FAIRification of Citizen Science Data Through Metadata-Driven Web API Development

Reynaldo Alvarez[1(✉)], César González-Mora[2], José Zubcoff[2], Irene Garrigós[2], Jose-Norberto Mazón[2], and Hector Raúl González Diez[1]

[1] University of Informatics Sciences, Havana, Cuba
{rluna,hglez}@uci.cu
[2] Department of Languages and Computing Systems, University of Alicante, Alicante, Spain
{cgmora,jose.zubcoff,igarrigos,jnmazon}@ua.es

Abstract. Citizen Science (CS) implies a collaborative process to encourage citizens to collect data in CS projects and platforms. Unfortunately, these CS initiatives do not follow metadata nor data-sharing standards, which hampers their discoverability and reusability. To improve this scenario in CS is crucial to consider FAIR (Findability, Accessibility, Interoperability and Reusability) guidelines. Therefore, this paper defines a FAIRification process (i.e. make CS initiatives more FAIR compliant) which maps metadata of CS platforms' catalogues to DCAT and generates Web Application Programming Interfaces (APIs) for improving CS data discoverability and reusability in an integrated approach. An experiment in a CS platform with different CS projects shows the performance and suitability of our FAIRification process. Specifically, the validation of the DCAT metadata generated by our FAIRification process was conducted through a SHACL standard validator, which emphasises how the process could boost CS projects to become more FAIR compliant.

Keywords: Citizen science · FAIR · DCAT metadata · Web APIs · Open data

1 Introduction

Nowadays, there is an emerging trend of democratising science, characterised as Citizen Science (CS) [19]. This term has different definitions depending on the scope, but it is mainly considered as a collaborative process to generate knowledge [15]. Interestingly, as stated in [7], CS is crucial in the production of relevant data to analyse and monitor certain natural, economic or social processes. Therefore, CS initiatives support the growth of research data, with millions of volunteers generating data from observations and sensors [2].

Data generated by CS is generally hosted by CS platforms, i.e., Web based portals which contain data from several CS projects obtained by volunteers. Those platforms, and their available projects, must follow the Ten Principles of

T. Di Noia et al. (Eds.): ICWE 2022, LNCS 13362, pp. 162–176, 2022.
https://doi.org/10.1007/978-3-031-09917-5_11

CS [22], which are an agreement to provide a common framework to evaluate and set up CS initiatives. Among of these principles, the importance of data is highlighted, since it is stated that "Citizen science project data and metadata are made publicly available and where possible, results are published in an open-access format". Consequently, CS data should follow the FAIR guidelines [32], aiming at making data Findable, Accessible, Interoperable and Reusable.

However, most CS platforms such as Scistarter[1] or Zooniverse[2] do not generally follow these FAIR guidelines [33]: regarding findability ("F" from FAIR), it is difficult to discover CS projects that already include CS data; considering accessibility ("A"), existing CS data is difficult to be accessed by data consumers, as they must download it as entire datasets (when available); then, the interoperability ("I") is really limited as CS metadadata is generally published in customised formats; and finally, the reusability ("R") is hampered, as CS platforms do not generally offer solutions such as Web APIs to facilitate the reuse of data. Although many CS platforms consider the Public Participation in Scientific Research (PPSR-Core) metadata standards to model CS projects, datasets and observations [9], FAIR guidelines are not considered.

To enforce FAIR, the W3C Data Catalog Vocabulary (DCAT[3]) is widely used in open data projects [29]. If DCAT is properly implemented, it facilitates the interoperability of dataset metadata and its consumption by using different applications [11]. However, DCAT is not generally adopted by CS platforms [14]. Indeed, a recent study [24] exposes the lack of metadata completeness as a general problem in CS field, describing in detail the behaviour in SciStarter as reference. Moreover, an assessment of current data practices in 36 CS projects highlights the lack of open access to data, metadata and documentation [5].

Consequently, our hypothesis is that the adoption of metadata standards like DCAT allows CS initiatives to become FAIR compliant, but the adoption of DCAT is not sufficient for this purpose. Therefore, in addition to adopting DCAT standard, more efforts are required to really achieve FAIR data by allowing also the data reuse and sharing by structures such as Web APIs [12,28]. In this sense, our paper aims to develop this FAIRfication –process of making data FAIR– by (i) mapping metadata from CS platforms and from DCAT, and (ii) providing access to CS data through Web APIs.

Adhering to the FAIR guidelines in CS initiatives will enhance contextualisation and data quality [24], as high data quality increases the value and reuse of the data. Therefore, FAIR adoption enables CS projects to be more successful and well recognised, which leads to the empowerment of the citizen scientists.

Therefore, the main contributions of this paper are:

- Review of existing CS platforms regarding FAIR guidelines' compliance.
- Development of a FAIRfication process for the enrichment of CS platforms by using DCAT as metadata standard for enabling Findability, Accesibility, Interoperability and Reusability of CS data.

[1] https://scistarter.org/.
[2] https://www.zooniverse.org/.
[3] https://www.w3.org/TR/vocab-dcat-3/.

– Integration of Web APIs as data services for CS platforms.
– Evaluation of our proposed approach for FAIRfication in a CS platform.

This article is structured as follows. In Sect. 2, the related work is described in detail. Then, in Sect. 3, a running example is presented in order to illustrate the process. After that, the full FAIRfication process definition is presented in Sect. 4. An evaluation of the approach is explained in Sect. 5, and finally, conclusions and future work are presented in Sect. 6.

2 Related Work

In this section, the most important related work is briefly described, highlighting existing CS platforms and solutions that support accessing CS data and metadata.

Table 1. Main CS platforms

Platform	Region	Projects	Metadata management	API for data
European Citizen Science Portal [30]	Europa	206	PPSR-Core	No
citizenscience.gov Portal	EEUU	493	PPSR-Core	No
SciStarter global citizen science hub [16]	Global	1591	PPSR-Core	No
Zooniverse [25]	Global	104	Own	No
CitSci [31]	Global	1040	PPSR-Core	Yes

CS platforms are recognised for serving as a discover point for CS projects. There are 5 types of CS platforms according to Liu et al. [18]: commercial platforms for CS initiatives, CS platforms for specific projects, CS platforms for specific scientific topics, national CS platforms and EU CS platforms. The most representative sample of CS projects are those non-commercial platforms with national and global scope, such as Scistarter. Table 1 summarises those target platform, and as can be seen, CS platforms generally adopt the PPSR-Core metadata model as practice in the CS field, although it is an ongoing work by the Data and Metadata Working Group[4] of The Citizen Science Association (CSA). Moreover, CS platforms hardly ever include an API, and it is even less frequent that a CS platform provides an API with query-level access to data, that is, a fine-grain API that allows to access not only to metadata, but also the dataset itself. Although there is a consensus about the adoption of PPSR-Core, in practice the platforms use different terms, structures, so that is not easy to harvest data from those different catalogues.

[4] https://citizenscience.org/get-involved/working-groups/data-and-metadata-working-group/.

In order to improve data sharing and adopt FAIR guidelines there are different related works: METAFair [28] consists of a metadata profile based on DCAT 2, a potential source schema due to the range of properties for describing different components of datasets, data services and their related resources. This metadata application profile is implemented for GenBank Metadata Analytics, not for general use or in CS as standard. Moreover, they do not approach the definition of data services in DCAT 2 as enabler of FAIR, particularly in the reuse feature. Another work includes FAIRification of the involved datasets in a research workflow [6], as well as applying semantic technologies to represent and store data about the detailed versions of the general protocol, of the concrete workflow instructions, and of their execution traces. They aim to establish a common framework between teams and experts. This approach shows good practices regarding DCAT for data sharing, but it is only developed in a closed environment and they do not address CS platforms and projects.

Also, when datasets are discoverable, the interoperability relies on the exposure of data through standard Application Programming Interfaces (APIs), improving the level of data sharing. Reviews of CS status identify as key aspect API developments to increase the capacity of the data generators apps and to help in the phase of analysis and visualization [10,17,23,27]. The previous work from the authors of this paper define an APIfication model-driven process for the automatic generation of a Web API from tabular data [13]. This process transforms a tabular dataset file to an standard Web API, allowing the reuse of dataset fields at query level. This approach is a first step in order to achieve the FAIRfication of data, but more efforts are required for improving metadata and APIs in the field of CS.

As pointed out, there is still a gap in terms of becoming FAIR, as there is a diversity of metadata types exposed in CS platforms, limiting the discoverability and reusability of CS projects out of their scope. Related research proposes different approaches, ones for specific domain solutions, others as IT solutions or models enablers for FAIR guidelines. Regarding generic solutions, they aim to create terms, processes or workflows to adopt FAIR capacities. The improvement of data repositories as FAIR is approached by different works, offering good practices, workflows and tools that try to become FAIR compliant (serving as proved good practices in their scenarios). Regarding standard adoption, different related research aim to use DCAT to increase the metadata discoverability. DCAT has created the path to do it through data services specifications, but they are not used in the main schemes reviewed, as far as the authors are concerned. Moreover, DCAT is yet not commonly used in CS projects and platforms. Therefore, these are the main reasons to contribute with our work to empower the use DCAT, APIs and data services as metadata, particularly in the field of CS.

3 Running Example

In order to illustrate the process, a sample project from SciStarter[5] is selected. SciStarter is one of the most recognised CS platform that currently includes more than 1500 CS projects [3]. Specifically, the project selected is "Street Story: Give Your Input on Safe Streets"[6], which offers public access to data about transportation collisions and other travel information since 2018, collected by non-scientist members [21]

Table 2. Fragment of "Street Story" dataset

Type	Crash/Near-miss	Involved	Was anyone injured?	Report date
incident	bike	driver	minor	2018/10/01
incident	bike	driver	none	2018/10/01
nearMiss	walk	driver	none	2020/08/02
nearMiss	ride	driver	none	2020/08/02
nearMiss	walk	ped	none	2021/10/17
incident	wheelchair	driver	sev	2021/10/25
nearMiss	walk	driver	none	2021/10/30
nearMiss	bike	ped	none	2021/10/30

The data generated in the Street Story project is public, and the dataset is accessible in tabular format, particularly in CSV, a non-proprietary format commonly used to publish and share data [12]. It contains 28 columns with different data types and more than 3500 records collected by CS volunteers. A fragment of those records is shown in Table 2. It includes information about the type of report (incident or near miss), the vehicle or activity involved in the incident (such as bike, walk or ride), who was involved (driver or pedestrian), the injured provoked (sever, minor or none) and the reporting date, among other fields.

As an example situation, a data engineer wants to perform an analysis of transportation incidents. First, the data engineer performs an exploration of the available data about the issue, using for example the search keywords "street story" or "street incidents" in the European Data Portal[7], a commonly used portal for open data. The search results obtained consists of 2893 datasets accessible in CSV format. The metadata catalogue of this open data portal is based on DCAT, including an API to harvest available metadata. However, when searching for this data in CS platforms such as SciStarter, or CS projects related to the same topic, it is hardly impossible to find, access and reuse CS data. In

[5] https://scistarter.org/.
[6] https://scistarter.org/street-story-give-your-input-on-safe-streets.
[7] https://data.europa.eu/.

order to integrate and analyse data from CS projects, specific procedures must be developed to harvest each platform metadata catalogue. This is the barrier about discovery approached in this work. Although DCAT adoption increase discovery, to become more FAIR compliant it is necessary to address reusability. Indeed, when the data engineer searches certain information in the data portals and CS platforms, the results indicate that none of the datasets are accessible via API services. Then, it requires an effort to develop those APIs. For this reason, approaching API generation and the annotation of data services as metadata is a higher desirable level in a FAIRification process.

As the impact of CS data in scientific and other communities has been proven, the launch in more spaces could gain attention from government or policies agencies. CS projects should be as much discoverable as possible. In this case, SciStarter and their projects should be aware of their potential and increase their discoverability, facilitating the access to their data sources and providing structures to obtain their metadata. With this running example we are approaching FAIR guidelines through an integrated process that enhances the growing and acknowledgement of CS projects out of this scope.

Therefore, this running example exposes the approach to become more FAIR, from the point of view of the data consumers and publishers. In fact, a recent study of 1020 projects from SciStarter shows an analysis of metadata completeness [24], exposing the number of projects that do no fully offer metadata. They also highlight that it is necessary for CS projects to include suitable data documentation to be able to produce scientific data [24]. Thus, CS platforms could be more integrated, sharing the same structure and metadata formats with other data repositories, supporting the harvesting of DCAT metadata, and allowing data consumers or developers to create value from CS data by facilitating the access to data through query-level Web APIs.

4 FAIRification of CS Platforms

Data publishers aim to become FAIR compliant by improving the discoverability and interoperability of CS data to enhance the engagement of new members and the recognition in open and formal sciences communities. In this sense, an important requisites for FAIRification are the adoption of established metadata standards and the inclusion of structures to access data. Therefore, standards as DCAT enable the capacity to link data services, increasing the capacity of generating value and applications; and providing Web APIs facilitate the easy access by consumers to the published data. Indeed, this integrated approach supports the use of *dcat:Dataservice*, a class included in the last DCAT version to easily provide access to data through the generated Web APIs [1].

In order to become FAIR compliant in the scope of CS initiatives, this research addresses the accomplishment of most relevant guidelines. The main goal is accordingly to contribute to the FAIRification of CS data, by a process that provides more accessible and standardised technologies and applications for improving data sharing in CS.

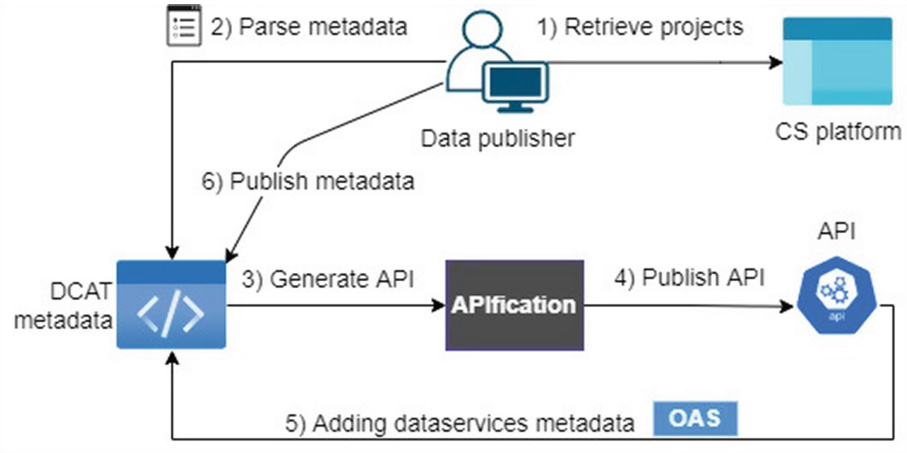

Fig. 1. General process for improving FAIR

This process for becoming FAIR in CS platforms is approached through metadata parsing and API generation. The whole process is represented by the schema in Fig. 1, and the corresponding programming code developed in Python is available in a Github repository[8].

The process begins when a data publisher or a platform manager retrieves the metadata of a CS platform (step 1 in Fig. 1). Then, the second step consists of parsing the relevant metadata fields and mapping them to the DCAT vocabulary, thus generating a DCAT metadata specification.

Once the DCAT specification is generated, the dataset distribution URL is accessible through the project metadata specification. This step aims to increase accessibility and interoperability of CS data by adopting standards and exposing data resources. Regarding reuse, the process continues with the step 3 from Fig. 1, which consists of an APIfication process to offer easy access to data at query level. This Web API generated is a desirable capacity to easily explore, reuse and access available datasets [23]. This part of the process comes from our previous work [13], which is now improved and integrated to work on CS platforms. Then, steps 4 and 5 are closely related. When the Web API is published, the specification of data services is thus added to the DCAT metadata. Finally, the process ends with the publication of the complete metadata defined as step 6.

A complete explanation of these steps to become FAIR compliant in the scope of CS, including relevant details and the inputs and outputs of each step, is described in the following subsections.

[8] https://github.com/ralvarezluna/csdatalab-apigen.

4.1 Retrieving the Projects

The SciStarter API[9] enables searching for CS projects within the platform. The metadata associated to the published projects is returned in a JSON file through the corresponding API call (e.g. Scistarter projects API finder). It includes a collection of objects corresponding to the existing projects that fulfil a matching condition. In order to retrieve those projects, our implementation will support JSON local files as input or API URLs.

Before the parser step, the general metadata of the catalogue must be available in a properties file. The mandatory fields for a DCAT catalogue are: identifier, language and homepage [1]. Also, for adding semantics, the catalogue "dcat:themeTaxonomy" attribute is defined by default as **citizen science**[10] from a public knowledge base, which allows to be denoted by a concept and published as Linked Data. The outputs of this step are the JSON file with the metadata retrieved and the initial configuration of the metadata catalogue.

4.2 Parsing the Metadata

The metadata retrieved in JSON format contains the attributes for projects defined in PPSR-Core [4]. PPSR-Core has four metadata components: (i) Core Attributes (Core Metadata Model - CMM), (ii) Projects (Project Metadata Model - PMM), (iii) Datasets (Dataset Metadata Model - DMM) and (iv) Observations (Observation Metadata Model - OMM).

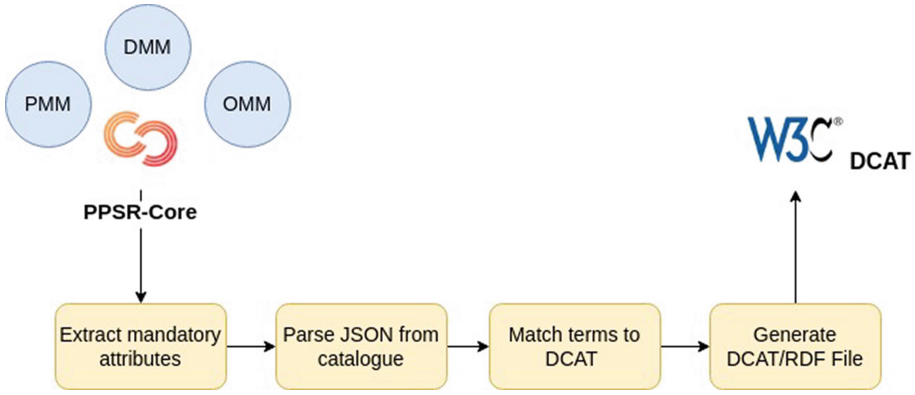

Fig. 2. Parsing metadata

The Fig. 2 shows the procedure to parse the metadata from the JSON to a DCAT compatible file. An analysis of the attributes from PPSR Core was conducted for extracting mandatory attributes to discover projects out of the

[9] https://scistarter.org/api.
[10] https://www.wikidata.org/wiki/Q1093434.

scope of CS. The second procedure defines how to parse the retrieved data catalogue in JSON format, how to read the objects and set up metadata for the catalogue in general. During the procedure for matching terms to DCAT each project entity will be mapped as a DCAT Resource, specifically as instances of the Dataset class. This class supports 23 of 36 attributes defined in PPSR-Core, which are enough to discovery purposes. In fact, 8 of the not supported attributes from PPSR-Core are optional, and the others 5 mandatory attributes not parsed are very specific domain attributes from projects, not relevant for discovering purposes.

Table 3. Abstract of the metadata attributes

PPSR	JSON	DCAT
projectId	guid	identifier
projectDateCreated	created	release date
projectLastUpdatedDate	Date Time	update/modification date
projectName	name	title
projectDescription	text	description
hasTag	search terms	keyword
keyword	keyword	keyword
projectStartDate	begin	release date
projectScienceType	fields_of_science	theme
projectUrl	signup_url	landing_page
projectResponsiblePartyName	presenter	publisher
contactPoint	url	contactpoint
projectGeographicCoverage	geographic_scope	spatial_coverage
Not defined	sust_dev_goals	theme/category

A subset of the attributes parsed are showed in Table 3, including information about the original attribute defined in PPSR-Core, how it is described in the JSON, and finally, how will be expressed in DCAT.

Additionally, the metadata retrieved of the projects includes references to Sustainable Development Goals (SDG) related to the project activities. Those references are mapped with the definition of the corresponding terms of the United Nations SDG Taxonomy[11]. During the mapping process, those terms enriched with semantics are included as themes (referring to attribute dcat:theme). After the matching of terms, the attributes of the entities from the JSON are annotated and converted in valid vocabulary terms of the DCAT standard. Finally, for the generation of the DCAT Resource Description Framework (RDF) catalogue the auxiliary library **datacatalogtordf**[12] is used, as

[11] http://metadata.un.org/sdg.
[12] https://github.com/Informasjonsforvaltning/datacatalogtordf.

it supports the DCAT specification classes and graph conversion to RDF. The resulting metadata is exported as RDF format ready to be shared as linked data.

4.3 Generating the API

In this step, the process continues generating the API, which is based in our previous work [13], a complete automatic process to generate an API from CSV datasets. In order to improve and extend the functionality of this API generator, in this work a feature to convert JSON files to CSV is included in order to consider more common formats of CS projects. Therefore, the APIfication process will be able to generate the code and documentation of an API for almost any dataset with resources (in both CSV and JSON formats). Then, data consumers only have to use these endpoints in their applications to allow users to collect and visualise data.

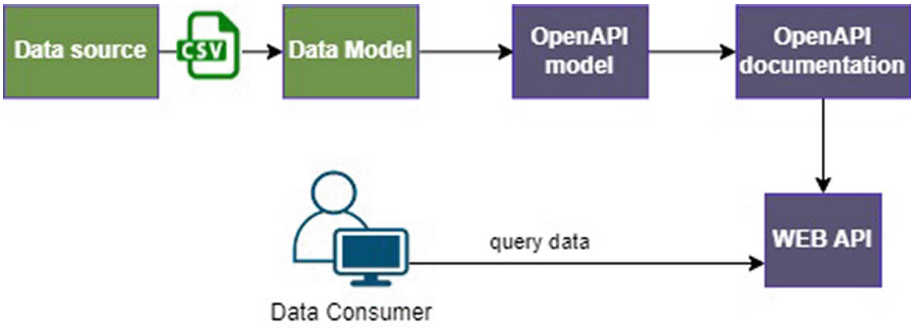

Fig. 3. APIfication process

Figure 3 shows the steps of this automatic generation process from the data source to the Web API. A data model is constructed from the data source, following this data model an OpenAPI model is generated. The OpenAPI documentation is built from the model following OpenAPI 3.0, the most common open standard regarding documentation and specification of APIs [12]. The output is a complete and running Web API accessible at localhost for development and test, which can be easily published online.

4.4 Publishing the Web API

This step depends on the infrastructure to publish data services, because OD portals do not generally cover this part itself as the growing number of requests could not be supported. Due to the existing freedom to manage their infrastructure for projects or catalogues, the API generated should be published by the owners. Therefore, the output of this step consists of a Web API with the Open

API Specification (OAS) for further addition of metadata for this endpoint as DCAT Dataservice Resource[13].

4.5 Updating Data Services Metadata

Once the API generation phase is completed and the Web API is published, the specification of data services as resources must be performed to extend the DCAT specification. Using a DCAT class for this type of resource, an instance of the class *dcat:DataService* will be created with reference to services endpoints (e.g. a URL of the published service). A feature to read the OpenAPI file or URL is developed within the process. The library Oastodcat[14] helps us with the creation of the Dataservice class from the OpenAPI specification file of the published Web API. Then, the object constructed will be able to be appended as data service in the CS metadata catalogue.

4.6 Publishing Metadata

To increase the discoverability of the projects, the DCAT catalogue obtained must be published in the CS platform. This step must be done by platforms' owners, as the open source code of the implemented process could be extended or integrated in automatic pipelines to keep this metadata catalogue updated. The way for publishing the graph database of the catalogue based in DCAT depends on the owner, but in any case, publishing it as a raw file enables the feature to be harvested or downloaded by third parties, achieving the desired discoverability. Open data portals, search engines and developers could access then the metadata and generate value, integrating data from similar projects.

The goal of FAIRification was approached by performing a parsing procedure to generate a DCAT compliant metadata catalogue, improving the interoperability of the CS metadata catalogue out of this environment. At project level the API generation enable the Reuse and integration features. Being Findable and Accessible depends of the implementation of the process for the CS platforms.

5 Evaluation

In order to perform an evaluation of our approach, we analysed the FAIRfication process for CS platforms. The evaluation aims to prove the performance of the process with different entries, and also, the compliance of the DCAT standard for metadata in RDF format. For this validation of DCAT the Shapes Constraint Language (SHACL)[15] is used, as W3C recommendation language for validating RDF data. Specifically, SHACL documents consists of constraints collections that enforce particular shapes on an RDF graph. SHACL is used in this research

[13] https://www.w3.org/TR/vocab-dcat-2/#Class:Data_Service.
[14] https://pypi.org/project/oastodcat/.
[15] https://www.w3.org/TR/shacl/.

for validating purposes as it is currently used for validating RDF files in recent related works [8,20,26].

The experiment was carried out in an Ubuntu 20.04 computer with an Intel i7 processor and 8 GB of RAM memory. The performance of the approach has been evaluated by analysing the parser time for different projects. The project metadata is the unit of information of the experiment, including 141 attributes for describing the project, related to: activities, resources, contacts, data, locations, topics and identification. Then, the entries are formed by different number of project metadata, which is remotely accessed from SciStarter. In this evaluation, the time measured does not include the time for the remote calls in order to avoid network issues.

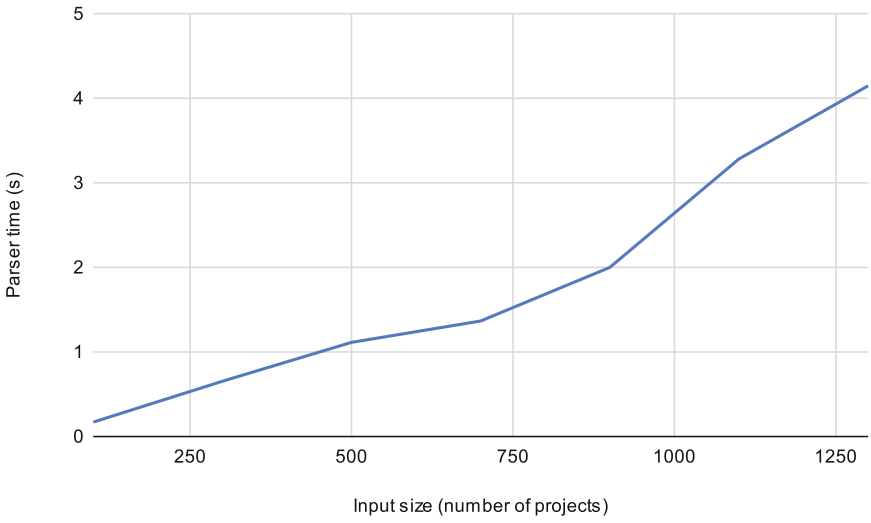

Fig. 4. Performance of the parser for diverse inputs

The results of the measurements with different size of projects are showed in Fig. 4. Beginning with 100 projects until 1300 in the X axis and from 0 to 5 s in the axis Y. The graphics shows a linear behaviour in the generation of DCAT metadata depending on the number of projects. The process with more than 300 projects takes less than 1 s and with 1300 projects takes around 5 s.

The generated DCAT file within the process is validated with a SHACL online validator[16], using the generic profile of DCAT. The validation report returns a positive conform result for the RDF file, generated using different random subsets of projects due to performance issues of the validator. The file conformity implies that the implemented parser generates DCAT complaints outputs ready to share as data catalogues.

[16] https://data.vlaanderen.be/shacl-validator/.

Moreover, the APIfication process was proved as part of the process, in a integration approach, because full validation of this step was exposed in our previous work [12]. The experiment exposed in this work concludes that the automatic generation process successfully achieves the objective of functionally and efficiently performing the APIfication process in an average time of 16 s using sample datasets from different sources and sizes.

Therefore, the evaluation results show that our FAIRfication process is functional and compliant with the DCAT standard, thus becoming FAIR through the adoption of established metadata standards and facilitating interoperability and access to the data. Indeed, according to the results of the experiment, a complete DCAT metadata catalogue from SciStarter is generated and made accessible in few seconds, providing also API access to the available data enforcing reuse capabilities.

6 Conclusions

The analysis of data repositories of CS projects shows the lack of metadata and sharing standards adoption, hampering the fulfilment of FAIR guidelines. Indeed, our running example shows, in a real scenario, the lack of discoverability regarding CS data, which is caused by the implementation of customised metadata solutions by CS initiatives. To overcome this pitfall, in this paper, we propose a FAIRification process to (i) enrich CS platforms by means of mapping elements from the PPSR-Core metadata (profusely used in CS projects) to their counterparts from the W3C standard DCAT metadata, as well as (ii) generating Web APIs to facilitate the access to the corresponding CS data. We argue in the paper that both mechanisms separately are not sufficient: just adopting DCAT is not enough, nor is it sufficient to provide Web API services for querying data. Consequently both mechanisms (metadata mapping and Web API generation) must be used together to make CS portals more FAIR-compliant.

In order to show the benefits of our approach, we have conducted an experiment to evaluate our FAIRfication process through performance and correctness tests. Results show that it performs efficiently and produces correct metadata in DCAT format, including suitable query-level Web APIs. Therefore, this proposal serves as driver to follow the path of the FAIRfication improvement in the field of CS.

As future work, the process will address the issue of missing metadata in most CS platforms and approaching FAIRfication from the early project set up phase, contributing from its inception in the compliance of FAIR guidelines. Also, the parser will be extended to support other CS platforms and further evaluation with user studies will be carried out.

Acknowledgement. This research work has been partially funded by the Proyecto Habana 2021 and by project GVA-COVID19/2021/103 funded by Conselleria de Innovación, Universidades, Ciencia y Sociedad Digital de la Generalitat Valenciana.

References

1. Albertoni, R., et al.: Data catalog vocabulary (DCAT)-version 2. World Wide Web Consortium (2020)
2. Baker, B.: Frontiers of Citizen Science: explosive growth in low-cost technologies engage the public in research. Bioscience **66**(11), 921–927 (2016)
3. Ben Zaken, D., Gal, K., Shani, G., Segal, A., Cavalier, D.: Intelligent recommendations for citizen science. Proc. AAAI Conf. Artif. Intell. **35**(17), 14693–14701 (2021)
4. Bowser, A.: Standardizing citizen science? Biodivers. Inf. Sci. Stand. **1**, e21123 (2017)
5. Bowser, A., et al.: Still in need of norms: the state of the data in citizen science. Citizen Sci. Theory Pract. **5**(1) (2020)
6. Celebi, R., et al.: Towards fair protocols and workflows: the OpenPredict use case. Peer J. Comput. Sci. **6**, e281 (2020)
7. Cooper, C.B., Rasmussen, L.M., Jones, E.D.: Perspective: the power (dynamics) of open data in citizen science. Front. Climate **3**, 57 (2021)
8. Corman, J., Reutter, J.L., Savković, O.: Semantics and validation of recursive SHACL. In: Vrandečić, D., et al. (eds.) ISWC 2018. LNCS, vol. 11136, pp. 318–336. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00671-6_19
9. Data, U.: Metadata WG. 2019. PPSR core data & metadata standards repository. Github (2019)
10. Ed-douibi, H., Cánovas Izquierdo, J.L., Daniel, G., Cabot, J.: A model-based Chatbot generation approach to converse with open data sources. In: Brambilla, M., Chbeir, R., Frasincar, F., Manolescu, I. (eds.) ICWE 2021. LNCS, vol. 12706, pp. 440–455. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-74296-6_33
11. Färber, M., Lamprecht, D.: The data set knowledge graph: creating a linked open data source for data sets. Quant. Sci. Stud. **2**(4), 1–30 (2021)
12. González-Mora, C., Garrigós, I., Zubcoff, J., Mazón, J.N.: Model-based generation of web application programming interfaces to access open data. J. Web Eng. **19**, 194–217 (2020)
13. González-Mora, C., Garrigós, I., Zubcoff, J.: An APIfication approach to facilitate the access and reuse of open data. In: Bielikova, M., Mikkonen, T., Pautasso, C. (eds.) ICWE 2020. LNCS, vol. 12128, pp. 512–518. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-50578-3_36
14. Hahnel, M., Valen, D.: How to (Easily) extend the FAIRness of existing repositories. Data Intell. **2**(1–2), 192–198 (2020)
15. Haklay, M.M., Dörler, D., Heigl, F., Manzoni, M., Hecker, S., Vohland, K.: What is citizen science? The challenges of definition. In: Vohland, K., et al. (eds.) The Science of Citizen Science, pp. 13–33. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-58278-4_2
16. Hoffman, C., Cooper, C.B., Kennedy, E.B., Farooque, M., Cavalier, D.: SciStarter 2.0: a digital platform to foster and study sustained engagement in citizen science. In: Analyzing the Role of Citizen Science in Modern Research, pp. 50–61 (2017)
17. Lemmens, R., Antoniou, V., Hummer, P., Potsiou, C.: Citizen science in the digital world of apps. In: Vohland, K., et al. (eds.) The Science of Citizen Science, pp. 461–474. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-58278-4_23
18. Liu, H.-Y., Dörler, D., Heigl, F., Grossberndt, S.: Citizen science platforms. In: Vohland, K., et al. (eds.) The Science of Citizen Science, pp. 439–459. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-58278-4_22

19. Mueller, M.P., Tippins, D., Bryan, L.A.: The future of citizen science. Democracy Educ. **20**(1), 2 (2011)
20. Pareti, P., Konstantinidis, G., Mogavero, F., Norman, T.J.: SHACL satisfiability and containment. In: Pan, J.Z., et al. (eds.) ISWC 2020. LNCS, vol. 12506, pp. 474–493. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-62419-4_27
21. Rebentisch, H., Wasfi, R., Piatkowski, D.P., Manaugh, K.: Safe streets for all? Analyzing infrastructural response to pedestrian and cyclist crashes in New York City, 2009–2018. Transp. Res. Rec. **2673**(2), 672–685 (2019)
22. Robinson, L.D., Cawthray, J.L., West, S.E., Bonn, A., Ansine, J.: Ten principles of citizen science. In: Citizen Science: Innovation in Open Science, Society and Policy, pp. 27–40 (2018)
23. de Sherbinin, A., et al.: The critical importance of citizen science data. Front. Climate **3**, 20 (2021)
24. Shwe, K.M.: Study on the data management of citizen science: from the data life cycle perspective. Data Inf. Manage. **4**(4), 279–296 (2020)
25. Simpson, R., Page, K.R., De Roure, D.: Zooniverse: observing the world's largest citizen science platform. In: Proceedings of the 23rd International Conference on World Wide Web, pp. 1049–1054 (2014)
26. Stani, E.: Metadata quality: generating SHACL rules from UML class diagrams. In: Proceedings of the 2018 International Conference on Dublin Core and Metadata Applications, pp. 63–64 (2018)
27. Sturm, U., et al.: Defining principles for mobile apps and platforms development in citizen science (2018)
28. Tompkins, V.T., Honick, B.J., Polley, K.L., Qin, J.: MetaFair: a metadata application profile for managing research data. Proc. Assoc. Inf. Sci. Technol. **58**(1), 337–345 (2021)
29. Vander Sande, M., Verborgh, R., Dimou, A., Colpaert, P., Mannens, E.: Hypermedia-based discovery for source selection using low-cost linked data interfaces. In: Information Retrieval and Management: Concepts, Methodologies, Tools, and Applications, pp. 502–537 (2018)
30. Wagenknecht, K., et al.: EU-citizen. Science: a platform for mainstreaming citizen science and open science in Europe. Data Intell. **3**(1), 136–149 (2021)
31. Wang, Y., Kaplan, N., Newman, G., Scarpino, R.: CitSci.org: a new model for managing, documenting, and sharing citizen science data. PLoS Biol. **13**(10), e1002280 (2015)
32. Wilkinson, M.D., et al.: The fair guiding principles for scientific data management and stewardship. Sci. Data **3**(1), 1–9 (2016)
33. Williams, J., et al.: Maximising the impact and reuse of citizen science data (2018)