Evaluation of the Text Reuse at Scale Explorer

About the new component

Thank you for agreeing to evaluate impresso Text Reuse at Scale alpha release for the exploration of text reuse data in semantically enriched newspaper data. We very much anticipate your feedback in the hope that it will lead to the improvement of future versions.

This new component was inspired by the outcomes of a workshop with historians, designers, computational linguists and digital humanists which discussed opportunities inherent in the integration of text reuse and other semantic enrichments such as the automated detection of named entities (persons, locations), topics, article types, and language.

We have pepared accompanying information about text reuse and its application on our corpus as part of the impresso FAQs which you can find here: https://impresso-project.netlify.app/faq

The new application offers three types of filters:

1) Metadata assigned during digitisation (e.g. newspaper titles and publication dates),
2) Semantic enrichments generated by the *impresso* project, and finally, and
3) Filters based on Text reuse specific data includes lexical overlap, cluster size and the time span between publication dates of passages within a text reuse cluster.

Interactive visualisations allow the exploration and comparison of text reuse data across time and titles in conjunction with these filters.

About the evaluation

At some you will need to be logged in to complete a task. We therefore ask you to create an impresso account
to access the full corpus. To do this, first complete the registration form and **then email us** this (digitally)
signed NDA form to info@impresso-project.ch.

The evaluation includes 5 tasks which guided the development of the application as well as a final assessment of the application.

Note: Some important tasks concerning e.g. virality detection, text passage-based search and data export are not yet supported by the application and are therefore not part of this evaluation.

For each task, we ask you to:

1. Carefully read the abstract task description.
2. Comment on its definition and scope below (optional).
3. Study how tasks are supported by the application using the descriptions, accompanying screenshots and links to example queries we provide.
4. Assign an overall score from 1 (easy) to 5 (hard).
5. Comment on your experience of completing the task together with suggestions for improvements in the comment box below.

Finally, we will ask you to score the individual interface elements together with additional comments on your overall user experience.

* Indicates required question

1. Email *

2. Please enter your first and last name *

Task 1: Obtain an overview of text reuse in a corpus, collection or query.

This task captures the need to gain orientation regarding the presence of text reuse data within a given set of documents and over time. Users need to determine whether or not a given corpus, corpus subset, query result or collection contains instances of text reuse.

Such occurrences of text reuse should be understood regarding their distribution over time and a first assessment of their characteristics should be made. This can be facilitated by overviews of the distribution of metadata and enrichments such as time, newspapers, countries, content types, languages and named entities but also text reuse-specific measures such as lexical overlap, time span between publication dates, cluster size and number of passages.

The ability to inspect outliers and to determine averages offers additional valuable insights about the distribution of text reuse data at different levels of granularity. This includes, for example, the inspection of largest/smallest clusters, clusters with the highest/lowest lexical overlap, the ability to filter for earliest/latest cluster in the corpus or the longest time span between publication dates and constellations of any of these measures.

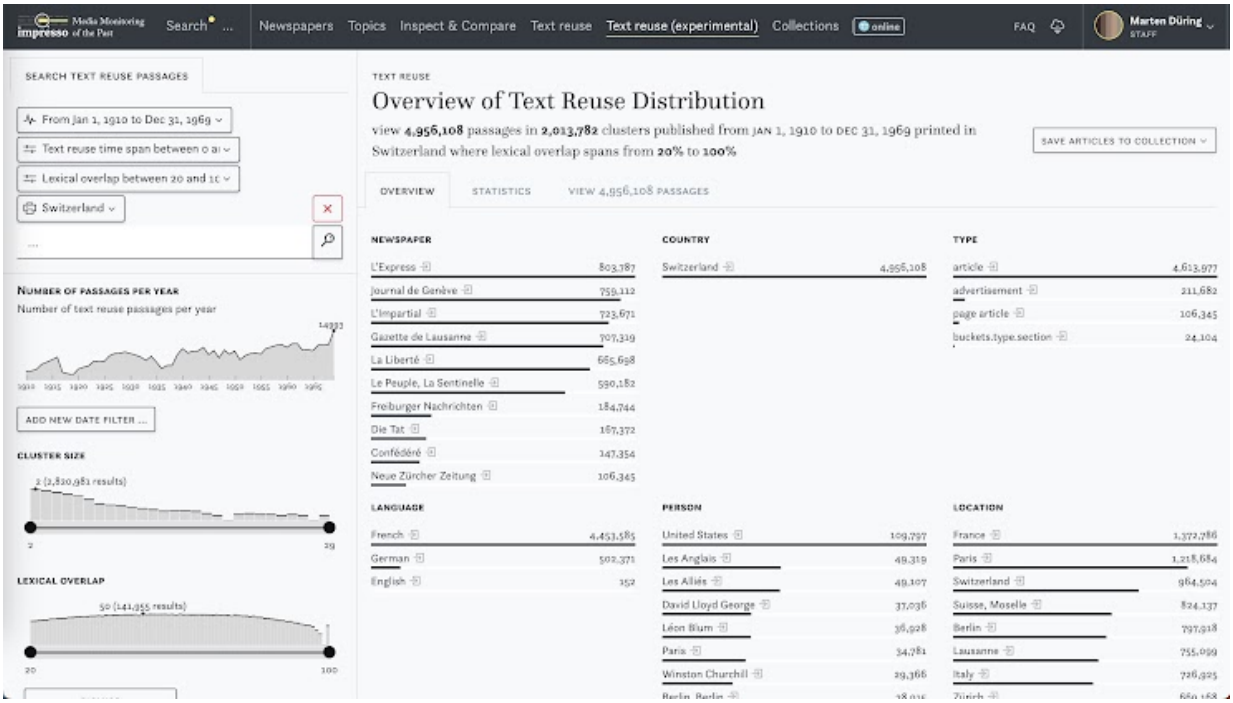
3. **What are your thoughts on Task 1?**

4. **How is Task 1a supported by the application?**

This first task will also help you familiarise yourself with the different components of the interface. Begin by opening the link to the first example query. Take a moment to orientate yourself within the interface. Note the small "i-buttons" which offer additional information.

Get an overview of the presence of text reuse in this query and use the following questions for orientation:

1. Which filter settings are currently applied?
2. What does the Overview tab reveal about the distribution of text reuse across article types, languages, and countries?



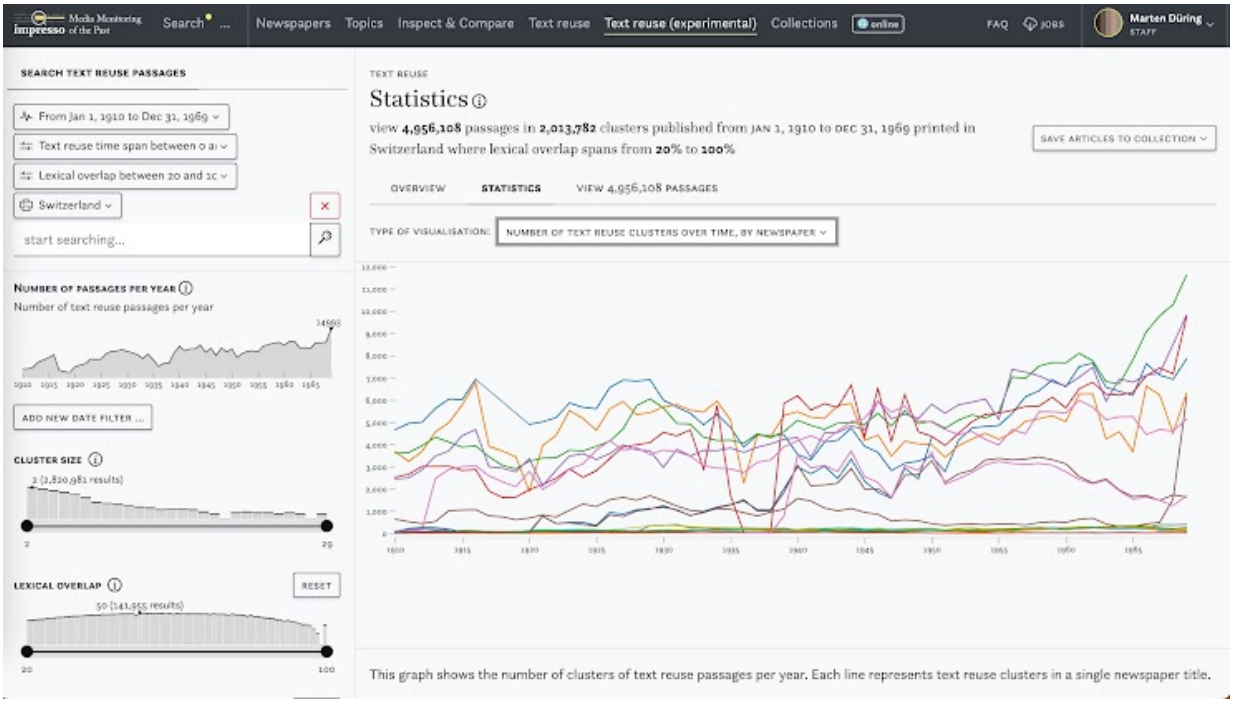
Mark only one oval.

- ◯ This is hard
- ◯ This is somewhat hard
- ◯ This is neither hard nor easy
- ◯ This is somewhat easy
- ◯ This is easy

5. **How is Task 1b supported by the application?**

Stay with the link to the first example query. Continue to get an overview of the presence of text reuse in this query:

1. Use the Statistics tab to explore the distribution of clusters and lexical overlap over time.
2. Use a different view in the Statistics tab to identify the size of the largest cluster in this query.
3. Move to the Passages tab. Sort the passages to find the cluster with the largest time span. How many days does it cover?
4. Change the country filter from Switzerland to Luxembourg. Which differences do you observe in cluster size over time?



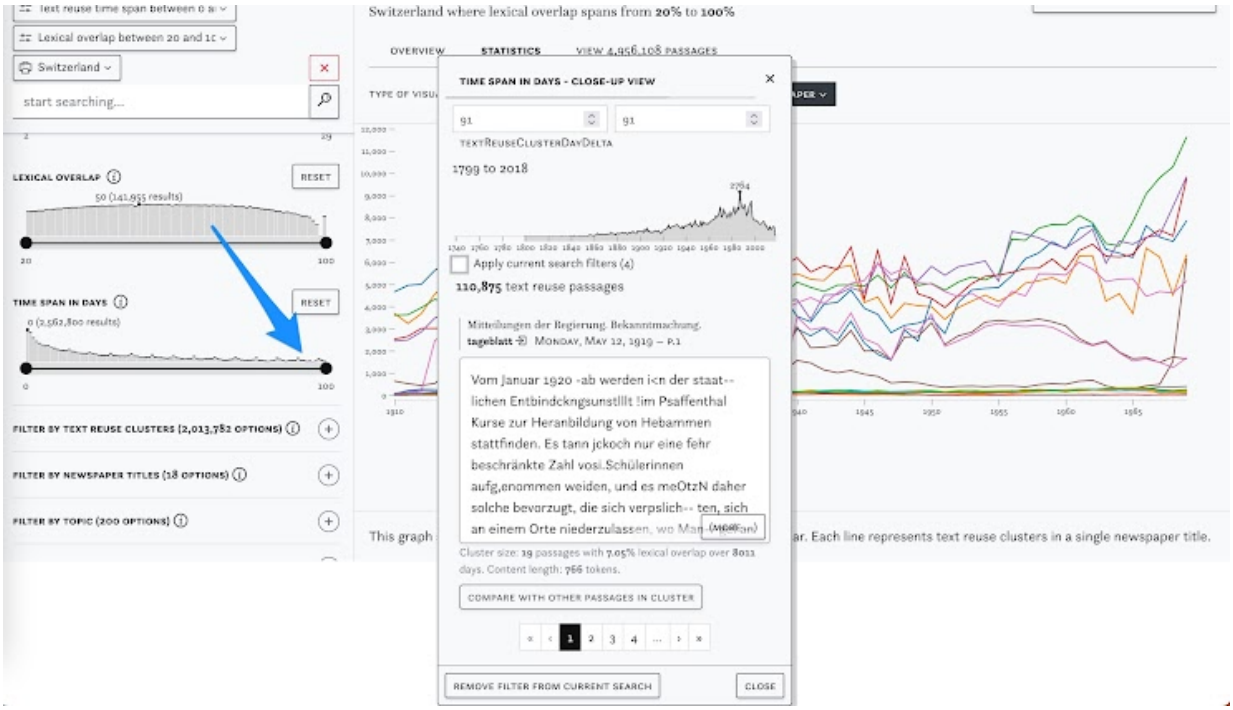
Mark only one oval.

- ◯ This is hard
- ◯ This is somewhat hard
- ◯ This is neither hard nor easy
- ◯ This is somewhat easy
- ◯ This is easy

6. **How is Task 1c supported by the application?**

Stay with the link to the first example query. Continue to get an overview of the presence of text reuse in this query:

1. Click on a slot in one of the histograms, e.g. lexical overlap, cluster size, or time span. The Close-up view lets you select ranges of interest, for example to take a closer look at the clusters with the largest time span as pictured below. Previews of the passages appear below.



Mark only one oval.

- This is hard
- This is somewhat hard
- This is neither hard nor easy
- This is somewhat easy
- This is easy

7. **What did you find easy / hard during the completion of Tasks 1a, b, and c? What should be improved?** *

Task 2: Obtain an overview of a single cluster.

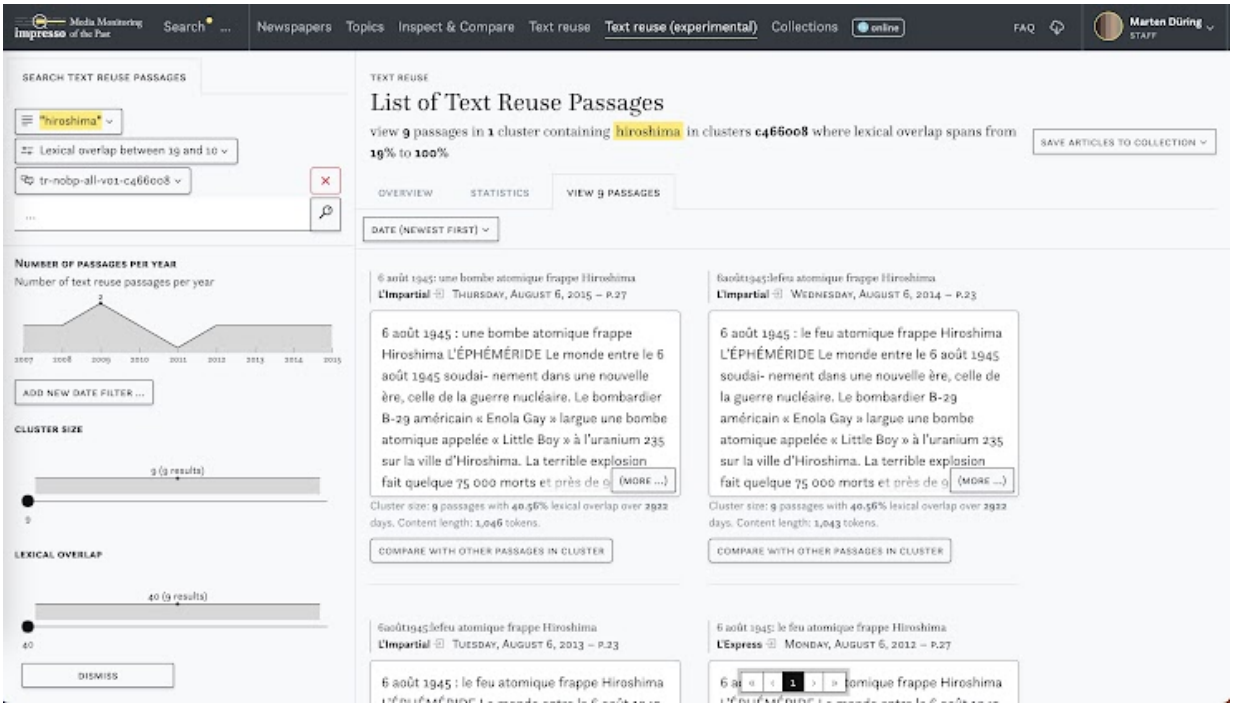
This task is similar to Task 1 but focuses on the main properties of a single cluster. This includes the number of passages, their content, the lexical overlap between them, the time span between their publication dates as well as the distribution of semantic enrichments and metadata.

8. **What are your thoughts on Task 2?**

9. **How is Task 2 supported by the application?**

[Open this example query (French)] or this [example query (German)] which filters for a single cluster.

1. Use the filters on the left as well as the tabs in the centre to get a sense of the different types of content of the underlying articles.
2. Find out about the cluster's lexical overlap, the number of passages it contains and time span between publication dates.
3. Click the red [X] to remove all current filters from Search. Now select other clusters using the left filter pane. What can you learn about their characteristics (e.g. lexical overlap, time span, or size)? Also try the Overview and Statistics tabs. Is there anything else you would like to know about a cluster?



Mark only one oval.

- ◯ This is hard
- ◯ This is somewhat hard
- ◯ This is neither hard nor easy
- ◯ This is somewhat easy
- ◯ This is easy

10. **What did you find easy / hard during the completion of Task 2? What should be improved?** *

Task 3: Compare differences between passages within a cluster.

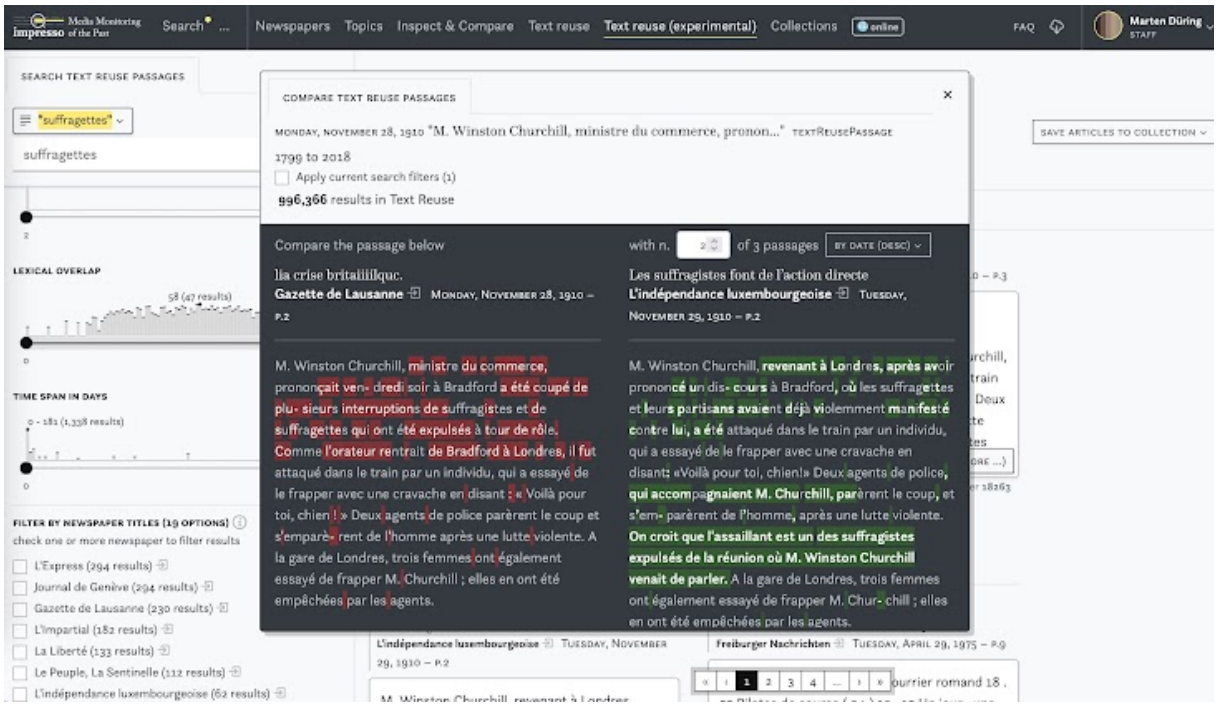
This task concerns the comparison of two or more passages to reveal differences and similarities. A common motivation for such a comparison are editorial edits of texts under circulation, such as press agency outputs. Comparisons can e.g. reveal adaptations to suit the political preferences of a newspaper's audience, clarifications - what is obvious to one set of readers may need additional explanation to others, but also unintended differences such as degeneration, for example caused by OCR errors.

11. **What are your thoughts on Task 3?**

12. **How is Task 3 supported by the application?**

This example query (in French, use the language filter for German-language passages) will lead you to passages which contain the keyword "suffragettes". The screenshot below highlights variations between articles which cover the same event. Characters present only in the selected passage on the left are highlighted in red, characters only present in the compared passage are highlighted in green.

1. Select a cluster of your own interest and explore other instances of how passages differ.



Mark only one oval.

- This is hard
- This is somewhat hard
- This is neither hard nor easy
- This is somewhat easy
- This is easy

13. **What did you find easy / hard during the completion of Task 3? What should be improved?** *

Task 4: Identify different types of text reuse.

Text reuse encompasses different forms of reiterated text, including e.g. co-publication, template-based content such as adverts or TV programmes, and press agency reports. But we can also distinguish text reuse based on temporal patterns, e.g. based on the duration (start - end date), virality, and rhythm of passage publication dates. Each of these phenomena maps to text reuse data characteristics, and semantic enrichments offer a highly versatile approach to segment text reuse data into meaningful categories.

We expect recurrent patterns for different types of text reuse. For instance, cinema adverts should be characterised by large number of named entities (persons), topics associated with media content, and a high lexical overlap, a large cluster size and a smaller time span.

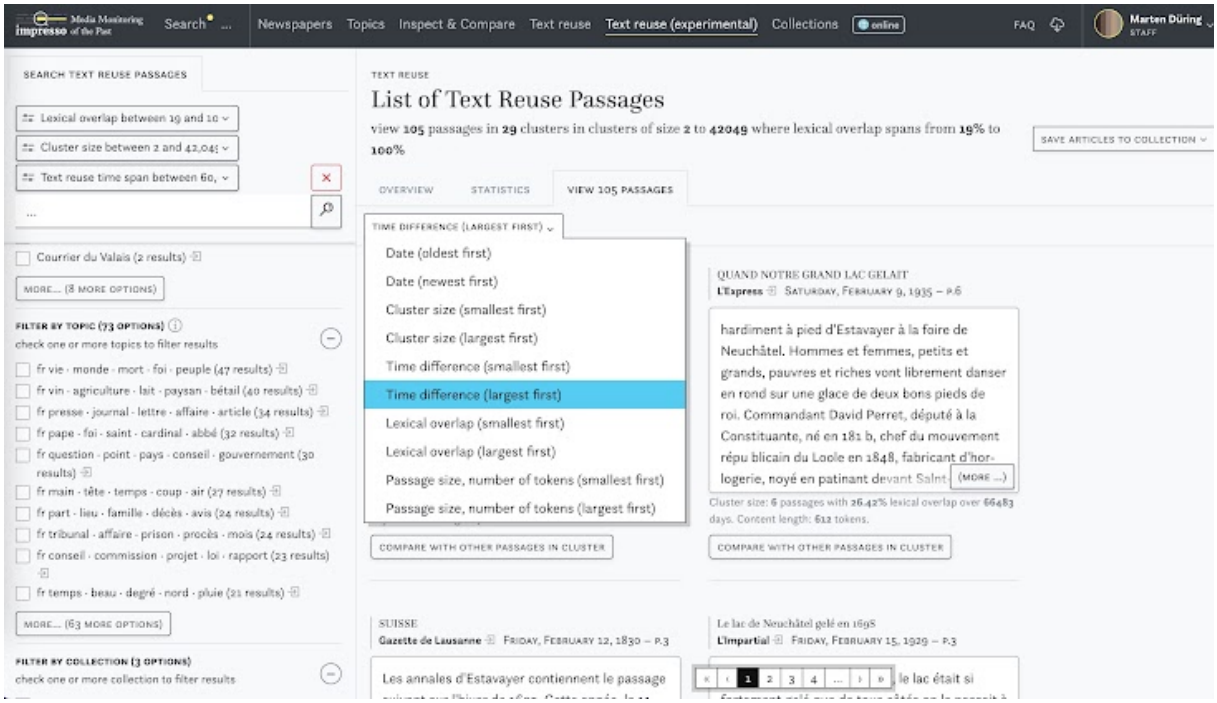
14. **What are your thoughts on Task 4?**

15. **How is Task 4 supported by the application?**

This example query filters for instances of slow text reuse with a time span of at least 60.000 days between the earliest and latest publication dates of passages.

1. Change the filter settings to e.g. 0-365 days and observe their distribution of clusters over time using the Statistics tab. Here, observe which newspaper title has the highest number of passages and how this evolves over time.
2. Continue to compile your own queries to identify different types of text reuse, for example based on article type, topic or keyword mentioned in a passage.

**) Note that this filter will obscure instances of rapid text reuse with clusters which contain many passages published in quick succession (e.g. in the first 365 days) together with one or more passages published later. Future work includes adding a dynamic timeline to display passage publication dates on a day level.*



Mark only one oval.

- ◯ This is hard
- ◯ This is somewhat hard
- ◯ This is neither hard nor easy
- ◯ This is somewhat easy
- ◯ This is easy

16. **What did you find easy / hard during the completion of Task 4? What should be improved?** *

Task 5: Generate research corpora based on text reuse clusters.

This task corresponds to the need for fine-grained content selection and the need to create meaningful subsets of text reuse clusters and associated passages based on all measures, enrichments and metadata.

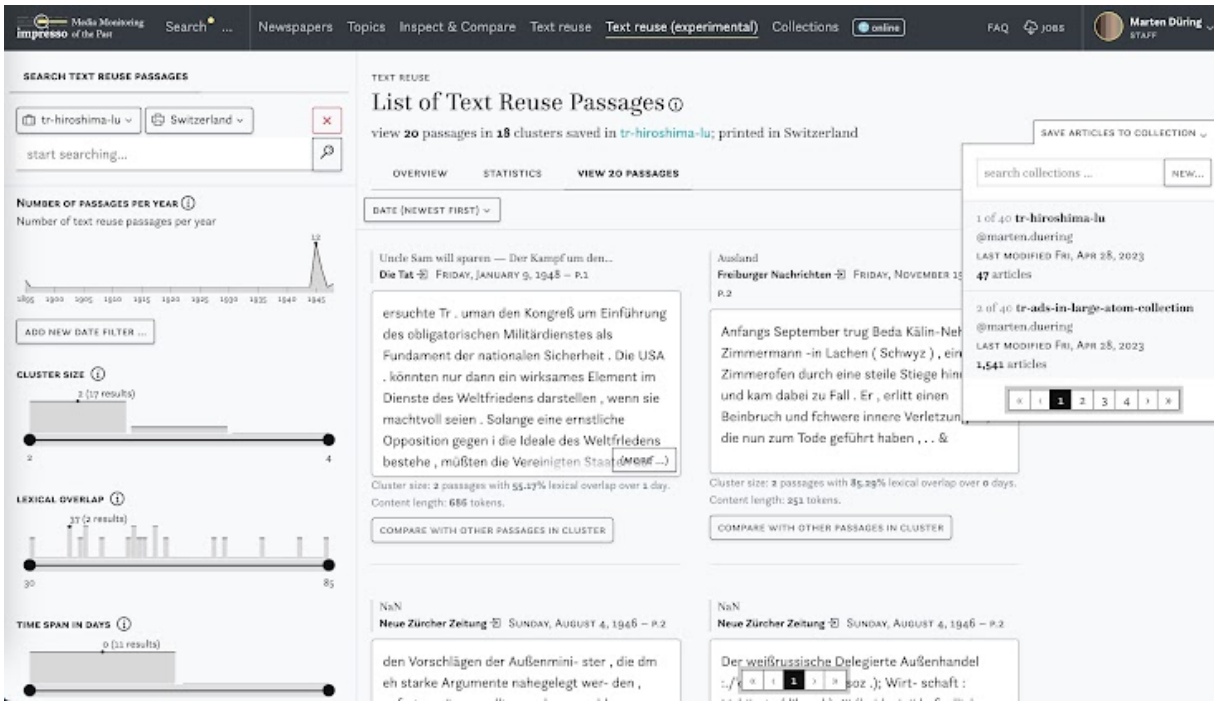
17. **What are your thoughts on Task 5?**

18. **How is Task 5 supported by the application?**

To complete this task, make sure that you are logged in with your impresso account. Within *impresso*, corpus subsets are called "Collections" and can be treated as research corpora in their own right. They can be searched, exported or compared to each other.

1. Generate a query for a meaningful subset of text reuse passages (note the current limit of 10.000 articles per collection). Save the articles to a new collection (see screenshot, right side)
2. Select the collection in the filter pane and filter for it. You can now focus your research on this subset of articles which contain text reuse passages and choose to export metadata and full text to csv (depending on copyright).

Note: A future release will allow comparative views using the Inspect & Compare explorer.



Mark only one oval.

- ◯ This is hard
- ◯ This is somewhat hard
- ◯ This is neither hard nor easy
- ◯ This is somewhat easy
- ◯ This is easy

19. **What did you find easy / hard during the completion of Task 5? What should be improved?** *

Wrapping up

In this final section we ask you to reflect on your overall experience with the application.

20. Please rate how you feel about the following statements. *

Mark only one oval per row.

	Strongly disagree	Somewhat disagree	Neither disagree nor agree	Somewhat agree	Fully agree
The application effectively supports the tasks presented in this evaluation.	◯	◯	◯	◯	◯
The usage examples and accompanying information give a clear understanding of the functionalities of the application.	◯	◯	◯	◯	◯
It is easy to navigate the application.	◯	◯	◯	◯	◯
Application loading times are acceptably fast/responsive	◯	◯	◯	◯	◯

21. The following element in the application facilitates effective exploration of text reuse data. *

Mark only one oval per row.

	Strongly disagree	Somewhat disagree	Neither disagree nor agree	Somewhat agree	Fully agree
Search					
Filter pane (left side)					
Overview tab					
Statistics tab and its views					
Passages tab					
Save articles to collection					
Close-up view					

22. Is there anything in the application that doesn't make sense? Does anything feel out of place? *

23. Future development of the application should focus on these tasks / features / overall improvements: *