PlanetData

Network of Excellence

FP7 – 257641

D18.3 Call 2: Linked Map Platform monitoring report

Coordinator: Jesús Barrera (GEOSLAB)
With contributions from: Francisco J Lopez-Pellicer (UNIZAR)

1st Quality reviewer: Loris Bozzato (FBK)
2nd Quality reviewer: Max Schmachtenberg (UMA)

Deliverable nature:	Report (R)
Dissemination level: (Confidentiality)	Restricted to group (RE)
Contractual delivery date:	M48
Actual delivery date:	M48
Version:	1.0
Total number of pages:	19
Keywords:	Linked data, geographic information, crowdsourcing, monitoring

Abstract

The Linked Map Platform enables users an intuitive way to comment on the quality of data belonging to different data sources and the links established between them. This deliverable describes the mechanisms to monitor the Linked Map Platform and collects the statistics on the use of this platform.

Executive summary

The purpose of this deliverable is to describe the mechanisms to monitor the platform and to present collected statistics on the use of the Linked Map Platform available at http://linkedmap.unizar.es/crowdsourcing-platform/. Monitoring helps us to understand how users are using the platform since its launch (July 2014) up to date and provides valuable data on crowdsourcing experiments of WP19 held in the platform.

Monitoring requires not only setting up a system that provides consistent, timely and accurate measurements but also addressing specific information needs. In our case, the monitoring procedure specifies the objectives of the task (measurement of audience, acquisition and behaviour), the measures and mechanisms for data collection, their implementation (server, client) and presents objective results in form of key indicators, tables, graphs, dashboards and maps.

There were 549 sessions to the website initiated by 295 unique users from 300 different hosts. As expected, due to the nature of the data offered in the project most of the users come from Spain (85.4%). These numbers are noteworthy given the limited time span available and the small size of the project. A review of acquisition data reveals that mails sent to mailing lists and personal contacts promoting the project are probably the main source of visits and, hence, participation in crowdsourcing activities.

Data show that the platform engages few users to consider editing resources although users explore its content. Nevertheless, monitoring has exposed different patterns of interaction with the platform and has helped the project to understand results of experiments of WP19.

Document Information

IST Project Number	FP7 - 257641	Acronym	PlanetData
Full Title	PlanetData		
Project URL	http://www.planet-data.eu/		
Document URL	http://wiki.planet-data.eu/web/D18.3		
EU Project Officer	Leonhard Maqua		

Deliverable	Number	D18.3	Title	Call 2: Linked Map Platform monitoring report
Work Package	Number	WP18	Title	Call 2: Linked Map Platform integration & development

Date of Delivery	Contractual	M48	Actual	M48
Status	version 1.0		final ☒	
Nature	prototype □ report ☒ demonstrator □ other □			
Dissemination level	public □ restricted to group ☒ restricted to programme □ consortium □			

Authors (Partner)	Jesús Barrera (GEOSLAB), Francisco J Lopez-Pellicer (UNIZAR)			
Responsible Author	Name	Jesús Barrera	E-mail	jesusb@geoslab.com
	Partner	GEOSLAB	Phone	+34 976 065152

Abstract (for dissemination)	This deliverable describes the mechanisms to monitor the Linked Map Platform and collects the statistics on the use of this platform which enables users an intuitive way to comment on the quality of data belonging to different data sources and the links established between them.
Keywords	Linked data, geographic information, crowdsourcing, monitoring.

Version Log			
Issue Date	Rev. No.	Author	Change
2014/07/08	0.1	Jesús Barrera Francés	Template instantiation
2014/08/26	0.2	Jesús Barrera Francés	First draft
2014/08/03	0.3	Francisco J Lopez-Pellicer	Revision and comments
2014/08/28	0.4	Jesús Barrera Francés	Adjustments
2014/09/01	0.5	Francisco J Lopez-Pellicer	First version ready for QA; Monitoring data up to 2014-08-29
2014/09/18	0.9	Francisco J Lopez-Pellicer	First version ready for AL; Monitoring data up to 2014-09-15
2014/09/24	1.0	Francisco J Lopez-Pellicer	Final version

Table of Contents

List of figures

List of tables

1 Introduction

The Linked Map Platform enables users an intuitive way to comment on the quality of data belonging to different data sources and the links established between them. This web application is available at http://linkedmap.unizar.es/crowdsourcing-platform/. The purpose of this deliverable is to describe the mechanisms to monitor the platform and to present collected statistics on the use of this platform. Monitoring helps us to understand how users are using the platform since its launch (July 2014), and provides valuable data on crowdsourcing experiments of WP19 held in the platform (August-September 2014). These experiments are detailed in the deliverable D19.2 [1].

Monitoring requires not only setting up a system that provides consistent, timely and accurate measurements but also addressing specific information needs. In our case, what should be measured is related to the crowdsourcing experiments (who, where, how). Each type of information need can be covered by different measurements implemented with different technologies. Nowadays, monitoring systems produce a deluge of data values. Thus, collected data require to be presented in an expressive way useful for decision taking.

Adapting ideas described in [2], an effective monitoring procedure includes:

- Specifying the objectives of the task.
- Specifying the measures and mechanisms for data collection.
- Implementing such mechanisms.
- Presenting objective results that can be used for decision taking.

In broad terms, following the structure outlined above, this document is divided as follows:

- Section 2 provides a rationale of our information needs, presents a set of meaningful measurements related to them and describes the implementation of the measurements in server and client side.
- Section 3 presents collected data in form of key indicators, tables, graphs, dashboards and maps.
- Section 4 concludes summarizing and analysing results.

Annex A and 0 complete the document by providing additional details about the measurements.

2 Monitoring procedure

The term *metric* is widely used and accepted for software measurement. However, its use is not consistent in standards and research proposals. For this reason, following [3], [4], we chose the word *measure* (and the companion term *measurement*) instead of *metric* for referring a defined measurement approach and its measurement scale. Table 8 and Table 9 in Annex A define each base and derived acquisition measure.

Measurement of the Linked Map platform is focused on addressing the following information needs:

- *Audience*. Who is the effective audience of the Linked Map platform?
- *Acquisition*. How the Linked Map platform is acquiring users?
- *Behaviour*. How users use the Linked Map platform?

2.1 Measurements

We have identified that the following measures would satisfy the information needs on audience:

- *Sessions*. Number of sessions or groups of interactions performed by users in a period.
- *Duration of sessions*. Accumulated duration of sessions.
- *Unique users*. Number of distinct users that have initiated a session.
- *Hosts*. Number of distinct hosts from which users have initiated a session.
- *Bounce users*. Number of distinct users that have initiated only one session.
- *Returning users*. Number of distinct users that have initiated more than one session.
- *Users per country*. Number of distinct users that have initiated a one session from a device located in a specific country
- *Page views*. Number of pages visited by users.
- *Pages per session* (derived measure). Ratio between *page views* and *sessions*.
- *Average session duration* (derived measure). Ratio between *duration of sessions* and *sessions*.

Table 8 and Table 9 in Annex A define each base and derived acquisition metric.

Regarding on acquisition, measuring the *source* and the *channel* from which new users reach the platform satisfies the information needs. Source is any point of origin for new traffic, e.g. a search engine, a newsletter, or a bookmark. Channel is a measure that groups new traffic according to the type of content of the source. That is, if the traffic arrives from a search result, it is classified as *organic search* traffic. If traffic arrives from social media, it is classified as *social* traffic. If it arrives from a web site, it is classified as *referral* traffic. Otherwise it is classified as *direct* traffic.

Table 10 in Annex A defines source and traffic in detail.

Finally, regarding on behaviour, the following measures have been identified as useful for assessing user behaviour:

- *Bounce sessions*. Number of sessions where there is a single interaction.
- *Content requested*. Number of requests per content class.
- *Server status code*. Number of responses per HTTP status code.
- *Client side events*. Number of actions performed by users in the web client per action class.
- *Bounce rate* (derived measure). Ratio between *bounce sessions* and *sessions*.

Table 11 and Table 12 in Annex A define these measures in detail.

2.2 Implementation

The measurements are implemented in both the server and the client side of the platform as follows:

- *Client side*. Google Analytics[1] is a service offered by Google that generates detailed statistics about a website's traffic, traffic sources and how visitors interact with web applications. Google Analytics can track visitors from all referrers, including search engines and social networks, direct visits and referring sites. Google Analytics' approach is to show high-level, dashboard-type data for the casual user, and more in-depth data further into the report set.
- *Server side*. GoAccess[2] is an open source real-time web log analyser for Unix-like systems that can parse Apache HTTP log files and produce useful statistics such as unique visitors per day, user location, requested files and HTTP status codes.

Google Analytics is implemented in the Linked Map platform with snippets of JavaScript code added to every page of the website. Snippets contain a tracking code provided by Google that runs in the client browser when the client interacts with the Linked Map platform. Tracking code collects data and user events and sends this information to the Google Analytics database for logging and post-processing.

Google Analytics only works if JavaScript is enabled and it is not blocked by ad filtering programs. The Linked Map platform requires running JavaScript code on to work. Therefore, only ad filtering programs may prevent users from being tracked. According to a recent report of PageFair [5], up to 30% of visitors may be using ad blocking software. Therefore, Google Analytics data should be considered a lower bound of the website's traffic.

The Linked Map platform uses an Apache HTTP server and GoAccess processes its logs. Resulting statistics validate Google Analytics reports and provide server side statistics unavailable in Google Analytics.

Table 13 in Annex B identifies for each base measurement its sources (server side, client side).

[1] http://www.google.com/analytics/

[2] http://goaccess.io/

3 Monitoring activities

Monitoring support was implemented in June. Next, we started monitoring user activities in the Linked Map platform on July 10th. We summarize below the results of the monitoring activity up to September 15th. Note that between August 7th-11th the platform was shut down (and intermittently between August 12th-14th) because of a planned hosting maintenance procedure that went wrong.

3.1 Audience

There were 750 sessions to the website initiated by 385 unique users from 385 different hosts in the period July-September 2014. As expected, 83.5% of the users come from Spain. 189 of 385 can be classified as bounce users. A bounce user is a user that has initiated only one session during the analysed period. Users have requested 2,852 pages. The average session duration was 10 minutes 12 seconds. Note that the value pages per session is low compared to the average session duration because the client side of the Linked Map platform uses JavaScript to dynamically interact with the user without requesting a new page. Table 1 and Table 2 show available audience measurements. Figure 1 presents an overview of the evolution of the audience. Figure 2 reveals the location of the users.

Table 1 – Base audience data. Period: July-September 2014. Source: Google Analytics

Measure	Value	Units
Sessions	750	*sessions*
Duration of sessions	459,000	*seconds*
Unique users	385	*users*
Hosts	385	*hosts*
Bounce users	189	*users*
Returning users	196	*users*
Page views	2,852	*pages*

Table 2 – Derived audience measurements. Period: July-September 2014.

Measure	Value	Units
Pages per session	3.80	$\frac{pages}{session}$
Avg. Session duration	612	$\frac{seconds}{session}$

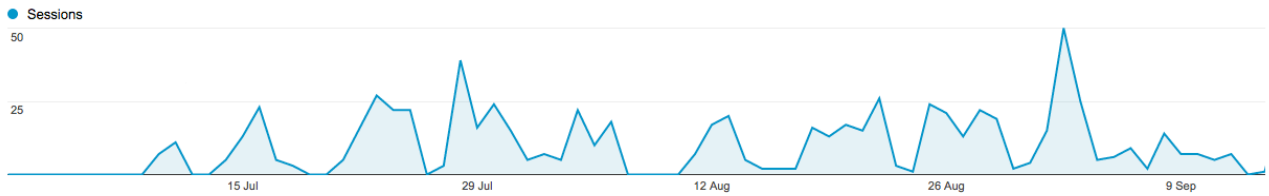

Figure 1 – Audience evolution. Period: July-September 2014. Source: Google Analytics

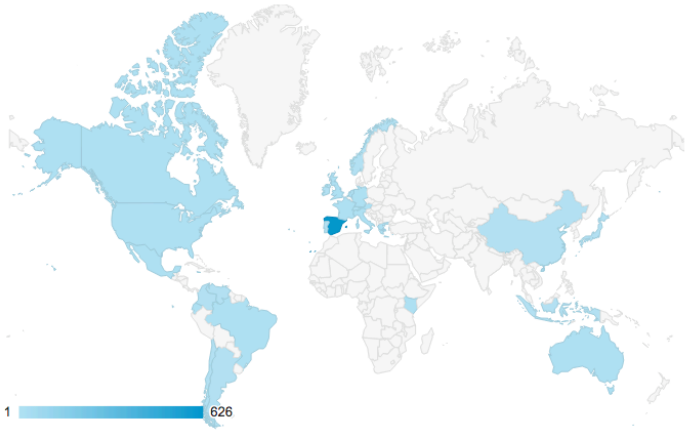

Figure 2 – Number of visitors by countries. Period: July-September 2014. Source: Google Analytics

3.2 Acquisition

Table 3 presents the source from which traffic to Linked Map originated. Most of 750 sessions are direct (83.1%). 6.9% of the visitors come from Google, that is, links to Linked Map on Google's results pages that appear because of their relevance to the search terms. 2.0% of visitors come from the IDEE blog[3] which is the official blog of the Spatial Data Infrastructure of Spain. Remaining visitors come from links in other sites. Source information can be retrieved both from Google Analytics and server log files.

Source information can be considered raw data for acquisitions analysis. Advanced monitoring tools such as Google Analytics are able to group distinct sources into channels. A channel identifies a set of sources of similar content (search results, social media, etc.) from which traffic to Linked Map originated. Table 4 presents channels from which traffic to Linked Map originated. This measure reveals that 6.9% of the visitors come from organic search and 3.2% of the visitors come from social channels. Results are similar to the source measure but semantically richer.

Table 3 – Source data. Period: July-September 2014. Source: Google Analytics

Source	Sessions	
Direct	623	83.1%
Google	52	6.9%
Blog IDEE (GIS blog)	15	2.0%
GEOSLAB	11	1.5%
UniGIS (GIS blog)	10	1.3%
Twitter	9	1.2%
W3C site	7	0.9%
Feedly (News aggregator)	6	0.8%
UNIZAR	6	0.8%
Mundogeo (GIS blog)	4	0.5%
Other sources	7	0.9%

[3] http://blog-idee.blogspot.com.es/

Table 4 – Channel data. Period: July-September 2014. Source: Google Analytics

Channel	Sessions	
Direct	623	83.1%
Organic Search	52	6.9%
Referral	51	6.8%
Social	24	3.2%

3.3 Behaviour

189 out of 750 sessions are bounce sessions (25.2% bounce rate). A bounce session is a session in which there is a unique interaction with the platform, namely requesting a page. 89.0% of the events triggered by users interacting with the platform are zoomings or pannings. 10.5% of the events are data reviews. 0.5% of the events are clicks on links to linked resources. These interactions generate asynchronous requests to the Linked Map platform. 7.9% of requests are GeoSPARQL requests to the SPARQL endpoint and 6.3% are requests to the Linked Map Service instance (see Table 6). It is remarkable that less than 0.1 % of all requests (includes requests to static content) to the site are requests related to content negotiation of RDF data (see Table 7). This result suggests that machine-to-machine interaction initiated by third parties have occurred very seldom during the experiments.

Table 5 – Client-side events. Period: July-September 2014. Source: Google Analytics

Events related to	Requests	%
Extent change	10,482	89.0%
Data reviews	1,233	10.5%
Links to resources	63	0.5%

Table 6 – Content requested restricted to API requests. Period: July-September 2014. Source: GoAccess

Operation	Requests	%
SPARQL SELECT	4,512	30.3%
SPARQL UPDATE	2,535	17.0%
SPARQL DESCRIBE	2,206	14.8%
LMS User edit	1,372	9.2%
SPARQL CONSTRUCT	1,280	8.6%
SPARQL SELECT SPATIAL	1,181	7.9%
LMS Service (includes proxy requests to remote WMSs)	942	6.3%
LMS Feature	848	5.7%

Table 7 – Server status codes. Period: July-September 2014. Source: GoAccess

Operation	Requests	%
200 OK	4,384,601	94.2%
500 Internal server error	183,621	3.9%
304 Not modified	65,010	1.4%
404 Document not found	6,448	0.1%
301 Moved permanently	3,210	0.1%
302 Moved temporarily	2,910	0.1%
206 Partial content	2,655	0.1%
408 Request timeout	2,482	0.1%
303 See other document	2,448	0.1%
Other	2,488	0.1%

4 Conclusions

This deliverable has given a detailed description of the monitoring procedure of the website and the crowdsourcing platform of the Linked Map subproject. The description ranges from a rationale of the monitoring procedure to the measurement of different dimensions required for the guidance of the project (who is the effective audience, how we acquire users, who users use the platform). The measurement infrastructure has captured data from the client and the server side using popular analytics tools such as Google Analytics and GoAccess.

The monitoring system has been running since July 2014. There were 549 sessions to the website initiated by 295 unique users from 300 different hosts. As expected due to the local nature of the data offered in the project, most of the users come from Spain (85.4%). Given the limited time span available and the small size of the project, these numbers are remarkable. A review of acquisition data reveals that mails sent to mailing lists and personal contacts promoting the project are probably the main source of visits and, hence, participation in crowdsourcing activities.

Data show that the platform engages few users to consider editing resources although users explore its content. Nevertheless, monitoring has exposed different patterns of interaction with the platform and has helped the project to understand results of experiments of WP19. The analysis of monitoring data has helped us to understand better how crowdsourced reviews were produced. It could be worth to analyse in future projects if merging monitoring data into crowdsourced data can increase the quality of results.

Annex A Measurement glossary

This annex defines each of measure used in this monitoring report, along with its measurement scale, measurement formula and measurement units when applicable. Measures here pertain only to two measurement scales: nominal and ratio. A measure with nominal scale assigns place each measured value into an unordered category. A measure with ratio scale works with ordered values with intervals of equal meaning and a concept of absolute zero.

Table 8 – Base audience measures

Measure	Method	Scale	Units
Sessions	Number of sessions or groups of interactions performed by users in a period; each user' session ends after 30 minutes of inactivity.	Ratio	*sessions*
Duration of sessions	Accumulated duration of sessions.	Ratio	*seconds*
Unique users	Number of distinct users that have initiated a session.	Ratio	*users*
Hosts	Number of distinct hosts from which users have initiated a session.	Ratio	*hosts*
Bounce users	Number of distinct users that have initiated only one session.	Ratio	*users*
Returning users	Number of distinct users that have initiated more than one session.	Ratio	*users*
Users per country	Number of distinct users that have initiated a one session from a device located in a specific country.	Ratio	*users*
Page views	Number of pages visited by users.	Ratio	*pages*

Table 9 – Derived audience measures

Measure	Method	Formula	Units
Pages per session	Ratio between page views and sessions.	$\frac{page\ views}{sessions}$	$\frac{pages}{session}$
Avg. Session duration	Ratio between duration of sessions and sessions.	$\frac{duration\ of\ sessions}{sessions}$	$\frac{seconds}{session}$

Table 10 – Acquisition measures

Measure	Method	Scale	Units
Source	Number of new users that arrive from the same point of origin. Each new user is assigned to a label as follows: • If it arrives from a search engine, its name • If it arrives from a site, its domain • Otherwise is classified as *direct*	Nominal	*users*

Measure	Method	Scale	Units
Channel	Number of new users that arrive from points of origin with the same content type. Each new user is assigned to a label as follows: • If it arrives from a search result, *organic search* • If it arrives from social media, *social* • If it arrives from a site, *referral* • Otherwise it is classified as *direct*.	Nominal	*users*

Table 11 – Base behaviour measures

Measure	Method	Scale	Units
Bounce sessions	Number of sessions where there is a single interaction.	Ratio	*sessions*
Content requested	Number of requests per content class. Each URL requested is assigned to a class as follows: • If the URL is a SPARQL request, the corresponding operation; if the request is a GeoSPARQL query the term *spatial* will be part of the label of the class. • If the URL is a LMS request, the corresponding kind and entity (see D15.1 [6]). • Otherwise, its URL path classifies it.	Nominal	*pages*
Server status codes	Number of responses per HTTP status code. Each response is assigned to a label according to its HTTP status code.	Nominal	*requests*
Client-side events	Number of actions performed by users in the web client per action class. Each action is classified as follows • If a user zooms or pans the map, *change map extent*. • If a user asks for link data, *get link*. • If a user updates link data, *review link*. • If a user navigates to a linked resource, *go to resource*.	Nominal	*events*

Table 12 – Derived behaviour measures

Measure	Method	Formula	Units
Bounce rate	Ratio between bounce sessions and sessions.	$\frac{bounce\ sessions}{sessions}$	

Annex B Data sources

This annex reports from where measures are collected.

Table 13 – Base measures: data collection source

Measure	Client	Server
Sessions	✔	
Duration of sessions	✔	
Unique users	✔	✔
Hosts		✔
New user	✔	
Returning user	✔	
Users per country	✔	✔
Page views	✔	✔
Source	✔	
Channel	✔	
Content requested	✔	✔
Client-side events	✔	
HTTP status code		✔

References

[1] F. J. Lopez-Pellicer and J. Barrera, “D19.2 Call 2: Linked Map Report on crowdsourcing trade-offs for geospatial data curation,” PlanetData, 2014.

[2] CMMI Product Team, “CMMI for Development, Version 1.2,” Software Engineering Institute, Carnegie Mellon University, CMU/SEI-2006-TR-008, Aug. 2006.

[3] F. García, M. F. Bertoa, C. Calero, A. Vallecillo, F. Ruiz, M. Piattini, and M. Genero, “Towards a consistent terminology for software measurement,” *Information and Software Technology*, vol. 48, no. 8, pp. 631–644, Aug. 2006.

[4] J. H. Allen and N. Davis, “Measuring Operational Resilience Using the CERT® Resilience Management Model,” SEI, CMU/SEI-2010-TN-030, Sep. 2010.

[5] The PageFair Team, “The rise of Adblocking,” PageFair, Aug. 2013.

[6] F. J. Lopez-Pellicer and J. Barrera, “D15.1 Call 2: Linked Map requirements definition and conceptual architecture,” PlanetData, 2014.