The Software Engineering Observatory Portal

Mívian M. Ferreira[1], Bruno L. Sousa[1], Kecia A. M. Ferreira[2], and Mariza A. S. Bigonha[1]

[1] *mariza, bruno.luan.sousa, mivian.ferreira@dcc.ufmg.br*
Federal University of Minas Gerais, Dept. of Computer Science, Minas Gerais (Brazil)

[2] *kecia@cefetmg.br*
Federal Center for Technological Education of Minas Gerais, Dept. of Computing, Minas Gerais (Brazil)

Abstract
Software Engineering is a 52 years vast field of knowledge. Along the time, a massive quantity of techniques, methods, and tools have been proposed by a vast academic and industrial community, involving a large spectrum of subjects. Aiming to compile the body of knowledge of Software Engineering, we created "The Software Engineering Observatory Portal". The portal allows analyzing software engineering evolution and researchers' contributions through data visualization. So far, the portal relies on data of the International Conference on Software Engineering (ICSE) and IEEE Transactions of Software Engineering (TSE), the premier venues in the area. A video explaining the usage of the portal is available at **https://youtu.be/57Y80z9ymVw**. The portal is available at **https://mivianferreira.github.io/docs/TheSoftwareEngineeringObservatoryPortal/**.

Introduction
Over the five decades of Software Engineering, a great deal of knowledge was produced. Some initiatives have been done to compile knowledge about Software Engineering and its evolution. An essential contribution in this context is the Guide to the Software Engineering Body of Knowledge (SWEBOK) of the IEEE Computer Society (Bourque et al., 2002; IEEE Computer Society et al., 2014). In a lecture at ICSE 2018, Briand Radell, the president of the NATO Software Engineering Conference (1968-1968), highlighted the importance of knowing Software Engineering history. He made a five-decade retrospective of the area and declared: "So I hope you'll forgive me for choosing to focus primarily on historical issues, but I hope these observations encourage at least some of you to pay a little more attention to the past". Besides knowing the past, it is essential to analyze the evolution of an area of human knowledge, including Software Engineering. Such analysis may aid in answering questions such as: *What subjects the academic community has mostly considered? How much effort has been made in research on such subjects? Which papers have been most cited by researchers? Who are the researchers that have extensively contributed to Software Engineering advances? Which subjects have been mainly investigated currently?*

We constructed a web-based tool called "The Software Engineering Observatory Portal" that aims to provide resources for Software Engineering evolution analysis through data visualization. So far, we based the portal on works published in the premier Software Engineering venues: the International Conference of Software Engineering (ICSE) and the IEEE Transactions on Software Engineering journal (TSE). We retrieved data of research papers published in Transactions on Software Engineering from 1975 to 2018, accomplishing 3,357 papers. From ICSE, we gathered data of papers published from 1988, the first year IEEE Xplore provides data of the conference, to 2018. As far as we know, this is the first initiative to compile historical data on Software Engineering with data visualization techniques. The audience of "The Software Engineering Observatory Portal" is the Software Engineering community as a whole. The portal may grow in many ways, such as: by providing other kinds of data visualizations, by considering data from other venues, and by providing updated data. Besides, the portal may be extended to other areas since the interest on investigating academic research status is global (Ioannidis et al., 2020). A video explaining

the usage of the portal is available at **https://youtu.be/57Y80z9ymVw**. The portal is available at **https://mivianferreira.github.io/docs/TheSoftwareEngineeringObservatoryPortal**/.

This paper presents "The Software Engineering Observatory Portal", its main features and its construction. Moreover, this article details how the data were retrieved, describes and exemplifies the visualizations provided by the portal, and presents the conclusion of the work.

Method

We built the database applied in portal by retrieving the metadata of ICSE (data from 1988 to 2017) and TSE (1975-2017) documents. To do so, we constructed a web crawler. We collected only documents available in the IEEE Xplorer digital library, as this library presents a complete metadata structure about ICSE and TSE documents. To include a document returned by the crawler in the database, we used the following criteria: full articles, short papers, research in progress, and works available in electronic format. We carried out the exclusion of documents from the database according to the following criteria: tutorials, panels, lectures, and keynote talks; call for workshops and symposium; proceedings and round tables; presentation of sessions and tracks. We stored the collected data in a CSV file that has the following fields: title of the work; name of the authors (separated by semicolons); affiliation of the authors (separated by semicolons); year of publication; the number of downloads; keywords of the authors; keywords IEEE, NON-INSPEC e INSPEC; publication venue (ICSE or TSE).

Data Pre-processing

After retrieving the documents, we performed a pre-processing stage of the data. At this stage, we carried out the following steps:

1. Data pre-processing: we removed the documents that met the exclusion criteria and mined the keywords not returned by the web crawler. To do so, we performed a manual inspection of the articles.

2. Keywords standardization: in this step, we performed a manual inspection of the data and joining keywords with the same semantic meaning as a single keyword. This step is important because the keywords are used in the portal to classify the subjects investigated by each work. Not performing this step will introduce bias in the data. We constructed a tool to automate this step.

3. Disambiguation of authors' names: the name of an author may appear in diverse ways in the papers. This step is essential to avoid introducing bias to the authors' data. We performed a semi-automatic disambiguation process by implementing scripts to pre-process the data. We intend to develop an automatic disambiguation process to make the portal update faster and less laborious.

4. Papers results classification. The portal provides visualization based on the kind of software engineering approach presented in the works. This step consisted of setting a category for each study by reading its abstract. We used the categories defined by Mary Shaw (Shawn, 2013) to classify the studies' results. The first and the second authors carried out his classification. After that, the results were discussed among the four authors to mitigate threats to this study. This process took a significant amount of time. Due to this, so far, the portal only provides such data for ICSE papers.

Visualization Techniques

We used D3.js, CanvasJS library, Javascript, and HTML to construct the portal. It provides five types of data visualization, described as follows.

- Data Table: we chose this visualization because it is considered the most appropriate to identify the most cited ICSE and TSE Engineering papers since their titles are usually long. Besides, we used the table to present other important data related to the articles, such as authors, number of citations, year of publication, and link to the article in the *IEEE Xplorer* database. The table is presented as a search engine, ordering (increasing and descending) of all fields, paging, and choice of entries viewed per page, ensuring a good usability.

- Dispersion Plot: we used the dispersion plot to show data of authors whose contributions have a significant impact. We consider that the author's contribution is determined by the number of citations of an author and the number of articles published by that author – thus establishing the need to use the dispersion plot. To determine the authors whose contributions have a high impact, we represented each author by a circle, whose area is proportional to the number of citations. i.e, the greater the number of citations of an author, the greater the circle's size. The x and y axes represent the number of citations of an author and the number of articles published by an author, respectively. To facilitate the visualization of the data, we subdivided the authors into three categories: All, with all authors; Top-10, with only the ten authors with the highest number of citations; and Top-100, with the 100 authors with the highest number of citations. Besides, there is a tool-tip in each of the circles where we displayed the author's name, number of citations, and articles.

- Keywords Cloud: this visualization shows a cloud with the subjects investigated in the area. In this cloud, the larger the font size of a keyword label, the larger the number of works considering the related subject. This visualization allows researching keywords according to a time frame chosen by the user. For instance, the user may want to see the cloud of works published from 2015 to 2018. When the user move the cursor on the word, the portal shows the corresponding number of papers. Hence, when the user clicks the word, the portal shows a data table with all the works in which the chosen keyword appears.

- Overlapping areas: we used this visualization to show the amount of research done on the top subjects that have been investigated in Software Engineering . In this visualization, each layer represents a keyword. The x and y axes represent the years used for data analysis and the number of occurrences of the keywords, respectively.

- Pareto Chart: we developed this visualization with CanvasJS. The portal provides a Pareto's (80/20) chart of the following data: number of citations/article, number of citations/author, and number of articles/author.

Results

In this section, we exemplify the use of "The Software Engineering Observatory Portal".

Figure 1. Keywords cloud of ICSE with data from 2015 to 2018.

The most investigated subjects in a given period.
The user selects the range of years and the portal shows the most investigated subjects in a keyword cloud. Figure 1 shows the keywords cloud of 1988 until 1990.

The most cited papers
The number of citations of a paper is an indicator of its relevance. The portal provides a table with data of the publications, including the number of citations. Figure 2 shows an example of this table.

Title	Authors	Year	#Citation
Recovering documentation-to-source-code traceability links using latent semantic indexing	A. Marcus; J.I. Maletic	2003	269
Experiments on the effectiveness of dataflow- and control-flow-based test adequacy criteria	M. Hutchins; H. Foster; T. Goradia; T. Ostrand	1994	268
Visualization of test information to assist fault localization	J.A. Jones; M.J. Harrold; J. Stasko	2002	225
Patterns in property specifications for finite-state verification	M.B. Dwyer; G.S. Avrunin; J.C. Corbett	1999	209
Mining version histories to guide software changes	T. Zimmermann; P. Weibgerber; S. Diehl; A. Zeller	2004	203

Figure 2. Table with the data of the most cited ICSE papers.

The researchers that have mainly contributed to Software Engineering advance. The portal provides visualizations of data about the number of articles and the number of citations by author. The visualization is a Dispersion plot, as shown in Figure 3. Figures 4 shows the Pareto chart of citations per article of ICSE and TSE, respectively. When the user move the mouse over the curve, the portal shows the cumulative number of articles corresponding to the percentage of citations.

The most investigated subjects along with Software Engineering evolution. The 11 most researched topics in ICSE were: test (236 occurrences); bugs (147 occurrences); software architecture (132 occurrences); measure/metric (87 occurrences); requirements (85 occurrences); education (74 occurrences); empirical study (59 occurrences); software quality (56 occurrences); software process(54 occurrences); software engineering (53 occurrences); and reuse (52 occurrences). We used an overlapping areas chart to show the evolution of the number of studies on those subjects. Figure 5 shows the chart for ICSE.

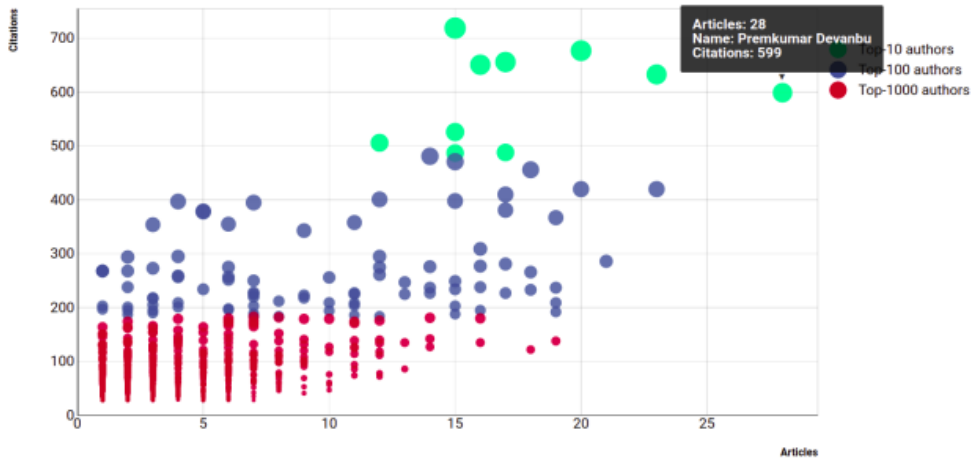

Figure 3. Dispersion plot with data of the 1000 most cited ICSE papers.

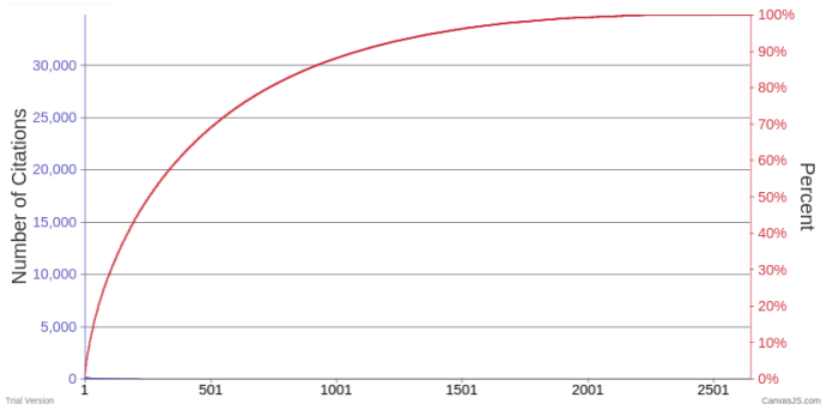

Figure 4. Pareto chart with number of citations of ICSE papers.

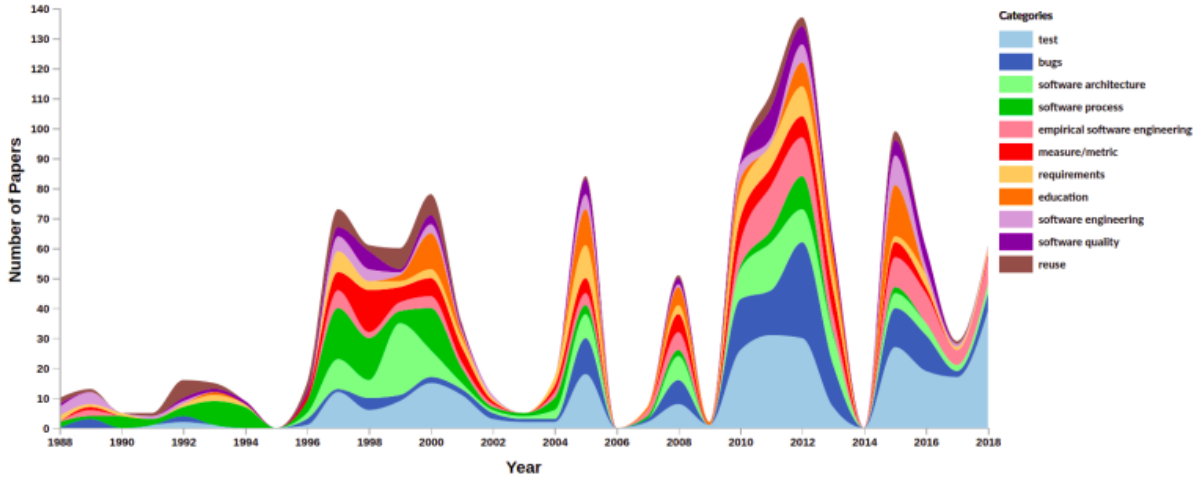

Figure 5. Overlapping area showing the evolution of number of papers considering the top-10 most investigated topics - data of ICSE.

Related Work

Visualization is a powerful resource to aid bibliometrics analysis. The study of Muñoz et al. (2020) shows that there are many tools for this purpose. The main difference between those tools and our portal is that the portal is an on-line and interactive tool containing pre-processed data of an specific area, Software Engineering. Besides, the visualizations resources provided by the portal are also different. Data visualization charts, such as co-autorship networks and keywords density map, were used by Liao et al. (2018) to analyze bibliometrics data in Medicine. The visualization they used are different from the ones we applied. Moreover, they did not constructed a portal as we did. Other initiatives have been taken to compile the body of knowledge on Software Engineering, such as the SWEBOK (Bourque et al., 2002; IEEE Computer Society et al., 2014) and the Boehm's work (Boehm, 2006). However, as far as we know, this work is the first one that compiles Software Engineering historical data by applying visualization techniques. We used the Overlapping Area chart in our work on Software Engineering evolution considering ICSE data (Sousa et al., 2019).

Conclusions

This paper presents a web-based tool called "The Software Engineering Observatory Portal" that provides resources for analyzing the evolution of Software Engineering, as well as data of contributions, number of citations, and subjects primarily investigated in the area. The portal applies data visualization and relies on papers publishes in ICSE and TSE, the premier software engineering discussion forums. The analysis provided by the portal is of interest to the Software Engineering community as a whole, specially the academia. As part of the pre-processing data was done manually and is exceptially laborious, we are developing an automatic tool to construct a thesaurus of Software Engineering and adapting the gathering data process to the new structure of IEEE Xplorer. Besides, we are investigating a proper approach to disambiguate authors' names. We believe that the portal could be applied to other Computer Science disciplines whose publications are available in IEEE Xplorer.

Acknowledgments

This work was supported by CAPES, CNPq and CEFET-MG.

References

B. Boehm. "A view of 20th and 21st century Software Engineering". In *Proceedings of the 28th International Conference on Software Engineering (ICSE '06)*, 2006, p. 12–29.

B. L. Sousa, M. M. Ferreira, K. A. M. Ferreira, and M. A. S. Bigonha, "Software engineering evolution: The history told by ICSE", in XXXIII *Brazilian Symposium on Software Engineering*, p. 17-21. 2019.

H. Liao, M. Tang, L. Luo, C. Li, F. Chiclana, and X. Zeng."A bibliometric analysis and visualization of medical big data research". Sustainability, vol. 10(1), p 166, 2018.

IEEE Computer Society, Pierre Bourque, and Richard E. Fairley. (2014). Guide to the Software Engineering Body of Knowledge (SWEBOK(R)): Version 3.0. *IEEE Computer Society Press.*

J. P. A. Ioannidis, K. W. Boyack, and J. Baas, "Updated science-wide author databases of standardized citation indicators", *PLoS Biol*, vol. 18(10): e3000918, 2020.

J. M. Muñoz, E. H. Viedma, A. S. Espejo, M. J. Cobo. "Software tools for conducting bibliometric analysis in science: An up-to-date review". El profesional de la información, v. 29, n. 1, 2020.

P.Bourque, J.-M. Lavoie, A. Lee, S. Trudel, T. C. Lethbridge *et al.*,"Guide to the software engineering body of knowledge (SWEBOK) and the software engineering education knowledge (SEEK)-a preliminary mapping," *10th International Workshop on Software Technology and Engineering Practice*. IEEE Computer Society, 2002, pp. 8–8.

M. Shaw, "Writing good software engineering research papers: Minitutorial", in *25th International Conference on Software Engineering (ICSE)*, 2003, pp. 726–736.