Unveiling Strategic Research Priorities: A Terminological Analysis of the ARCHE SRIA Using Word Rain

Elisa Squadrito[1,*], Francesca Frontini[2], Vania Virgili[3] and Monica Monachini[2]

[1]University of Macerata, Italy
[2]Institute of Computational Linguistics "A. Zampolli", National Research Council (CNR-ILC), Italy
[3]Institute of Heritage Science, National Research Council (CNR-ISPC), Italy

Abstract

This paper presents an experiment conducted through collaboration between the CLARIN and E-RIHS Research Infrastructures to analyse the Strategic Research and Innovation Agenda (SRIA) of the Alliance for Cultural Heritage Research in Europe (ARCHE). Using the Word Rain tool, semantically structured word clouds were generated to uncover key domain conceptualizations within the SRIA preparatory documents. The experiment aims to reveal the terminology that shapes the research coverage of the agenda and the conceptual framework that guides future initiatives in the field. Through this approach, we highlight the utility of corpus-based analysis in enhancing strategic policy development.

Keywords

Cultural Heritage, ARCHE, Distant Reading, Terminology, Word Cloud

1. Introduction

The need for accurate and precise terminology in institutional and technical documents cannot be overstated. In these contexts, the choice of terms contributes to conceptualizing domain knowledge, ensuring effective communication, and minimizing the margin of error. Although this is true, in policy documents such as Strategic Research and Innovation Agendas (SRIAs), terminology not only ensures domain accuracy but also helps define research priorities, influence policy development, and guide strategic decision-making. According to the ERA-LEARN portal[1], a SRIA can be defined as a strategy document produced by a partnership, for which it "identifies its objectives, impact areas and expected outcomes, portfolio of activities, outputs, and milestones within a certain timeline." Resulting from a complex co-creation process, such agendas strive to unite stakeholders from the most diverse backgrounds in a domain and collaboratively outline research trajectories to be pursued in future calls for proposals. This is the case of the ARCHE Strategic Research and Innovation Agenda, the SRIA which is currently being prepared for the domain of Cultural Heritage in Europe. In a multidisciplinary environment such as that of Heritage Science, the strategic use of shared terminology that is commonly accepted and representative becomes not only useful but imperative. From this necessity stems the present contribution and the idea of putting terminology research to use in Cultural Heritage, through a joint collaboration between the co-creators of the agenda from E-RIHS[1] and CLARIN[2], respectively, the research infrastructures for Heritage Science and language as social and cultural data.

3rd International Conference on "Multilingual digital terminology today. Design, representation formats and management systems" (MDTT) 2025, June 19-20, 2025, Thessaloniki, Greece.

*Corresponding author.
e.squadrito@unimc.it (E. Squadrito); francesca.frontini@ilc.cnr.it (F. Frontini); vania.virgili@cnr.it (V. Virgili); monica.monachini@ilc.cnr.it (M. Monachini)
0009-0008-1975-9240 (E. Squadrito); 0000-0002-8126-6294 (F. Frontini); 0000-0002-7948-0136 (V. Virgili); 0000-0003-3356-3988 (M. Monachini)

[1]Further information at: ERA-LEARN portal, What is a SRIA?, https://www.era-learn.eu/support-for-partnerships/cross-cutting-issues-and-additional-activities/strategy-and-foresight/what-is-a-sria

2. The ARCHE Project and SRIA Development

The socioeconomic, technological, and environmental challenges that our society has faced in recent years have made it necessary to redefine classical paradigms for Cultural Heritage management, protection, and restoration. To this end, the Alliance for Research on Cultural Heritage in Europe[2] (ARCHE) was established. Funded by Horizon Europe, the project revolves around the creation of a holistic network of stakeholders in the Cultural Heritage domain, including researchers, heritage professionals, organizations, and institutional bodies. Operating at the academic, governmental, and public levels, ARCHE aims to address current gaps in Heritage Science research and to develop multidisciplinary and sustainable approaches to innovation in the field. The primary tool for achieving these objectives has been identified as the drafting of a SRIA. Building on the JPI CH 2020 SRIA[3], the forthcoming agenda will ensure the implementation of the Alliance's vision into tangible and actionable goals, thus serving as a driving force of innovation in the domain. More specifically, it will act as a "roadmap with research priorities that will form the basis of calls for projects and other activities due to start in 2026, within the European Partnership for Resilient Cultural Heritage (RCH)"[4]. Started in 2024, the development of the ARCHE SRIA was rooted in a mapping and assessment phase aimed at providing a comprehensive initial overview of innovative research areas in the Cultural Heritage domain. Building on this foundation, all stakeholders involved in the ARCHE project played an active role in drafting the agenda through multiple consultations in the form of workshops and surveys. The direct involvement of experts was crucial to capturing the perspectives of the many communities of practice targeted by the policy document, as well as ensuring its accuracy and precision. It was nevertheless believed that adopting a distant reading approach[3] to explore the specialized documents resulting from these consultations could provide valuable insight into both linguistic and extralinguistic issues. Examining the terminology of documents produced by ARCHE stakeholders from a broader perspective proved useful for identifying recurrent patterns, prevalent concepts, and how they clustered.

This paper presents a distant reading experiment conducted on a corpus of documents drafted within the ARCHE project, including the *Key Messages and Preliminary Findings* that will form the basis of the future SRIA. The tool selected for the analysis was Word Rain[5], an advanced data visualization tool capable of generating semantically structured word clouds. Given the terminologically rich nature of such specialized texts, the rationale for the experiment was that examining how terms cluster and appear in the word cloud could help uncover recurrent themes and reveal differences in their distribution between early and later project documents.

3. Methodology

In this section, we will briefly give an overview of the tool chosen for conducting the analysis, that is, Word Rain, as well as a brief description of its key functionalities. A few lines will be devoted to the comparison of the tool against the backdrop of classic word cloud visualisation tools. Finally, the reasons for choosing to use Word Rain in the context of this study and the opportunities it offers will be presented, without forgetting to highlight the limitations in its application.

3.1. Introducing Word Rain

Word Rain[4] is an advanced data visualisation tool that generates a semantically structured word cloud, referred to as "word rain". It enables users to visualise word distributions within a text, adjusting the size and placement of terms based on their frequency and semantic relevance. Word Rain was developed through a collaboration between the Centre for Digital Humanities and Social Sciences at Uppsala (CDHU), the National Language Bank of Sweden/CLARIN Knowledge Centre for the Languages of

[2] https://www.heritageresearch-hub.eu/arche-home/about-arche/
[3] https://www.heritageresearch-hub.eu/strategic-research-and-innovation-agenda-2020-sria/
[4] For further information, see https://www.heritageresearch-hub.eu/event/arche-2nd-stakeholders-workshop-to-take-place-in-florence-on-septem
[5] Accessible at: https://wordrain.isof.se/

Sweden (SWELEN) and the iVis group at Linköping University. The tool is available as a web-based application and has also an open source version of GitHub[6] for those interested in customising or integrating it into other projects.

3.2. Word Rain compared to Classic Word Clouds

Word clouds are popular tools for visualising text content in an immediate and intuitive manner, due to their compact and static layout. However, traditional word clouds typically feature tightly packed words without a semantically motivated positioning, and prominence is solely indicated by font size, based on the word frequency within the text. Word Rain draws from classic word clouds but features a few salient innovations. As highlighted by its creators, it enhances the traditional word cloud model by incorporating a distributional semantics-based approach, reduced to one dimension, to position words along a semantically meaningful x-axis[5]. Colour-coded bars further enhance the visual grouping of related terms. While font size continues to indicate prominence, additional indicators, such as bar height and vertical positioning along the y-axis, allow for a more nuanced interpretation of the data. Less prominent words are positioned lower on the y-axis, creating a sense of "falling," hence the term "word rain."

3.3. Word Rain within the ARCHE project: opportunities and limitations

Because of their ability to give an intuitive and immediate snapshot of a text or a collection of texts, word cloud visualisations outlets are frequently incorporated into main corpus manager software, such as Sketch Engine and Voyant Tools. However, such simplicity and intuitiveness comes at the expense of depth and granularity in the type of information provided. A similar situation can be observed when analysing the purpose of wordlists. Such instruments are frequently used by corpus linguists and terminologists to evaluate the fitness and obtain an initial picture of a corpus. Nevertheless, they are exhaustive enough for a comprehensive terminological analysis, The same limitations can be stated for Word Rain. However, because of its distributional semantics-based approach, the tool adds a layer of complexity to classic word clouds. One of the main reasons for which it was selected, other than keywords and n-grams extraction, was indeed its ability to visually group related terms into easily readable semantic clusters. These clusters can both give a quick view on narratives in the corpus, and be used as indicators of the main topics covered in the document, thus providing guidance on how to later proceed with the terminological analysis. As an example, the terminology extracted can be later analysed, subdivided, and compared in accordance with the thematic areas highlighted from the Word Rain visualisations. In doing so, imbalances in representation might be detected and motivated.

Word Rain, despite its enhanced and improved features, is still a word cloud. However, alongside its semantically motivated visualisations, its ease in use by non-experts too is the second main reason for which it was selected for this analysis. The main goal of this study was to allow experts in the Cultural Heritage domain involved with the preparation of the ARCHE SRIA to learn how to make sense of their specialised documents, without requiring any prior background in linguistics. Although the Word Rain analysis was performed by CLARIN experts and then validated by ARCHE experts, its intuitive character allowed domain experts to follow its reasoning and, hopefully, replicate it in the future. In short, findings from the Word Rain can 1) provide an initial picture of the documents at hand by unveiling possible hidden narratives 2) highlight keyword clusters that might guide further term extractions and analysis of the corpus and 3) help inform experts involved in the ARCHE SRIA writing for the preparation of the agenda.

[6]Accessible at : https://github.com/CDHUppsala/word-rain

3.4. Word Rain Key Functionalities

Word Rain allows its visualisations to be easily customised, according to one's goals and needs. Its main key functionalities[7] can be listed as such:

1. Term Frequency-Inverse Document Frequency (TF-IDF): Word Rain allows users to visualise the distribution of words in a text by adjusting word size and placement based on frequency and semantic relevance, using techniques such as TF-IDF.
2. N-gram extraction: This parameter allows users to extract and visualise common n-grams, either in addition to or instead of individual words. As specified by developers of the tool, this feature works best for languages where compounds are constructed as: more common word followed by specifying word.
3. Background corpus selection: Users can upload a background corpus, refining TF-IDF values and creating more nuanced relevance for terms in the main document. The tool is handy for exploring and comparing corpora, as it can visualise language patterns and shifts within datasets, such as climate change reports or other specialised texts.
4. Customised visualisation: Users can modify settings, including font size and word vertical position, to optimise readability. It is also possible to regulate the maximum font size and decide how to arrange words when they overlap on the vertical axis. These settings allow to control the visualisation's airiness or density, depending on the desired appearance.

4. Analysis and Results

Word Rain was chosen to visually analyse, extract meaningful information from and compare the following documents:

1. ARCHE D2.1 Future Trends on Cultural Heritage (Foresight Analysis);
2. ARCHE D2.4 Vision and Mission;
3. ARCHE D2.5 SRIA Key Messages and Preliminary Findings;
4. WGs forms on ARCHE SRIA.

The main objectives of the experiment were to: a) Identify similarities and differences between the four documents' discourse; b) Determine if any themes originally identified in the WGs forms are missing or appear less emphasised in the SRIA Key Messages and preliminary findings; c) Examine how keywords clustered. To get an initial general picture of the recurring themes in the reports, it was decided to generate four word rains, one per document. By generating four visualisations, it will be possible not only to have a look at each document individually and understand its structure but also to compare every one of them against each other, retrieve shifts in argumentation and analyse themes that were not covered homogeneously.

4.1. Preprocessing of data: conversion and cleaning of the reports

In order to generate the four Word Rains, source reports in the .pdf format were converted to the .txt format and cleaned of any unnecessary elements that could introduce noise into the word rain visualisation. The noise was removed in two phases:

1. Documents were first cleaned in a standardised manner, following general rules commonly used for cleaning data in textual corpora before processing.
2. A second cleanup was carried out after having generated four mock-up word rains, one per document. Words of no interest that frequently appeared in the clouds were removed from the documents, such as: https, www, .com, homepage, org, pdf, core theme, title, keyword, ch, hub.

[7]For any in depth discussion on the tool and its key functionalities, please refer to Skeppstedt, Maria, et al. "From word clouds to Word Rain: Revisiting the classic word cloud to visualize climate change texts." Information Visualization (2024): 14738716241236188.

For instance, “core theme" frequently appeared in the word rains but held no meaning in this type of analysis. For this experiment, the Word Rain web application was used. However, for more fine-grained analysis, it would be possible to generate the clouds in a coding environment, allowing the use of one or more stop-word lists. This would eliminate the need for modifications to the data.

4.2. Processing of data: generating word rains

Once the documents were cleaned, word rains were generated using the web application tool. The parameters selected for plotting the clouds were the following:

1. Language selected: English
2. Word count: 300
3. Frequency: TF-IDF
4. Word combinations: extraction of n-grams
5. Background corpus: none
6. Word size fall-off: 0.7
7. Bar height: 40

The rationale for choosing 300 word count instead of the 600 option was to create an airy and easily interpretable word cloud. Similarly, a 0.7 fall and a 40% bar height will prevent the vertical y-axis from becoming clogged and difficult to read. However, no background corpus was selected at this stage. This choice was motivated by the goal of the analysis, which was to compare only term occurrences contained in the four ARCHE reports. Figures 1 to 4 display the Word Rain visualizations corresponding to the four ARCHE documents processed.

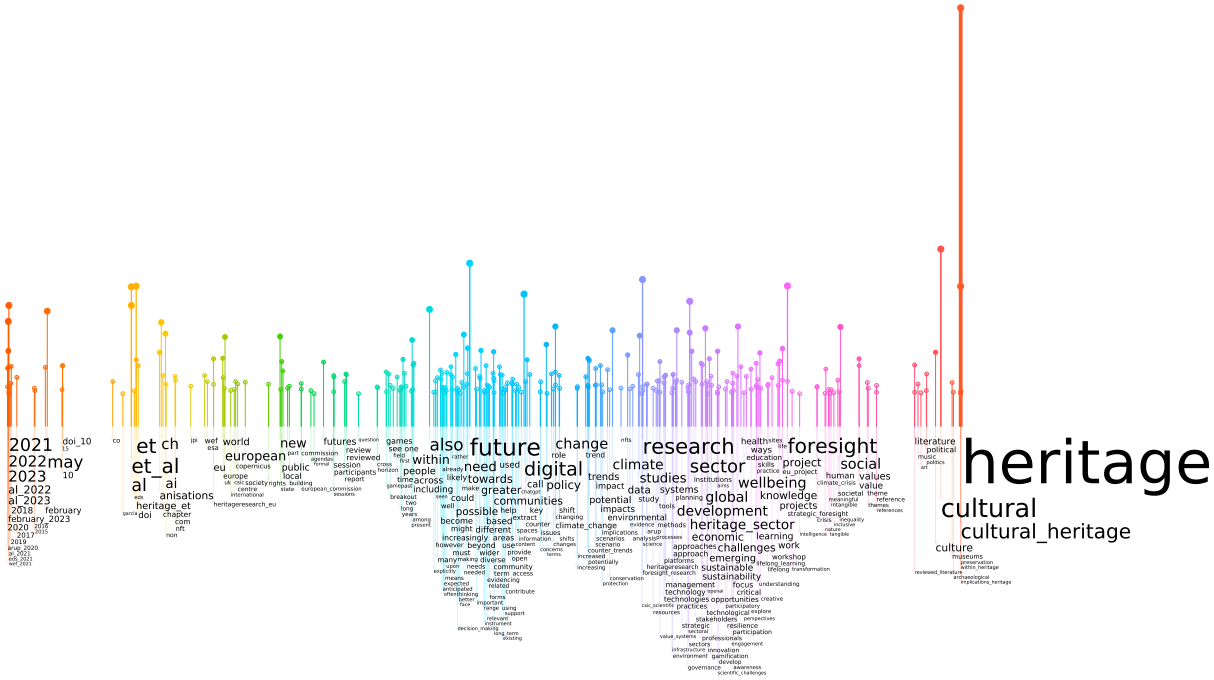

Figure 1: Word Rain of ARCHE D2.1 Future Trends on Cultural Heritage (Foresight Analysis)

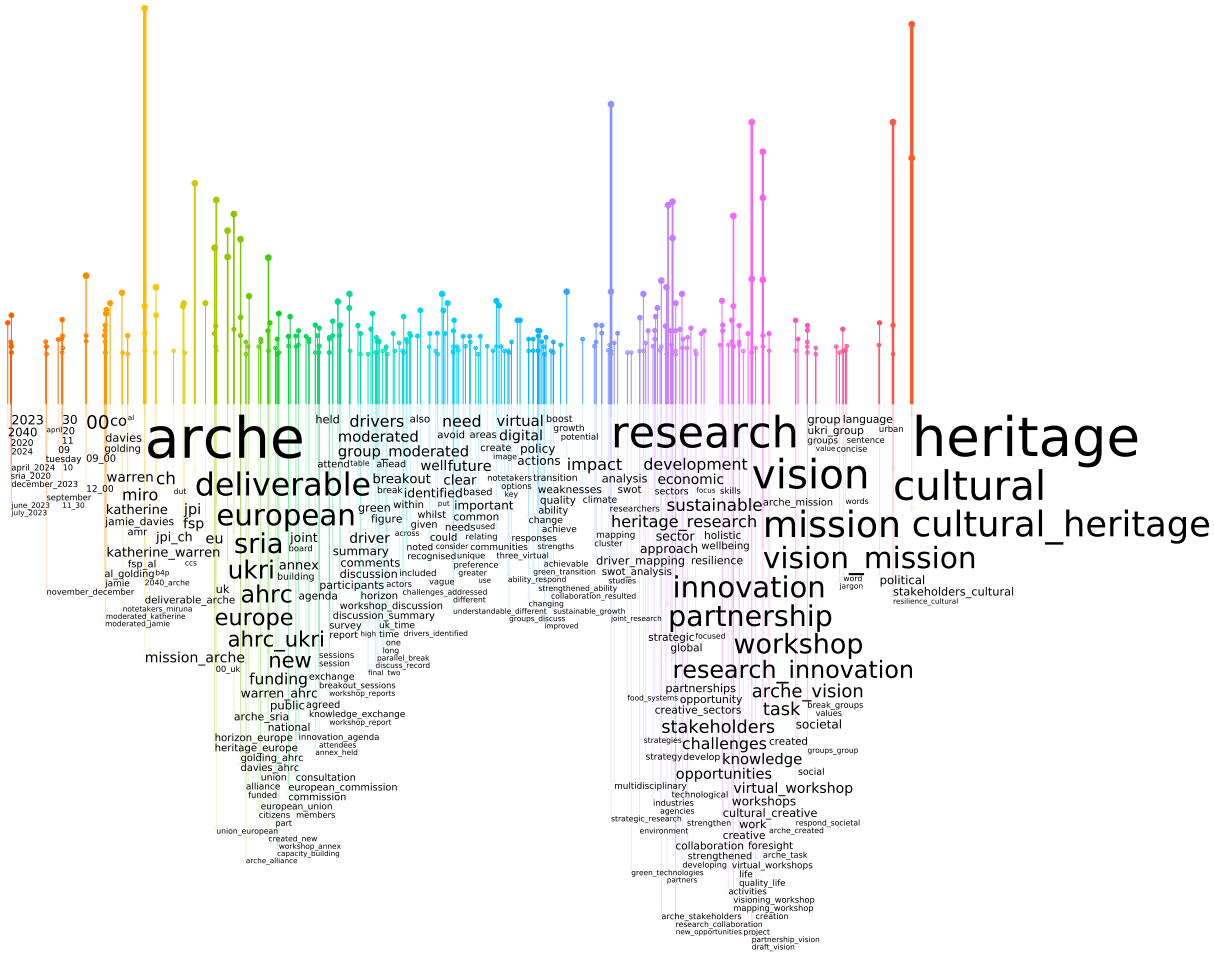

Figure 2: Word Rain of ARCHE D2.4 Vision and Mission

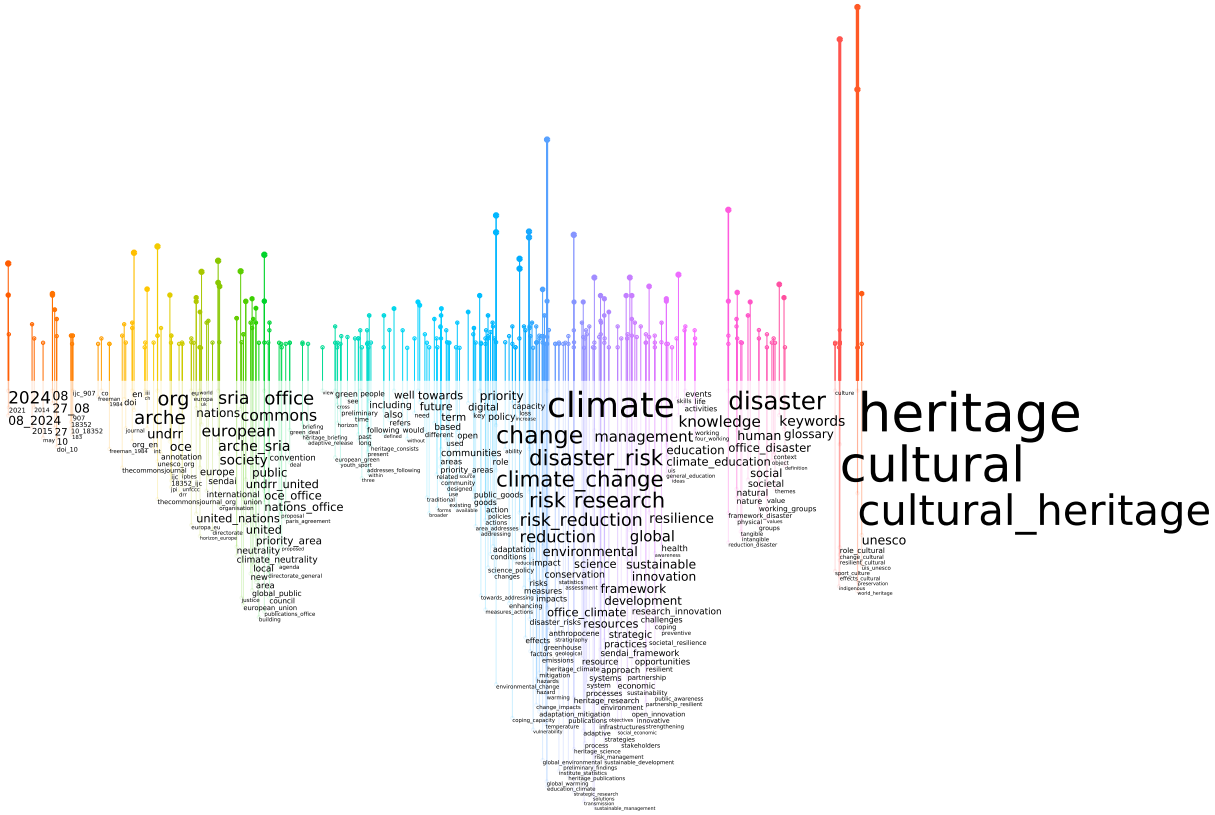

Figure 3: Word Rain of ARCHE D2.5 SRIA Key Messages and Preliminary Findings

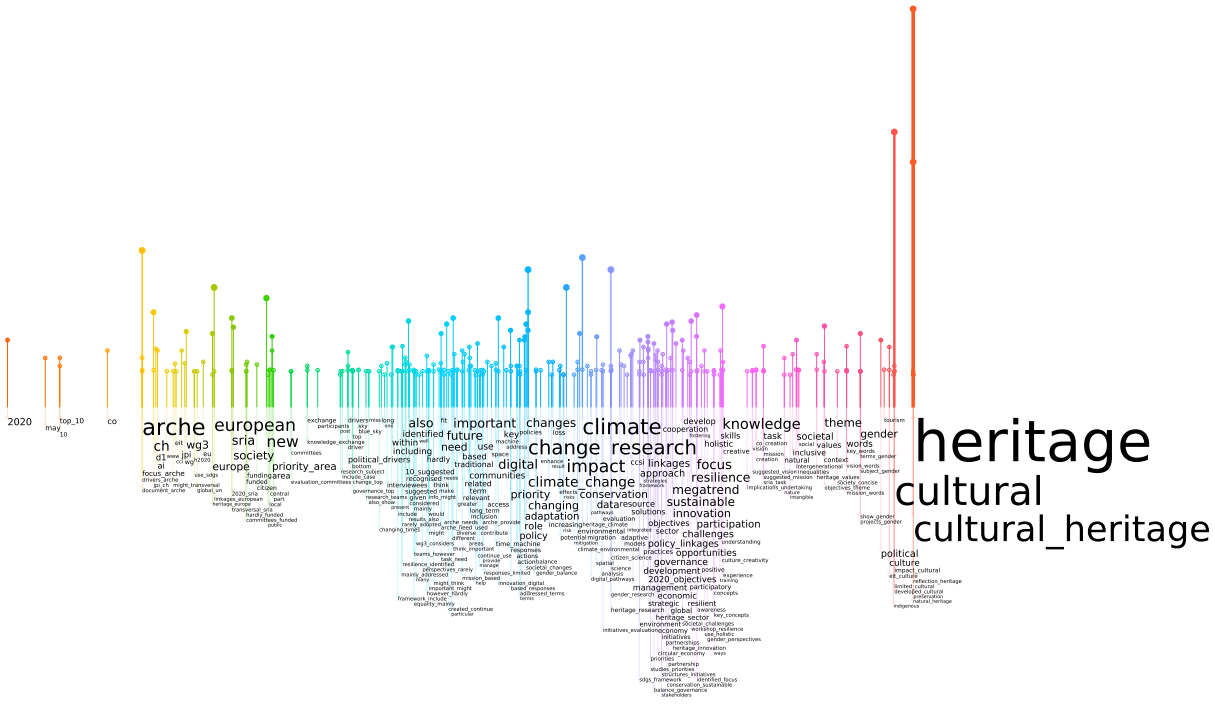

Figure 4: Word Rain of ARCHE WGs forms

4.3. Commenting on Word rains: a few considerations

By uploading ARCHE documents to the tool in one session, it was possible to generate comparable word clouds. The first thing that catches the eye is, to the far right of the infographic, the Cultural Heritage thematic cluster. The biggest terms, hence with the highest semantic value, are as expected fundamental ones, such as *heritage, cultural* and *cultural heritage*. Although this information is not of particular interest, it was decided not to exclude "heritage" and "cultural" from the word cloud. Excluding these terms would have affected all n-grams containing "heritage". In the adjacent section, a few meaningful terms start to emerge. What is interesting here is the comparison of word clusters for the SRIA preliminary findings and WGs forms. Both word rains share similarities in the themes of *cultural resilient, natural heritage, heritage preservation*. The word indigenous appears in both, but that of tourism only appears in the WGs forms. Notably, *tourism* does not appear in any other word cluster along the horizontal axis of the SRIA graphic, and might thus be an indicator of a possible topic that was missed to include.

Another point worth further exploration is the representation of gender issues within the reports. While ***gender*** appears eight times in the Word Cloud for the WG forms — both as a standalone term and within n-grams like *gender perspective, gender balance, and gender research*— it is notably absent in other visualisations. A quick review of the preliminary SRIA document reveals only two occurrences of *gender*, confirming its lack of visibility in the Word Cloud.

Climate is a fundamental theme in all reports, appearing at least once in every word cloud, with increasing frequency from the *Vision and Mission* documents to the *preliminary SRIA*. In the latter, the word cluster related to climate change occupies a significant portion of the entire vector graphic. Next to *climate change* and *climate education* are n-grams such as *risk reduction, resilience, temperature, warming, mitigation* and *adaptation*. Very interestingly, the concept of climate change is semantically close on the x-axis to that of *societal resilience*, and *public awareness* (right) and to the cluster related to *community* and *public goods* (left). At the core of the SRIA we find *geological, stratigraphy, greenhouse, emissions temperature, environmental change, mitigation, adaptation, risk research*, and *risk reduction*. These findings align closely with the report, reflecting its preliminary focus on a comprehensive and multidisciplinary approach to climate change. Terms like *Stratigraphy* resonate with SRIA's emphasis on understanding climate change within a long-term geological and human context, acknowledging humanity's impact on Earth's systems. Key scientific terms such as *greenhouse, temperature*, and *emissions* directly address the priorities of SRIA's around the core mechanisms driving climate change, while terms like *environmental change, mitigation*, and *adaptation* highlight the agenda's commitment to developing strategies that respond to these shifts. The inclusion of *risk research* and *risk reduction* underscores the SRIA's proactive approach to managing climate-related risks to society and ecosystems. Together, these findings mirror SRIA's focus on bridging scientific understanding with practical actions in multiple fields to effectively address climate change challenges.

Within the WGs Forms Word Rain, climate change appears as a theme, albeit less prominently. Notably, this cluster emphasises societal and political participation in climate change discourse, along with heritage preservation. Terms such as *policy linkages, governance, management, innovation*, and *participatory* suggest an emphasis on community involvement and strategic governance in addressing climate issues. Additional topics that surface here include *citizen science, potential migration, gender research, gender perspectives*, and *circular economy*. The distinct presence of *circular economy*—absent from other word clouds — highlights a potential area for inclusion in the final SRIA, underscoring sustainable resource use in response to climate challenges. Similarly, *citizen science*, found only in the WG forms word cloud, highlights the role of societal engagement and collective knowledge building in climate action. *Potential migration* also emerges as a significant factor, recognizing migration as a likely impact of climate change on both cultural heritage and society, warranting deeper analysis. These findings collectively suggest important dimensions of climate change that could further enrich SRIA's focus. In all four word clouds, greenish clusters highlight themes related to Europe and ARCHE. Expectedly, ***ARCHE*** is a prominent keyword across most word clouds, except for the one focused on Future Trends. Since it does not contribute to the identification salient collocations and is not part of

any multiword expression, adding it to a stop-word list could be beneficial. Other recurring keywords include *European Commission, European Union, United Nations, Horizon Europe*, and *heritage research-EU* (with *hub* intentionally removed). It is unclear whether the presence of proper names might be useful to ARCHE experts or just a noise source. Always in the same cluster, ARCHE SRIA's commitment to climate themes reappears in the SRIA word cloud, with keywords such as *climate neutrality, Green Deal,* and *adaptive release* emphasising this ongoing priority.

4.4. Validating results: the ARCHE experts assessment

Once the Word Rain visualisations were generated, they were first interpreted by CLARIN experts and subsequently evaluated in consultation with three E-Rihs experts involved in the ARCHE consortium. Their feedback was particularly useful for 1) identifying lexical noise; 2) excluding terms that, although frequent, were not relevant to the domain; and 3) contextualising why such terms were considered misleading or insignificant in this specific context. For instance, the term *Anthropocene* emerged prominently in the visualisation, initially suggesting thematic relevance. However, experts clarified that its occurrence was related to ongoing debates surrounding the formal recognition of the Anthropocene as a geological epoch, an issue that had recently been resolved with a negative verdict. As a result, the term was deemed irrelevant to the agenda's core research concerns. Following this first consultation, the results obtained were shared with the wider consortium of experts involved in shaping the agenda, allowing them to be involved in the process and provide their feedback. The present terminological experiment was further assessed in comparison to another analysis, which is beyond the scope of this paper, that focused on the responses to a survey launched by the ARCHE consortium to the broader Cultural Heritage community.

5. Conclusions and Future Steps

This experiment demonstrated how looking at keywords and n-grams of a specialized corpus can provide valuable insights for extralinguistic analysis. By offering an immediate and accessible way to examine policy documents and interpret their discourse, Word Rains enabled ARCHE experts to assess their work and reflect on the conceptualizations earlier produced. However, this analysis should be viewed as a preliminary experiment, which requires further research and refinement. Consulting with ARCHE to remove unnecessary stop words in a more systematic fashion might be very useful in observing any shifts in the overall thematic structure. To achieve this, running a custom-based experiment within a vector-based environment using the Word Rain GitHub code appears to be an optimal approach. Pairing the results obtained with insights from other analysis tools, such as Sketch Engine, and using a general language reference corpus for Automatic Term Extraction (ATE), will be necessary to conduct an extensive terminological analysis. Ultimately, the ARCHE corpus will be integrated into a larger corpus of European Cultural Heritage and Climate Change policy reports. The ARCHE documents used for this analysis comprise a variety of textual types, including a foresight analysis, working group discussions on thematically relevant topics, and a formalised output summarising the "preliminary" findings intended to shape the final Strategic Research and Innovation Agenda (SRIA) for Cultural Heritage in the coming years. Although these documents do not conform to conventional policy formats, such as green papers, white papers, or policy briefs—they nonetheless fall within the category of strategic documentation, as they aim to guide and structure collaborative research and action within the domain to which they pertain. Their inclusion in the corpus is therefore justified, particularly within a designated subcorpus focused on documents produced by projects addressing the intersection of Cultural Heritage and Climate Change. This body of grey literature, although often overlooked in mainstream analyses, is notably rich in domain-specific terminology and conceptual formulations. The practice of annotating these documents according to project, document type, and communicative intent aligns with the broader objective of enriching the corpus. This, in turn, contributes directly to the overarching aim of developing a **transdisciplinary glossary** for the Cultural Heritage domain — one that reflects both the expert language and the strategic direction of the field.

6. Acknowledgements

This work is partly supported by the H2IOSC Project - Humanities and cultural Heritage Italian Open Science Cloud funded by the European Union NextGenerationEU - National Recovery and Resilience Plan (NRRP) - Mission 4 “Education and Research” Component 2 “From research to business” Investment 3.1 “Fund for the realization of an integrated system of research and innovation infrastructures” Action 3.1.1 “Creation of new research infrastructures strengthening of existing ones and their networking for Scientific Excellence under Horizon Europe” - Project code IR0000029 – CUP B63C22000730005. Implementing Entity CNR.

The authors express their sincere gratitude to the Alliance for Research on Cultural Heritage in Europe (ARCHE) project, funded by the European Union under the Horizon Europe programme (Grant Agreement No. 101060054). This research would not have been possible without the valuable collaboration of experts from the Alliance, whose willingness to share data and validate results has been crucial.

7. Declaration on Generative AI

During the preparation of this work, the author(s) used Chat-GPT-4 in order to: perform grammar and spelling check. After using this tool, the author(s) reviewed and edited the content as needed and take full responsibility for the publication’s content.

References

[1] E. Cano Díaz, M. Castillejo, B. Ramírez Barat, M. Sanz, M. Martín Gil, M. Bueso, I. Sarró, The european research infrastructure for heritage science (e-rihs): an infrastructure for an interdisciplinary scientific domain, Unpublished (2019). Add journal name or type if known.

[2] F. d. Jong, D. Van Uytvanck, F. Frontini, A. van den Bosch, D. Fišer, A. Witt, Language matters. the european research infrastructure clarin, today and tomorrow, in: D. Fišer, A. Witt (Eds.), CLARIN. The Infrastructure for Language Resources, volume 1 of *Digital Linguistics*, De Gruyter, Berlin, Boston, 2022, pp. 31–58. doi:10.1515/9783110767377-002.

[3] S. Jänicke, G. Franzini, M. F. Cheema, G. Scheuermann, On close and distant reading in digital humanities: A survey and future challenges, in: EuroVis (STARs), 2015, pp. 83–103.

[4] M. Ahltorp, M. Skeppstedt, Word rain as a service, in: CLARIN Annual Conference, 2024, pp. 22–25.

[5] M. Skeppstedt, M. Ahltorp, F. Johansson, S. Velupillai, From word clouds to word rain: Revisiting the classic word cloud to visualize climate change texts, Information Visualization (2024). doi:10.1177/14738716241236188.