This project has received funding from the European Union's Digital Europe programme under grant agreement No 101100625

D4.7 – QUALITY AND VALIDATION PRINCIPLES FOR THE OME2 LARGE-SCALE TOPOGRAPHIC DATASET

Document authors	**Noémie GREMEAUX (IGN-France)** **Nathalie DELATTRE (IGN-Belgium)**
Document contributors	Dominique LAURENT (IGN-France) Karin MERTENS (IGN-Belgium)

Abstract

This document is a first version of the quality and validation requirements defined for the OME2 topographic large-scale dataset. It describes the general requirements and principles for checking the quality of the data created from national large-scale datasets using the OME2 topographic production workflow. It also describes the necessary validation steps to be included within the workflow in order to ensure that it can run smoothly and provide successful results.

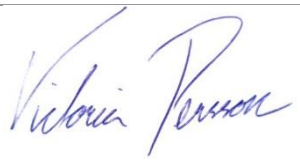
Keywords

Validation procedure: is based on specific measures ensuring that a dataset fulfils the required specifications enabling the production workflow to run smoothly by ensuring compliancy with the data specifications and cross-border continuity.

Quality assessment: is based on specific measures evaluating the quality level of a dataset as regards data content and harmonisation. It aims mainly at evaluating the usability of the dataset according to user needs.

Information Table

Project information	
PROJECT ID	101100625
PROJECT FULL TITLE	Open Maps for Europe 2
PROJECT ACRONYM	OME2
FUNDING SCHEME	Digital Europe Programme (DIGITAL)
START DATE OF THE PROJECT	1st January 2023
DURATION	36 months
CALL IDENTIFIER	DIGITAL-2022-CLOUD-AI-02

Deliverable information	
DELIVERABLE No AND TITLE	D4.7 – Quality and validation principles for the OME2 large-scale Topographic dataset
TYPE OF DELIVERABLE[1]	R
DISSEMINATION LEVEL[2]	PU
BENEFICIARY NUMBER AND NAME	999646111 - INSTITUT NATIONAL DE L'INFORMATION GEOGRAPHIQUE ET FORESTIERE - IGN
AUTHORS	Noémie GREMEAUX, Nathalie DELATTRE
CONTRIBUTORS	Dominique LAURENT, Karin MERTENS
WORK PACKAGE No	4
WORK PACKAGE LEADER **WP LEADER VALIDATION DATE**	Noémie GREMEAUX (IGN-France) 30/03/2023
COORDINATOR VALIDATION DATE	3rd April 2023
COORDINATOR SIGNATURE	

[1] Use one of the following codes: R=Document, report (excluding the periodic and final reports)
DEM=Demonstrator, pilot, prototype, plan designs
DEC=Websites, patents filing, press & media actions, videos, etc.
OTHER=Software, technical diagram, etc.
ORDP : Open Research Data Pilot
[2] Use one of the following codes: PU=Public, fully open, e.g. web
CO=Confidential, restricted under conditions set out in Model Grant Agreement
CI=Classified, information as referred to in Commission Decision 2001/844/EC.

Table of Contents

1. Scope

This document is a first version of the quality and validation requirements defined for the OME2 topographic large-scale dataset. It describes the general requirements and principles for checking the quality of the data created from national large-scale datasets using the OME2 topographic production workflow. It also describes the necessary validation steps to be included within the workflow in order to ensure that it can run smoothly and provide successful results.

This first version is based on the quality and validation principles defined for the current EuroGeographics pan-European products, namely EBM (EuroBoundaryMap), ERM (EuroRegionalMap) and EGM (EuroGlobalMap). These principles will be detailed in the final version of the document so as to be fully adequate to the OME2 user needs.

The data model and specifications of the OME2 central topographic large-scale database will be more precisely defined in the course of the project. Since the quality and validation rules are very dependent on the final data model, this document will be completed once the data model and specifications are fully stabilised.

The OME2 topographic large-scale dataset will be created and updated using national large-scale datasets, with their own quality management procedures defined and applied by NMCAs. The OME2 quality and validation requirements aim at measuring the general quality level of the aggregated data created by assembling these national components, but the resulting data will of course be largely dependent on the quality of the source data.

2. General information on the production workflow

The OME2 topographic large-scale dataset will be created and updated using national large-scale datasets from the NMCAs.

The OME2 production workflow has been conceived as a centralised process. All the procedures will be centralised and will run on a dedicated cloud infrastructure under the maintenance of the product manager of the OME2 large-scale dataset.

The following figure illustrates the target OME2 production workflow:

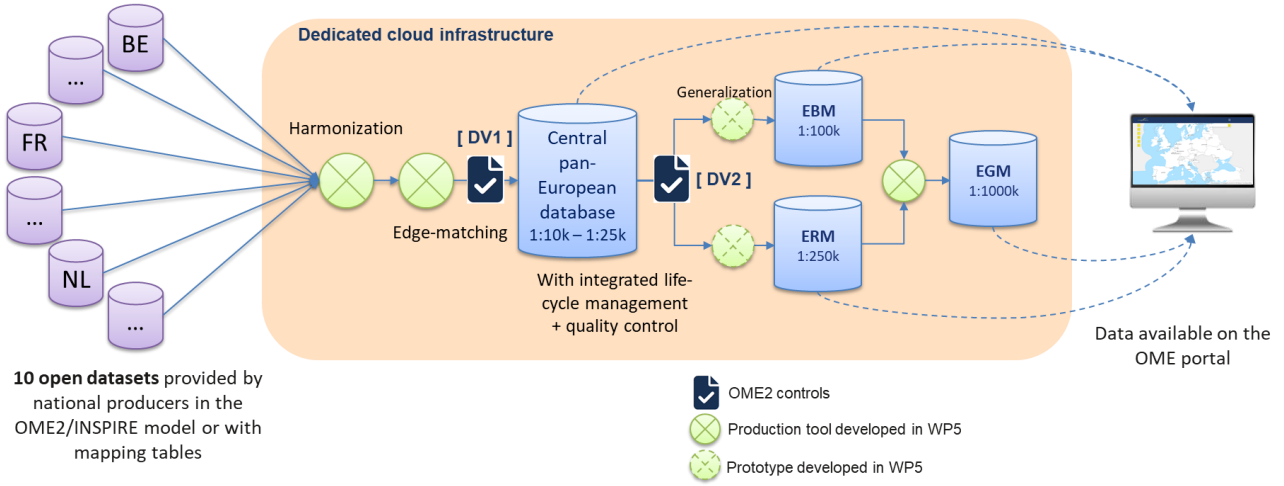


Fig 1: OME2 production workflow

The OME2 production workflow will apply a number of transformations on national datasets in order to reach a suitable level of harmonisation for pan-European use, which include:

- a harmonisation step, consisting of a conversion from national specifications to the OME2 data specifications; geometric conversions might be considered if necessary. A mapping table between the national specifications and the OME2 data specifications will be provided by the NMCAs and the conversion tools will be implemented and run on each national dataset in the dedicated cloud infrastructure;
- an edge-matching step, aiming at connecting and/or aligning features along the international boundaries in order to reach a seamless coverage. For that purpose, a database gathering agreed and harmonised international boundaries between countries will be set up.

The result will be a centralised pan-European dataset.

A first data validation step [DV1] will be performed on national component with the purpose of ensuring a well-running integration process. This first step will validate the compliancy of the national data component against the OME2 large-scale data specifications and measuring the good conditions for a proper edge-matched network. [DV1] will be performed at the end of the harmonisation and edge-matching steps before integrating new national data or updates into the central database.

During the integration phase, the validated national data components will be submitted to an integrated life-cycle maintenance, ensuring the unicity, stability and consistency of the unique identifiers and providing harmonised life-cycle information on what has been changed, added or deleted compared to the previous version of the OME2 central dataset. The adopted principles for managing life-cycle information and change-only updates will be based on the comparison between two national data releases but OME2 will apply its own life-cycle rules and unique identifiers management, and not directly reuse those of the source national datasets.

The second validation step [DV2] will then be applied on the integrated pan-European dataset and will more aim more specifically at checking unique identifiers and life-cycle information and validating the continuity of the networks at cross-border level.

3. Data validation

This chapter describes the general principles of this validation. Detailed validation specifications will be provided in a second version of this document.

3.1. Before data integration [DV1]

The following controls will be applied before integrating a new dataset in the OME2 database, in case of data updates or if a new country is added. The data to be integrated will first go through the harmonisation (model and geometric conversion) and edge-matching steps. The following checklist aims at ensuring the quality of the data resulting from these two steps.

3.1.1. General content

- All feature classes are part of the OME2 data schema.
- All feature class names, attribute names and attribute values are properly spelled.
- The data is in in the horizontal and vertical Coordinate Reference Systems defined in the OME2 data specifications.

3.1.2. Geometric resolution

- All geometries are valid according to OGC standards.

3.1.3. Data model and attribute structure

- Each feature class has the appropriate geometrical definition(s).
- The data schema is consistent with the OME2 data schema
- All attributes and attribute values are consistent with the OME2 data schema.
- All features have a valid country code (possibly indicating more than one country).

3.1.4. Topology

- All feature classes comply with the topological relationships described in the data specifications.
- The data is topologically connected along international boundaries (15% of errors are allowed). The means to measure this error percentage will be fully documented and delivered with the first prototype of the database at the end of the first year of the project.

3.1.5. Edge-matching

- The data is harmonised with the common technical international boundaries defined by OME2.
- No data is located outside of these boundaries.

3.2. After data integration [DV2]

The following controls will be performed once the new dataset has been integrated into the central database. They aim at ensuring the consistency of the newly integrated data with the existing content of the database.

3.2.1. Inter-theme consistency

- Transport and Hydrography data is located within administrative boundaries.
- Other inter-theme consistency rules defined in the data specifications are complied with.

3.2.2. Edge-matching

- Transport and Hydrography networks are continuous across international boundaries.
- A 15% error margin is tolerated.

3.2.3. Specific checks

- Unique identifiers and life-cycle information are correctly filled on all objects.

4. Quality assessment

As mentioned before, the OME2 topographic large-scale dataset will be creating by assembling national large-scale datasets, provided by NMCAS, who produce and update them for their own national needs and with their own constraints. Therefore, the data harmonisation level will depend on the discrepancies or similarities between countries in matter of data content, selection criteria and classification. The resulting OME2 pan-European dataset will also be largely dependent of the information completion rate of the national datasets.

As the OME2 large-scale topographic dataset aims at gathering what is available among NMCAs, there are no requirements regarding mandatory information or completion rates. The objective is to include as much data as possible, even if not all countries can provide the information. However, the OME2 project cannot be held responsible for the quality of national datasets provided by NMCAs

The OME2 quality assessment will therefore consist in providing statistical information about the existence or absence of features in national datasets compared to the OME2 data specifications and about the attribute completion rate. These statistics should also be validated by national data providers in the sense that they should be compared with the real world. For example, in Malta, railways will be missing but this corresponds to the real-world situation and should be documented as such.

The main purpose of this quality assessment will be to evaluate and document the quality level of the central dataset in terms of data harmonisation and completion compared with the OME2 data specifications content. The quality assessment specifications might evolve according user needs and applications and will be detailed in the second version of this document.