An architectural proposal to explore the data of a private community through visual analytic

Jorge Durán-Escudero
GRIAL Research Group,
Research Institute for Educational Sciences, University of Salamanca
37008 Salamanca, Spain
jorge.d@usal.es

Francisco J. García-Peñalvo
GRIAL Research Group,
Research Institute for Educational Sciences, University of Salamanca,
37008 Salamanca, Spain
fgarcia@usal.es

Roberto Therón-Sánchez
GRIAL Research Group,
Faculty of Sciences,
University of Salamanca
37008 Salamanca, Spain
theron@usal.es

ABSTRACT

In this document*, a proposal is made to study the data that will be generated in the private and anonymous community of the WYRED project, in order to extract knowledge about how their users interact, both between them, and with the platform. To do this, it is started with the creation of a system that will generate a set of test data, as close as possible to the original. With this information and considering the impact of privacy when dealing with the data of the project, a flexible and complete architecture has been proposed for the development of interactive visualizations that will allow to visualize the previously generated data. Finally, a use case is presented where the suitability of the visual analytic is demonstrated to perform analysis of the data of the project and to extract knowledge, in a simple way.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI) → Interactive systems and tools**

KEYWORDS

Visual Analytic, Social Network, Software Architecture, Data Generation, Users Interaction, Interaction in Social Networks, WYRED

1 INTRODUCTION

Today, social networks are one of the types of communities that are experiencing higher growth, thanks to the wide diffusion of the technologies of information and communication [1]. However, they continue presenting some problems, like the management of the privacy or the analysis of the data, to increase the knowledge of what is happening within them. In addition, experts find that, due to the volume of information that they generate, it is not currently possible to realize manual analysis of what occurs in them.

The WYRED project consists in the development of a technological ecosystem [2], in order to know in greater depth the interests and problems of young people, the way they have to face them and, ultimately, to be a place where their voice is heard and taken into account [3, 4].

A technological ecosystem is a set of technological elements that allow to cover all the needs of a project, for it is necessary the management of the users and the generated information, the support for the diffusion of these data, the integration with other technological ecosystems and the ability of each of these aspects to evolve to fit the project changes [5-7]. In the case of the WYRED project, this ecosystem consists of four distinct parts: a service that is responsible for anonymizing users, a private platform where dialogues with young people take place, a system for dissemination in social networks and a public web to know the project.

The architectural proposal that is going to be presented in this paper is centered in the community of WYRED, that is similar to a forum, where the users organize the discussions in communities, threads and comments, but it can also host social dialogues and research projects. However, the project has a number of characteristics that distinguish it from the others [8], such as its use in an international context (several languages, different sociological characteristics and very different points of view) or the need to safeguard users' privacy, firstly because they may be minor and secondly, because it is sought to make the platform a place where they can interact freely, for which a high degree of anonymity is required. Due to the large number of data that the project is going to generate, the use of visual analytics is proposed as an effective technique for representing and extracting knowledge [9].

The main objective of this work is to propose, in these early stages of the WYRED project, an architectural proposal of a system that allows to support the development of interactive visualizations that help to better understand the data, to anticipate the future needs of the project.

This architecture has to be flexible enough to be able to adapt to the diverse characteristics of the project, allowing

TEEM 2017, October 18–20, 2017, Cádiz, Spain

ACM ISBN 978-1-4503-5386-1/17/10…$15.00
https://doi.org/10.1145/3144826.3145398

also to build on it any type of visualization that is required, at this moment or in the future. It must help researchers in two main tasks:

- Know how the community evolves and the content that is being generated.
- Assist in the decision-making process.

Therefore, although the main topic of study of the project are the youth, the architecture will have as end users the project's researchers. In line with the final objective of the project, what is ultimately sought is to influence in the decisions of public representatives, to develop actions that help improve the lives of young people and, ultimately, to take advantage of their contributions.

This article is organized in the following sections: firstly, the WYRED project is introduced, then the architectural proposal is shown and the way in that a testing dataset has been generated, later a use case is presented followed by the main conclusions.

2 WYRED PROJECT

WYRED (netWorked Youth Research for Empowerment in the Digital society - https://wyredproject.eu/) is a European Project (Ref. 727066.) funded by the Horizon 2020 Programme in its "Europe in a changing world – inclusive, innovative and reflective Societies (HORIZON 2020: REV-INEQUAL-10-2016: Multi-stakeholder platform for enhancing youth digital opportunities)" Call.

Project that aims to provide a framework for research in which children and young people can express and explore their perspectives and interests in relation to digital society, but also a platform from which they can communicate their perspectives to other stakeholders effectively through innovative engagement processes. It will do this by implementing a generative research cycle involving networking, dialogue, participatory research and interpretation phases centred around and driven by children and young people, out of which a diverse range of outputs, critical perspectives and other insights will emerge to inform policy and decision-making in relation to children and young people's needs in relation to digital society.

WYRED aims to give young people a voice, and a space to explore their concerns and interests in relation to digital society and share their perspectives and insights to stakeholders with other strata of society.

3 ARCHITECTURAL PROPOSAL

An architectural proposal consists in defining each of the elements of a system and what is going to be the way in which they interact. This type of work becomes necessary when it is proposed to carry out a project of a certain size, since in it are present a multitude of requirements that must be fulfilled, to reach a high degree of satisfaction of the users. In case of not establishing it, there is a risk that the project will not achieve all the proposed objectives and/or the quality of the result is very low. In the case of this project, it has to support a large number of requirements, the main ones being:

- The ability to work with different data sources.
- Support to manage their privacy.
- Automatic analysis of data (as far as possible).
- The ability to represent data through interactive visualizations.

To support these requirements, it has been decided to use an architecture called microkernel [10]. This architecture is based on providing minimal functionality in the kernel, and complementing it with a set of components that are the ones that perform the tasks required by the users. This model presents a change of philosophy with respect to the layers-pattern, characterized by stacking the layers horizontally, each having a specific role within the application.

The great advantage of applying this architecture in this case is that the core will only be in charge of obtaining the data and anonymizing them, each of the components will be in charge of processing that data and perform the corresponding visualization. This also allows to achieve a very flexible architecture, where you can easily add new visualizations or eliminate any existing ones, in case the results were not satisfactory [11].

Taking the Docker architecture as a reference, due to is one of the most known examples of microkernel architecture (https://goo.gl/LGk7vj), shown in the Fig. 1, this proposal has been designed, consisting of two layers that form the micronucleus and two main layers for each of the components, which will lead to the generation of interactive visualizations, Fig. 2.

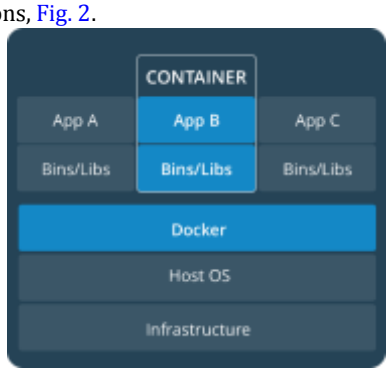

Figure 1: Docker architecture

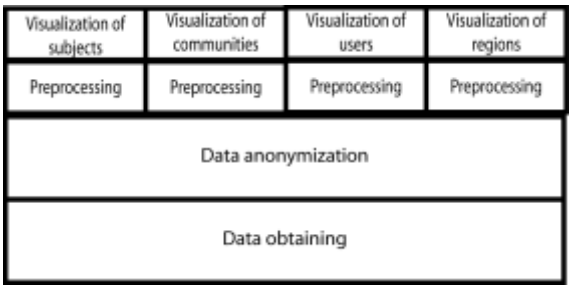

Figure 2: Architectural proposal for the WYRED project

3.1 Data sources

Obtaining data in this project involves more than just querying a database. This is because the information of the same is distributed among several services, present in several machines.

The private information of the users is stored in a CAS, Central Authentication Service (https://goo.gl/xD4Jkg), following the trail of other studies that have faced this problem [12, 13].

In the case of the public information of the users, this is part of the WYRED platform and is available in its database. Finally, the information of the users' interaction with the platform is stored in a NoSQL database, in order to satisfactorily deal with the problems of scalability [14, 15].

This layer of the microkernel, therefore, will have to be in charge of merging the data from the different data sources, in addition to the retrieval of the information.

3.2 Data anonymization

The layer responsible for anonymizing the data is of vital importance in this work, because it's handling data that contain personal information of the users. In addition, many of them are minor, so this process is obligatory to comply with the current data protection legislation.

The way to work with some of this data is simple, since issues such as name, surname or email can be eliminated without losing representative information. However, this is not enough to ensure that the data are already anonymous, since by combining the remaining data it may be possible to identify the initial user [16]. This type of data that is not unique, but has values that are not usually repeated (or its repetition rate is low) in a dataset, are called quasi-identifiers.

The proposal for the anonymization of the data consists of analyzing and detailing the quasi-identifier attributes that are going to be, and try to reduce them:

- In the case of the date of birth, it is proposed to transform this data into the year of birth. In this way, the number of users with a unique value for this field will be very small or zero.
- In the case of the place of residence, a similar process is planned, reducing the information to the province from where it takes part.

In addition to these transformations, it is proposed that the results be always k-anonymous with a value of k=2. This means that there cannot be registers with unique values, since at least, each record must have 2 users with equal values. The use of this value ensures the anonymity of the data, which will be published openly, so that other researchers can use them as a source of information in their research, as stipulated by the European Union for projects funded under the Horizon 2020 project (https://goo.gl/b24XP9).

3.3 Module for the analysis of the most frequent subjects

The analysis of the most frequent subjects is one of the questions most repeated by the different researchers. Some focus only on the temporal evolution of these, however, other researchers also consider it very important to be able to explore the use of these themes according to the characteristics of individuals (age, gender, country, etc.).

Thanks to the architecture proposed above, this module is able to access the data of the platform to be able to preprocess them. In this case, it is proposed to perform an automatic analysis of the most frequent subjects using LDA (Latent Dirichlet Allocation) [17]. One of the problems with this method is that it is not intended to work in multilingual systems, a very important issue because it is one of the characteristics of the context of use, however, some authors have proposed different methods to support it [18, 19]. Another of the handicaps of this mechanism is that it is able to group the words that are part of the same theme, but not to associate a representative name to each theme.

To carry out the visualization proposal, the first thing that has been taken into account are its main associated tasks:

- Knowing the evolution of a theme: maximum, minimum, patterns, etc.
- Being able to compare the evolution of several topics.
- Being able to know how users' attributes influence the evolution of the themes.

Considering the above, it had to select one type of chart from among the many existing [20]. Because of the importance of the temporal characteristic, the first decision was to use a representation that had a horizontal axis to show each of the temporal instants. But it was still necessary to indicate how the frequency of a subject was to be coded, for which there were several possibilities such as line graphs, areas, or histograms.

At first it was thought to use a visualization based on the concept of Theme River [21]. This system has already been used effectively in other research [22, 23], since it allows you to easily identify the most important trend changes. However, it has been shown unsuitable for detecting minor trend changes and, moreover, does not allow for a large number of subjects. To reduce these disadvantages, this representation has been combined with another one based on representing each topic individually, on parallel timelines [24]. This allows us to take advantage of the Theme River representation, when making comparisons, and of representations with parallel time lines, in order to know in greater depth the temporal evolution of the subject and to allow the representation of a greater number of them.

Regarding the interaction and adaptation capabilities of the chart, the following are proposed:

- Ability to select the themes to represent.
- Possibility of making comparisons, choosing the attribute of the users by which it is compared.
- Supporting to rearrange themes, as it is easier to compare those that are closest to each other.

- Ability to know the level of relevance of that topic at a specific time.
- Possibility to zoom automatically.
- Ability to restrict the selection to a temporary period.

3.4 Module for the detection of communities

Another of the most important aspects when exploring a community is to detect the communities implicitly created by users. For this purpose, a large number of techniques have been used, such as hierarchical clustering, the detection of central nodes or the centrality measure [25], but these requires the execution of complex algorithms. It is therefore proposed to address this task through interactive visualization.

The main task to be addressed with this visualization is to discover how users interact, in order to intuit the implicit communities that they form. For this reason, graph representation has been highlighted as the best system for visualizing this type of data [26]. This representation is composed of two main components, the nodes or vertices and the arcs or links, representing the users and their relations, respectively. In the specific case of the visualization proposed, the relationships refer to the number of comments they exchange.

Regarding the visualization, it is proposed to code each node with a size relative to the number of messages that it has published in the platform. In addition, the length of the links should represent the proximity or distance of one node relative to another, taking into account the interactions they have had. To do this, it is introduced the concept of relative distance d_r between two nodes A and B, as a value proportional to the number of total links between the number of links they share:

$$d_r(A,B) = k * \frac{g(A) + g(B)}{E(A,B)}$$

g(A), degree of A, is the number of links that have origin or destination A, and E(A, B) the number of links that share both.

Regarding the interaction characteristics implemented, to solve the task of subcommunity detection in a simple and effective way, the following have been established:

- Possibility of knowing in detail the characteristics of a user, when visiting a node.
- Ability to zoom to be able to analyze the graph in greater depth and to help that this visualization can still be useful with a high number of nodes.
- Supporting to select a set of nodes and know the average value of their attributes.
- Possibility to move and analyze in detail each of the communities that are formed.

3.5 Module for the exploration of users

The problem of representing the attributes of the users of a platform is quite complex, due to the large number of users and features to be shown. For these reasons, it is necessary to use a visualization that scales on demand, compact and easy to interpret. For this reason, the parallel coordinates have been chosen [27, 28], since they allow us to represent n dimensions or attributes in a two-dimensional context.

With respect to interaction characteristics, the following are proposed:

- Possibility of reordering the attributes to be visualized, in order to be able to detect if there is correlation between them or not.
- Ability to filter through each of the attributes, supporting multiple filtering.
- Possibility of restricting the time period to be studied.

3.6 Module for the geographical exploration of the project

This module is responsible for answering the need to know which countries are the most active and how this dimension affects the analysis of the data of the platform, the visualization that best represents this concept is the map. However, there are many types of maps, both attending to the characteristics they represent and the projection that they use. In the proposal, it has been chosen to use the Mercator projection, because it's the most familiar, to represent the countries and regions of the world. In addition, the color will be used to represent the number of messages that have been generated by the users of each of the territories. With respect to interaction characteristics, the following are proposed:

- Possibility to move around the map.
- Ability to have semantic zoom, so that when the zoom level is high, the map stops representing the countries and goes on to show their provinces.
- Support to refocus the map.
- Ability to know the exact number of messages in each country.
- Possibility to filter data by country.

4 DATASET GENERATION

One of the problems that has occurred in this work has been the lack of a dataset with which to develop an architectural proposal, due to the WYRED project's community had not enough activity at this moment. For this reason, the decision of try to generate a testing dataset as close as possible to the real datasets that the project will generate was taken. The main approaches to make it are the following:

- Use data from a similar community.
- Use other data sources that have common characteristics.
- Generate data artificially.

Extracting data from a community close to the one that is being studied is the simplest and fastest process for obtaining a set of data. This can be done using the largest social networks (Twitter, Facebook, Flickr, etc.), which have been analysed in depth by many authors who, in most cases,

have made available to other researchers their data [29, 30]. But this solution is not valid in all cases, because these are too generic communities and the data is usually anonymized.

Other authors [31] have proposed to use some data that are easier to obtain, such as entries in the log files, to generate the dataset. In such way that those characteristics that are present in the records and in the target dataset, are maintained and those that do not appear, are generated from the combination of others that do form part of them. This system has the advantage that part of the data corresponds to real information and, therefore, it is possible to study it to find patterns and verify hypotheses, while the rest of the data can serve to add context to them.

Other researchers focus their studies on generating the dataset completely, artificially. Within this field, we must highlight those who focus on simulating the interaction and those who in addition to the above, try to generate the content that would occur. In the first case, they have worked in mathematical modeling the growth and the evolution of the interactions in a network [32], which allows them to reach a set of data whose behavior is representative. In the second case, the authors face the high complexity involved in the generation of content, for example, textual type content, along with the assignment of representative attributes to each individual and their interactions. The main work that tackle this is LDBC-SNB Data Generator [33], which is a program developed to generate community datasets for LDBC (Linked Data Benchmark Council) [34]. To assign attributes to values logically, the authors rely on S3G2 [35] a framework that defines the correlation that exists between certain attributes. For the choice of values, the software has a set of dictionaries where the different values that the attributes can take are selected, selecting the final value through various functions that model the probability of an event.

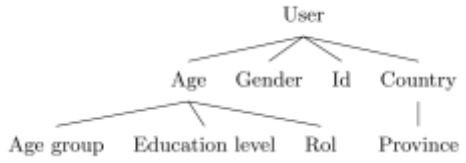

Figure 3: Dependency between the attributes of a user

At first, to build the dataset was tried to use LDBC-SNB Data Generator, however, this was not feasible, when generating a dataset that is not customizable and does not contain some of the necessary attributes. That is why it has been decided to build the dataset from scratch, for this have taken the following steps:

1. Analysis of the entities to be simulated.
2. Identification of its main attributes.
3. Creation of the dependency graph between them according to the model described in S3G2 [35]. In Fig. 3 an example of how to make this is shown.
4. Assignment of values for each attribute according to the attributes on which they depend and the values they present in the usual way. For this purpose, different indicators have been used, such as the population of a country, the expectations of use or different sociological studies [36, 37].

5 RESULTS

To develop the proposed architecture, it has been used to use web programming languages and technologies. This decision mainly allows two things: to make the developments accessible to a greater public and to equip them with a greater degree of interactivity.

At the moment of developing each module, there is an aspect that has taken great importance, the ability to filter the data that you want to study. For this purpose, controls have been established at the top for this purpose, which helps to comply with the mantra of the visual analytics enunciated by Ben Shneiderman [38] and expanded by Keim et al. [39]: Analyze first, show the important, zoom, filter and analyze further, details on demand.

The use of a modular architecture does not necessarily imply the use of each of the components separately. For this reason, they have been combined through the linked views technique [40], to form a monitoring panel that allows exploring all facets of the project, at same time, as shown in Fig. 4. The dashboard can be seen in the following URL https://goo.gl/CrBnni.

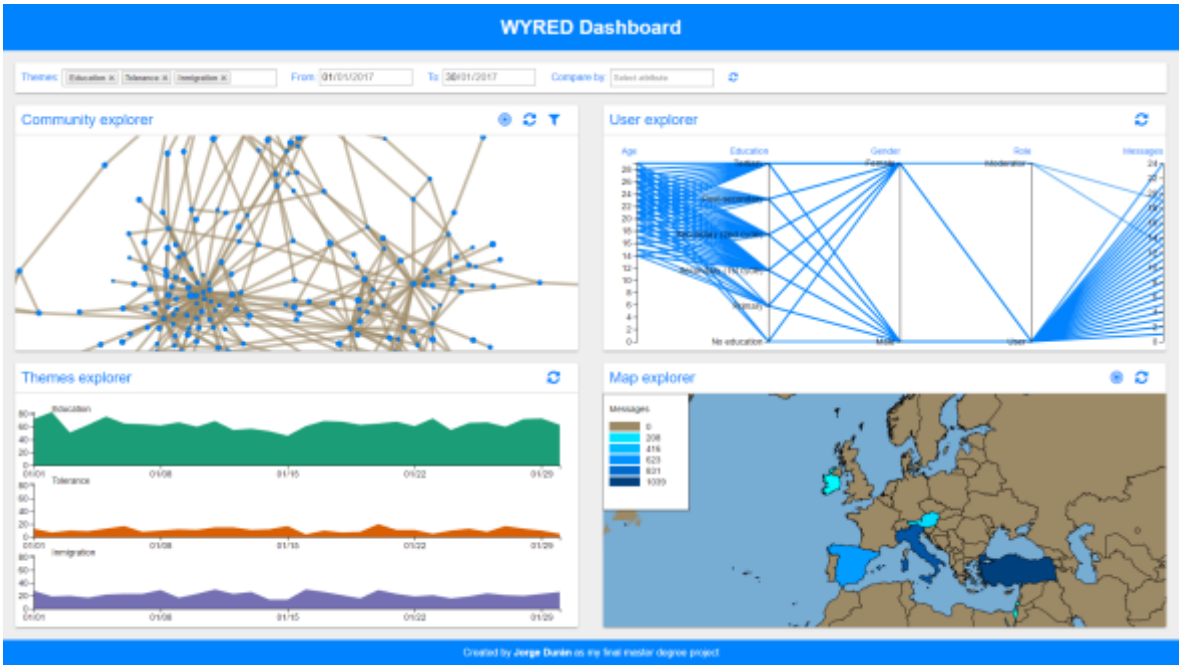

Figure 4: Dashboard to explore the WYRED's data

Figure 5: Selection of themes for the use case

To demonstrate how this proposed system could be used, a use case is described. The research question of this use case is: *what are the main communities on education and employment and what are their characteristics?* So, the first thing that a research has to do, is identify which are the themes that are presented in the research question, in this case, education and employment. For this reason, the researcher has to select them in the selector of themes (Fig. 5). Then, in the community explorer can be identified the main communities formed about these themes. This can be seen in the Fig. 6, where each point is a user and they have been grouped into three communities. To explore a community, the researcher has to select the users that form it dragging the selection rectangle that it's shown when he clicks into the community explorer. To improve the usability, the selected users maintain their color while the unselected users turns brown.

Figure 6: Identification of communities

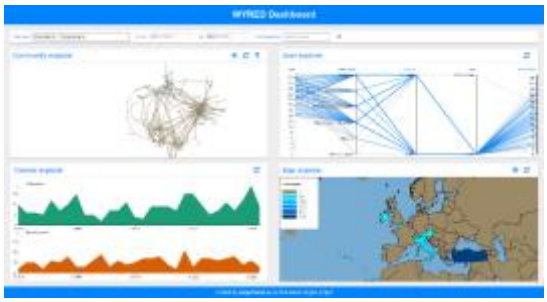

Figure 7: Data of the first community

Figure 8: Data of the second community

Figure 9: Data of the third community

If the data of the first community is analyzed (Fig. 7), it can be seen that the main users are Turkish, because this country has the darkest color in the map explorer, who talk more about education that employment, as the themes explorer shown. In the case of the second community, which data can be consulted in the Fig. 8, the users are mainly from Spain, who in the specific days talked a lot about employment, although their most common theme is education. Finally, the third community (Fig. 9) is made up by Italians whose behavior is similar to the second community. However, there is an unusual low number of postsecondary students, this can be appreciated because the majority of the lines whose destination is postsecondary are in grey, in the user explorer.

To show the interactive characteristics of the development and how a researcher can use these visualizations to solve this use case, a video has been recorded (https://goo.gl/js3hkp).

5 CONCLUSIONS

Due to the current lack of data generated by the WYRED project, the automatic generation of the data has been analyzed and a proposal has been developed to construct a dataset as similar as possible to the real datasets of the project.

In this research, the architecture proposal has also been presented to elaborate a set of interactive visualizations that allow to explore the data of the WYRED project. This modular architecture, based on the microkernel architecture, consists of 2 basic layers: data acquisition and anonymization, and 4 modules: exploration of the main themes, representation of the communities, visualization of the characteristics of the users and geographic exploration. Therefore, it can be affirmed that this work has fulfilled the goals set out in the beginning:

- Showing how to extract knowledge through interactive visualizations of the WYRED project data.
- Keeping the information anonymous.
- Analyzing the problems of working with large and complex dataset.
- Creating a flexible architecture to fit all requirements of WYRED and allowing the adaptation to the WYRED ecosystem evolution.

Regarding the future lines of research, it is considered that there are some aspects in which the research could be continued to enhance and expand this work:

- Conducting a study with users of the usability of the proposed system. To do this, users should be selected, which could be limited to 5 according to the Nielsen study [41].
- Studying and implement the collaborative use of the visualizations, so that different researchers can cooperate in the analysis, both synchronously and asynchronously [42].
- Addressing the integration of the proposed system with other systems, to favor the research work [42].

ACKNOWLEDGMENTS

With the support of the EU Horizon 2020 Programme in its "Europe in a changing world – inclusive, innovative and reflective Societies (HORIZON 2020: REV-INEQUAL-10-2016: Multi-stakeholder platform for enhancing youth digital opportunities)" Call. Project WYRED (netWorked Youth Research for Empowerment in the Digital society) (Grant agreement No 727066). The sole responsibility for the content of this webpage lies with the authors. It does not necessarily reflect the opinion of the European Union. The European Commission is not responsible for any use that may be made of the information contained therein.

This work has been partially funded also by the Spanish Government Ministry of Economy and Competitiveness throughout the DEFINES project (Ref. TIN2016-80172-R).

REFERENCES

[1] J. Schultz. 2016. Title. Micro Focus Blog. Stories and updates from our team, partners and supporters. https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day/

[2] F. J. García-Peñalvo. 2016. Technological Ecosystems. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje* 11, 1, 31-32. DOI:10.1109/RITA.2016.2518458.

[3] F. J. García-Peñalvo. 2016. The WYRED Project: A Technological Platform for a Generative Research and Dialogue about Youth Perspectives and Interests in Digital Society. *Journal of Information Technology Research* 9, 4, vi-x.

[4] F. J. García-Peñalvo and N. A. Kearney. 2016. Networked youth research for empowerment in digital society. The WYRED project. In *Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'16) (Salamanca, Spain, November 2-4, 2016)*, F.J. García-Peñalvo Ed. ACM, New York, NY, USA, 3-9. DOI:10.1145/3012430.3012489.

[5] A. García-Holgado and F. J. García-Peñalvo. 2013. The evolution of the technological ecosystems: An architectural proposal to enhancing learning processes. In *Proceedings of the First International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'13) (Salamanca, Spain, November 14-15, 2013)*, F.J. García-Peñalvo Ed. ACM,

New York, NY, USA, 565-571. DOI:10.1145/2536536.2536623.

[6] A. García-Holgado and F. J. García-Peñalvo. 2014. Architectural pattern for the definition of eLearning ecosystems based on Open Source developments. In *Proceedings of 2014 International Symposium on Computers in Education (SIIE), Logrono, La Rioja, Spain, 12-14 Nov. 2014*, J.L. Sierra-Rodríguez, J.M. Dodero-Beardo and D. Burgos Eds. Institute of Electrical and Electronics Engineers, USA, 93-98. DOI:10.1109/SIIE.2014.7017711.

[7] A. García-Holgado and F. J. García-Peñalvo. 2016. Architectural pattern to improve the definition and implementation of eLearning ecosystems. *Science of Computer Programming* 129, 20-34. DOI:10.1016/j.scico.2016.03.010.

[8] F. J. García-Peñalvo and J. Durán-Escudero. 2017. Interaction design principles in WYRED platform. In *Learning and Collaboration Technologies. Technology in Education. 4th International Conference, LCT 2017. Held as Part of HCI International 2017, Vancouver, BC, Canada, July 9–14, 2017. Proceedings, Part II*, P. Zaphiris and A. Ioannou Eds. Springer International Publishing, Switzerland, 371-381. DOI:10.1007/978-3-319-58515-4_29.

[9] D. A. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann (Ed.). 2010. *Mastering the Information Age Solving Problems with Visual Analytics*. Eurographics Association, Goslar, Germany.

[10] R. Richards. 2015. *Sofware architecture patterns*. O'Reilly Media, Sebastopol, CA.

[11] K. Matković, W. Freiler, D. Gračanin, and H. Hauser. 2008. ComVis: A coordinated multiple views system for prototyping new visualization technology. In *12th International Conference Information Visualisation, IV08*, London, 215-220. DOI:10.1109/IV.2008.87.

[12] Z. A. Pardos and K. Kao. 2015. MoocRP: An open-source analytics platform. In *2nd ACM Conference on Learning at Scale, L@S 2015* Association for Computing Machinery, Inc, 103-110. DOI:10.1145/2724660.2724683.

[13] F. Huang, C. X. Wang, and J. Long. 2011. Design and Implementation of Single Sign on System with Cluster CAS for Public Service Platform of Science and Technology Evaluation. In *2011IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications*, 732-737. DOI:10.1109/TrustCom.2011.95.

[14] J. Pokorny. 2013. NoSQL databases: a step to database scalability in web environment. *International Journal of Web Information Systems* 9, 1, 69-82. DOI:https://doi.org/10.1108/17440081311316398.

[15] R. Cattell. 2011. Scalable SQL and NoSQL data stores. *ACM SIGMOD Record* 39, 4, 12-27. DOI:10.1145/1978915.1978919.

[16] L. Sweeney. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 5, 557-570.

[17] D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research* 3, 993-1022.

[18] J. Boyd-Graber and D. M. Blei. 2009. Multilingual topic models for unaligned text. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI '09) Montreal, Quebec, Canada — June 18 - 21, 2009* AUAI Press, Arlington, Virginia, United States, 75-82.

[19] J. Jagarlamudi and H. Daumé. 2010. Extracting Multilingual Topics from Unaligned Comparable Corpora. In *Advances in Information Retrieval. ECIR 2010*, C. Gurrin, Y. He, G. Kazai, U. Kruschwitz, S. Little, T. Roelleke, S. Rüger and K. Van Rijsbergen Eds. Springer, Berlin, Heidelberg, 444-456. DOI:https://doi.org/10.1007/978-3-642-12275-0_39.

[20] K. Kucher and A. Kerren. 2015. Text visualization techniques: Taxonomy, visual survey, and community insights. In *2015 IEEE Pacific Visualization Symposium (PacificVis)*, 117-121. DOI:10.1109/PACIFICVIS.2015.7156366.

[21] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. 2002. ThemeRiver: visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics* 8, 1, 9-20. DOI:10.1109/2945.981848.

[22] W. Dou, X. Wang, R. Chang, and W. Ribarsky. 2011. ParallelTopics: A probabilistic approach to exploring document collections. In *2nd IEEE Conference on Visual Analytics Science and Technology 2011, VAST 2011*, Providence, RI, 231-240. DOI:10.1109/VAST.2011.6102461.

[23] S. F. Liu, M. X. Zhou, S. Pan, W. Qian, W. Cai, and X. Lian. 2009. Interactive, topic-based visual text summarization and analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management, CIKM '09, Hong Kong, China — November 02 - 06, 2009* ACM, New York, NY, USA, 543-552. DOI:10.1145/1645953.1646023.

[24] W. Ribarsky, D. X. Wang, and W. Dou. 2014. Social media analytics for competitive advantage. *Computers and Graphics* 38, 1, 328-331. DOI:http://dx.doi.org/10.1016/j.cag.2013.11.003.

[25] S. Fortunato. 2010. Community detection in graphs. *Physics reports* 486, 3-5, 75-174. DOI:https://doi.org/10.1016/j.physrep.2009.11.002.

[26] S. Wasserman and K. Faust. 1994. *Social network analysis: Methods and applications*. Cambridge University Press, Cambridge, UK.

[27] J. Heinrich, D. Weiskopf, and 95-116. .. . In Eurographics (Stars). 2013. State of the Art of Parallel Coordinates. In *Eurographics 2013 - State of the Art Reports*, M. Sbert and L. Szirmay-Kalos Eds. The Eurographics Association, 95-116. DOI:http://dx.doi.org/10.2312/conf/EG2013/stars/095-116.

[28] A. Inselberg and B. Dimsdale. 1987. Parallel Coordinates for Visualizing Multi-Dimensional Geometry. In *Computer Graphics 1987.*, T.L. Kunii Ed. Springer, Tokyo. DOI:10.1007/978-4-431-68057-4_3.

[29] R. Zafarani and H. Liu. 2009. Social computing data repository at ASU. http://socialcomputing.asu.edu/

[30] J. Leskovec and A. Krev. 2014. SNAP Datasets: Stanford Large Network Dataset Collection. https://snap.stanford.edu/data/

[31] J. Yee, R. F. Mills, G. L. Peterson, and S. E. Bartczak. 2005. *Automatic Generation of Social Network Data from Electronic-Mail Communications.* Defense Technical Information Center.

[32] H. Pérez-Rosés and F. Sebé. 2015. Synthetic generation of social network data with endorsements. *Journal of Simulation* 9, 4, 279-286. DOI:https://doi.org/10.1057/jos.2014.29.

[33] A. Prat and X. Sánchez. 2017. Ldbc-snb data generator. https://github.com/ldbc/ldbc_snb_datagen

[34] O. Erling, A. Averbuch, J. Larriba-Pey, H. Chafi, A. Gubichev, A. Prat, M.-D. Pham, and P. Boncz. 2015. The LDBC Social Network Benchmark: Interactive Workload. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data (SIGMOD '15). Melbourne, Victoria, Australia — May 31 - June 04, 2015* ACM, New York, NY, USA, 619-630. DOI:10.1145/2723372.2742786.

[35] M.-D. Pham, P. Boncz, and O. Erling. 2012. S3G2: A Scalable Structure-Correlated Social Graph Generator. In *Selected Topics in Performance Evaluation and Benchmarking. TPCTC 2012*, R. Nambiar and M. Poess Eds. Springer, Berlin, Heidelberg. DOI:10.1007/978-3-642-36727-4_11.

[36] A. Lenhart. 2015. *Teen, Social Media and Technology Overview 2015. Smartphones facilitate shifts in communication landscape for teens.* Pew Research Center.

[37] S. Greenwood, A. Perrin, and M. Duggan. 2016. *Social Media Update 2016. Facebook usage and engagement is on the rise, while adoption of other platforms holds steady.* Pew Research Center.

[38] B. Shneiderman. 1996. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of 1996 IEEE Symposium on Visual Languages. (3-6 Sept. 1996, Boulder, CO, USA, USA)* IEEE, EEUU, 336-343. DOI:10.1109/VL.1996.545307.

[39] D. A. Keim, F. Mansmann, and J. Thomas. 2010. Visual analytics: how much visualization and how much analytics? *ACM SIGKDD Explorations Newsletter* 11, 2, 5-8. DOI:10.1145/1809400.1809403.

[40] J. C. Roberts. 2007. State of the Art: Coordinated & Multiple Views in Exploratory Visualization. In *Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV 2007)*, 61-71. DOI:10.1109/CMV.2007.20.

[41] J. Nielsen and T. K. Landauer. 1993. A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems. Amsterdam, The Netherlands — April 24 - 29, 1993* ACM, New York, NY, USA, 206-213. DOI:10.1145/169059.169166.

[42] W. A. Pike, J. Stasko, R. Chang, and T. A. O'connell. 2009. The science of interaction. *Information Visualization* 8, 4, 263-274. DOI:10.1057/ivs.2009.22.