UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE MEDELLÍN
FACULTAD DE MINAS

DYNA
http://dyna.medellin.unal.edu.co/

Mining unstructured data to support requirements elicitation by using controlled vocabularies: A systematic mapping study

José L. Barros-Justo [a]

[a] *Escuela Superior de Ingeniería Informática, Universidad de Vigo, Ourense, España. jbarros@uvigo.es*

Received: December 17th, 2014. Received in revised form: May 20th, 2015. Accepted: May 27th, 2015.

Abstract
This paper presents a work-in-progress that deals with the assessment of the use of controlled vocabularies during the processes of requirements engineering, as a means to mine data from different sources (interviews, contracts, schemas and diagrams). By doing this the requirements description, analysis and comprehension is facilitated for both developers and end users. As a research methodology, we decided to use a systematic mapping study covering the last fourteen years (2000 - 2014). As far as we know, such studies have not yet been done; however, the cost incurred from errors in the requirements elicitation phase is one of the problems that is most commonly reported by the practitioners. Our study includes data on the processes of building the controlled vocabulary and assesses the productivity and quality. We are also interested in tools and techniques to classify and retrieve information. Our first findings suggest that this is an under-research area.

Keywords: Unstructured data; requirements elicitation; controlled vocabularies; systematic mapping study.

Minería de datos sin estructura para el soporte de la adquisición de requerimientos mediante el uso de vocabularios controlados: Un estudio de mapeo sistemático

Resumen
Este trabajo muestra un trabajo en progreso relacionado con la evaluación del uso de vocabularios controlados durante los procesos de ingeniería de requisitos, como un medio para obtener datos de muy diferentes fuentes (entrevistas, contratos, esquemas y diagramas) y, al hacer esto, facilitar la descripción de requisitos, el análisis y la comprensión por los desarrolladores y usuarios finales. Como metodología de investigación se decidió utilizar un estudio sistemático de mapeo que abarca los últimos catorce años (2000-2014). Por lo que sabemos no existen estudios previos de este tipo, sin embargo, el coste producido por errores en la fase de obtención de requisitos es uno de los problemas más mencionados por los profesionales. Nuestro estudio incluye datos sobre los procesos de construcción del vocabulario y evalúa la productividad y la calidad. También estamos interesados en las herramientas y técnicas utilizadas para clasificar y recuperar información. Nuestros primeros resultados sugieren que esta es un área de investigación sub-estudiada.

Palabras clave: Datos no estructurados; obtención de requerimientos; vocabularios controlados; estudio de mapeo sistemático.

1. Introduction

Over the last few decades, Software Engineering (SE) has become increasingly popular among software developers as an effective method to plan and document all the software development processes that have led to high quality products. Using the Unified Modeling Language (UML), programmers can deal with different types of software artifacts, employing a suitable and widely accepted representation formalism [1], which may incorporate: use case descriptions, diagrams and other relevant artifacts [2]. The use of common definitions, terms and keys, including dictionaries, vocabularies and ontologies are also an extended practice. All these strategies that comprise requirement analysis, planning and project

documentation, can reduce most of the error costs that emerge during the software development process [3].

One of the most important problems that may arise when developing software is the fact that, on many occasions, documentation is constructed employing vocabulary and term descriptions that are not as standardized as is desirable and not all the stakeholders are familiar with them. Moreover, a term may denote more than one concept, and this can lead to ambiguity, as well as the opposite (synonymy): more than one term can denote a unique concept [4].

In order to unify the definitions and terms for a specific project, several authors have proposed the use of controlled vocabularies, as a way to create narratives that could be used for a wide range of individuals involved in the software project. This could lead to the creation of more usable and maintainable software applications [5-7].

However, a problem remains; there is an abundance (sometimes overwhelming) of unstructured data that should, in some way, be transformed to become useful and understandable. This abundance is due to three main reasons:

- A group of heterogeneous participants with (frequently) divergent goals: end users, developers, business people, competitors, market experts, governments, etc.
- Many data/information sources: text documents, forms, video feeds, interviews recordings, regulations, books, rules, methodologies, bug reports, international standards, code comments, user manuals, maintenance documentation, etc.
- The lack of widely accepted techniques and associated tools to deal with this mass of data. Researchers have offered all kinds of theoretical tools such as: data and text mining, information retrieval, natural language processing, controlled vocabularies and ontologies, and modeling languages, but none of these has won the favor of all stakeholders.

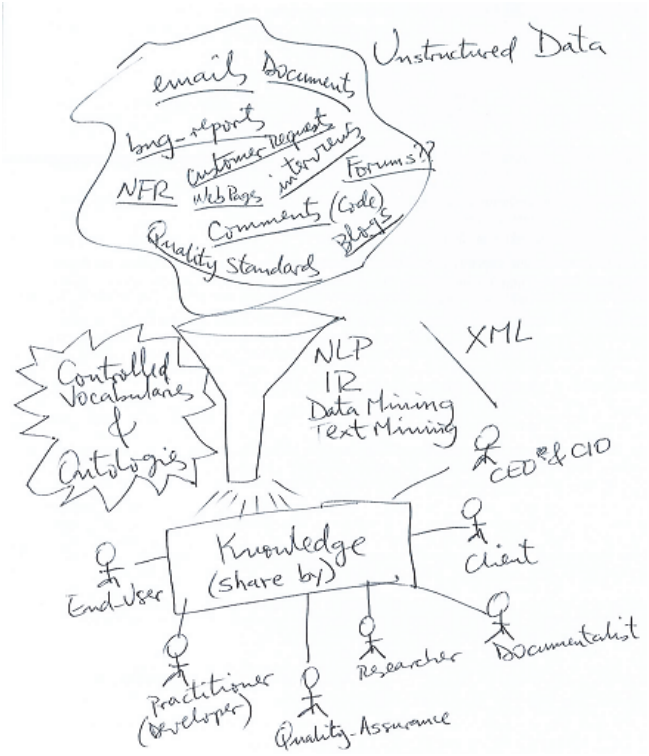

Figure 1 Transforming Unstructured Data into Useful Knowledge
Source: The Author

Keeping this in mind, we present this preliminary work aimed at assessing the use of controlled vocabularies as a tool to transform unstructured data, collected during the stage of requirements engineering (mainly requirements elicitation) into useful knowledge, contributing to develop high quality software and documentation, thus avoiding some costs due to misunderstandings and shortened development time.

The rest of the paper is organized as follows: Section 2 presents several prior studies dedicated to the topics of interest (mapping studies, unstructured data, requirements engineering and controlled vocabularies); Section 3 explains the goal and proposed methodology; finally, Section 4 presents main conclusions of this ongoing work, as well as guidelines for further investigation.

2. Related work

In recent years, the problem of facilitating the understanding of the requirements by all stakeholders and reducing the risk of misunderstandings has been thoroughly dealt with in the literature. However, it still remains a real problem and more detailed research is need from academia. Requirements are often ambiguous, incomplete, or even contradictory. One cause of this problem is miscommunication between analysts and users. In their work, Meth, Brhel and Maedche [8] analyze in detail 36 relevant publications in the area of automatic elicitation of requirements. They prove that there is a lot of text that is not clearly structured, coming from different stakeholders and part of large software development projects. All these documents must be analyzed and, finally, transformed into well-structured specification requirements.

As Al-Fedaghi stated, [9] UML is clearly inappropriate for non-technical users (clients and end users in the business sector), even for developers, as the specification of UML still contains ambiguous terms and shows a lack of a conceptual model that provides a common basis for understanding.

Michael H. Hugos uses five diagrams, among which four are classic and one is modern, in order to confront the problems in communication between developers and users (business users). The problem lies not so much in the text itself, but in that ambiguous mass of words and abstractions called UML [10].

Moreover, Durdik and Reussner [11] highlight the fact that during the system design, many decisions are made without sufficient information, due to lack of documentation. This leads to wrong decisions, which, in turn, affect the construction phase of the system. These design errors are the causes of customer and end user dissatisfaction , low quality applications and cost overruns in the development process. A similar case, involving students doing a Master's in computer science, is reported in the work of Scanniello et al. [12]. Researchers carried out a controlled experiment to verify the impact of the documentation available on design patterns for software maintenance activities.

Although specialized vocabularies have repeatedly proven their utility in the recovery of information, the

organization of the documentation and the understanding of natural language, their construction and the underlying taxonomies and ontologies is a complex process that consumes time and effort. This is not only during the initial phase of construction, but throughout their lifecycle; they require regular and frequent maintenance and upgrades. Current research is therefore focused on the automation of these creation and maintenance processes, by means of the semi-automatic extraction of information from collections of documents, databases, the Internet and the web, expert opinions and from many other sources [4]. Ontologies have also been proposed as a mechanism to recover information, in particular for the retrieval and potential reuse of UML class diagrams [13].

Pagano and Bruegge [14] reported on the need to have tools that can provide support to consolidation, structuring, analysis and monitoring processes of the information feedback, which occurs once the system is delivered to the end users. One recent work [5] highlights the fact that language, used in the specification of the system's functional requirements, can determine the value of the use cases that are employed to design the user interface.

Finally, Zapata and Vargas [24] recognized that organizational problems during requirements elicitation need to be traced to organizational goals. However, at the present time, this trace is based on stakeholder's experience. They propose a method composed by semantic and syntactic rules to express the organizational problem (requirement) in terms of goal statements.

3. Goals and methodology

Our work has two aims. First, we want to assess whether the use of controlled vocabularies during requirements elicitation is efficient, i.e., whether it is worth the effort that needs to be dedicated to building the vocabulary regarding expected revenue potential for improvements in the quality of the final product (software system). Additionally, we want to know the state of the art in relation to the existence of techniques and tools to automate (or at least facilitate) the mining of unstructured data and its transformation into structured and useful knowledge that will be shared by the stakeholders through all phases of the software development life cycle.

We conducted a systematic mapping study [15,16] in order to analyze the current published literature relating to the use of controlled vocabularies to facilitate the requirements in the elicitation process, especially when dealing with unstructured data.

The essential process steps of our systematic mapping study, as suggested by Petersen [17], are: defining research questions, conducting the search for relevant papers, screening papers, keywording of abstracts and data extraction and mapping. Each process step has an outcome, the final outcome of the process being the systematic map.

We would like to gain deeper knowledge by answering the following research questions:

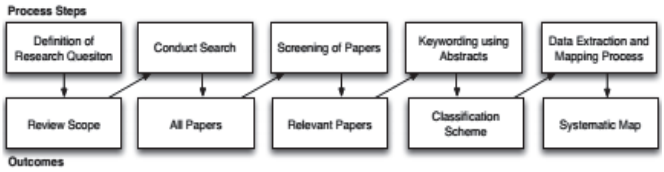

Figure 2. Systematic mapping process steps
Source: Petersen, 2008

1. What are the current techniques and tools for mining unstructured data during requirements elicitation? What is their level of automation?
2. Can controlled vocabularies really facilitate the requirements elicitation process when dealing with unstructured data? Is there any empirical evidence? Is the effort dedicated to building vocabularies worthwhile given the expected benefits?

We included all the literature produced between 2000 and 2014 in our study that was published in scientific journals, magazines, conferences and workshops. It is especially interesting to assess the method, technique or tool used by researchers to validate their own work [18], so we checked whether this validation process was carried out, and, if so, how [19-20]. For example, before selecting the evaluation method to be used to prove the correctness and completeness [21] of each particular vocabulary, some of the techniques reported in the literature have been analyzed. Hevner et al. [22] undertook a controlled experiment, and the method was validated in a controlled environment. A simulation process was also used by the same authors, who employed outputs created by different expert groups. This second option was also used by Tena et al. [5], who employed evaluation measures based on experts' opinions.

The main contributions of our work are:

1. A systematic mapping study of the use of controlled vocabularies to support the requirements elicitation when dealing with unstructured data, and structuring research from the last fourteen years by analyzing published literature (specialized journals and conferences, peer-reviewed sources).
2. A detailed research protocol that includes: search strategy, definition of inclusion/exclusion criteria and data extraction guidelines to properly classify and analyze the retrieved works.
3. A set of bibliometric measures to map the current research field (state of the art), both at a research space level and at a publication space level.
4. Detection of research gaps and suggestions for further studies and promising new areas of research.

4. Threats to validity

There are many limitations to this study. From the viewpoint of the search process:

- It is based on a restricted set of indexing systems. We have used SCOPUS[1], a particularly useful system as it indexes publications from a large number of publishers

[1] http://www.scopus.com

including ACM, IEEE, Elsevier, and Springer. We also used Web of Knowledge (WoS[2]) and IEEE Xplore[3], both having significant overlap. Finally, Google Scholar[4] pretends to index all the sources covered by the others systems, but after some tests we found this not to be true.

- The search was based on a restricted set of terms and perhaps missed topics that might normally be considered to be unstructured data or mining techniques. We tried to limit the possible threat by using synonyms, but it is clearly not an exhaustive solution.
- The search string was applied to a restricted set of elements in the paper: title, abstract and keywords. Although it is reasonable to think that if the search terms are not present in these elements then the content of the paper is not relevant to the study, this cannot be formally proved.
- Publication bias for internal validity: Successful cases are probably published more often than failures, and significant results may be published more often than when the results are not considered to be so significant.

To obtain a set of relevant papers that ensured good coverage of the area of interest, we constructed the search strings incrementally, starting with the broader term controlled vocabularies. The search was then refined by applying more specialized terms, particularly those which made it possible to reduce the search space to the area of interest (unstructured data and requirements elicitation). These terms were agreed upon by the authors as being the most representative; however, the list of search terms may not be complete, and additional or alternative terms (synonyms) in the search string could affect the number of papers retrieved by the databases.

After completing the search process, retrieved papers were distributed randomly among two external reviewers and the author. Each author received an even number of papers, half exclusively and the other half in duplicate (another reviewer received the same sample), as the extraction of data needed to happen independently, without communication between reviewers. Duplicate documents allow us to cross-check data retrieved from each paper. The cross-check also helped us increase the reliability of data and mitigate, as far as possible, threats to the validity of our study.

Regarding the external validity and the conclusions, we believe it is important to offer other researchers the data and details on the selection and extraction processes, so as they can replicate the study and check the results. Thus, we plan to provide researchers with an online repository with all the data of our mapping. Given the systematic approach we have followed for all the processes, we believe our study is easily repeatable.

5. Conclusions and future research

This is an ongoing research project and we are still in the first stages of extracting and analyzing data, and expanding the initial search by including references from previous selected papers and citations that other authors have made to papers in the original selected set (Snow balling technique). We have followed an extended version of the formal process proposed by Kitchenham [16] and Wohlin [23] when carrying out a systematic mapping study.

So far we have been able to identify three important elements for our study: frequent uses of controlled vocabularies, potential sources of data (and vocabulary terms) and mining techniques frequently used when dealing with unstructured data.

Our initial set of pre-selected studies consists of 226 papers (from specialized journals and main international conferences), but we were unable to find any systematic mapping study related to unstructured data. This encourages us to continue with our aim of carrying out this type of formal research. We are now working in the inclusion/exclusion criteria and the classification schema.

References

[1] Rumbaugh, J., Jacobson, I. and Booch, G., The unified modeling language: Reference manual. Addison Wesley, 2005.

[2] Sommerville, I., Software engineering. Pearson education, 2008.

[3] Urquhart, C., Themes in early requirements gathering: The case of the analyst, the client and the student assistance scheme. Information Technology & People, 12, pp. 44-70, 1999. DOI: 10.1108/09593849910250547

[4] Medelyan, O., Witten, I., Divoli, A. and Broekstra, J., Automatic construction of lexicons, taxonomies, ontologies, and other knowledge structures. WIREs Data mining Knowl Discov, 00, 2013. DOI: 10.1002/widm.1097

[5] Tena, S., Díez, D., Díaz, P. and Aedo, I., Standardizing the narrative of use cases: A controlled vocabulary of web user tasks. Information and Software Technology, 55, pp. 1580-1589, 2013. DOI: 10.1016/j.infsof.2013.02.012

[6] Bleik, S., Mishra, M., Huan, J. and Song, M., Text categorization of biomedical data sets using graph kernels and a controlled vocabulary. IEEE Transactions on Knowledge and Data Engineering, 10, pp. 1211-1217, 2013. DOI: 10.1109/TCBB.2013.16

[7] Liu, M., Halper, M., Geller, J. and Perl, Y., Using OODB modelling to partition a vocabulary into structurally and semantically uniform concept groups. IEEE Transactions on Knowledge and Data Engineering, 14, pp. 850-866, 2002. DOI: 10.1109/TKDE.2002.1019218

[8] Meth, H., Brhel, M. and Maedche, A., The state of the art in automated requirements elicitation. Information and Software Technology, 55, pp. 1695-1709, 2013. DOI: 10.1016/j.infsof.2013.03.008

[9] Al-Fedaghi, S., A method for modeling and facilitating understanding of user requirements in software development. Journal of Next Generation Information Technology (JNIT), 4 (3), pp. 30-38, 2013. DOI: 10.4156/jnit.vol4.issue3.4

[10] Hugos, M., Five diagrams beat a victorian novel. Computerworld, 41 (39), pp. 23- 23, 2007.

[11] Durdik, Z. and Reussner, R., On the appropiate rationale for using design patterns and pattern documentation, Proceedings of QoSA'13, 2013. DOI: 10.1145/2465478.2465491

[12] Scanniello, G., Gravino, C., Risi, M. and Tortora, G., A controlled experiment for assessing the contribution of design pattern documentation on software maintenance, Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement, 2010. DOI: 10.1145/1852786.1852853

[13] Robles, K., Fraga, A., Morato, J. and Llorens, J., Towards an ontology-based retrieval of UML Class Diagrams. Information and

[2] http://www.webofkonwledge.com

[3] http://ieeexplore.ieee.org

[4] http://scholar.google.es/

Software Technology, 54, pp. 72-86, 2012. DOI: 10.1016/j.infsof.2011.07.003
[14] Pagano, D. and Bruegge, B., User involvement in software evolution practice: A case study, Proceedings of ICSE 2013, 2013. DOI: 10.1109/ICSE.2013.6606645
[15] Kitchenham, B. and Brereton, P., A systematic review of systematic review process research in software engineering. Information and Software Technology, 55, pp. 2049-2075, 2013. DOI: 10.1016/j.infsof.2013.07.010
[16] Kitchenham, B. and Charters, S., Guidelines for performing systematic literature reviews in software engineering. EBSE, 1, 2007.
[17] Petersen, K., Feldt, R., Mutjaba, S. and Mattsson, M., Systematic mapping studies in software engineering, Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering (EASE), 2008.
[18] Mader, C. and Haslhofer, B., Quality criteria for controlled web vocabularies, Proceedings of the 10th European Networked Knowledge Organisation Systems Workshop, NKOS, 2011.
[19] Dyba, T., Kitchenham, B. and Jorgensen, M., Evidence-based software engineering for practitioners. IEEE Software, 22 (1), pp. 58-65, 2005. DOI: 10.1109/MS.2005.6
[20] N.I.S. Organization., Guidelines for the construction, format, and management of monolingual controlled vocabularies. ANSI/NISO Z39.19-2005, NISO Press, 2005.
[21] Zowghi, D. and Gervasi, V., On the interplay between consistency, completeness, and correctness in requirements evolution. Information and Software technology, 45, pp. 993-1009, 2003.
[22] Hevner, A., March, S., Park, J. and Ram, S., Design science in information systems research. MIS Quarterly, 28, pp. 75-105, 2004.
[23] Wohlin, C., Runeson, P., Anselmo da Mota, P., Engström, E., do Caro, I. and Santana de Almeida, E., On the reliability of mapping studies in software engineering. The Journal of Systems and Software, 86, pp. 2594-2610, 2013. DOI: 10.1016/j.jss.2013.04.076
[24] Zapata-J., C.M. and Vargas-Agudelo, F.A., Specification of problems from the business goals in the context of early software requirements elicitation. DYNA, 81 (186), pp. 193-199, 2014. DOI: 10.15446/dyna.v81n186.39910

J.L. Barros-Justo, received his PhD in Informatics Eng. in 2006, from the Universidade de Vigo, Ourense, Spain. Since 1990 he has worked in the Universidade de Vigo. Previously he worked for several companies including Hewlett-Packard, I.D.E.A., Inforpyme and ALBA insurance. Currently, he is a Professor in the Informatics Department, Escuela Superior de Ingeniería Informática, Universidade de Vigo, Spain. His research interests include: software engineering, software reuse, development methodologies and evidence-based software engineering.
ORCID: orcid.org/0000-0003-2046-2643#sthash.Myllb7NP.dpuf