New Tools for Terrain Gravimetry
NEWTON-g
Project number: 801221

Deliverable 3.1

Database Structure

Lead beneficiary:	Koninklijk Nederlands Meteorologisch Instituut
Dissemination level:	Public
Version:	Final

NEWTON-g has received funding from the EC's Horizon 2020 programme, under the FETOPEN-2016/2017 call (Grant Agreement No 801221)

DOCUMENT INFORMATION

Grant Agreement Number	801221
Acronym	NEWTON-g
Start date of the project	1 June 2018
Project duration (months)	48
Deliverable number	D3.1
Deliverable Title	Database Structure
Due date of deliverable	31 May 2019
Actual submission date	27 May 2019
Lead Beneficiary	Koninklijk Nederlands Meteorologisch Instituut
Type	ORDP: Open Research Data Pilot
Dissemination level	PU - Public
Work Package	WP3 – On-field application

Version	**Date**	**Author**	**Comments**
v.0	23/01/2019	M. Koymans	Creation
v.1	26/03/2019	A. Messina, F. Cannavò, D. Carbone, R. Middlemiss	1st Revision
v.2	22/05/2019	A. Messina, D. Carbone, J. Lautier-Gaud, E. Rivalta	2nd Revision
Final	22/05/2019	M. Koymans, E. de Zeeuw - van Dalfsen	Validation

TABLE OF CONTENTS

1. EXECUTIVE SUMMARY

This is deliverable D3.1 – Database Structure – of the H2020 FET-Open project NEWTON-g (GA No 801221). This document was produced in the framework of WP3 (On-field application). As part of the EC Horizon 2020 research and innovation programme, the consortium of NEWTON-g has committed to ensure findability, accessibility, interoperability, and reusability of data generated and/or collected during the project (FAIR principles).
The aim of this document is to present our strategy for storing and sharing NEWTON-g data, in compliance with the FAIR guiding principles.

2. INTRODUCTION

This document describes the approach to archiving and distributing data generated and/or collected within the NETWON-g project. In particular we present:

1) the selected method of data storage;
2) the design of a metadata catalogue that will serve as an index to the data;
3) the chosen services for data distribution;
4) a plan for data back-up and recovery.

In our design we store file related metadata in a database with a reference to the respective data archived on disk. This approach facilitates the discovery of data through metadata (e.g. by spatial or temporal constraints) at an acceptable granularity and follows the FAIR data management principles (i.e. the findability, accessibility, interoperability, and reusability of data). In the framework of the project, data will be produced by the *gravity imager* consisting of two types of gravimeters: one AQG (Absolute Quantum Gravimeter) and an array of MEMS gravimeters. The two types of instruments feature different output formats and number of complementary sensors. Data coming from the *gravity imager* will be stored in an acquisition archive hosted at INGV-OE. At the end of every day, a file for each sensor (e.g. gravity, temperature, pressure) will be created from the raw data transmitted by the sensors and will be stored in a curated archive hosted by KNMI. Each row in a curated file contains a timestamp followed by a value that represents the parameter sampled at a constant sampling rate. The data curation step does not affect the overall size of the raw data transmitted by the *gravity imager* to INGV-OE.

3. METHOD OF STORAGE

3.1. Storage strategy

Data collected from the deployed instruments are processed and saved in text delimited files, encoded as ASCII and written to file system storage. The file system will include a single NEWTON-g root directory that contains the archive with a folder for each station deployed in the field. In turn, each station folder contains a directory for each specific sensor type. The files are stored on a per-day basis and following the filename convention reported below (figure 1):

NG/MEMS1/GRAV/2020/NG.MEMS1.GRAV.2020-01-01.txt

Signifying the NEWTON-g network (NG), the specific station identifier (e.g. MEMS1, MEMS2, AQG), the particular measurement (e.g. GRAV; vertical gravity acceleration) and the particular year (2020) of the data covered in the file. The last part of the filename indicates the date of the file.

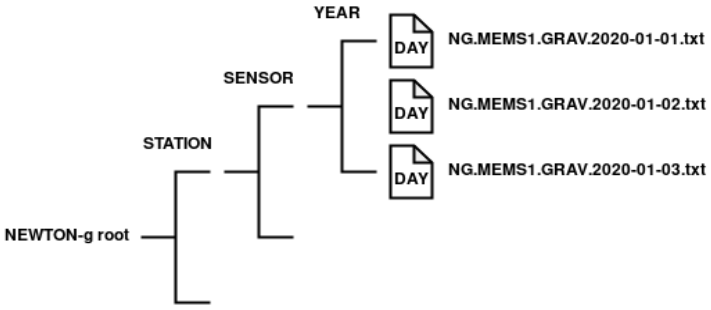

Figure 1: Schematic overview of the archive directory of the NEWTON-g data

3.2. Storage Size

We expect that about 7 parameters will be recorded at each MEMS station at a sampling rate of 1Hz. Assuming each sample requires 36 bytes to be stored (ISO8601 timestamp followed by 10 remaining bytes for the parameter), we obtain a rough estimate of 25MB/station/day. The gravity imager will include a few tens of pixels (individual MEMS stations), implying that it will produce less than 1GB of data per day. This is a manageable amount of data that may be compressed when the lack of storage space becomes a pressing issue.

4. METADATA CATALOGUE

Metadata associated with every archived file will be stored in a relational PostgreSQL database and will be attributed a persistent identifier that can be used to reference a single data file. The persistent identifier will be chosen as the SHA256 hash of the contents of the file and will make public reference to a particular file. By using the checksum as a persistent identifier we can inherently guarantee data integrity. The relational database schema is illustrated in figure 2. The relational database of NEWTON-g will include seven tables and will use relational identifiers to link e.g. data objects to their parent stations. For example: every data object is inherently related to a station installed within the NETWON-g *gravity imager* through a parent station identifier (stat_id) and through a sensor (sens_id). A complete and searchable metadata catalogue will allow to make the NEWTON-g database available to a broader community through HTTP web service (see Section 6 – Data distribution), once the data embargo is lifted (see deliverable D1.2). The metadata catalogue described in this report is in accordance with our previously defined data management plan, described in deliverable D1.2.

5. RELATIONAL DATABASE TABLES

5.1. Tables list

The tables in the relational database of NEWTON-g are listed below.

- DataObject (Metadata describing a single data file saved in the file system)
- SensorObject (Configuration of the sensor that recorded the data)
- SensorTypeObject (Constant sensor properties)
- StationObject (Station metadata, geographical location)
- PublisherObject (The publisher of metadata: information on the NEWTON-g consortium)
- MetricsObject (Waveform metrics metadata)
- DeploymentObject (Additional station installation information)

5.2. Tables description

5.2.1. DataObject

This table contains the metadata for a single data file covering one day from one sensor from one station and is identified by a unique persistent identifier and the temporal coverage of its content. The entry in this table points to the location of the file on disk and has a many-to-one relationship to an entry in the sensor table (SensorObject). This table also contains a one-to-one relationship to the file sample metrics (MetricObject) that serves as quality indicator.

5.2.2. SensorObject

The SensorObject table describes the collection of available sensors that produce the data. Sensor measurements may include but are not limited to: gravity acceleration, X & Y tilt, (external) temperature, pressure, humidity, and, for the AQG, laser polarization. Each row represents a single sensor and is referenced by multiple data objects. All sensors also reference a single parent station.

5.2.3. SensorTypeObject

The SensorTypeObject table describes constant properties of the model of the sensor that cannot change throughout the deployment. The SensorObject table references this parent table with a many-to-one relationship.

5.2.4. StationObject

The station table contains information on all stations deployed within the network, e.g.: geographical location, start & end time of station operation, etc.

5.2.5. MetricObject

The metric table has a one-to-one relationship with the DataObject table. It serves as a quality catalogue for the data that is collected by the gravity imager. Through the summary of sample metrics for a given day file, poor quality data can be identified, or data with a specific characteristic can be found. These metrics include the average, minimum, maximum, standard deviation, quadratic mean, and 25, 50, and 75 percentiles of the time series of a given measurement channel.

To illustrate: file **NG.MEMS1.GRAV.2020-01-01.txt** has one entry in the metric table with a column for each of the aforementioned metrics. This file may potentially show an anomalously high quadratic mean of the vertical gravity as would be visible from the table entry. By storing summarized sample information in our database, we allow users to easily check the quality of data produced by each sensor.

5.2.6. DeploymentObject

The deployment table serves as a table that references a station and contains additional information on the deployment (e.g. transmission type, installation facilities, etc.) and a filepath that may include additional resources that are important for the station deployment (e.g. photographs, maintenance logs, etc.).

5.2.7. PublisherObject

The publisher table is a supplementary table with static metadata that describes metadata concerning the owner and publisher of the data (i.e. the NEWTON-g consortium). This table contains information that is needed for future compatibility with EPOS. The publisher metadata is complemented with the fields defined in the

semantic vocabularies of the EPOS DCAT-AP data model to facilitate the integration of the data in larger European research infrastructures.

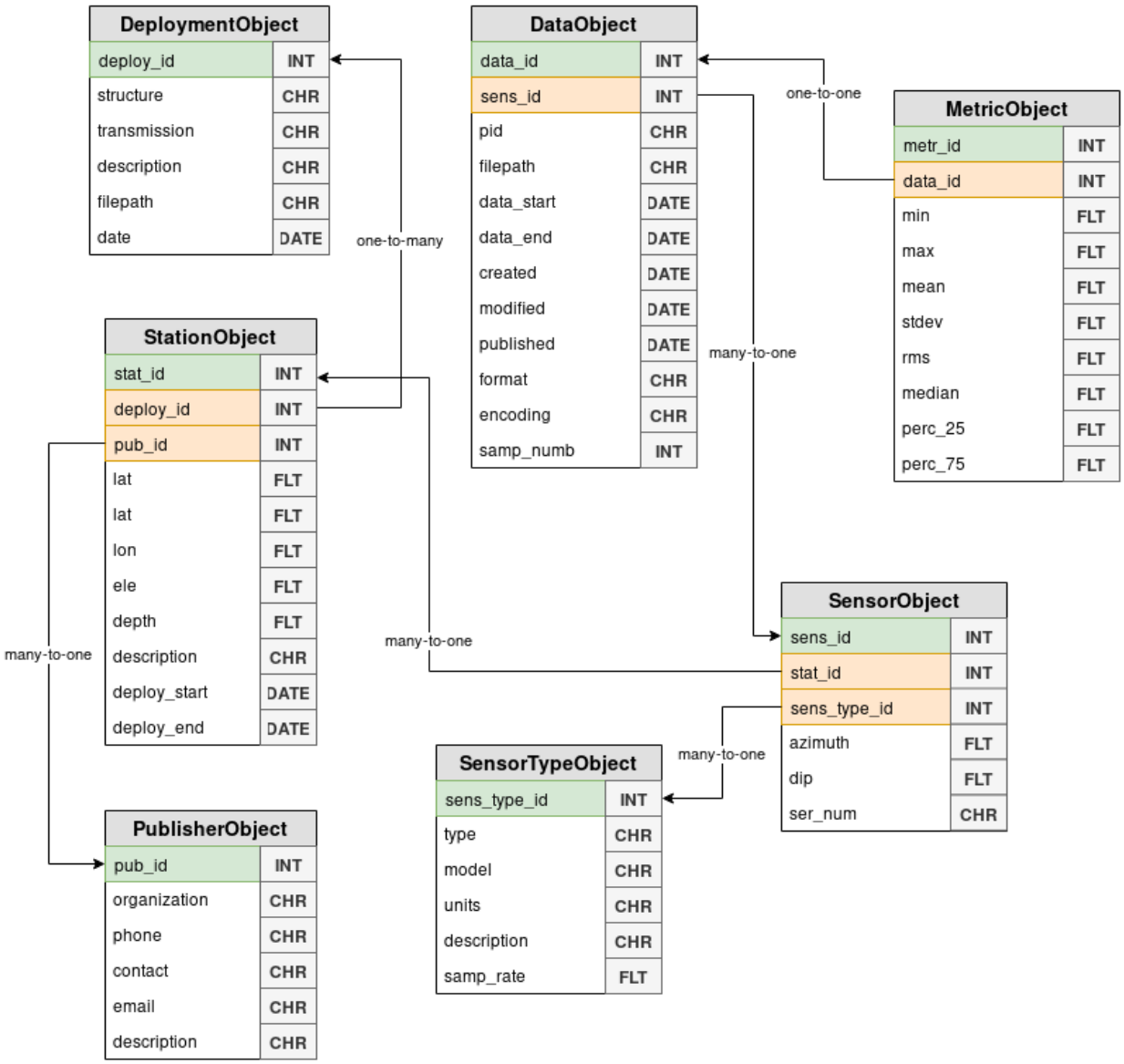

Figure 2 – Relational database schema designed for the NEWTON-g database. Green and orange colours represent primary and a foreign keys of a table, respectively. Grey fields indicate the type of field. Object relations are illustrated by arrows with their respective relationships indicated.

6. DATA DISTRIBUTION

A conventional way of data distribution is done through the FTP protocol and we will offer this way of data access to project partners during the embargo period. After the embargo is lifted (see D1.2 for details), we will publicly offer access to the data through an HTTP API. This method of distribution is the de facto standard supported by wider communities, including the EPOS infrastructure and has many advantages over alternative methods. This interface will allow any client to programmatically request part of the NEWTON-g dataset of interest. The API can be queried with a particular set of request options (e.g. start, end time of data) and return the data that matches the submitted constraints. For

example: to request gravity data for the first week of 2020 for the station identified by MEMS1 a request can be made to the API:

https://api.newton-g.eu/newton-g/data?start=2020-01-01&end=2020-01-07&station=MEMS1&type=gravity

Note: the hostname of the API request is an example. The precise specifications of the API are outside the scope of this document.

The service will use the metadata catalogue described in this document to locate the data files, trim it to the requested extent, and return it to the user in a standardized format (e.g. JSON, CSV). Metadata associated with the data will always be publicly available and give an overview of the available data collected by the gravity imager. The chosen approach promotes the findability and provides easy accessibility of the data, in compliance with FAIR guidelines. This implementation is powerful and facilitates the discovery of data through spatio-temporal constraints, which is a common workflow for many researchers.

7. DATA BACK-UP AND RECOVERY

Besides the raw data stored at INGV-OE, the archived data will be stored off-site with a cloud provider, which will serve as a back-up and will be maintained by KNMI. This archive will be synchronized daily with the raw archive hosted at INGV-OE. In case of catastrophic failure of any infrastructure the data can be recovered from the copy. KNMI will maintain the archive, metadata catalogue, and develop the web service that distributes the data to the clients until the project is terminated in June 2022. Afterwards, the database produced by the *gravity imeger* during its field test at Mt. Etna (year 3 and 4 of the project) will be transferred to another appropriate archive to allow data re-use and sharing after the end of the project, under the terms set out in D1.2.