Supporting open dataset publication decisions based on Open Source Software reuse

Alvaro E. Prieto
Universidad de Extremadura
Cáceres, Spain
aeprieto@unex.es

Jose-Norberto Mazón
Universidad de Alicante
San Vicente del Raspeig, Alicante, Spain
jnmazon@dlsi.ua.es

Adolfo Lozano-Tello
Universidad de Extremadura
Cáceres, Spain
alozano@unex.es

Luis-Daniel Ibáñez
University of Southampton
Southampton, United Kingdom
l.d.ibanez@southampton.ac.uk

ABSTRACT

Publishing and maintaining open data is a costly task for public institutions, that becomes even more challenging in the context of Smart Cities, where large amounts of varied data are generated from different domains. To optimize resources, they should prioritize the publication and maintenance of datasets most likely to generate social and economic impact. However, there is currently a lack of decision-support tools to help public sector data publishers to evaluate datasets on the light of their particular reuse goals. In this paper, we propose to suggest to data publishers the dataset categories with most potential impact, based on the impact of already published datasets of the same category. To measure impact, we propose a set of indicators based on the amount and quality of Open Source Software projects that use datasets. To aggregate indicators according to specific reuse goals, we provide an Analytic-Hierarchy-Process based tool.

1 INTRODUCTION

One of the most important challenges faced by Smart Cities is creating an ecosystem of public and private actors that reuse open data in order to produce IT services and products that both (i) would improve citizens' quality of life and (ii) would contribute to economic growth [32]. However, few open data portals in cities currently track data usage and consider the impact of data on deciding which datasets to maintain or what complementary datasets publish. Cities are not even aware of what kinds of apps are developed, using what data, and how many there are. Answering these questions is a significant research issue [30] that would allow prioritizing which categories of data must be published and maintained with respect to the applications that use them (i.e., impact that a category of open data generates).

To reverse this situation, publishing datasets as open data requires a decision support system to select those categories of datasets that offer higher potential to generate value [12]. Such a system must consider indicators about the impact of the already published open datasets, as well as the strategy of the Smart City. E.g., a small town could provide 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 city could be to extend the use of the open data portal by prioritizing those datasets that belong to categories 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 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.

Unfortunately, to the best of our knowledge, Smart Cities lack such decision support system, mainly because the process of calculation of those indicators that would use the system is not a trivial task. According to Janssen et al. [14] , "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 open data is reused, thus lacking knowledge about the impact generated by reusing the published open data.

More reasonable indicators of the use of open datasets could help to identify which categories of datasets have more possibilities of being reused and, in this way, generate some type of economic impact to people or enterprises. In this sense, good indicators could come from the reuse of datasets within the open source community. The Tenth Annual Future of Open Source Survey [11] reflects the increasing adoption of pen source and highlights the abundance of organizations participating in the open source community. Concretely, this survey estimates that 65% of companies currently participate in open source projects. Open Source Software (hereon OSS) is considered to encourage the creation of SMEs and jobs, by providing a skills development environment valued by employers and retaining a greater share of generated value locally [8]. Focusing in Europe, a study estimated that the contribution of OSS to its economy was of 450 billion euro per year [7].

Based on these figures, an estimation of the use of the different categories of datasets by the OSS community could be a good indicator of their potential impact. Therefore, when Smart Cities make decisions on which data to publish, they could prioritize publication of data which allows a community of developers to generate impact and effectively release benefits of open data through OSS projects.

In this paper, we present an approach based on the estimation of indicators of the use of open datasets in OSS projects. The goal of this approach is to provide Smart Cities with a Decision Support System which provides an ordered list of categories of datasets most suitable to be published or maintained in their open data portal. To do so, we have carried out a set of actions aimed at estimating useful impact indicators related to the datasets of the same category already published by open data portals of other cities. Concretely, to calculate our proposed indicators we needed two kinds of data sources: (i) already published Smart City datasets (and their metadata) and (ii) OSS projects (together with information about them) which referenced the gathered datasets; i.e., we needed to know which open datasets were being used in which OSS projects. To collect already published open datasets, we chose Socrata [26] because it is one of the most used open data repositories, and notably by some of the most important US cities. We also measured the existence of potential reuses within a community in order to measure open data impact. To do this, we used GitHub [9], because it is 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. [33].

Using the indicators obtained from these sources, we provide an Analytic Hierarchy Process (hereon AHP)-based [24] tool[1] that allows decision makers weigh these indicators, taking into account the reuse objectives of the city, to offer an ordered list of categories of datasets recommended to publish.

This paper is structured as follows: section 2 describes a new approach to select the most relevant categories of data to be published in a smart city open data portal. Section 3 presents toy samples of two different stereotypical smart cities using our approach and, to finish, section 4 summarizes other work related to the publishing of open data in Smart Cities.

2 USING REUSE INDICATORS BASED ON DATA FROM OSS PROJECTS IN GITHUB FOR SELECTING DATASETS TO OPEN

This section describes the steps that have been carried out to get an AHP process that allows classifying categories of dataset based on the preferences of the decision-maker. These preferences are applied to a set of useful indicators obtained from data about their reuse in OSS Projects of GitHub repositories. Concretely, these steps[2] are detailed in the following subsections and are summarized below:

(1) From GitHub repositories, studying the characteristics of OSS projects that use open datasets. This information was analyzed to establish a set of reuse indicators.
(2) Gathering datasets from 32 cities of the United States (such as San Francisco, Chicago or New York) which use Socrata as an open data repository. With respect to this point, it should also be noted that, although these cities are from the same country, United States, they have different cultural, social and economic characteristics that make us consider that the results obtained from their data are enough scalable to other Smart Cities located in different countries.
(3) Classifying the datasets according to a set of categories specifically designed for Smart Cities.
(4) Searching for references to the datasets obtained from Socrata in GitHub to calculate the indicators.
(5) With the reuse indicators established in step 1 as criteria, and the values from step 4, we have created a Google Spreadsheet [w3] based on AHP that allows decision makers to prioritize the most relevant categories of datasets that must be published in a smart city open data portal.

2.1 A proposal of indicators of reuse based on GitHub

Smart Cities should follow a strategy for opening data as described in [17]. This strategy should prioritize publication of data which allows a community of developers to generate impact and effectively release benefits of open data through OSS projects [37]. 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. However, due to "open-data by default" idiosyncrasy [23], data is usually published without establishing specific goals and without imposing utilization or authentication restrictions to the infomediaries and end users. As a result, collecting the usage information and measuring impact generated by open datasets may become very complex.

To overcome this situation, our approach is based on considering that the more used an open dataset is by OSS projects, the more impact is generated. Therefore, we borrowed some well-known indicators that measure the success of OSS projects and we have used as starting point to develop our indicators to measure such success when open data is reused. Then, these indicators allow Smart Cities to measure which categories of open data have more reuse potential and decide which data must be released according to the requirements of each city. The following **indicators** from existing research literature on OSS are considered [27] [28]. First of all, 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. 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 [28]. There is another measure for the success of OSS projects [27] as the (iii) age of an active project that is positively related to OSS progress toward completion, as well as the experience of the community of developers.

Based on these three indicators described in the literature about success of OSS projects, we developed a set of three indicators that measure the success of open source projects that reuse open datasets (they are summarized in Table 1). The aim is to compare projects that use different categories of datasets and how successful they are. First of all, we define the **reputation** among a community of developers of OSS projects that reuse open data from a category. Some projects that reuse open data from some specific categories can be perceived by developers as being highly appealing projects. Smart Cities are interested in opening data that will be reused in these kinds of projects in view of creating a community around open data, thus allowing an open data portal to attract the attention of potential developers. Therefore, the reputation indicator measures how well-known projects reusing data from some specific category are (within the community of developers). Furthermore, the **size of the community** involved in projects that use data from a category is defined in terms of the size of the community of developers that use open data from a given category. A city needs to adapt the size of the community to the budget and available infrastructure. Finally, **maturity** of projects that use an open data category is

[1] https://goo.gl/HcUc1e
[2] A repository containing all the scripts and detailed instructions needed to carry out a functional application of our approach is available at GitHub https://goo.gl/TDp1xi

Table 1: Proposed indicators and their definitions

Indicator	Description
Reputation	Average number of subscribers of each repository that references datasets of the category
Community size	Average number of contributors of every repository that references datasets of the category
Maturity	Average maturity of every repository referencing datasets of the category. Maturity is computed using 2 lifetimes, project lifetime (PL) and last update lifetime (LUL). Thus, the resulting formula is: PL/LUL
Efficiency	Proportion of datasets of each category referenced in GitHub

proposed. Maturity means that the community has been working on the project for some time without the project being abandoned. A Smart City 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.

An additional indicator has been developed in order to assess the impact of a dataset category, i.e. the likelihood of datasets from each category of being used. To do so, we defined efficiency of an open data category, as 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 Cities will use this indicator to know which categories of open data are most likely to be reused. Therefore, in a scenario where the Smart City has the chance of opening a large number of datasets, the **efficiency** indicator will become secondary to the publishing efforts regarding a wide a variety of datasets.

As aforementioned, these indicators come from well-known indicators from the OSS community, being thus completely generalizable to be used in any OSS repository. It is worth noting that our proposal of indicators is not set in stone, consequently more indicators could be created and checked to be used by Smart Cities according to their requirements.

2.2 Search of smart city datasets on Socrata

Once the impact measuring indicators have been established and defined, information should be gathered. This gathering of information focuses on datasets specifically related to the smart cities so as to obtain a more accurate assessment of the collected data.

Socrata 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 [22] for the development of open data strategies. 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 possible to collect precisely 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. Every Socrata open data portal provides a list of its published datasets containing the identifier of every dataset and useful metadata about it, such as the theme or the keyword of the dataset. These metadata of open datasets are important because they are needed to facilitate the categorization step that comes next. To collect the data from Socrata, we followed these steps:

(1) Retrieve data from Socrata on institutions which use its Open Data Platform. 106 institutions were recovered.
(2) Gather and filter the identifier and the minimal metadata needed to categorize them (theme or keyword) from every dataset published by US cities using Socrata. 8960 datasets from 32 different US cities met these conditions.

Table 2: G8 Open Data Categories

Id	Data Category	Example Datasets
1	Companies	Company/business register
2	Crime and Justice	Crime statistics, safety
3	Earth observation	Meteorological/weather, agriculture, forestry, fishing, and hunting
4	Education	List of schools; performance of schools, digital skills
5	Energy and Environment	Pollution levels, energy consumption
6	Finance and contracts	Transaction spend, contracts let, call for tender, future tenders, local budget, national budget (planned and spent)
7	Geospatial	Topography, postcodes, national maps, local maps
8	Global Development	Aid, food security, extractives, land
9	Government Accountability and Democracy	Government contact points, election results, legislation and statutes, salaries (pay scales), hospitality/gifts
10	Health	Prescription data, performance data
11	Science and Research	Genome data, research activity, experiment results
12	Statistics	National Statistics, Census, infrastructure, wealth, skills
13	Social mobility and welfare	Housing, health insurance and unemployment benefits
14	Transport and Infrastructure	Public transport timetables, access points broadband penetration

2.3 Categorization fo datasets

In this step, we had to choose the taxonomy of dataset categories to be analyzed. There is no common agreement on the best way of classifying Smart City open datasets. However, a 14 high-value data categories is suggested by the G8 Open Data Charter [10]. These categories, together with example datasets for each one, are shown in Table 2.

These categories seem to be a good way to classify Smart City datasets, however, some of these categories, such as Global Development and Science and Research, might not be used in the Smart City context. Thus, specific domains which can generate data within a Smart City must be taken into account. In this sense,

Table 3: G8 Open Data Categories

Id	Domain	Subdomain
A	Natural resources and energy	1.-Smart grids 2.-Public lighting 3.-Green/renewable energies 4.-Waste management 5.-Water management 6.-Food and agriculture
B	Transport and mobility	7.-City logistics 8.- Info-mobility 9.- People mobility
C	Buildings	10.-Facility management 11.-Building services 12.-Housing quality
D	Living	13.-Entertainment 14.-Hospitality 15.-Pollution control 16.-Public safety 17.-Healthcare 18.-Welfare and social inclusion 19.-Culture 20.-Public spaces management
E	Government	21.-E-government 22.-E-democracy 23.-Procurement 24.-Transparency
F	Economy and people	25.-Innovation and entrepreneurship 26.-Cultural heritage management 27.-Digital Education 28.-Human capital management

Table 4: 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, Community, 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	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

a survey [21] about Smart City initiatives proposes a classification divided in domains and subdomains show in Table 3

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 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 Table4 together with example datasets for each category.

Once the categories were established we had to classify the collected datasets according to such categories. Due to its characteristics, this step requires the participation of experts to execute it adequately. The research groups that have developed this approach includes 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
(5) As a result of this process, 8949 datasets were adequately categorized and 11 were discarded due to their unclear fit.

2.4 Collecting data from GitHub to calculate indicators

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, as mentioned previously, is a platform for collaborative development of software based on a Git repository. It is used by individuals, communities and businesses alike to develop software projects. GitHub is free to use for public and open source projects, and it is profusely used in studies on Software Engineering. Therefore, it offers useful data about open source software projects, including information on whether they are using open data.

GitHub has been used for collecting data and calculating indicators related to OSS success in several works such as [3] [19], where GitHub allows researchers to collect several measures regarding open source projects, for example, forks, stars, etc. GitHub has an API that is used to collect all required data from an open source software 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 as follows: repository_id, user_id, stargazers_count, watchers_count, language, forks_count, subscribers_count, network_count, created_at, updated_at, pushed_at, total_contributors, total_contributions. GitHub user data also provide interesting data to be considered, such as followers_user, following_user, public_repos_user, location_user, updated_at_user, created_at_user. The indicators used in our approach are based on these data. We established a process for identifying 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 [5] 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 needed data.

After this process, we made an estimation of the indicators in order to be used with AHP. 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 formulas previously presented in Table 1.
(4) Normalizing the 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.

2.5 Use of AHP to weight indicators

The method of decision-making, which our model is based on, is named Analytic Hierarchy Process, hereinafter referred to as AHP [25]. It is a powerful and flexible tool for decision-making in complex multi-criteria problem situations and is useful for comparing several alternatives when several objectives need to be borne in mind at the same time.

Following this method, the evaluator can directly assign a normalized weight to a criterion that will indicate the importance which that criterion has with regard to the final objective. Firstly, the AHP method compares the relative importance that each criterion has in relation to all the others; this assessment enables the relative weights of the criteria to be calculated, and finally the method normalizes the weights in order to obtain the measures for the existing alternatives; for this reason, AHP constitutes one of the best options to assist multi-criteria decision making. This method allows people to gather knowledge about a particular problem, to quantify subjective opinions and to force the comparison of alternatives in relation to established criteria. The method consists in the following steps:

(1) Define the problem and the main objective in making the decision.
(2) If required, build a hierarchy tree in this way: the root node is the objective of the problem, the intermediate levels are the criteria, and the lowest level contains the alternatives.
(3) At each level, build a pairwise comparison matrix with the brothers (sons of the same node). The matrix contains the weights of pairwise comparisons between brother nodes. This provides us with a pairwise comparison matrix (see a simple example in Table 5) for each parent node.
(4) For each comparison matrix, an eigenvector must be calculated, using the equation: $|A - \lambda I| = 0$, where A is the comparison matrix, I is the identity matrix and λ is the eigenvector. This calculus must be performed for each level of the tree.
(5) Rate each alternative (leaf nodes) with a previously calculated fixed value for every criteria. The scales for rating alternatives should be established and described in a precise way.
(6) Determine the value of each alternative using a weighted addition formula, with the weights from the previous steps. These results ascend up the tree to calculate the final value

of the objective (root). This final value is used to make a decision about the alternative to choose.

Using this method, as final stage, we have created a Google Spreadsheet based on AHP that uses the reuse indicators as criteria of the process. Concretely, this spreadsheet is composed of three sheets:

(1) 'Indicators'. This sheet provides the normalized indicators that were calculated from GitHub in the previous step.
(2) 'AHP Criterion Pair Comparison'. This sheet allows assessing the relative importance between pairs of indicators using AHP. Thereby, a decision maker could weigh the importance of the indicators set out in the previous steps, taking into account the characteristics and objectives of the city. These weights can be assigned according to the institution's strategic reuse objectives. Thus, different 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 reuse objectives. Cities can be characterized by their size, the importance of the tourism sector, or its residential, commercial or industrial sectors, etc. And also, cities may have different priorities for publishing datasets depending on the type of reuse they want to promote. The result of this step will be the eigenvectors of each matrix, meaning the relative importance of the established indicators.
(3) Finally, the 'AHP Direct Results' shows a suitability ranking list of dataset categories to publish according to the weights introduced in the second sheet and the indicators calculated from GitHub shown in the first sheet. That is, the value used to elaborate such ranking is the result of multiplying the relative importance of each indicator, calculated in the second sheet, by the values of the indicators in the corresponding categories shown in the first sheet.

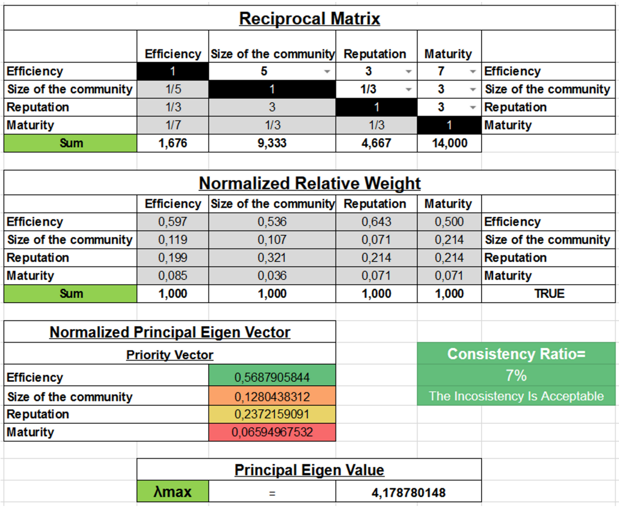
Thus, the use of this tool allows Smart Cities to prioritize datasets in a reasonable way based on the data collected from well-known cities, the indicators taken into account and the open data strategy of the city.

3 SIMULATING THE BEHAVIOUR OF THE TOOL ON STEREOTYPICAL CITIES

In order to check our proposal according to different motivations in the weighting process, we have simulated the behavior of the tool taking into account the different prospects of two stereotypical cities. We asked three experts to agree on the importance assignment of the indicators, with the assumptions of the two cities.

On one hand, a medium-sized town located in a rural region, with small software companies in its zone rather than big ones, that is starting to develop its own open data portal. On the other hand, a big city with a well-known open data portal and a lot of cutting edge software companies in its area of influence.

In the first case, we have guessed that the town could be interested, mainly, in getting reuses of its different datasets through the development of simple applications by small local enterprises. Hence, the town would assign high weights to efficiency whereas reputation, size of the community and maturity would perform a secondary role.



Reciprocal Matrix

	Efficiency	Size of the community	Reputation	Maturity	
Efficiency	1	5	3	7	Efficiency
Size of the community	1/5	1	1/3	3	Size of the community
Reputation	1/3	3	1	3	Reputation
Maturity	1/7	1/3	1/3	1	Maturity
Sum	1,676	9,333	4,667	14,000	

Normalized Relative Weight

	Efficiency	Size of the community	Reputation	Maturity	
Efficiency	0,597	0,536	0,643	0,500	Efficiency
Size of the community	0,119	0,107	0,071	0,214	Size of the community
Reputation	0,199	0,321	0,214	0,214	Reputation
Maturity	0,085	0,036	0,071	0,071	Maturity
Sum	1,000	1,000	1,000	1,000	TRUE

Normalized Principal Eigen Vector

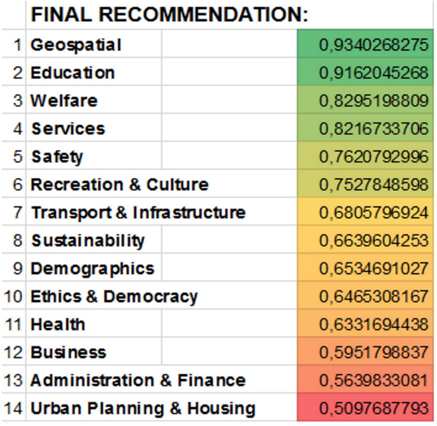
Priority Vector	
Efficiency	0,5687905844
Size of the community	0,1280438312
Reputation	0,2372159091
Maturity	0,06594967532

Consistency Ratio=
7%
The Incosistency Is Acceptable

Principal Eigen Value

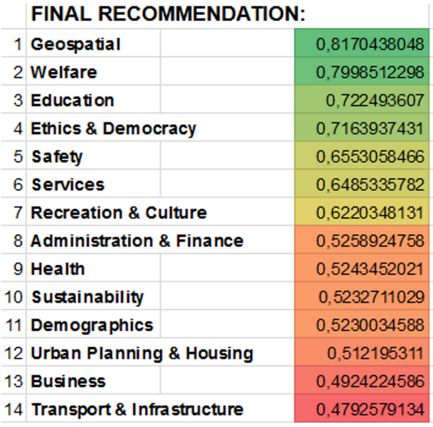
λmax	=	4,178780148

Figure 1: Simulated weights of a medium-sized town.



	FINAL RECOMMENDATION:	
1	Geospatial	0,9340268275
2	Education	0,9162045268
3	Welfare	0,8295198809
4	Services	0,8216733706
5	Safety	0,7620792996
6	Recreation & Culture	0,7527848598
7	Transport & Infrastructure	0,6805796924
8	Sustainability	0,6639604253
9	Demographics	0,6534691027
10	Ethics & Democracy	0,6465308167
11	Health	0,6331694438
12	Business	0,5951798837
13	Administration & Finance	0,5639833081
14	Urban Planning & Housing	0,5097687793

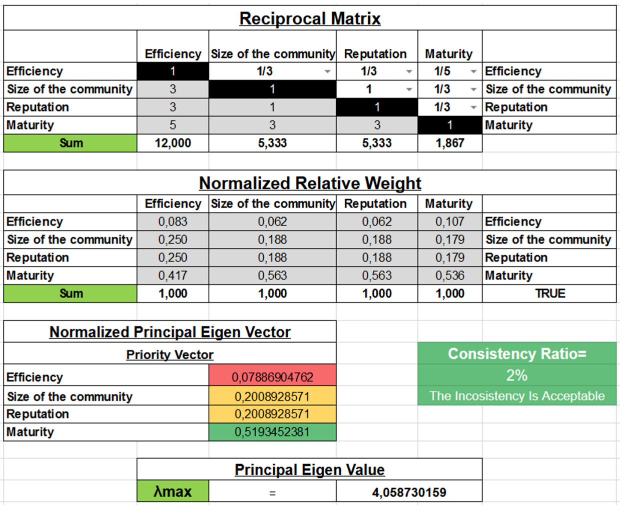
Figure 2: Medium-sized town rankings



	FINAL RECOMMENDATION:	
1	Geospatial	0,8170438048
2	Welfare	0,7998512298
3	Education	0,722493607
4	Ethics & Democracy	0,7163937431
5	Safety	0,6553058466
6	Services	0,6485335782
7	Recreation & Culture	0,6220348131
8	Administration & Finance	0,5258924758
9	Health	0,5243452021
10	Sustainability	0,5232711029
11	Demographics	0,5230034588
12	Urban Planning & Housing	0,512195311
13	Business	0,4924224586
14	Transport & Infrastructure	0,4792579134

Figure 3: Default ranking

The weights applied with this philosophy are shown in Figure 1, and the resulting in the ranking shown in Figure 2. The first position of 'Geospatial' does not change with respect to the default ranking (same weights for all the indicators) shown in Figure 3 but the rest of the ranking suffers some variations.



Reciprocal Matrix

	Efficiency	Size of the community	Reputation	Maturity	
Efficiency	1	1/3	1/3	1/5	Efficiency
Size of the community	3	1	1	1/3	Size of the community
Reputation	3	1	1	1/3	Reputation
Maturity	5	3	3	1	Maturity
Sum	12,000	5,333	5,333	1,867	

Normalized Relative Weight

	Efficiency	Size of the community	Reputation	Maturity	
Efficiency	0,083	0,062	0,062	0,107	Efficiency
Size of the community	0,250	0,188	0,188	0,179	Size of the community
Reputation	0,250	0,188	0,188	0,179	Reputation
Maturity	0,417	0,563	0,563	0,536	Maturity
Sum	1,000	1,000	1,000	1,000	TRUE

Normalized Principal Eigen Vector

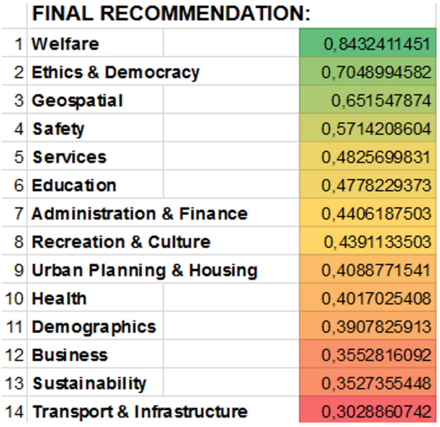
Priority Vector	
Efficiency	0,07886904762
Size of the community	0,2008928571
Reputation	0,2008928571
Maturity	0,5193452381

Consistency Ratio=
2%
The Incosistency Is Acceptable

Principal Eigen Value

λmax	=	4,058730159

Figure 4: Simulated weights of a big-sized city



FINAL RECOMMENDATION:

1	Welfare	0,8432411451
2	Ethics & Democracy	0,7048994582
3	Geospatial	0,651547874
4	Safety	0,5714208604
5	Services	0,4825699831
6	Education	0,4778229373
7	Administration & Finance	0,4406187503
8	Recreation & Culture	0,4391133503
9	Urban Planning & Housing	0,4088771541
10	Health	0,4017025408
11	Demographics	0,3907825913
12	Business	0,3552816092
13	Sustainability	0,3527355448
14	Transport & Infrastructure	0,3028860742

Figure 5: Big city ranking

In the second case, we have conjectured that, due to its portal is well-known, it does not search for more reuses, that is, efficiency, but for mature projects with good reputation and bigger communities behind them. The weights applied with this philosophy are shown in Figure 4.

The ranking obtained with these weights is shown in Figure 5 Here, 'Geospatial' changes to third position and 'Welfare' takes the first one. As can be seen, the indicators obtained from GitHub produces that some categories of the ranking tend to have a stable position regardless of the weights assigned with AHP but, even so, different combinations of weights may change this ranking.

4 RELATED WORK

This section gives a description of (i) some relevant studies about the use of GitHub to measure different indicators about Open Source Software projects, (ii) applications of AHP in Smart Cities as well as (iii) the most relevant studies about how (local) governments publish open data.

Firstly, GitHub is used by individuals, communities and businesses alike to develop software projects. GitHub is free to use for public and OSS projects, and it is profusely used in studies on Software Engineering related to OSS success in several works. Thus, Bissyande et al. uses GitHub [3] to study a possible relation between programming languages and projects success. Marlow et al. [19] analyze metadata projects of GitHub to find how its users decide whom and what to keep track of, or where to contribute next. Sheoran et al. [25] investigate what kind of contributors can be the "watchers" of GitHub. Jarczyk et al. [15] study the relation between popularity of a project in GitHub and its quality. Muthukumaran et al. [20] uses GitHub to propose change metrics that can predict possible bugs. As far as we know, this is the first time GitHub has been used to estimate indicators related to reuse of open data in OSS projects.

Secondly, AHP is a multiple criteria decision making method that has been used in many different applications related to decision making [31]. Some works specifically use AHP in Smart Cities and e-government. In this context, Bartolozzi et al. [2] present a DSS which uses AHP for supporting the decision-making process related to Smart City issues. Sultan et al. [29] suggest the use of AHP to decide the most appropriate technology for the development of e-government projects in Smart Cities. Boselli et al. [4] 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. [18]. The authors propose considering 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, our approach proposes AHP to recommend the most appropriate datasets to be published.

Finally, with respect to how (local) governments publish open data, Conradie & Choenni [6] explain 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. [13] show that benefits associated with opening data are ill-understood. In their systematic review of open government data initiatives, Attard et al. [1] 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 have been no initiatives related to strategies 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 [34], 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. [36] 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. Our approach is related to principle (iii) since we provide a decision support framework based on activities of data consumers. We also contribute to principle (v) since our approach is useful for monitoring how datasets are being reused in OSS applications. Additionally, Jetzek et al. [16] 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.

Furthermore, Zuiderwijk & Janssen [35] state 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.

Therefore, there are several methods that support opening data, but to the best of our knowledge no approaches focus on supporting Smart Cities in selecting and prioritizing which datasets should be open according to their preferences and the context of the city they work for. To fill this gap, we presented our approach based on obtaining useful indicators from Socrata and GitHub and use them with AHP.

5 CONCLUSIONS

Smart Cities usually have a limited budget and insufficient time to release and maintain all available open data. In this paper, we have presented an approach whose goal is to provide an AHP tool that allows weighting different indicators of reuse, calculated using Socrata and GitHub as sources of information, in order to combine them taking into account objective criteria. This approach 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 4 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 the cities.

This approach is completely functional and reproducible. We provide a public repository containing the data obtained from Socrata and GitHub, the scripts to collect and analyze the information and the AHP tool in order to users can use or modify these processes. So, Smart Cities or any other public institution 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. In this sense, further alternative applications of our approach 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 in order to get more data.
(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 impact of open datasets in mass media, social media, blogs, etc. by searching the references to the datasets in these sites.
(5) A set of controlled experiments to demonstrate the effectiveness of our approach in different scenarios.

In summary, a successful publication of open datasets should be based on the proper combination of the objectives of the open data portal and the analysis of the impact of already available open datasets. This approach provides a useful method for Smart City decision makers to carry out this task in an objective and analytic way.

6 ACKNOWLEDGEMENTS

We would like to thank GitHub that allowed us to use its API without limitations and Socrata that provides a way to collect precisely all the datasets published using its tools. This work has been developed with the support of (i) TIN2015-69957-R and TIN2016-78103-C2-2-R (MINECO/ERDF, EU) project, (ii) POCTEP 4IE project (0045-4IE-4-P), (iii) Consejería de Economía e Infraestructuras/Junta de Extremadura (Spain) - European Regional Development Fund (ERDF)- GR15098 project and IB16055 project, and (iv) Consejería de Educación y Empleo/Junta de Extremadura (Spain) - Becas de Movilidad al Personal Docente e Investigador Curso 2016/2017.

REFERENCES

[1] Judie Attard, Fabrizio Orlandi, Simon Scerri, and Sören Auer. 2015. A systematic review of open government data initiatives. *Government Information Quarterly* 32, 4 (2015), 399–418. https://doi.org/10.1016/j.giq.2015.07.006

[2] Marco Bartolozzi, Pierfrancesco Bellini, Paolo Nesi, Gianni Pantaleo, and Luca Santi. 2015. A Smart Decision Support System for Smart City. In *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*. IEEE, 117–122. https://doi.org/10.1109/SmartCity.2015.57

[3] Tegawende F. Bissyande, Ferdian Thung, David Lo, Lingxiao Jiang, and Laurent Reveillere. 2013. Popularity, interoperability, and impact of programming languages in 100,000 open source projects. In *Proceedings - International Computer Software and Applications Conference*. IEEE, 303–312. https://doi.org/10.1109/COMPSAC.2013.55

[4] Roberto Boselli, Mirko Cesarini, Fabio Mercorio, and Mario Mezzanzanica. 2015. 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*. SCITEPRESS, 354–361. https://doi.org/10.5220/0005580003540361

[5] Hitachi Vantara Community. 2018. Data Integration - Kettle. (2018). http://community.pentaho.com/projects/data-integration/

[6] Peter Conradie and Sunil Choenni. 2014. On the barriers for local government releasing open data. *Government Information Quarterly* 31, SUPPL.1 (2014), S10–S17. https://doi.org/10.1016/j.giq.2014.01.003

[7] Carlo Daffara. 2012. Estimating the Economic Contribution of Open Source Software to the European Economy. In *The First Openforum Academy Conference Proceedings*. OpenForum Europe LTD, 11–14.

[8] Rishab Aiyer Ghosh. 2006. *Economic impact of open source software on innovation and the competitiveness of the Information and Communication Technologies (ICT) sector in the EU.* Technical Report. Maastricht: UNU-MERIT. http://stuermer.ch/blog/documents/FLOSSImpactOnEU.pdf

[9] Github. 2018. Github: The world's leading software development platform. (2018). https://www.github.com/

[10] Group of Eight. 2013. G8 Open Data Charter. (2013). https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/207772/Open_Data_Charter.pdf

[11] Jeffrey Hammond, Paul Santinelli, Jay Jay Billings, and Bill Ledingham. 2016. *The Tenth Annual Future of Open Source Survey.* Technical Report. Black Duck Software and North Bridge. https://www.blackducksoftware.com/2016-future-of-open-source

[12] Anders Hjalmarsson, Niklas Johansson, and Daniel Rudmark. 2015. Mind the gap: Exploring stakeholders' value with open data assessment. In *Proceedings of the Annual Hawaii International Conference on System Sciences*. IEEE, 1314–1323. https://doi.org/10.1109/HICSS.2015.160

[13] Mohammad Alamgir Hossain, Yogesh K Dwivedi, and Nripendra P. Rana. 2016. State of the Art in Open Data Research: Insights from Existing Literature and a Research Agenda. *Journal of Organizational Computing and Electronic Commerce* 26, 1-2 (apr 2016), 14–40. https://doi.org/10.1080/10919392.2015.1124007

[14] Marijn Janssen, Yannis Charalabidis, and Anneke Zuiderwijk. 2012. Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information Systems Management* 29, 4 (sep 2012), 258–268. https://doi.org/10.1080/10580530.2012.716740 arXiv:arXiv:1011.1669v3

[15] Oskar Jarczyk, Blazej Gruszka, Szymon Jaroszewicz, and Leszek Bukowski. 2014. GitHub Projects. Quality Analysis of Open-Source Software. In *SocInfo 2014: The 6th International Conference on Social Informatics*. Springer, Cham, 80–94. https://doi.org/10.1007/978-3-319-13734-6_6

[16] Thorhildur Jetzek, Michel Avital, and Niels Bjorn-Andersen. 2014. Data-driven innovation through open government data. *Journal of Theoretical and Applied Electronic Commerce Research* 9, 2 (aug 2014), 100–120. https://doi.org/10.4067/S0718-18762014000200008

[17] Maxat Kassen. 2013. A promising phenomenon of open data: A case study of the Chicago open data project. *Government Information Quarterly* 30, 4 (2013), 508–513. https://doi.org/10.1016/j.giq.2013.05.012

[18] Sylvain Kubler, Jérémy Robert, Yves Le Traon, Jürgen Umbrich, and Sebastian Neumaier. 2016. Open Data Portal Quality Comparison using AHP. In *Proceedings of the 17th International Digital Government Research Conference on Digital Government Research - dg.o '16*. ACM Press, New York, New York, USA, 397–407. https://doi.org/10.1145/2912160.2912167

[19] Jennifer Marlow, Laura Dabbish, and Jim Herbsleb. 2013. Impression Formation in Online Peer Production : Activity Traces and Personal Profiles in GitHub. In *16th ACM Conference on Computer Supported Cooperative Work*. ACM Press, New York, New York, USA, 117–128. https://doi.org/10.1145/2441776.2441792

[20] K. Muthukumaran, Abhinav Choudhary, and N.L. Bhanu Murthy. 2015. Mining GitHub for Novel Change Metrics to Predict Buggy Files in Software Systems. In *2015 International Conference on Computational Intelligence and Networks*. IEEE, 15–20. https://doi.org/10.1109/CINE.2015.13

[21] Paolo Neirotti, Alberto De Marco, Anna Corinna Cagliano, Giulio Mangano, and Francesco Scorrano. 2014. Current trends in smart city initiatives: Some stylised facts. *Cities* 38 (2014), 25–36. https://doi.org/10.1016/j.cities.2013.12.010

[22] National League of Cities. 2018. National League of Cities. (2018). https://www.nlc.org/

[23] Monica Palmirani, Michele Martoni, and Dino Girardi. 2014. Beyond Transparency Introduction : OGA Beyond Transparency. *Electronic Government and the Information Systems Perspective (EGOVIS 2014)* 8650, 2014 (2014), 275–291. https://doi.org/10.1007/978-3-319-10178-1_22

[24] T.L. Saaty. 1980. *The Analytic Hierarchy Process.* McGraw-Hill, New York.

[25] Jyoti Sheoran, Kelly Blincoe, Eirini Kalliamvakou, Daniela Damian, and Jordan Ell. 2014. Understanding "watchers" on GitHub. In *MSR 2014: Proceedings of the 11th Working Conference on Mining Software Repositories*. ACM Press, New York, New York, USA, 336–339. https://doi.org/10.1145/2597073.2597114

[26] Socrata. 2018. Socrata: Data-driven innovation of government programs. (2018). https://www.socrata.com/

[27] Katherine J. Stewart, Anthony P. Ammeter, and Likoebe M. Maruping. 2006. Impacts of license choice and organizational sponsorship on user interest and development activity in open source software projects. *Information Systems Research* 17, 2 (jun 2006), 126–144. https://doi.org/10.1287/isre.1060.0082

[28] Chandrasekar Subramaniam, Ravi Sen, and Matthew L. Nelson. 2009. Determinants of open source software project success: A longitudinal study. *Decision Support Systems* 46, 2 (jan 2009), 576–585. https://doi.org/10.1016/j.dss.2008.10.005 arXiv:arXiv:cond-mat/0402594v3

[29] Abobakr Sultan, Khalid A. AlArfaj, and Ghassan A. AlKutbi. 2012. Analytic hierarchy process for the success of e-government. *Business Strategy Series* 13, 6 (nov 2012), 295–306. https://doi.org/10.1108/17515631211286146

[30] Jeffrey Thorsby, Genie N.L. Stowers, Kristen Wolslegel, and Ellie Tumbuan. 2016. Understanding the content and features of open data portals in American cities. *Government Information Quarterly* 34, 1 (2016), 53–61. https://doi.org/10.1016/j.giq.2016.07.001

[31] Omkarprasad S. Vaidya and Sushil Kumar. 2006. Analytic hierarchy process: An overview of applications. *European Journal of Operational Research* 169, 1 (2006), 1–29. https://doi.org/10.1016/j.ejor.2004.04.028

[32] Nils Walravens, Jonas Breuer, and Pieter Ballon. 2014. Open Data as a Catalyst For The Smart City as a Local Innovation Platform. *Communications & Strategies* 96, 4th quarter 2014 (2014), 15–33. https://ssrn.com/abstract=2636315

[33] Liguo Yu, Alok Mishra, and Deepti Mishra. 2014. 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., 457–466. https://doi.org/10.1007/978-3-662-45550-0_46

[34] Anneke Zuiderwijk and Marijn Janssen. 2013. 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., 38–49. https://doi.org/10.1007/978-3-642-40358-3_4

[35] Anneke Zuiderwijk and Marijn Janssen. 2014. Barriers and Development Directions for the Publication and Usage of Open Data: A Socio-Technical View. In *Open Government*. Vol. 4. Springer New York, New York, NY, 115–135. https://doi.org/10.1007/978-1-4614-9563-5_8 arXiv:arXiv:1011.1669v3

[36] Anneke Zuiderwijk, Marijn Janssen, Sunil Choenni, and Ronald Meijer. 2014. Design principles for improving the process of publishing open data. *Transforming Government: People, Process and Policy* 8, 2 (may 2014), 185–204. https://doi.org/10.1108/TG-07-2013-0024

[37] Anneke Zuiderwijk, Iryna Susha, Yannis Charalabidis, Peter Parycek, and Marijn Janssen. 2015. 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*. Edition Donau-Universität Krems, 197–208.