Priorities for Social and Humanities Projects Based on Text Analysis^①

Ülle Must[†]

Archimedes Foundation, Väike-Turu Street 8, Tartu 51004, Estonia

Citation: Must, Ü.
“Priorities for social and
humanities projects
based on text analysis.”
*Journal of Data and
Information Science*,
vol. 5, no. 4, 2020,
pp. 116–125. [https://doi.
org/10.2478/jdis-2020-
0036](https://doi.org/10.2478/jdis-2020-0036)

Received: Jan. 20, 2020
Revised: Jun. 15, 2020
Accepted: Jul. 31, 2020

Abstract

Purpose: Changes in the world show that the role, importance, and coherence of SSH (social sciences and the humanities) will increase significantly in the coming years. This paper aims to monitor and analyze the evolution (or overlapping) of the SSH thematic pattern through three funding instruments since 2007.

Design/methodology/approach: The goal of the paper is to check to what extent the EU Framework Program (FP) affects/does not affect research on national level, and to highlight hot topics from a given period with the help of text analysis. Funded project titles and abstracts derived from the EU FP, Slovenian, and Estonian RIS were used. The final analysis and comparisons between different datasets were made based on the 200 most frequent words. After removing punctuation marks, numeric values, articles, prepositions, conjunctions, and auxiliary verbs, 4,854 unique words in ETIS, 4,421 unique words in the Slovenian Research Information System (SICRIS), and 3,950 unique words in FP were identified.

Findings: Across all funding instruments, about a quarter of the top words constitute half of the word occurrences. The text analysis results show that in the majority of cases words do not overlap between FP and nationally funded projects. In some cases, it may be due to using different vocabulary. There is more overlapping between words in the case of Slovenia (SL) and Estonia (EE) and less in the case of Estonia and EU Framework Programmes (FP). At the same time, overlapping words indicate a wider reach (culture, education, social, history, human, innovation, etc.). In nationally funded projects (bottom-up), it was relatively difficult to observe the change in thematic trends over time. More specific results emerged from the comparison of the different programs throughout FP (top-down).

Research limitations: Only projects with English titles and abstracts were analyzed.

Practical implications: The specifics of SSH have to take into account—the one-to-one meaning of terms/words is not as important as, for example, in the exact sciences. Thus, even in co-word analysis, the final content may go unnoticed.

[†] Corresponding author: Ülle Must (E-mail: ulle.must@archimedes.ee).

^① This is an extended version of a paper presented at the ISSI 2019 Conference on Scientometrics and Informetrics with a special STI Indicators Conference track. Rome, Sapienza University of Rome, 2–5 September 2019 (Must, Ü., 2019).

Originality/value: This was the first attempt to monitor the trends of SSH projects using text analysis. The text analysis of the SSH projects of the two new EU Member States used in the study showed that SSH's thematic coverage is not much affected by the EU Framework Program. Whether this result is field-specific or country-specific should be shown in the following study, which targets SSH projects in the so-called old Member States.

Keywords Text analysis; SSH; Estonian Research Information System (ETIS); Slovenian Research Information System (SICRIS); Community Research and Development Information Service (CORDIS)

1 Introduction

Much research has been written on the nature of social sciences and the humanities (SSH). It is acknowledged that their focus is often local or national, while agreed that it creates opportunities for international comparisons and exchange. At the beginning of this century, three young researchers, Jakob Edler, Stefan Kuhlmann, and Maria Behrens, stated in the book about Changing Governance of Research and Technology Policy: *“European semantics, be it in RTDI policy or other fields, is dominated by the wish to be something really big in the global arena,”* and further on: *“there is a need for clearer roles and definitions: what should be done on the European level and what on national level”* (Edler, Kuhlmann, & Behrens, 2003). It was a time of rapid expansion of the European Union. Already at that time, the EU Framework Program (FP) was the biggest funding instrument that supported research and innovation. For newcomers, it was also an excellent tool for selecting national Research and Technology Development (RTD) priorities. It happened that the similarity in choosing science and technology policy priorities was remarkable (Must, 2006). The beginning of the 1990s was also a time for newly independent countries to become acquainted with competitive research funding and to start setting up their own funding systems (Jonkers & Zacharewicz, 2016). Although the FP was not yet officially open to these countries, many researchers participated in the individual level projects. Thus, it was also one of the first options to get acquainted with competitive financing.

Even now, external funding (including FP) plays a vital role in research funding of the “new” member states. For instance, during the period 2008–2015, external funding for the Estonian R&D activities accounted for 9–12% of the total Gross domestic expenditure on research and development (GERD) (which included participation in the Framework Programmes but did not include Structural Funds like in some other countries). In 2015, 12% of Estonian total GERD was funded from abroad (circa 50% of which was FP funding) (Kattel & Stamenov, 2018). Against this background, we wanted to analyze to what extent the thematic orientations of the FP have influenced social and humanities projects in selected

Research Paper

countries. It is known that SSH is characterized by individuality and locality—both in terms of the research object and the terminology. We chose Slovenia and Estonia as the countries of comparison because of their overall good performance in the FP and well-developed national Research Information Systems (RIS). We decided to use a full-text analysis to conduct the study. A similar method has been used by van den Besselaar (van den Besselaar, 2016) in text analysis for deriving quality indicators of project proposals. Lexical analysis has also been used in FP Gender Equality Interim Evaluation (de Cheveigné et al., 2017). This survey used funded project titles and abstracts derived from the EU FP, Slovenian, and Estonian RIS. As the texts used (especially in case of national funding) were not written in most cases by English native speakers, we were aware of the risks involved. For some domains, this information may not be enough to achieve high performance in text mining tasks (Gonçalves et al., 2018). Also, the bias risk caused by abstract’s quality and completeness exists (Gómez-García et al., 2017). The full text analysis has shown that substantial differences in profile and orientation can occur within the categories, such as methodological or empirical research (Glenisson et al., 2005).

Thus, we aimed to investigate the SSH thematic pattern through three funding instruments during 2007–2018:

- a) **SSH oriented EU Framework Programme projects** (FP 7 Cooperation. Theme 8: Socio-economic Sciences and Humanities and HORIZON 2020 Societal Challenge 6. Europe in a changing world—inclusive, innovative and reflective societies);
- b) **Slovenian SSH projects** funded by the Slovenian Research Agency, and
- c) **Estonian SSH projects** funded by the Estonian Research Council (until 2012 the Estonian Science Foundation).

2 Methods

We used publicly available tools (datasets and software) to conduct the survey (data collection and analysis).

2.1 Data collection

Table 1. SSH funded projects in 2007–2018.

Source	No. of Projects
FP 7 Cooperation. Theme 8: Socio-economic Sciences and Humanities.	255
HORIZON 2020 Societal Challenge 6. Europe in a changing world—inclusive, innovative and reflective societies (as of March 2018)	277
The Slovenian Research Information System (SICRIS)	663
The Estonian Research Information System (ETIS)	489

The Community Research and Development Information Service (CORDIS) is the European Commission's primary source of results from projects funded by the EU framework programs for research and innovation (FP1 to Horizon 2020). From among collaborative projects, we searched by the thematic programs: "*FP 7 Cooperation. Theme 8: Socio-economic Sciences and Humanities*", and "*Horizon 2020 Societal Challenge 6. Europe in a changing world—inclusive, innovative, and reflective societies*". For this funding instrument, we have to bear in mind that this is a top-down funding model where the topic results from a long-term consultation process involving experts from all participating countries. There has been a huge competition in these thematic programs throughout history (2007–March 2018, the average success rate was 9%). Thus, the number of people affected by the Framework Program is much wider than the number of successful participants may show. In the case of Estonia, for example, 535 organizations participated as applicants, of which 59 were successful, and 641 organizations from Slovenia, of which 39 were successful.

The Estonian Research Information System (ETIS) collects data about Estonian research institutions and researchers working in Estonia (CVs, publications, supervisions, patents, research projects, and contracts). Data are available since 2006. All Estonian national research financing is processed via ETIS: researchers submit applications (and later financial or final reports) and the funding bodies process them. We performed a search by the project field and the Frascati Manual specialties "*Social Sciences*" and "*Humanities and the Arts*." From the collection, we selected research projects funded by the Estonian Research Council (until 2012 the Estonian Science Foundation).

The Slovenian Research Information System (SICRIS) collects data about research organizations, research groups, researchers, research projects, research and infrastructure programs, and research equipment. The Projects database contains data on projects partly financed by the Slovenian Research Agency from 1998 onwards. The project classification scheme is based on CERIF. We performed searches using the science categories "*Social Sciences*" and "*Humanities*." Due to the language barrier, we could not analyze all the projects funded: in the humanities, 251 of the 291 funded projects (86.25%), and in the social sciences, 221 of the 372 funded projects (59.4%) were with English summaries.

2.2 Data cleaning

Table 2. The number of words in ETIS, SICRIS, and CORDIS projects titles and abstracts.

Database	Total no. of words	No. of words after cleaning	Proportion of Top 200 words from cleaned words (%)
ETIS	9,132	4,854	46
SICRIS	12,813	4,421	45
CORDIS	28,819	3,950	44

It is said that data pre-processing, acquisition, and cleansing jointly represent up to 80% of the overall effort distribution from a text analysis survey (Glenisson et al., 2005). Pre-processing steps include the removal of punctuation marks, numeric values, articles (a, the), prepositions (on, at, in), conjunctions (and, or, but) and auxiliary verbs, such as “to be” (*am, are, is, was, were, being*), “do” (*did, does, doing*), “have” (*had, has, having*). In the current survey, we did not remove word suffixes, such as plurals, verb tenses, and deflections during the first phase, because in English, nuances also play a role in differentiating the content.

The final analysis and comparisons between different datasets were made based on the 200 most frequent words.

3 Findings

After removing punctuation marks, numeric values, articles, prepositions, conjunctions, and auxiliary verbs, 4,854 unique words in ETIS, 4,421 unique words in SICRIS, and 3,950 unique words in FP were identified (Table 2). Co-occurrence analyses were made on the basis of the top 200 words. Across all funding instruments, about a quarter of the top words constitute half of the word occurrences (Figure 1).

Figure 1. The curve of the Top 200 most frequent words of projects in FP, Estonia, and Slovenia during 2007–2018.

Word frequency is an important measure in content analysis. This measure is used to identify the most important research topics or concepts in a field by focusing on the most frequently occurring words (Milojević et al., 2011). As one aim of this paper was to examine to what extent FP affects national programs, we can see from Figure 2 that in the majority of cases words do not overlap. There is more overlapping between words in the case of SL and EE (20.85%), and also in the case of SL and FP (15.6%), in the case of EE and FP it is almost half of these (8.05%).

Figure 2. The overlapping and unique words in Top 200 of CORDIS, ETIS, SICRIS.

As stated by Milojević et al. (2011), all words are specific or nonspecific to some degree, depending on the context. Adjectives and nouns play a major role in understanding the content of the text. Three types of words were distinguished in the sample. The first group consists of the so-called **project classics**. Since project preparation is subject to certain standards, there are terms in the text that are not related to the content but are at the same time necessary to achieve the given criteria (*project, research, study, analysis, focus, deliver, evaluate, implement, network, develop, publish, area, findings, process, platform, etc.*). They make up the majority of the top 200 words. The most common pairs of words are formed from them (*research project, proposed project, proposed research, long term, project aims*). The second group is **content words** which form the core of the projects and enable them to follow the research trends in the given time frame. It is worth considering in the future to analyze data based solely on content words. **Geographic locations** form a separate group. It looks like good practice of project writing to mention the target area (Europe, Estonia, and Slovenia). At the same time, research topics are influenced by the past and proximity—in the case of Estonia, top 200 words include “Russia,” “Baltic,” “Livonia,” and “German”; “Yugoslavia” in the case of Slovenia; “Mediterranean,” “China,” and “Africa” in the case of FP.

Research Paper

From the point of view of our study, the content words deserve closer examination. There is a definite set of content words that overlap in all datasets throughout the period: *culture, education, environment, history, human, identity, innovation, national, policy, social*.

Throughout the period some words overlap between different databases: EE-FP (*age, east, public*), SL-EE (*individual, language, literature*). For the most part, overlapping words in different periods are specific to a database. In the case of Estonia: *children, school, student, linguistic*; in the case of Slovenia: *art, legal, spatial*; in the case of FP: *citizen, employment, migration, mobility, urban, young*. Unique words form a separate group, in the case of Estonia: *dialogue, interdisciplinary, music, semiotics, teacher, ancient, collection, genetic, infrastructure, medieval*; in the case of Slovenia: *family, minority, territory, tradition, values, memory, tourism, war*. The largest number of unique words appears in FP where periods also differ. FP7 (2007–2013): *carbon, exclusion, foresight, gender, humanities, lifestyle, peace, poverty, rural, security, SSH, transition, unemployment, welfare*; H2020 (2014–20): *investment, job, justice, cohesion, crises, emergence, inclusion, inequalities, reflective, responsible, transparency*.

However, using co-word analysis, we see that the meaning of the overlapping words varies from one database to another (Figure 3). Thus, taking for example one of the most commonly used words “culture,” we see differences throughout different datasets. The most commonly used co-words in the case of FP are “*heritage*,” “*scientific*,” “*program*,” “*Europe*,” “*identity*,” in case of Slovenia: “*national*,” “*media*,” “*relations*,” “*religion*,” “*history*,” “*saints*”; in case of Estonia: “*narrative*,” “*conflicts*,” “*semiotic*,” “*social*,” “*history*,” “*Estonian*,” “*memory*,” etc. This also applies to words which show the geographic location. For example, the word “Europe” is related in the Estonian dataset to “*Central Eastern European*,” “*union*,” “*public administration*”; in the Slovenian dataset to “*political*,” “*court*”; in FP dataset to “*union*,” “*research*,” “*policy*,” and “*integration*.”

At the same time, we have to take into account the specifics of SSH again and again—the one-to-one meaning of terms/words is not as important as, for example, in the exact sciences. Thus, even in co-word analysis, the final content may go unnoticed. For example, one of the most commonly used co-words with “culture” was “landscape.” Going back to the original data, we selected out projects with these keywords. The words were interpreted in many ways: “*Where Land Meets the Sea. Maritime Cultural Landscapes in Prehistoric and Medieval Estonia*,” or “*Positioning life-writing on Estonian literary landscapes*,” or “*The Politics of Peace and Conflict Knowledge: Syria and the Diverse Landscape of Local Knowledge/Experience*,” or “*Orthodox People in Estonia and Orthodox Churches in Estonian Landscape (18th-21st Century)*.” Projects belonged to a variety of fields, starting from

Figure 3. The most commonly used co-words in the three datasets for the terms “Culture” and “Innovation.”

archaeology, cultural anthropology, ethnology, general and comparative literature, literary criticism, and literary theory, political and administrative sciences, ending with social geography.

4 Conclusions

In this study, we assume that textual analysis is one way to track changes over time. As one aim of this paper was to examine to what extent FP affects national programs, the results show that in the majority of cases words do not overlap. In some cases, it may be due to using different vocabulary. There is more overlapping between words in the cases of SL and EE, and less in the cases of EE and FP. At the same time, overlapping words indicate a wider reach (culture, education, social, history, human, innovation, etc.) than, for example, unique words in national projects (in Estonia: ancient, medieval, semiotics, etc.). We also concluded that there are words in the project applications that do not show thematic focus but are part of the obligatory vocabulary of the project writer (project, research, analysis, study, etc.). They make up the majority of the top words. The content words form the core of

the projects and enable to follow the research trends in the given time frame. In the case of national databases, it was relatively difficult to observe the change in thematic trends over time. More specific results emerged from the comparison of the different programs throughout FP. The text analysis of the SSH projects of the two new EU Member States used in the study showed SSH's thematic coverage is not much affected by the EU Framework Program. Whether this result is field-specific or country-specific should be shown in the following study, which targets SSH projects in the so-called old Member States.

References

- de Cheveigné, S., Knoll, B., Bustelo, M., Engebretsen, E., & Sandström, U. (2017). Interim Evaluation: Gender equality as a crosscutting issue in Horizon 2020. Luxembourg: Publications Office of the European Union.
- Edler, J., Kuhlmann, S., & Behrens, M. (Ed.) (2003). Changing Governance in European Research and Technology Policy. Possible Trajectories and the European Research Area. In *Changing Governance of Research and Technology Policy: The European Research Area*. Edward Elgar Publishing Ltd.
- Estonian Research Information System (ETIS). Manual for Research Staff. Retrieved from: https://www.etag.ee/wp-content/uploads/2012/05/ENG2_User_manual_long.pdf
- Glenisson, P., Glanzel, W., Janssens, F., & De Moor, B. (2005). Combining full text and bibliometric information in mapping scientific disciplines. *Information Processing & Management*, 41(6), 1548–1572.
- Gómez-García, F., Ruano, J., Aguilar-Luque, M., Alcalde-Mellado, P., Gay-Mimbrera, J., Hernández-Romero, J.L., Sanz-Cabanillas, J.L., Maestre-López, B., González-Padilla, M., Carmona-Fernández, P.J., Vélez García-Nieto, A., & Isla-Tejera, B. (2017). Abstract analysis method facilitates filtering low-methodological quality and high-bias risk systematic reviews on psoriasis interventions. *BMC Medical Research Methodology series*, 17, 180. <https://doi.org/10.1186/s12874-017-0460-z>
- Gonçalves, C., Iglesias, E.L., Borrajo, L., Camacho, R., Seara Vieira, A., & Gonçalves, C.T. (2018). LearnSec: A Framework for Full Text Analysis. In: de Cos Juez F. et al. (eds) *Hybrid Artificial Intelligent Systems*. HAIS 2018. Lecture Notes in Computer Science, vol 10870. Springer, Cham. https://doi.org/10.1007/978-3-319-92639-1_42
- Jonkers, K., & Zacharewicz, T. (2016). Research performance based funding systems: A comparative assessment, publications office of the European Union, Luxembourg. EUR 27837 EN; doi:10.2791/659483
- Kattel, R., & Stamenov, B. (2018). RIO Country Report 2017: Estonia. EUR 29153 EN, Publications Office of the European Union, Luxembourg, 2018, ISBN 978-92-79-81198-2, doi:10.2760/312784, JRC111256
- Milojević, S., Sugimoto, C. R., Yan, E., & Ding, Y. (2011). The cognitive structure of library and information science: Analysis of article title words. *Journal of the American Society for Information Science and Technology*, 62(10), 1933–1953.

- Must, Ü. (2006). “New” countries in Europe—Research, development and innovation strategies vs bibliometric data. *Scientometrics*, 66(2), 241–248.
- Must, Ü. (2019). Priorities for Social and Humanities Projects Based on Text Analysis. In *Proceedings of the ISSI 2019 Conference on Scientometrics and Informetrics with a special STI Indicators Conference track*. Rome, Sapienza University of Rome 2–5 September 2019, 2, 2555–2556.
- Slovenian Current Research Information System (SICRIS). Retrieved from: <https://www.sicris.si/public/jqm/cris.aspx?lang=eng&opdescr=home&opt=1>
- The Community Research and Development Information Service (CORDIS). Retrieved from <https://cordis.europa.eu/projects/en>
- Van den Besselaar, P.A.A. (2016). Predicting panel scores by linguistic analysis. In *Peripheries, Frontiers and Beyond: Ismael Rafols et al, Peripheries, Frontiers and Beyond; Proceedings STI 2016* (pp. 1081–1088).

This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivs License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).