David Schwelien
Stadtbachstrasse 58, 3012 Bern
davidschwelien@gmail.com

Data Science Project

Reducing Duplication and Accelerating Multilingual Publishing with AI

Using the Swissinfo article corpus to build an AI-based assistant for story idea comparison and language-specific adaptation at SRG

Conceptual Design Report

September 2025

Abstract

SRG newsrooms face increasing pressure to produce more with fewer resources, reflecting broader trends in the media industry. Swissinfo, with its unique multilingual setup, maintains a large corpus of multilingual content that can be leveraged to train and apply patterns in editorial adaptations. This project proposes an AI-based assistant to support editors in two key areas: (1) identifying overlapping story ideas by comparing new pitches with the existing corpus, and (2) generating multilingual adaptation drafts based on patterns learned from past articles. The goal is to reduce redundancy, accelerate publishing, and preserve editorial control.

Table of Contents

01 Project Objectives

SRG newsrooms, including Swissinfo, are under increasing pressure to produce high-quality content with fewer resources. By 2029, SRG plans to cut CHF 270 million—17% of its 2024 budget—highlighting the urgency of developing more efficient workflows (Blick, 2025; SRF, 2025).

One major inefficiency is the duplication of editorial effort across language services. SRG is particularly affected, as it operates four business units, each producing content tailored to Switzerland's different language regions. In the past, redundancies often arose, making it clear that reducing overlaps and streamlining production processes across language boundaries is a key area for potential savings.

However, this raises two crucial questions: First, how can SRG continue to preserve the distinctiveness and long-standing expertise of RTS, RSI, RTR, Swissinfo, and SRF while also cutting costs? Second, how can SRG ensure that a story covered in one part of Switzerland is not redundantly produced in another? In this context, Swissinfo's corpus is especially valuable. Publishing in ten languages and maintaining a large multilingual archive, Swissinfo provides unique resources for reducing duplication while safeguarding diversity.

This project aims to leverage Swissinfo's assets and proposes an AI-based assistant to support editorial teams at two critical stages:

- **Story idea comparison**: New pitches are automatically compared with the existing SRG corpus to identify thematic overlaps early—without requiring journalists to manually search the archive.
- **Multilingual adaptation**: Using patterns learned from Swissinfo's multilingual corpus, the system generates draft translations tailored to each language's style and audience expectations.

The envisioned tool will highlight overlapping coverage, suggest multilingual drafts, and route them for editorial approval—reducing duplication, accelerating publishing, and preserving editorial control. Ultimately, the goal is to support journalists, not replace them, and to ensure that quality journalism remains viable under resource constraints.

02 Methods

The project is developed in Databricks on Microsoft Azure, where the Swissinfo/SRG article corpus is already available as SQL tables. This environment enables a combination of SQL queries and Python-based text analysis.

The following tools and libraries will be used:

- **pandas** and **PySpark** for data cleaning and structuring.
- **NLTK** / **spaCy** for text preprocessing (tokenization, stopwords, POS tagging).
- **scikit-learn** for clustering and similarity detection.
- **Hugging Face Transformers** for multilingual adaptation suggestions.
- **matplotlib** / **seaborn** for visualizations.
- Optionally, the **DeepL API** will serve as a benchmark for translation quality.

The methodological approach is iterative:

1. **Data exploration and cleaning** to prepare the corpus.
2. **Baseline models** for similarity search and clustering.
3. **Pattern learning** from multilingual adaptations to generate draft translations.
4. **Evaluation** using translation metrics (e.g., BLEU, ROUGE) and editorial feedback, with a focus on practical usefulness over purely quantitative scores.

The final goal is to prototype an assistant tool that identifies duplicate story ideas and suggests multilingual drafts, integrated into editorial workflows.

03 Data

The project uses the Swissinfo/SRG article corpus, stored in Databricks on Microsoft Azure as SQL tables. The dataset contains thousands of published articles across ten languages, enriched with metadata and structured in a nested format.

Each article is represented as a complex object with fields such as:

- ***ID**: URN and SourceId*
- ***Title**: multilingual, stored as arrays*
- ***Content**: paragraphs in array format*
- ***Lead/Kicker**, **Contributors**, **Genres/Keywords***
- ***Publication and Modification Dates***
- ***Linked Resources**: images, videos, external links*
- ***Technical Metadata**: versioning hashes, timestamps*

Due to the nested JSON-like structure, preprocessing is required—e.g., flattening content fields and extracting language-specific titles—to make the data usable for analysis.

For a better understanding of the data, I am sharing some cells of an example entry of this database:

Data Base example entry

ID: urn:pdp:cms_swi:article:73104714
Title: Parliament dashes hopes of Swiss whistleblowers
Release Date: 2024-02-27 17:18:13
Content (excerpt): “Parliament dashes hopes of Swiss whistleblowers. Proposed laws to grant extra protection for Swiss whistleblowers have again been voted down by parliament. To understand the debate today, we need to look back: Four years ago [...] Translated from German by DeepL/mga”
Web link: https://www.swissinfo.ch/eng/swiss-politics/parliament-dashes-hopes
etc...

04 Metadata

All metadata is stored in Databricks Unity Catalog as part of the Delta table history and schema documentation. Additional metadata—such as analysis steps and parameter settings—is documented in the project's Databricks notebooks, which are versioned via GitHub.

All analyses and changes are transparently available in a public Git repository [link]

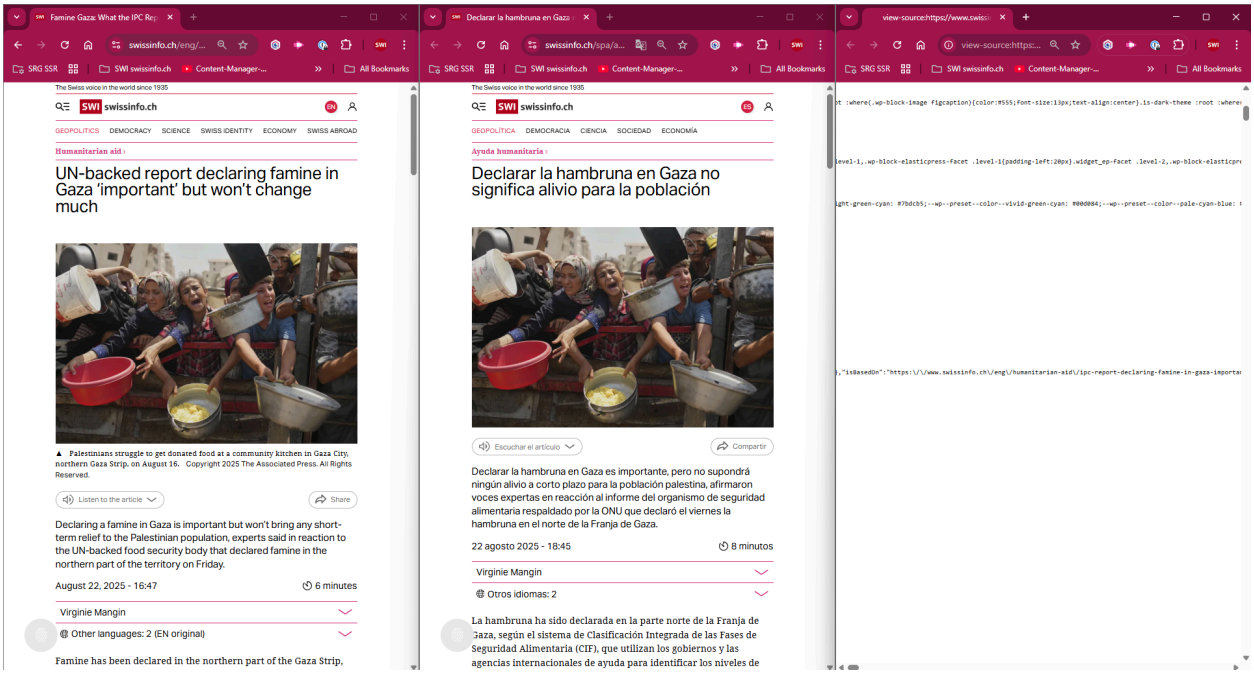
Note on missing metadata

A critical metadata element currently missing from the Databricks dataset is the reference to the original article on which an adaptation is based. This link is essential for the final product, as the goal (see *Section 1: Project Objectives*) is to generate multilingual drafts based on editorial patterns. Without knowing which article an adaptation is based on, these patterns cannot be reliably learned or applied.

For articles adapted within swissinfo.ch, this metadata is available in the HTML source code of the published page. Specifically, the "isBasedOn" field in the embedded JSON-LD structure contains the URL of the original article. This information is not present in the Databricks corpus but can be extracted via web scraping.

This screenshot shown in Figure 01 at the end of this chapter shows three browser tabs side by side: the original English article [link], the Spanish adaptation [link], and the HTML source of the Spanish adaptation [view-source on link]. In the source view, the "isBasedOn" field clearly links the adaptation to its original. This metadata is crucial for training adaptation models and should be systematically added to the dataset.

Figure 01: Screenshot of Three Browser Tabs Showing Source Mapping in Swissinfo Adaptations

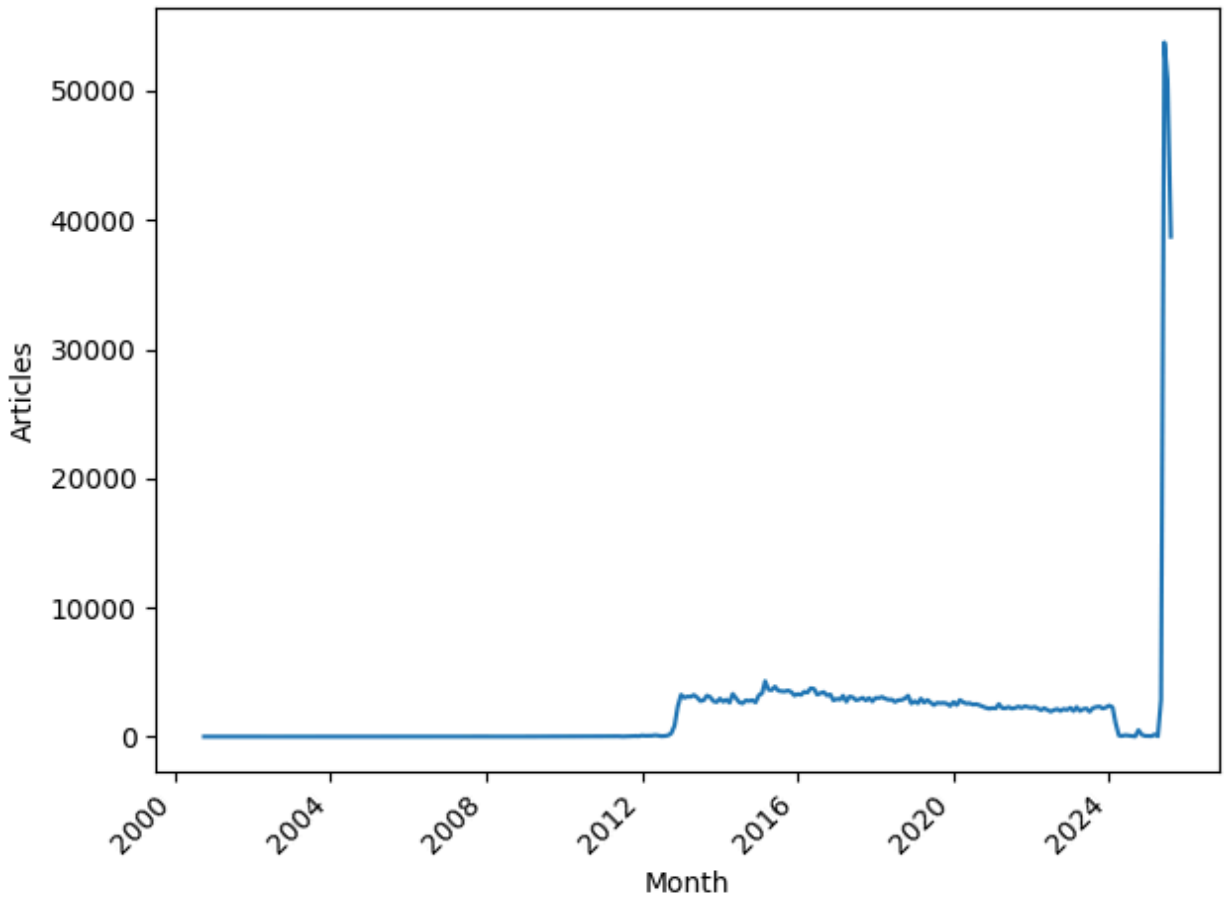


05 Data Quality

Initial data analysis reveals that the dataset is currently incomplete and not automatically updated. Articles published after August 2025 are missing. This is clearly visible in Figure 02, which shows the number of articles on the y-axis and the publication date on the x-axis. A realistic monthly volume (over 5,000 articles across SRG) is only observed in August 2025. This does not indicate a spike in publication, but rather suggests that the current dataset is corrupted or truncated. The issue must be resolved in collaboration with the data owners—specifically, the CMP team—to ensure completeness and reliability.

Other quality aspects such as missing values, outliers, and inconsistent metadata (e.g., empty title arrays or extremely long chat transcripts) are present but manageable through filtering and preprocessing. However, the lack of completeness and regular updates poses a more serious limitation and must be addressed before production use.

Figure 02: Article Count by Publication Date (per month)



06 Data Flow

The data flow in this project begins with the SRG/Swissinfo article corpus, stored in the Databricks table articles_v2. This table contains nested fields such as multilingual titles, paragraph-based content, and metadata.

From there, a series of Databricks-hosted Jupyter notebooks perform steps shown in figure 03 at the end of this chapter. The diagram illustrates how the SRG article corpus is processed in Databricks through a sequence of Jupyter notebooks for cleaning, enrichment, translation, and adaptation. The enriched corpus feeds into the editorial frontend, where journalists propose and produce stories, language heads perform quality checks, and distribution teams publish content across multiple platforms. To successfully do so, a series of Databricks-hosted Jupyter notebooks perform the following steps:

- **Web scraping notebook**: Extracts the isBasedOn field from the HTML source of published articles to identify original sources of adaptations. This metadata is not available in the SQL table and must be added manually.

- **Cleaning notebook**: Normalizes and filters the data, removes outliers, and handles missing values.

- **Flattening notebook**: Converts nested structures into a flat, analysis-ready format.

- **Enrichment notebook**: Adds derived features such as word count, language distribution, and similarity embeddings.

The output of these steps is a **cleaned and enriched article table**, which serves as the foundation for advanced analysis and downstream applications.

On top of this, additional notebooks extend the corpus with AI-based functionality:

- **Vector Indexing / Cognitive Search notebook**: Computes similarity embeddings and indexes articles, enabling efficient retrieval of semantically related texts.

- **Translation notebook**: Uses APIs such as DeepL or Google Translate to automatically translate articles into target languages.

- **Adaptation training notebook**: Learns language- and style-specific editorial patterns from past Swissinfo adaptations (e.g. headlines, tone, phrasing).

- **Adaptation application notebook**: Applies these learned patterns to new content, producing language- and audience-specific drafts.

- **Text multiplication notebook**: Multiplies articles across all languages and generates platform-specific text blocks tailored for different output channels (e.g. SRF Sport, Facebook, newsletters).

The results are stored back into an **extended corpus**, which contains both original and newly generated rows of multilingual, platform-specific content.

Finally, this enriched corpus is connected to the **editorial frontend process**:

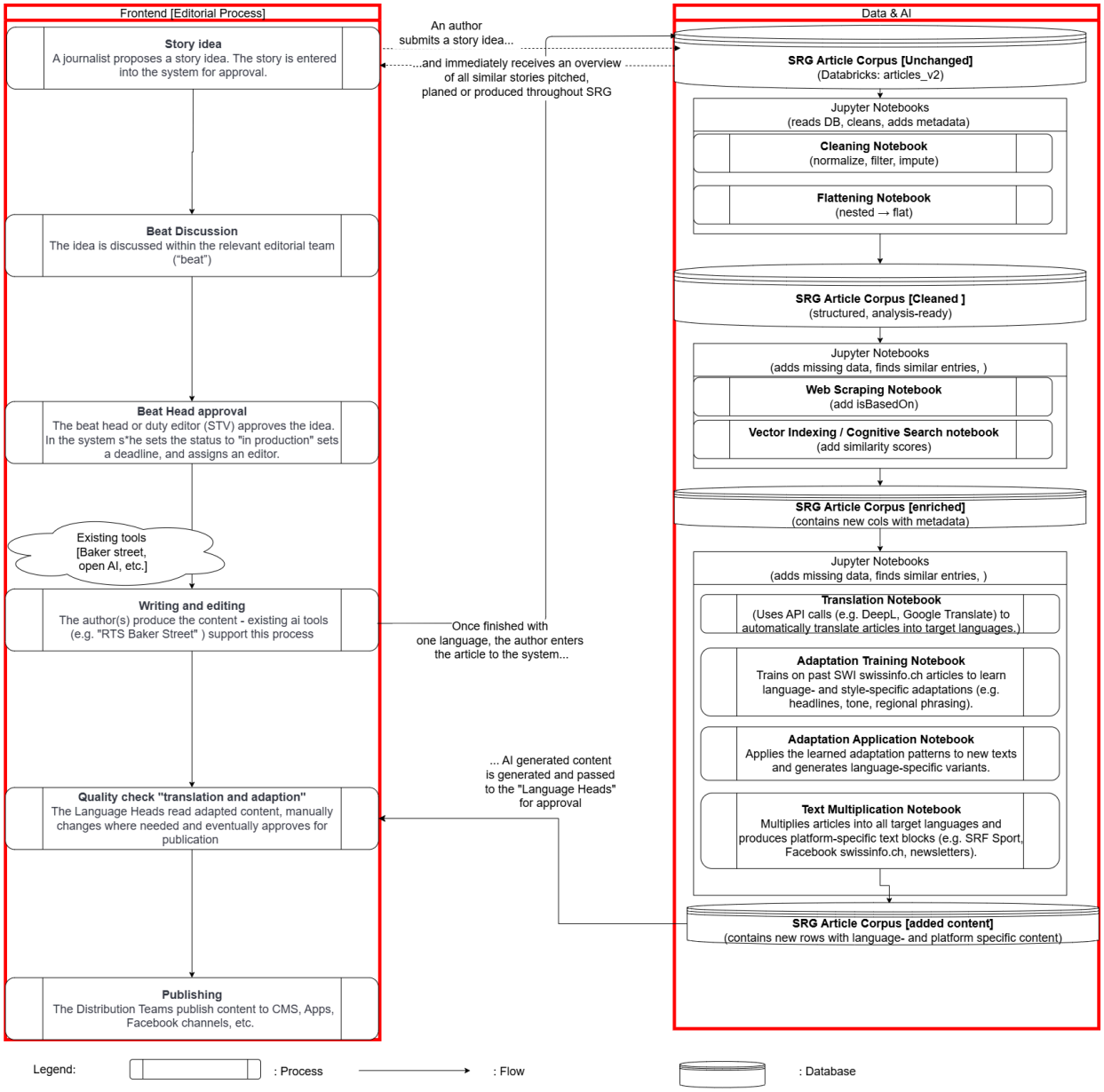
- Journalists can submit new story ideas, which are automatically compared with the SRG corpus to reveal overlaps with existing or planned content.

- Approved stories move into production, where authors draft and edit texts with the support of AI tools.

- Language heads review and approve AI-generated translations and adaptations.

- After quality checks, distribution teams publish the final content to CMS, apps, and social media channels.

Figure 03: Data Flow of the SRG/Swissinfo Editorial AI Pipeline



07 Data Model

Conceptual Level

At the core of the model is the **article** as the central object. Each article consists of multilingual titles, paragraph-based content, metadata (e.g. author, publication date, keywords), and linked resources. Articles may be adaptations of other articles, which is captured via the "isBasedOn" field (see Section 4).

Logical Level

The logical model defines which fields are used for analysis and application development. These include:

- **Identifiers**: URN, SourceId
- **Text fields**: title, content, lead/kicker
- **Metadata**: language, releaseDate, modificationDate, isBasedOn
- **Derived features**: wordCount, embedding vectors, cluster ID, adaptation flags

This layer also includes enriched fields added during preprocessing, such as similarity scores and adaptation links.

Physical Level

Physically, the data resides in **Databricks Delta tables**, structured in three layers:

- **Bronze**: Raw imported data from the CMS, nested and unprocessed.
- **Silver**: Cleaned and flattened data, with normalized fields and added metadata.
- **Gold**: Enriched data with analytical features, scraped metadata (isBasedOn), and embeddings.

This final table serves as the foundation for both use cases: story idea comparison and multilingual adaptation. It is accessed via notebooks and integrated into editorial workflows through APIs and pipelines

08 Documentation

All project steps are documented in versioned **Databricks notebooks**, which include data cleaning, enrichment, scraping, and modeling. These notebooks are stored in a public GitHub repository for transparency and reproducibility: [link]

For reproducibility, all changes to the tables are versioned in **Delta Lake**, ensuring that every transformation step can be traced and reproduced.

09 Risks

This project involves several risks—both technical and organizational—that could impact its success:

Editorial acceptance
There is a risk that journalists may perceive the tool as a threat to their professional autonomy. To mitigate this, the tool is positioned as an assistant, not a replacement. Early involvement, training, and feedback loops are essential.

Quality of journalism
Over-reliance on automated drafts could lead to formulaic or culturally insensitive content. Editorial control must remain central, supported by clear guidelines and quality checks.

Data completeness and freshness
The current corpus is incomplete and not automatically updated. Without regular updates, the tool risks being trained on outdated data. Collaboration with the CMP team is needed to establish a reliable update pipeline.

Misaligned expectations
Stakeholders may expect full automation, while the goal is editorial support. Clear communication and realistic scoping are key to managing expectations.

Resource constraints
SRG's financial pressure could affect project continuity. A modular approach ensures that even partial implementations deliver value.

Technical dependencies
Reliance on external libraries (e.g. Hugging Face, DeepL) and the Databricks environment introduces risks of vendor lock-in. Alternatives and fallback strategies should be documented.

10 Preliminary Studies

No external studies or prior experiments were conducted beyond the work documented in this report. All analyses, data preparation steps, and prototype development were carried out within the scope of this project.

For reproducibility, all notebooks are versioned and publicly available in the GitHub repository: https://github.com/Tao-Pi/CAS-Applied-Data-Science-GroupWork/tree/main/Module1 (last access 19.09.2025)

11 Conclusions

This project addresses a key challenge in multilingual journalism: how to reduce duplication and accelerate adaptation workflows without compromising editorial quality. In the context of financial pressure and limited resources, Swissinfo and SRG must find smarter ways to work.

The proposed assistant tool supports two critical editorial tasks: identifying overlapping story ideas and generating multilingual adaptation drafts. It is built on a structured and enriched article corpus, enhanced through scraping, cleaning, and feature engineering. The tool is designed to integrate into existing workflows and preserve editorial control.

However, the project is not yet ready for production. Data completeness issues, missing metadata (e.g. "isBasedOn"), and the complexity of editorial processes require further development. Collaboration with the CMP team and newsroom stakeholders will be essential.

If these challenges are addressed, the tool has the potential to deliver real value: faster publishing, reduced redundancy, and more consistent multilingual output. Most importantly, it can help journalists focus on what matters—telling stories—while technology handles the repetitive parts.

12 Acknowledgements

I would like to thank my colleagues **Jessica** and **Isabelle** for their valuable input throughout this project. Their insights into editorial workflows and their sharp analysis of challenges and opportunities were essential in shaping the direction and scope of this work.

13 Statement

„Ich erkläre hiermit, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls die Arbeit als nicht erfüllt bewertet wird und dass die Universitätsleitung bzw. der Senat zum Entzug des aufgrund dieser Arbeit verliehenen Abschlusses bzw. Titels berechtigt ist. Für die Zwecke der Begutachtung und der Überprüfung der Einhaltung der Selbstständigkeitserklärung bzw. der Reglemente betreffend Plagiate erteile ich der Universität Bern das Recht, die dazu erforderlichen Personendaten zu bearbeiten und Nutzungshandlungen vorzunehmen, insbesondere die schriftliche Arbeit zu vervielfältigen und dauerhaft in einer Datenbank zu speichern sowie diese zur Überprüfung von Arbeiten Dritter zu verwenden oder hierzu zur Verfügung zu stellen.“

Date:

19th September 2025

Signature(s):

14 Appendix

Swissinfo Production Workflow for Pitching and Producing a Feature Story (status quo and challenges)

1. Story idea

A journalist proposes a story idea.

- *Challenge / Improvement*: In a large, decentralized publishing house like SRG, the same idea may already be covered elsewhere. Today, there is no efficient way to check the vast SRG corpus, which leads to redundant work.

2. Beat discussion

The idea is discussed within the relevant editorial team ("beat").

3. Approval

The beat head or duty editor (STV) approves the idea, sets a deadline, and assigns an editor. The story is then entered into the Storyline system for production.

4. Writing

If possible, draft versions and key developments are shared with the assigned editor and the Multimedia (MM) Department.

5. Content translation and adaptation

The language heads translate the original article into swissinfo's 10 languages. Where necessary, they add background or contextual information (e.g., explaining what the "Bundesrat" is for the Russian audience).

- *Challenge / Improvement*: Current translation relies on off-the-shelf tools like DeepL plus manual editing. These tools do not use the rich knowledge stored in swissinfo's archive of multilingual adaptations. The opportunity is to:
    - Apply patterns from swissinfo's multilingual corpus to improve quality and speed.
    - Extend adaptation support to other SRG units, which today rarely adapt content from each other.

15 References and Bibliography

Blick. (2025, June 30). *Sparen bei Sport, Filmen und Verwaltung: Die SRG gibt massive Kürzungen bekannt*. Online available following URL: https://www.blick.ch/politik/sparen-bei-sport-filmen-und-verwaltung-die-srg-gibt-massive-kuerzungen-bekannt-id21007755.html (last access 19th September 2025)

SRF. (2025, June 30). *SRG stellt sich neu auf und rückt näher zusammen*. Online available following URL: https://www.srf.ch/news/schweiz/in-eigener-sache-srg-stellt-sich-neu-auf-und-rueckt-naeher-zusammen (last access 19th September 2025).