

STARS4ALL

Deliverable

D4.11

Data Management Plan (Intermediate Version)

Esteban González
Oscar Corcho
Universidad Politécnica de Madrid

Irene Celino
Andrea Fiano
Gloria Re Calegari
CEFRIEL

Christopher Phethean
UNIVERSITY OF SOUTHAMPTON

June 30th, 2017









Document Information

Contract Number	H2020-688135	Acronym	STARS4ALL
Full title	STARS4ALL		
Project URL	www.stars4all.eu		
Document URL	https://figshare.com/s/ade9b8a217868ca8377d		
EU Project Officer	Fabrizio Sestini		

Deliverable	Number	4.11	Name	Data Management Plan (Intermediate)
Task	Number	4.7	Name	Data Management Plan
Work package	Number	4	WP Code	
Date of delivery	Contract	30/06/2017	Actual	30/06/2017
Code name	4.2		Status	Draft <input type="checkbox"/> Final <input checked="" type="checkbox"/>
Nature	Prototype <input type="checkbox"/> Report <input checked="" type="checkbox"/> Specification <input type="checkbox"/> Tool <input type="checkbox"/> Other <input type="checkbox"/>			
Distribution Type	Public <input checked="" type="checkbox"/> Restricted <input type="checkbox"/> Consortium <input type="checkbox"/>			
Authoring Partner	Universidad Politecnica de Madrid			
QA Partner	Universidad Politecnica de Madrid			
Contact Person	Esteban Gonzalez			
	Email	egonzalez@fi.upm.es	Phone	+34913367473
Abstract (for dissemination)	This document describes the intermediate version of the Data Management Plan. The objective of this document is to describe the project's datasets, including a brief description of their lifecycle. We have analysed in this document datasets generated by the photometer's network, community health module and the game Nightknights.			
Keywords	DMP, data			

Version log / Date	Change	Author
0.1 /28.06.2017	Document creation and inclusion of DMPs generated by the DMPOnline Tool	Esteban Gonzalez
0.2 / 30.06.2017	Quality assurance	Oscar Corcho

Project Information

Partner	Acronym	Contact
Universidad Politécnica de Madrid (Coordinator)	UPM 	Prof. Dr. Óscar Corcho Facultad de Informática Departamento de Inteligencia Artificial Campus de Montegancedo, sn Boadilla del Monte 28660 Spain #@ ocorcho@fi.upm.es #t , #f
Societa' Consortile A Responsabilita	CEFRIEL 	Mrs. Irene Celino
University of Southampton	SOTON 	Assoc. Prof Elena Simperl
European Crowdfunding Network AISBL	ECN 	Mr. Oliver Gajda
Chambre de commerce et d'industrie de region Paris – L'Ile-de-France – CCIP L'Ile de France	ESCP – EUROPE 	Mr. Miguel Palacios
Instituto de Astrofísica de Canarias	IAC 	Dr. Miquel Serra-Ricart
Leibniz Institute of Freshwater Ecology and Inland Fisheries	IGB-BERLIN 	Prof. Dr. Franz Hölker
Universidad Complutense de Madrid	UCM 	Prof. Jaime Zamorano

Contents

1.	Introduction.....	5
2.	Datasets.....	6
2.1.	Photometers' Network (Initial Version).....	6
2.2.	Photometers' Network (Detailed Version).....	7
2.3.	Community Health (Initial Version).....	9
2.4.	Games - NightKnights (Initial Version).....	10

1. Introduction

This document is the intermediate version of the Data Management Plan (DMP). We have used the tool DMPOnline¹ for generating the partial DMPs for each dataset. They have been written with the template specially created for H2020 projects. Regarding previous version of the document (D4.2), some modifications have been done:

- In the dataset generated by the **photometer network**, we have included in each measurements the photometer's coordinates (latitude, longitude).
- We have removed the crowdsourcing dataset, integrating its data in the **community health** and **games** datasets

Also, we have added the following datasets:

- A **detailed version** of the photometers' dataset.
- **Games dataset (initial version)**: In STARS4ALL, games with a purpose will be developed to acquire and process data from LPs. These data will generate two datasets: i) one with the results of the tasks execution, and ii) another with the gamification process results like points, badges, etc ... The plan for the first dataset is equivalent to the plan generated for the toolbox (annex). The plan for the second dataset will be ready in the next version.
- **Community Health dataset (initial version)**: STARS4ALL is a project oriented to create awareness among citizens, so for us, it is very important to measure the impact of CAs on the community. It goes without saying that this data must be open and shared with the community in order to engage newcomers and increase their participation.

¹ <https://dmponline.dcc.ac.uk/>

2. Datasets

2.1. Photometers' Network (Initial Version)

Data set description

Light pollution samples will be continuously taken by the photometer network deployed in the project. The specification of the photometers is provided in deliverable D4.1 (<https://figshare.com/s/76f01dec468f8286f781>). Each photometer will send data to a message broker, which will send it to the subscribers. In our case, the subscribers will insert data in our data portal (also described in deliverable D5.4).

Raw data will be acquired every 5 minutes. This default data acquisition rate may be reduced to 1 minute if necessary. Raw data will be generated as a 10-row CSV in ASCII format.

This means that each photometer will generate 288K per day in the worst scenario (1 sample / minute). Following the STARS4ALL Key Performance Indicators, which are described in the Document of Action (Section 2.1.1.2), at the end of the project, 250 photometers will be deployed. This amounts to a volume of data of 72MB per day, 2.16 GB per month and 25.96GB per year. This size has been calculated considering uncompressed data.

Standards and metadata

Night sky brightness data will be archived using the "NSBM Community Standards for Reporting Skyglow Observations", which was officially adopted at the 12th European Symposium for the Protection of the Night Sky and endorsed by the International Dark Sky Association (IDA) and by the International Astronomical Union (IAU) in Beijing 2012 (SpS17: "Light Pollution: Protecting Astronomical Sites and Increasing Global Awareness through Education"). More information about the header files can be found in the following link: <http://darksky.org/light-pollution/measuringlight-pollution/>

The fields present in the dataset are:

- **name**
 - Name of the device
- **mag**
 - Magnitude measured by the photometer
- **tsky**
 - Temperature of the sky measured by an infrared sensor placed in the photometer.
- **tamb**
 - Ambient temperature
- **latitude**
- **longitude**
- **tstamp**
 - Timestamp of the measurement in ISO-8601 format

Data sharing

Users can access to the datasets in three different manners:

- From our data portal (<http://ckan.stars4all.eu>). A description of this data portal can be found in deliverable D5.4.
- Using our data API (D4.9), which is used by our dashboards to visualize data.

- From our Zenodo Community². Monthly datasets will be published and a Digital Object Identifier will be generated for each of them.

We are using the UPM's own servers for data storage, and in the long-term we will hire hosting services with the funds raised by the foundation.

All data generated by this network will be open access and we do not consider any restriction, including embargo periods.

Archiving and preservation (including storage and backup)

As discussed in the data sharing section, data will be archived and preserved using Zenodo. The amount and preservation of the data will depend of the policy applied by the Zenodo consortium. According with its current policy, there is no limitation in the public space and in the preservation time. Also, data files are backed up nightly.

2.2. Photometers' Network (Detailed Version)

Scientific research data should be easily:

Discoverable

Datasets with the measurements generated by photometers are accessible through two portals:

- STARS4ALL data portal based on CKAN (<http://ckan.stars4all.eu>)
- STARS4ALL community in Zenodo (<https://zenodo.org/communities/stars4all/>)

Datasets in Zenodo have their own DOI.

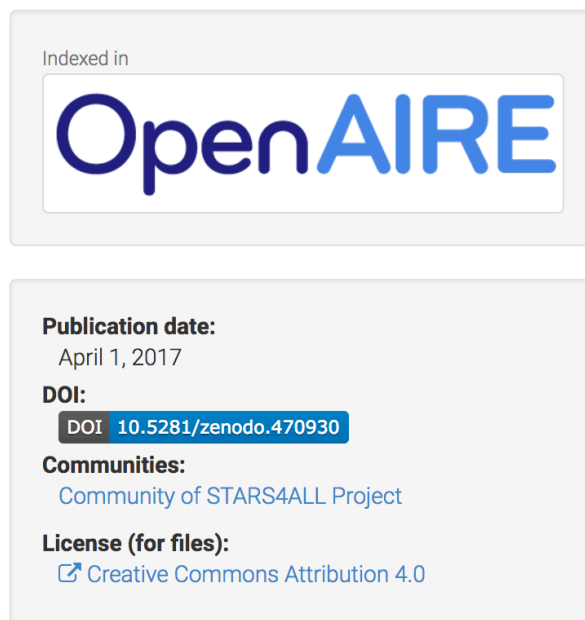


Figure 1: DOI of the dataset with April's measurements

² <https://zenodo.org/communities/stars4all/>

Accessible

Are the data and associated software produced and/or used in the project accessible and in what modalities, scope, licenses?

Apart from using the data portals, the last 30 days of measurements from photometers are accessible through an API (see <http://docs.photometer.apiary.io>).

The associated software is available and accessible at:

- Data extraction: <https://github.com/STARS4ALL/tess-adapter>
- API: <https://github.com/STARS4ALL/photometer-api>
- CKAN image (forked from CKAN project): <https://github.com/STARS4ALL/ckan>

The license of the datasets and the associated software is Creative Commons Attribution 4.0 (CC-BY-4.0).

See <http://opendefinition.org/licenses/cc-by/>

Assessable and intelligible

Are the data and associated software produced and/or used in the project assessable for and intelligible to third parties in contexts such as scientific scrutiny and peer review?

Each dataset is generated with a DOI and the devices, which take the measurements, are conveniently identified (id & location). The interpretations of each measure, as well as the technical components of the device, are open and available to general public in the following link:

<https://figshare.com/s/66d11c0d5a1c9ac81160>

Software is available in Github (see previous point) and the API properly documented in Apiary:

<http://docs.photometer.apiary.io>

Usable beyond the original purpose for which it was collected

Are the data and associated software produced and/or used in the project useable by third parties even long time after the collection of the data?

Yes, this data can be used to generate future mathematical models that can predict or allow understand, light pollution.

Many domains can benefit:

- Urban planners to measure the impact of new light sources or to analyse the urban growth.
- Biologists can study the impact of light pollution on animals or plants
- Medical doctors can use them to deep in the study of light pollution in human health.

These photometers and their measurements will provide the community an historical dataset that shows the evolution of the light pollution. This will be a valuable resource for scientists and public institutions.

Interoperable to specific quality standards

Are the data and associated software produced and/or used in the project interoperable allowing data exchange between researchers, institutions, organisations, countries, etc?

Night sky brightness data will be archived using the "NSBM Community Standards for Reporting Skyglow Observations", which was officially adopted at the 12th European Symposium for the Protection of the Night Sky and endorsed by the International Dark Sky Association (IDA) and by the International Astronomical Union (IAU) in Beijing 2012.

2.3. Community Health (Initial Version)

Data set description

In order to carry out the community health analysis in WP3, we use data from the various LPIs to calculate a set of metrics as set out in D3.4. We are primarily concerned with analysing the number of classifications or tasks carried out, how these are distributed among the community, and how many people in total are participating. Since we cover a range of project types in the LPIs, we for now focus on a set of data that is particular to the classification-based projects, however much of this can be easily adapted to other types of project, too.

The dataset will continuously evolve throughout the project, both as the current LPIs continue to produce data, and as new LPIs are run. This will allow us to compare the metrics from various projects to determine what types attract ‘healthier’ communities, while it will also allow us to manage those communities better by being aware of their current health at a particular moment in time.

The dataset consists of the results of a range of metrics used to measure various aspects of community health. It is therefore derivative data from the raw data obtained by each LPI.

We are generating four metrics per project per day.

Standards and metadata

As data comes from numerous different projects, we do not put restrictions of the format of the data itself, but instead require that certain attributes are contained. Using this, we then carry out a range of calculations on the data to make our own secondary dataset that consists of the metrics that we require to measure community health on the dashboard discussed in D3.4.

The most critical fields required in the data provided to us by an LPI are regarding ‘task runs’, or records of which player has completed which tasks, as follows:

- PlayerID
 - The unique identifier of the player/volunteer carrying out a specific task. This is so we can determine how many contributions each player is making, and how these are distributed among the whole community, as well as the lifetime of a player’s activity. This ID will not include reference to any kind of personal data to maintain the user’s anonymity.
- TaskID
 - The unique identifier of a specific tasks, e.g. an image that is to be classified. This allows us to determine which task was completed by the player on this particular task run.
- Timestamp
 - The time at which the task run was completed. This means we can calculate metrics such as the user’s lifetime (time from first task run to their last task run), and the active period of the project and each task.

Data sharing

Data is shared to the community health analysis dashboard from each individual LPI, as discussed above. In some cases, this is a download of the most recent data dump available from the project (such as from the crowdfunding platform for Dark Skies, Lost at Night and Night Cities), or through APIs (such as for the recently released NightKnights game).

We share the results of the community health metrics via the dashboard that will position projects on a two-dimensional matrix, and show their path over time to indicate their changing health. Where possible, we will release project-specific snapshots of these metrics as datasets that can be downloaded for additional study. Once the dashboard is stable we will aim to release a monthly snapshot of the metrics calculated as part of this analysis.

We will ensure that datasets will not include personal data from users, maintaining their anonymity.

Archiving and preservation (including storage and backup)

The data for the community health analysis is currently hosted at the University of Southampton in the UK, where it will be backed up to allow recovery as necessary. The computed metrics data is stored in a database and therefore it may be moved easily should we deem that it requires hosting elsewhere during the project.

We will ensure that the data will remain available throughout the project duration (until December 2018). We will archive the data regarding the computed metrics for community health measures and will then be stored and preserved by the STARS4ALL foundation servers.

Furthermore, monthly datasets will be accessible from our data portal and our Zenodo community.

2.4. Games - NightKnights (Initial Version)

Data set description

Night Knights (<https://www.nightknights.eu/>) is a Game With A Purpose (GWAP) aimed at Data Linking; more specifically, by playing the game, players help in classifying images taken from the International Space Station (ISS) with respect to a given taxonomy (hence, links are created between images and categories). In short, all actions by all players are weighted and aggregated by the game, producing a “crowdsourced” classification of ISS pictures that can be used for subsequent analysis of the light pollution phenomenon. Details on the Night Knights application and its internal functioning are given in deliverables D4.4, D4.6 and D4.10.

The data set consists of the output data of the Night Knights game. The data are stored in a MySQL database securely accessible by means of a dedicated API, as described in the following.

Standards and metadata

The format of the dataset was defined in a way to facilitate its subsequent reuse, i.e. the game output contains all information needed for further processing and evaluation of the data.

Night Knights data gives information about three main themes:

- Classification results (classified photo, assigned classification, number of players needed to come to a classification agreement)
- Players’ actions (“log” of all classification actions done by anonymized game users)
- Game evaluation data (including KPIs like number of tasks available/started/completed, number of players, total played time, throughput, average life play, etc.)

No personal information about game players is given, in order to respect user privacy, including the application privacy policy available here: <https://www.nightknights.eu/#/privacy>.

We are also evaluating the possibility to make available the classification results (first bullet point of the above list) in RDF format according to the Human Computation Ontology (<http://swa.cefriel.it/ontologies/hc#>), which preserves the provenance of the collected information. In this case, each “consolidated information” consists of an RDF triple (with subject=ISS photo, object=aggregated classification and predicate=link) enriched with metadata about the consolidation process.

Finally, it is worth noting that the ISS images are provided by the NASA’s Gateway to Astronaut Photography Of Earth (courtesy of the Earth Science and Remote Sensing Unit, NASA Johnson Space Center, <https://eol.jsc.nasa.gov/>), free of any copyright restrictions. The Night Knights dataset contains links to the original photo location.

Data sharing

Data sharing is implemented through a secure Web API exposed on the same server where the game is running. The full detailed documentation to access the API is documented according to the API

Blueprint format and available online at <http://docs.crowdtaskmanagement.apiary.io>. All API responses are described, including their output JSON format that contains the specified information. The actual URL to access the API is not public given, the access restrictions explained below.

Access is regulated through token-based authentication. Token can be obtained at the administrator authentication endpoint by providing username and password; once obtained the token it can be used to access the API through the Authorization header field with the ‘Bearer’ scheme. Token expires after 60 minutes; the administrator credentials can be requested by writing to the following e-mail address: contact@stars4all.eu. Data access is currently restricted to project partners, also to prevent potential malicious use during the online competition based on the use of the Night Knights game. Decision on the actual access by third parties will be evaluated by STARS4ALL partners on a case-by-case basis. Nevertheless, datasets with the solved tasks will be uploaded periodically to our data portal and to Zenodo.

Currently the data is used by the University of Southampton, in order to analyse metrics around community health, so that the engagement levels with NightKnights may be compared with other LPIs, and allow us to notice when intervention is required to re-engage with users. Details about the community health dashboard are available in D3.4, and the dataset used in this is discussed in the ‘[Community Health data model](#)’.

The photos classified as “city”, “stars” and “black” will also be used to improve research on light pollution, in order to extend the coverage of the analysis of the phenomenon. The “city” photos can be used to map the light pollution effect and “stars” and “black” photos are used for calibration in order to measure that effect.

Archiving and preservation (including storage and backup)

Universidad Politecnica de Madrid is responsible for backup and recovery of the dataset. Night Knights data is stored in a MySQL database hosted in an Aurora DB Instance (Amazon Web Services). The class of instance is db.t2.medium (2 virtual CPUs and 4GB)

The dataset will remain available and accessible through the aforementioned API until the actual game will be available online, which means at least until the end of the STARS4ALL project (December 2018).

As previously commented, datasets will be preserve in our data portal and our Zenodo community. After the end of the project, the STARS4ALL foundation will have the responsibility of the maintenance of the data servers.