# Analysis of the STI Policy Database "STIP COMPASS"

**Pierfrancesco Diella**[1]

[1]Politecnico di Bari, Italy - Management Engineering

pierfra.diella@gmail.com

## 1. Introduction

STIP Compass is a joint initiative of the European Commission (EC) and the OECD. Its goal is to collect qualitative and quantitative data on national trends in science, technology and innovation (STI) policies and provide a single centralized platform for research and policy decision support.The main data come from countries' responses to the survey conducted every two years. The responses are provided by national government officials responsible for STI policies.

The survey covers a wide range of issues. It consists of 57 questions covering 6 'core' policy areas, plus a policy area added this year:

1. Governance
2. Public research system policy area
3. Innovation in firms and innovative entrepreneurship
4. Knowledge exchange and co-creation
5. Human resources for research and innovation
6. Research and innovation for Society
7. *Net-zero transition.

The territorial coverage of the survey is 57 countries. In practice, government officials respond, for each Policy Area, first by indicating the main policy debates *(e.g. Briefly, what are the main ongoing policy debates around government support for the public research system?)*. Then, answering the remaining questions, they point out the policy initiatives by providing a set of details for each initiative, such as the organisation responsible for the initiative, the instruments used to implement the plan, data on the budget estimate, the start and completion date, etc.

## 2. Objectives

In this project, therefore, these 2 forms of data were analysed separately with different methodologies: debates and initiatives. In particular, the analysis focused around 3 objects extrapolated from the survey: textual data from debates, Countries involved in the initiatives and instruments indicated by countries.

The main objectives were:

- to fix a model or methodology that can extract useful patterns and information from the textual data of debates;
- to understand the scale and proportions to each other of the initiatives conducted by different countries;
- to identify pattern and combination of tools across countries.

In this case, the 2023 edition, administered between February and April 2023, was considered.

## 3. Methodology

The first step was the analysis of the debates. Each country indicated 7 Policy area responses by answering a simple question. The raw data available from the Compass STIP site was imported via XLSX file and merged. The pre-processing activities led to the creation of a single dataset in which the data was transformed into a tidy format and organised into the columns 'ID', 'Country', 'Policy area' and 'description', which represents the column containing the responses and thus the textual data to be analysed.

An initial approach to understanding the nature of this data was the generation of simple Word-clouds of all debates (Fig. 1) and one for each policy area (e.g. Fig. 2). Preliminary activities to this were tokenisation of the corpus of this data and other cleaning activities such as removal of punctuation, numbers and stop-words.


**Figure 1. Wordcloud considering all debates**


**Figure 2. Wordcloud of debates on policy area Governance**

This approach generally highlighted keywords considering all debates, such as 'innovation', 'development', 'research', 'policy', 'national'. However, the level of detail of characterisation of the policy area did not increase when analysing the individual groups with this technique. Only with regard to the last module, the one concerning zero emissions, did the word-cloud highlight different words compared to those encountered in the others and thus appear to be more characterising

Then, to explore a further approach, the 'tf-idf' of the most recurring words for a single Country policy area was calculated: Italy was chosen as a case study (Fig. 3).

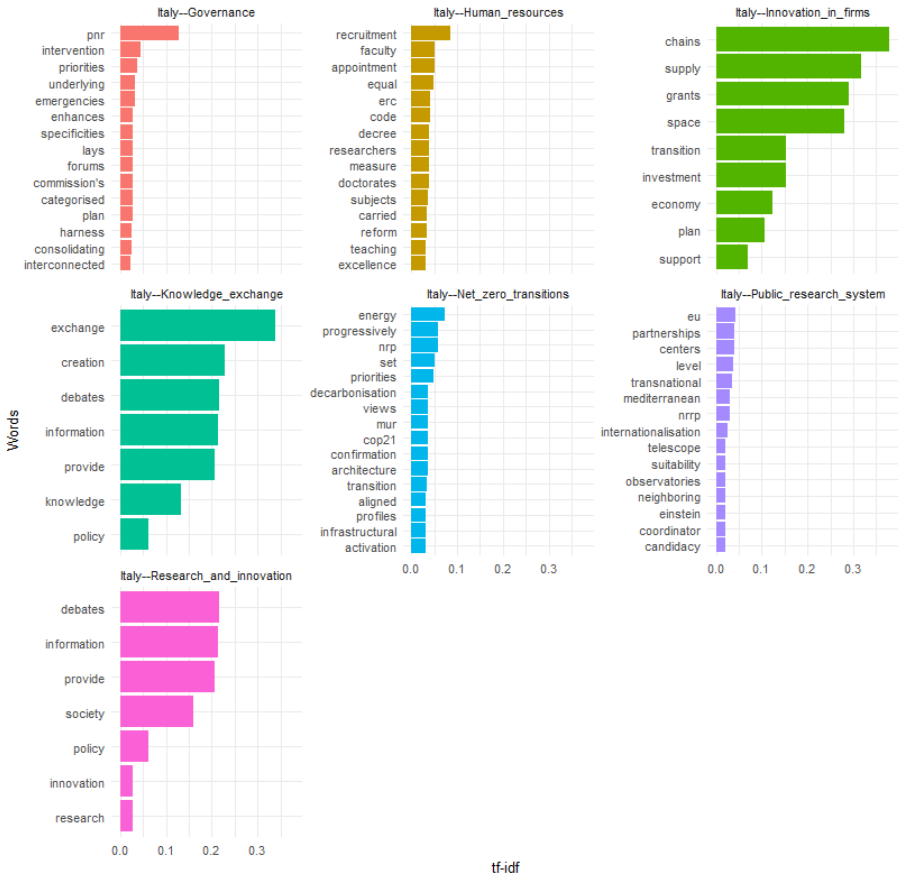


**Figure 3. 'tf-idf' of Italy debates**

Indeed, there is a fairly good correspondence between words with a high 'tf-idf' value and the related topic of the document. For example, words such as 'PNRR' (National Recovery and Resilience Plan) or 'recruitment' are better able to differentiate documents related to governance or human resources than others. This trend and these results are evident in all analysed policy areas.

We went on to analyse the initiatives. the data came from a CSV file. Once imported, several cleaning and pre-processing operations were carried out; some of these were the saving of the column headings in an additional dataset called 'codebook', which contains the explanation of the individual variables: this was a crucial operation for the understanding of this dataset.

In the main dataset, for each initiative a large amount of further information is recorded, e.g. concerning the original names, dates, budgets, countries, objectives, national organisation responsible, but above all a range of information on the instruments used. In fact for each initiative OECD provides a complex taxonomy of instruments to be matched to the initiatives. A further pre-processing activity was therefore to simplify this by considering only the type of instrument used.

7293 initiatives were analysed, of which 1949 were structural reforms. Within

these initiatives, 10102 individual instruments were used. There were 51 reference themes and finally, 57 countries with 2273 national organisations were registered. A study was conducted on the countries and national organisations and then an analysis of the instruments was carried out, in particular how several combined instruments are used within the same initiative.

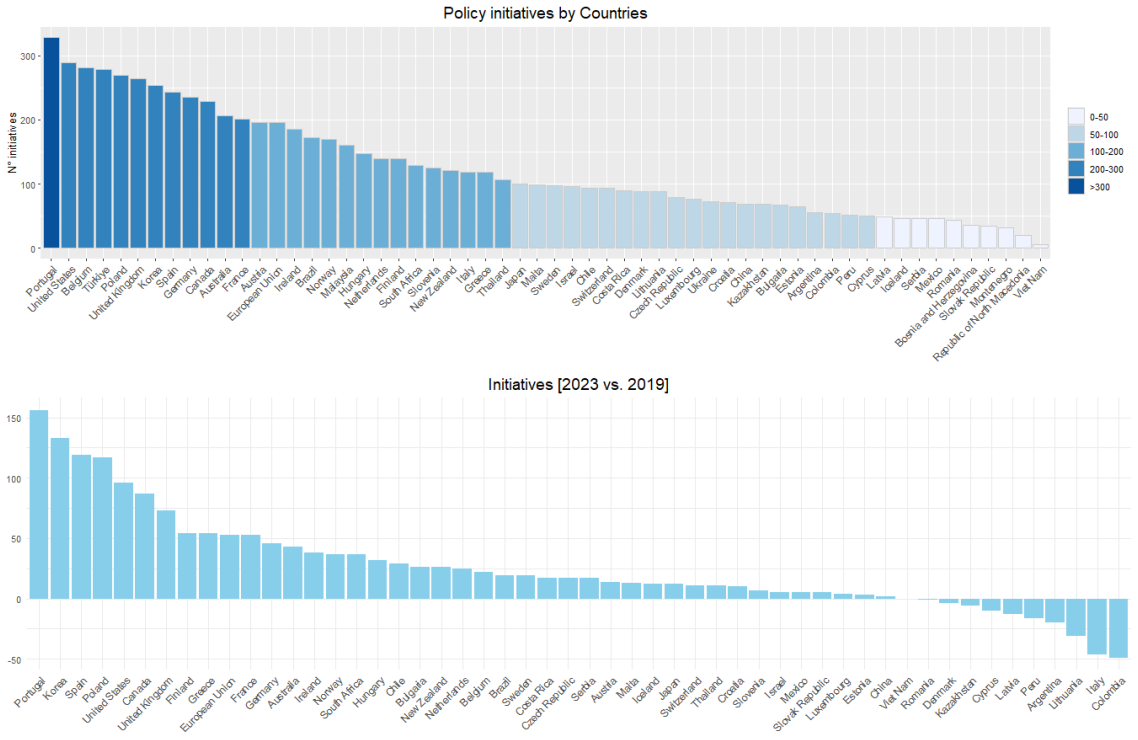


**Figure 4. Number of Country Initiatives / Variation since the last survey**

Fig. 4 shows the number of initiatives reported by the various countries. These have been arbitrarily divided into 5 groups, based on their number of initiatives. The variation since the last survey in 2019 is reproduced below. The combined analysis of these two graphs shows that some countries, such as Portugal above all, but also Poland and the United States for example, have achieved these results in the last four years. On the other hand, there are other countries such as Italy and Denmark that have reported fewer initiatives than in the previous survey.

As far as the analysis of national organisations is concerned, the 10 countries whose domestic organisations had the highest number of initiatives were taken into consideration. It would appear (Fig. 5) that in these countries there are organisations, and therefore ministries, responsible for a large number of initiatives.

This evidence is confirmed by Fig. 6, which clearly shows that for Brazil European Union Germany Hungary and Ukraine there is one organisation that contributes to more than 40% of all initiatives. In other countries, on the other hand, it is assumed that there is a more evenly distributed contribution of the responsible organisations

Next came an analysis of the instruments mentioned within the initiatives. The most frequently used type of instrument emerged to be 'strategies, agendas and plans',

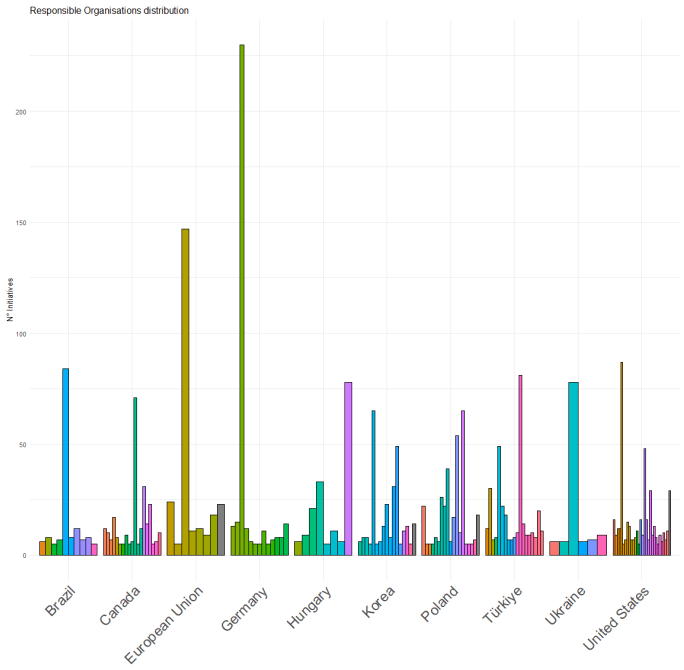


**Figure 5. Responsible Organisations per Country and Number of Initiatives**

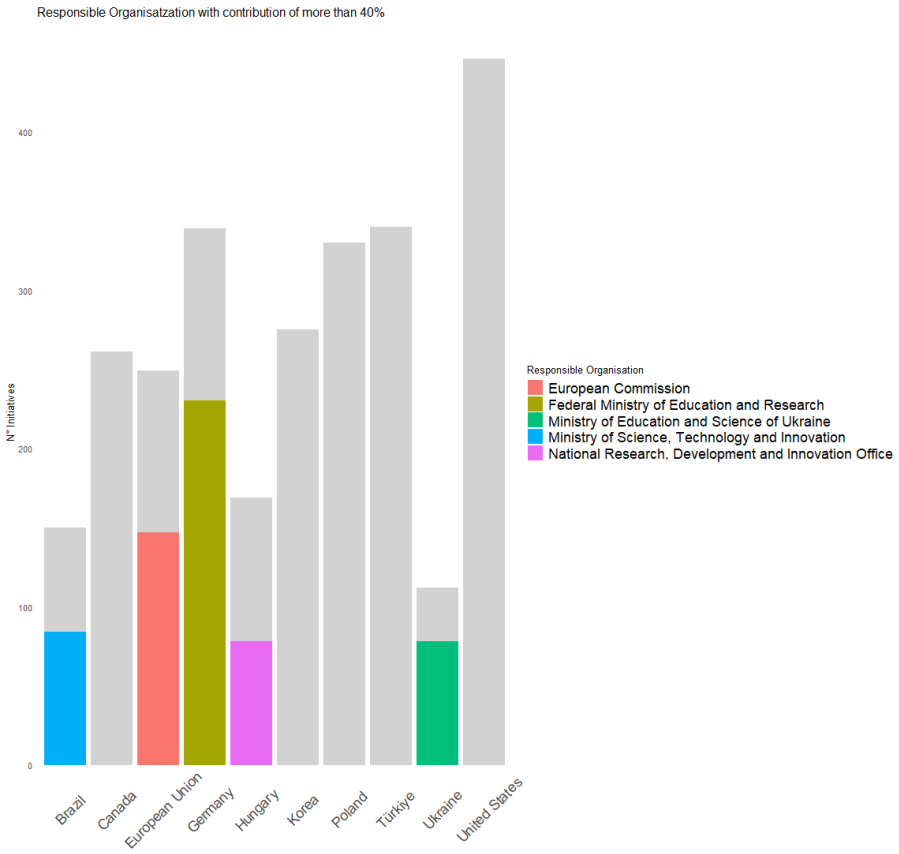


**Figure 6. Responsible Organisations with a contribution of more than 40%**

followed by 'Project grants for public research' and 'Grants for business R&D and innovation'.

Finally, we moved on to explore how several instruments are combined within an initiative. The cases of Portugal United Kingdom Brazil Luxembourg and Romania were considered, chosen respectively one for each group of countries considering the number of initiatives. For each country, a network was generated in which each node is a type of instrument and an edge between two nodes represents an initiative in which they were used in a combined manner. A weight was assigned to each edge associated with the number of times the two tools were used together. Finally, the network was distributed according to a Fruchterman-Reingold layout and a colour was assigned to the nodes according to degree value, from lightest to darkest. Parallel to the creation of these networks, few network measures of these.

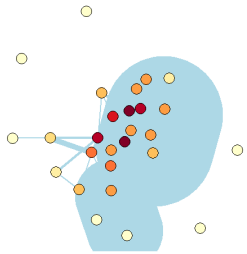

**Figure 7. Portugal, Instrument Network**

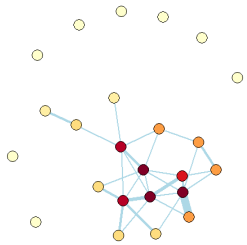

**Figure 8. UK, Instrument Network**

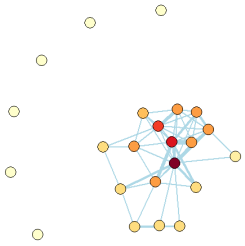

**Figure 9. Brazil, Instrument Network**

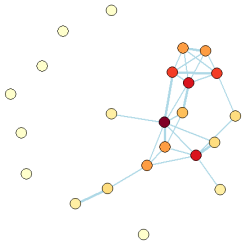

**Figure 10. Luxembourg, Instrument Network**

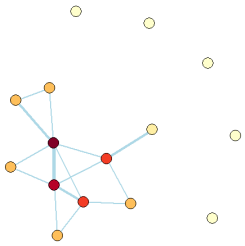

**Figure 11. Romania, Instrument Network**

Portugal's network is the largest in terms of number of nodes and edges, while Romania's network has the fewest nodes and edges. Other network measures taken into account were diameter, average path length and density: interestingly, the UK and Luxembourg are the networks with the highest diameter and average path length values, while Brazil and Portugal seem to have the densest networks. From the point of view of the number of components, Portugal has the network with the smallest number of components and the biggest 'large component' compared to the others, and this can also be seen from the fact that it has fewer isolated nodes overall: there have been few initiatives in which the combination of several instruments has not been opted for. Finally, the last measures analysed were the number of bridges, inclusivity and transitivity; in fact, the Luxembourg network has four bridges, while the Brazilian network has none at all: this can certainly be associated with the different vulnerability of these two networks, in the

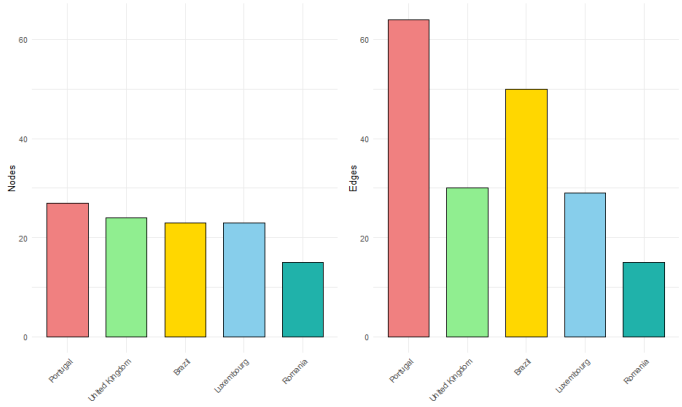


**Figure 12. Network measures 1/4**

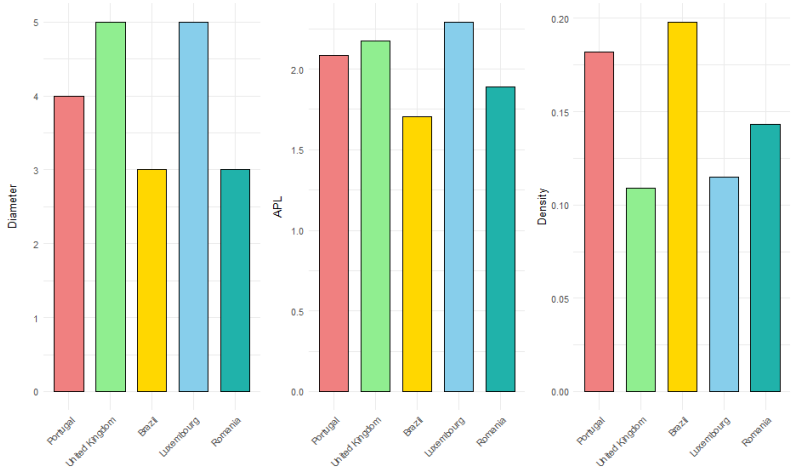


**Figure 13. Network measures 2/4**

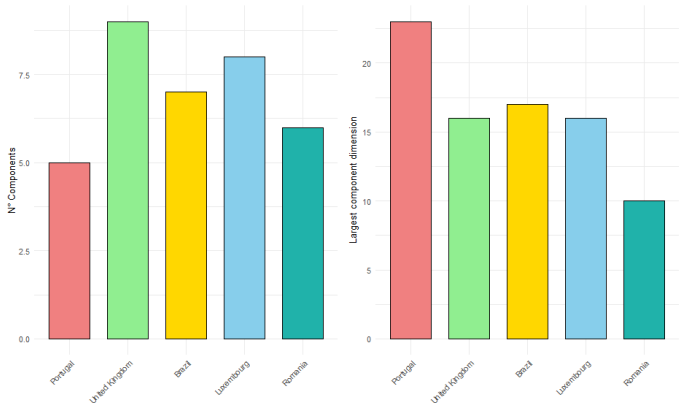


**Figure 14. Network measures 3/4**

case of Brazil there are no combinations of instruments that if removed would generate an isolated node. Portugal's network recorded the best inclusiveness value also due to the lower number of isolated nodes certainly, while Brazil's had the highest transitivity value, which again proves what was said before about the strong solidity of this network.

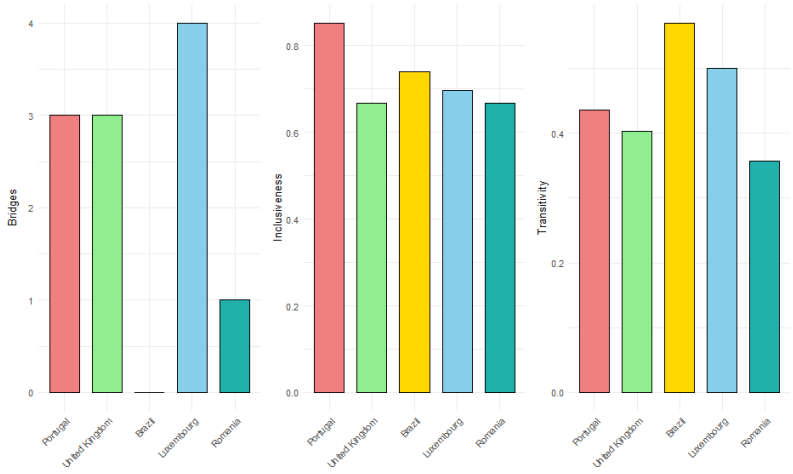


**Figure 15. Network measures 4/4**

## 4. Conclusions

The analysis of the debates revealed the validity of the 'tf-idf' approach to characterise groups of documents, as opposed to other seemingly more immediate methods.

The study of initiatives has clarified the scale and proportions at international and national level, showing how it is possible to relate countries to the number of initiatives by also considering historical data and thus identifying some sort of improvement or deterioration in their policies.

Above all, it is evident that for some countries there is a tendency to have internal organisations with a high production of initiatives.

Finally, when looking at the instruments, it emerged that most initiatives, especially for countries with a high number of initiatives, tend to promote policies by combining several types of instruments. Future developments could certainly focus on identifying these recurring patterns in the combination of multiple instruments.

## 5. Appendix

The following link provides the script in R with the analysis : https://github.com/PierDiella/STIP

## References


- EC-OECD (2023), STIP Compass: International Database on Science, Technology and Innovation Policy (STIP), edition [02-01-2024]
- EC/OECD (2023), EC/OECD Science, Technology and Innovation Policy (STIP) Survey, edition 2023, https://stip.oecd.org. [02-01-2024]
- OECD, A. Berreneche, Overarching analysis of the 2023 EC-OECD STIP Survey data, https://stiplab.github.io/R4r/main.html [02-01-2024]