



JUNTA DE EXTREMADURA

Consejería de Economía e Infraestructuras



Unión Europea

DOLAP 2018

Supporting open dataset publication decisions based on Open Source Software Reuse

Álvaro E. Prieto¹, José-Norberto Mazón², Adolfo-Lozano-Tello¹, Luis-Daniel Ibáñez³

¹U. de Extremadura (Spain), ²U. de Alicante (Spain), ³U. of Southampton (UK)



Software Engineering Group

QUERCUS



<http://quercusseg.unex.es/>

aeprieto@unex.es

Fondo Europeo de Desarrollo Regional

Una manera de hacer Europa.

Index

The problem

What datasets should be opened by Smart Cities?

Our approach

Using reuse of Open Datasets in Open Source Software projects to prioritize them



Software Engineering Group

QUERCUS



<http://quercusseg.unex.es/>

aeprieto@unex.es

◀ ▶

The problem

What datasets should be opened by Smart Cities?



Software Engineering Group
QUERCUS



<http://quercusseg.unex.es/>
aeprieto@unex.es

What datasets should be opened by Smart Cities?

Open Data in Smart Cities

- enable the creation of new services
 - by reusing and combining open datasets
 - in different novel ways that Smart Cities might not have foreseen
 - by third parties as journalists, software developers, data scientists, etc.



Open Data is considered an important source of raw material for innovation and both economic and social impact



Software Engineering Group
QUERCUS



<http://quercusseg.unex.es/>
aeprieto@unex.es



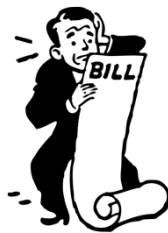
What datasets should be opened by Smart Cities?

Open Data in Smart Cities

The main challenge is that open data has no value in itself; **it only becomes valuable when used**

Maximizing their chances of being reused requires stability and maintenance over time

Extra cost that most cities **cannot afford** to publish all their datasets



Software Engineering Group
QUERCUS



<http://quercusseg.unex.es/>
aeprieto@unex.es



What datasets should be opened by Smart Cities?

Open Data in Smart Cities

- Currently Smart Cities usually release
 - mandatory data (transparency laws)
 - data that is easier (or cheaper) to release
 - privacy issues
 - technical formats

If a smart city wants to generate economic impact, it must **prioritize** opening the **most demanded** datasets for reusers, not the **easiest** to open



Software Engineering Group
QUERCUS



<http://quercusseg.unex.es/>
aeprieto@unex.es



What datasets should be opened by Smart Cities?

So far



Do Smart Cities have some way of measuring the reuse of their open datasets?

Do Smart Cities have some tool or method that use these data to support open dataset publication and maintenance decisions?



Software Engineering Group
QUERCUS



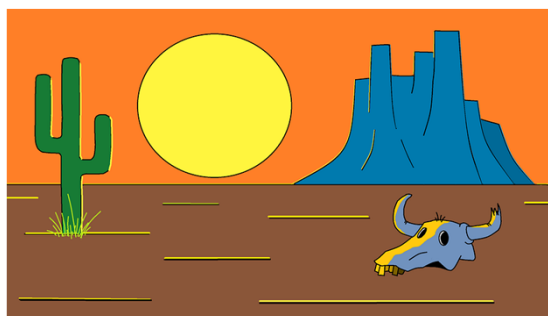
<http://quercusseg.unex.es/>
aeprieto@unex.es



What datasets should be opened by Smart Cities?

So far

To the best of our knowledge:



Software Engineering Group
QUERCUS



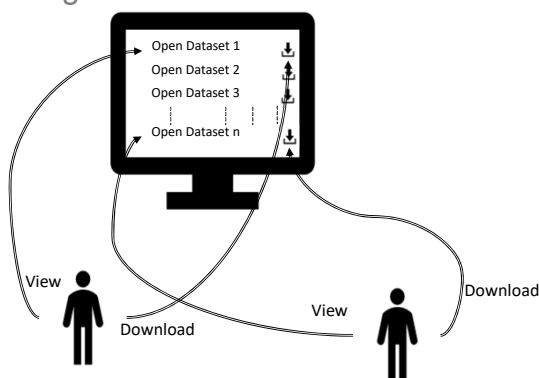
<http://quercusseg.unex.es/>
aeprieto@unex.es



What datasets should be opened by Smart Cities?

So far

And something similar?



Software Engineering Group
QUERCUS



<http://quercusseg.unex.es/>
aeprieto@unex.es

What datasets should be opened by Smart Cities?

So far

And something similar?

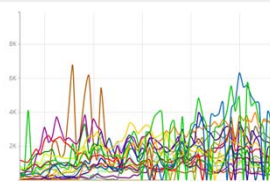
Data.gov.uk Usage All datasets

Download as CSV

Graph Legend

Click on a series below to isolate the graph.

- Road Safety Data
- MOT vehicle testing stations in Great Britain
- Planned road works on the H1 road network
- Live traffic information from Highways England (previously Highways Agency)
- HM Land Registry Price Paid Data
- English Indices of Deprivation 2010
- Building Price and Cost Indices
- Lower Layer Super Output Area (LSOA) boundaries
- HM Land Registry Property
- Statistics on Crime, Criminal Activity and
- Crime Victims Unpublished Estimates and
- LSOA Community DSM - 1m
- National average GDP by metropolitan
- All Road Traffic Counts
- National Public Transport Access Nodes
- All Driving License Data
- Airborne Monitoring Agency and
- C500 vehicle test stations
- Learning and Performance Service
- Health Survey for England



STATISTICS FOR ALL MONTHS

All months

All publishers

Export

Dataset	Views	Downloads
Road Safety Data	179415	100829
MOT vehicle testing stations in Great Britain	122764	24331
Planned road works on the H1 road network	121554	27361
Live traffic information from Highways England (previously Highways Agency)	113753	20736
HM Land Registry Price Paid Data	103740	58933
English Indices of Deprivation 2010	102891	35149
Building Price and Cost Indices	98791	28356
Lower Layer Super Output Area (LSOA) boundaries	95639	20515



Software Engineering Group
QUERCUS



<http://quercusseg.unex.es/>
aeprieto@unex.es

What datasets should be opened by Smart Cities?

So far

But

- What did the users that viewed the dataset do?
 - Did they speak with friends about it?
 - Did they read it to know about the busy roads? A raw CSV???
- What did the users that downloaded the dataset do?
 - Did they create a visualization?
 - Did they develop a mobile app?
 - Or are they suffering from some kind of Digital Diogenes Syndrome?



Software Engineering Group
QUERCUS



<http://quercusseg.unex.es/>
aeprieto@unex.es



Our approach

Using reuse of Open Datasets in Open Source
Software projects to prioritize them



Software Engineering Group
QUERCUS



<http://quercusseg.unex.es/>
aeprieto@unex.es

Using reuse of OD in OSS projects to prioritize them

Why Open Source Software?

Encourages the creation of SMEs and jobs



Providing a skills development environment valued by employers and retaining a greater share of generated value locally



Software Engineering Group
QUERCUS



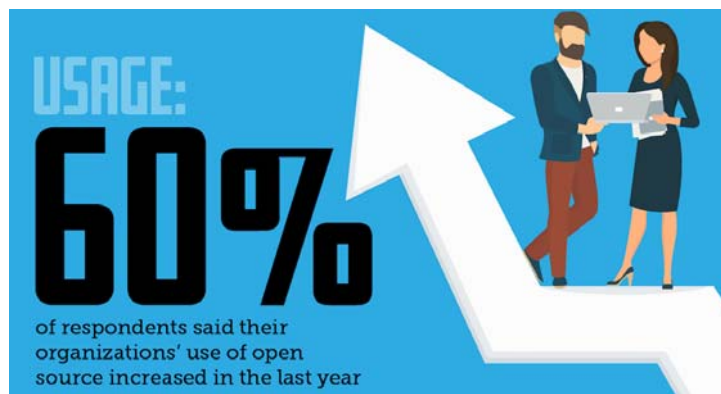
<http://quercusseg.unex.es/>
aeprieto@unex.es



Using reuse of OD in OSS projects to prioritize them

Why Open Source Software?

In 2017:



Source: 2017 Open Source 360° Survey by Black Duck's Center for Open Source Research and Innovation (COSRI)



Software Engineering Group
QUERCUS



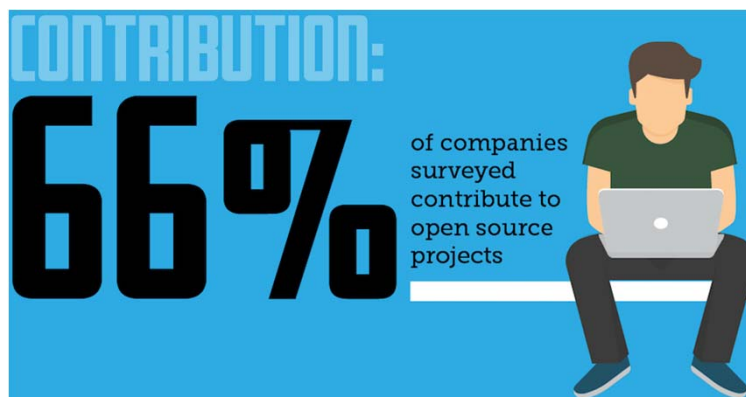
<http://quercusseg.unex.es/>
aeprieto@unex.es



Using reuse of OD in OSS projects to prioritize them

Why Open Source Software?

In 2017:



Source: 2017 Open Source 360° Survey by Black Duck's Center for Open Source Research and Innovation (COSRI)



Software Engineering Group
QUERCUS



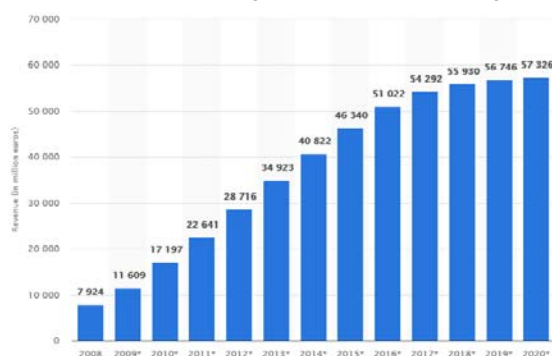
<http://quercusseg.unex.es/>
aeprieto@unex.es



Using reuse of OD in OSS projects to prioritize them

Why Open Source Software?

- Projected revenue of open source software from 2008 to 2020 (in million euros)



Data visualized by +

© Statista 2018



Software Engineering Group
QUERCUS



<http://quercusseg.unex.es/>
aeprieto@unex.es



Using reuse of OD in OSS projects to prioritize them

Why Open Source Software?

- Why don't use an estimation of the reuse in OSS of the different categories of datasets as an indicator of their potential impact?
- Why don't use this information in Smart Cities to make decisions on which data to publish?

So, 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.



Software Engineering Group
QUERCUS



<http://quercusseg.unex.es/>
aeprieto@unex.es



Using reuse of OD in OSS projects to prioritize them

Steps of the proposal

- 1st A proposal of indicators of reuse
- 2nd Taxonomy of dataset categories for Smart Cities
- 3rd Gathering datasets
- 4th Classifying collected datasets
- 5th Collecting data from GitHub to calculate indicators
- 6th Estimation of the indicators
- 7th Use of AHP to weight the indicators
- 8th Simulating the Behaviour



Software Engineering Group
QUERCUS



<http://quercusseg.unex.es/>
aeprieto@unex.es



Using reuse of OD in OSS projects to prioritize them

1st A proposal of indicators of reuse

- We borrowed some well-known indicators that measure the success of OSS projects:
 - 1. **Reputation**
 - number of people who agree to receive information about the project because they find it interesting
 - reveal a deeper interest in the OSS project

Smart Cities could be interested in opening datasets of categories that have been reused in high reputation projects in view of creating a community around their open data



Software Engineering Group
QUERCUS



<http://quercusseg.unex.es/>
aeprieto@unex.es



Using reuse of OD in OSS projects to prioritize them

1st - A proposal of indicators of reuse

- We borrowed some well-known indicators that measure the success of OSS projects:
 - 2. **Size of the community**
 - number of people who actually work on the OSS project
 - is critical to its success, since survival of an OSS project depends on their continued contribution

Smart Cities could be interested in opening datasets of categories that have been reused in projects of different number of developers according to the size of the companies in their area of influence



Software Engineering Group
QUERCUS



<http://quercusseg.unex.es/>
aeprieto@unex.es



Using reuse of OD in OSS projects to prioritize them

1st A proposal of indicators of reuse

- We borrowed some well-known indicators that measure the success of OSS projects:

– 3. **Maturity**

- age of an active project
 - is positively related to OSS progress toward completion, as well as the experience of the community of developers

A Smart City may want to select the dataset categories that help in promoting fewer projects stretching over longer periods of time, rather than promoting a larger number of short-term projects



Software Engineering Group
QUERCUS



<http://quercusseg.unex.es/>
aeprieto@unex.es



Using reuse of OD in OSS projects to prioritize them

1st A proposal of indicators of reuse

- An additional indicator has been developed in order to assess the impact of a dataset category:

– 4. **Efficiency**

- the likelihood of datasets from each category of being reused
 - based on the proportion of datasets of each category currently reused

Smart Cities will use this indicator to know which categories of open data are most likely to be reused



Software Engineering Group
QUERCUS



<http://quercusseg.unex.es/>
aeprieto@unex.es

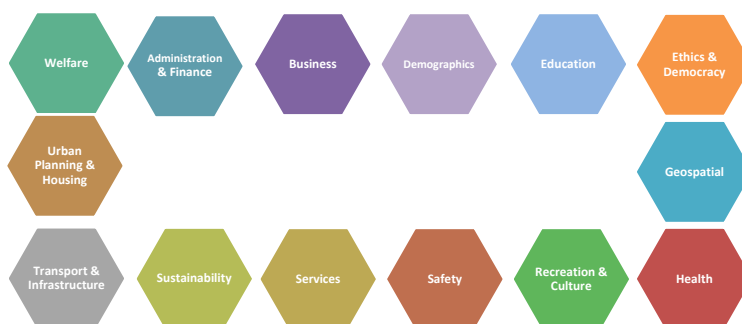


Using reuse of OD in OSS projects to prioritize them

2nd Taxonomy of dataset categories for Smart Cities

Our proposal for Smart Cities:

- As close as possible to the G8 Open Data Charter
- Incorporates modifications to encompass domains and subdomains proper to Smart Cities



Software Engineering Group
QUERCUS



<http://quercusseg.unex.es/>
aeprieto@unex.es



Using reuse of OD in OSS projects to prioritize them

3rd Gathering datasets

32 US cities



Software Engineering Group
QUERCUS



<http://quercusseg.unex.es/>
aeprieto@unex.es



Using reuse of OD in OSS projects to prioritize them

3rd Gathering datasets

8960 open datasets

Open Dataset 1
 Open Dataset 2
 Open Dataset 3
 Open Dataset 4
 Open Dataset 5
 Open Dataset 6
 Open Dataset 7
 Open Dataset 8
 Open Dataset 9
 Open Dataset 10
 Open Dataset 11
 Open Dataset 12
 Open Dataset 13
 Open Dataset 14
 Open Dataset 15
 Open Dataset 16
 Open Dataset 17
 Open Dataset 18
 Open Dataset n

Software Engineering Group
QUERCUS

<http://quercusseg.unex.es/>
aeprieto@unex.es

Using reuse of OD in OSS projects to prioritize them

4th Classifying collected datasets

215 different themes

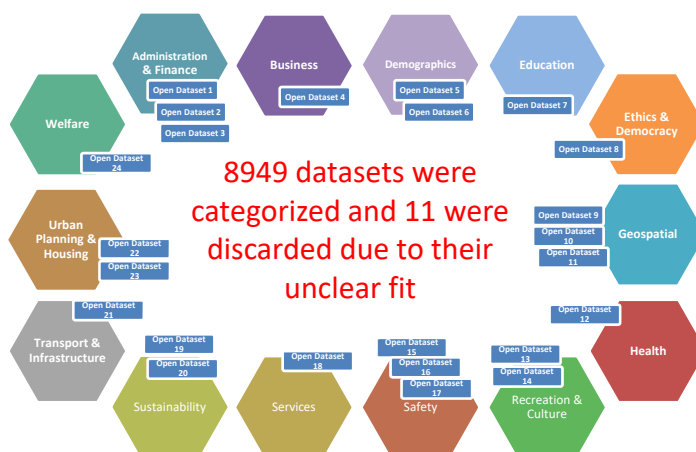
Open Dataset 1
 Open Dataset 2
 Open Dataset 3
 Open Dataset 4
 Open Dataset 5
 Open Dataset 6
 Open Dataset 7
 Open Dataset 8
 Open Dataset 9
 Open Dataset 10
 Open Dataset 11
 Open Dataset 12
 Open Dataset 13
 Open Dataset 14
 Open Dataset 15
 Open Dataset 16
 Open Dataset 17
 Open Dataset 18
 Open Dataset n

Software Engineering Group
QUERCUS

<http://quercusseg.unex.es/>
aeprieto@unex.es

Using reuse of OD in OSS projects to prioritize them

4th Classifying collected datasets



Software Engineering Group
QUERCUS



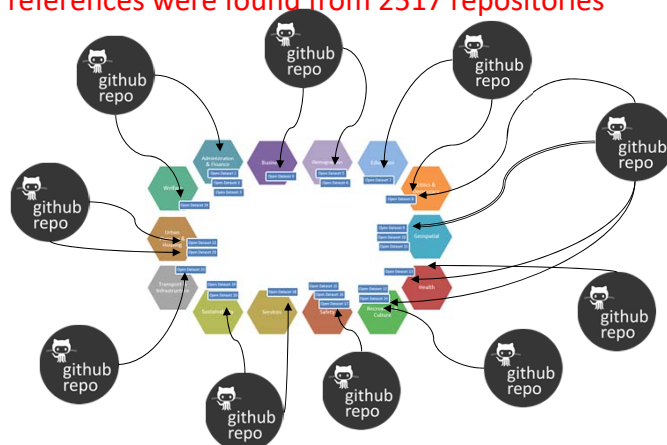
<http://quercusseg.unex.es/>
aeprieto@unex.es



Using reuse of OD in OSS projects to prioritize them

5th Collecting data from GitHub to calculate indicators

350644 references were found from 2517 repositories



to 5874 of the 8949 categorized datasets



Software Engineering Group
QUERCUS



<http://quercusseg.unex.es/>
aeprieto@unex.es



Using reuse of OD in OSS projects to prioritize them

6th Estimation of the indicators

1. Discarding repositories that do not have all the required data
 - Only 2501 repositories remained.
2. Discarding all repeated references to a specific dataset from a specific repository.
 - 32551 unrepeated references remained.
3. Making an estimation of the indicators and normalizing them to a 0-1 range

	Efficiency	Size of the community	Reputation	Maturity
Administration & Finance	0,541	0,634	0,665	0,264
Business	0,560	0,390	0,933	0,087
Demographics	0,760	0,556	0,577	0,199
Education	1,000	0,829	1,000	0,061
Ethics & Democracy	0,489	0,860	0,923	0,593
Geospatial	0,997	1,000	0,904	0,366
Health	0,718	0,603	0,563	0,213
Recreation & Culture	0,817	0,801	0,748	0,122
Safety	0,877	0,631	0,637	0,477
Services	0,947	0,608	0,797	0,242
Sustainability	0,726	0,568	0,734	0,065
Transport & Infrastructure	0,890	0,506	0,438	0,083
Urban Planning & Housing	0,363	0,680	0,875	0,131
Welfare	0,967	0,721	0,512	1,000



Software Engineering Group
QUERCUS



<http://quercusseg.unex.es/>
aeprieto@unex.es



Using reuse of OD in OSS projects to prioritize them

7th Use of AHP to weight the indicators

Fractional Scale	Decimal Scale	Degree of preference
9	9	Extremely more important
8	8	
7	7	Very strongly more important
6	6	
5	5	Strongly more important
4	4	
3	3	Moderately more important
2	2	
1	1	Equal importance
1/2	0,5	
1/3	0,33333	Moderately less important
1/4	0,25	
1/5	0,2	Strongly less important
1/6	0,166667	
1/7	0,142857	Very strongly less important
1/8	0,125	
1/9	0,11111	Extremely less important

Reciprocal Matrix					
	Efficiency	Size of the community	Reputation	Maturity	
Efficiency	1	1	1	1	Efficiency
Size of the community	1	1	1	1	Size of the community
Reputation	1	1	1	1	Reputation
Maturity	1	1	1	1	Maturity
Sum	4,000	4,000	4,000	4,000	

Normalized Relative Weight					
	Efficiency	Size of the community	Reputation	Maturity	
Efficiency	0,250	0,250	0,250	0,250	Efficiency
Size of the community	0,250	0,250	0,250	0,250	Size of the community
Reputation	0,250	0,250	0,250	0,250	Reputation
Maturity	0,250	0,250	0,250	0,250	Maturity
Sum	1,000	1,000	1,000	1,000	TRUE

Normalized Principal Eigen Vector			
Priority Vector			
Efficiency	0,25		
Size of the community	0,25		
Reputation	0,25		
Maturity	0,25		

Consistency Ratio=
0%
The inconsistency Is Acceptable



Software Engineering Group
QUERCUS



<http://quercusseg.unex.es/>
aeprieto@unex.es



Using reuse of OD in OSS projects to prioritize them

8th Simulating the Behaviour

- Medium-sized town
 - From a rural region
 - Small SW companies around
 - Starting its Open Data portal
- We have guessed that the town
 - Wants reuses of its datasets
 - through the development of simple applications by small local enterprises.

Reciprocal Matrix				
	Efficiency	Size of the community	Reputation	Maturity
Efficiency	1	6	3	7
Size of the community	1/6	1	1/3	3
Reputation	1/3	3	1	3
Maturity	1/7	1/3	1/3	1
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
Size of the community	0.119	0.107	0.071	0.214
Reputation	0.189	0.321	0.214	0.214
Maturity	0.085	0.036	0.071	0.071
Sum	1,000	1,000	1,000	1,000

Normalized Principal Eigen Vector				
	Efficiency	Size of the community	Reputation	Maturity
Efficiency	0.5687905844			
Size of the community	0.120438312			
Reputation	0.2372159091			
Maturity	0.0654967532			

Principal Eigen Value				
	Efficiency	Size of the community	Reputation	Maturity
Amax	=	4.178780148		

Consistency Ratio=	7%
The Inconsistency Is Acceptable	



Software Engineering Group
QUERCUS



<http://quercusseg.unex.es/>
aeprieto@unex.es

Using reuse of OD in OSS projects to prioritize them

8th Simulating the Behaviour

- Big city
 - A well-known open data portal
 - Many cutting edge SW around
- We have guessed that the city
 - Wants mature projects with good reputation and bigger communities

Reciprocal Matrix				
	Efficiency	Size of the community	Reputation	Maturity
Efficiency	1	1/3	1/3	1/6
Size of the community	3	1	1	1/3
Reputation	3	3	1	1/3
Maturity	6	3	3	1
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
Size of the community	0.250	0.188	0.188	0.179
Reputation	0.250	0.188	0.188	0.179
Maturity	0.417	0.563	0.563	0.536
Sum	1,000	1,000	1,000	1,000

Normalized Principal Eigen Vector				
	Efficiency	Size of the community	Reputation	Maturity
Efficiency	0.07880904702			
Size of the community	0.2008928571			
Reputation	0.2008928571			
Maturity	0.5193852561			

Principal Eigen Value				
	Efficiency	Size of the community	Reputation	Maturity
Amax	=	4.068730159		

Consistency Ratio=	2%
The Inconsistency Is Acceptable	



Software Engineering Group
QUERCUS



<http://quercusseg.unex.es/>
aeprieto@unex.es

Using reuse of OD in OSS projects to prioritize them

8th Simulating the Behaviour

FINAL RECOMMENDATION:

1	Geospatial	0.8170438048
2	Welfare	0.7996512298
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

a. Default ranking

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

b. Medium-sized town ranking

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

c. Big city ranking



•
•
•

Software Engineering Group
QUERCUS



•
•
•

<http://quercusseg.unex.es/>
aeprieto@unex.es



Summary



•
•
•

Software Engineering Group
QUERCUS



•
•
•

<http://quercusseg.unex.es/>
aeprieto@unex.es

Summary

Our goal

Provide an AHP tool that allows weighting different indicators of reuse, calculated using Socrata and GitHub as sources of information



Software Engineering Group
QUERCUS



<http://quercusseg.unex.es/>
aeprieto@unex.es



Summary

This approach is characterized by

1. A definition of 4 indicators based on the reuse of datasets in OSS
2. A classification of 14 categories for Smart City open datasets based on the G8 Open Data Charter and the Smart City domain.
3. Almost 9000 open datasets collected 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.



Software Engineering Group
QUERCUS



<http://quercusseg.unex.es/>
aeprieto@unex.es



Summary

Future works

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.



Software Engineering Group
QUERCUS



<http://quercusseg.unex.es/>
aeprieto@unex.es



Thank you very much.
Questions and suggestions?



Software Engineering Group
QUERCUS



<http://quercusseg.unex.es/>
aeprieto@unex.es