

Distant and Close Reading Approaches to Specialised Language in the ARCHE Strategic Research and Innovation Agenda Corpus

Elisa Squadrito^{1,2*}, Francesca Frontini^{2,3} and Vania Virgili⁴

^{1*}Department of Humanities, University of Macerata, Via M. Morbiducci 40, Macerata, 62100, Italy.

²CNR Institute for Computational Linguistics “A. Zampolli” (CNR-ILC), Via G. Moruzzi 1, Pisa, 56124, Italy.

³CLARIN-IT, Via G. Moruzzi 1, Pisa, 56124, Italy.

⁴E-RIHS ERIC, Manifattura Tabacchi, Via delle Cascine 35, Florence, 50144, Italy.

*Corresponding author(s). E-mail(s): e.squadrito@unimc.it

Contributing authors: francesca.frontini@cnr.it; dg@e-rihs.eu

Abstract

This paper reports on a collaboration between the Common Language Resources and Technology Infrastructure (CLARIN) and the European Research Infrastructure for Heritage Science (E-RIHS), exploring how language-based methods can support the interpretation of Cultural Heritage policy documents. The study focuses on documents drafted within the Alliance for Research on Cultural Heritage in Europe (ARCHE) project, forming the basis of the future Strategic Research and Innovation Agenda (SRIA). Combining distant and close reading, the analysis investigates the specialised language of the corpus through Word Rain visualisations of salient lexical items and lexical clusters, complemented by concordance analysis of selected items in context.

Keywords: Cultural Heritage, Distant Reading, Word Clouds, Word Rain, Terminology

1 Introduction

In multidisciplinary policy settings, terminology plays a central role in stabilising concepts, ensuring consistency, and enabling communication among stakeholders from different professional and institutional backgrounds. This is particularly relevant in Strategic Research and Innovation Agendas (SRIAs), where the choice of terms not only ensures domain accuracy, but also contributes to defining research priorities, shaping policy directions, and guiding future strategic action. According to the ERA-LEARN portal,¹ a SRIA is a strategy document produced by a partnership, which identifies its objectives, impact areas, expected outcomes, portfolio of activities, outputs, and milestones within a given timeline. As documents resulting from complex co-creation processes, SRIAs bring together stakeholders from different backgrounds and translate their perspectives into shared research trajectories for future calls and policy actions. As such, the language and terminology they employ need to be widely shared and accepted, and to convey specialised knowledge precisely, especially in transdisciplinary contexts. This is especially relevant in

¹<https://www.era-learn.eu/support-for-partnerships/cross-cutting-issues-and-additional-activities/strategy-and-foresight/what-is-a-sria>

Cultural Heritage and Heritage Science, a broad and multidisciplinary field bringing together different research traditions, professional communities and institutional actors, and addressed here through the SRIA developed by the Alliance for Research in Cultural Heritage (ARCHE).

This paper presents an experiment carried out on a corpus of documents drafted within the ARCHE project, including *Key Messages and Preliminary Findings* (hereafter *Key Messages*), which formed the basis of the SRIA. The work results from a collaboration between experts involved in the ARCHE Consortium, E-RIHS and CLARIN, and explores how distant and close reading methods can support the analysis of specialised language and discourse in Cultural Heritage policy.

The objectives of the analysis are twofold: first, to identify recurrent topics and narratives in the corpus by using salient lexical items, recurrent n-grams and visual lexical clusters as access points to specialised discourse; second, to investigate whether themes from the *Working Group Forms* (hereafter *WG Forms*), representing inputs from the Cultural Heritage community, are absent, less visible, or differently framed in the other documents, especially in the draft SRIA represented by the *Key Messages*.

Section 2 outlines the context of the analysis: SRIAs in the Cultural Heritage sector and the collaboration between CLARIN and E-RIHS. It also provides a brief theoretical clarification of foundational concepts in terminology in order to frame the scope of the analysis. Section 3 outlines the methodological framework of the analysis, introducing word clouds, Word Rain and the concordance-based approach used to complement the visualisations. Section 4 presents the corpus compilation and preprocessing, before moving to the results and discussion in Section 5.

2 Background

2.1 SRIAs in Cultural Heritage and the ARCHE Partnership

Strategic Research and Innovation Agendas (SRIAs) are policy documents aimed at defining research priorities and aligning European research activities around shared societal challenges. In the Cultural Heritage sector, they have been closely connected to the Joint Programming Initiative on Cultural Heritage and Global Change (JPI CH), launched in 2010 to promote a coordinated European approach to Cultural Heritage research and to support the development of strategic agendas for the sector.

To date, three strategic agendas have been published in this field: the JPI CH Strategic Research Agenda in 2014, the JPI CH Strategic Research and Innovation Agenda in 2020, and the ARCHE Strategic Research and Innovation Agenda, published in 2025. The latter was developed by the Alliance for Research on Cultural Heritage in Europe (ARCHE),² a Horizon Europe project centred on the creation of a network of researchers, heritage professionals, organisations, 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 was identified as the drafting of an SRIA intended to translate the Alliance's vision into tangible and actionable goals. More specifically, the SRIA is expected to act as a roadmap for research priorities that will form the basis of calls for projects and other activities within the European Partnership for Resilient Cultural Heritage (RCH), due to start in 2026.³

Starting in 2024, the development of the ARCHE SRIA was rooted in a mapping and assessment endeavour aimed at providing an initial overview of innovative research areas in the Cultural Heritage domain. Building on this foundation, ARCHE project stakeholders contributed to the drafting of the agenda through multiple consultations, including workshops and surveys. The direct involvement of experts was crucial to capturing the perspectives of the many communities of practice addressed by the policy document, as well as to ensuring its accuracy and precision.

2.2 CLARIN and E-RIHS Shared Collaboration

The experiment presented in this paper is the result of a collaboration involving the Italian nodes of the CLARIN (Fišer & Witt, 2022) and E-RIHS (Cano Díaz et al., 2019; Striova & Pezzati, 2017) Research Infrastructures (RIs),⁴ with the aim of applying text-analysis methods as tools for the interpretation of Cultural

²<https://www.heritageresearch-hub.eu/arche-home/about-arche/>

³<https://www.heritageresearch-hub.eu/event/arche-2nd-stakeholders-workshop-to-take-place-in-florence-on-september-25/>

⁴For research infrastructures, see (Monachini & Frontini, 2023)

Heritage policy documents. Foregrounding an understanding of language as “a carrier of social and cultural information” (de Jong et al., 2022, p. 34), the Common Language Resources and Technology Infrastructure, also known as CLARIN, plays a central role in fostering the sharing, use, and sustainability of language data and tools for researchers in linguistics and language technology, as well as in the humanities and social sciences at large (de Jong et al., 2022). In this sense, CLARIN’s mission is rooted in the acknowledgement that language data can support the comparative study of social and cultural phenomena across linguistic, disciplinary, and national boundaries.

The present collaboration with the European Research Infrastructure for Heritage Science (E-RIHS) should be understood within this framework. Experts from E-RIHS IT involved in the development of the SRIA within the ARCHE consortium sought CLARIN expertise for the linguistic analysis of texts produced in preparation for the final strategic agenda. The analysis of the ARCHE SRIA therefore represents a useful point of contact between the two infrastructures: through the adoption of a distant reading approach (Jänicke, Franzini, Cheema, & Scheuermann, 2015), it explores the specialised documents resulting from these consultations, so as to provide valuable insights into both linguistic and extralinguistic issues.

2.3 Terminology, Specialised Communication and Units of Analysis

Before discussing the methods in Section 3, a few theoretical clarifications are necessary regarding terminology as a field of study and the units of analysis it deals with. Following Sager’s definition, terminology can be understood as “the study of and the field of activity concerned with the collection, description, processing and presentation of terms, i.e. lexical items belonging to specialised areas of usage in one or more languages” (Sager, 1990, p. 2). When dealing with specialised knowledge, terms as defined above cannot be properly considered in isolation from concepts. This relation is reflected in ISO 1087:2019, where a *concept* is defined as a “unit of knowledge created by a unique combination of characteristics” (International Organization for Standardization, 2019, 3.2.7), while a *term* is defined as a “designation that represents a general concept by linguistic means” (International Organization for Standardization, 2019, 3.4.2).

As Pecman and Kübler observe, terms are closely connected to specialised knowledge and are therefore crucial to specialised communication and to the study of languages for specific purposes (LSPs) (Pecman & Kübler, 2022, p. 264). LSPs can be understood as “subcodes (that partially overlap with the subcodes of the general language), each of which can be ‘specifically’ characterized by certain particulars such as subject field, type of interlocutors, situation, speakers’ intentions, the context in which a communicative exchange occurs, the type of exchange, etc.” (Cabrè Castellví, 1999, p. 59). In line with this definition, the present study focuses on a specialised language, namely the specialised policy communication of Cultural Heritage and Heritage Science. This communication is situated in a domain-specific institutional context, involves expert and stakeholder communities, and concerns the definition of research priorities, practices, and future strategic actions for the field.

As it is based on the investigation of a corpus of specialised policy texts, the present work does not claim to access concepts directly. Rather, it starts from observable linguistic evidence: lexical items, recurrent multiword expressions, and their contextual patterns in specialised policy documents. In operational terms, the units first identified through Word Rain are treated as salient lexical items. These outputs may include terms or candidate terms, but they are not considered validated terminological units in themselves. Their relevance to the specialised discourse is assessed through frequency information, concordance analysis, Word Sketch inspection, and discussion with domain experts.

3 Methods

The following section outlines the methodological framework adopted for the analysis, which combines distant and close reading. Section 3.1 introduces word clouds and Word Rain as the distant-reading visualisations used to identify salient lexical items and visual clusters, while Section 3.2 describes the concordance-based approach used to contextualise and complement the visual interpretation.

3.1 Distant Reading through Word Rain

The increasing availability of digital collections and computational tools has deeply affected humanities research, contributing to the emergence and consolidation of the Digital Humanities as a field (Jänicke et al.,

2015). Among the methodological shifts associated with this development, visualisation techniques have played a central role, as they allow textual and cultural heritage data to be explored through abstraction, aggregation and visual representation.

In this context, *distant reading*, introduced by Moretti (2005, 2013) in the field of literary studies, opened the way to approaches that shift attention from the direct interpretation of individual passages to the exploration of broader patterns across one or multiple texts (Jockers, 2013; Underwood, 2019). In contrast to *close reading* (Boyles, 2012), which is based on detailed textual interpretation, distant reading aims to make texts accessible through aggregated views of their global features (Jänicke et al., 2015; Jänicke, Franzini, Cheema, & Scheuermann, 2017). This approach can help researchers identify recurrent themes, semantic linkages and broader patterns that may be difficult to detect through manual reading alone, especially across larger or specialised text collections.

As Jänicke et al. (2015) observe, visualisation techniques such as maps, heat maps, graphs and word clouds can support distant reading by facilitating the comprehension and interpretation of complex textual information, and have therefore seen many applications in the Digital Humanities. Among these techniques, word clouds are particularly relevant for the present study because they offer an immediate visual access point to salient lexical items, although, especially in their classical forms, they present limitations in terms of semantic interpretability. For this reason, the present analysis adopts Word Rain, an improved form of word cloud visualisation designed to make word positioning distributionally motivated and more semantically interpretable.

3.1.1 Introducing Word Clouds

Word clouds were inspired by *tag clouds* (Khusro, Jabeen, & Khan, 2021), a visualisation technique that became popular in community-oriented websites such as Flickr⁵ or Delicious (del.icio.us)⁶, where they were used to summarise and browse folksonomies through cloud-like graphical representations (Castella & Sutton, 2013). The technique later moved from social tagging systems to the field of text summarisation, where both *word cloud* and *tag cloud* are often used interchangeably (Heimerl, Lohmann, Lange, & Ertl, 2014). In this paper, the term *word cloud* refers to a graphical representation of a text or collection of documents in which frequently occurring or relevant keywords are displayed in a cloud arrangement.

In word clouds, word prominence is usually encoded through font size, whereas spatial arrangement, colour, orientation and other features may vary depending on the algorithm or tool used to generate the cloud. Their goal is to provide an immediate and intuitive overview of the main contents and topics of a text, which explains their use in contexts ranging from media and education to information retrieval and text analysis (Berson & Berson, 2009; Gottron, 2009; Heimerl et al., 2014; McNaught & Lam, 2010; Viegas, Wattenberg, & Feinberg, 2009). As exploratory tools, word clouds can help researchers gain an initial overview of a text or collection of texts, by making its vocabulary visually accessible. In the case of specialised texts, they may guide terminologists or domain experts towards salient lexical items, recurrent expressions, and possible areas of interest for further analysis.

However, such simplicity and intuitiveness come at the expense of depth and granularity in the type of information provided. Among the most discussed limitations are the difficulty of comparing multiple documents, the representation of temporal variation,⁷ and, most importantly for the present study, the lack of semantically motivated positioning. In classical word clouds, words are often arranged randomly, alphabetically, or according to aesthetic criteria. As Wu, Provan, Wei, Liu, and Ma (2011) observe, this makes it difficult to understand how items connected to a specific topic relate to one another, or how they vary from those related to different topics.

3.1.2 Semantic Word Clouds

In response to this limitation, semantic word clouds have been proposed and evaluated as a way to make spatial arrangement more meaningful, by grouping or positioning words according to semantic relatedness rather than arbitrary placement. Drawing on the distributional hypothesis, according to which linguistic items occurring in similar contexts tend to have similar meanings (Firth, 1957; Harris, 1954), earlier

⁵<https://www.flickr.com/>

⁶The service was acquired by Pinboard and discontinued in 2017.

⁷For a more detailed discussion of works addressing such limitations, see Burch et al. (2014); Castella and Sutton (2013); Collins, Viegas, and Wattenberg (2009); Lee, Henry Riche, Karlson, and Carpendale (2010); Lohmann, Heimerl, Bopp, Burch, and Ertl (2015).

approaches estimated semantic similarity through co-occurrence matrices (Barth, Kobourov, & Pupyrev, 2014; Cui et al., 2010; Wu et al., 2011), while more recent approaches use word embeddings to model semantic relations in distributional space (Ahltorp & Skeppstedt, 2025; Xu, Tao, & Lin, 2016). This development is particularly relevant for the present study, since it provides the methodological basis for choosing Word Rain (Skeppstedt & Aangenendt, 2024; Skeppstedt, Aangenendt, & Söderfeldt, 2024; Skeppstedt, Ahltorp, Kucher, & Lindström, 2024), the visualisation technique adopted for this analysis.

3.1.3 Word Rain

Word Rain is an information visualisation tool, available both as a web application⁸ and as open-source code,⁹ for the generation of semantic word clouds. It 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 Sweden (SWELEN), and the iVis group at Linköping University. Ahltorp and Skeppstedt (2025) conceived the tool with the aim of addressing two major shortcomings shared by most classical word clouds available online, namely: 1) the lack of semantically meaningful spatial arrangement of words, and 2) the over-reliance on font size as the sole indicator of word prominence.

As the authors suggest, word clouds where word positioning is determined by random or aesthetic criteria do not allow for the inference of relations between prominent words; this limitation becomes even more relevant when outputs for multiple documents are compared. Building on the distributional approach, Word Rain addresses this issue by employing word2vec embeddings (Le & Mikolov, 2014), projected onto one dimension with t-SNE (van der Maaten & Hinton, 2008), to encode distributional similarity along the x-axis.

Additionally, the colour grading of words in the bar chart further reinforces and highlights the presence of distinct clusters in the cloud. Word Rain maintains font size as an indicator of word prominence, but does not rely on it as the sole visual cue for its interpretation. Word prominence, which can be measured using either TF-IDF (Spärck Jones, 2004) or raw term frequency depending on the configuration selected by the user, is also encoded through two additional indicators: vertical positioning along the y-axis and a bar chart whose bars indicate the prominence of each word. More specifically, words that would overlap on the x-axis are vertically positioned according to prominence, with less prominent words moved downwards, creating the sense of words “raining down” from which the name Word Rain derives. Moreover, the tool also allows users to customise parameters such as the number of displayed words, application of stop-word lists, extraction of n-grams and the measure used for word prominence.

Overall, Word Rain was chosen because it preserves the intuitive visual logic of classical static word clouds, while also encoding additional information usually lacking in these types of visualisations, such as distributionally motivated word positioning and user-defined visualisation parameters. This spatial organisation can support the interpretation of how lexical items, including terms, are used in related contexts and how they cluster with other items in the corpus. However, spatial proximity in the visualisation is not interpreted here as direct evidence of conceptual proximity in a terminological sense. It is treated instead as an indication of contextual and distributional relatedness between lexical items, which requires verification through close reading.

From a practical standpoint, ease of access and ease of use were also decisive factors in the choice of Word Rain. Compared with other semantic word-cloud tools, the Word Rain algorithm and underlying code are accessible and up to date, thus allowing reproduction and replication of the methodology.

A methodological clarification is necessary regarding the role assigned to Word Rain in this workflow. Unlike automatic term extraction (ATE) approaches, which aim to produce candidate term lists through linguistic, statistical, and neural criteria (Di Nunzio, Marchesin, & Silvello, 2023; Drouin, 2003; Kageura & Umino, 1996), Word Rain is not used here as a substitute for term extraction or terminological validation. Although the tool can rely on weighting measures such as TF-IDF to select salient lexical items for visualisation, its output is not a ranked list of candidate terms, nor a structured terminological resource. The lexical items displayed in the graph may include terms or candidate terms, but they cannot be treated as an exhaustive representation of the terminology of the domain. Systematic candidate-term extraction or

⁸<https://wordrain.org/>

⁹<https://github.com/CDHUppsala/word-rain>

glossary construction would therefore require an additional dedicated extraction, selection and validation workflow.

In the present study, Word Rain is used as an exploratory visual aid to access the vocabulary of the specialised corpus by identifying salient lexical items, recurrent n-grams, and distributionally related lexical clusters. Its main benefit lies in guiding further lexical and terminological analysis when combined with close reading methods. Its visual immediacy also makes it useful for presenting results to domain experts, who can use the visualisations as a starting point for discussing and clarifying salient lexical items. A previous exploratory use of Word Rain with ARCHE domain experts is discussed in [Squadrito, Frontini, Virgili, and Monachini \(2025\)](#).

3.2 Close Reading through Concordance Analysis

As noted above, word-cloud-based visualisations are generally intended as exploratory methods and starting points for analysis, rather than as independent tools for textual interpretation ([McNaught & Lam, 2010](#)). For this reason, they are often incorporated within wider text-analysis environments in the Digital Humanities ([Svensson, 2013](#)), such as Voyant Tools ([Sinclair & Rockwell, 2016](#)), although this is not currently the case for Word Rain ([Ahltorp & Skeppstedt, 2025](#)).

As anticipated above, a combined distant and close reading approach can address this limitation. Whereas Word Rain is used as a distant reading method to identify salient lexical items, distributional clusters and possible thematic patterns, corpus analysis in Sketch Engine ([Kilgariff et al., 2014](#)) is used here as a closer reading stage to validate, contextualise, and complement the visual findings. More specifically, concordance reading ([Tribble, 2010](#); [Wulff & Baker, 2020](#)) was chosen because it allows selected items to be examined in context, while the Word Sketch functionality supports the analysis of recurrent word combinations.

This combination was deemed particularly appropriate for the present corpus. Since ARCHE-WD is a small specialised corpus, some relevant items may have low overall frequency but still become noticeable through their position within broader visual and lexical patterns. Concordance analysis and Word Sketch inspection were therefore used to examine whether the visual patterns suggested by Word Rain were confirmed, refined, or contradicted by textual evidence.

3.2.1 From Distant to Close Reading: Selection and Analysis Protocol

To test the methodology presented above, the experiment focuses on two complementary types of lexical items: recurrent items that are consistently visible across the corpus and relevant to the ARCHE SRIA process, and emerging items that surface in the *WG Forms* visualisation but remain less prominent in the Word Rain outputs of the other subcorpora.

For the first group, selection criteria included recurrence across the four Word Rain visualisations, visual prominence, occurrence in relevant n-grams, and relevance to ARCHE macro-topics. To allow for comparison across subcorpora, selected items had to occur at least once in each Word Rain visualisation. Their centrality to the ARCHE process was instead checked through consultation with project experts, with preference given to items related to themes explicitly discussed as relevant for the final SRIA.

For the second group, the main criterion was visual emergence only in the Word Rain visualisation of the *WG Forms*. This does not imply that the items are absent from the other subcorpora, but that they did not appear in the other subcorpus graphs. In this case, visual prominence was not treated as a necessary condition, since the aim was to identify less represented but potentially meaningful themes. Expert consultation was used here only to avoid the inclusion of items resulting from noise, conflation, or preprocessing artefacts.

Although the Word Rain visualisations were initially inspected across word classes and n-grams, only nouns and noun phrases were retained as focal items for the close reading stage.

For reasons of space, the experiment focuses on four selected items. These were first examined in relation to their distribution and visual behaviour in Word Rain. In the close reading stage, they were then analysed through Word Sketch and concordance inspection. Word Sketches were used to identify frequent collocates and recurrent grammatical patterns, while concordance lines were inspected both in full and in relation to the patterns identified through Word Sketch.

Table 1 Composition of the ARCHE-WD corpus by subcorpus

Subcorpus	Tokens	Share of corpus (%)
ARCHE D2.1 Future Trends on Cultural Heritage (Foresight Analysis)	69,112	68.5
ARCHE D2.4 Vision and Mission	15,775	15.6
ARCHE D2.5 SRIA Key Messages	8,557	8.5
ARCHE SRIA WG Forms	7,468	7.4
Total	100,912	100.0

The contextual features considered included raw and relative frequency, recurrent n-grams, collocational patterns involving the selected item, modifiers, and grammatical patterns. Where relevant, dedicated CQL queries were used to filter concordance lines in which the selected item occurred in specific grammatical positions, especially as subject or object. This made it possible to examine verbs associated with the selected item, as well as nouns modified by, or modifying, the item.

Shifts in framing were identified by comparing these features across subcorpora. In this way, close reading was used to evaluate the visual hypotheses suggested by Word Rain, determining whether they were confirmed, refined, or contradicted by contextual evidence.

4 Data and Analysis

This section introduces the corpus data and the main processing steps used to generate the Word Rain visualisations. After describing the dataset composition in Section 4.1, it details the preprocessing and processing of the texts in Word Rain (Section 4.2).

4.1 Data Collection

The corpus compiled for the analysis, called ARCHE-WD, was provided by E-RIHS partners from the ARCHE Partnership and is composed of four deliverables produced at different stages during the consultation process leading to the publication of the final ARCHE SRIA. The corpus was divided into four subcorpora, one per document, for a total of 100,912 tokens (see Table 1).

The four documents correspond to different moments in the development of the ARCHE SRIA. *Foresight Analysis* represents the output of the initial round of the consultation process leading towards the ARCHE SRIA; *Vision and Mission* defines the shared orientation of the partnership; *Key Messages* constitutes the draft basis for the new SRIA; and the *WG Forms* collect additional inputs received from the Cultural Heritage community through the online public consultation launched in September 2024.

Treating each document as a separate subcorpus was necessary in order to compare the distribution and framing of salient lexical items across different stages of the SRIA process. A second layer of analysis focuses specifically on the comparison between the *WG Forms* and the other documents, most notably *Key Messages*, in order to identify themes raised by the Cultural Heritage community that may be absent, less visible, or differently framed in the SRIA draft.

Although the corpus distribution is unbalanced, with the *Foresight Analysis* accounting for 68% of the total number of tokens, this imbalance was addressed in the concordance analysis through relative frequency measures. As for the Word Rain visualisations, the use of TF-IDF as a measure of word prominence partly mitigates the imbalance, since prominence is not based on raw frequency alone, but also on the distribution of terms across the documents under comparison. Nevertheless, the imbalance was taken into account when interpreting the visualisations.

4.2 Corpus Processing

In Squadrito et al. (2025), the authors presented a first exploratory experiment on the same corpus, relying on the Word Rain web application to generate the outputs. In the present analysis, by contrast, the final visualisations were produced using the code provided by the authors of Word Rain, thereby allowing for finer customisation of the results.

To generate the Word Rain visualisations, the code requires a word2vec embedding space, a stop-word list and a corpus folder. For the embedding space, the pre-trained *word2vec-google-news-300* model

(Mikolov, Chen, Corrado, & Dean, 2013) was selected. This model was chosen because it offers broad lexical coverage, is widely used as a pre-trained word2vec model, and is readily available through the Gensim data repository (Řehůřek & Sojka, 2010).¹⁰

The use of a general-domain pre-trained embedding model entails a methodological limitation. The spatial arrangement produced by Word Rain may partly reflect distributional associations learned from the Google News corpus,¹¹ rather than relations specific to Cultural Heritage policy discourse. Nevertheless, a domain-specific embedding model trained on Cultural Heritage policy documents could not be developed at this stage, as this would have required a larger training corpus than ARCHE-WD.

For stopword filtering, the NLTK stopword list, used by default in the Word Rain web application, was combined with a custom list created for this analysis. The latter included frequent words that did not provide relevant semantic information, such as proper nouns, dates, and noise words appearing in headings and subheadings, e.g. *core theme*, *title*, *keyword*, *ch*, and *hub*. This step significantly improved the final outputs compared to previous experiments based on the web application.

The parameters used for Word Rain generation included TF-IDF, lowercase normalisation, and the inclusion of unigrams and bigrams (n-grams = 1–2), with a minimum document frequency of 1. Visualisation settings included a unified graph layout, compact display mode, a maximum of 300 words per plot, and automatic title generation. Four dotted vertical lines were used to divide the visualisation into visual distributional regions and guide the interpretation of the results.

The subcorpora were run together in order to obtain a unified graph layout, enabling the projection of all documents within a shared semantic space and ensuring that word positions remained consistent across visualisations. The documents were sequentially ordered by date of writing to allow comparison across different stages of the SRIA process. Moreover, the additional “new word” parameter was applied to mark words appearing for the first time in the unified visualisation.¹² Overall, these settings were selected to make clusters directly comparable across graphs, thereby facilitating the identification of thematic similarities and shifts across documents.

The resulting aligned Word Rain visualisations are shown in Figures 1 and 2. The figures included in the article use the vector PDF version of the visualisations in order to preserve readability and allow detailed inspection within the constraints of the publication format. However, they are intended as an overview of the aligned outputs rather than as the sole basis for the interpretation of the results.¹³ Given the limits of static figures in an article, the full Word Rain outputs and related supporting files, including the lists of newly appearing words and exported metrics, are available in the supplementary repository archived on Zenodo (Squadrito, 2026).

5 Results and Discussion

This section combines the distant-reading evidence provided by the Word Rain visualisations with frequency-based observations and close concordance analysis. It first focuses on recurrent items that structure the ARCHE SRIA discourse, namely *climate* and *resilience*, before considering less visible but thematically relevant items such as *gender* and *migration*.

To complement the Word Rain visualisations, Table 2 reports the raw and relative frequencies of the selected lexical items across the subcorpora. Although Word Rain exports tool-generated scores used to rank and visualise lexical items, these scores are dependent on the tool settings and preprocessing pipeline. The close reading analysis therefore reports raw and relative frequencies as transparent reference measures, while the full Word Rain output files, including the exported scores, are made available as supplementary material.

In accordance with the criteria outlined above, four lexical items were selected for analysis: *climate* and *resilience*, as recurrent items central to the ARCHE SRIA process, and *gender* and *migration*, as items occurring only in the *WG Forms* Word Rain graph. The analysis first examines the distribution and visual prominence of *climate* and *resilience* across the subcorpora, with particular attention to neighbouring lexical items, clusters, and recurrent n-grams. Close reading is then used to determine whether the visual

¹⁰<https://github.com/piskvorky/gensim-data>

¹¹<https://code.google.com/archive/p/word2vec/>

¹²This feature is highly dependent on corpus ordering, since words that already appear in a previously generated graph are not marked as new in subsequent visualisations.

¹³The visualisations should be interpreted alongside close-reading results, rather than as self-standing evidence.



Fig. 1 Aligned Word Rain visualisations for the ARCHE-WD subcorpora, part 1.

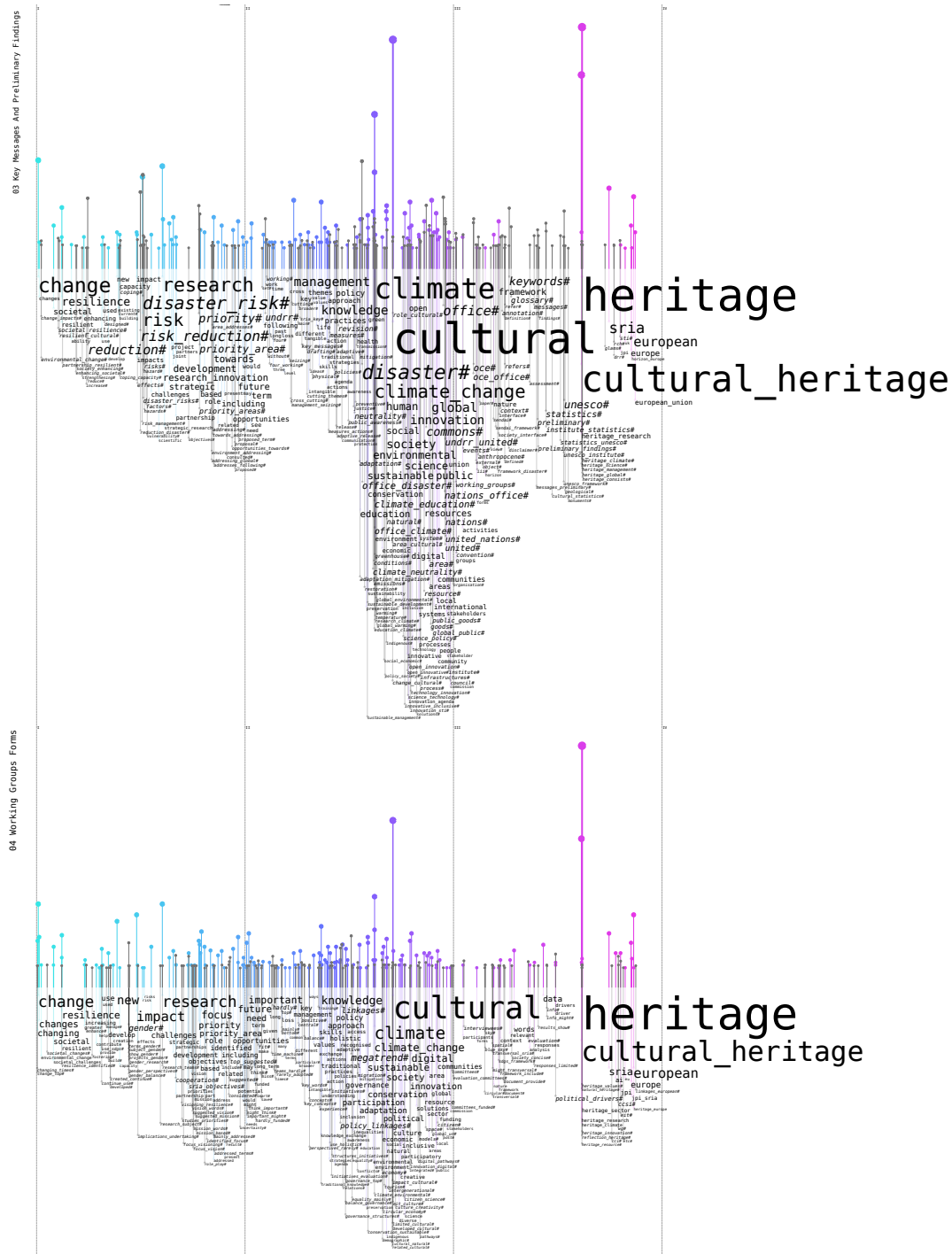


Fig. 2 Aligned Word Rain visualisations for the ARCHE-WD subcorpora, part 2.

Table 2 Distribution of the selected lexical items across the ARCHE-WD subcorpora. Values report raw frequency followed by relative frequency per 10,000 tokens in parentheses. FA = Foresight Analysis; VM = Vision and Mission; KM = Key Messages; WGF = WG Forms.

Item	FA	VM	KM	WGF	Total
<i>Recurrent items</i>					
<i>climate</i>	132 (19.05)	12 (7.61)	85 (99.33)	41 (54.90)	270 (26.76)
<i>resilience</i>	49 (7.09)	20 (12.68)	23 (26.88)	20 (26.78)	112 (11.10)
<i>Emerging items</i>					
<i>gender</i>	6 (0.87)	1 (0.63)	2 (2.34)	16 (21.42)	25 (2.48)
<i>migration</i>	15 (2.17)	5 (3.17)	2 (2.34)	7 (9.37)	29 (2.87)

hypotheses suggested by Word Rain are confirmed, refined, or contradicted by concordance and Word Sketch evidence. For readability, the aligned Word Rain visualisations are shown in two parts in Figures 1 and 2. The panels derive from the same aligned Word Rain output and are separated only to improve readability. Information on the corresponding subcorpus is provided to the left of each graph.

Section 5.3 focuses on *gender* and *migration*, examining their occurrence in the *WG Forms* and their absence or limited visibility in the other subcorpora, in order to assess whether they point to potentially under-represented themes in the policy discussion.

5.1 Climate: From Crisis and Uncertainty to Policy-oriented and Participatory Action

As shown in Table 2, *climate* appears 132 times in the *Foresight Analysis* subcorpus, corresponding to 19.05 occurrences per 10,000 tokens. Although this is the highest raw frequency, it is the second-lowest relative frequency among the four subcorpora.

Climate-related lexical items are consistently present across the four Word Rain visualisations, both as the standalone item *climate* and as part of recurrent n-grams, such as *climate change*, *climate crisis*, *climate neutrality*, *climate education*, and *heritage climate*. Rather than treating *climate* only as a keyword, the visualisations make it possible to observe how the item enters different distributional neighbourhoods and how its lexical environment changes according to the communicative function of each document. Across the graphs, climate-related vocabulary intersects with several recurrent domains, including sustainability, environmental change, governance, public action, education, risk, and social transformation. However, the distribution and density of these associations vary from one document to another.

5.1.1 Foresight Analysis: Climate as Crisis, Uncertainty and Heritage Transformation

In the *Foresight Analysis* Word Rain (Figure 1), *climate* is represented by three items: the standalone form *climate*, and the n-grams *climate change* and *climate crisis*. These appear within a broad lexical environment in which neighbouring items point not only to sustainability and the environment, but also to social and political dimensions, including *governance*, *political*, *social*, *human*, *society*, and *wellbeing*. This suggests a framing of climate as both an environmental issue and part of a broader socio-political and sustainability-oriented discourse.

The presence of the n-gram *climate crisis* is particularly relevant since, although not visually dominant, the *Foresight Analysis* is the only graph in which *crisis* appears. Its surrounding lexical region includes items such as *conflict*, *health*, *wellbeing*, *pandemic*, *inequality*, *protection*, and *inequalities*, suggesting a distributional environment in which climate is connected to broader conditions of instability, vulnerability, and social disruption. This does not mean that all these items are directly syntactically related to *climate*; rather, their visual co-location points to a discursive area in which climate, crisis, health, inequality, and conflict participate in a wider vocabulary of uncertainty and transformation.

A further element emerging from the *Foresight Analysis* graph is the proximity of climate n-grams to items associated with public action and knowledge-building, such as *awareness*, *recognition*, *understanding*, *participation*, *engagement*, *learning*, *education*, and *lifelong learning*. This suggests that climate is also positioned within a vocabulary of public response, social learning, and participatory adaptation. Overall,

Table 3 Top collocates of *climate* in the *Foresight Analysis* subcorpus, L3–R3

Collocate	Co-occurrence	MI3	LogDice
<i>crisis</i>	39	19.91	12.81
<i>change</i>	50	18.90	12.29
<i>Change</i>	18	17.13	11.82

the Word Rain visualisation suggests a broad framing of climate that combines sustainability, crisis, social vulnerability, governance, and learning.

This visual interpretation was then checked through Sketch Engine.¹⁴ A lemma-based search for *climate* returned 132 occurrences in the *Foresight Analysis* subcorpus. Collocates were computed within an L3–R3 window using co-occurrence frequency, MI3, and LogDice. The results show that *crisis* and *change* are among the strongest collocates of *climate*, confirming the relevance of the n-grams *climate crisis* and *climate change* in the Word Rain visualisation (Table 3).

Concordance reading of *climate change* shows that the n-gram appears in association with environmental degradation, global uncertainty, political tensions, economic instability, wellbeing, inequality, and social change. This confirms the visual proximity of climate-related items to words such as *social*, *political*, *economic*, *wellbeing*, *society*, and *governance* in the graph.

The prominence of *crisis* as a collocate further indicates that *climate* is repeatedly framed through the language of emergency and disruption. Concordance lines connect *climate crisis* to conflict, migration, health, poverty reduction, technology governance, and global cooperation, supporting the interpretation that, in the *Foresight Analysis*, climate is embedded in a broader vocabulary of instability and polycrisis:

“increased transnational *migration* – largely induced by the **climate crisis**, *wars* and *conflict* – is likely, creating progressively more heterogeneous societies”

“This isolationism hinders global cooperation in tackling common challenges such as the **climate crisis**, *health*, *poverty reduction* and *technology governance*.”

At the same time, the concordances show that the crisis and change framing is not entirely negative or defensive. Some passages present climate change as a condition that may require a redefinition of heritage values and management practices, through the recognition of different knowledge systems, the creation of new forms of heritage, and a shift from protective approaches to a more dynamic management of change:

“With heritage understood here as a range of human (and living) archives of tangible and intangible heritage, the **climate crises** can [...] create new forms of heritage [...] through adaptation to new climate-based norms; inspire the need to recognise different knowledge systems and their contributions to climate change responses.”

“Change is unavoidable to CH, so move away from defensive protection of CH against threat to dynamic management of change that meets environmental demands and changing interests; focus on preservation of values. [...] Seeing change (e.g., **climate change**) as part of CH, not just something that acts on it.”

Close reading therefore confirms the main visual hypothesis suggested by Word Rain, while also refining it. In the *Foresight Analysis*, *climate* is framed both as a source of risk and instability and as a catalyst for rethinking the role of cultural heritage in uncertain futures.

5.1.2 Vision and Mission: Low Salience and Green-transition Vocabulary

In the *Vision and Mission* Word Rain (Figure 2), *climate* is much less prominent than in the other visualisations: it appears only once as a standalone item and is placed low along the vertical bar chart in the second region. This limited visual salience is consistent with the frequency data, since *climate* has the lowest relative frequency in the *Vision and Mission* subcorpus. Although *climate change* is not visible in the graph, neighbouring lexical items such as *green*, *green transition*, *green technologies*, *transition*, *sustainable*, *sustainable growth*, *environment*, *policy*, and *economic* suggest a broader focus on environmental and sustainability-related concerns.

¹⁴<https://www.sketchengine.eu/>

Compared with the *Foresight Analysis*, where *climate* is framed through uncertainty, crisis and social change, the *Vision and Mission* Word Rain places climate-related concerns within a wider lexical environment related to transition, sustainability, and institutional action. The focus is therefore not on *climate* as a central item, but on the actions and frameworks through which green-related transitions are to be addressed. N-grams such as *green transition* and *green technologies* suggest a more programmatic and policy-oriented vocabulary, where environmental concerns are embedded in the language of innovation, action, and strategic planning.

The presence of *holistic* and *holistic approach*, positioned near items such as *green*, *agenda*, *quality*, and *actions*, also suggests an attempt to present transition-related issues as integrated and cross-sectoral. Similarly, the proximity of *quality* and *quality life* to *health*, *wellbeing*, *growth*, and *knowledge* indicates that environmental transition is connected to a wider vocabulary of social benefit. The graph therefore suggests that *Vision and Mission* does not foreground climate as a discrete item, but incorporates climate-related concerns into a broader institutional vocabulary of transition, innovation, quality of life, and holistic action.

Close reading refines this interpretation. In the *Vision and Mission* subcorpus, *climate* occurs 12 times, seven of which appear in the expression *climate change*. This low number of occurrences confirms that climate is not central to the document. However, when climate is explicitly mentioned, it is associated with *climate change*, *heritage loss*, *climate neutrality*, *mitigation*, and *adaptation*. These references occur alongside a broader policy vocabulary related to *green technologies*, the *European Green Deal*, *green and digital transitions*, *innovation*, *capacity-building*, and research partnerships, including references to Horizon Europe frameworks. In this sense, close reading confirms the limited salience of *climate* in the Word Rain, while also showing that its occurrences are aligned with the green-transition and strategic-action vocabulary suggested by the visualisation.

5.1.3 Key Messages: Specialised Climate-policy and Educational Vocabulary

In the *Key Messages* Word Rain (Figure 2), *climate* is more visually prominent than in the other graphs and appears in a wider range of n-grams, including *climate change*, *climate education*, *climate neutrality*, *research climate*, *education climate*, *office climate*, and *heritage climate*. These n-grams are reported as they appear in the Word Rain output: while some correspond to recognisable expressions, such as *climate change* and *climate neutrality*, others require concordance checking to determine whether they point to thematic patterns, institutional references, or artefacts of n-gram extraction. Surrounding items are related to environment, sustainability, public awareness, science and society, including *environmental*, *global environmental*, *education*, *public awareness*, *sustainable*, *sustainability*, *sustainable development*, *society*, *social*, *public*, *human*, *global*, *people*, and *public goods*. This suggests continuity with the broad social and sustainability framing already observed in the *Foresight Analysis*.

At the same time, several items marked as new through the Word Rain #newword function point to a more specialised climate-policy and environmental vocabulary: *emissions*, *greenhouse*, *warming*, *temperature*, *global warming*, *adaptation*, *adaptation mitigation*, and *climate neutrality*. These items suggest that, in this graph, climate is not only connected to general sustainability or social change, but also to policy, environmental science, mitigation, and adaptation. The presence of *measures actions* and related items further points to an operational dimension.

The appearance of *Anthropocene* and *geological* in the third region of the graph may point to a broader contextualisation of climate change. However, as discussed in Squadrito et al. (2025), consultation with heritage experts from the ARCHE Consortium clarified that these terms were mainly present because of ongoing discussions concerning the rejection of the Anthropocene as a geological epoch.

Other visually salient items also require contextual verification. For example, *climate education*, *education climate*, *public awareness*, and *OCE Office* suggest a possible link between climate, education, and public understanding. Similarly, the n-gram *heritage climate* suggests a possible connection between climate-related and heritage-related vocabulary, although the nature of this relation cannot be established from the visualisation alone. Overall, the Word Rain suggests that *Key Messages* contains the most diversified climate-related vocabulary in the corpus, combining environmental, scientific, educational, policy-oriented, and heritage-related dimensions.

Close reading confirms this diversified lexical pattern, while also refining some of the visual indications. In the Word Sketch for *climate*, nouns modified by the term include *change*, *neutrality*, *education*,

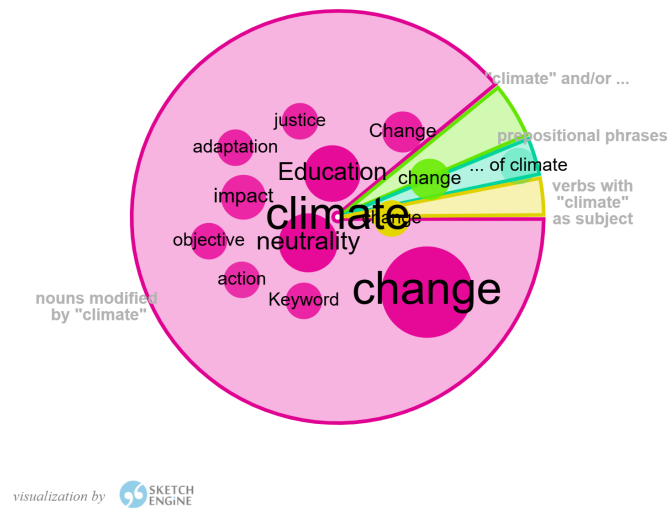


Fig. 3 Word Sketch for *climate* in Key Messages.

Details	Left context	KWIC	Right context
1	to expected climate change impacts (OCE, Office for	Climate Education , n.d.).</s><s>Adaptive release A proposed term for th	
2	i, years, decades, centuries or more) (OCE, Office for	Climate Education , n.d.).</s><s>Climate change Climate change refers	
3	ie term "global warming" is also used (OCE, Office for	Climate Education , n.d.).</s><s>Climate justice This term is used to ack	
4	lar, social and environmental justice) (OCE, Office for	Climate Education , n.d.).</s><s>Climate neutrality The term refers to the	
5	landslides, droughts and heatwaves (OCE, Office for	Climate Education , n.d.).</s><s>Gender lens Gender lens takes the exis	
6	e, methane, nitrous oxide and ozone (OCE, Office for	Climate Education , n.d.).</s><s>Hazard A process, phenomenon or hun	
7	generations to meet their own needs (OCE, Office for	Climate Education , n.d.).</s><s>Sustainable management It is defined a	
8	tures and few economical resources (OCE, Office for	Climate Education , n.d.).</s><s>The conditions determined by physical,	
9	hange.</s><s>(n.d.).</s><s>Retrieved from OCE, for	Climate Education : Glossary of IPBES - Anthropocene.</s><s>(n.d.).</s>	

Fig. 4 KWIC concordance lines for *climate education* in Key Messages.

impact, and *action* (Figure 3). These patterns support the interpretation that *climate* is more central in this subcorpus than in *Vision and Mission*, and that it is connected to both policy-oriented and educational vocabulary.

Some associations, however, need to be interpreted cautiously. For instance, *climate education* appears visually relevant in the graph, but concordance analysis shows that its prominence is partly due to repeated references to the “Office for Climate Education”,¹⁵ which also appears in the graph as *OCE Office* (Figure 4). This refines the Word Rain output: the n-gram points to climate-related educational activity, but part of its salience is due to references to a specific institution rather than to a broader thematic discussion of climate education.

By contrast, *climate neutrality* appears as a more stable policy expression (Figure 5). Concordance analysis shows that, out of 17 occurrences of *neutrality* in the corpus, 11 occur in *Key Messages*, all in the expression *climate neutrality*. This pattern is reinforced by related items such as *emissions*, *greenhouse*,

¹⁵<https://www.oce.global/en/>

Details	Left context	KWIC	Right context
1	ovative and Inclusive Cultural Heritage Research for	Climate Neutrality	2.3.1. 2.3.2. 2.3.3.</s><s>Annexes Communication,
2	he objectives of the Paris Agreement ¹⁵ and attaining	climate neutrality	by 2050. ¹⁶ Such a role highlights the functions of cult
3	omy and ultimately contribute to the goal of achieving	climate neutrality	in different sectors by 2050.</s><s>It is therefore imp
4	egative effects of environmental change, to attain the	climate neutrality	objectives, and to allow for the positive results of activ
5	aming heritage in all areas of society and meeting the	climate neutrality	objectives.</s><s>Priority Area C addresses the follo
6	ovative and Inclusive Cultural Heritage Research for	Climate Neutrality	The SRIA aims at integrating the preservation of cultu
7	: of climate change with the broader goal of achieving	climate neutrality	.</s><s>Reaching climate neutrality requires a comp
8	goal of achieving climate neutrality.</s><s>Reaching	climate neutrality	requires a comprehensive set of actions that establish
9	nitatives', thereby contributing to the advancement of	climate neutrality	.</s><s>The cross-cutting themes will be taken into a
10	ovative and Inclusive Cultural Heritage Research for	Climate Neutrality	". 2.3.1.</s><s>Communication, Education and Public
11	e) (OCE, Office for Climate Education, n.d.).</s><s>	Climate neutrality	The term refers to the idea of achieving net zero gree

Fig. 5 KWIC concordance lines for *climate neutrality* in *Key Messages*.

global warming, and *warming*, and supports the interpretation of a stronger climate-policy vocabulary in this subcorpus.

The close reading also identifies a broader vocabulary of climate-related social and political action, including *climate justice*, *climate action*, and *climate activism*. These expressions indicate that climate is framed not only as an environmental or policy issue, but also as a social and political concern. Conversely, the items *Anthropocene* and *geological*, although visible in the Word Rain, mostly occur in relation to the debate around the proposed formalisation of the Anthropocene as a geological unit. They therefore broaden the contextual range of climate-related discussion, but should not be over-interpreted as evidence of a stable geological framing.

Overall, close reading confirms that *Key Messages* contains the most terminologically dense treatment of *climate* in the corpus, while also refining the Word Rain interpretation. Compared with the *Foresight Analysis*, where *climate* is strongly associated with uncertainty and crisis, this subcorpus frames climate through a more specialised vocabulary of climate neutrality, emissions, mitigation, education, public awareness, and policy action.

5.1.4 WG Forms: Applied and Participatory Climate Response

In the *WG Forms* graph (Figure 2), *climate* appears as *climate*, *climate change*, and *climate environmental*, always in the second region. A closer look at neighbouring items reveals the presence of governance-related vocabulary, including *policy linkages*, *governance*, *balance governance*, and *governance structure*. This suggests that, in the *WG Forms*, climate is approached through frameworks of coordination, policy alignment, and institutional organisation.

At the same time, the visual area around *climate* includes a more visible participatory vocabulary than in the previous documents. Items such as *participation*, *participatory*, *citizen science*, *citizen*, and *inclusion*, together with a smaller cluster related to equality, including *equality*, *equality mainly*, and *inequalities*, point to civic engagement and social justice as relevant dimensions in the *WG Forms*. Given their proximity to climate-related items, these lexical patterns suggest a possible connection between climate, participation, and equality. This marks a shift from the more specialised policy and scientific vocabulary of the *Key Messages* graph towards a vocabulary of participation and collective action.

The presence of *adaptation* and *mitigation* confirms the continuity of the climate-policy vocabulary already observed in the *Key Messages*. However, the surrounding items suggest that these concepts are embedded here in a more socially and operationally oriented framework. In the *WG Forms*, climate is

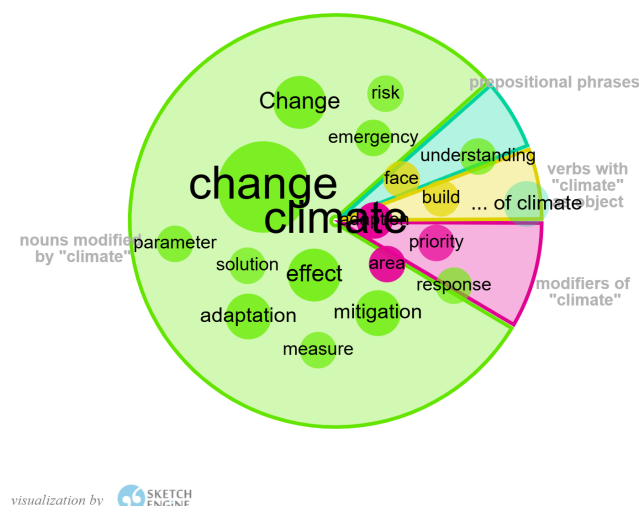


Fig. 6 Word Sketch for *climate* in *WG Forms*.

therefore not only linked to environmental transformation or policy response, but also to participation, inclusion, equality, and citizen involvement. This suggests that climate-related vocabulary is shaped by the consultative and participatory function of the document.

The Word Sketch for *climate* in the *WG Forms* (Figure 6) refines the Word Rain findings. As in the *Key Messages*, *climate* modifies nouns such as *change*, *mitigation*, *adaptation*, *emergency*, *effect*, *measure*, and *solution*, pointing to a practical framing centred on response, adaptation, and mitigation. These results are consistent with the *WG Forms* Word Rain, where *adaptation* and *mitigation* appear near *climate* in the same visual region of the visualisation.

Concordance analysis also serves to clarify the relationship between heritage and climate-related vocabulary (Figure 7). The n-gram *heritage climate* appears only twice in the Word Rain visualisation, namely in the *Key Messages* and *WG Forms* graphs. A CQL query run on both subcorpora for occurrences of *heritage* and *climate* within a span of three tokens yields a total of 13 lines (see Figure 7 for selected KWIC concordance lines).

The results confirm the twofold framing of heritage observed above: heritage is represented both as exposed to climate and environmental change and, at the same time, as a resource for anticipating, mitigating, and adapting to climate-related effects. In Figure 7, heritage is presented as facing climate and environmental change, while also playing a central role in climate change adaptation and mitigation strategies. While the examples refer to narrative excerpts or to priority areas and core themes of the strategy, this representation echoes the *Foresight Analysis*, where climate change was presented not only as a threat to heritage, but also as a condition requiring new forms of heritage management. What emerges is the inextricable relationship between heritage and climate action in the ARCHE policy discourse.

Overall, close reading confirms the main Word Rain interpretation that *climate* is a recurrent item across the corpus, while also showing that it is framed differently across subcorpora. It refines the visual evidence in at least three ways: it shows that *climate crisis* in the *Foresight Analysis* is connected to wider social instability and heritage transformation; that *climate education* in *Key Messages* is partly driven by institutional references; and that *heritage climate* in the *WG Forms* points to a twofold representation of heritage as both vulnerable to and useful for climate response.

5.2 Resilience

A second recurrent lexical item across the Word Rain visualisations is *resilience*, which repeatedly appears in relation to a vocabulary of change, adaptation, strengthening, societal response, and capacity-building. The visualisations suggest that *resilience* is consistently associated with change, but that this association

	Details	Left context	KWIC	Right context
1	uth, Sport and Culture, Strengthening cultural	, <u>heritage</u>	resilience for climate	change – Where the European Green Deal m
2	je for resilient society - WG2.</s><s>Cultural	, <u>heritage</u>	facing climate	and environmental change - WG3.</s><s>Cul
3	efore important to demonstrate that cultural	, <u>heritage</u>	is central to climate	change understandings, responses and action
4	inerability.</s><s>A.2.</s><s>Role of cultural	, <u>heritage</u>	for climate	change adaptation and mitigation strategies K
5	<s>Nexus between geopolitical crisis, cultural	, <u>heritage</u>	, and climate	change Keywords: wars and conflicts, resourc
6	munity, inclusion.</s><s>B.5.</s><s>Cultural	, <u>heritage</u>	and climate	change intertwined narratives Keywords: clima
7	on.</s><s>C.4.</s><s>Economics of cultural	, <u>heritage</u>	for climate	action Keywords: circular economy, resource
8	aty.</s><s>Biodiversity WG2.</s><s>Cultural	, <u>heritage</u>	facing climate	and environmental change Priority Area: Cultu
9	nvironmental changes on cultural and natural	, <u>heritage</u>	Climate	change has an impact on our environment, so
10	ion and prediction; * demonstrate that cultural	, <u>heritage</u>	is central to climate	change understandings, responses and action
11	responses it should not be limited to cultural	, <u>heritage</u>	and climate	change.</s><s>This is because the top 10 su

Fig. 7 Selected KWIC concordance lines for *heritage climate* in *WG Forms* and *Key Messages*. CQL query: [lemma="heritage"] []{0, 3} [lemma="climate"]].

is framed differently depending on the communicative function of each subcorpus. As with *climate*, the relevance of *resilience* lies not only in its frequency, but also in the lexical environments in which it appears across the four subcorpora.

As reported in Table 2, *resilience* occurs 112 times across the corpus: 49 times in the *Foresight Analysis* (7.09 per 10,000 tokens), 20 times in *Vision and Mission* (12.68 per 10,000 tokens), 23 times in *Key Messages* (26.88 per 10,000 tokens), and 20 times in the *WG Forms* (26.78 per 10,000 tokens). In several cases, occurrences of *resilience* appear in headings, titles, references, or names of cross-cutting themes. These occurrences were considered when interpreting the distributional patterns, but the close reading focused mainly on concordance lines where the item contributed to the framing of the documents.

5.2.1 Foresight Analysis: Resilience, Change and Uncertain Futures

In the *Foresight Analysis* Word Rain (Figure 1), *resilience* appears once and is positioned near lexical items related to change, including *change(s)*, *changing*, and *shift(s)*. Its limited visibility in the graph, together with the relatively low density of the surrounding cluster when compared with the other aligned visualisations, suggests that *resilience* is less visually prominent in this graph. On the basis of the Word Rain alone, however, it is not possible to determine what type of resilience is being represented. The presence of *societal* near *resilience* may suggest a connection with societal resilience, but this requires concordance checking.

A further observation concerns the visual area around *resilience*, in the central region of the first quadrant, where several verbs and degree-related forms appear, including *increased*, *increasing*, *enhancing*, *superior*, and *greater*. These items point to a vocabulary of intensification and improvement, suggesting that *resilience* may be framed not as a static condition, but as something that can be increased, enhanced, or strengthened. This visual pattern therefore lends itself to closer inspection through concordance analysis.

Concordance reading confirms and refines the pattern suggested by the Word Rain. In the *Foresight Analysis*, *resilience* is repeatedly linked to uncertainty, adaptation, and change. A particularly relevant point is that change is not represented only as a threat, but also as a possible condition for resilience. This is visible in passages where change is described as a value to increase resilience, or even as a means to increase community resilience:

Details	Left context	KWIC	Right context
1	ovation actions.	, strengthen the resilience	of the cultural heritage of Europe and benefit futu
2	rofessional practice, policy and societal impacts,	, strengthen the resilience	of the cultural heritage of Europe in response to t
3	hange and capacity building activities in order to	, strengthen the resilience	of cultural heritage in Europe during the twin digit

Fig. 8 Selected KWIC concordance lines from *Vision and Mission*. CQL query: [lemma="strengthen|strengthening"] [] {0, 3} [lemma="resilience"].

"The climate crisis will likely shape a persistently uncertain future where communities must adapt to and accept change, loss and transformation to a greater extent. [...] Change itself may be prioritised as a value to increase **resilience**."

"Embracing change in the heritage sector refers to approaching change in cultural heritage as inevitable, and even as a means to increase community **resilience**."

These examples confirm that the visual proximity between *change* and *resilience* corresponds to a contextual pattern in the text. In this document, *resilience* is associated with adaptation to uncertainty, but also with a possible revaluation of change as a resource for heritage practice. Concordance lines also clarify the types of resilience involved, including *community resilience*, *cultural resilience*, and *climate resilience*, which were not immediately identifiable from the Word Rain graph alone.

Close reading therefore confirms the Word Rain cue linking *resilience* to change, while refining its interpretation. In the *Foresight Analysis*, *resilience* is not simply framed as resistance to threat, but as a capacity that can be increased through adaptation, acceptance of change, and new approaches to heritage management.

5.2.2 Vision and Mission: Resilience as Strengthened Capacity

Compared with the *Foresight Analysis*, the *Vision and Mission* Word Rain (Figure 1) gives greater visibility to *resilience* and related forms. The item appears both as a noun, in *resilience* and in the Word Rain n-gram *resilience cultural*, and in the adjectival form *resilient*. The graph shows a closer association between resilience, cultural heritage, and societal response, suggesting that the subcorpus deals with cultural and societal resilience. Nevertheless, the interpretation of forms such as *resilience cultural* requires concordance checking, since they reflect the output of n-gram extraction rather than necessarily stable expressions.

The visual area surrounding *resilience* contains several items related to strengthening, capacity-building, and response. References to change, through *change* and *changing*, appear alongside n-grams such as *respond societal*, *ability respond*, and *strengthened ability*. These patterns suggest that *resilience* is framed less in its broader sense condition and more as an ability to be strengthened in response to societal change. Such an interpretation is reinforced by the cluster of strength-related forms positioned near *resilience*, including *strength*, *strengthen*, *strengthened*, and *strengthened ability*. Other nearby verbs, such as *stimulate*, *promote*, *boost*, *achieve*, *respond*, *develop*, *build*, *enable*, and *benefit*, further point to an action-oriented vocabulary of activation, support, and capacity-building.

However, the graph alone does not clarify whether resilience enables these processes or is itself produced by them. This makes concordance analysis necessary, especially for expressions involving *ability*, *respond*, and *strengthen*. The KWIC concordance lines for *strengthen* and *resilience* in the subcorpus show recurrent formulations such as *strengthen the resilience* and *strengthened resilience*, especially in relation to cultural heritage in Europe (Figure 8). These patterns support the interpretation that, in *Vision and Mission*, resilience is framed as a capacity that can be actively supported through collaboration, funding actions, knowledge exchange, and capacity-building.

The connection between *resilience*, *ability*, and *response* also helps explain the Word Rain cluster around *strengthened ability*, *respond*, *societal*, and *resilience*. At the same time, concordance analysis shows that these items are not always syntactically attached to *resilience*; rather, they belong to a recurring institutional formulation of capacity-building and societal response. Close reading therefore confirms the Word Rain association between *resilience* and strengthening, while refining it by showing that the pattern is connected to the institutional and strategic role of the ARCHE partnership.

Details	Left context	KWIC	Right context
1	tural heritage, particularly in the context of developing its	<u>,resilience</u>	against global climate change.</s><s>This evolving draft
2	Youth, Sport and Culture, Strengthening cultural heritage	<u>,resilience</u>	for climate change – Where the European Green Deal m
3	/stems, the aim of adaptation is to reduce risks, increase	<u>,resilience</u>	or seize on beneficial opportunities.</s><s>In natural sys
4	and reconstruction phases after a disaster to increase the	<u>,resilience</u>	of nations and communities through integrating disaster r
5	ociety to manage and reduce disaster risks and strengthen	<u>,resilience</u>	</s><s>Annotation: Capacity may include infrastructure,
6	ging residual risk, all of which contribute to strengthening	<u>,resilience</u>	and therefore to the achievement of sustainable develop
7	ngthening of economic, social, health and environmental	<u>,resilience</u>	</s><s>A global, agreed policy of disaster risk reduction

Fig. 9 Selected KWIC concordance lines for *resilience* in *Key Messages*.

Compared with the *Foresight Analysis*, where *resilience* is linked to uncertainty and the acceptance of change, *Vision and Mission* frames resilience more explicitly as a cultural and societal capacity to be strengthened. This creates continuity with the earlier document, since resilience remains associated with change, but the emphasis shifts from adaptation to uncertain futures towards institutional support, collaboration, and strategic capacity-building.

5.2.3 Key Messages: Resilience, Risk and Climate Governance

In the *Key Messages* Word Rain (Figure 2), *resilience* appears near items related to disaster, risk, change, and society. This positioning suggests a shift towards a more policy-oriented and risk-related lexical environment. The item appears alongside *change*, *changes*, *change impacts*, and *environmental change*, while also occurring near *societal*, *societal resilience*, and the Word Rain n-gram *resilient cultural*. Compared with *Vision and Mission*, where the dominant pattern concerned strengthening and capacity-building, the *Key Messages* graph places resilience more explicitly in relation to risk, preparedness, and climate-related response.

The presence of items such as *reduce*, *strengthen*, *increase*, *coping*, *enhancing societal*, and *strengthening* confirms the continuity of the action-oriented vocabulary already observed in the previous documents. However, in this graph these forms are more closely connected to risk management and societal response. The proximity of *resilience* to *disaster* and *risk* suggests that resilience is not only associated with general change, but also with more specific forms of vulnerability, impact, and preparedness. The n-gram *partnership resilient* also points to the role of the ARCHE partnership within this lexical environment, suggesting that resilience is framed not only as a property of cultural heritage or society, but also as an objective of partnership-building and coordinated research action.

Concordance analysis confirms and refines what was observed above, showing that resilience is strongly connected to disaster risk reduction, climate change, and governance frameworks. The selected KWIC lines bring together the main contexts in which *resilience* occurs in the subcorpus, particularly in relation to disaster risk reduction, climate adaptation, governance frameworks, and societal resilience (Figure 9). These patterns show that, in the *Key Messages*, earlier framings of resilience as a cultural and societal capacity are not replaced; rather, they are integrated with a more explicit vocabulary of climate governance, risk reduction, and research impact.

This reading also creates a link with the analysis of *climate* in the same subcorpus. In both cases, the *Key Messages* document develops a more specialised policy-oriented vocabulary than the previous subcorpora: *climate* is associated with neutrality, emissions, mitigation, and education, while *resilience* is associated with disaster risk reduction, climate adaptation, governance, and societal response.

In general, distant and close readings show that *resilience* is specified through different contextual and terminological uses, including environmental, societal, economic, and institutional resilience. The Word Rain points to a denser risk- and governance-oriented lexical environment around the term, while concordance reading shows continuity with the *Vision and Mission*. In particular, societal resilience remains connected to cultural heritage, well-being, and research impact, suggesting that *Key Messages* does not

replace the earlier framing of resilience as a societal and institutional capacity, but integrates it into a more explicit framework of climate governance and risk reduction.

5.2.4 WG Forms: Resilience as Participatory and Community-based Response

In the *WG Forms* Word Rain (Figure 2), *resilience* appears in connection with several patterns already observed above. As in the *Foresight Analysis*, degree-related forms such as *increasing* and *greater* appear near *resilience*, together with action-oriented forms such as *fostering* and *enhancing*. These items suggest continuity with the earlier representation of resilience as something that can be developed, strengthened, or increased.

At the same time, the *WG Forms* graph brings back the societal dimension already visible in *Key Messages*, through items such as *societal*, *societal changes*, and *societal challenges*. The proximity between *resilience* and vocabulary of change again confirms a recurring association across the corpus. In this subcorpus, however, change appears in more specific forms, including *changing times*, *environmental change*, and *societal changes*. This suggests that resilience is located at the intersection of environmental and societal transformation.

The Word Rain n-gram *visioning resilience* also appears in this graph, but it is positioned near items such as *vision words*, *suggested vision*, and *focus visioning*. This suggests that part of the resilience-related vocabulary may reflect the structure of the consultation and feedback process, rather than a fully developed thematic discussion of resilience. Concordance analysis is therefore needed to distinguish substantive patterns from procedural or formulaic language linked to the form-based nature of the document.

Concordance reading confirms and refines the visual interpretation, with *resilience* remaining associated with climate change, adaptation, and environmental transformation, as shown by expressions such as *resilience and adaptation*, *developing resilience through risk, mitigation and adaptation*, and *pathways to resilience and drive positive change*. This confirms continuity with the risk- and climate-governance vocabulary observed in *Key Messages*. However, the *WG Forms* also place resilience in a more collaborative and participatory context, where it is connected to communities, stakeholders, and co-creation processes.

This participatory framing is particularly visible in the following concordance line:

“Co-designed and co-created solutions can also enhance public awareness, strengthen mitigation and adaptation, and build **resilience**.”

Another concordance line links resilience to the broader social and cultural function of heritage:

“Heritage as a shared space for dialogue and mutual understanding, identity continuity and change, fostering inclusivity and **resilience**.”

These examples show that *resilience* is not framed only as a technical or policy objective, but also as a capacity connected to participation, dialogue, and collective action. As with *climate*, the *WG Forms* retain the vocabulary of adaptation and mitigation, but embed it in a more applied and participatory frame. Compared with *Key Messages*, where resilience is more explicitly connected to disaster risk, governance, and climate adaptation, the *WG Forms* present a more socially grounded framing, in which heritage, communities, and stakeholders are part of the process of building resilience.

5.3 Emerging Themes: Key Messages vs. WG Forms

In this section, two terms that were present only in the *WG Forms* graph, *gender* and *migration*, are explored in order to assess whether they point to under-explored themes in need of integration in the final SRIA. Given the low overall frequency of *gender* and *migration*, statistical collocation measures were not calculated for these items. Their analysis therefore relied on Word Rain patterns and concordance inspection, which were used to identify recurrent contextual and terminological patterns.

5.3.1 Gender: Gender Balance and Gender Perspectives

Gender is one of the emerging items identified through the *WG Forms* Word Rain. Unlike *climate* and *resilience*, it does not appear as a recurrent visual pattern across all subcorpora, but becomes visible in the *WG Forms* graph. It occurs both as a standalone item and as part of several Word Rain n-grams, including *terms gender*, *subject gender*, *show gender*, *projects gender*, *gender perspectives*, and *gender balance*.

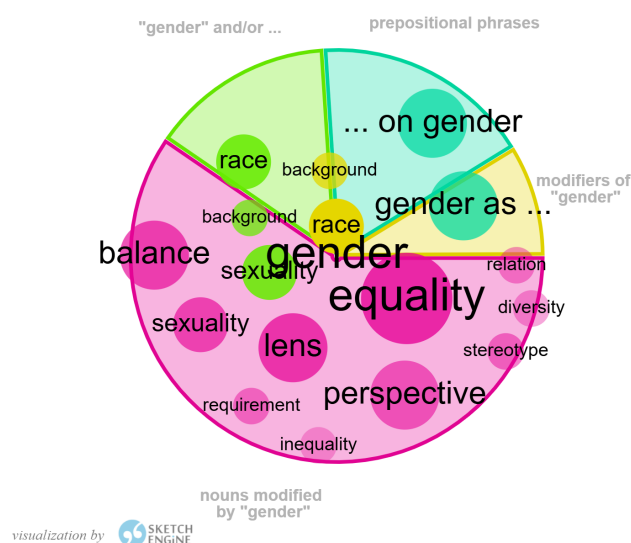


Fig. 10 Word Sketch for *gender* in the whole corpus.

This makes the item relevant not only because of its concentration in one graph, but also because the surrounding n-grams point to institutional and research-related uses of *gender*.

As reported in Table 2, *gender* occurs 25 times in the corpus (2.48 per 10,000 tokens), with 16 occurrences concentrated in the *WG Forms* (21.42 per 10,000 tokens). This distribution confirms that *gender* is not absent from the other subcorpora, but that it is much more concentrated in the *WG Forms*. For this reason, concordance analysis was necessary to determine which contextual and terminological uses were associated with the item.

The Word Sketch for *gender* (Figure 10) shows that the item is mainly used as a modifier in multiword expressions such as *gender balance*, *gender perspective*, *gender lens*, *gender stereotype*, and *gender equality*. These patterns suggest that *gender* is not treated as a broad standalone topic, but is specified through recurrent terminological combinations. Two main uses emerge from the concordance lines. The first relates to institutional and organisational representation, especially through expressions such as *gender balance* and *gender equality*. The second relates to research and analytical orientation, through expressions such as *gender perspective*, *gender perspectives*, and *gender lens*.

Concordance reading confirms this distinction. Two lines are particularly relevant:

“The results also show that **gender** equality has been mainly addressed in terms of **gender** balance in the governance structures of the initiatives, the evaluation committees, and the funded research teams.”

“However, there are hardly any funded projects on **gender** as a research subject and **gender** perspectives have rarely been adopted.”

In the *WG Forms*, *gender equality* is mainly discussed in relation to *gender balance* in governance structures, evaluation committees, and funded research teams. At the same time, expressions such as *gender perspectives* and *gender as a research subject* indicate a second use of the item, connected to the integration of gender as an analytical dimension in Cultural Heritage research. Close reading therefore refines the Word Rain output by showing that the relevance of *gender* lies less in its general thematic presence than in the specific terminological patterns through which it is used.

The occurrences in the other documents are fewer and less central. Outside the *WG Forms*, *gender* occurs 9 times, six of which are in the *Foresight Analysis*. In these cases, it is generally not treated as a separate theme, but as one dimension within broader discussions of inclusivity, inequality, social justice, and heritage policy, as shown in Figure 11. For instance, the *Foresight Analysis* refers to possible inequalities in “race, gender, sexuality and class”, and to the rootedness of heritage narratives in “ethnicity, nation,

Details	Left context	KWIC	Right context
1	irly relevant should we see exacerbated inequalities in race,	gender	, sexuality and class.</s><s>Further manifestations could p
2	.</s><s>'Rising inequalities among various groupings (race,	gender	, sexuality, class) and multidimensional poverty (wealth, hea
3	mate crisis and ongoing patterns of inequity via colonialism,	gender	inequality, LGBTQIA+ suppression and industrialisation (Sh
4	ourses of ethnicity, nation, centre–periphery, class divisions,	gender	relations and migrations, among others...</s><s>[to] enable
5	above, applying heritage towards greater social justice and	gender	equality.</s><s>As facilitators, the heritage sector can creat
6	's><s>A selection was then made to ensure geographic and	gender	diversity.</s><s>Reviewers were asked three questions: ••

Fig. 11 KWIC concordance lines for *gender* in the *Foresight Analysis* subcorpus.

centre–periphery, class divisions, gender relations and migrations”. In another passage, *gender equality* appears as part of a wider human rights-based approach to heritage and social justice.

The *Vision and Mission* document also mentions gender equality, but mainly in relation to the European Commission’s gender equality requirements and their integration into the SRIA process. This reinforces the pattern identified in the *WG Forms*: *gender* is associated both with institutional equality requirements and with research-oriented expressions such as *gender perspectives*.

Overall, *gender* shows how Word Rain can guide close reading beyond the most frequent lexical items. In a frequency list, the item might have seemed marginal because of its low overall frequency. However, its concentration in the *WG Forms* and its appearance in specific n-grams made it visible as a potentially relevant emerging theme. Concordance analysis then clarified the nature of this theme: not *gender* in general, but the recurrent use of expressions such as *gender balance*, *gender equality*, *gender perspectives*, and *gender lens*. In this sense, what is observed for *gender* also connects with the broader pattern observed in the *WG Forms*, where social, participatory, and equality-related vocabulary becomes more visible than in the other subcorpora.

5.3.2 Migration: Demographic Change and Terminological Clarification

Similarly to *gender*, *migration* becomes visually salient only in the *WG Forms* Word Rain. In this graph, *migration* is located in the upper part of the second region, near *mitigation*, *sustainable*, and *society*. Its proximity to *climate change* suggests a possible association between migration and climate or environmental change, but this visual indication requires concordance checking.

The visibility of *migration* in the subcorpus composed of inputs from the Cultural Heritage community makes it particularly relevant to analyse, since its distribution may help clarify how social and demographic change is addressed in the Cultural Heritage policy area. Other items in the same broader area, such as *inequalities*, *circular economy*, and *linkages*, also point to social and policy-oriented concerns in the *WG Forms*. This does not mean that such concerns are absent from the earlier subcorpora; for example, *inequalities* is also visible in the *Foresight Analysis*. Rather, the *WG Forms* appear to bring some of these concerns into a more explicit consultative and policy-oriented context.

A corpus query for *migration* in ARCHE-WD yielded 29 occurrences overall (2.87 per 10,000 tokens), distributed as follows: 15 in the *Foresight Analysis* (2.17 per 10,000 tokens), 5 in *Vision and Mission* (3.17 per 10,000 tokens), 2 in *Key Messages* (2.34 per 10,000 tokens), and 7 in the *WG Forms* (9.37 per 10,000 tokens). This distribution is relevant because it shows that *migration* is not exclusive to the *WG Forms*; however, it reaches its highest relative frequency there and becomes visible in the corresponding Word Rain output.

In the *WG Forms*, *migration* is explicitly mentioned as a concept requiring clearer definition within the SRIA, especially in relation to demographic change:

“The SRIA would benefit from a clearer distinction between **migration** and (other) demographic changes, and include a more balanced view of migration in terms of host country and those on the move.”

This concordance line is particularly important because it does not simply mention *migration* as a background social phenomenon. Rather, it identifies *migration* as a conceptual and terminological issue

requiring clarification in the agenda. In this sense, the *WG Forms* do not only introduce migration as a theme, but also point to the need for a more precise treatment of the term within the SRIA.

A possible connection between migration and climate-related change also appears in the *WG Forms*, although the evidence is limited and should not be over-interpreted:

“On the basis of CSA, the Time Machine Organisation was founded, which continues to implement the Time Machine Initiative including the changing relations between areas as a result of climatic changes and disasters, **migration** paths [...]”

This supports the initial Word Rain observation that *migration* may be connected to climate and environmental change, while also showing that this connection is not extensively developed in the *WG Forms*.

Although the *Foresight Analysis* has the lowest relative frequency of *migration*, it contains the highest number of raw occurrences. It is therefore useful to inspect its concordance lines, where the item is framed more broadly in relation to social change, instability, and the transformation of heritage meanings. For example, in the line below, migration is linked to climate crisis, war, and conflict:

“increased transnational **migration** – largely induced by the climate crisis, wars and conflict – is likely, creating progressively more heterogeneous societies”

This example connects with patterns already observed for *climate*. Here, migration is presented as partly induced by the climate crisis, wars, and conflict, showing how a lower-frequency item can complement the interpretation of more recurrent terms and themes.

Another line associates *migration* with population growth, diversity, and demographic transformation:

“population growth and **migration**, resulting in societies that are large and diverse”

Finally, the *Foresight Analysis* also connects *migration* to the changing spatial anchoring of heritage:

“As a result of widespread **migration** and displacement, heritage may become less geographically anchored and more dispersed and movable across borders”

Considered together, these concordances contextualise *migration* in relation to climate crisis, demographic transformation, and changing understandings of heritage. Close reading therefore refines the Word Rain output in two ways. First, it confirms that *migration* is relevant in the *WG Forms* as an emerging item requiring clearer terminological and policy treatment. Second, it shows that the broader corpus, especially the *Foresight Analysis*, already connects migration to climate crisis, social transformation, and the changing spatial meanings of heritage. Although the low overall frequency of *migration* might have limited its visibility in a frequency-based analysis, its appearance in the *WG Forms* Word Rain helped identify it as a theme requiring further contextual and terminological examination.

Overall, both *gender* and *migration* show that the *WG Forms* make visible social and policy-oriented concerns that are less prominent in the other Word Rain visualisations, while concordance analysis clarifies the specific terminological and contextual patterns associated with each item.

6 Conclusions and Future Steps

This experiment demonstrated how a combined distant and close reading approach can provide valuable insights for the analysis of a specialised corpus. Word Rain allowed for an intuitive and structured overview of the vocabulary, narratives and topics present across the documents, highlighting clusters of related lexical items and enabling comparison between aligned visualisations. The analysis of *climate* and *resilience* showed how recurrent items across all graphs may point to shared thematic and terminological areas, while also revealing different framings depending on the type and communicative function of each document.

Concordance analysis made it possible to confirm, refine, or qualify the initial hypotheses advanced through the interpretation of the Word Rain outputs. At the same time, Word Rain guided the selection of lexical items to be explored through concordance reading, especially with regard to micro-themes that might have been difficult to identify through frequency-based approaches alone. This was particularly visible in the analysis of *gender* and *migration*, whose concentration in the *WG Forms* allowed the comparison with *Key Messages* to identify themes requiring clearer integration or treatment in the SRIA process.

The terminological contribution of the study should therefore be understood in these methodological terms. Word Rain is not presented here as a substitute for automatic term extraction, glossary building, or terminological validation. Rather, when combined with close reading, it can support the preliminary exploration of specialised vocabulary by helping researchers identify salient lexical items, possible candidate terms, recurrent n-grams, and distributional patterns across subcorpora. Although the present experiment did not combine Word Rain with automatic term extraction, such an integration could be explored in future work. In this sense, the method could be used before, alongside, or after automatic term extraction: before extraction, to obtain an overview of the corpus and guide the selection of areas of interest; alongside extraction, to interpret candidate terms in relation to broader lexical clusters; and after extraction, to support the discussion and contextualisation of terminological patterns.

This is particularly relevant for collaborative and policy-oriented settings such as the one presented in this paper. Because Word Rain visualisations are intuitive and allow aligned comparison between documents, they can facilitate discussion with domain experts, making it easier to identify possible gaps and assess whether specific items require further terminological clarification. Moreover, functions such as the identification of new words are especially useful in this respect, since they help locate items that appear only in specific parts of the corpus and may therefore point to emerging or under-represented themes. A discussion of this collaborative dimension is beyond the scope of the present paper and is reported in [Squadrito et al. \(2025\)](#); however, it is worth noting here that this aspect was also tested in the ARCHE-WD workflow. Guided by CLARIN experts, the Word Rain outputs were presented to E-RIHS experts, who helped assess whether visually salient lexical items were relevant to the domain and, where necessary, moderate their interpretation. After this first stage of expert consultation, the visualisations were shared with the wider ARCHE consortium, where they served as an accessible basis for presenting corpus-based findings, involving experts in the interpretive process and collecting feedback from those involved in shaping the agenda.

Acknowledgements

This work was 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 also express their 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 study was made possible by the availability of ARCHE working documents and by the collaboration of experts from the Alliance, who assisted the authors during the initial inspection of the visualisations.

Authors Contribution

This contribution statement follows the CRediT taxonomy to clarify the different roles involved in the development of the study.

Vania Virgili: Resources, Project Administration. Vania Virgili contributed to the establishment of the initial collaboration between E-RIHS and CLARIN that enabled the study. She made the ARCHE data available for analysis and provided expertise on the research infrastructures and resources involved in the collaboration.

Francesca Frontini: Conceptualization, Methodology. Francesca Frontini contributed to the conceptualisation and methodological framing of the preliminary work on the same dataset. In particular, she proposed the adoption of a combined distant and close reading approach for the linguistic analysis of the corpus and contributed to the early methodological discussions that formed the basis of the present study.

Elisa Squadrito: Conceptualization, Methodology, Data Curation, Formal Analysis, Investigation, Visualization, Writing – Original Draft, Writing – Review & Editing. She participated in the conceptualisation of

the study and, for the present article, designed and implemented the analytical workflow and the methodological approach. She curated and processed the data, configured Word Rain and extracted relevant visual outputs, carried out the concordance-based analysis, interpreted distant and close reading results, and selected the examples and figures discussed in the article. She wrote the original draft and was responsible for the revision, expansion and finalisation of the manuscript.

References

- Ahlthrop, M., & Skeppstedt, M. (2025). Word rain as a service: Making semantically structured word clouds available to everyone. V. Vandeghinste & T. Kontino (Eds.), *Selected papers from the CLARIN Annual Conference 2024* (pp. 25–36). Linköping University Electronic Press. <https://doi.org/10.3384/ecp216.03>
- Barth, L., Kobourov, S.G., Pupyrev, S. (2014). Experimental comparison of semantic word clouds. J. Gudmundsson & J. Katajainen (Eds.), *Experimental Algorithms. SEA 2014*. (Vol. 8504, pp. 247–258). Springer International Publishing. https://doi.org/10.1007/978-3-319-07959-2_21
- Berson, I.R., & Berson, M.J. (2009). Making sense of social studies with visualization tools. *Social Education*, 73(3), 124–126, Retrieved from <https://www.socialstudies.org/making-sense-social-studies-visualization-tools>
- Boyles, N. (2012). Closing in on close reading. *Educational Leadership*, 70(4), 36–41, Retrieved from <https://api.semanticscholar.org/CorpusID:143325342>
- Burch, M., Lohmann, S., Beck, F., Rodriguez, N., Di Silvestro, L., Weiskopf, D. (2014). RadCloud: Visualizing multiple texts with merged word clouds. E. Banissi et al. (Eds.), *2014 18th international conference on information visualisation* (pp. 108–113). IEEE. <https://doi.org/10.1109/IV.2014.72>
- Cabré Castellví, M.T. (1999). *Terminology: Theory, Methods and Applications*. John Benjamins Publishing Company. <https://doi.org/10.1075/tlrp.1>
- Cano Díaz, E., Castillejo, M., Ramírez Barat, B., Sanz, M., Martín Gil, M., Bueso, M., ... Sarró, I. (2019). The European Research Infrastructure for Heritage Science (E-RIHS): An infrastructure for an interdisciplinary scientific domain. F. Pinto Puerto (Ed.), *TechnoHeritage 2019: IV International Congress Science and Technology for the Conservation of Cultural Heritage: Book of abstracts* (p. 56). Red de Ciencia y Tecnología para la Conservación del Patrimonio. Retrieved from <https://digital.csic.es/handle/10261/207299>
- Castella, Q., & Sutton, C. (2013). *Word Storms: Multiples of Word Clouds for Visual Comparison of Documents*. arXiv. <https://doi.org/10.48550/arXiv.1301.0503>
- Collins, C., Viegas, F.B., Wattenberg, M. (2009). Parallel Tag Clouds to explore and analyze faceted text corpora. J. Stasko & J.J. van Wijk (Eds.), *2009 IEEE Symposium on Visual Analytics Science and Technology* (pp. 91–98). IEEE. <https://doi.org/10.1109/VAST.2009.5333443>
- Cui, W., Wu, Y., Liu, S., Wei, F., Zhou, M.X., Qu, H. (2010). Context preserving dynamic word cloud visualization. S. North, H.-W. Shen, & J.J. van Wijk (Eds.), *2010 IEEE Pacific Visualization Symposium (PacificVis)* (pp. 121–128). IEEE. <https://doi.org/10.1109/PACIFICVIS.2010.5429600>
- de Jong, F., Van Uytvanck, D., Frontini, F., van den Bosch, A., Fišer, D., Witt, A. (2022). Language matters: The European research infrastructure CLARIN, Today and Tomorrow. D. Fišer & A. Witt (Eds.), *CLARIN: The Infrastructure for Language Resources* (pp. 31–58). De Gruyter. <https://doi.org/10.1515/9783110767377-002>
- Di Nunzio, G.M., Marchesin, S., Silvello, G. (2023). A systematic review of automatic term extraction: What happened in 2022? *Digital Scholarship in the Humanities*, 38(Supplement_1), i41–i47, <https://>

doi.org/10.1093/llc/fqad030

- Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 9(1), 99–115, <https://doi.org/10.1075/term.9.1.06dro>
- Firth, J.R. (1957). A synopsis of linguistic theory, 1930–1955. *Studies in Linguistic Analysis* (pp. 1–32). Basil Blackwell.
- Fišer, D., & Witt, A. (Eds.). (2022). *CLARIN: The infrastructure for language resources*. De Gruyter.
- Gottron, T. (2009). Document Word Clouds: Visualising web documents as tag clouds to aid users in relevance decisions. M. Agosti, J. Borbinha, S. Kapidakis, C. Papatheodorou, & G. Tsakonas (Eds.), *Research and Advanced Technology for Digital Libraries. ECDL 2009*. (pp. 94–105). Springer. https://doi.org/10.1007/978-3-642-04346-8_11
- Harris, Z.S. (1954). Distributional structure. *Word*, 10(2–3), 146–162, <https://doi.org/10.1080/00437956.1954.11659520>
- Heimerl, F., Lohmann, S., Lange, S., Ertl, T. (2014). Word Cloud Explorer: Text analytics based on Word Clouds. R. H. Sprague Jr. (Ed.), *Proceedings of the 2014 47th Hawaii International Conference on System Sciences* (pp. 1833–1842). IEEE. <https://doi.org/10.1109/HICSS.2014.231>
- International Organization for Standardization (2019). *Terminology work and terminology science — Vocabulary (ISO Standard No. 1087:2019)*. Retrieved from <https://www.iso.org/standard/62330.html>
- Jockers, M.L. (2013). *Macroanalysis: Digital Methods and Literary History* (1st ed.). University of Illinois Press. <https://doi.org/10.5406/illinois/9780252037528.001.0001>
- Jänicke, S., Franzini, G., Cheema, M.F., Scheuermann, G. (2015). On close and distant reading in digital humanities: A survey and future challenges. R. Borgo, F. Ganovelli, & I. Viola (Eds.), *Eurographics Conference on Visualization (EuroVis) - STARs* (pp. 83–103). The Eurographics Association. <https://doi.org/10.2312/eurovisstar.20151113>
- Jänicke, S., Franzini, G., Cheema, M.F., Scheuermann, G. (2017). Visual text analysis in digital humanities. *Computer Graphics Forum*, 36(6), 226–250, <https://doi.org/10.1111/cgf.12873>
- Kageura, K., & Umino, B. (1996). Methods of automatic term recognition: A review. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 3(2), 259–289, <https://doi.org/10.1075/term.3.2.03kag>
- Khusro, S., Jabeen, F., Khan, A. (2021). Tag Clouds: Past, present and future. *Proceedings of the National Academy of Sciences, India. Section A: Physical Sciences*, 91(2), 369–381, <https://doi.org/10.1007/s40010-018-0571-x>
- Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., ... Suchomel, V. (2014). The Sketch Engine. *Lexicography*, 1(1), 7–36, <https://doi.org/10.1007/s40607-014-0009-9>
- Le, Q.V., & Mikolov, T. (2014). Distributed representations of sentences and documents. E.P. Xing & T. Jebara (Eds.), *Proceedings of the 31st International Conference on Machine Learning* (Vol. 32, pp. 1188–1196). JMLR.org. <https://doi.org/10.48550/arXiv.1405.4053>
- Lee, B., Henry Riche, N., Karlson, A., Carpendale, S. (2010). SparkClouds: Visualizing trends in tag clouds. *IEEE Transactions on Visualization and Computer Graphics*, 16(6), 1182–1189, <https://doi.org/10.1109/TVCG.2010.194>
- Lohmann, S., Heimerl, F., Bopp, F., Burch, M., Ertl, T. (2015). Concentri Cloud: Word cloud visualization

- for multiple text documents. E. Banissi et al. (Eds.), *2015 19th International Conference on Information Visualisation* (pp. 114–120). IEEE. <https://doi.org/10.1109/iV.2015.30>
- McNaught, C., & Lam, P. (2010, May). Using Wordle as a supplementary research tool. *The Qualitative Report*, 15(3), 630–643, <https://doi.org/10.46743/2160-3715/2010.1167>
- Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv. <https://doi.org/10.48550/arXiv.1301.3781>
- Monachini, M., & Frontini, F. (2023). Infrastrutture digitali per le scienze umane e sociali. C. Fabio (Ed.), *Digital Humanities. Metodi, strumenti, saperi* (p. 197 - 213). Carocci Editore.
- Moretti, F. (2005). *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso.
- Moretti, F. (2013). *Distant Reading*. Verso Books.
- Pecman, M., & Kübler, N. (2022). Text genres and terminology. P. Faber & M.-C. L’Homme (Eds.), *Theoretical Perspectives on Terminology: Explaining Terms, Concepts and Specialized Knowledge* (Vol. 23, pp. 263–290). John Benjamins Publishing Company. <https://doi.org/10.1075/tlrp.23.12pec>
- Sager, J.C. (1990). *A Practical Course in Terminology Processing*. John Benjamins Publishing Company. <https://doi.org/10.1075/z.44>
- Sinclair, S., & Rockwell, G. (2016). *Voyant tools*. Retrieved from <https://voyant-tools.org/>
- Skeppstedt, M., & Aangenendt, G. (2024). Using the Word Rain technique to visualize longitudinal changes in periodicals from the Swedish Diabetes Association. T. Yousef & K. Al-Khatib (Eds.), *Vis4NLP 2024: Workshop on visualization for natural language processing*. The Eurographics Association. <https://doi.org/10.1177/14738716241236188>
- Skeppstedt, M., Aangenendt, G., Söderfeldt, Y. (2024). Visualizing longitudinal trends in digitized periodicals from the swedish diabetes association. *Huminfra conference (HiC 2024): Abstract submissions—talks* (p. 35). Retrieved from <https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-522107>
- Skeppstedt, M., Ahlthorp, M., Kucher, K., Lindström, M. (2024). From word clouds to Word Rain: Revisiting the classic word cloud to visualize climate change texts. *Information Visualization*, 23(3), 217–238, <https://doi.org/10.1177/14738716241236188>
- Spärck Jones, K. (2004). IDF term weighting and IR research lessons. *Journal of Documentation*, 60(5), 521–523, <https://doi.org/10.1108/00220410410560591>
- Squadrito, E. (2026). *ARCHE-WD Word Rain Supplementary Materials*. Zenodo. <https://doi.org/10.5281/zenodo.21245203>
- Squadrito, E., Frontini, F., Virgili, V., Monachini, M. (2025). Unveiling strategic research priorities: A terminological analysis of the ARCHE SRIA using word rain. F. Vezzani, G.M. Di Nunzio, E. Loupaki, G. Meditskos, & M. Papoutsoglou (Eds.), *Proceedings of the 4th International Conference on Multilingual Digital Terminology Today (MDTT 2025)*. CEUR-WS.org. Retrieved from <https://ceur-ws.org/Vol-3990/paper5.pdf>
- Striova, J., & Pezzati, L. (2017). The European Research Infrastructure for Heritage Science (ERIHS). *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2-W5, 661–664, <https://doi.org/10.5194/isprs-archives-XLII-2-W5-661-2017>
- Svensson, P. (2013). Humanities computing as digital humanities. M. Terras, J. Nyhan, & E. Vanhoutte (Eds.), *Defining digital humanities: A reader* (p. 159-186). Routledge. <https://doi.org/10.4324/9781315576251-10>

- Tribble, C. (2010). What are concordances and how are they used? A. O’Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 167–183). Routledge. <https://doi.org/10.4324/9780203856949-16>
- Underwood, T. (2019). *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press. <https://doi.org/10.7208/chicago/9780226612973.001.0001>
- van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86), 2579–2605, Retrieved from <https://www.jmlr.org/papers/v9/vandermaaten08a.html>
- Viegas, F., Wattenberg, M., Feinberg, J. (2009). Participatory visualization with Wordle. *IEEE Transactions on Visualization and Computer Graphics*, 15(6), 1137–1144, <https://doi.org/10.1109/TVCG.2009.171>
- Wu, Y., Provan, T., Wei, F., Liu, S., Ma, K.-L. (2011). Semantic-Preserving word clouds by Seam Carving. *Computer Graphics Forum*, 30(3), 741–750, <https://doi.org/10.1111/j.1467-8659.2011.01923.x>
- Wulff, S., & Baker, P. (2020). Analyzing Concordances. M. Paquot & S.T. Gries (Eds.), *A Practical Handbook of Corpus Linguistics* (pp. 161–179). Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_8
- Xu, J., Tao, Y., Lin, H. (2016). Semantic word cloud generation based on word embeddings. C. Hansen, I. Viola, & X. Yuan (Eds.), *2016 IEEE Pacific Visualization Symposium (PacificVis)* (pp. 239–243). IEEE. <https://doi.org/10.1109/PACIFICVIS.2016.7465278>
- Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. R. Witte et al. (Eds.), *Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks* (pp. 46–50). ELRA.