International Cartographic Association
Association Cartographique Internationale

Integrating Statistical Databases with Geospatial Datasets

Pekka Latvala [a],*, Kim Huuhko [b], Matti Kokkonen [b]

[a] *Finnish Geospatial Research Institute FGI, pekka.latvala@nls.fi*
[b] *Statistics Finland, kim.huuhko@stat.fi, matti.kokkonen@stat.fi*

* Corresponding author

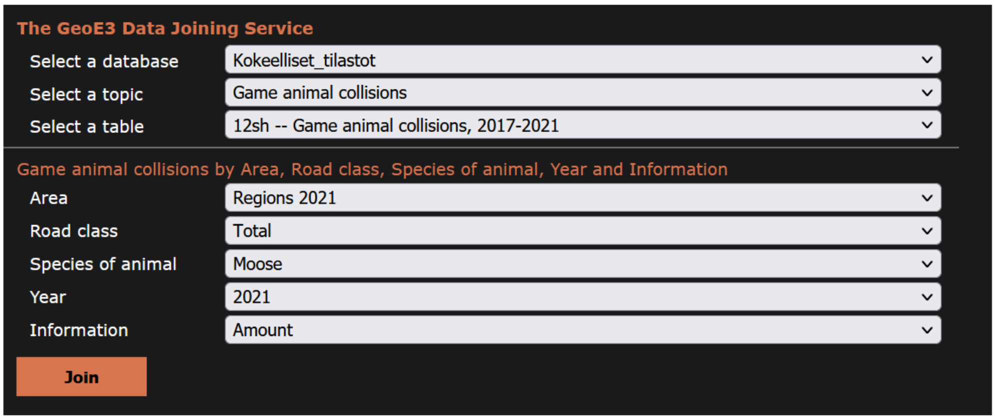
Keywords: OGC API Joins, PxWeb API, data integration

Abstract:

The Open Geospatial Consortium (OGC) is currently specifying a new set of interface standards. One of the emerging standards is the OGC API – Joins specification that is currently in a draft stage. The core module of the OGC API – Joins specification contains functionalities for providing metadata on feature collections and their key values. Data can be joined from the CSV files with the collections by using key values that are available in both datasets. The join response can contain information on the successfulness of the data joining operation and the joined dataset outputs can be provided in several formats. In addition, the OGC API – Joins specification contains a functionality for joining data from the CSV files directly with the GeoJSON files.

The PxWeb is an application for sharing statistical tables online. It is a widely used data delivery system by the National Statistical Institutes in the Nordic countries. The PxWeb API is a service interface that can be used for querying the PxWeb statistical tables. The PxWeb API supports various output formats, including px and CSV.

The Geospatially Enabled Ecosystem for Europe (GeoE3) is a project that aims at creating services that integrate dynamically various datasets with geospatial data. We created a proof-of-concept service in the GeoE3 project that integrates data from the Statistics Finland’s statistical databases with geospatial datasets. The selected statistical theme for the proof-of-concept service was “Game animal collisions in mainland Finland”. The used geospatial datasets were the Finnish municipalities and regions datasets from the years 2017 – 2022. The geospatial datasets were imported into the PostgreSQL / PostGIS database, and their metadata were configured to be served via an OGC API – Joins service as collections. The retrieving of the statistical data and the execution of the data joining process are performed from web browser-based user interface (Figure 1).



The GeoE3 Data Joining Service

Select a database: Kokeelliset_tilastot
Select a topic: Game animal collisions
Select a table: 12sh -- Game animal collisions, 2017-2021

Game animal collisions by Area, Road class, Species of animal, Year and Information

Area: Regions 2021
Road class: Total
Species of animal: Moose
Year: 2021
Information: Amount

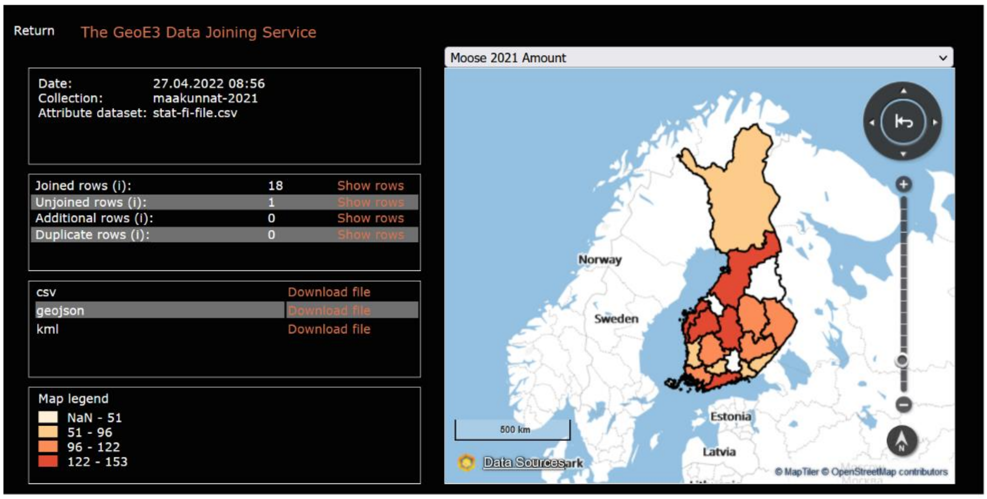
Figure 1. The front page of the user interface.

The statistical data retrieval process consists of two phases. In the first phase, the user selects a statistical table in the user interface (Figure 1) and the table’s metadata are retrieved dynamically in the JSON format from the PxWeb API. The metadata contains information on the variables and their values that can be selected when retrieving the data. The selectable geographical areas are gathered into two groups: municipalities and regions, where selecting a specific group, all areas that belong to that group will be selected. The information on the yearly version of geospatial area classification, to which the data corresponds to, is included in the ‘map’ element in the metadata response. The OGC API – Joins specification requires that the values that are related to a single geographical area are listed in one row in the CSV file.

Abstracts of the International Cartographic Association, 5, 2022.
European Cartographic Conference – EuroCarto 2022, 19–21 September 2022, TU Wien, Vienna, Austria.
https://doi.org/10.5194/ica-abs-5-121-2022 | © Author(s) 2022. CC BY 4.0 License.

Therefore, the selection of other variables than the geographical area was limited to a single value to prevent the PxWeb API for creating multiple rows for each geographical area. The second phase of the statistical data retrieval process consists of formulating the data retrieval query based on the user's selections and fetching the data from the PxWeb API in the CSV format. Depending on the query, the PxWeb API returns reference to the geographical areas either as area names, area codes or as strings that contain both the area names and the area codes. The last form required editing in the client before making the OGC API– Joins query to create matching key values for the data joining operation.

The CSV data are joined with the corresponding collection by executing an OGC API – Joins query. The results of the join operation are visualized in the user interface's result page (Figure 2) that contains a map functionality, information on the successfulness of the join operation and download links for the joined dataset. The map functionality was implemented with the Oskari application's RPC functionality where a GeoJSON file containing the joined data is uploaded together with a styling information to the Oskari server. The map styling was created by classifying the data values into four classes.



Return The GeoE3 Data Joining Service

Date: 27.04.2022 08:56
Collection: maakunnat-2021
Attribute dataset: stat-fi-file.csv

Joined rows (i):	18	Show rows
Unjoined rows (i):	1	Show rows
Additional rows (i):	0	Show rows
Duplicate rows (i):	0	Show rows

csv	Download file
geojson	Download file
kml	Download file

Map legend
- NaN - 51
- 51 - 96
- 96 - 122
- 122 - 153

Moose 2021 Amount

Figure 2. The result page of the user interface.

The PxWeb user interface does not currently enable the user to visualize the statistical data on a map or to download them in a geospatial format, which makes it difficult to examine the spatial distribution of the phenomenon and to create custom analysis for the data. The ultimate end-product of this work could be a user interface where anyone could get any given statistical dataset related to a geospatial area classification extracted from the statistical database, visualized on a map, downloaded in a geospatial format, and saved as an embedded view on a webpage. This kind of user interface should be generic and allow users to dynamically query and filter all relevant datasets according to the corresponding metadata and user needs.

Acknowledgements

The GeoE3 project is co-financed by the Connecting Europe Facility (CEF) of the European Union with the grant agreement number INEA/CEF/ICT/A2019/2063390.

The OpenStreetMap data is available under the Open Database License and the map data under the CC-BY-SA Licence: https://www.openstreetmap.org/copyright.

The OpenMapTiles is available at: http://openmaptiles.org/.

The MapTiler copyright is available at: https://www.maptiler.com/copyright/.

The authors wish to acknowledge CSC – IT Center for Science, Finland for computational resources.

https://doi.org/10.5194/ica-abs-5-121-2022 | © Author(s) 2022. CC BY 4.0 License.