QROWD - Because Big Data Integration is Humanly Possible

Innovation Action

Grant agreement no.: 732194

D10.1 – Data Management Plan

Due Date	30 May 2017
Actual Delivery Date	30 May 2017
Document Author/s	Luis-Daniel Ibáñez (University of Southampton)
Version	1.0
Dissemination level	PU
Status	Final
Document approved by	Claus Stadler (InfAI)

Project co-funded under Horizon 2020 Research and Innovation EU programme

TABLE OF CONTENT

LIST OF ABBREVIATIONS
(if there is need for such)

FAIR	Findable, Accessible, Interoperable and Re-usable
DMP	Data Management Plan
OASC	Open Access Smart Cities Initiative

EXECUTIVE SUMMARY

The Data Management Plan (DMP) describes QROWD's data management life cycle for the data to be collected, processed and/or generated, as part of making its research data findable, accessible, interoperable and re-usable (FAIR), following the guidelines for FAIR Data Management[1]. FAIR data management guarantees that the advancements and results developed on top of these data can be replicated and exploited by future EU-funded initiatives and the community in general.

The key target readers of the DMP are the warrants of the Open Research Data Pilot program that will check the compliance of the project with H2020 guidelines, and members of the research community interested in replicating and/or reproducing the results of QROWD's research and development. Technical partners of the consortium, that will use it as a guide for publishing and maintaining data associated to bespoke research.

In this deliverable we describe the processes, policies and tools the QROWD consortium will use to ensure FAIR data management, that can be summarized as follows:

- Findability: Published datasets will be assigned a DOI. We will use the DataCite and DCAT-AP vocabularies to describe them, and consider the use of relevant description vocabularies issued from the OASC initiative.
- Accessibility: Research datasets will be published in the institutional repositories of the University of Southampton and University of Trento, that comply with the relevant european directives on research data management. Active data research will be managed by the leading partner.
- Interoperability: In addition to the descriptions used for accessibility, we will select the most appropriate standards to format the data. We will pay particular attention to those issued from the OASC initiative.
- Re-usability: All research data outputs will be published with an open license with the exception of those corresponding to the TomTom use case.

[1] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

1 DATA SUMMARY

Data in QROWD is divided in three groups, according to its use and its relation with other WPs:

1. Data collected and generated for testing and improving analytics and hybrid discovery capabilities of the QROWD's platform. This comprises the input data in RDF format on which the analysis/discovery processes will be run, the "ground truth" data used to assess the effectiveness of the newly developed approaches, and the crowdsourced data used either as part of the "ground truth", or to improve the analysis/discovery processes. All these data will be made openly accessible.
2. Data collected and generated for the Trento municipality use case (cf. D2.2). This comprises data transformed from municipality's data sources (static datasets and sensor data) and data crowdsourced from citizens as a complement to sensors. Transformations from Open Data will remain open. An appropriately curated and anonymized subset of the crowdsourced data that allows the reproduction of experiments conducted in connection with this use case will be published.
3. Data collected and generated for the TomTom use case (cf D1.1). This use case re-uses some data from the Trento municipality, but also includes data proprietary to TomTom that cannot be released due to being core to TomTom's business model.

Data used for use cases is detailed in the Data Catalog deliverable (c.f. D4.1) and in each use case description (cf. D1.1 and D2.2). During the project, Data will be under the custody of the research partner leading the work package, in the case of WP2, UniTN will take the responsibility. In the case of WP1, the custody is shared between TomTom and InfAI. During this phase, access to the data will be restricted to consortium partners.

For each publication, the leading research partner will be in charge of annotating and storing the associated dataset following their research data management policies (cf. section 2.1).

2 FAIR DATA

2.1 FINDABILITY

Each dataset made available during the project will be assigned a Digital Object

Identifier and annotated using the latest version of the DataCite schema[2]. We will also describe them using the DCAT-AP extension for scientific datasets recently proposed by the European Commission Joint Research Centre[3]. The extension is a close direct map to DataCite, but in RDF format, ensuring the link of the dataset with the Web of Data and improving its discoverability.

Following our collaboration the Open and Agile Smart Cities Initiative (OASC), we will study the applicability of any vocabulary for describing Smart Cities data issued from them. This plan will be updated accordingly.

Internal versioning will follow the incremental 0.x format until publication of the linked scientific contribution, point from which the numeration will follow the 1.x format for minor corrections or improvements.

2.2 ACCESSIBILITY

Data corresponding to research conducted for the Trento municipality use case will be made available after undergoing an anonymisation process (cf. deliverable 11.1). Data corresponding to the TomTom use case will not be publicly available due to the reliance of the business model of TomTom on it. The possibility of releasing a subset of the data is currently being discussed internally. This plan will be updated according to the final decision made.

Data corresponding to research conducted for WP5 and WP6 will be released to the community with an open license.

While research is being undertaken, accessibility to datasets will be limited to the members of the consortium. During this stage, partners leading the specific research will be in charge of the storage and accessibility for the rest of the partners that require it.

Data linked to scientific publications will be published in the institutional repository of either the University of Southampton or the University of Trento, following their research data management policies. University of Southampton will use the PURE[4] repository. The Research Data Management policy of the University is publicly available[5]. UniTN will use its IRIS[6] repository, that complies with the European Commission Recommendation on access to and preservation of scientific information (July 17th 2012) the H2020/ERC Model Grant Agreement and the UNITN Open Access Policy[7].

[2] https://schema.datacite.org/meta/kernel-4.0/
[3] https://www.w3.org/2016/11/sdsvoc/SDSVoc16_paper_27
[4] https://pure.soton.ac.uk
[5] http://www.calendar.soton.ac.uk/sectionIV/research-data-management.html
[6] https://iris.unitn.it
[7] http://eprints.biblio.unitn.it/4258/1/policy-ateneo-open-access-2912014.pdf

At the time of the first version of this plan, InfAI is implementing an infrastructure for the management of research code and data. As research in which InfAI will be conducted in collaboration with either Southampton or Trento, research data will be published by one of the two. This plan will be updated when InfAI's infrastructure is ready for publishing research data in compliance with the required directives.

Concerning the tools required to access QROWD's research data. At the time of the first version of this plan, we expect that all datasets will be available in RDF format. Any RDF Graph-Store and SPARQL engine can be used to load them and query them. This plan will be updated if a research data output is in other format.

2.3 INTEROPERABILITY

To ensure inter-operability, datasets will be annotated using the DataCite schema, complemented with the DCAT-AP extension for scientific datasets developed by the Joint Research Centre of the European Commission.

As mentioned in section 2.1, following our collaboration with OASC, the consortium will study the vocabularies and standards issued from them, and align produced datasets when appropriate.

2.4 RE-USABILITY

Datasets connected to the Trento Municipality use-case (WP2) will be re-usable according to one of the following schemes

1. Data transformed from existing data sources (curated or not) will have the same license as the original source.
2. Data collected and generated from the project that is connected to a scientific publication will be made available with a ODC-ODbL[8] license. In the case of data coming from crowdsourcing, appropriate anonymisation processes will be apply before release (cf. D11.1).

Datasets connected to the TomTom use-case (WP1) will not be re-usable. At the time of submission of the deliverable, the possibility of providing sample datasets that could be shared with the research community is under discussion. This plan will be updated accordingly after a final decision is made.

Datasets published in connection with a scientific publication will be made available under a ODC-ODbL license.

[8] https://opendatacommons.org/licenses/odbl/summary/

3 RESOURCE ALLOCATION

The archiving infrastructure and resources will be provided by Southampton and UniTN. Successive updates of the DMP will be led by Soton, as coordinating partner.

4 DATA SECURITY

Southampton, has a secure enterprise scale coherent storage solution for active research data. The data stored within this facility is regularly backed up and a copy of the back-up, regularly off-sited to a secure location for disaster recovery purposes. The research data storage platform is solely for the storage of research data. For data deposition, Southampton will make use of its institutional repository (as described in section 2.2) Data will be held for at least 10 years.

UniTN's infrastructure abides to the European Commission Recommendation on access to and preservation of scientific information (July 17th 2012) the H2020/ERC Model Grant Agreement, that has all the security measures to avoid data loss and intrusion.

Data for the TomTom business case will be hold by TomTom, following their industry-grade security measures. Concerning location-identifying data, we cite the following excerpt from TomTom's policy:

"*Within 24 hours of you shutting down your device or app, TomTom automatically and irreversibly destroys the data that would allow you or your device to be identified from the location data we received.*

For Traffic, SpeedCameras, Danger Zones and Weather we delete the information within 20 minutes after you have stopped using the service by shutting down your device or app. We do not know where you have been and cannot tell anyone else, even if we somehow were forced to.

This, now anonymous, information is used to improve TomTom's products and services, such as TomTom maps, Traffic, products based on traffic patterns and average speeds driven, and for search queries to inform businesses how well-received their information is. These products and services are also used by government agencies and businesses."