifier, learns from hand-annotated text boxes which label to assign to un-annotated text boxes. For our case, we used a random forest classifier (Breiman, 2001) with a conditional random field (CRF) layer (Lafferty et al., 2001). This CRF layer takes previous and next elements into consideration when assigning labels. The classification model additionally incorporates regularization mechanisms to control for overfitting, and an exhaustive hyperparameters' tuning step that ensures achieving the best possible performance (see SI).

Assisted annotation and supervised learning. In principle, the more hand-annotated labels of text boxes we train our model with, the better the model is at separating the text boxes. To ease the work of the annotator, we implement a semi-automatic labelling approach where the user receives suggestions (see Section 3.2 in the SI for details on the pdf2data software tool). With the help of our pdf2data tool (Salamanca et al., 2023) we incrementally teach our model to discern subtle differences between the elements, improving step by step the global performance and accelerating the manual annotation process.

Assessing the model performance during annotation. It is impossible to specify beforehand the number of annotated text boxes we need in order to make satisfactory predictions of unseen text boxes. However, while annotating, we can perform cross-validation (CV) to assess how good our model is at distinguishing elements. CV first requires to split the annotated elements into a specific number of K folds, 5 in our case. Then, we select one fold to test the performance on, and we train the model with the data from the other $K - 1$ folds. We perform this process with all K folds, finally obtaining a prediction for each of the annotated elements in the training set. By comparing these prediction with the hand-annotated labels, we can check how well the model predicts the labels of the text boxes, e.g., how good the model is at labeling a speech text box with the label ‘speech’. With this estimate, a natural stopping point can be identified for the annotation process. Cross-validation also helps in identifying structural changes in the documents. Since the model learns to identify commonalities in both structure of the documents and content of the text boxes, changes in layout or formatting automatically lead to drops in prediction performance. Whenever these changes occur, the model has to be re-trained. In our case, our documents span over 100 years and layouts have changed slightly over the years. We have grouped our documents into similar layouts (based on model performance scores).

Generating predictions for all data. Once the annotator is satisfied with the model's performance, the model is used to predict the labels of all the text boxes that have not been annotated by hand yet. Combined with the hand-annotated text boxes, every text box on every page is thus assigned a label.

Assessing the final accuracy. In order to ensure a satisfactory performance of the model we also hand-annotate a separate test set, which helps identify over-fitting (Bishop, 2006). We predict the labels of the text boxes in the test set with our model and compare them to the hand-annotated labels. This is similar to assessing model performance above, with the distinction that the text boxes in the test set have not been used to train the model. It, therefore, makes for a more stricter test of our model's performance. With this, we can validate that the learning model is correctly trained by comparing the accuracy obtained in the CV with the accuracy on the test set, and that the model is not over-fitting to the training data.

We report model performance for the Swiss parliamentary records in the SI.

STEP 4: POST-PROCESSING

After classifying elements it may be worth doing some post-processing on the extracted text before loading it into a database. This post-processing step is reserved for source-specific tasks and is thus the least generalizable in our pipeline. In our use case we perform three post-processing steps: assessing page overlaps (reported in the SI), merging text blocks and identifying parliamentary speakers.

Improving Merging of Discussions and Assessing Accuracy

With the help of the methodology described above, we are able to assign a label to each text box. Yet, for speeches, legislative propositions and votes, it is necessary to merge

consecutive text boxes into coherent blocks. For instance, a speech that spans over multiple pages is broken up into multiple text boxes (i.e., by paragraphs or page breaks). While the learning model provides label-predictions for each of these text boxes, it does not automatically merge all these text boxes into a single speech. Furthermore, if two speeches follow each other, we *cannot* merge all these speech boxes together, but rather have to separate the speeches by their speakers. For this, we implement an additional, source-specific, post-processing step (see the SI).

Linking Speeches to Speakers

We also identify the speaker of each speech and try to match him or her with a database of MPs and Federal Council members. Out of 3371 speeches analyzed during manual validation on a small subset of documents, 252 (7.6%) have speakers that cannot be linked precisely to one of our MPs in our names database.

STEP 5: VALIDATION OF EXTRACTED DATA

The final step in any processing chain has to be an extensive validation of the extracted data. We propose to hand-curate a database and compare it to the data extracted from the semi-automated processing. That way, the accuracy of the extracted data can be assessed effectively.

In our case, we hand-labeled speeches, votes and legislative propositions on a random sample of documents. We sampled 1% of all documents, with a minimum of 2 documents per year. In total, we evaluated 398 documents (3,251 pages). For each document we assess (i) whether the number of speeches (counted by the number of times different MPs speak during a discussion) is correct, (ii) how many speeches are not identified, and whether the (iii) legislative propositions and (iv) votes are identified correctly (see [Figure 5](#)).

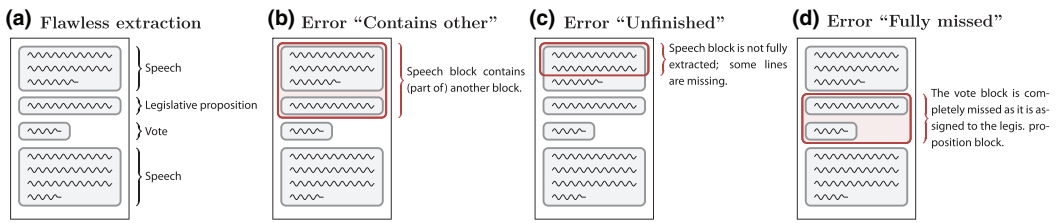
Out of 3,371 speeches, 3205 (95%) are extracted without error. 75 speeches are unfinished, i.e., not all text blocks are assigned to the same speech. For legislative proposals (LP), only 87% are extracted without error. Here, too, most errors stem from LPs being only partially extracted. For votes we reach an accuracy of 92%. Here, most errors are due to votes being completely missed. This is common as sometimes votes consist only of the word ‘Angenommen’ (accepted), which is easily missed (see SI for details).

DISCUSSION AND CONCLUSION

Quantitative and computational political science relies on high-quality data. This rings especially true for legislative studies. While Western democracies often have long traditions of recording activities of their legislative branch, these records have not (yet) been brought to their full potential. With our pipeline, we were able to overcome three limitations in parliamentary data available for research today.

First, our data extraction embraces data complexity by linking different parliamentary activities with each other, e.g., linking speeches to bills. Existing databases often focus on a single aspect of the legislative process. The Hansard databases,⁴ for instance, reports proceedings and assigns them to the speakers but does not link speeches to bills. The US Congress provides access (from 1994 onward) to all political speeches on a variety of bills discussed on specific days through so-called

⁴For the British parliament, see <https://hansard.parliament.uk>.



	#	a Flawless	b Contains other	c Unfinished	d Fully missed
Speeches	3371	95.1% (3205)	1.3% (44)	2.2% (75)	0.5% (18)
LP	990	86.8% (859)	2% (20)	7.8% (77)	3.5% (35)
Votes	1128	91.5% (1032)	1.1% (12)	1.4% (16)	6.1% (69)

FIGURE 5 Validation of information extraction. Flawless extraction refers to entities that are fully and correctly identified with the correct label. We record three types of errors in our validation process: if an element contains (parts of) another entity; if an element is unfinished or if an element is fully missed. Assessment is based on visual inspection of 398 randomly sampled documents. Multiple errors per entity are possible.

‘daily editions’.⁵ However, these daily editions do not disambiguate the speakers, nor relate the speeches to a bills database. A third example is the French National Assembly which provides records on parliamentary debates as far back as 1958 in the form of unstructured XML-documents.⁶

Second, all available records, to the best of our knowledge, neglect the legislative propositions under discussion. Our extraction pipeline pays special attention to this aspects of parliamentary discussions. Parliamentary speeches are often highly structured and follow a given logic. While some parliaments know the format of ‘free-form’ speeches (such as the one-minute speeches in the US House), most speeches revolve around legislative changes that are discussed at a detailed level. Often, these proposed changes are ignored in the creation of databases on parliamentary speeches. In this paper, we make a case for separating these elements from one another in the extraction and then linking them back together in a structured database.

Third, with our (semi-)automated pipeline we can process large corpora, resulting in a database that spans over 100 years. Most available databases today do not cover such a long time range. The US Congress provides extensive data on legislative activities (cosponsorship, committee activities, bills), starting with the 97th Congress (1981). While they provide some data before 1981, their records are not complete, e.g., not reporting the full list of proposed amendments (see Fowler, 2006, p. 460). The French National Assembly reports on MPs as far back as 1997, yet their reports on bills and votes are relatively new, starting only from 2017.

We processed over 100 years worth of documents and structure them into a coherent database. We are able to extract over 95% of all parliamentary speeches without error. These include speeches that are disrupted by plenary votes or records of legislative propositions. They also reflect speeches of differing lengths, from speeches that are 2 lines long to speeches that span multiple pages. With our pipeline, we are able to not only extract the individual elements, we are also able to link elements together. As such, our database comprises not simply a list of bills that Swiss MPs have tackled over the years, but it also links those bills to the parliamentary speeches, legislative propositions or to votes. We believe the potential of the extracted dataset is yet to be unveiled.

We hope our paper provides guidance for similar efforts in processing archival records and spurs researchers into updating databases to better account for the complexity of the political process.

⁵<https://www.congress.gov/congressional-record>

⁶<https://data.assemblee-nationale.fr/travaux-parlementaires/debats>

ACKNOWLEDGMENTS

We would like to thank the Swiss Federal Archives (Dr. Stefan Nellen) as well as the Parliamentary Services for their efforts in digitizing the records of the Swiss parliament. We also thank our student assistant Clemens Hutter for his engagement and effort in helping us parse the documents. We further thank the Editor and the reviewers for helpful criticism and guidance.

FUNDING INFORMATION

This article is supported by the SDSC Grant entitled ‘A Research Platform for Data-Driven Democracy Studies in Switzerland’ as well as the SNF-Grant entitled ‘Analyzing Co-Sponsorship Networks from 127 Years of the Swiss Federal Assembly’ (grant nr. 184963).

OPEN RESEARCH BADGES



This article has earned an Open Materials badge for making publicly available the components of the research methodology needed to reproduce the reported procedure and analysis. All materials are available at <https://renkulab.io/projects/luis.salamanca/processing-large-records>.

DATA AVAILABILITY STATEMENT

The data and script that support the findings of this paper and contain hands-on guidance for researchers working on processing archival records are openly available in our online repository (<https://renkulab.io/projects/luis.salamanca/processing-large-records>).

ORCID

Luis Salamanca  <https://orcid.org/0000-0001-9314-8466>

Laurence Brandenberger  <https://orcid.org/0000-0003-0392-9766>

Sophia Schlosser  <https://orcid.org/0000-0003-1107-928X>

Vincent Jung  <https://orcid.org/0009-0005-6731-1553>

Fernando Perez-Cruz  <https://orcid.org/0000-0001-8996-5076>

Frank Schweitzer  <https://orcid.org/0000-0003-1551-6491>

REFERENCES

- Bishop, C. M. (2006). Pattern recognition. *Machine learning*, 128(9).
- Bornschieer, S. (2019). Historical polarization and representation in south american party systems, 1900–1990. *British Journal of Political Science*, 49(1), 153–179.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32.
- Childs, S. (2023). Feminist institutional change: the case of the uk women and equalities committee. *Parliamentary Affairs*, 76(3), 507–531.
- Fowler, J. H. (2006). Connecting the congress: A study of cosponsorship networks. *Political Analysis*, 14(4), 456–487.
- Goet, N. D. (2019). Measuring polarization with text analysis: Evidence from the uk house of commons, 1811–2015. *Political Analysis*, 1–22.
- Grimmer, J., Roberts, M. E., & Stewart, B. M. (2021). Machine learning for social science: An agnostic approach. *Annual Review of Political Science*, 24, 395–419.
- Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6), 395–405.
- Lafferty, J. D., McCallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning* (p. 282–289). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.
- Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., the Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvic, P., Orwant, J., Pinker, S., Nowak, M. A., & Lieberman Aiden, E. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014), 176–182.
- Owens, T., & Padilla, T. (2021). Digital sources and digital archives: historical evidence in the digital age. *International Journal of Digital Humanities*, 1(3), 325–341.

- Salamanca, L., Meyer, M., Brandenberger, L., Schlosser, S., Campos-Schweitzer, J., Schweitzer, F., & Perez-Cruz, F. (2023). *Pdf2data: A tool for extracting information from pdf documents*. London, UK: Paper presented at the 4th European Political Methodology Meeting (PolMeth Europe), King's College, London, UK, June 19-20, 2023.
- Schlosser, S., Minder, J., & Brandenberger, L. (2023). The evolution of parliamentary polarization in Switzerland: An over century-long perspective. *Paper presentation at the ECPR Standing Group of Parliaments Conference, Vienna, AU, July, 6–9(2023)*, 1–11.
- Schlosser, S., Brandenberger, L., Minder, J., Russo, G., Salamanca, L., & Schweitzer, F. (2023). From expertise to versatility: The evolution of issue engagement in the Swiss Parliament over 130 years. *Paper presentation at the European Political Science Association 13th Annual Conference, EPSA, Glasgow, UK, June 22–24, 2023.*, 1–12.
- Sieberer, U., Müller, W. C., & Heller, M. I. (2011). Reforming the rules of the parliamentary game: Measuring and explaining changes in parliamentary rules in Austria, Germany, and Switzerland, 1945–2010. *West European Politics*, 34(5), 948–975.
- Smith, R. (2007). An overview of the tesseract ocr engine. In *Ninth international conference on document analysis and recognition (icdar 2007)* (Vol. 2, pp. 629–633).
- Solberg, J. (2012). Googling the archive: Digital tools and the practice of history. *Advances in the History of Rhetoric*, 15(1), 53–76.

DATASET REFERENCE

- [dataset] Salamanca, Luis, Brandenberger, Laurence, Lilian Gasser, Sophia Schlosser, Marta Balode, Vincent Jung, Fernando Pérez-Cruz and Frank Schweitzer. 2024. Files for “Processing Large-Scale Archival Records: The Case of the Swiss Parliamentary Records”. Online repository, available at: <https://renkulab.io/projects/luis.salamanca/processing-large-records>.

AUTHOR BIOGRAPHIES

Luis Salamanca is Lead Data Scientist at the Swiss Data Science Center, ETH Zürich. He holds a Ph.D. in Electrical Engineering from the University of Seville. His research focuses on machine learning and its applications to various fields, such as social sciences and design for architecture and engineering.

Laurence Brandenberger is a Senior Scientist at the Chair of Systems Design at ETH Zürich. She holds a Ph.D. in Political Science from the University of Bern and her research focuses on legislative politics, political networks and political methodology.

Lilian Gasser is a data scientist at the Swiss Data Science Center. She holds an MSc in Chemical Engineering from ETH Zürich and her research focuses on applied data science and visualization.

Sophia Schlosser is a Ph.D. candidate at the Chair of Systems Design at ETH Zürich. Her research focuses on legislative politics and political methodology.

Marta Balode holds a MSc in Neural Systems and Computation from ETH Zürich and the University of Zürich. She supported the Chair of Systems Design as a student data scientist.

Vincent Jung holds a MSc in Data Science from ETH Zürich. He supported the Chair of Systems Design as a student data scientist.

Fernando Perez-Cruz is a Titular Professor in the Computer Science Department at ETH Zürich and Deputy Executive Director and Chief Data Scientist at the Swiss Data Science Center. His research focuses on machine learning and AI.

Frank Schweitzer is a full professor for Systems Design at ETH Zürich. His research focuses on applications of complex systems theory in the dynamics of social and economic organizations.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Salamanca, L., Brandenberger, L., Gasser, L., Schlosser, S., Balode, M., Jung, V., Perez-Cruz, F. & Schweitzer, F. (2024). Processing Large-Scale Archival Records: The Case of the Swiss Parliamentary Records. *Swiss Political Science Review*, 00, 1–14. <https://doi.org/10.1111/spsr.12590>