



Redesigning Equality and Scientific Excellence Together



Project Information

Topic:	SwafS-09-2018-2019-2020 Supporting research organisations to implement gender equality plans
Funding Scheme:	EU H2020 - Coordination and Support Action
GA Number:	101006560
Start date:	01/01/2021
Duration in months:	48
Project Coordinator:	UNIVERSITE DE BORDEAUX

RESET aims to address the challenge of Gender Equality in Research Institutions in a diversity perspective, with the objective to design and implement a user-centred, impact-driven and inclusive vision of scientific excellence.

Consortium partners





Redesigning
Equality and
Scientific
Excellence
Together

D3.1 Report on Qualitative Crowdsourced and Open Data Filtering Methodology



This project has received funding from the European Union's Horizon
2020 Framework Program for Research and Innovation under
Grant Agreement no 101006560.

UNIVERSITÉ
BORDEAUX

UNIVERSIDADE
DE ALCANTARA

UNIVERSITY OF
PORTO

UNIVERSITY
OF LIège

UNIVERSITY OF
DULWICH

UNIVERSITY OF
DULWICH

UNIVERSITY OF
DULWICH

UNIVERSITY OF
DULWICH

UNIVERSITY OF
DULWICH

Document Information

Title	Report on Qualitative Crowdsourced and Open Data Filtering Methodology
Deliverable No.	3.1
Version	1.0
Type	<input checked="" type="checkbox"/> Report <input type="checkbox"/> Demonstrator <input type="checkbox"/> ORDP <input type="checkbox"/> Ethics <input type="checkbox"/> Other
Work Package	3
Work Package Leader	Aristotle University of Thessaloniki
Issued by	Aristotle University of Thessaloniki
Issued date	20/12/2022
Due date	31/12/2022
Dissemination Level	<input checked="" type="checkbox"/> Public <input type="checkbox"/> Confidential <i>only for members of the consortium (including the EC)</i>

LEGAL NOTICE

The information and views set out in this report are those of the authors and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein.

Copyright

© Copyright 2021 The RESET Consortium

Consisting of:

UNIVERSITE DE BORDEAUX
ARISTOTELIO PANEPISTIMIO THESSALONIKIS
UNIwersytet LODZKI
UNIVERSIDADE DO PORTO
RUHR-UNIVERSITAET BOCHUM
OULUN YLIOPISTO
FONDATION NATIONALE DES SCIENCES POLITIQUES

This document may not be copied, reproduced, or modified in whole or in part for any purpose without written permission from the RESET Consortium. In addition, an acknowledgement of the authors of the document and all applicable portions of the copyright notice must be clearly referenced.

All rights reserved.

This document may change without notice.

Main Authors	
Name	Organization
Athena Vakali, Kelly Malerou, Areti Ampatzoglou, Stylianos Karamanidis	Aristotle University of Thessaloniki

Contributors and Quality Reviewers	
Name	Organization
Marion Paoletti, Maryna Radchuk, Ninon Junca	University of Bordeaux
Netta Livari	University of Oulu
Marisa Matias, Sara Magalhães and Jorge Peixoto Freitas	University of Porto

Abbreviations

GE	Gender Equality
GEP	Gender Equality Plan
HR	Human Resources
EC	European Commission
PSI	Public Sector Information
EU	European Union
EIGE	European Institute for Gender Equality
UIS	UNESCO Institute for Statistics
WDI	World Development Indicators
QS	Quacquarelli Symonds
R&I	Research and Innovation
ERA	European Research Area
STEM	Science, Technology, Engineering, Mathematics
AUTH	Aristotle University of Thessaloniki
UBx	University of Bordeaux
LU	University of Lodz
UPorto	University of Porto



Redesigning
Equality and
Scientific
Excellence
Together

D3.1 Report on Qualitative Crowdsourced and Open Data Filtering Methodology

API	Application Programming Interface
R&D	Research and Development



This project has received funding from the European Union's Horizon 2020 Framework Program for Research and Innovation under Grant Agreement no 101006560.



Executive Summary

Deliverable “D3.1 Report on Qualitative Crowdsourced and Open Data Filtering Methodology” is part of WP3 “Supporting data-driven GE and diversity policy-making in designing qualitative assessment tools and processes”. In particular, D3.1 is associated with the activities of collecting and analysing data, performed under Tasks 3.1 “GE data harvesting throughout RESET activities” and 3.2 “Text and statistical analysis to design GEPs”. The output of Tasks 3.1 and 3.2 has served as input for Task 3.3 “Establishing a strong GE repository and dashboard to support policy-making”, concerning the development of the RESET GE Awareness Platform. As described in detail in D3.2 “GE Data Collection and Processing Pipeline”, the platform consists of two interconnecting parts, namely, the RESET Dashboard and the RESET Forum.

In this context, the datasets selected during the processes of Tasks 3.1 and 3.2 contain both crowdsourced data generated through the RESET Forum activity and static data collected from institutional and open sources. In D3.1 we document the primary data sources as well as the filtering methodology of the data that are eventually visualised in the platform.

Table of contents

1.	Introduction	10
2.	RESET Data Source Types	11
2.1.	Open and Other Data Sources (Static Data)	11
2.1.1.	Open Data	11
2.1.2.	HR Data	12
2.2.	Discussion Forum (Dynamic Data)	14
3.	Dynamic Content Filtering	17
3.1.	RESET Dashboard: Dynamic Content Management	17
3.1.1.	Data Collection and Evaluation	17
3.1.2.	Data Analysis Methodologies	18
3.1.3.	Dashboard Statistics Outlines	26
3.1.4.	Future Actions & Maintenance	29
3.2.	RESET Forum: Dynamic Content Management	30
3.2.1.	Personalized Statistics Outlines	30
3.2.2.	Administrative Statistics Outlines	34
4.	Open Data Processing Pipeline	44
4.1.	Data Collection and Evaluation	44
4.2.	Data Pre-processing – Cleaning	47
4.3.	Data Reformatting	49
4.4.	Data Visualisations	51
5.	HR Data Processing Pipeline	55
5.1.	Data Collection	55
5.2.	Data Pre-processing – Filtering	56
5.3.	Data Analysis	57
5.4.	Data Visualisations	58
6.	Conclusions and Future Work	60
	References	61

List of Figures

Figure 1: Workshop Use Case Scenarios (December 2021)	13
Figure 2: Workshop Brainstorming Activity (December 2021)	14
Figure 3: Dynamic Data Overview	15
Figure 4: RESET forum registration process	19
Figure 5: RESET forum unsuccessful topic creation	20
Figure 6: RESET forum unsuccessful post creation	21
Figure 7: RESET forum registration institution field	21
Figure 8: RESET forum registration gender field	22
Figure 9: RESET forum word cloud	23
Figure 10: Single Post Sentiment abstract framework	24
Figure 11: Overall Sentiment abstract framework	26
Figure 12: RESET forum - Total Users, Posts & Topics	27
Figure 13: RESET forum - Active Users per Institution	27
Figure 14: RESET forum - Active Users per gender (%)	28
Figure 15: RESET forum - Overall Sentiment	28
Figure 16: RESET forum - Most used forum words	28
Figure 17: RESET forum - Most replied posts	29
Figure 18: RESET Forum - User Summary	30
Figure 19: RESET Forum - User Summary Generic Activity stats	31
Figure 20: RESET Forum - User Summary Top Replies & Top Topics	31
Figure 21: RESET Forum - User Summary Top Links	31
Figure 22: RESET Forum - User Summary Most replied to	32
Figure 23: RESET Forum - User Summary Most liked & Most liked by	32
Figure 24: RESET Forum - User Summary Top Categories	33
Figure 25: RESET Forum - User Summary Top Categories – Selection of particular category	33
Figure 26: RESET Forum - User Summary Top Badges	33
Figure 27: RESET Forum - Trigger User Summary for other users	34
Figure 28: RESET Forum - Admin option	34
Figure 29: RESET Forum - Admin dashboard	35
Figure 30: RESET Forum - General Admin statistics	36
Figure 31: RESET Forum - General Admin charts	38
Figure 32: RESET Forum - General Admin metrics	38
Figure 33: RESET Forum - Moderation Admin statistics	39
Figure 34: RESET Forum - Security Admin statistics	40
Figure 35: RESET Forum - Security Admin statistics	42
Figure 36: Open data pipeline	44
Figure 37: Data Filtering	47
Figure 38: QS Original source state snapshot	48
Figure 39: Country alpha2 and alpha3 codes	49
Figure 40: JSON file description	50
Figure 41: JSON simple example	51
Figure 42: Visualization of the Gender Equality Index	52
Figure 43: Visualization of the Gender Employment Gap	53

Figure 44: Percentage of women of boards in the participating countries	53
Figure 45: Men and Women in a typical academic career, students, and academic staff	54
Figure 46: HR Data Pipeline	55
Figure 47: HR Data Template (sex ratio per faculty)	56
Figure 48: HR Data Template (sex ratio per scientific field)	56
Figure 49: Part of Unified Spreadsheet with HR Data	57
Figure 50: Part of JSON File – AUTH HR Data	58
Figure 51: Percentages of male/female administrative and academic staff per faculty	59
Figure 52: Percentages of male/female academics per faculty and age group	59
Figure 53: Comparison of male/female academics percentage per faculty	59
Figure 54: Comparison of statistics at institutional level	60

1. Introduction

The design, implementation, monitoring, and upgrade of Gender Equality Plans (GEPs) in the context of the RESET project is being supported by the collection and analysis of relevant data throughout the project's life cycle. Different sets of static and dynamic data have been collected and are presented on the RESET platform – i.e., the targeted Gender Equality (GE) awareness platform – in order to support institutions' gender equality policy making [for more information on the platform see D3.2 - GE Data Collection and Processing Pipeline].

Selected datasets mainly consist of two types of data, namely (a) static data and (b) dynamic data. Static data refer to the data collected either online or from institutional sources and depict the current state of metrics and indices, both national and institutional, at the time of collection. These measures evolve over time and information may need to be updated periodically, e.g., annually. On the other hand, dynamic data are created from the use of the RESET Forum. RESET forum is an online discussion board dedicated to RESET, where questions and experiences can be expressed, discussions can be held and events can be shared [for more information on the RESET Forum, see D3.2 - GE Data Collection and Processing Pipeline]. Posts, comments, and users' participation feed datasets with information such as number of users in the forum, number of users per institution, or the most discussed topics, as well as with data on sentiment analysis of the posts. This type of data is continuously updated through the activities performed in the forum.

The aim of this document is to describe and evaluate the developed semantic analysis and overall data analysis methodology. Specifically, the data collection and processing pipeline, both for static and dynamic data used in the tasks of WP3 are presented. The rest of the deliverable is structured as follows:

- Chapter 2 presents the types of data sources used for static and dynamic data collection.
- Chapter 3 refers to the processes of data collection and filtering of dynamic data, which originates from the RESET Forum.
- Chapters 4 and 5 provide a presentation of the data collection and filtering procedures of open and HR data respectively.
- Chapter 6 concludes the document and describes future activities under the WP3.

2. RESET Data Source Types

In this section, we provide a detailed presentation of the sources used for primary data collection activities performed under the tasks of WP3, in order to feed the development of the databases for the Data Dashboard of the RESET platform. Our data consist of static data, collected from open and or other data sources presented in section 2.1, and dynamic data, collected through the activities performed in the RESET Forum, as presented in section 2.2.

2.1. Open and Other Data Sources (Static Data)

Static data represent a considerable part of the data collected and analysed in the RESET project, to the end of supporting the implementation and monitoring of the GEPs in corresponding partner institutions. Specifically, the static data collected originate from (a) open data sources and (b) the HR departments of the GEP implementing partners.

2.1.1. Open Data

Open data is data that is easily accessible and re-used without permission barriers (Murray-Rust, 2008). The European Commission (EC), with its digital strategy for open data (EC, 2022) strongly supports the opening of data in general and of Public Sector Information (PSI), in particular. Moreover, EC's open data policy is closely related to its open research data policy, as both engage with data that either is publicly funded or result from public funding. The EC, in its digital strategy for open data, clearly states that the open data creates extra value for both the economy and the society. Open data accessibility and reusability are supported by open data portals, which are web platforms that contain datasets metadata and facilitate the search for PSI (EC, 2022).

Under this perspective, open data represent an important part of the data collected, analysed, and presented in the RESET platform. Primarily, the most relevant sources of GE data through the reliable European Union (EU) portals were specified. For GE related data, the sources employed were SheFigures, European Institute for Gender Equality (EIGE), and the UNESCO Institute for Statistics (UIS) Women in Science database, while for more generic country-related statistical data, the datasets of Eurostat and the World Development Indicators (WDI) of the World Bank have been adopted. Finally, QS Quacquarelli Symonds was the main source used for university rankings.

While gender equality is one of the founding values of the European Union, EC's focus on gender equality in Research and Innovation (R&I) has been increasing since its 2012 Communication 'A Reinforced European Research Area Partnership for Excellence and Growth' laid out the current approach for realising a European Research Area (ERA) (EC, Directorate-General for Research and Innovation, 2021). In this context, She Figures is a EC periodical publication that presents comparable European and worldwide statistics on GE in research and innovation. Particularly, approximately 88 indicators depict information concerning women's participation in doctoral studies, labour market, acquisition of decision-making positions, and differences in women's and men's working

conditions and research and innovation output (EC, Directorate-General for Research and Innovation, 2021).

European Institute for Gender Equality is an independent service of the EU, aiming at promoting gender equality as well as gender mainstreaming in the policies of EU and the national policies of the Member States, by delivering high-level expertise to the European Commission, the European Parliament, the Member States and Enlargement countries. Moreover, EIGE is working towards raising citizens' awareness of gender equality¹. Among other projects, EIGE has developed the Gender Equality Index, as a tool to assess progress of gender equality in the European Union and the Member States (EIGE, 2022).

In the same context, UNESCO Institute for Statistics (UIS), has started to develop a set of indicators concerning the gender gap in research and academia, trying to identify the factors, such as family constraints or other social aspects that prevent women from following a career in the STEM field. These statistics are published in the Women in Science database, which also constitutes a main source of our data on GE issues (UIS, 2019).

Eurostat is the European office for statistics and its main responsibility is to provide the EU with reliable, comparable, high quality European statistics. Eurostat collaborates with national statistics offices of the member states' data, enabling comparisons between countries and regions, supporting decision-making and the design and monitoring of European policies (Eurostat, 2021).

The World Bank, a cooperation of 189 member countries, is acting towards sustainability goals to fight poverty in the developing world, provides access to free and open datasets on various topics, focusing on global development data. The World Development Indicators (WDI) of the World Bank have also served as a source for RESET datasets².

QS Quacquarelli Symonds is one of the leading providers of services and analytics concerning the global higher education sector. The QS World University Rankings portfolio, inaugurated in 2004, has grown to become the world's most popular source of comparative data about university performance and has been used as an input for university related information³.

2.1.2. HR Data

Since the RESET platform is dedicated to support decision-making in the context of implementing, monitoring, and redesigning the GEPs, the current state of GE issues should be depicted in the dashboard, so that stakeholders have an overall perspective of the existing situation in all GEP implementing institutions. As a matter of course, the

¹ <https://eige.europa.eu/>

² <https://databank.worldbank.org/source/world-development-indicators>

³ <https://www.topuniversities.com/about-qs>

respective data should be delivered by institutional sources. To this end, it was decided to collect GE related data from the HR departments of the corresponding partners.

Subsequently, AUTH organised two workshops with the consortium, in May 2021 and December 2021, in order to describe the data that would be important for designing and monitoring the GEPs and should be presented on the RESET platform. During the workshops, different use case scenarios for the RESET platform were discussed among the partners. All possible stakeholders were identified, along with their potential activities and the reason they could be using the platform, as depicted in the indicative examples of Figure 1.

RESET Platform Use Case Scenarios			
	Stakeholder	Action	Goal
	<i>Who is going to visit/use the platform</i>	<i>How they are going to use it</i>	<i>for which reason, what is their goal</i>
1	<i>I am responsible for policy making of a Research and Technology Organisation</i>	<i>and I would like more information</i>	<i>related to Gender Equality Plans.</i>
2	<i>I am a top-level manager at a University of interest/or GEB (Gender equality Board) member</i>	<i>and I would like to track the progress of the 1st version of the Gender Equality Plan imolementation in mv oraanization</i>	<i>to make changes towards the 2nd version.</i>
3	<i>I am a member in the community of practitioners/staff/professor/researcher</i>	<i>and I would like to discuss with other members of CoPs</i>	<i>topics related to the implementation of the Gender Impact Assessment.</i>
4	<i>I am a member in the community of practitioners/staff/professor/researcher</i>	<i>and I would like to have access to available data</i>	<i>related to the implementation of Gender Equality Plans.</i>
5	<i>I am a top-level manager at a University of interest/or GEB (Gender equality Board) member</i>	<i>and I would like to have access (and/or download) to GE related graphs and reports</i>	<i>to understand the current situation in my institution.</i>

Figure 1: Workshop Use Case Scenarios (December 2021)

Moreover, in a brainstorming activity, partners wrote down their suggestions on the indices they would expect to be supportive for GEP monitoring, implementation, and redesigning, as depicted in the indicative screenshot of Figure 2 below.

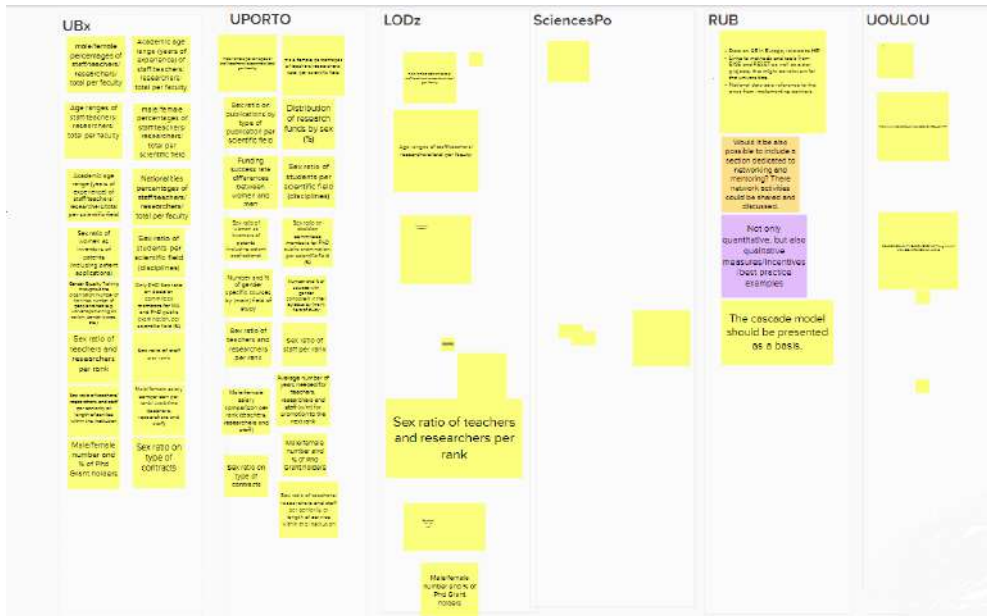


Figure 2: Workshop Brainstorming Activity (December 2021)

2.2. Discussion Forum (Dynamic Data)

As mentioned above, the dynamic data are produced by the activities taking place in the RESET Forum and are classified in four categories, namely: generic statistics, word cloud, sentiment analysis, and thematic analysis as presented in Figure 3 and analysed next.

Dynamic Data Overview

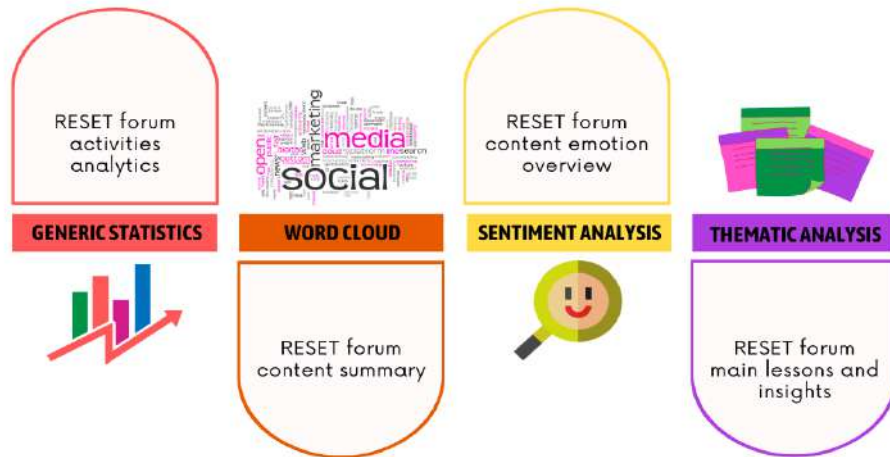


Figure 3: Dynamic Data Overview

- **Generic Statistics** describe specific metrics that are constantly produced by the current state of the RESET Forum. These statistics are automatically calculated through the RESET forum data provider and contain statistics such as:
 - Total active users
 - Total number of users per institution
 - Total number of users per gender
 - Total number of posts
 - Total number of topics
 - Total 100 topics with most replied posts

- The **Word Cloud** summarises the RESET forum textual content. It essentially depicts a representation of the mostly appeared words in the posts of the RESET community.

- **Sentiment Analysis** provides a deeper understanding of the emotions behind the texts. It is performed in the texts of main topics and in total to compute a sentiment score for each post, which gives a brief idea about the content emotion, and categorises the post as sentimentally positive, negative, or neutral.

- **Thematic Analysis** refers to a qualitative analysis that is additionally applied to the posts partners upload in the forum. RESET partners have been asked to upload descriptive posts, reporting their experiences and/or reflections with respect to a specific topic. Thus, a thematic analysis is applied to posts, in order to detect core themes and patterns within and across data collected from the forum. Apart from being representative of the forum content, these patterns and experience-based themes reflect the main lessons and insights partners gained through RESET, as well as potential needs and aspirations. They are complemented by the aforementioned word cloud and will be additionally exploited for the following: a) potentially upscaling some project tools based on partners' identified needs; b) providing content for some (policy) reports to be developed in T3.4 and reporting the RESET contribution to the ERA and EC relevant goals and priorities.

3. Dynamic Content Filtering

In this chapter, we provide a detailed presentation of the dynamic data collection and processing activities performed under the tasks of WP3, in order to feed the development of the databases for the Data Dashboard of the platform. Apart from the data represented in the dashboard, statistics offered straight by the RESET forum are also further analysed.

3.1. RESET Dashboard: Dynamic Content Management

As stated in the previous chapter, there are two types of primary data sources:

- static, to which open and HR data belong
- discussion forum, to which dynamic data belong

Dynamic data attract great interest since they are real-time generated. Moreover, they are directly correlated with the activities of the RESET members in the forum and contribute to the extraction of meaningful conclusions. In this section, all the selected metrics and the reasons for their selection are firstly explained in detail in 3.1.1. Afterwards in 3.1.2, the developed methodologies are examined and analysed. In subsection 3.1.3 the corresponding representations in the dashboard are presented, while 3.1.4 contains the following steps for the enrichment and maintenance of dynamic data.

3.1.1. Data Collection and Evaluation

Discussion forums are probably the earliest form of social media platform. From a data analysis point of view, discussion forums are truly significant since content created within these communities can be utilised to identify trends, sentiment, and other meaningful relations. Registration details, messages exchanges and general activities of the discussion forums members form a rich source of information for research.

Taking into consideration the scope and the intentions of the use of an online discussion forum for the RESET project, the AUTH team has developed a list of metrics to be monitored, presented it to the partners during consortium meetings, and gained their approval. Therefore, the following list of metrics was formed:

- **Total active users** → gain overall idea of the number of users registered
- **Total posts** → gain overall idea of the number of posts created
- **Total topics** → gain overall idea of the number of topics created
- **Total active users per institution** → gain knowledge of the distribution of the users across the partner universities
- **Total active users per gender** → gain knowledge of the distribution of the users across genders (male/female/other)
- **Overall forum sentiment** → gain overall idea of the sentiment developed in forum

- **Most used forum words – wordcloud** → gain overall idea of the most used words in the forum
- **Top 100 most replied posts list** → gain knowledge of the exact list of most replied posts
- **Top 100 most replied post sentiment score and sentiment category** → gain knowledge of the sentiment scores and sentiment category of most replied posts

For the collection of the RESET forum data, the [Discourse API](#) is used.

In general, API stands for Application Programming Interface, which is a set of definitions and protocols for building and integrating application software.

Here's how an API works: (<https://www.ibm.com/cloud/learn/api> - add source)

1. **A client application initiates an API call** to retrieve information—also known as a *request*.
2. **After receiving a valid request**, the API makes a call to the external program or web server.
3. **The server sends a response** to the API with the requested information.
4. **The API transfers the data** to the initial requesting application.

While the data transfer will differ depending on the web service being used, this process of requests and response all happens through an API.

In particular, for RESET, requests are initiated to the Discourse API (1) through the dashboard – e.g., get a list of users⁴. Discourse API receives the request and examines the validity (2). According to possible request parameters – e.g., gender or institution, data are filtered by Discourse API and are sent back as a response (3). Finally, a response is received from the dashboard and the data are shown accordingly (4).

3.1.2. Data Analysis Methodologies

Each metric in the list presented in 3.1.1 requires a different approach. This subsection encloses all the proper methodologies developed for the computation and analysis of the defined measures.

Total active users

It concerns the calculation of the active users of the forum. In order to include a member in the computation, two aspects are taken into consideration:

⁴ <https://docs.discourse.org/#tag/Users/operation/adminListUsers>

- the registration process (Figure 4) has been successfully completed and a forum admin has accepted the new member's request to register in the RESET forum
- the account is active - not deleted/suspended. If a user has an online behaviour that harms the community and quality of conversation after being warned, moderators have the option to suspend this user for a fixed period of time. Suspended users cannot log into the site and the reason for the suspension is displayed on their user page. These users are not calculated in the total active users' number.

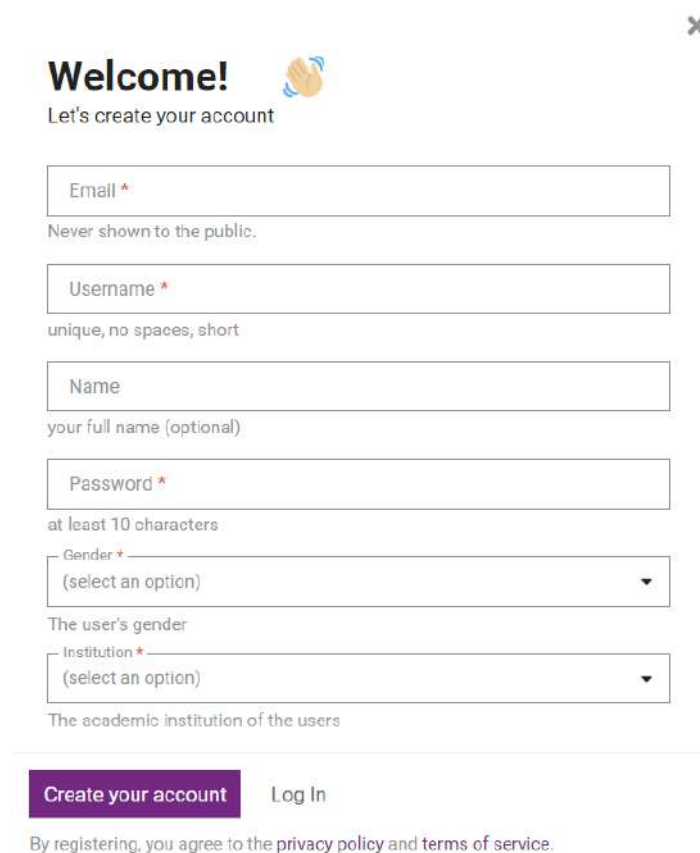


Figure 4: RESET forum registration process

Total topics

It concerns the calculation of the total number of topics created in the RESET forum. A topic is a collection of messages grouped together in a meaningful conversation, with a title, listed in a category, beginning with an original post. In order to include a topic in the computation of total topics, it is necessary that the user has successfully created the topic. For the successful creation of a topic, the following prerequisites must be fulfilled:

- Min content length: 20 characters

- Max content length: 32000 characters
- Min topic title length: 3 characters
- Max topic title length: 255 characters
- Max emojis in title: 1
- No exact duplication of title in another topic
- Max user mentions: 10 @name

Figure 5 shows an incomplete topic creation. If a user has successfully created a topic, they are instantly redirected to the newly created topic page.

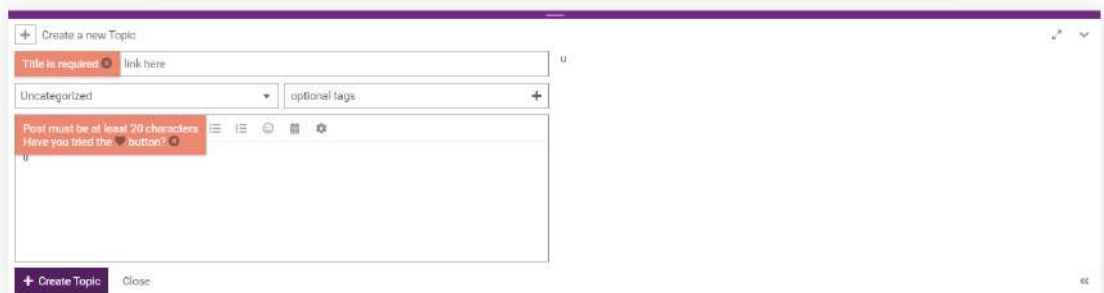


Figure 5: RESET forum unsuccessful topic creation

Total posts

It concerns the calculation of the total amount of posts created in the RESET forum. Each element of a topic is referred to as a post - can also be thought of as a comment. In order to include a post in the computation of total posts, it is necessary that the user has successfully created the post. For the successful creation of a post, the following prerequisites must be fulfilled:

- Min content length: 20 characters
- Max content length: 32000 characters
- Max user mentions: 10 @name

Figure 6 presents an unsuccessful post creation.

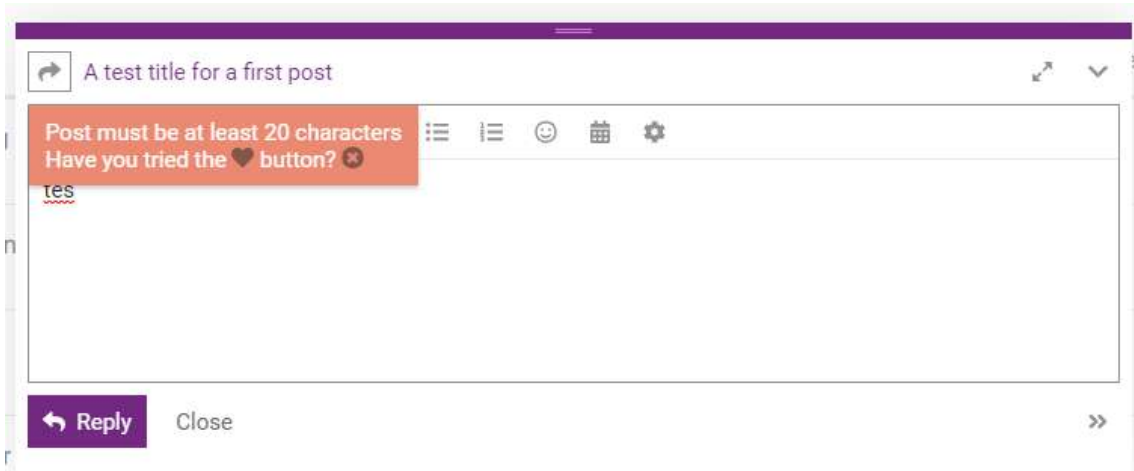


Figure 6: RESET forum unsuccessful post creation

Total active users per institution

It concerns the distribution of the active users of the forum by partner institutions. During the registration the user must fill in the institution, in order to sign up (a required field to complete the process), as shown in Figure 7.

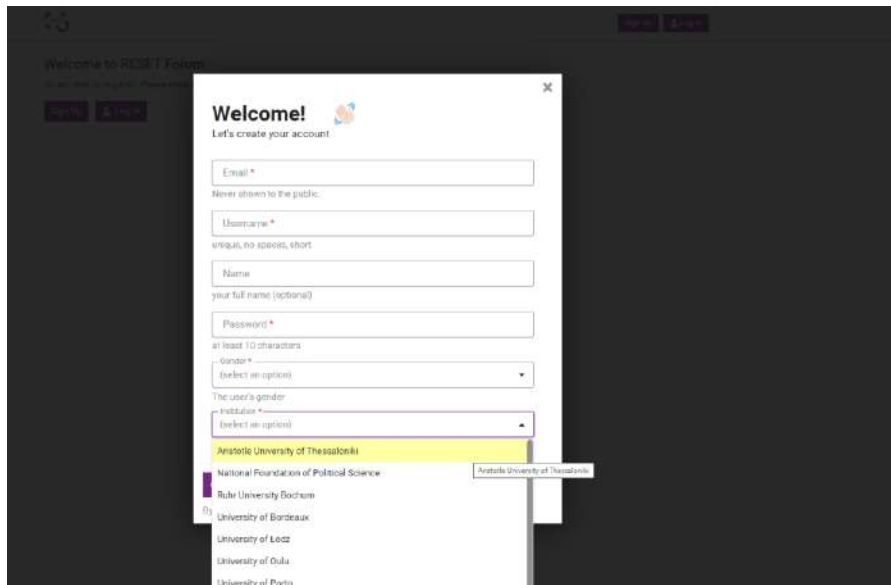


Figure 7: RESET forum registration institution field

The possible institution options are:

- Aristotle University of Thessaloniki
- National Foundation of Political Science
- Ruhr University Bochum

- University of Bordeaux
- University of Łódź
- University of Oulu
- University of Porto

Each time a registration is completed successfully, the corresponding institution counter is increased. Having calculated the total active users in general, and the total active users per institution, it is simple to calculate the users' distribution by institution in percentages.

Total active users per gender

It concerns the distribution of the active users of the forum by gender. During the registration the user must fill in the gender in order to sign up, since it is a required field to complete the process (Figure 8).

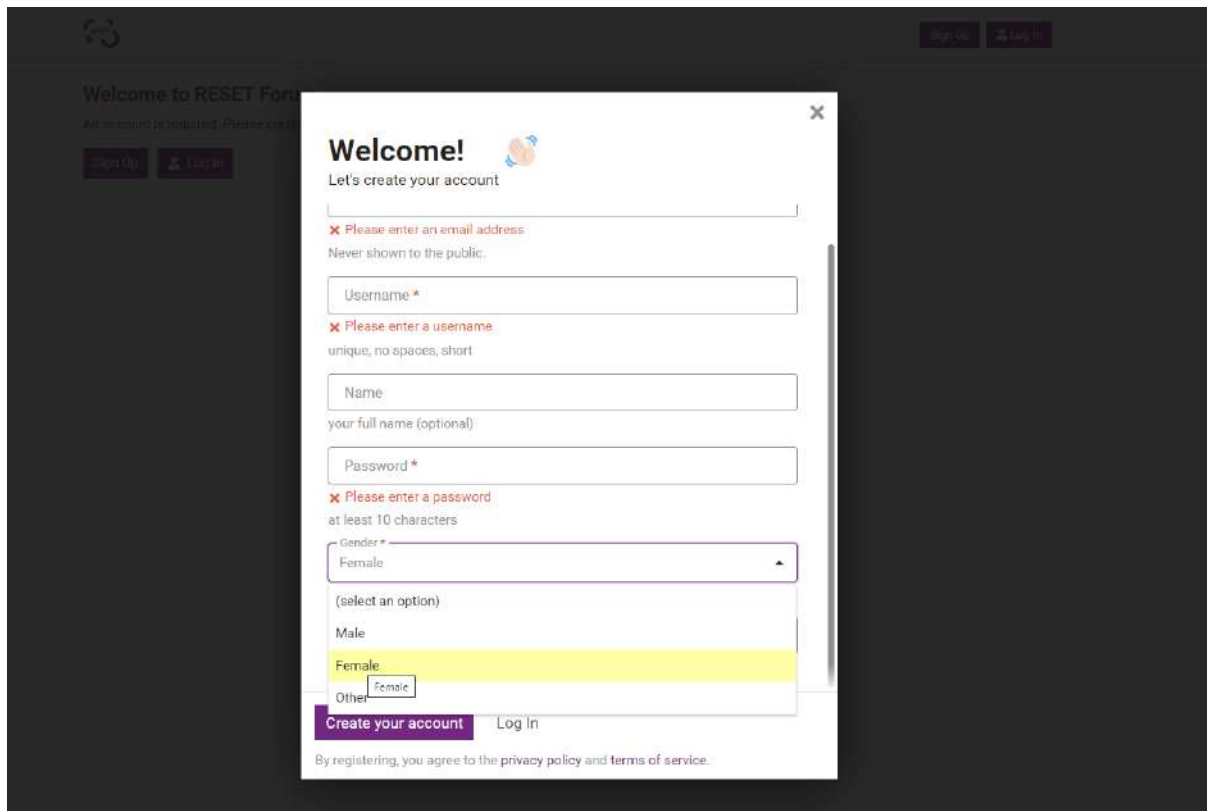


Figure 8: RESET forum registration gender field

The possible gender options are:

- Male
- Female

- Other

Each time a registration is completed successfully, the corresponding gender counter is increased. Having calculated the total active users and the total active users per gender, it is simple to calculate the users' distribution by gender in percentages.

Most used forum words – Word Cloud

It concerns the calculation and the distribution of the most used forum words. A word cloud has been chosen for this representation: a word cloud is a collection, or cluster of words depicted in different sizes. The bigger and bolder the word appears, the more often it is mentioned within a given text and the more important it is. For the RESET forum word cloud, the system includes the calculation words from all posts, including those in PMs. The size of the word list is 300 words, while the minimum length for a word to be included for calculations is 3 characters.

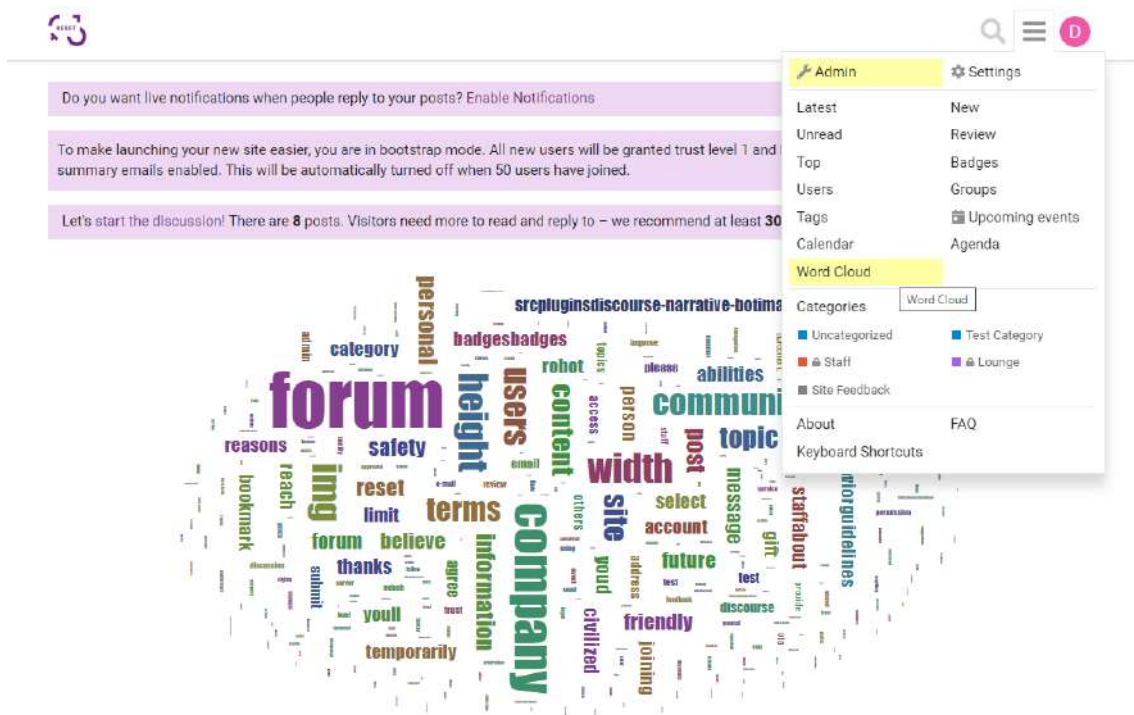


Figure 9: RESET forum word cloud

Top 100 most replied topics

It concerns the retaining of a list with the top 100 most replied topics. As mentioned before, posts in topics can be thought of as comments. It is sensible to maintain this list for the ease of identifying the topics which mostly awaken the interest in the RESET forum. Each time a new post is created under a topic, the corresponding replies count for the topic is increased and, if needed, the order in the list is changed accordingly.

Top 100 most replied topics sentiment score and sentiment category

Concerns the calculation of a single post sentiment score along with its sentiment category. An abstract framework of the calculation process is shown below:

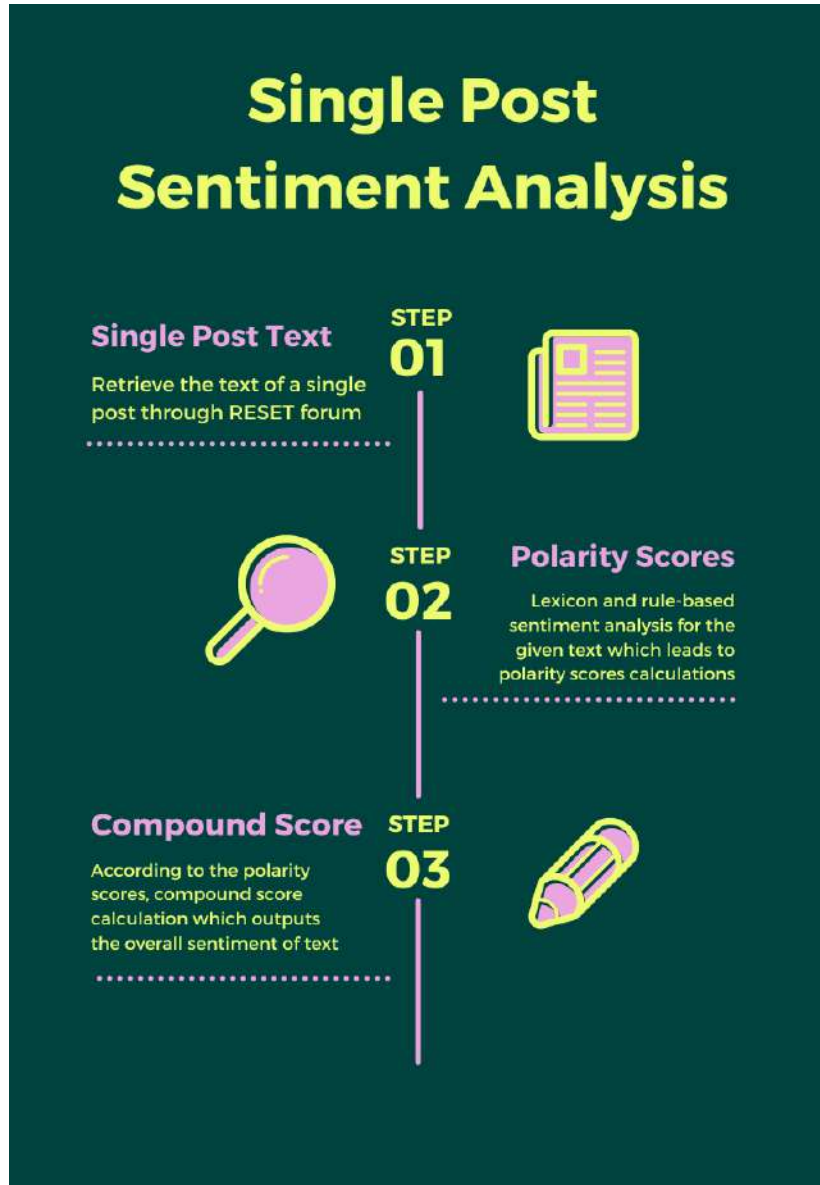


Figure 10: Single Post Sentiment abstract framework

As shown in Figure 10 above, the calculation is basically divided into three steps and it is implemented for the 100 most replied posts:

- *Single Post Text*: first, using the [Discourse API](#) and the topic id we retrieve the text of the post

- *Polarity Scores*: the next step is about the calculation of the polarity scores. The calculation is implemented with the usage of VADER, a popular library for social media sentiment analysis in the programming language Python.
- *Compound Score*: the last step is to determine the degree of the sentiment with the calculation of the compound score.

More information on VADER and steps 2 and 3 can be found below.

About VADER (Keita, 2022)

Valence Aware Dictionary and sEntiment Reasoner or VADER for short is a lexicon and simple rule-based model for sentiment analysis. It can efficiently handle vocabularies, abbreviations, capitalizations, repeated punctuations, emoticons (😄, 😊, 😞, etc.), usually adopted on social media platforms to express one's sentiment, which makes it a great fit for social media sentiment text analysis.

VADER has the advantage of assessing the sentiment of any given text without the need for previous training, as often needed for Machine Learning models.

Given a text, the result generated by VADER is an object/dictionary composed of 4 keys *neg*, *neu*, *pos* and *compound*:

- *neg*, *neu*, and *pos* mean negative, neutral, and positive respectively. Their sum should be equal to 1 or close to it with float operation.
- *compound* corresponds to the sum of the valence score of each word in the lexicon and determines the degree of the sentiment rather than the actual value as opposed to the previous ones. Its value is between -1 (most extreme negative sentiment) and +1 (most extreme positive sentiment).

Using the compound score can be enough to determine the underlying sentiment of a text, because for:

- a positive sentiment ($\text{compound} \geq 0.05$)
- a negative sentiment ($\text{compound} \leq -0.05$)
- a neutral sentiment (compound is between -0.05, and 0.05).

Overall Forum Sentiment

Concerns the calculation of the overall forum sentiment using the single posts' sentiment category.

In particular, having completed the single post process for every topic and identification of the corresponding category, as described in the previous subsection, the calculation of the overall forum sentiment is a simple procedure. Three types of counters are needed:

- *Positives Counter*: number of topics in category “Positive”
- *Negatives counter*: number of topics in category “Negative”
- *Neutrals Counter*: number of topics in category “Neutral”

By dividing each counter value with the total number of topics and by multiplying with 100%, one can get the overall sentiment percentages.

An abstract framework of the calculation process is shown below:

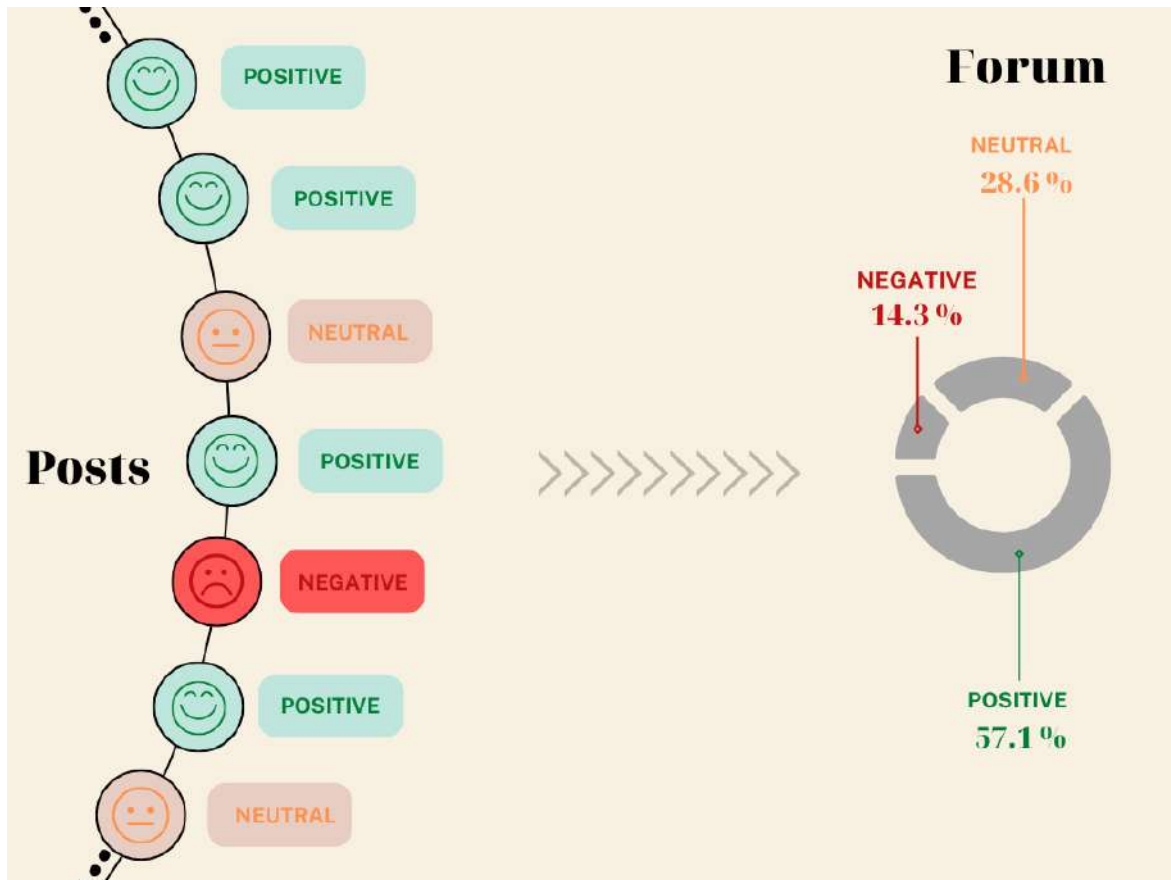


Figure 11: Overall Sentiment abstract framework

3.1.3. Dashboard Statistics Outlines

The outlines of the statistics described on the previous subsection can be found in the *Forum Stats* page of the Dashboard - <https://toolkit.wereset.eu/#/chart/12>.

Total Users, Posts & Topics

The total amounts of users, posts and topics are depicted in three different rectangles:



Figure 12: RESET forum - Total Users, Posts & Topics

Active Users per Institution (%)

The active users per institution are shown in the form of a vertical bar chart. Each bar represents an institution, while the vertical axis represents the values of active users percentages. On hover on each bar, one can check at a glance the exact value of active users % for the corresponding institution.

Active users per institution (%)

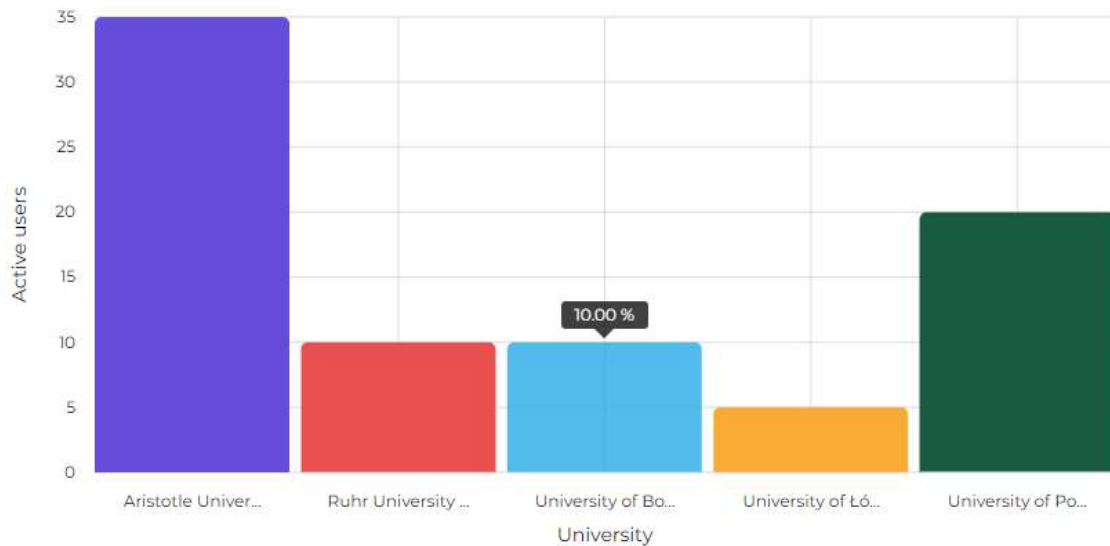


Figure 13: RESET forum - Active Users per Institution

Active Users per Gender (%)

The active users per gender are shown in the form of a pie chart. Each separate piece in the pie represents a gender. On hover on each separate piece, one can check at a glance the exact value of active users % for the corresponding gender.

Active users per gender(%)

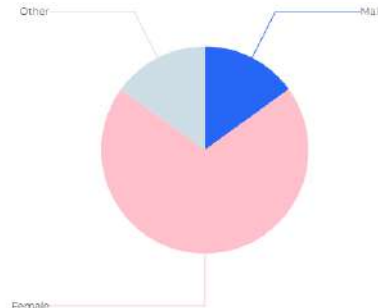


Figure 14: RESET forum - Active Users per gender (%)

Overall Sentiment

The overall forum sentiment is depicted in three different rectangles:

- Positive sentiment, with text colour in a green shade
- Negative sentiment, with text colour in a red shade
- Neutral sentiment, with text colour in a yellow shade

Overall Sentiment



Figure 15: RESET forum - Overall Sentiment

Most used forum words

The most used forum words are shown in the form of a word cloud. All values have the same colour. However, the more a word is used, the more the font size increases.

Most used forum words



Figure 16: RESET forum - Most used forum words

Most replied posts

The most replied posts are displayed within a table. The table consists of the following columns: Post ID, Post Title, Number of replies in the post, Link to the post in RESET forum, Post Sentiment Category, Post Analytical Sentiment Score. Moreover, a user can filter the table, change the items per page selection and navigate through the table's pages.

Most replied posts

ID	Title	Number of posts ↓	Link	Sentiment	Sentiment Score
37	A test title for a first post	3	visit	Positive	79% Positive, 2% Negative, 19% Neutral
46	Can I attach the files here? for example the pdf files?	3	visit	Positive	100% Positive, 0% Negative, 0% Neutral
11	This is the latest topic in the forum	2	visit	Positive	97% Positive, 1% Negative, 2% Neutral
13	Another topic here!	2	visit	Positive	45% Positive, 28% Negative, 27% Neutral
31	Why should title be at least 15 characters?	2	visit	Positive	48% Positive, 11% Negative, 40% Neutral
35	This is a test post for the new category	2	visit	Negative	18% Positive, 61% Negative, 21% Neutral
12	A topic with calendar	1	visit	Neutral	25% Positive, 25% Negative, 50% Neutral
24	Discourse RESET test topic	1	visit	Positive	58% Positive, 13% Negative, 31% Neutral
28	Test event for forum	1	visit	Neutral	6% Positive, 5% Negative, 85% Neutral
34	About the Test Category category	1	visit	Neutral	24% Positive, 22% Negative, 54% Neutral

Items per page: 10 1 - 10 of 21 < >

Figure 17: RESET forum - Most replied posts

3.1.4. Future Actions & Maintenance

Apart from demonstrating the above measures, it is important to maintain and implement future actions for dynamic content management.

Evolution of the statistics in time: It will be of great interest to keep track of the statistics and discover their evolution in time. Depending on this development, it is simple to identify the impact of the forum from its start till the end of the project.

Thematic analysis on forum posts: As previously mentioned, a thematic analysis will be applied to the forum posts in order to detect core patterns with respect to the experiences and reflections of the RESET community. All posts from the forum will be extracted to the qualitative software NVivo, so as to be afterwards analysed using an inductive approach.

Potential update on the forum measures: Given the fact that the traffic along with the number of online activities to the RESET forum will be increased, it is essential to examine the idea of including additional measures and statistics. Further data analysis should be up for consideration accompanied by the corresponding appropriate representations.

3.2. RESET Forum: Dynamic Content Management

The RESET forum is developed using [Discourse](#), an open-source discussion forum. Discourse by itself is a very powerful tool and among various features, it offers statistics for every member and additional analytics for each admin.

3.2.1. Personalized Statistics Outlines

The RESET forum provides several statistics to each member according to their activities on the forum. The statistics can be found in the toolbar user icon menu (triggered by clicking on user avatar/image), under the preferences (👤 icon), at the option **Summary**, as shown in Figure 18, below.

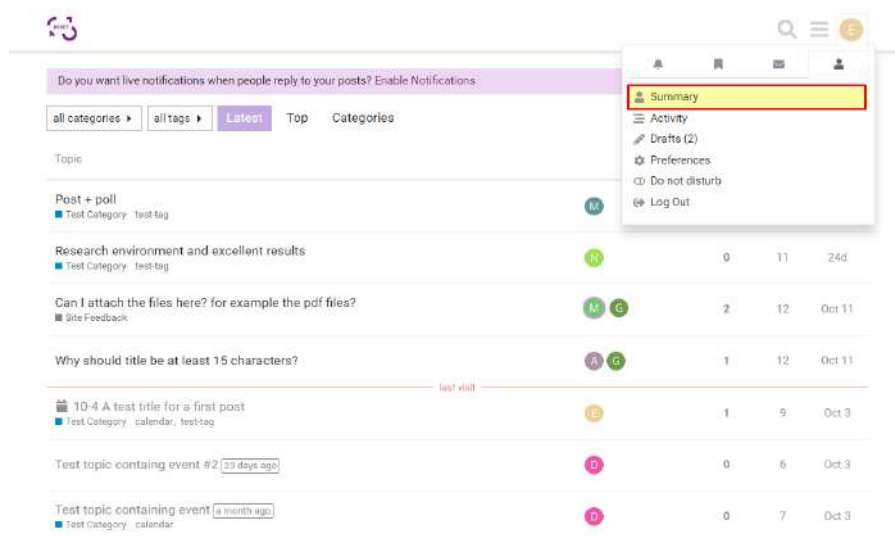


Figure 18: RESET Forum - User Summary

At the top of the Summary page, a list of generic activity statistics is presented (Figure 19), which in particular are:

- Total days the user visited the RESET forum
- Total minutes of reading time
- Recent minutes of reading time – last 60 days
- Total number of topics the user viewed
- Total number of posts the user read
- Total number of likes the user gave
- Total number of likes the user received
- Total number of topics the user created

- Total number of posts the user created

STATS

18 days visited 8m read time 3m recent read time 11 topics viewed 16 posts read 0 ❤️ given 1 ❤️ received 1 topic created

1 post created

Figure 19: RESET Forum - User Summary Generic Activity stats

Following the generic activity stats, three topic-related statistics are presented to the user:

- Top replies: a list of the topics, to which the user replied the most
- Top topics: a list of the most active topics where the user was the original poster
- Top links: a list of the most active topics where the user was the original poster and has shared a link



Figure 20: RESET Forum - User Summary Top Replies & Top Topics



Figure 21: RESET Forum - User Summary Top Links

Apart from the above, three more statistics are presented, which derive from the activity related to other users, such as the number of replies a user gives or the number of likes a user gives or receives, as follows:

- Most replied to: a list of the RESET forum members to whom the user gave more replies

- Most liked by: a list of the RESET forum members from whom the user received the most of likes
- Most liked: a list of the RESET forum members to whom the user gave the most of likes



Figure 22: RESET Forum - User Summary Most replied to



Figure 23: RESET Forum - User Summary Most liked & Most liked by

Finally, two statistics related to categories and badges are offered as well:

- Top categories: a table with the top categories, created according to the number of times each category is used in a topic or reply (Figure 24). By clicking on each table row, the user is navigated to a page showing the exact list of topics/replies where the particular category is used (Figure 25).
- Top badges: a list of the top badges the user has been granted. Each user can earn a badge related to their activity on the forum, e.g., first emoji badge, the first time they add an emoji to their post.

TOP CATEGORIES

	Topics	Replies
■ Uncategorized	3	1
■ 🗨 Lounge	2	1
■ Test Category	1	–

Figure 24: RESET Forum - User Summary Top Categories

2 results for @dev1fs #lounge

@dev1fs #lounge Topics/posts Search

Advanced filters

Sort by Relevance

- Discourse RESET test topic
🗨 Lounge calendar
Jun 23 - Test Discourse RESET test post
- This is the latest topic in the forum
🗨 Lounge calendar
Mar 1 - Some details for this!

No more results found.

Figure 25: RESET Forum - User Summary Top Categories – Selection of particular category

TOP BADGES






 <p>Basic Granted all essential community functions</p>	 <p>First Emoji Used an Emoji in a Post</p>	 <p>First Like Liked a post</p>
 <p>Editor First post edit</p>	 <p>Welcome Received a like</p>	

Figure 26: RESET Forum - User Summary Top Badges

It must be highlighted that the list of the above statistics for each member is **public**, i.e., every member has access to the statistics summary of each RESET forum member. Statistics summary of other RESET forum members can be triggered by clicking on the avatar of the specific member:

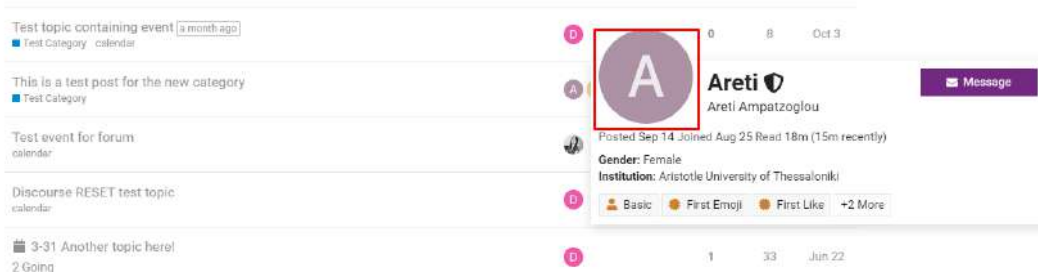



Figure 27: RESET Forum - Trigger User Summary for other users

3.2.2. Administrative Statistics Outlines

Admins have a crucial role for the RESET forum and therefore they are provided with a variety of additional statistics and reports. The further statistics for the admins can be accessed through the toolbar menu button, at the option  Admin (Figure 28) and then at the option **Dashboard** (Figure 29).

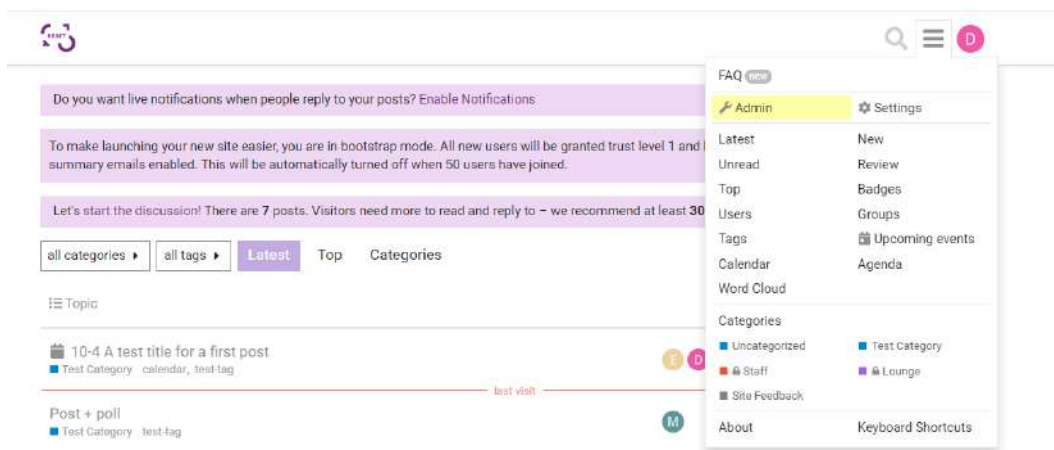


Figure 28: RESET Forum - Admin option

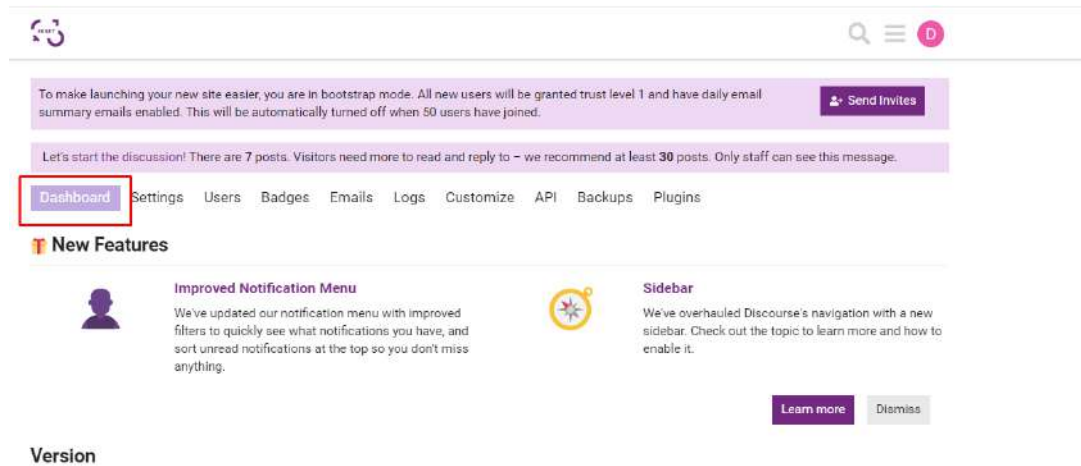


Figure 29: RESET Forum - Admin dashboard

At the Dashboard, four types of analytics are shown, namely, General, Moderation, Security and Reports, and are analysed next.

General Admin Statistics

An overview of the general admin statistics page is shown in Figure 30. The admins can select the period of their choice among last week, last month, last quarter and last year and the data presented get updated accordingly.

The exact list of charts of the general type is the following:

- Consolidated Pageviews: pageviews for logged-in users, anonymous users and crawlers in the form of a bar chart
- Signups: new account registrations for this period in the form of a line chart along with the absolute number
- Topics: new topics created during this period in the form of a line chart along with the absolute number
- Posts: new posts created during this period in the form of a line chart along with the absolute number
- DAU/MAU: number of members that logged in in the last day divided by the number of members that logged in in the last month – returns a %, which indicates community 'stickiness' (aim should be >30%) in the form of a line chart along with the calculated percentage for the selected period
- Daily Engaged Users: number of users that have liked or posted in the last day in the form of a line chart along with the absolute value for the selected period
- New Contributors: number of users who made their first post during this period in the form of a line chart along with the absolute number.

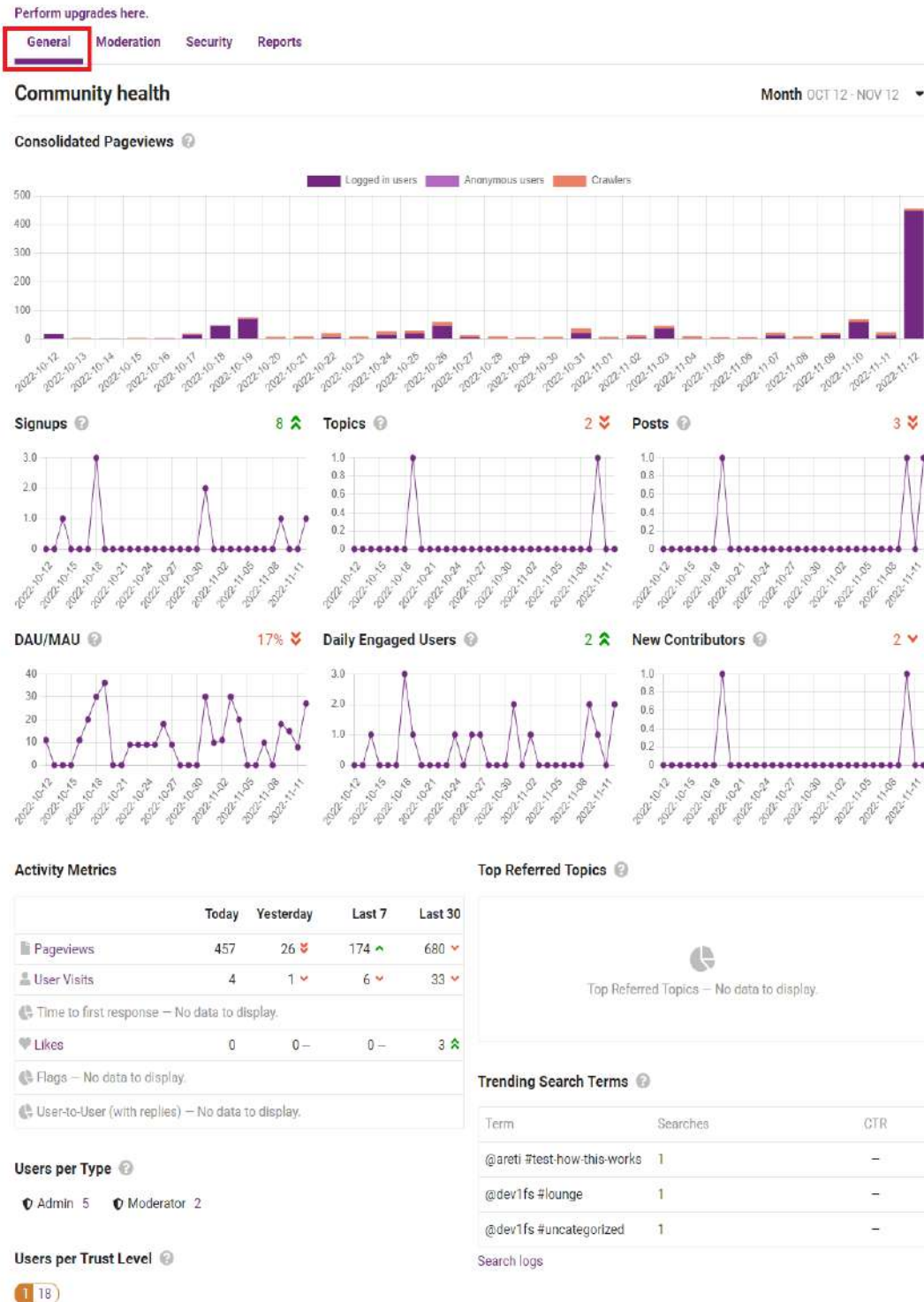


Figure 30: RESET Forum - General Admin statistics

The rest of the community health statistics, as presented in Figures 31 and 32, are the following:

- **Activity Metrics:** a list of metrics (e.g., pageviews, user visits, likes etc) in the form of a table for the current day, the previous day, the last 7 days, and the last 30 days
- **Top Referred Topics:** a list of topics that have received the most clicks from external sources in the form of a table
- **Users per Type:** absolute values of number of users grouped by admin, moderator, suspended, and silenced
- **Users per Trust Level:** absolute values of number of users grouped by trust level, where trust level is earned through user's activity, reflects the trust of the community and the user's abilities to assist in governing their community.
- **Trending Search Terms:** most popular search terms with their click-through rates in the form of a table.

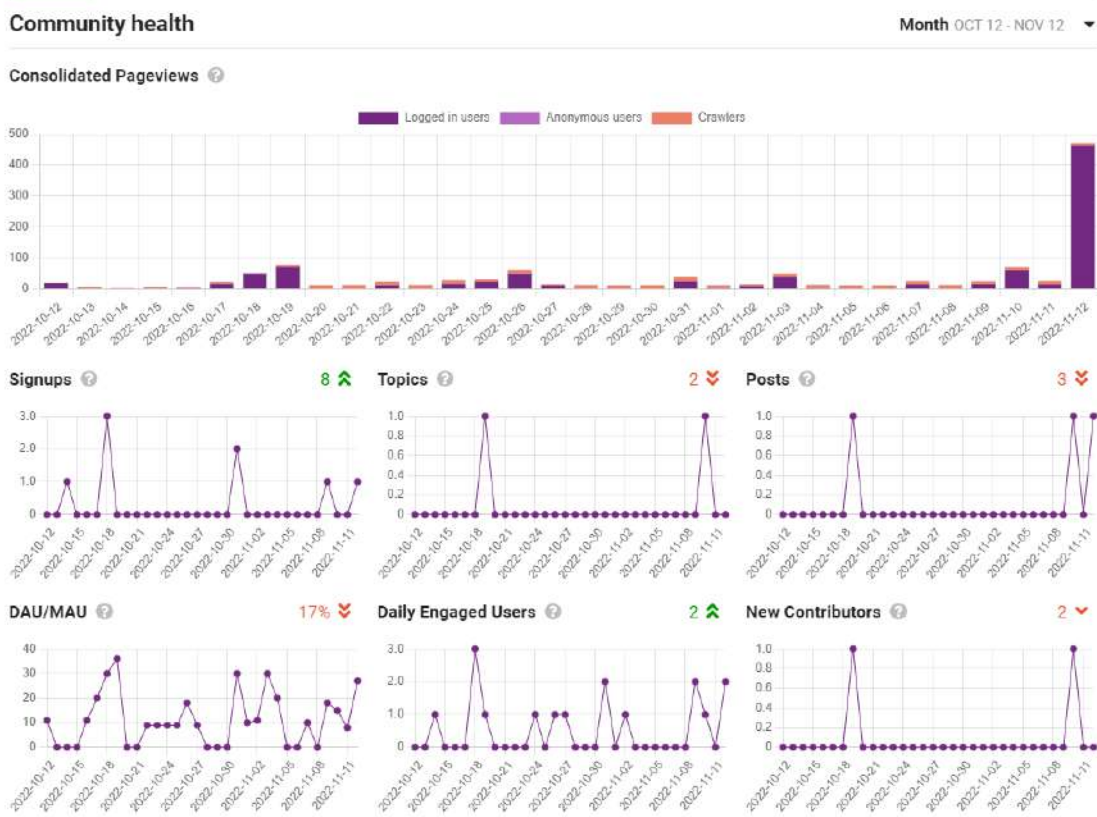


Figure 31: RESET Forum - General Admin charts

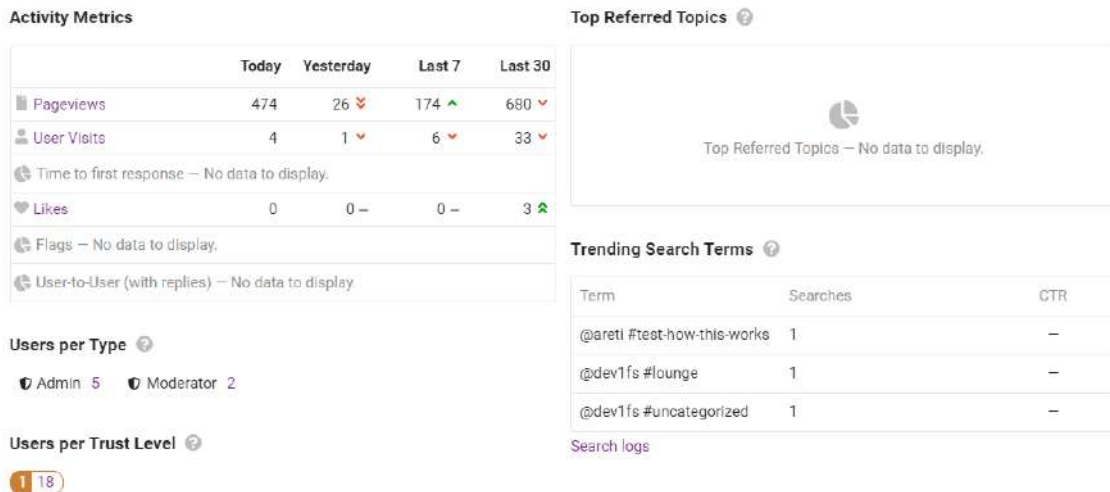


Figure 32: RESET Forum - General Admin metrics

Moderation Admin Statistics

An overview of the moderation admin statistics page is shown in Figure 33. The admins can select the period of their choice among last week, last month, last quarter and last year and the data presented get updated accordingly.

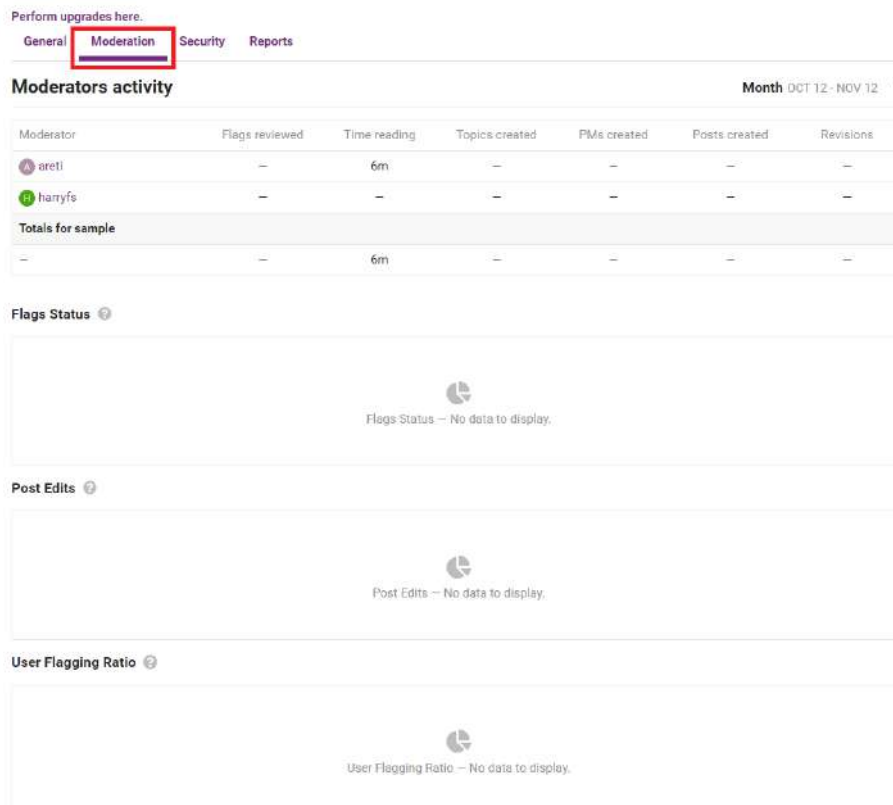


Figure 33: RESET Forum - Moderation Admin statistics

At the top, a table with the users belonging to the group of moderators is shown along with several information such as Topics and Posts created, Time reading, and PMs created for the selected period of time.

The rest of the analytics in the Moderation option are:

- **Flags Status:** a list of flags' statuses including type of flag, poster, flagger, and time to resolution
- **Post Edits:** number of new post edits
- **User Flagging Ratio:** list of users ordered by the ratio of staff response to their flags (disagreed to agreed).

Security Admin Statistics

An overview of the security admin statistics page is shown in Figure 34.

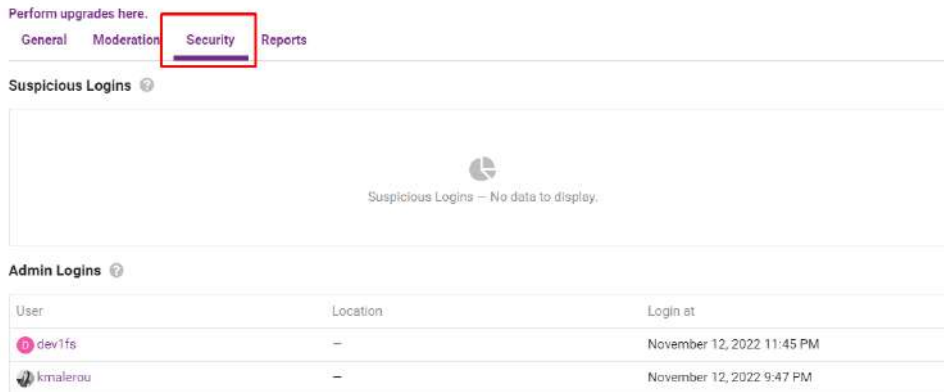


Figure 34: RESET Forum - Security Admin statistics

Two basic statistics are shown in the Security option:

- Suspicious Logins: a detailed list of new logins that differ suspiciously from previous logins
- Admin Logins: list of admin login times in the form of table.

Reports Admin Statistics

An overview of the reports admin statistics page is shown in Figure 35. When the admins select the report of their preference, they are navigated to a new page which displays the results. The results can be presented in the form of a table or a line chart. Moreover, a custom selection for the results period of time is available as well. Finally, the admins have the option to export the report data in the form of CSV.

A detailed list of all the reports is:

- Admin Logins: list of admin login times with locations
- Anonymous: number of new pageviews by visitors not logged in to an account
- Bookmarks: number of new topics and posts bookmarked
- Consolidated Pageviews: pageviews for logged-in users, anonymous users and crawlers
- DAU/MAU: number of members that logged in in the last day divided by number of members that logged in in the last month – returns a % which indicates community 'stickiness'
- Daily Engaged Users: number of users that have liked or posted in the last day
- Emails Sent: number of new emails sent
- Flags: number of new flags

- Flags Status: list of flags' statuses including type of flag, poster, flagger, and time to resolution
- Likes: number of new likes
- Logged In: number of new pageviews from logged in users
- Moderator Activity: list of moderator activity including flags reviewed, reading time, topics created, posts created, personal messages created, and revisions
- Moderator Warning: number of warnings sent by personal messages from moderators
- New Contributors: number of users, who made their first post during this period
- Notify Moderators: number of times moderators have been privately notified by a flag
- Notify User: number of times users have been privately notified by a flag
- Pageviews: number of new pageviews from all visitors
- Post Edits: number of new post edits
- Posts: new posts created during this period
- Signups: new account registrations for this period
- Suspicious Logins: details of new logins that differ suspiciously from previous logins
- System: number of personal messages sent automatically by the System
- Time to first response: average time (in hours) of the first response to new topics
- Top Ignored / Muted Users: users who have been muted and/or ignored by many other users
- Top Referred Topics: topics that have received the most clicks from external sources
- Top Referrers: users listed by number of clicks on links they have shared
- Top Traffic Sources: external sources that have linked to this site the most
- Top Uploads: list of all uploads by extension, file size and author
- Top Users by likes received: top 10 users who have received likes
- Top Users by likes received from a user with a lower trust level: top 10 users in a higher trust level being liked by people in a lower trust level
- Top Users by likes received from a variety of people: top 10 users who have had likes from a wide range of people
- Topics: new topics created during this period

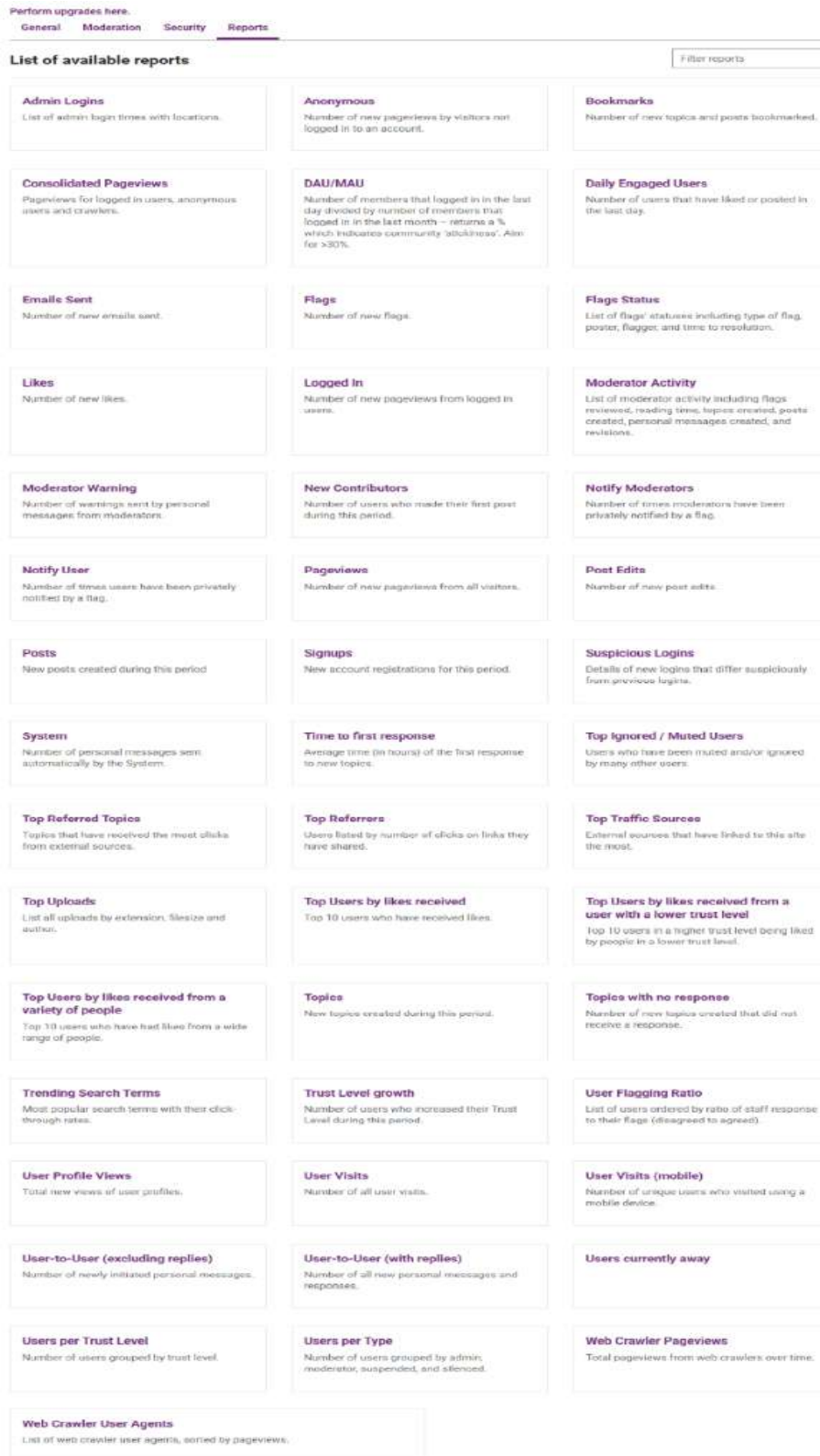


Figure 35: RESET Forum - Security Admin statistics

- Topics with no response: number of new topics created that did not receive a response
- Trending Search Terms: most popular search terms with their click-through rates
- Trust Level growth: number of users who increased their Trust Level during this period
- User Flagging Ratio: list of users ordered by ratio of staff response to their flags (disagreed to agreed)
- User Profile Views: total new views of user profiles
- User Visits: number of all user visits
- User Visits (mobile): number of unique users who visited using a mobile device
- User-to-User (excluding replies): number of newly initiated personal messages
- User-to-User (with replies): Number of all new personal messages and responses.
- Users per Trust Level: number of users grouped by trust level
- Users per Type: number of users grouped by admin, moderator, suspended, and silenced
- Web Crawler Pageviews: total pageviews from web crawlers over time
- Web Crawler User Agents: list of web crawler user agents, sorted by pageviews.

4. Open Data Processing Pipeline

Open data represent a considerable proportion of the project’s data and extensive online research has been conducted to identify the most appropriate open data sources, as described in Chapter 2. For the effective filtering and processing of these data, AUTH defined a methodology depicted in Figure 36. In this chapter the filtering process of the open data collected is being presented in detail.

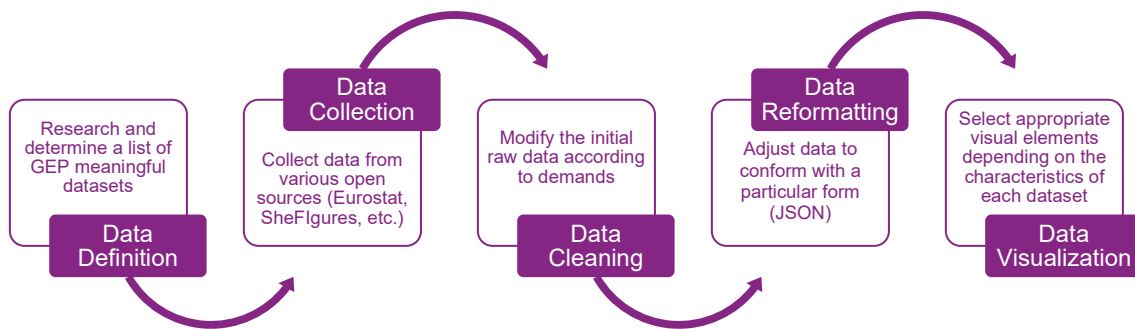


Figure 36: Open data pipeline

4.1. Data Collection and Evaluation

Among numerous available open databases, a list of specific meaningful GE datasets was decided, such as to serve the goals of the project. Specifically, the selected datasets were those that would address the status of gender equality, inclusiveness, and diversity policies in partner institutions as well as the overall situation of participating countries, other EU member states, and other countries. To this end, the datasets described below were collected from SheFigures, UIS Women in Science, the EIGE Gender Statistics Database, Eurostat, the WDI of the World Bank and QS University Rankings (see section 2.1.1).

Ratio of women to men among active authors, in all fields of R&D, per seniority level, 2015-2019

The source of the specific dataset is SheFigures 2021, and the calculations are computed by Elsevier using Scopus data.

The fields of R&D in the dataset are divided into Natural Sciences, Engineering and Technology, Medical and Health Sciences, Agricultural and Veterinary Sciences, Social Sciences, Humanities and the Arts, while active authors are defined as those that produced 10 or more papers in the last 20 years (2000-2019) and at least one paper in the last five years or those who produced four or more papers in last five years.

Seniority level is estimated via the time elapsed since an author's first publication in a journal indexed in Scopus and is categorised as follows:

- < 5 years or 'early-stage': authors whose first paper in Scopus was published up to and including the years 2015-2019.
- 5 to 10 years or 'middle-stage': authors whose first paper in Scopus was published in the years 2010-2014.
- >10 years or 'senior' authors: authors whose first paper in Scopus was published in the year 2009 or earlier.

The value for each seniority level and each field states the gender parity. Gender parity between women and men is indicated by a ratio of 1.0. A ratio of 1.0 indicates as many active women authors as men authors at a given seniority level. If the ratio is above 1.0, it means that the number of active women authors in the group exceeds the number of active men authors and if it is below 1.0, it means that the number of men authors in the group exceeds the number of women authors.

Country GDP, 2000-2019

The source of this dataset is World Development Indicators (WDI). The source values of the were in USD but were converted in EUR for the needs of the project, with the rate at the time of conversion being 1.00 USD = 0.9188 EUR.

Country gender pay gap and/or gender employment gap in time

The source of these datasets is Eurostat.

Gender Employment Gap: The indicator measures the difference between the employment rates of men and women aged 20 to 64, while the employment rate is calculated by dividing the number of individuals aged 20 to 64 in employment by the total population of the same age group. The indicator is based on the *EU Labour Force Survey* (EU-LFS).

Gender Pay Gap (in unadjusted form): In general, the gender pay gap measures the difference between the average earnings of women and men in the workforce. The gender pay gap is an internationally established measure of women's position in the economy in comparison to men. Eurostat defines the unadjusted gender pay gap as the difference between the average gross hourly earnings of men and women expressed as a percentage of the average gross hourly earnings of men. The particular indicator is calculated for enterprises with 10 or more employees. The indicator has been defined as unadjusted because it gives an overall picture of gender inequalities in terms of pay and measures a concept which is broader than the concept of equal pay for equal work. All employees working in firms with ten or more employees, without restrictions for age and hours worked, are included. The gender pay gap is based on the methodology of the *Structure of Earnings Survey* (SES), which is conducted every four years.

Country gender equality indexes

The source of the dataset is European Institute for Gender Equality. Gender Equality Index (EIGE, 2022) is a tool that measures how far (or close) the EU and the Member States are from achieving a gender equality society and produces a score between 1 and 100. It also provides the respective results for each one of the six core domains: *work, money, knowledge, time, power, and health*. Furthermore, for specific countries such as Albania and Serbia, data were not available in the generic European list. For these cases, data were collected by the separate country publications (Marija Babović, 2021).

Country Population, 1990-2020

The source of this dataset is World Development Indicators (WDI). According to WDI, total population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship.

R&D personnel by sector of performance, professional position, and sex, 2015-2019

The source of this dataset is Eurostat - Statistics on research and development. The dataset employs two units of measure, i.e., headcount – where each individual counts as one employee whether their appointment is full time or part time – and full time equivalent – that represents the actual workload hours as an equivalent to the status of the worker as full time.

Proportion (%) of men and women in a typical academic career, students, and academic staff, 2007-2013-2015-2018

The source of this dataset is the Women in Science database and the statistics are about EU Members. Academics are classified by grade, i.e.,

Grade A: the single highest grade/post at which research is normally conducted

Grade B: researchers working in positions not as senior as top position (A) but more senior than newly qualified PhD holders

Grade C: the first grade/post into which a newly qualified PhD graduate would normally be recruited.

Regarding students, the International Standard Classification of Education (ISCED) is employed, where:

ISCED 6 and 7: denote tertiary programmes that provide sufficient qualifications to enter advanced research programmes and professions with high skills requirements

ISCED 8: refers to tertiary programmes which lead to an advanced research qualification (PhD) (Eurostat, 2022).

QS World University Rankings 2022

The source of this dataset is QS Quacquarelli Symonds. The QS World University Rankings feature over 1,400 universities from around the world. Institutions are assessed across six categories (or indicators) to effectively capture university performance – including academic and employer reputation, faculty/student ratio and research citations (for more information on the methodology followed by QS the reader can refer to <https://www.topuniversities.com/users/laural>)

Proportion (%) of women on boards, members, and leaders, 2010-2019

The source of this dataset is Women in Science database and refers to the percentage of women on scientific and administrative boards, or advisory boards of a research organization, publicly or privately managed and financed. By boards is meant scientific and administrative boards, or advisory boards of a research organization, publicly or privately managed and financed.

4.2. Data Pre-processing – Cleaning

Data collection through the selected databases generated a list of large datasets in various forms, such as .xls or .csv files. Consequently, each dataset had to be thoroughly examined in order to determine any possible need for filtering. Data filtering is a widespread practice and can be performed on any existing attribute value of the database (Inmon, Linstedt, & Levins, 2019). Filtering data may be intended to exclude invalid cases from the dataset, to analyse results for a specific data subset of interest or for a specific time frame, or to train statistical models (Facer, 2018).

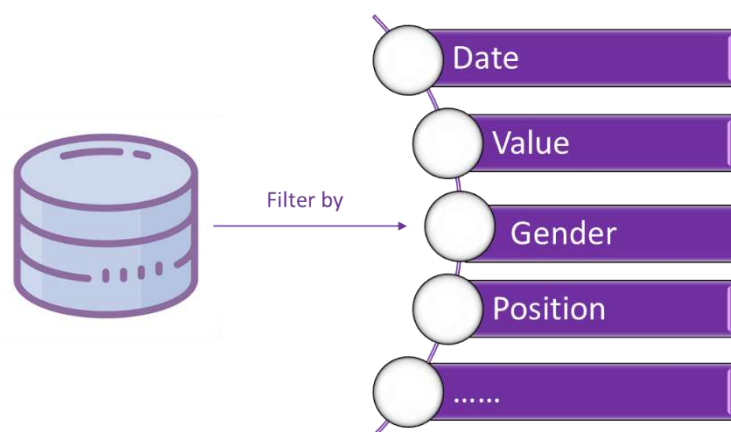


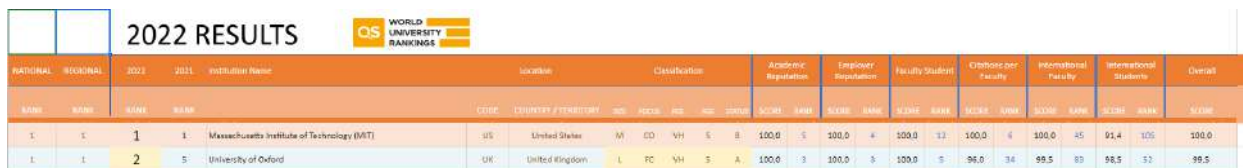
Figure 37: Data Filtering⁵

⁵ Database icon by [icons8](https://icons8.com/icon/42904/database) <https://icons8.com/icon/42904/database>

Filtering involves the employment of a set of rules to specify the observations to include or exclude from the analysis. In the case of the selected datasets, the inclusion rule was for the data to meet the demands of the project and serve the goals of the RESET platform. This procedure led to several modifications which had to be performed to “clean” the initial raw data so that they met the criteria.

Examples of data filtering followed for RESET platform are the following:

- removal of non-European countries from the datasets
- removal of previous years data according to the definitions made for each dataset
- removal of institutions classification data and scores across each assessment category (only institution ranks in each assessment category were stored), as shown in Figure 38.



NATIONAL		REGIONAL		2022	2021	Institution Name	Location	Classification	Academic Reputation	Employer Reputation	Faculty Staff	Citations per Faculty	International Faculty	International Students	Overall	
RANK	RANK	RANK	RANK				CODE	COUNTRY / TERRITORY	QS	AR	ER	FS	CF	IF	IS	SCORE
1	1	1	1			Massachusetts Institute of Technology (MIT)	US	United States	M	CD	VH	S	B			100.0
1	1	2	5			University of Oxford	UK	United Kingdom	L	RC	VH	S	A			100.0

Figure 38: QS Original source state snapshot

Apart from irrelevant data removals, further pre-processing essential techniques were applied as well. More precisely, in some datasets, countries were represented by full name, in others by the ISO 3166-1 alpha2, while in others by ISO 3166-1 alpha3 code – examples are shown in Figure 39. In order to be able to make comparisons and preserve consistency in the data, the appropriate mapping had to be developed so that the same structure is followed in every case.












Country	ISO 3166-1 alpha2	ISO 3166-1 alpha3
 Afghanistan	AF	AFG
 Åland Islands	AX	ALA
 Albania	AL	ALB
 Algeria	DZ	DZA
 American Samoa	AS	ASM
 Andorra	AD	AND
 Angola	AO	AGO
 Anguilla	AI	AIA
 Antarctica	AQ	ATA
 Antigua and Barbuda	AG	ATG
 Argentina	AR	ARG

Figure 39: Country alpha2 and alpha3 codes

An alternative example is the dataset merging. Data merging is the process of combining two or more data sets into a single data set. This process makes it easier and faster to analyse data stored in multiple locations, worksheets, or data tables. Merging data into a single point is necessary in certain situations, especially when there are needs to add new cases, variables, or data based on the lookup values. However, data merging needs to be performed with caution; otherwise, it can lead to duplication, inaccuracy, or inconsistency issues (What is Data Merging?, 2022). Data merging was utilized in the dataset of “R&D personnel by sector of performance, professional position and sex”.

4.3. Data Reformatting

As data are presented on the dashboard of the RESET platform, it was necessary to extract them to a file form that would comply with the platform requirements. A generic form for data storage is JavaScript Object Notation format, i.e., JSON files, which are mainly used for data transmission in web applications. Figure 40 presents a visualised concise description of JSON files. Specifically, JSON is a text-based, readable file format that can be opened on any text editor software.

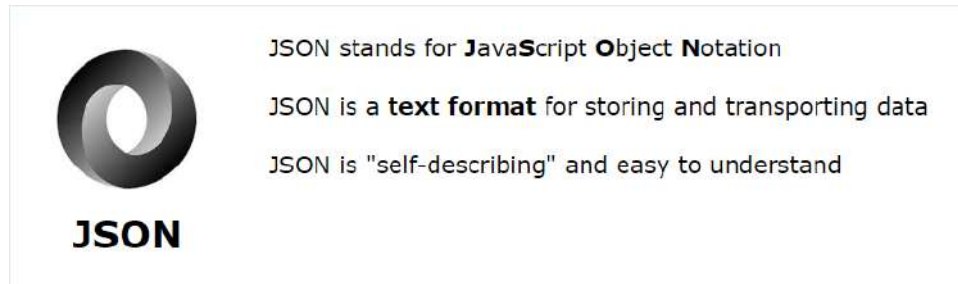


Figure 40: JSON file description
(Source: https://www.w3schools.com/js/js_json_intro.asp)

It is a collection of key-value pairs where the key must be a string type, and the value can be of any of the following types:

- Number
- String
- Boolean
- Array
- Object
- Null.

A couple of important rules to note: (ADHIKARY, 2021)

- In the JSON data format, the keys must be enclosed in double quotes.
- The key and value must be separated by a colon (:) symbol.
- There can be multiple key-value pairs. Two key-value pairs must be separated by a comma (,) symbol.
- No comments (// or /* */) are allowed in JSON data.

JSON is mostly used to transmit data from the server to a client, so that they appear on a webpage, or the opposite. As initial data were stored in various formats (e.g., pdf, xlsx, csv, txt), they were all converted to be compatible with the JSON format.

Figure 41 depicts a simple JSON example.

```
[
  {
    "_id": "638fc1f68f86b2d353a740de",
    "age": 27,
    "name": "Susanne Serrano",
    "favoriteFruit": "strawberry"
  },
  {
    "_id": "638fc1f66ac56c40615d4523",
    "age": 31,
    "name": "Millicent Robbins",
    "favoriteFruit": "apple"
  },
  {
    "_id": "638fc1f63055592fa3beee8d",
    "age": 34,
    "name": "Livingston Anderson",
    "favoriteFruit": "apple"
  },
  {
    "_id": "638fc1f6aaa79b98609c7009",
    "age": 29,
    "name": "Mccullough Santana",
    "favoriteFruit": "strawberry"
  },
  {
    "_id": "638fc1f684707e5ae0b9f1ff",
    "age": 28,
    "name": "Leona Taylor",
    "favoriteFruit": "banana"
  },
  {
    "_id": "638fc1f69f680ece63eefc3a",
    "age": 31,
    "name": "Jannie Salas",
    "favoriteFruit": "strawberry"
  },
  {
    "_id": "638fc1f6deab733ec2609ea0",
    "age": 28,
    "name": "Dunlap Walters",
    "favoriteFruit": "banana"
  }
]
```

Figure 41: JSON simple example

4.4. Data Visualisations

Based on the features of each dataset, various visualisation approaches were considered, to choose the most appropriate ones to depict their content. Visual elements, such as charts and graphs, can provide a straightforward representation of

data that supports the identification of data patterns and outliers. Representative examples of data visualisations on the RESET platform are shown in the Figures below. Figures 42 and 43 appear on the home page of the dashboard and depict the latest data on Gender Equality Index and Gender Employment Gap, respectively, for the countries represented in the consortium. On the other hand, Figure 44 appears when selecting **Women On Boards**, under the option **Gender Gap**, on the side navigation menu and depicts the trajectory of percentage of women on boards, i.e., the percentage of women – both members and leaders – on boards, where boards are scientific and administrative boards, or advisory boards of a research organisation, publicly or privately managed and financed. The selected time period in the screenshot is for the years 2015-2019, while the selected countries are the six countries participating in the project. Figure 45 appears with the selection of **Academic Career**, under the title **Academy**, in the side navigation menu and shows the proportion of men and women in a typical academic career, students and academic staff, for years 2007-2013-2015-2018, in EU Member States.

Gender Equality Index (2019)

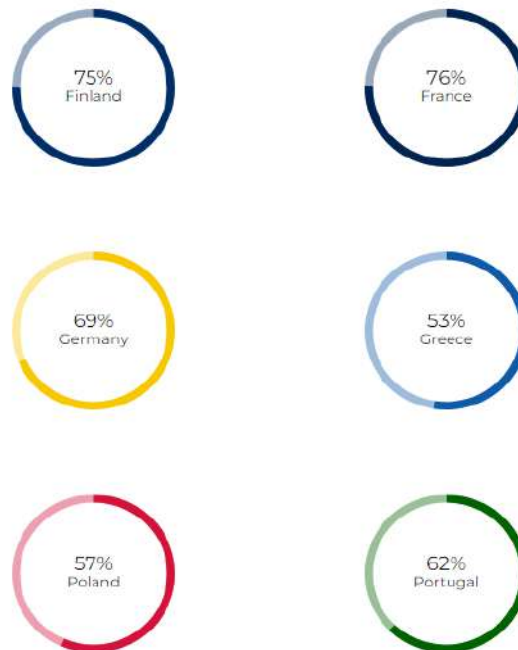


Figure 42: Visualization of the Gender Equality Index

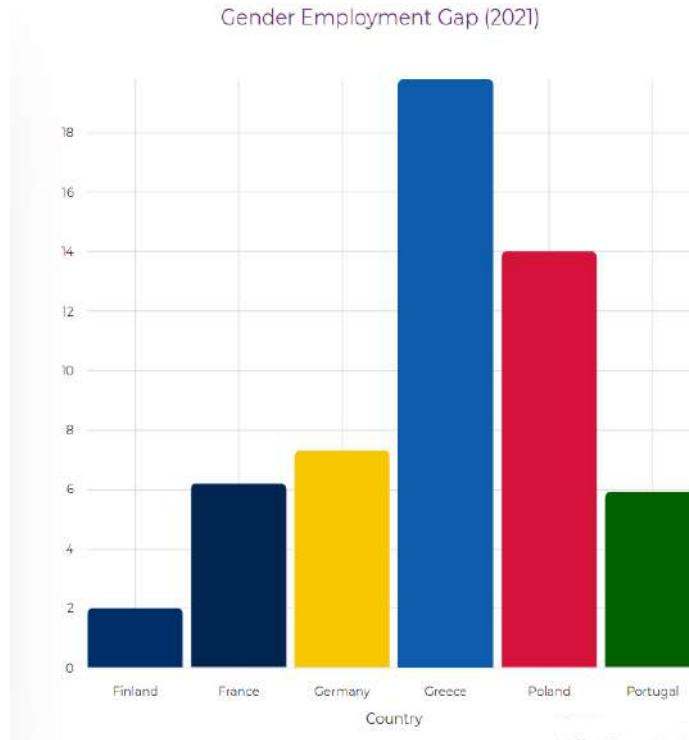


Figure 43: Visualization of the Gender Employment Gap

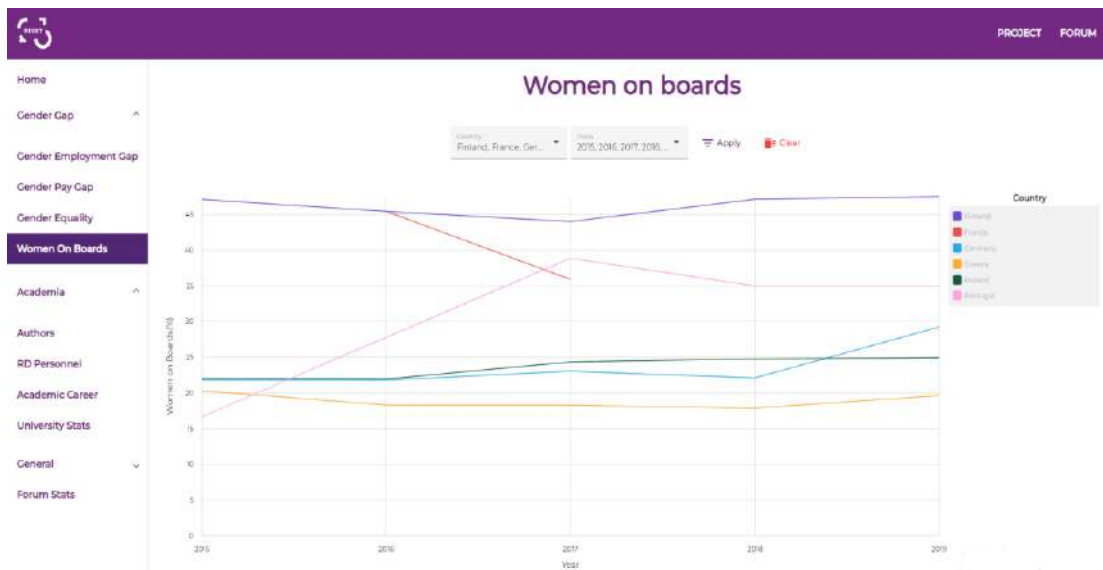


Figure 44: Percentage of women of boards in the participating countries

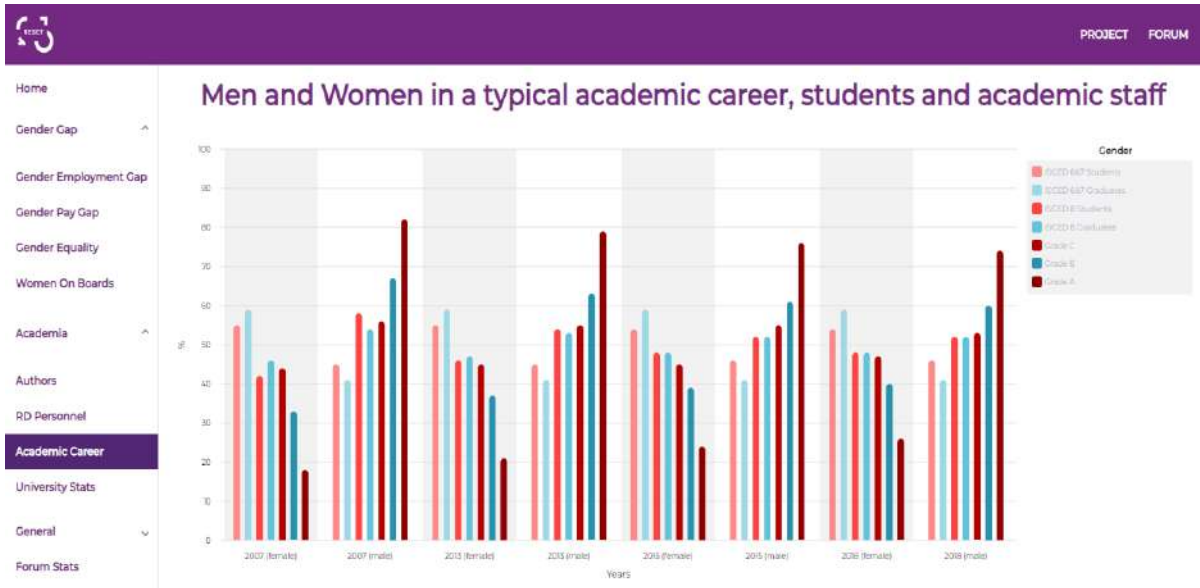


Figure 45: Men and Women in a typical academic career, students, and academic staff

5. HR Data Processing Pipeline

As it has been discussed in section 2.1.2, in collaboration with all the GEP implementing partners, data concerning GE issues from Human Resources (HR) departments of the institutions has been collected, while Figure 46 depicts the corresponding data collection and processing pipeline.

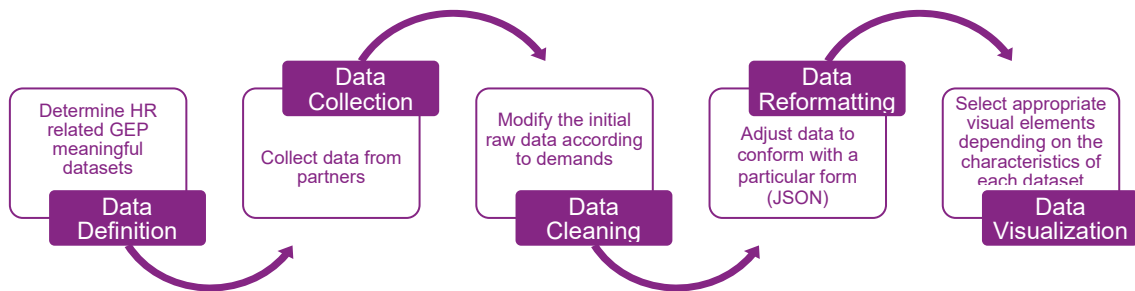


Figure 46: HR Data Pipeline

5.1. Data Collection

Based on the outcomes of the workshops, as described in section 2.1.2, AUTH defined the HR related datasets that would be needed from the HR departments and designed a unified template to be filled in with all required information by the HR departments of UBx, AUTH, LU, and UPorto. Specifically, the data requested referred to:

- Percentages of male/female administrative staff, teachers, researchers per faculty/scientific field
- Percentages of teachers/researchers per age group, gender, and faculty
- Sex ratio of teaching/researching staff per scientific field
- Male/female percentages per rank
- Proportion of women in top leadership and unit head positions
- Percentages of inventorships and patents per gender.

Examples of the spreadsheet template in demand are presented in Figures 47 and 48.

	A	B	C	D	E	F	G	H	I	J	K
		%									
		Men	Women								
3	Faculty of Theology	0,00%	100,00%								
5	Faculty of Philosophy	20,51%	79,49%								
7	Faculty of Law	10,00%	90,00%								
9	Faculty of Economics and Political Sciences	14,29%	85,71%								
11	Faculty of Sciences	11,11%	88,89%								
13	Faculty of Engineering	22,22%	77,78%								
15	Faculty of Health Sciences	29,55%	70,45%								
17	Faculty of Fine Arts	14,29%	85,71%								
19	Faculty of Agriculture Forestry and Natural Environmen	27,78%	72,22%								
21	Faculty of Physical Education and Sport Sciences	50,00%	50,00%								
23	Faculty of Education	37,50%	62,50%								

Figure 47: HR Data Template (sex ratio per faculty)

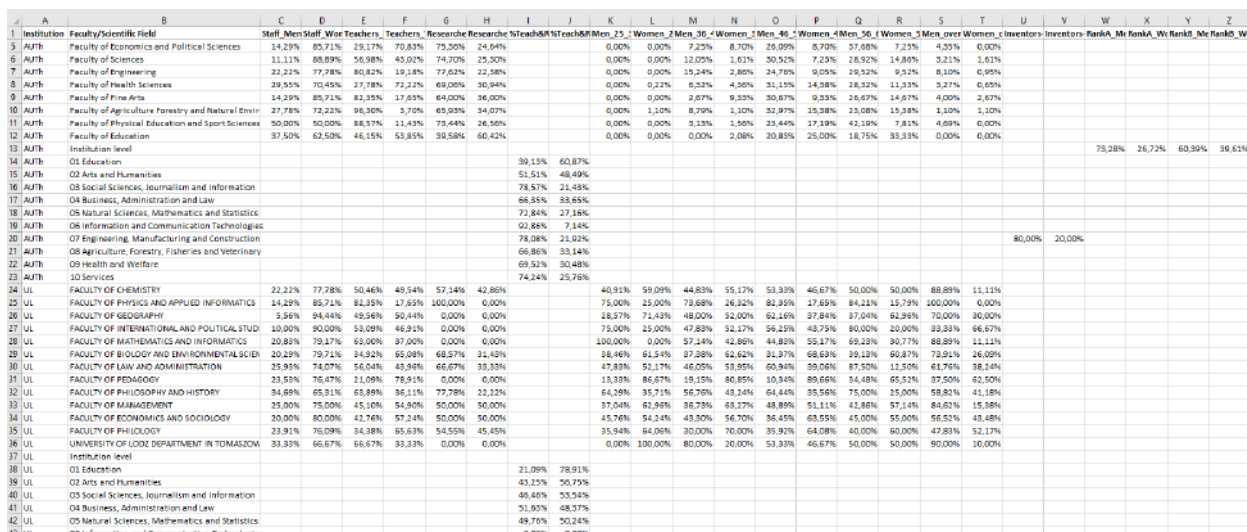
	A	B	C	D	E	F	G	H	I	J	K
		%									
		Men	Women								
3	01 Education	39,13%	60,87%								
4	02 Arts and Humanities	51,51%	48,49%								
5	03 Social Sciences, Journalism and Information	78,57%	21,43%								
6	04 Business, Administration and Law	66,35%	33,65%								
7	05 Natural Sciences, Mathematics and Statistics	72,84%	27,16%								
8	06 Information and Communication Technologies	92,86%	7,14%								
9	07 Engineering, Manufacturing and Construction	78,08%	21,92%								
10	08 Agriculture, Forestry, Fisheries and Veterinary	68,86%	33,14%								
11	09 Health and Welfare	69,52%	30,48%								
12	10 Services	74,24%	25,76%								

Figure 48: HR Data Template (sex ratio per scientific field)

5.2. Data Pre-processing – Filtering

After the completion of the data collection, as described above, it was necessary to filter the initial raw data. Available data from each partner’s HR department did not match all requested fields. Therefore, several modifications needed to be made, so that our data would have a cohesive structure and inter-institutional comparisons would be possible.

Thus, the data were combined into a single Excel worksheet, as shown in Figure 49, and the dataset was checked for any missing or invalid values. Concerning the cases where certain values were not available for an institution, it was decided to include in the dataset the values for the other partner universities and only available data for the institution in question. However, cases were excluded when data were available only for one partner.



The table displays HR data for 43 institutions, categorized by faculty/scientific field. The columns represent various demographic and professional metrics such as Staff, Men, Women, Teachers, Researchers, and Inventors, along with their respective percentages and counts. The data is organized in a grid format with columns labeled A through Z.

Figure 49: Part of Unified Spreadsheet with HR Data

5.3. Data Analysis

The next step was the analysis of the data, to draw interesting conclusions and meet the need for visualisation, as described in the following session. The data to be presented on the platform were categorised in two types, shown in two different tabs. The first one included data to be displayed for each university, i.e., *the percentages of male and female administrative staff, teachers, and researchers per faculty for all universities and the percentages of male and female professors/researchers per faculty and per age group (i.e., 25-35, 36-45, 46-55, 56-65, over 65) for all universities.* The second one contained data to be presented comparatively, which were further divided into data presented per scientific field for each university, i.e., *the percentages of male and female professors/researchers, and data presented at institutional level, i.e., the percentages of male and female professors/researchers per rank, the percentages of women in top leadership and unit head positions, and the percentages of male and female inventors and patent owners.*

```
{
  "Auth": [
    {
      "Faculty/Scientific Field": "Faculty of Theology",
      "Data": {
        "Staff_Men": "0.00%",
        "Staff_Women": "100.00%",
        "Teachers_Men": "42.86%",
        "Teachers_Women": "57.14%",
        "Researchers_Men": "73.77%",
        "Researchers_Women": "26.23%",
        "Men_25_35": "0.00%",
        "Women_25_35": "0.00%",
        "Men_36_45": "8.20%",
        "Women_36_45": "0.00%",
        "Men_46_55": "36.07%",
        "Women_46_55": "9.84%",
        "Men_56_65": "27.87%",
        "Women_56_65": "16.39%",
        "Men_over_65": "1.64%",
        "Women_over_65": "0.00%"
      }
    },
    {
      "Faculty/Scientific Field": "Faculty of Philosophy",
      "Data": {
        "Staff_Men": "20.51%",
        "Staff_Women": "79.49%",
        "Teachers_Men": "17.95%",
        "Teachers_Women": "82.05%",
        "Researchers_Men": "37.13%",
        "Researchers_Women": "62.87%",
        "Men_25_35": "0.00%",
        "Women_25_35": "0.50%",
        "Men_36_45": "5.94%",
        "Women_36_45": "9.90%",
        "Men_46_55": "15.84%",
        "Women_46_55": "28.22%",
        "Men_56_65": "14.36%",
        "Women_56_65": "22.77%",
        "Men_over_65": "0.99%",
        "Women_over_65": "1.49%"
      }
    }
  ]
}
```

Figure 50: Part of JSON File – AUTH HR Data

As in the case of open data, the HR data also had to comply with the platform’s storage demands. Consequently, HR data were aggregated per institution, to be presented more clearly and were converted to JSON files to meet the platform’s requirements. Figure 50 shows an example of a JSON file, i.e., a part of the file that contains the HR data of AUTH. As discussed in section 4.3, JSON files are text-based files that can be opened with any text editing application.

5.4. Data Visualisations

According to the internal characteristics of the datasets, meaningful visualisations were planned. The goal set for the visualisation was to provide visual insights to the most interesting data outcomes. The created HR data representations can be accessed through the Dashboard’s side navigation menu on the RESET platform, by selecting **University Stats**, under the option **Academia**. Representative visualisation examples are presented in Figures 51-54.

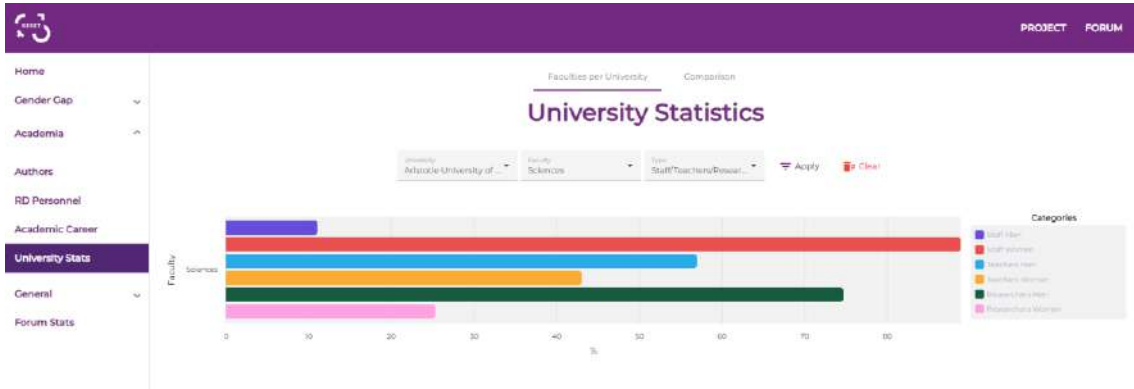


Figure 51: Percentages of male/female administrative and academic staff per faculty

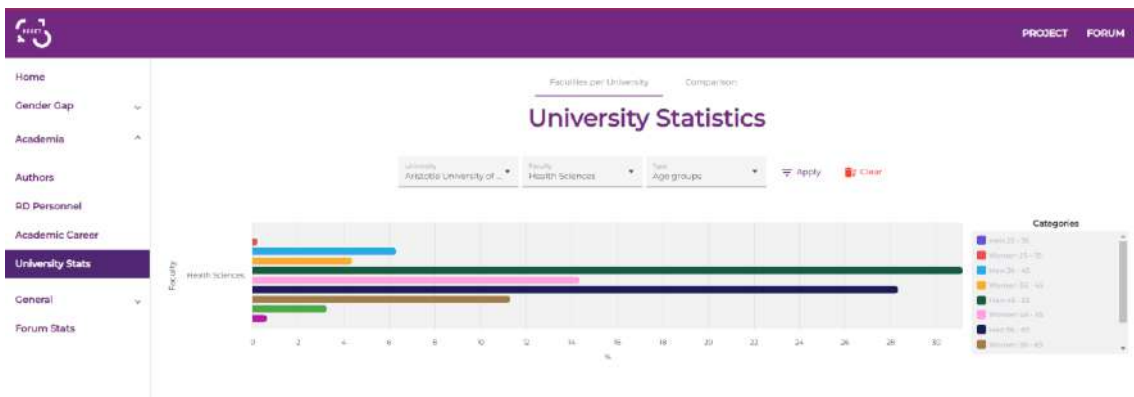


Figure 52: Percentages of male/female academics per faculty and age group



Figure 53: Comparison of male/female academics percentage per faculty



Figure 54: Comparison of statistics at institutional level

6. Conclusions and Future Work

Under the tasks of WP3 “Supporting data-driven GE and diversity policy-making in designing qualitative assessment tools and processes”, a plethora of data is collected and analysed. The data filtering methodology that is described in detail in this document refers to qualitative and quantitative data, static or dynamic, that were collected and processed to be visualised on the RESET GE awareness platform (see D3.2). The datasets have been carefully collected and thoroughly filtered, to the end of assisting the implementation and monitoring of the GEPs and of supporting their redesign.

In the following months until the end of the project, as dynamic data are going to be collected with the RESET Forum, data analysis on the Forum activities will be continued, along with the Natural Language Processing (NLP) performed on the Forum content, and the addition of the thematic analysis to achieve insight to the community’s lessons learnt and good practices gained through the project. Moreover, the inclusion of new forum metrics and statistics and the respective graphic representations will be considered, depending on the increase in the Forum activity. Finally, static data are going to be updated and additional data collection and analysis activities will be planned, in respect with the needs created by ongoing project activities. Nevertheless, co-design actions, concerning all types of data, data analysis, and visualisations can take place. These actions can involve partner universities and/or other stakeholders and include their envision in decision making about the data management activities both through the RESET Dashboard and Forum and generally throughout the project.

References

- ADHIKARY, T. (2021, November 29). *JavaScript Object Notation Explained in Plain English*. Retrieved from freecodecamp:
<https://www.freecodecamp.org/news/what-is-json-a-json-file-example/>
- EC. (2022). *Open Data*. Retrieved from <https://digital-strategy.ec.europa.eu/en/policies/open-data>
- EC. (2022). *Open data portals*. Retrieved from <https://digital-strategy.ec.europa.eu/en/policies/open-data-portals>
- EC, Directorate-General for Research and Innovation. (2021). *She figures 2021: gender in research and innovation: statistics and indicators*. Publications Office. Retrieved from <https://data.europa.eu/doi/10.2777/06090>
- EIGE. (2022). *Gender Equality Index 2022*. Retrieved from <https://eige.europa.eu/gender-equality-index/2022>
- Eurostat. (2021). *Eurostat and the European Statistical System*. Retrieved from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Eurostat_and_the_European_Statistical_System#Eurostat
- Eurostat. (2022). *International Standard Classification of Education (ISCED)*. Retrieved from [https://ec.europa.eu/eurostat/statistics-explained/index.php?title=International_Standard_Classification_of_Education_\(ISCED\)](https://ec.europa.eu/eurostat/statistics-explained/index.php?title=International_Standard_Classification_of_Education_(ISCED))
- Facer, C. (2018). *What is Data Filtering?* Retrieved November 25, 2022, from <https://www.displayr.com/what-is-data-filtering/>
- Inmon, W., Linstedt, D., & Levins, M. (2019). *Data Architecture: A Primer for the Data Scientist*. Academic Press. doi:10.1016/C2018-0-01666-7
- Keita, Z. (2022, 03 02). *Social Media Sentiment Analysis In Python With VADER*. Retrieved from Towards Data Science: <https://towardsdatascience.com/social-media-sentiment-analysis-in-python-with-vader-no-training-required-4bc6a21e87b8>
- Marija Babović, M. P. (2021). *Gender Equality Index for the Republic of Serbia 2021*. Social Inclusion and Poverty Reduction Unit of the Government.
- Murray-Rust, P. (2008). Open data in science. *Nature Precedings*, 1(1). Retrieved from <https://www.nature.com/articles/npre.2008.1526.1.pdf?origin=ppub>
- UIS. (2019). *Women in Science*. Retrieved from <https://uis.unesco.org/sites/default/files/documents/fs55-women-in-science-2019-en.pdf>