Towards a FAIR Dataset for Spanish Non-Functional Requirements

María Isabel Limaylla-Lunarejo, Nelly Condori-Fernandez, and Miguel R. Luaces

Database Lab, Fac. Informática, Universidade da Coruña, 15071 A Coruña, Spain
Citius, Universidad de Santiago de Compostela, 15071 A Coruña, Spain
Centro de Investigación CITIC, Universidade da Coruña, 15071 A Coruña, Spain
Correspondence:
maria.limaylla@@udc.es,n.condori.fernandez@usc.es,miguel.luaces@udc.es

DOI: *https://doi.org/10.17979/spudc.000024.30*

Abstract: Supervised Machine Learning algorithms (ML) have enhanced the performance of the automatic non-functional requirements (NFR) classification in the Requirements Engineering domain. However, the lack of public datasets, dealing with imbalanced datasets and reproducibility are current concerns in ML experiments. We conducted a quasi-experiment to generate a dataset of NFR in the Spanish Language, following the FAIR Principles. We collected 109 requirements from an open access repository of the University of A Coruña, and performed a labeling process based in the categories and subcategories of the ISO/IEC 25010 quality model. Using a Fleiss' Kappa test we obtained a substantial agreement (0.78) at the category level and a moderate agreement (0.48) when the classification is per subcategory.

1 Introduction

Supervised Machine Learning (ML) algorithms and Natural Language Processing techniques have been used to improve the performance of the automatic non-functional requirements classification. However, the lack of publicly datasets for requirements categorized in sub-classes of non-functional classification is still one concern when conducting ML experimentsAhmad et al. (2020); Binkhonain and Zhao (2019). Moreover, the lack of diverse datasets in languages other than English is also currently challenges. Reproducibility is another concern when speaking of ML experiments, in part due to several barriers like accessibility and availability, capability of reuse, among others. To avoid these barriers, Wilkinson el al.(2016) propose the FAIR Guiding Principles, that consist of four principles: Findable, Accessible, Interoperable, and Reusable, guiding how data should be managed to be more easily accessible, understood, exchangeable, and reusable Wilkinson et al. (2016).

In this research we present a summary of the quasi-experiment conducted in a previous work Limaylla-Lunarejo et al. (2023). The main objective of the quasi-experiment was generate a dataset of non-functional requirements written in the Spanish Language, following the FAIR Guiding Principles for facilitating reuse. The Fleiss' Kappa test was used to assess the inter-rater reliability with multiple annotators.

2 Adapted FAIRification process

Some research have explored the application of the FAIR principles to datasets. One example is the FAIRification process propose by GO FAIR[1], an initiative to coordinate and collaborate on the global Internet of FAIR Data & Services (IFDS). Figure 1 presents the FAIRification process adopted by GO FAIR, consisting on seven steps: 1.Retrieve non-FAIR data, 2.Analyse the retrieved data, 3.Define the semantic model, 4.Make data linkable, 5.Assign license, 6.Define metadata for the dataset, and 7.Deploy FAIR data resource. This process have been used as a base on several studies Kochev et al. (2020); Sinaci et al. (2020).

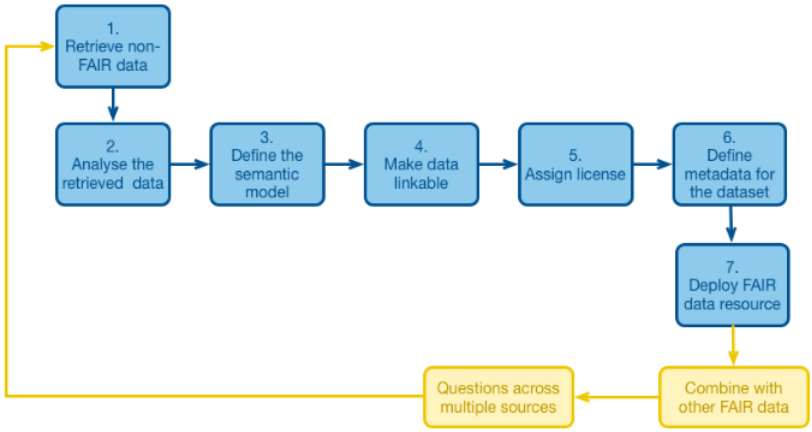

Figure 1: FAIRification Process GO FAIR International Support and Coordination Office. (2022)

While the original process primarily focused on making FAIR existing data (and metadata), our adaptation allowed the generation of FAIR datasets. In order to utilize the GO FAIR Process (FAIRification) for dataset generation, we made adaptations by incorporating, modifying, and excluding certain tasks. Figure 2 shows the adapted FAIRification process, that consist also in seven tasks: 1.Define Semantic Model, 2.Data Definition, 3.Data Collection, 4.Data Labeling, 5.Define Metadata, 6.FAIR Validation, and 7.Data Publishing.

The first two steps perform several definitions before the data collection and labeling. A semantic model involves defining the meaning of entities, their relationships, vocabulary, and ontologies/taxonomies. The Certified Professional for Requirements Engineering (CPRE) glossary Glinz (2011) was used for concepts like 'requirement', 'non-functional requirement', and 'stakeholders'. The ISO/IEC 25010 quality model ISO (2011) was chosen as taxonomy for the NFRs, and an entity-relationship model was selected to represent data with two entities: 'requirements' and 'projects'. We also established the data's structure for each entity, aligning it with the previously established semantic model. The next two tasks are focused on data collection and labeling. We review several Bachelor projects and selected 19 projects with has at least three NFR and collected the information in two Excel files. The labeling process will be present in more detail in the following section. The data collection and labeling could be an iterative process. Once the data is collected and labeled, the metadata is describe, such as the authors, the description, the license, the language used, the process, etc. Finally, the last two tasks is about carrying out an evaluation of the FAIR principles before, during and/or after publishing, the last task. A preliminary version of Metadata and datasets was released on Zenodo CERN and OpenAIRE (2013), an open publishing repository aligned with the FAIR principles. Four

[1] https://www.go-fair.org/fair-principles/fairification-process/

files were published[2], including the data structure, projects list and the requirements with the final label for the categories and subcategories, each one in a CSV file; and an RDF ontology model.

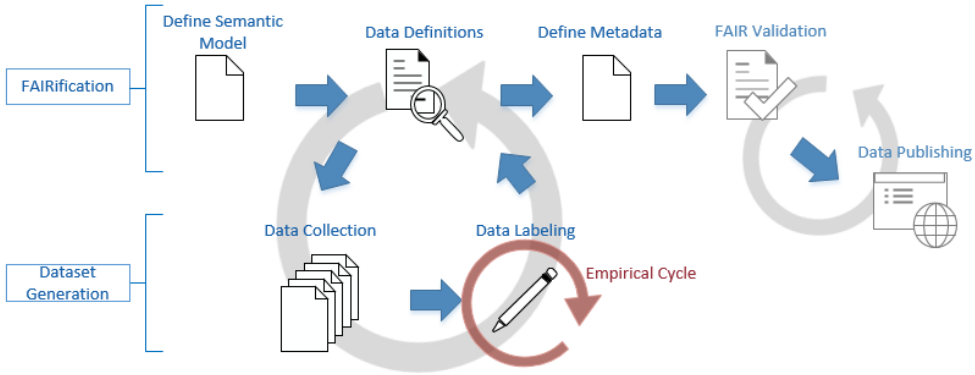

Figure 2: Adapted FAIRification Process for Dataset Generation

3 Experiment and Results

This section introduces the research questions, the instrumentation and process of the quasi-experiment, and the data analysis and results.

3.1 Research Questions

- **RQ1**: Does the adapted FAIRification process contribute to improve the reliability of agreement between multiple annotators?
- **RQ2**: Which NFR categories/subcategories are less reliable due to the high number of disagreements in several requirements?

3.2 Instrumentation and process

We collected 109 requirements from the free repository of the University of A Coruña[3], and performed a labeling process based in the categories and subcategories of the ISO/IEC 25010 quality model[4]. The labeling process was accomplished by seven annotators (four PhD students and three professionals), all native Spanish speakers. The requirements were divided in four groups and assigned one or two groups to each annotators. We performed a meeting with the annotators to explain the process and also prepared a document with the Spanish definitions of software requirements: functional and non-functional, and the definitions of all categories (characteristics) and subcategories (sub-characteristics) of the ISO/IEC 25010. All this definitions were based on the semantic model, the vocabulary, and the taxonomy defined in the first two task of the adapted FAIRification process.

[2] https://doi.org/10.5281/zenodo.7124407
[3] https://ruc.udc.es
[4] https://iso25000.com/index.php/en/iso-25000-standards/iso-25010

3.3 Data analysis and Results

Each requirement obtained between three and five labels from the annotators. For define the final label we based on unanimity or at least a prevailing vote among all the labels (majority). The outcomes of the label consensus for categories are displayed in Table 1. For the 109 labeled requirements a 63% obtained a label by unanimity, a 35% by majority and a 2% didn't agreed. From this two requirements, one was identified as Constraint (not a quality concern) and the other mentions a compact device and it could lead to confusion.

Table 1: Numbers of requirements belonging to categories

Categories	Unanimity	Majority	No agreement	Total
Usability	26	6	-	32
Security	19	6	-	25
Performance	11	8	-	19
Reliability	10	6	-	16
Functional Suitability	1	4	-	5
Maintainability	2	3	-	5
Portability	0	5	-	5
No agreement	0	0	2	2
Total	69	38	2	109

The identical analysis was carried out for the 107 requirements regarding the subcategories label. Around 37% of the requirements were labeled with unanimity in the subcategory level and 42% were labeled with majority agreement. The annotators disagreed on 21% of the classification of the subcategory of requirements. Table 2 presents the numbers of requirements belonging to these three subcategories (Unanimity, Majority, No agreement). Annotators encountered challenges and achieved limited consensus when categorizing requirements under subcategories associated with Usability (Operability, Learnability, User interface aesthetics, Accessibility, User error protection, and Appropriateness recognizability), Security (Confidentiality, Authenticity, and Integrity), and Performance efficiency (Time behavior, Capacity, and Resource utilization). Recurring disagreements were observed for subcategory groups like Operability and Learnability, Time behavior and Capacity, and Confidentiality, Authenticity, and Integrity.

Using a Fleiss' Kappa test Fleiss (1971) we obtained a substantial agreement in the category level (0.78) and a moderate agreement (0.48) when the classification is per subcategory. This indicate that it was possible to label almost all requirements based on unanimity or majority agreement. Regarding subcategories, there was a 20% of requirements without consensus, probably due to numerous subcategories Gut and Bayerl (2004), intersection of some meaning (such as Confidentiality and Integrity subcategories Samonas and Coss (2014)) and a lack of detail in some requirements.

4 Conclusions

In summary, this experiment has made two main contributions. Firstly, it has addressed the prevalent lack of requirement datasets in the Spanish language by successfully generating a new Spanish dataset. Secondly, it has provided valuable insights into the FAIRification process, offering an adapted framework that incorporates dataset creation from its inception. The reliability assessment of agreement among multiple annotators, using Fleiss' Kappa, has demonstrated substantial agreement when the classification is conducted at the category level (0.78) and moderate agreement (0.48) when the classification is done for subcategories. Future work will include additional experiments in NFR labeling and a FAIR validation process.

Table 2: Numbers of requirements belonging to subcategories

Categories	SubCategories	U	M	N	Total
Usability	Accessibility	0	1	-	1
	User error protection	1	1	-	2
	Learnability	1	2	-	3
	Operability	5	6	-	11
	User interface aesthetics	4	4	-	8
	No Agreement	0	0	7	**7**
Portability	Adaptability	0	5	-	5
Security	Authenticity	1	4	-	5
	Confidentiality	1	6	-	7
	Integrity	5	3	-	8
	No Agreement	0	0	5	**5**
Performance efficiency	Capacity	0	1	-	1
	Time behaviour	9	0	-	9
	Resource utilization	1	2	-	3
	No Agreement	0	0	6	**6**
Reliability	Recoverability	2	0	-	2
	Fault tolerance	2	2	-	4
	Availability	6	1	-	7
	Maturity	0	1	-	1
	No Agreement	0	0	2	**2**
Maintainability	Modifiability	1	3	-	4
	Modularity	1	0	-	1
Functional Suitability	Functional completeness	0	1	-	1
	Functional correctness	0	1	-	1
	Functional appropriateness	0	1	-	1
	No Agreement	0	0	2	**2**
	Total	40	45	22	107

Legend: U: Unanimity, M: Majority, N: No agreement

Acknowledgements

CITIC is funded by the Xunta de Galicia through the collaboration agreement between the Consellería de Cultura, Educación, Formación Profesional e Universidades and the Galician universities for the reinforcement of the research centres of the Galician University System (CIGUS).

Bibliography

A. Ahmad, C. Feng, M. Khan, A. Khan, A. Ullah, S. Nazir, and A. Tahir. A systematic literature review on using machine learning algorithms for software requirements identification on stack overflow. *Security and Communication Networks*, 2020, 2020.

M. Binkhonain and L. Zhao. A review of machine learning algorithms for identification and classification of non-functional requirements. *Expert Systems with Applications: X*, 1:100001, 2019.

CERN and OpenAIRE. Zenodo, 2013. URL *https://www.zenodo.org/*.

J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76 (5):378, 1971.

M. Glinz. A glossary of requirements engineering terminology. *Standard Glossary of the Certified Professional for Requirements Engineering (CPRE) Studies and Exam, Version*, 1:56, 2011.

GO FAIR International Support and Coordination Office. Fairification process, 2022. URL *https://www.go-fair.org/fair-principles/*. [Online; accessed 02-September-2023].

U. Gut and P. S. Bayerl. Measuring the reliability of manual annotations of speech corpora. In *Speech prosody 2004, international conference*, 2004.

ISO. Iso/iec 25010:2011 systems and software engineering - systems and software quality requirements and evaluation (square) - system and software quality models, 2011.

N. Kochev, N. Jeliazkova, V. Paskaleva, G. Tancheva, L. Iliev, P. Ritchie, and V. Jeliazkov. Your spreadsheets can be fair: A tool and fairification workflow for the enanomapper database. *Nanomaterials*, 10(10):1908, 2020.

M.-I. Limaylla-Lunarejo, N. Condori-Fernandez, and M. R. Luaces. Towards a fair dataset for non-functional requirements. In *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, SAC '23, page 1414–1421, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450395175.

S. Samonas and D. Coss. The cia strikes back: Redefining confidentiality, integrity and availability in security. *Journal of Information System Security*, 10(3), 2014.

A. A. Sinaci, F. J. Núñez-Benjumea, M. Gencturk, M.-L. Jauer, T. Deserno, C. Chronaki, G. Cangioli, C. Cavero-Barca, J. M. Rodríguez-Pérez, M. M. Pérez-Pérez, et al. From raw data to fair data: the fairification workflow for health research. *Methods of information in medicine*, 59(S 01):e21–e32, 2020.

M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.