This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2019.2893016, IEEE Transactions on Emerging Topics in Computing

Framework for Prioritization of Open Data Publication: An Application to Smart Cities

Alvaro E. Prieto, Jose-Norberto Mazón, and Adolfo Lozano-Tello

Abstract— Public Sector Information is considered to play a fundamental role in the growth of the knowledge economy and improvements in society. Given the difficulty in publishing and maintaining all available data, due to budget constraints, institutions need to select which data to publish, giving priority to data most likely to generate social and economic impact. Priority of publication could become an even more significant problem in Smart Cities: as huge amounts of information are generated from different domains, the way data is prioritized and thus reused, could be a determining factor in promoting, among others, new and sustainable business opportunities for local entrepreneurs, and to improve citizen quality of life. However, people in charge of prioritizing which data to publish through open data portals (such as Chief Data Officers, or CDOs) do not have available any specific support in their decision-making process. In this work, a proposal of a framework for prioritization of open data publication as well as its application to Smart Cities is presented. This specific application of the framework relies on OSS (Open Source Software) indicators to help making decisions on the most relevant data to publish focused on developers and businesses operating within the Smart City context.

Index Terms— Decision support, dataset reuse indicators, Open Data, Smart City application

1 Introduction

Governments, citizens, companies, journalists, software developers, researchers, NGOs and other stakeholders are increasingly acknowledging the importance of Public Sector Information (PSI) as an important source of raw material for innovation and both economic and social impact [1][2]. Importantly, the European Union [3] and the US government [4] recognizes the role of PSI to stimulate economic growth and promote social engagement. Public institutions have thus started to open their data so that they are reused with the aim of generating economic and social impact. To maximize possibilities of reuse, open datasets need to have a certain level of stability and maintenance over time [5]. This comes at an extra cost in time and money which means that public institutions cannot release all their available data because they usually have finite resources, so trying to publish all data at a time is not a good strategy [6]. Attard et al. [7] state that, "while there exist a huge number of government data portals that enable data producers to publish their data, there are not many tools aiding data publishers in this task". Although there are no such tools, the European Union has published two reports with recommendations about high value data domains for commercial use [8][9]. However, according to Serra [10], the most widely used strategy is, in fact, based on how easy it is to release data according to privacy issues, legal issues (such as transparency laws) and technical formats. This scenario means that public institutions publish large number of datasets but only a limited number of those datasets are being reused [11].

This problem becomes critical in local governments that are currently developing Smart City projects, because of their intensive use of IT, which generates huge amounts of data from different domains such as transportation, sustainability, tourism and so on [12]. The existence of open data portals providing these data has been acknowledged as a fundamental attribute of Smart Cities [13] for enhancing innovation in order to boost novel business opportunities and sustainable economic growth [14], as well as improving transparency, accountability and civil engagement [15].

In Smart Cities, not all data users have the same preferences. Zuiderwijk & Janssen [16] assert that different types of users of open data are often interested in different types of data, therefore, publication of data can be improved by taking into account preferences for certain types of data for certain open data users. That is, journalists or NGOs can be interested in reusing certain datasets related to transparency such as budgets or demography while software developers or IT consultants can be interested in reusing datasets on which to develop apps such as monuments, culture events or public transport timetables.

However, in accordance with Thorsby et al [17], few of the portals are tracking the usage of the data by applications, citizens or other city agencies in order to know how many reuses there are, what is their purpose or whether

- *Corresponding author: A.E. Prieto is with the Quercus Software Engineering Group, Escuela Politécnica, Universidad de Extremadura, Cáceres, 10003, Spain. E-mail: aeprieto@unex.es.*
- *J.N. Mazón is with the WaKe research group,DLSI & IUII, Universidad de Alicante, carretera San Vicente s/n 03690 San Vicente del Raspeig, Alicante (Spain). E-mail: jnmazon@ua.es.*
- *A. Lozano-Tello is with the Quercus Software Engineering Group, Escuela Politécnica, Universidad de Extremadura, Cáceres, 10003, Spain. E-mail: alozano@unex.es.*

2168-6750/0x/$xx.00 © 2018 IEEE Published by the IEEE Computer Society

they are effective or not.

Unfortunately, in any case, the potential value for each kind of users that a concrete PSI dataset might generate is not easy to estimate [5]. Janssen et al. [2] state that, "there is no way to predict and calculate the return of investment (ROI) in advance. […] The main challenge is that open data has no value in itself; it only becomes valuable when used". Therefore, the main problem is that data owners have limited understanding on how the published open data is reused and about its impact, in the sense of awareness, usefulness and interest generated by those reuses.

According to Zuiderwijk et al. [18], a clear process for publishing data of interest to users is required so that the benefits of open data are shown. This process should contain decision making steps enabling to prioritize which new data should be made open, based on monitoring how previously published open data is reused. In the same manner, Hjalmarsson et al. [19] proposes that publishing PSI as open data (i.e. in an open data portal) requires a decision support system to prioritize publication of those datasets offering higher potential to generate value.

This decision making process, which consists in selecting which data to publish, is now carried out by a person occupying a new position that has recently emerged in many public institutions, the Chief Data Officer (hereon CDO)[1]. This novel C-suite position is borrowed from the private business domain where CDOs are responsible for developing a data management strategy to achieve a company's goals considering (i) internal structure and external context of the company, (ii) useful dataspaces for the company, and (iii) generated value impact from data [20]. Specifically, a CDO, as stated by Kassen [15], is in charge of developing strategies for implementing, managing and supervising an open data project.

In the process of publishing data, there may be different criteria to prioritize the datasets in which to invest more resources. Broadly speaking, a public institution may have different objectives and policies to decide what data to publish, taking into account different issues such as services for citizens, the promotion of businesses in their environment, social inclusion, health and so on. When a regulation or policy has been established (which can be local, regional, national, international or thematic), the CDO must first comply with these established policies and guidelines. But, once the datasets that meet these policies have been published, the CDO should observe the interest that may exist in a wider range of datasets so that they could be reused in as many cases as possible.

Thus, one of the duties of a CDO is applying a decision making process for choosing those most desirable datasets to release according to their expected potential of being reused beyond policies and, so, to be useful and interesting for the citizens and to generate some kind of economic or social impact. To do this, it would be desirable to have the possibility to have some indicators estimating in some way the interest that the different open data reusers have in different types of datasets. Moreover, the importance of these indicators may vary for each CDO because public institutions may have different reuse goals related to their open data portals. In other words, some CDOs may place more emphasis on social impact using indicators related to reuse in view of transparency issues or social media, while other CDOs may give more importance to economic impact using indicators related to reuse in software applications. Whatever the case, an additional difficulty is that of establishing which indicators may be more or less relevant according to one or more of the potential benefits.

Therefore CDOs should be able to rely on a decision support system that appropriately relates different indicators on the use of already published open data on the one hand, with the assessment of these indicators made by CDOs according to their own strategy on the other. This problem thus becomes a multicriteria decision making (hereon MCDM) problem [21].

Unfortunately, to the best of our knowledge, CDOs of open data portals lack a decision making process for prioritizing datasets taking into account all the dimensions mentioned above. To overcome this pitfall, the POPSC (Prioritization of Open data Publication in Smart Cities) framework is described in this work. The goal of this framework is to provide CDOs of with a Decision Support System which recommends categories of datasets most suitable to be published in their open data portal. To do so, the framework proposes a set of actions aimed at estimating some kind of indicators about the awareness, usefulness and interest about the reuses of the datasets of the same category already published by similar open data portals. These indicators are then weighed, using some multi-criteria decision making method, taking into account the objectives of open data portals to offer an ordered list of categories of datasets to publish. Moreover, we present a fully functional AHP-based application of this framework oriented to support CDOs of Smart Cities.

It should be noted that, the proposed framework is not only applicable to Smart Cities but also to any other organization involved in an open data process, regardless of its type or size.

This paper is structured as follows: section 2 presents the research approach and summarizes other works related to the publishing of open data. Section 3 describes the POPSC Framework. Section 4 then details the characteristics of the specific application of the POPSC Framework using information from Github and Socrata. Finally, section 5 presents a test case of POPSC in a city.

[1] Although we advocate the figure of CDO as a key stakeholder in making decisions to support open data publication in a Smart City, it is also possible to rely on a board of experts

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2019.2893016, IEEE Transactions on Emerging Topics in Computing

2 Research Approach and Related Work

In this section we explain how we conducted our research as well as the fundamental background literature on publishing open data.

2.1 Action Research

Our research approach is based on action research, since it has been proved a good method for involving together researchers and practitioners on identifying a problem, resolving it and checking the success of the solution [22]. It consists on identifying a problem, resolving it and checking the success of the solution. According to [23] there are four types of participants in a research:

1. The researcher. In this work we are the researchers.
2. The researched. In this work, it is the specification of a framework for decision making support when publishing open data.
3. The researched for, in the sense of being who have the problem that will be researched. In our case, the organizations that are willing to open their data (specifically, the CDOs working on those organizations).
4. The researched for, in the sense of resulting beneficiary, even though not participating in the research. In our case, reusers that will use data for creating novel and added-value services and applications based on reusing open data, as well as citizens that will have access to larger amount of useful data about their cities, as well as interesting applications and services that allow them to improve their daily life. The datasets were classified according to a set of categories specifically designed for Smart Cities.

An action research plan will follow the tasks in an iterative and incremental way, creating a research process with the following stages [24]:

1. Identify the issues that guide the research, related with "the researched".
2. Gather all the information available to deal with the previous issues.
3. Analyse the information gathered to obtain suitable solutions to the issues.
4. Share the results with the other participants in order to improve the solutions and lead to new issues, which allows starting the cycle again.

In summary, the process defined by action research is iterative and incremental. It means that after applying that process, the solution to the issues are refined and improved. This succession of cycles characterizes action research as a process of searching solutions based on both feedback from practitioners, and our own experience.

Thus, following this action research process, we noticed that the advantages that AHP provides for prioritization processes could be useful for CDOs. Since AHP is based on the use on decision criteria to prioritize different alternatives, we defined the steps for the POPSC framework accordingly. On one hand, one step of the POPSC framework is to establish the alternatives, i.e., the different categories of datasets to prioritize. On the other hand, another step of the POPSC framework is to set the decision criteria, i.e., the indicators that can estimate the awareness, usefulness and interest of the reuses of the different categories. On the basis of these two fundamental steps, the next step is to do the actions oriented to get the real data needed to estimate the chosen indicators for the different categories. Finally, in the last step, AHP is applied using the indicators previously defined.

Showing the application of action research is made by using a test case of the POPSC framework [25], which is shown in section 5 .

2.2 Related Work on Publishing Open Data

In order to identify relevant related work on how open data is published by governments, we have considered a backward and forward snowball method [26] by navigating citations and references from a starting set of research papers in an iterative fashion. Our starting set of papers is composed of a couple of review papers on open data research, namely:

- Attard et al [7] with 84 references and 170 citations.
- Hossain et al [27] with 113 references and 43 citations

From these two papers we started the first iteration conducting backward and forward snowballing. Backward snowballing means using the reference list to identify new papers to include. Forward snowballing refers to identify new papers based on those papers citing the paper being examined. Next step is to go through the list of papers and exclude those that are not related with publication of open data (also we remove papers from the list that have already been examined in previous iterations). Iterations are repeated until no new papers are selected. Next, the result of this process is shown by providing a brief description of some examined papers.

Conradie & Choenni [6] state that data release by local governments is still a novel task, thus knowledge is lacking as to its benefits and barriers. Therefore, they conduct a participatory action research approach to get a better understanding of how internal processes of local governments influence data release. The authors found that the following indicators needed to be addressed by local governments to overcome barriers to releasing public sector information: (i) Data Storage, i.e., is data stored centrally, or is it decentralized?; (ii) Use of data, i.e., the way data is used by the department; (iii) Source of data, i.e., how is a set of data obtained?; and (iv) Suitability of data for release, i.e., are there rules and regulations that determine whether a dataset may be released or not, such as privacy or copyright.

Notwithstanding, these indicators are related to current data but do not address the actual use of the data and its benefits. For example, Hossain et al. [27] show that benefits associated with opening data are ill-understood. In their systematic review of open government data initiatives, Attard et al. [7] explore open data initiatives of a large number of governments, as well as existing tools and approaches. They found that while efforts have focused on developing tools for helping data publishers to open data, there has been less effort in developing strate-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2019.2893016, IEEE Transactions on Emerging Topics in Computing

gies for supporting decisions on which data to release. This means that public entities may end up publishing data with no value, rather than focusing on the relevance of the data they are publishing. Therefore, success in opening data is not a matter of the amount of data published, but of understanding how data is reused. As highlighted by Zuiderwijk & Janssen [28], since providers of open data are not concerned with needs of open data users, they do not know how their data are reused, and business related issues (such as creation of added-value services or products based on open data) are not widely used as a decision criterion.

Furthermore, Zuiderwijk et al. [18] argue that the publication of open data is often cumbersome so standard procedures and processes for opening data are required. They found a series of barriers preventing easy and low-cost publication of open data, leading them to propose a set of five design principles for improving the open data publishing process of public organizations: (i) start thinking about the opening of data at the beginning of the process; (ii) develop guidelines, especially about privacy and policy sensitivity of data; (iii) provide decision support by integrating insights into the activities of other actors involved in the publishing process; (iv) make data publication an integral, well-defined and standardized part of daily procedures and routines; and (v) monitor how the published data are reused. The goal of our approach is addressing principles (iii) and (v), since we provide a decision support framework based on activities of data consumers, which is useful for monitoring how datasets are being reused. Additionally, Jetzek et al. [29] propose a framework to explain how value is generated from open data. This framework is useful for governments to understand the value of their open data. Their framework is based on assessing the impact of open data based on two dimensions: (i) how openness generates value, and (ii) how society as a whole can get value from openness. The authors identify four different archetypical generative mechanisms (cause-effect relationship between open data and value) in their framework: transparency (open data helps to improve visibility to ensure socially responsible resource allocation), participation (open data as a mechanism for engaging stakeholders who help in solving social problems), efficiency (open data to improve how resources are used) and innovation (open data as a cornerstone for generating new ideas, processes, services and products). The authors claim that their framework can help governments in the development of their strategy for opening data by considering factors that can enable the generation of value from open data through the mechanism of innovation.

Therefore, there are several methods that support the CDOs in opening data, but to the best of our knowledge no approaches focus on supporting CDOs in selecting and prioritizing which datasets should be open according to their preferences. To overcome this drawback, in this paper, a novel approach for prioritization of open data publication is presented.

3 The POPSC (Prioritization of Open Data Publication in Smart Cities) Framework

This section proposes and describes the POPSC framework, which aims to serve as a model for supporting CDOs in the selection of the most appropriate datasets to publish in a Smart City open data portal. Concretely, the framework (shown in Fig 1) provides a way of prioritizing datasets according to the category in which they are classified, by taking into account different indicators of reuse of the datasets of the same category already published in similar open data portals and the objectives of the open data portal where data will be published. Thus, the framework identifies the steps to be followed to ensure that CDOs have a methodology that guides this decision-making process. These steps are described next.

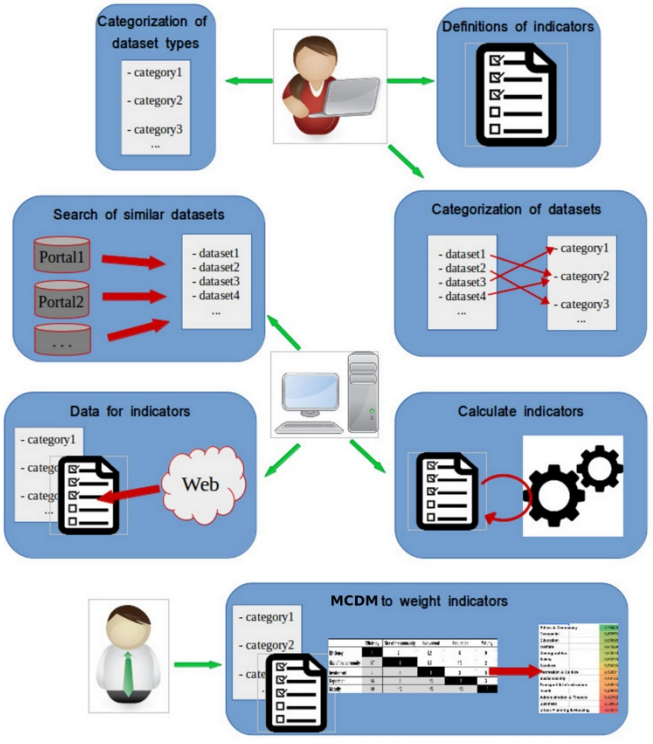

Fig. 1. General overview of the POPSC framework.

3.1 Categorization of the types of datasets

In this first stage, the CDO should define the taxonomy of dataset categories to analyze. This process involves information classification work, so the CDO can ask a knowledge engineer to help perform this task. It is advisable that this stage is carefully executed because results of the whole process are based on this taxonomy. Therefore, it is recommended that the categories composing this taxonomy follow some criteria of standardization that is as normalized as possible. In this sense, if a common shared worldwide recommendation of dataset categories existed, it could be the first choice. Alternatively, a good choice may also be a national or a regional recommendation. If none were found, a categorization might be collected from existing nearby institutions with similar characteristics. In any case, it might be

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2019.2893016, IEEE Transactions on Emerging Topics in Computing

possible that some adaptations are necessary to apply to the final selected taxonomy. Finally, if CDO eventually realized that the selected categories are not appropriate, then new categorization should be proposed and easily used in the rest of the framework.

3.2 Definitions of indicators for categories

Beyond institution policies, one of the main objectives of open data philosophy is to provide public data that could be reused by society and can generate some kind of "impact". In this sense, a dataset has "impact" when it produces some kind of benefit (either social or economic). Thus, "impact" criteria could be defined from various points of view: the point of view of the institution which publishes the dataset, the reusers who try to generate some kind of profit using them, or the end users who consume either raw open data or some of their reuses. The problem is that there is no standard way to measure dataset "impact". However, it would be desirable that CDOs understand how open data is reused, trying to align the goals of all of the stakeholders.

In this sense, the goal of this stage is to establish indicators to measure, in some way, the awareness, usefulness and interest of the reuses of the datasets and, consequently, a plausible comparison of datasets can be carried out. Thus, for every dataset, it would be very interesting to be able to measure criteria such as popularity, economic profit, transparency improvement, etc. However, it is not easy to establish reliable ways of measuring these factors. Due to open data initiative idiosyncrasy, data is usually published without establishing specific goals and without imposing utilization or authentication restrictions to the reusers. As a result, collecting the usage information and measuring this awareness, usefulness or interest generated by the reuses of open datasets may become quite complex. Therefore, it may be better to define only those indicators that can be obtained and measured in a reasonable way without spending more time or resources than in the process of publishing open data itself.

3.3 Search, categorization and estimation of dataset use

This stage is composed of four interrelated tasks: search similar organizations' datasets, categorize collected datasets, search data to calculate indicators and estimation of the use of the collected datasets in the studied context.

3.3.1 Search similar organizations' datasets

Once the indicators about the reuses have been established and defined, it is necessary to gather information. As in any scientific field, the more data collected for estimating an indicator, the more accurate the assessment.

This task may require a considerable amount of effort, it can either be done manually or executed by implementing a software tool which automates or at least, facilitates this task. Moreover, if the number of categories is high and a broad study of datasets is required, the effort required for collecting and classifying these datasets could be even greater. Obviously, without a software tool, this stage could be even more demanding.

In any case, this gathering of information could focus on datasets specifically related to the analyzed organization so as to obtain a more accurate assessment of the collected data. That is, for a Smart City portal, this search could be made in other city open data portals.

For every collected dataset, it would be necessary to obtain at least some kind of unique identifier and some kind of metadata containing a theme, a description, a keyword or something similar. On the one hand, the unique identifier will be necessary to find out the references to the dataset in the studied context. On the other hand, the metadata are necessary to classify the dataset according to the predefined categories.

3.3.2 Categorize collected datasets

After collecting the datasets, it is necessary that the CDOs classify them according to the categories established at stage 3.1. It should be noted that, in most cases, organizations describe themes or keywords of their datasets using their own words. Therefore, this task can hardly ever be done quickly or easily and may require thesauri search techniques, semantic matching techniques or manual supervision.

3.3.3 Search and acquire data to calculate indicators

After classifying the datasets, the next step is searching and acquiring the references to every dataset relating to the studied context. Depending on the nature of indicators, data can be acquired from different sources. For example, if indicators measure the awareness or the interest in reuses of datasets on social networks, then data can be acquired by searching references of datasets within Twitter [w2] or Facebook [w3]. If they measure the reuses on the web, then data may come from PublicWWW [w4] or NerdyData [w5], and if they measure the use of open data in OSS projects, then references can be searched (and data can be acquired) from Github[w6], etc.

3.3.4 Calculate indicators

When all possible references have been located and corresponding data have been acquired, the final step at this stage is to compute previously defined indicators (see Section 3.2) by grouping references (and corresponding data) according to the dataset categories. In order to use the ideal mode of any MCDM method, these values must be normalized.

3.4 Use of MCDM to weight the indicators

Once the indicators are set, in this final stage the framework proposes that CDOs use a MCDM method that allows them to establish the weights of importance of these indicators for their institutions and return a ranking of the different categories according to weights and the indicators. MCDM is the field of operational research

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2019.2893016, IEEE Transactions on Emerging Topics in Computing

wherein the decision alternatives are analysed with respect to a set of multiple (and often conflicting) criteria [30].

The CDOs could assign these weights according to the institution's strategic objectives. Thus, CDOs of Smart Cities may have different objectives, strategies and target audiences when deciding which datasets should have priority of publication. Each city has its own idiosyncrasy defining what is most important or of particular interest, and it is unlikely two cities share the same priorities with regard to their respective objectives. Cities can be characterized by their size, the importance of the tourism sector, or its residential, commercial or industrial sectors, etc. Smart Cities can orient policy making towards employment, welfare of citizens, quality of life, accessibility or towards promoting transparency [31].

For example, in cities focused on software reuses of their open data and according to each city's strategic choices, the publication of certain datasets may have priority in order to arouse greater interest amid a large number of developers, with the aim of developing numerous applications, or conversely, they may envision few applications but stable ones with guaranteed durability. Some of them may want to widely disseminate published datasets but other ones may encourage small groups of developers and provide them with more personalized support.

As a result of this final stage, the MCDM method allows CDOs to prioritize datasets in a reasonable way based on the data collected from similar organizations, the indicators taken into account and the open data strategy of the organization.

4 Framework Application using GitHub and Socrata

Our previously-described POPSC framework is only useful when is instantiated in a specific scenario. In this section, a specific application characterized by proposing the use of indicators based on the reuse of open datasets in Open Source Software (hereon OSS) projects is presented. These indicators are estimated using existing references to open datasets of some of the most important US cities retrieved from Github repositories. Concretely, this application of the POPSC framework has the following characteristics:

1. The datasets were classified according to a set of categories specifically designed for Smart Cities.
2. The proposed indicators attempted to measure the potential reuse of open datasets within OSS.
3. Similar organization datasets were obtained from 32 cities of the United States such as San Francisco, Chicago or New York which use Socrata as an open data repository.
4. The data about the use of the datasets within OSS source projects were obtained from Github.
5. The use of the Analytic Hierarchy Process (AHP) as MCDM method.

A repository containing all the scripts and detailed instructions needed to carry out a functional application of the POPSC Framework is available at Github[w7]. In this way, CDOs can replicate the whole process explained in this section using this repository and create their own up-to-date version of this application. Moreover, this concrete framework application to Smart Cities is not written on stone: although this application is completely functional and has been developed with the aim of being helpful to Smart City CDOs, it can be adapted to the necessities and preferences of each particular case. Consequently, any CDO could decide to add or delete some categories or indicators or even apply the framework completely differently using other categories, indicators, sources from gathered data or MCDM method. On the other hand, if CDOs simply want to use this application, with the original data, categories, indicators and AHP explained hereafter, they may do so directly [w8].

4.1 Open data categorization

There is no common agreement on the best way of classifying Smart City open datasets. However, in June 2013, the G8 [w9] proposed an Open Data Charter which, among other interesting points about open data, suggested 14 high-value data categories [32]. The G8 Open Data Charter is currently becoming the International Open Data Charter, which is supported by an increasing number of countries and institutions [w10]. Due to this, these categories seem to be a good way to classify Smart City datasets. Nevertheless, some of these categories, such as Global Development and Science and Research, might not be used in the Smart City context. Therefore, these G8 categories can be a good starting point but not the final one.

TABLE 1
Proposal of Open Data categories for Smart Cities

Id	Data Category	Example Datasets
1	Administration & Finance	Audits and Reports, City Finance and Budget, City Government, Fees, Liabilities and Assets, Purchasing, Revenue
2	Business	City Businesses, Community & Economic Development, Growing Economy, Regulated Industries
3	Demographics	Census, CitiStat, Forecasts, Neighborhoods, Statistics
4	Education	Schools, Youth
5	Ethics & Democracy	City Management and Ethics, Elections, Ethics, Expenditures, General Information, Governance, Government, Human Relations, Human Resources, Legislation, People, Permitting, Public Works, Taxes
6	Geospatial	Geographic Locations and Boundaries, Mapping, Location, GIS
7	Health	Public Health, Human Services, Social Services
8	Recreation & Culture	Arts and Culture, Events, Greenways, Historic Preservation, Library, Parks, Recreation, Tourism
9	Safety	Crime, Emergency, Fire, Police, Public Safety
10	Services	311 Call Center, City Services, Communi-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2019.2893016, IEEE Transactions on Emerging Topics in Computing

		ty, Customer Service, Facilities, Government Buildings and Structures, Inspectional Services, Public Property, Public Services, Service Requests
11	Sustainability	Energy and Environment, Natural Resources, Sustainability, Waste Management, Food, Agriculture
12	Transport & Infrastructure	Airports, City Infrastructure, Transportation, Parking, Streetcar, Traffic
13	Urban Planning & Housing	Area Plans, Buildings, City Facilities, City Parks and Tree Data, Construction, Development, Housing, Land Use, Urban Planning
14	Welfare	Insurance, Life Enrichment, Quality of Life, Pension, Retirement, Sanitation, Social Services

Thus, specific domains which can generate data within a Smart City must be taken into account. In this sense, a survey [33] about Smart City initiatives proposes a classification composed of 6 domains and 28 subdomains.

Establishing an exhaustive classification of open data categories for Smart Cities is beyond the scope of this paper. However, this work proposes an initial classification of open data categories for Smart Cities (shown in Table 1) aimed to be as close as possible to the G8 Open Data Charter but incorporating modifications to encompass the aforementioned domains and subdomains proper to Smart Cities. This proposed classification is given in Table 1 together with example datasets for each category.

4.2 Indicators for selecting open data to release

Strategy for opening data could prioritize publication of data which allows a community of developers to generate some kind of impact and effectively release benefits of open data through software projects [34].

A Smart City could in fact prioritize publication of open data with more reuse potential depending on the category to which the data belong to.

In this sense, one approach to measure which datasets are most likely to be reused, may consist in counting the number of times that they are reused within software applications and measuring some indicators (number of users, earned money, etc.). Interestingly, a good approach could be analyzing the reuse of datasets within the open source software (OSS) community, since OSS is considered to encourage the creation of SMEs and jobs [35]. Actually, the Tenth Annual Future of Open Source Survey [36] reflects the increasing adoption of open source and highlights the abundance of organizations participating in the open source community (e.g., 65% of companies currently participate in open source projects).

Consequently, it seems that an estimate of the use of the different categories of datasets by the OSS community could be a good indicator of their potential benefits. In the Smart City scenario, when CDOs make decisions on which data to publish, they could prioritize publication of data which allows a community of developers to create applications and services that are useful and interesting for citizens and effectively release benefits of open data through OSS projects.

However, it is not possible to know the exact number of OSS projects that reuse a dataset. Fortunately, many OSS projects make their source code available in public repositories that provide statistics about their use. Among the OSS repositories, one of the most used is Github [w1], the largest web-based distributed revision control and source code repository in the world, and the source of several empirical studies such as in Yu et al [37].

As a result, we developed a set of indicators so that Smart City CDOs measure which category of open data has more reuse potential and decide which data must be released according to the reuse requirements of each city. This is an initial proposal that can be refined by Smart City CDOs according to their requirements. The aim is to compare projects that use different categories of datasets and how successful they are.

Firstly, based on [38], we included (i) number of people who agree to receive information about the project because they find it interesting (subscribers), and (ii) number of people who actually work on the OSS project (developers). On the one hand, subscribers to OSS choose to obtain information on the project and thus reveal a deeper interest in the OSS project. The subscriber indicator not only measures interest within the project but the reputation of the project within the community and the dissemination of the project through the community. Thus, our first indicator is the **reputation** among a community of developers of OSS projects that reuse open data from a category. This indicator measures how well-known projects reusing data from some specific category are (within the community of developers). Smart City CDOs could be interested in opening data that will be reused in these kinds of projects in view of creating a community of developers around open data. On the other hand, the number of developers working on a project is critical to its success, since survival of an OSS project depends on continued contribution from developers [39]. Thus, our second indicator is the **involvement** which measures how often developers contribute to OSS projects that reuse some specific category of open data. The CDO of a Smart City with enough infrastructure could be willing to involve developers that actively collaborate in the development of projects and to support their success.

Secondly, based on [40], we included (iii) the amount of development activity (which is an important indicator to measure how involved a community is), and (iv) age of an active project that is positively related to OSS progress toward completion, as well as the experience of the community of developers. On one hand, based on the amount of development activity, our third indicator is the **size of the community** involved in projects that use data from a category. This indicator measures the number of developers that use open data from a given category. A city CDO could need to adapt the size of the community to the budget and available infrastructure. On the other hand, our fourth indicator is the **maturity** of projects that use an open data category. Maturity means that the community

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2019.2893016, IEEE Transactions on Emerging Topics in Computing

has been working on the project for some time without the project being abandoned. A Smart City CDO may want to select the datasets that help in promoting fewer projects stretching over longer periods of time, rather than promoting a larger number of short-term projects.

Finally, an additional indicator called **efficiency** has been developed in order to assess the probability of reuse of a dataset category, i.e. the likelihood of datasets from each category of being used. So, this indicator measures the probability of datasets of one category to be referenced by an OSS project. This indicator determines how relevant a category of datasets is. Smart City CDOs will use this indicator to know which categories of open data are most likely to be reused.

4.3 Search, categorization and estimation of dataset reuse

Once dataset indicators have been determined, we have to select which data sources will be used to measure them. As we are measuring the reuse potential of open data within OSS projects, we need two kinds of data sources: (i) already published Smart City datasets (and their metadata) and (ii) OSS projects (together with information about them) which reference the gathered datasets; i.e., we need to know which open datasets are used in which OSS projects. Assuming that our scenario is a Smart City, then we have to search open data from municipalities. To perform this search, we chose Socrata because it is one of the most used open data repositories, and notably by some of the most important US cities. We also tried to measure the existence of potential reusers within a community in order to measure open data reuses. To do this, we used Github, because, as mentioned earlier, it is probably the largest repository of OSS projects in the world.

4.3.1 Data from Socrata open data catalogs

Socrata[w11] is a software company focused "exclusively on democratizing access to public sector data around the world". It provides an Open Data Platform for allowing local, regional or national governments to release data. Socrata is a partner of the USA National League of Cities[w12] for the development of open data strategies[w13]. Nowadays, the Socrata Open Data Platform is used by some of the most important US cities such as New York, Chicago, San Francisco or Los Angeles. In this respect, Socrata is very useful as a proof-of-concept of our approach, since it is easier to collect open dataset identifiers and their metadata. In this sense, every Socrata dataset has its own endpoint and each is designated by a unique dataset identifier (a sort of primary key consisting of a code with eight alphanumeric characters split into two four-character phrases by a dash). Moreover, every Socrata open data portal provides an easy way to access a list of its published datasets. This list not only contains the identifier of every dataset but also useful metadata about it, such as the theme or the keyword of the dataset. We must remember that the metadata of open datasets are important because they are needed to facilitate the categorization step that comes next. On this basis, we decided to choose Socrata as our source for open datasets.

After making this decision, we drew up the process of gathering all the data needed from Socrata. In essence, this process was composed of the following steps:

1. Retrieving data from Socrata on institutions which use its Open Data Platform. Concretely, when this step was applied, 106 institutions were recovered.
2. Gathering and filtering the identifier and the minimal metadata needed to categorize them (theme or keyword) from every dataset published by US cities using Socrata. Concretely, when this step was performed, 8960 datasets from 32 different US cities met these conditions.

4.3.2 Categorization

As mentioned in section 3.3.2, gathered datasets must be categorized according to the categories pre-determined for the concrete application of the framework. In this case, the classification proposed in section 4.1 was used.

Due to its characteristics, this process requires the participation of experts to execute it adequately. The two research groups that developed this framework included researchers working in related fields such as open data and knowledge representation. These researchers were responsible for classifying the datasets following the steps described below:

1. Extracting different themes from US city datasets. In our case, 215 different themes were extracted.
2. Mapping every theme to one of the available categories. Themes without a clear fit had to be classified as 'Others' in order to be discarded later. When we performed this step, 211 themes could be mapped to the established categories and 4 were classified as 'Others'.
3. Automatically classifying datasets with a theme according to the mapping in step 2. In our case, 8299 datasets were classified according to the established categories, 11 were categorized as 'Others' and 650 were not categorized due to their lack of theme.
4. Optionally, trying to categorize datasets that have no theme manually, using other metadata such as keywords. This step can be carried out when the number of datasets without a theme is considered high enough to distort the value of the indicators. In our case, although the datasets without a theme represented less than 10% of the total, the experts decided to classify them manually, one by one.

As a result of this process, 8949 datasets were adequately categorized and 11 were discarded due to their unclear fit.

4.3.3 Collecting data from GitHub

In order to calculate the above-described indicators on the success of OSS projects that reuse open data, we decided to collect data from Github. Github is used by individu-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2019.2893016, IEEE Transactions on Emerging Topics in Computing

als, communities and businesses alike to develop software projects. GitHub is free to use for public and OSS projects, and it is used in studies on Software Engineering related to OSS success in several works such as [38] [41] [42]. Github has an API that is used to collect all required data from an OSS project. More specifically, the data can be acquired from repositories and from users. A repository is a kind of software project folder that contains all the project files. Valuable data from a repository that can be collected by using the API, apart from the code itself, are, among others, repository_id, stargazers_count, watchers_count, forks_count, subscribers_count, created_at, updated_at, total_contributors, total_contributions, etc. The indicators used in our implementation of the POPSC Framework are based on these data.

We established a process for determining which OSS projects were using open datasets from Socrata US Cities. Our process consists in the following steps (it was implemented by using the GitHub API within a Pentaho Data Integration [w14] process):

1. Searching every eight-character code from existing Socrata datasets belonging to USA cities (obtained as described in Section 3.3.1) based on code from OSS repositories hosted on Github in order to know which projects are reusing open data. When we performed this step, 350644 references were found from 2517 repositories to 5874 of the 8949 categorized datasets.
2. Gathering required data from Github on the repositories that reference open datasets to make an estimation of the indicators. In our case we found that 2501 of the 2517 repositories had all the data.

4.3.4 Calculating indicators

At this stage, we made an estimation of the indicators defined in section 4.2. So, this was the last step before applying AHP. To the best of our knowledge, this is the first time Github was used to estimate indicators related to reuse of open data in software projects.

We defined a process consisting in the following steps:

1. Discarding repositories that do not have all the required data to make an estimation of the indicators. When we performed this step, only 2501 repositories remained.
2. Discarding all repeated references to a specific dataset from a specific repository. When we performed this step, 32551 unrepeated references from 2501 repositories remained.
3. Making an estimation of the indicators. When we performed this step, we applied the following formulas:
   a. **Efficiency**. Understood as the proportion of datasets of each category referenced in Github.
   b. **Size of the community**. Understood as the average number of contributors of every repository that references datasets of the category.
   c. **Involvement**. Understood as the average number of contributions of every repository that references datasets of the category.
   d. **Reputation**. Understood as the average number of subscribers of each repository that references datasets of the category.
   e. **Maturity**. Understood as the average maturity of every repository referencing datasets of the category. Maturity is computed using 2 lifetimes, project lifetime (PL) and last update lifetime (LUL) and the formula is: PL/LUL.
4. Normalizing the estimated indicators in order to use the ideal mode of AHP. When we applied this step to our case, the indicator of each category was divided by the maximal value obtained by a category in the indicator. Thus, all the indicators of each category were normalized to a 0-1 range.

4.4 Using AHP to weight the indicators

AHP is a multiple criteria decision making method that has been used in many different applications related to decision making [43]. Some works specifically use AHP in Smart Cities and e-government. In this context, Bartolozzi et al. [44] present a DSS which uses AHP for supporting the decision-making process related to Smart City issues. Sultan et al. [45] suggest the use of AHP to decide the most appropriate technology for the development of e-government projects in Smart Cities. Boselli et al . [46] use AHP to rank the factors for innovating a smart-mobility service in the city of Milan. A very interesting use of AHP to evaluate open data portal quality can be found in Kubler et al. [47]. The authors propose taking into account different dimensions: completeness, openness, addressability and retrievability to assess the quality of 146 open data portals.

Although there are several applications of AHP to the domains of Smart Cities and e-governments, they all aim at assessing Smart City strategies and the quality of open data portals. Instead, POPSC Framework proposes AHP to recommend the most appropriate datasets to be published.

Concretely, this stage implicates, using AHP, taking the rows and columns of indicators from the previous phase, and assessing the relative importance between pairs of indicators. The result of this step will be the eigenvectors of each matrix, meaning the relative importance of each indicator. Subsequently, and for each category, the weights of importance of indicators calculated in the previous step, multiplied by the values of the indicators in the corresponding categories obtained at the end of stage 3.3.4 are used for calculating the suitability of publication of datasets in each category.

This value corresponds to a measure that takes into account the strategic criteria of the institution together with the indicators obtained for every dataset category. This assessment will produce a suitability ranking list of dataset categories to publish.

Different CDOs may have to address a diversity of contexts in their cities implying an array of different strategic

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2019.2893016, IEEE Transactions on Emerging Topics in Computing

objectives. Cáceres for example, a town in Spain, provides an open data portal with many high-quality datasets but the portal is rather unknown, and the technological fabric of the city is composed of small IT companies. Therefore, the goal of the CDO could be to extend the use of the open data portal by prioritizing those datasets that are likely to generate a large number of projects -though simpler ones that involve fewer people. On the other hand, a big city with consolidated open data portals (such as Madrid or Barcelona) may prefer opening datasets that could be used in complex and mature software applications that involve big teams, since it is more relevant to their specific technological industry context.

Therefore, taking into account the characteristics and objectives of the city, the CDO should weigh the importance of the indicators set out in the previous steps in order to determine their relative importance. CDOs may logically wish to assign maximum values to each of the five indicators but, in this phase, they should make the effort of assessing the importance they have relatively to each other.

5 A Test Case: Decision Making on Opening Datasets In a Medium-Size Smart City

To test the applicability of the framework, we wanted to use them in a real case of a medium-sized city council. This took place in a Spanish city of around one hundred thousand inhabitants. The city has had an open data portal since 2014. Although they are interested in publishing as much information as possible, decision criteria for prioritizing datasets had not been determined at any time since the inception of the portal.

The city employs a manager in charge of integrating the council data and the data from the GIS department. He plays the role of the city's CDO. His department is composed of six software and geography engineers. His budget for publishing and maintaining open datasets is very low. To date, the CDO does not take into account the global demand for data coming from the community of developers, or external interest in each data category, nor does he analyze and justify priority of publication of some datasets over others. He uses his intuition and external requests (via Twitter or emails from developers) to take the decision of dataset publication.

The context was thus conducive to involve this CDO in applying the POPSC framework. Thanks to the fact that datasets and indicators had already been categorized as described in the previous section, the manager only had to execute the step of assigning weights of importance to the indicators.

The CDO of the city had a highly technical profile and sufficient knowledge to understand these concrete five indicators. He needed time to analyze and assimilate each of the concepts, but before starting the assessment of the AHP matrix, he said he was able to establish the relative levels of importance of the five indicators. He used the spreadsheet that implements the AHP method, producing the values shown in Fig 2.

Reciprocal Matrix

	Efficiency	Size of the community	Involvement	Reputation	Maturity
Efficiency	1	7	1/2	5	9
Size of the community	1/7	1	1/5	1/3	2
Involvement	2	5	1	3	5
Reputation	1/5	3	1/3	1	3
Maturity	1/9	1/2	1/5	1/3	1
Sum	3,454	16,500	2,233	9,667	20,000

Normalized Relative Weight

	Efficiency	Size of the community	Involvement	Size of the community	Involvement
Efficiency	0,290	0,424	0,224	0,517	0,450
Size of the community	0,041	0,061	0,090	0,034	0,100
Involvement	0,579	0,303	0,448	0,310	0,250
Reputation	0,058	0,182	0,149	0,103	0,150
Maturity	0,032	0,030	0,090	0,034	0,050
Sum	1,000	1,000	1,000	1,000	1,000

Consistency Ratio=
9%
The Incosistency Is Acceptable

	Normalized Principal Eigen Vector
	Priority Vector
Efficiency	0,3809772919
Size of the community	0,06520027043
Involvement	0,3780360885
Reputation	0,1284849202
Maturity	0,04730142908

Fig. 2. Assignment by the CDO of the relative importance of indicators.

As shown, the CDO did assign the relative importance of the indicators with moderation, avoiding extreme values, resulting in an eigenvector with high values in Involvement and Efficiency and low values in Size of Community and Maturity.

The final result of these weights multiplied by the values of indicators in each category according to data from Github, is shown in Fig 3. In this case, application of the POPSC led to the recommendation to publish datasets related to "Ethics & Democracy" first, to "Geospatial", second, etc.

After going through the evaluation process, we asked him about the answers given, in an attempt to analyze the importance values assigned by peers. Following the interview, we understood that his department had a very limited budget, but that it was stable over time, so data useful to citizens through the development of applications by small local enterprises was given preference. Hence, he assigned high weights to effectiveness.

FINAL RECOMMENDATION:	
Ethics & Democracy	0,7672003498
Geospatial	0,6994274703
Education	0,6832903108
Welfare	0,6476995047
Safety	0,6392093807
Demographics	0,6389276539
Services	0,6055289323
Recreation & Culture	0,5297485997
Sustainability	0,5037035936
Transport & Infrastructure	0,4955444574
Health	0,4711124567
Administration & Finance	0,4371655657
Business	0,4136070592
Urban Planning & Housing	0,3902801185

Fig. 3. Final recommendation of the POPSC framework for the open data portal of the city.

However, taking into account these groups of local developers, local municipal politics aim at boosting em-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2019.2893016, IEEE Transactions on Emerging Topics in Computing

ployment at the regional level, rather than favoring use by a large community of developers. The portal is relatively young, so they are not particularly looking for the continued use of apps, but rather to encourage the development of new apps. Due to all the above, high weights were also given to Involvement.

The CDO decided to establish his dataset publishing policy prioritizing the category, recommended by the POPSC framework, related to municipal contracts (Ethics & Democracy). It was also decided that the publication and maintenance of their abundant geospatial datasets would be kept, postponing other datasets which were on the list of potential pending publication datasets.

The CDO's feedback revealed that assessing indicators is complicated and quite subjective; however, the CDO also commented that, once ideas were understood, the framework helped to justify decisions made during the process, helped to carefully analyze publication strategy, and to weigh up the advantages and obstacles in choosing one dataset category over another.

6 Conclusions

Institutional CDOs, who are responsible for publishing information in open data portals, usually have a limited budget and insufficient time to release and maintain all available data. Simply trying to publish "all" data is not even considered a good strategy. At present, CDOs usually rely on their intuition concerning potential data reuse to decide what datasets must be prioritized. Obviously, if there are specific policies and guidelines in their institution, they should publish the specific datasets to meet this objective. But, once these mandatory datasets are published, it is desirable they have available some kind of methodology that helps them to choose those datasets most likely to generate a given kind of interest. Importantly, the CDO should take into account different relevant user groups (citizens, companies, software application developers, journalists, NGOs, etc.) that may be interested in the reuse of the data. Each of them can have different objectives such as improving transparency, doing business, improving public services, etc., and the CDO should take this whole context into account when prioritizing data publication. It is worth noting that our proposal focuses specifically on the interest of using datasets for application developers, but the framework could be expanded to consider other points of view as future work.

Therefore, in this paper, we presented the POPSC (Prioritization of Open data Publication in Smart Cities) Framework. Its goal is to provide CDOs with a generic and methodological way of making decisions about what type of datasets should be given priority of publication in an open data portal. Although in the first place the CDO shoul prioritize those datasets that are established in the regulations and strategies of their institutions, an objective process to measure the demand for the use of the data can be a very useful tool for the CDO. The process is designed to take into account objective criteria rather than relying on the intuition of CDOs. To do this, this framework specifies actions to perform in order to provide a decision support system for CDOs for prioritizing datasets.

In addition, we have developed a fully functional application of the POPSC Framework that is oriented to the software developer community of reusers and that is characterized by:

1. A classification of 14 categories for Smart City open datasets based on the G8 Open Data Charter and the Smart City domain.
2. A definition of 5 indicators based on the reuse of datasets in OSS projects.
3. Almost 9000 open located datasets of many of the most important US cities.
4. A catalogue of these US city datasets classified according to the proposed categories.
5. Around 32000 distinct references from 2500 different Github projects referencing two thirds of the categorized datasets found, based on a search performed over all OSS projects in Github.
6. An estimation of the defined indicators of reuse of every Smart City dataset category.
7. An AHP-based Decision Support System to recommend Smart City dataset categories to prioritize, taking into account the estimated indicators and the importance of each indicator for CDOs.

Regarding application domain, our framework has been aimed at helping CDOs to prioritize the publication of datasets in a Smart City, but as work in progress, we want to apply the same idea to the universities domain. Universities are complex institutions with a variety of heterogeneous data that must be collected, classified and opened, ranging from (anonymized) data about students and staff, to geographical campus data to or financial data and so on. These data can be consumed by several actors: researchers, former students, citizenship, companies (or entrepreneurs) that reuse data to add value, and the own university. We are working on categorizing the datasets of the universities, identifying the sources of data from where we are going to collect the information, and establishing the appropriate indicators to be useful in the prioritization process.

In the Smart Cities domain, we want to emphasize that the specific implementation of this application of the proposed framework is completely reproducible. So, if they wish, CDOs can reuse and adapt them to their concrete requirements regardless of whether they work in a Smart City or in any other type of institution. With our framework, if a CDO has possibilities to access additional information about the use and interest of datasets (such as access to the source code that uses datasets, mentions in social media, news, strategic reports, etc.), could apply the idea of the POPSC method to prioritize the publication of datasets using this information. Similarly, a CDO

can consider other measures of use and interest apart from the five indicators we propose. In this case, CDO should link the formulas of the indicators with the information sources where the data resides. Considering other useful indicators and the corresponding data sources to calculate them remains as an interesting challenge for future work.

Regarding our proposal of open data categories for Smart Cities, if a CDO wants to modify them, he/she should only assign datasets to the new categories and apply the POPSC framework.

Further alternative future work from our framework that can be considered as a continuation of this research may include:

1. Searching and categorizing open datasets of different cities, regions, countries, companies or any other kind of institutions.
2. Developing semantic-based software tools for automatic classification of datasets.
3. Analyzing the reuse of open datasets in proprietary software projects, for instance, by developing an app web repository where developers could register their applications that use open data and indicating which particular datasets are reused.
4. Analyzing the reuses of open datasets in mass media, social media, blogs, etc. by searching the references to the datasets in these sites.

In summary, a successful publication of open datasets should be based on the proper combination of the objectives of the CDOs in charge of open data portals and the analysis of the reuses of already available open datasets. The POPSC framework has been designed with the objective of providing this basis to CDOs.

ACKNOWLEDGMENT

We would like to thank GitHub that allowed us to use its API without limitations. This work has been developed with the support of (i) Ministerio de Economía y Competitividad- European Regional Development Fund (ERDF): Project TIN2015-6957-R and Project TIN2016-78103-C2-2-R, (ii) POCTEP 4IE (0045-4IE-4-P) and (iii) Consejería de Economía e Infraestructuras/Junta de Extremadura - Fondo Europeo de Desarrollo Regional (FEDER)- IB16055 project and GR15098.

REFERENCES

[1] G. Vickery, "Review of Recent Studies on PSI-Re-use and Related Market Developments," *Inf. Econ. París*, pp. 1–44, 2011.

[2] M. Janssen, Y. Charalabidis, and A. Zuiderwijk, "Benefits, Adoption Barriers and Myths of Open Data and Open Government," *Inf. Syst. Manag.*, vol. 29, no. 4, pp. 258–268, Sep. 2012.

[3] European Parliament and Council of the European Union, "Directive 2013/37/EU of the European Parliament and of the Council of 26 June 2013 amending Directive 2003/98/EC on the re-use of public sector information," *Official Journal of the European Union*, 2013. .

[4] E. O. of the P. of the U. States, "Executive Order -- Making Open and Machine Readable the New Default for Government Information," 2013.

[5] S. Martin, M. Foulonneau, S. Turki, and M. Ihadjadene, "Risk Analysis to Overcome Barriers to Open Data," *Electron. J. e-Government*, vol. 11, no. 2, pp. 348–359, 2013.

[6] P. Conradie and S. Choenni, "On the barriers for local government releasing open data," *Gov. Inf. Q.*, vol. 31, no. SUPPL.1, pp. S10–S17, 2014.

[7] J. Attard, F. Orlandi, S. Scerri, and S. Auer, "A systematic review of open government data initiatives," *Gov. Inf. Q.*, vol. 32, no. 4, pp. 399–418, 2015.

[8] L. Bargiotti, M. De Keyzer, S. Goedertier, and N. Loutas, "Value based prioritisation of Open Government Data investments," 2014.

[9] W. Carrara, W. San Chan, S. Fischer, and E. van Steenbergen, "Creating Value through Open Data: Study on the Impact of Re-use of Public Data Resources," 2015.

[10] L. E. C. Serra, "The mapping, selecting and opening of data: The records management contribution to the Open Data project in Girona City Council," *Rec. Manag. J.*, vol. 24, no. 2, pp. 87–98, Jul. 2014.

[11] A. Zuiderwijk, M. Janssen, and Y. K. Dwivedi, "Acceptance and use predictors of open data technologies: Drawing upon the unified theory of acceptance and use of technology," *Gov. Inf. Q.*, vol. 32, no. 4, pp. 429–440, 2015.

[12] A. Cocchia, "Smart and Digital City: A systematic Literature Review," in *Smart City: How to Create Public and Economic Value with High Technology in Urban Space?*, Springer International Publishing, 2014, p. 239.

[13] C. E. A. Mulligan and M. Olsson, "Architectural implications of smart city business models: An evolutionary perspective," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 80–85, Jun. 2013.

[14] S. Zygiaris, "Smart City Reference Model: Assisting Planners to Conceptualize the Building of Smart City Innovation Ecosystems," *J. Knowl. Econ.*, vol. 4, no. 2, pp. 217–231, Jun. 2013.

[15] M. Kassen, "A promising phenomenon of open data: A case study of the Chicago open data project," *Gov. Inf. Q.*, vol. 30, no. 4, pp. 508–513, 2013.

[16] A. Zuiderwijk and M. Janssen, "Barriers and Development Directions for the Publication and Usage of Open Data: A Socio-Technical View," in *Open Government*, vol. 4, New York, NY: Springer New York, 2014, pp. 115–135.

[17] J. Thorsby, G. N. L. Stowers, K. Wolslegel, and E. Tumbuan, "Understanding the content and features of open data portals in American cities," *Gov. Inf. Q.*, 2016.

[18] A. Zuiderwijk, M. Janssen, S. Choenni, and R. Meijer, "Design principles for improving the process of publishing open data," *Transform. Gov. People, Process Policy*, vol. 8, no. 2, pp. 185–204, May 2014.

[19] A. Hjalmarsson, N. Johansson, and D. Rudmark, "Mind the gap: Exploring stakeholders' value with open data assessment," in *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2015, pp. 1314–1323.

[20] Y. Lee, S. E. Madnick, R. Y. Wang, F. Wang, and H. Zhang, "A Cubic Framework for the Chief Data Officer (CDO): Succeeding in a World of Big Data," *MIS Q. Exec.*, vol. 13, no. 1, 2014.

[21] M. Köksalan, J. Wallenius, and S. Zionts, "An Early History of Multiple Criteria Decision Making," *J. Multi-Criteria Decis. Anal.*, vol. 20, no. 1–2, pp. 87–94, Jan. 2013.

[22] D. E. Avison, F. Lau, M. D. Myers, and P. A. Nielsen, "Action research," *Commun. ACM*, vol. 42, no. 1, pp. 94–97, Jan. 1999.

[23] W. Yoland, "What is Participatory Action Research?," *Action Res. Int.*, vol. Paper 2, 1998.

[24] N. Padak and G. Padak, "Guidelines for Planning Action Research Projects. Research to Practice," Oct. 1994.

[25] P. Runeson and M. Höst, "Guidelines for conducting and reporting case study research in software engineering," *Empir. Softw. Eng.*, vol. 14, no. 2, pp. 131–164, 2009.

[26] C. Wohlin, "Guidelines for snowballing in systematic literature studies and a replication in software engineering," in *Proceedings of the 18th International Conference on Evaluation and Assessment in*

Software Engineering - EASE '14, 2014, pp. 1–10.
[27] M. A. Hossain, Y. K. Dwivedi, and N. P. Rana, "State of the Art in Open Data Research: Insights from Existing Literature and a Research Agenda," *J. Organ. Comput. Electron. Commer.*, vol. 26, no. 1–2, pp. 14–40, Apr. 2016.
[28] A. Zuiderwijk and M. Janssen, "A Coordination Theory Perspective to Improve the Use of Open Data in Policy-Making," in *Proceedings of the 12th IFIP WG 8.5 International Conference on Electronic Government - Volume 8074*, Springer-Verlag New York, Inc., 2013, pp. 38–49.
[29] T. Jetzek, M. Avital, and N. Bjorn-Andersen, "Data-driven innovation through open government data," *J. Theor. Appl. Electron. Commer. Res.*, vol. 9, no. 2, pp. 100–120, Aug. 2014.
[30] A. Ishizaka and S. Siraj, "Are multi-criteria decision-making tools useful? An experimental comparative study of three methods," *Eur. J. Oper. Res.*, vol. 264, no. 2, pp. 462–471, Jan. 2018.
[31] T. Bakici, E. Almirall, and J. Wareham, "A Smart City Initiative: The Case of Barcelona," *J. Knowl. Econ.*, vol. 4, no. 2, pp. 135–148, Jun. 2013.
[32] Group of Eight, "G8 Open Data Charter," 2013.
[33] P. Neirotti, A. De Marco, A. C. Cagliano, G. Mangano, and F. Scorrano, "Current trends in smart city initiatives: Some stylised facts," *Cities*, vol. 38, pp. 25–36, 2014.
[34] A. Zuiderwijk, I. Susha, Y. Charalabidis, P. Parycek, and M. Janssen, "Open data disclosure and use : critical factors from a case study," in *In: CeDEM 2015: Proceedings of the International Conference for E-Democracy and Open Government 2015*, 2015, pp. 197–208.
[35] R. A. Ghosh, "Economic impact of open source software on innovation and the competitiveness of the Information and Communication Technologies (ICT) sector in the EU," 2006.
[36] J. Hammond, P. Santinelli, J. J. Billings, and B. Ledingham, "The Tenth Annual Future of Open Source Survey," 2016.
[37] L. Yu, A. Mishra, and D. Mishra, "An Empirical Study of the Dynamics of GitHub Repository and Its Impact on Distributed Software Development," in *Proceedings of the Confederated International Workshops on On the Move to Meaningful Internet Systems: OTM 2014 Workshops - Volume 8842*, Springer-Verlag New York, Inc., 2014, pp. 457–466.
[38] R. Sen, S. S. Singh, and S. Borle, "Open source software success: Measures and analysis," *Decis. Support Syst.*, vol. 52, no. 2, pp. 364–372, 2012.
[39] C. Subramaniam, R. Sen, and M. L. Nelson, "Determinants of open source software project success: A longitudinal study," *Decis. Support Syst.*, vol. 46, no. 2, pp. 576–585, Jan. 2009.
[40] K. J. Stewart, A. P. Ammeter, and L. M. Maruping, "Impacts of license choice and organizational sponsorship on user interest and development activity in open source software projects," *Inf. Syst. Res.*, vol. 17, no. 2, pp. 126–144, Jun. 2006.
[41] T. F. Bissyande, F. Thung, D. Lo, L. Jiang, and L. Reveillere, "Popularity, interoperability, and impact of programming languages in 100,000 open source projects," in *Proceedings - International Computer Software and Applications Conference*, 2013, pp. 303–312.
[42] J. Marlow, L. Dabbish, and J. Herbsleb, "Impression Formation in Online Peer Production : Activity Traces and Personal Profiles in GitHub," in *16th ACM Conference on Computer Supported Cooperative Work*, 2013, pp. 117–128.
[43] O. S. Vaidya and S. Kumar, "Analytic hierarchy process: An overview of applications," *Eur. J. Oper. Res.*, vol. 169, no. 1, pp. 1–29, 2006.
[44] M. Bartolozzi, P. Bellini, P. Nesi, G. Pantaleo, and L. Santi, "A Smart Decision Support System for Smart City," in *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, 2015, pp. 117–122.
[45] A. Sultan, K. A. AlArfaj, and G. A. AlKutbi, "Analytic hierarchy process for the success of e-government," *Bus. Strateg. Ser.*, vol. 13, no. 6, pp. 295–306, Nov. 2012.
[46] R. Boselli, M. Cesarini, F. Mercorio, and M. Mezzanzanica, "Applying the AHP to Smart Mobility Services: A Case Study," in *Proceedings of 4th International Conference on Data Management Technologies and Applications - Volume 1: KomIS*, 2015, pp. 354–361.
[47] S. Kubler, J. Robert, Y. Le Traon, J. Umbrich, and S. Neumaier, "Open Data Portal Quality Comparison using AHP," in *Proceedings of the 17th International Digital Government Research Conference on Digital Government Research - dg.o '16*, 2016, pp. 397–407.

Web References

[w1] https://github.com/
[w2] https://dev.twitter.com/rest/public/search
[w3] https://developers.facebook.com/docs/graph-api/using-graph-api
[w4] https://publicwww.com/
[w5] https://nerdydata.com/search
[w6] https://developer.github.com/v3/search/
[w7] https://goo.gl/tfoUMT
[w8] https://goo.gl/uNKOtU
[w9] http://www.g8.utoronto.ca/what_is_g8.html
[w10] http://opendatacharter.net/history/
[w11] https://socrata.com/
[w12] http://www.nlc.org/about-nlc
[w13] https://goo.gl/6Jd5Gw
[w14] http://community.pentaho.com/projects/data-integration/

Álvaro E. Prieto is member of the Quercus Software Engineering Group and assistant professor of Computer Languages and Systems at the University of Extremadura (Spain). He received his BSc in Computer Science from the University of Extremadura in 2000 and a PhD in Computer Science in 2013. His research interests include open data, linked open data, ontologies and workflows. He is currently involved in various R&D&I projects.

Jose-Norberto Mazón is member of the WaKe research group at the University of Alicante (Spain). His research work focuses on open data, data integration and business intelligence within "big data" scenarios. He has published his research in international journals, such as Decision Support Systems, Information Sciences or Data & Knowledge Engineering. Finally, he is doing tasks of CDO in the open data project at the University of Alicante (http://datos.ua.es).

Adolfo Lozano-Tello is member of the Quercus Software Engineering Group and assistant professor of Computer Languages and Systems at the University of Extremadura (Spain) and Ph.D. (2002) at Computer Science. He is Director of International LINUX Center since 2006. His research interests include ontological engineering, semantic web and open linked data. He has published more than 100 papers on the above issues on Software Engineering and Knowledge Engineering. He is CDO in the open data project at the University of Extremadura (http://opendata.unex.es/).