GeoE3_A4_3 Validated guidelines for including tabular data in the Geospatially Enabled Ecosystem

Co-financed by the Connecting Europe Facility of the European Union

Publishing Information/Date

Revision History

Version	Status	Author	Modifications
1.0	Final version	Panu Muhli	Updated version
0.2	Updated	Pekka Latvala	Updated version
0.1	First version	Panu Muhli	Initial version

Distribution

Version	Status	Recipient
1.0	Final	INEA

Table of Contents

Figures

1. Introduction

1.1. Document scope

This document presents user-validated guidelines for integrating tabular data from statistical APIs with geospatial datasets by using the proof-of-concept (PoC) data joining service [1], developed in the GeoE3 project.

Chapter 2 "Guidelines" contains updated technical and functional descriptions of the PoC service and guidelines for its use. The updates reflect changes due to further development of the PoC service since the release of the original implementation guidelines presented in the project deliverable Milestone 8 Table Joining Service (TJS) implementation 30/09/2022 Implementation guidelines for including tabular data in the Geospatially Enabled Ecosystem.

Furthermore, in chapters 3 "Validation" and 4 "Conclusions" we present the method and outcome of user validation of the PoC service and summarize the final conclusions and prospectives for further work.

1.2. Data joining service in a nutshell

The data joining process is managed from the web-based user interface. The user interface contains functionalities for retrieving the statistical table metadata and data from the statistical API together with functionalities for performing the data joining operation. The data integration process is executed via a prototype application that is based on the OGC API – Joins draft standard [2]. The results of the data joining operation are visualized on the user interface that includes a map-based preview together with links where the joined dataset can be downloaded.

2. Guidelines

This section contains guidelines for the use of the proof-of-concept data joining service, information on the used geospatial and statistical datasets, and technical implementation of the service. This section also contains instructions on the use of the PxWeb API of Statistics Finland and the SDMX API of the Eurostat.

2.1. Geospatial Data

The following geospatial datasets are used in the GeoE3 data joining service:

- Finnish municipalities (2017-2023).
- Finnish regions (2017-2023).
- NUTS level 0 areas.

The Finnish datasets have yearly versions available from the years 2017-2023. The geospatial datasets were loaded into PostgreSQL / PostGIS database, and their metadata were configured to be served via the OGC API – Joins service as collections. The Finnish datasets were named in the OGC API – Joins service so that the names contain the area division name in Finnish and the year, to which the dataset applies to, separated by hyphen (example: 'kunnat-2021', for the 2021 municipalities dataset).

The following dataset fields were configured for the Finnish datasets to be key fields in the OGC API – Joins service. The key fields contain unique values, and they can be used for joining the tabular data with the geospatial data.

- Area codes.
- Area names in English.
- area names in Finnish.
- area names in Swedish.

The following key fields were configured for the NUTS areas:

- Nuts identifiers.
- Names written in Latin alphabet.
- Names written in local alphabets.

2.2. Statistical Data

This section contains information on the statistical data and statistical APIs that were used in the proof-of-concept service.

2.2.1. PxWeb API

The PxWeb is an application for publishing statistical tables online. The PxWeb API is a web interface that can be used for querying the PxWeb statistical tables. The statistical table metadata and data from the Statistics Finland's PxWeb API [3] were included to the GeoE3 data joining service.

The database options from the PxWeb API were limited to 'Kokeelliset_tilastot' (experimental statistics), 'Kuntien_avainluvut' (municipality key figures) and 'StatFin'. The experimental statistics database contains statistics relating to traffic network coverage, game animal collisions, population by type of activity and deaths by week. The municipal key figures database contains various key figures relating to Finnish municipalities, including population, share of Swedish speakers of the population, number of families and employment rate. The 'StatFin' database contains statistics relating to first registration of motor vehicles, population projections, population structure and population statistics.

A list of available databases in the PxWeb API of Statistics Finland can be retrieved by making a HTTP GET request to address:

https://pxweb2.stat.fi/PxWeb/api/v1/{language}

where {language} is the one of the following: 'en' for English, 'fi' for Finnish and 'sv' for Swedish.

A list of topics in a specific statistical database can be retrieved by making a HTTP GET request to address

https://pxweb2.stat.fi/PxWeb/api/v1/{language}/{database-id}

where {language} is the requested language and the {database-id} is the identifier of the statistical database from the response of the statistical databases query.

A list of statistical tables under a topic level can be retrieved by making a HTTP GET request to address:

https://pxweb2.stat.fi/PxWeb/api/v1/{language}/{database-id}/{levels} where {language} is the requested language, {database-id} is the identifier of the statistical database and {levels} is the identifier of a topic.

The metadata of a specific statistical table can be retrieved by making a HTTP GET request to address

https://pxweb2.stat.fi/PxWeb/api/v1/{language}/{database-id}/{levels}/{table-id}

where {language} is the requested language, {database-id} is the identifier of the statistical database, {levels} is the identifier of the topic and {table-id} is the identifier of the statistical table.

Some of the statistical topics in the PxWeb API may contain additional statistics level that is related to a topic. A list of statistics levels can be retrieved by making a HTTP GET request to address (not used in the poc service):

https://pxweb2.stat.fi/PxWeb/api/v1/{language}/{database-id}/{levels}

where {language} is the requested language, {database-id} is the identifier of the statistical database and {levels} is the identifier of a topic.

A list of available statistical tables in a statistics level can be retrieved by making a HTTP GET request to address (not used in the poc service):

https://pxweb2.stat.fi/PxWeb/api/v1/{language}/{database-id}/{levels}/{levels}

where {language} is the requested language, {database-id} is the identifier of the statistical database, the first {levels} is the identifier of the topic and the second {levels} is the identifier of the statistics level.

The metadata of a specific table under a statistics level can be retrieved by making a HTTP GET request to address (not used in the poc service):

https://pxweb2.stat.fi/PxWeb/api/v1/{language}/{database-id}/{levels}/{levels}/{table-id}

where {language} is the requested language, {database-id} is the identifier of the statistical database, the first {levels} is the identifier of the topic, the second {levels} is the identifier of the statistics level and {table-id} is the identifier of a statistical table.

More detailed instructions for the use of the Statistics Finland's PxWeb API can be found from the PxWeb API Help document [4].

2.2.2. SDMX API

The SDMX (Statistical Data and Metadata eXchange) is an initiative for standardizing the exchange of statistical data and metadata. The SDMX API is a web interface that can be used for querying the statistical data and metadata.

The statistical table metadata and data from the Eurostat's SDMX 2.1 API [5] were included to the GeoE3 data joining service. The statistical topic options were limited to 'Solar thermal collector's surface', 'Energy efficiency', ‘Population on 1st January’, ‘Population projections’, ‘Population density’ and ‘Share of zero emission vehicles in newly registered passenger cars’.

A list of all structural metadata available in the SDMX API of the Eurostat can be retrieved by making a HTTP GET query to the address:

https://ec.europa.eu/eurostat/api/dissemination/sdmx/2.1/{resource}/{agencyId}/all?detail={details}

where {resource} is one of the following: [dataflow, datastructure, codelist, conceptscheme], {agencyID} is one of the following: [ESTAT, COMP, EMPL, GROW] and {details} is one of the following: [full, allstubs, referencestubs].

Statistical data can be retrieved by making a HTTP GET query to the address:

https://ec.europa.eu/eurostat/api/dissemination/sdmx/2.1/{resource/{flowRef}/{key}/{parameters}

where the value of {resource} parameter is ‘data’, {flowRef} is the identifier of the dataflow, {key} is optional parameter for dimensional filtering, and {parameters} can contain the following parameters separated by ‘&’ character: {format} for requesting output in the specific format where supported values are: [SDMX_2.1_STRUCTURED, SDMX-CSV, JSON, TSV], {startPeriod] for filtering data on start time, {endPeriod} for filtering data on end time, firstNObserservations and lastNObservations for filtering data based on observations, {detail} for getting detailed response and {compression} for getting the results in compressed format.

More detailed instructions for the use of the Eurostat’s SDMX API can be found from the Eurostat’s SDMX API documentation page [6].

2.3. User Interface

This section contains instructions on how to use the user interface of the proof-of-concept application.

The use of the data joining service begins with the user selecting a statistical database from the user interface (Figure 1). When the ‘Eurostat’ option is selected, the queries are made to the Eurostat’s SDMX API and when any other options is selected, the queries are made to the Statistics Finland’s PxWeb API.

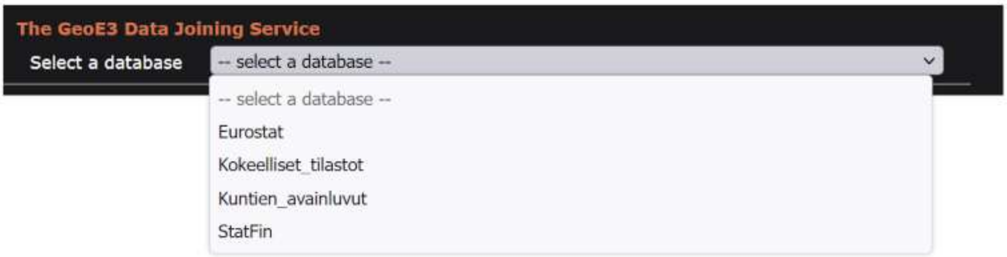
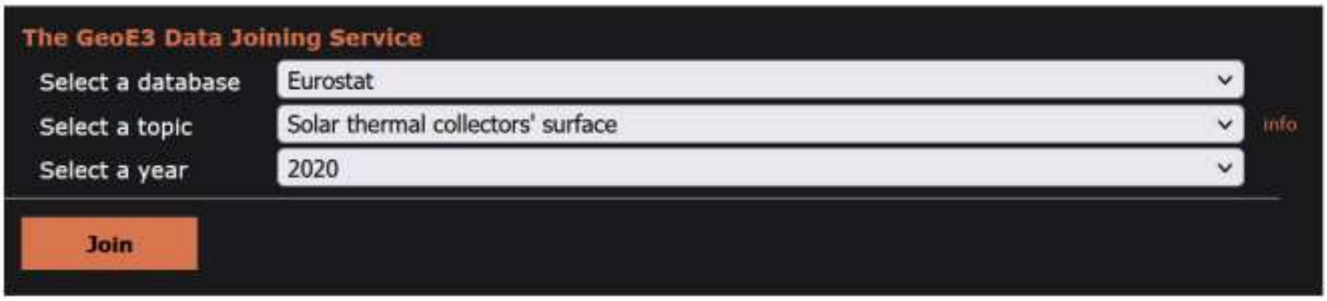


Figure 1: Selecting a database.

All metadata relating to the statistical tables are retrieved in the JSON format from both of the background APIs.

2.3.1. SDMX API

When the ‘Eurostat’ value is selected for the database, the process continues by selecting statistical topic and year from the user interface (Figure 2). The ‘info’ link contains links to the metadata descriptions when they are available from the Eurostat.



The GeoE3 Data Joining Service

Select a database Eurostat

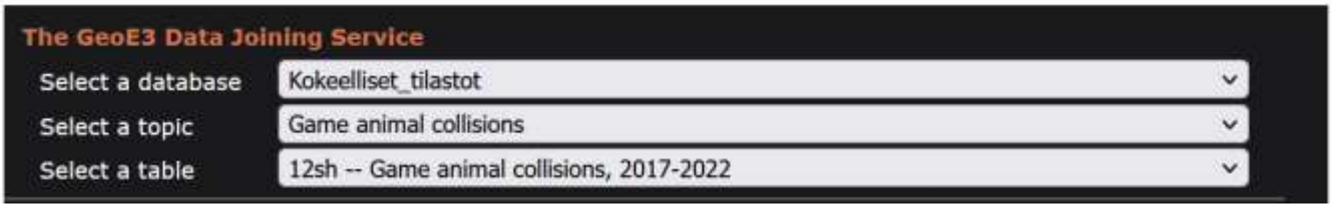
Select a topic Solar thermal collectors' surface

Select a year 2020

Figure 2: Selecting statistical topic and year from the SDMX API.

2.3.2. PxWeb API

When the user selects any other database value than ‘Eurostat’, the process continues by selecting a statistical topic and a statistical table (Figure 3).



The GeoE3 Data Joining Service

Select a database Kokeelliset_tilastot

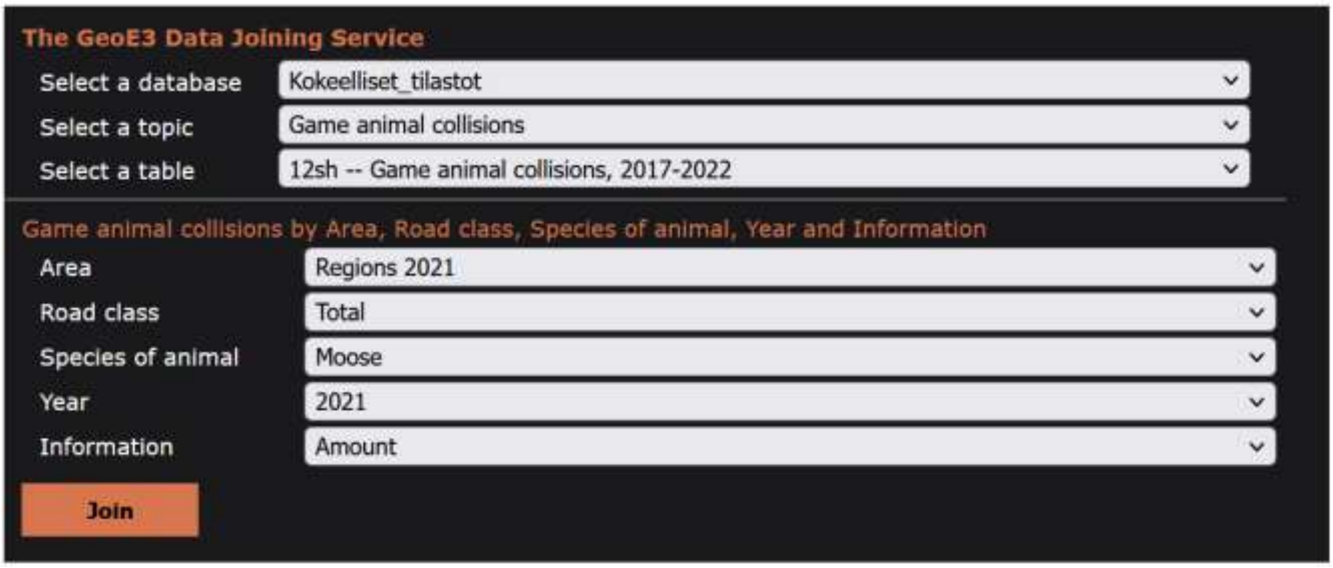
Select a topic Game animal collisions

Select a table 12sh -- Game animal collisions, 2017-2022

Figure 3: Selecting a statistical topic and a statistical table from the PxWeb API.

2.3.3. Selecting Statistical Table Variables

After the user has selected the statistical table and its metadata have been retrieved from the PxWeb API, the contents of the metadata are listed in the user interface (Figure 4). The metadata contains a title and a list of statistical variables with their selectable values.



The GeoE3 Data Joining Service

Select a database Kokeelliset_tilastot

Select a topic Game animal collisions

Select a table 12sh -- Game animal collisions, 2017-2022

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 4: Selecting statistical variable values from the PxWeb API.

The values of the statistical variables determine the statistical data that can be retrieved from the PxWeb API. A ‘map’ element was added to the area variable response of the statistical table metadata for indicating the yearly version of the area division, to which the statistical data corresponds to. The options for the area selection were limited to municipalities and regions that correspond to the geospatial datasets that are available in the service. The individual area codes that belong to those area divisions are grouped under these options, so that when a specific area division is selected, all area codes that belong to that group are selected.

The OGC API – Joins service requires that the values that are related to an individual geographical area unit, are presented in a single row in the CSV file. If multiple values are selected for other statistical variables that area code variable, the PxWeb API of Statistics Finland returns the values in multiple rows. Therefore, the value selection for other statistical variables was limited to a single value.

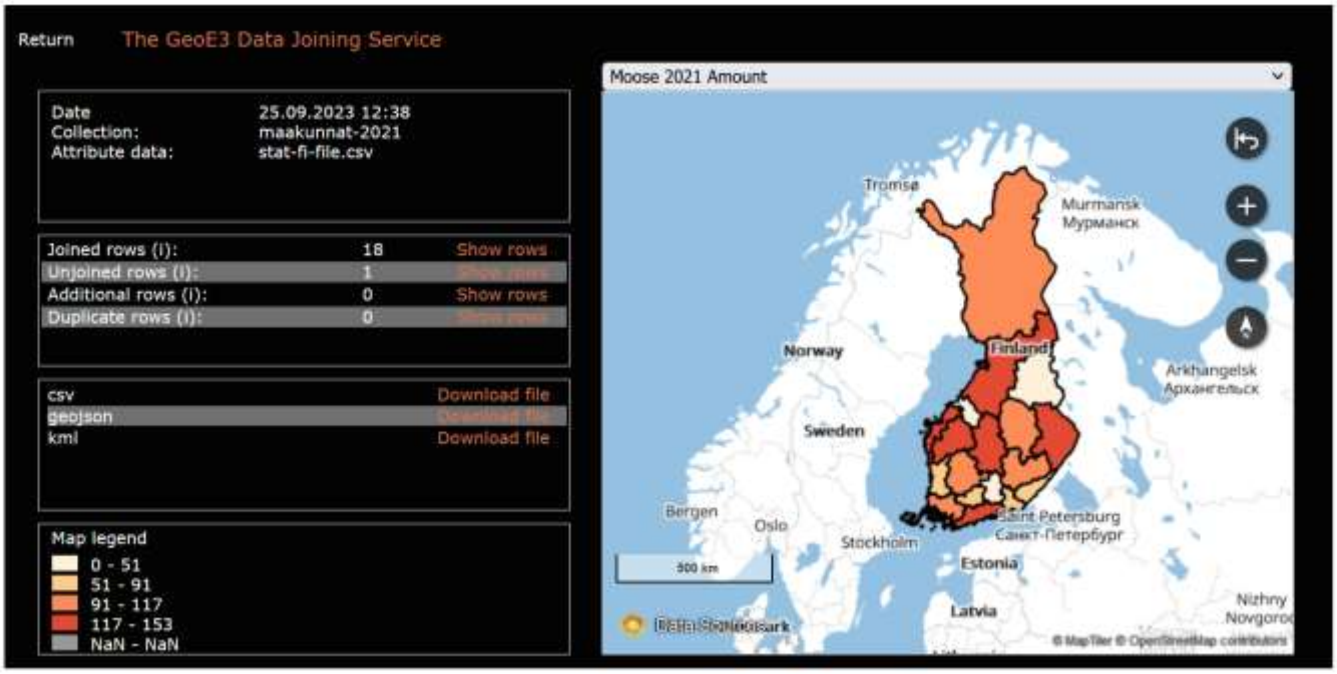
2.3.4. Data Joining (both APIs)

After the user has selected values for the statistical variables and clicked the 'Join'-button, the statistical data are retrieved from the background APIs in the CSV format. The PxWeb API request contains information on the selected values for each statistical variable.

In some cases, the key values of the returned statistical areas don't' match with the key values used in geospatial datasets and need to be processed to create matching key values before the data joining query can be executed.

After the potential editing of the key values, a data joining request is made to the OGC API - Joins service that joins the statistical data with the geospatial dataset. For the PxWeb API, the geospatial dataset used in the operation can be determined automatically, based on the selected area division and the value of the 'map' variable that indicates the yearly version of the dataset.

The response of the OGC API – Joins service is parsed in the user interface and the results page is created from the response's contents (Figure 5). The results page contains general information about the join, information on the successfulness of the join operation, map preview, download links for the joined dataset in various formats and a map legend. The map preview was created with the Oskari application's RPC functionality [7]. The map styling was created by classifying the data values into four classes.



Return The GeoE3 Data Joining Service

Date	25.09.2023 12:38
Collection:	maakunnat-2021
Attribute data:	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

■	0 - 51
■	51 - 91
■	91 - 117
■	117 - 153
■	NaN - NaN

Moose 2021 Amount

Figure 5: The Results Page of the User Interface.

2.4. Service Architecture

The architecture of the data joining service is presented in (Figure 6). The user interface is used for managing the entire data joining process. The Oskari application is used for producing the map preview functionality and it is maintained by the National Land Survey of Finland. The PxWeb API is used for requesting the statistical metadata and data and it is maintained by Statistics Finland. The SDMX API is also used for requesting the statistical metadata and data and it is maintained by the Eurostat.

The geospatial datasets are hosted in the PostgreSQL / PostGIS database. The metadata of the geospatial datasets are published via the OGC API – Joins Service. The joined data outputs are published from the database to the GeoServer application where they can be retrieved in several output formats.

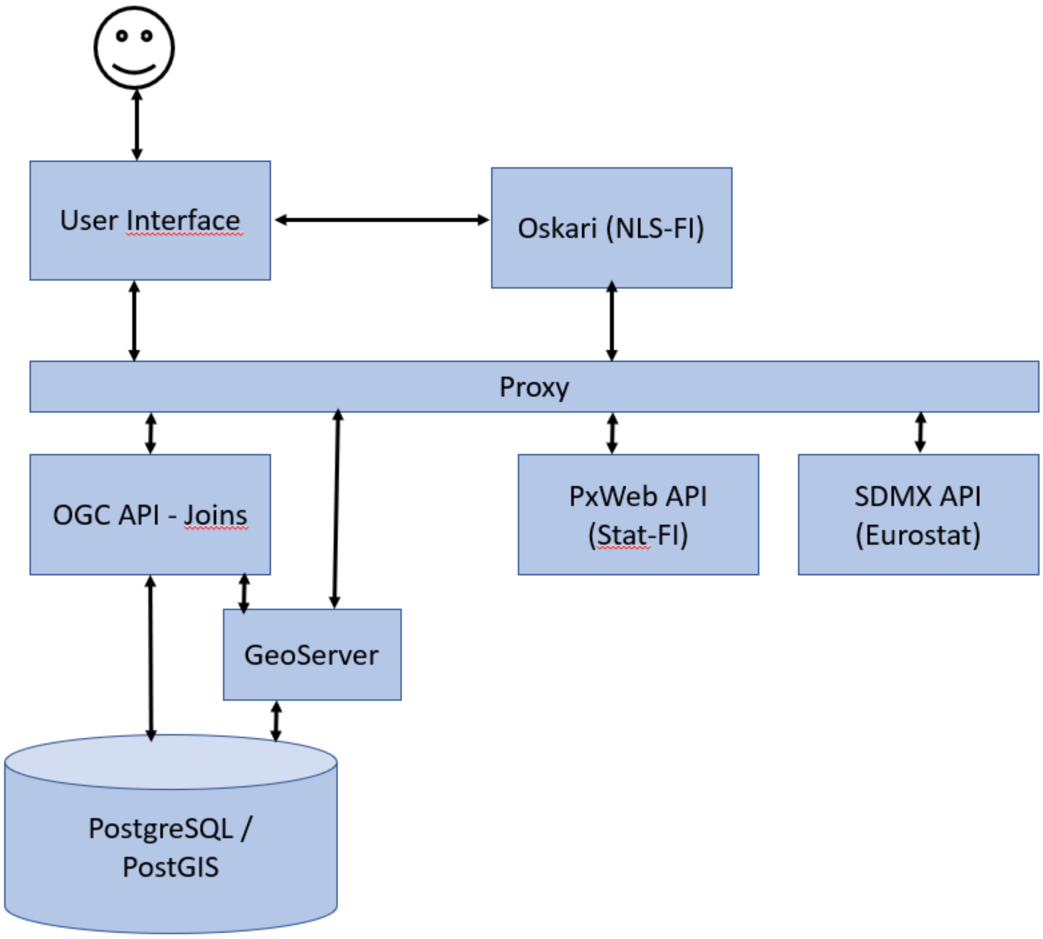


Figure 6: Architecture of the Data Joining Service.

3. Validation

User testing of the PoC service was organized at the end of the service development activity to collect feedback on the quality, usability and usefulness of the data joining service and to acquire ideas for further development in future projects.

A questionnaire for collecting user feedback was sent to stakeholders (outside the GeoE3 consortium) which could potentially benefit from utilizing the data joining service. The recipients of the testing call were Eurostat, European Forum for Geography and Statistics Steering Committee and statistical institutes of Portugal, Poland, Slovenia, Sweden, Norway, and Ireland. The questionnaire is presented in Appendix A. The call for user validation took place in June-July 2023. A total of six replies to the call were received.

All respondents agreed unanimously that the service met the backbone quality expectations presented in the questionnaire, see Appendix A.

As for the usefulness of the service most of the respondents found the service useful or potentially useful as this kind of generic API based service makes life easier and simplifies data sharing using a standard implementation. Access to more detailed levels of information in the service would be even more beneficial. One respondent did not find the data joining web service implementation useful at all.

Several respondents provided ideas for improving the usability and usefulness of the PoC service. For instance, allowing file upload of data in CSV or other format in addition to API access and ability to join more than one statistical variable onto a geography were the primary wishes for further development. Additionally, a number of more detailed suggestions for improving the functionalities and visual representation of the user interface were proposed. However, as stated in the original call for user testing, the user interface implementation was not the primary target of user validation. Instead, the data joining service concept was the main priority of the evaluation.

Finally, the respondents provided feedback on the naming convention of the service. Currently, the service is generally entitled as "data joining service" reflecting the joining operation of tabular content with geospatial features as the service backbone. Generally, the naming was evaluated to be quite descriptive and adequately easy to understand. However, some respondents criticized the lack of "geo" dimension in the name and the use of the term "joining" as not sufficiently intuitive for users.

4. Conclusions

Currently, the CSV output of the statistical APIs often require lots of manual editing before it can be utilized in the OGC API – Joins service. The OGC API – Joins service sets certain requirements for the CSV output and the statistical APIs should format the CSV output according to those requirements in order to minimize the extra effort.

The CSV files should also contain a header row that includes the attribute names in the columns. They should also contain only one row for each geographical feature and the attribute values should be stored in the columns. The geographical identifiers should be provided in the same formats that are used in the spatial datasets.

The metadata response should also contain information on the geographical area division, to which the data relates to. If the statistical data is related to a specific yearly version of geographical units, the metadata should contain also that information.

In summary, presently statistical APIs provide quite diversely formatted and variable content. Hence, a generic standard "plug'n'play" data joining service which could tap into statistical APIs without extra development effort, is still not quite feasible. Clearly, more standardization effort on the statistical APIs is required. In addition, interoperable persistent identifiers allowing for unambiguous linking of geographical features with statistical data need further standardization from data integration perspective.

As an outcome of the PoC development work and user feedback no particular further development ideas for the OGC API Joins draft standard emerged. However, in our opinion it is very important to finalise the standarisation process and approve the draft Joins standard as soon as possible to set a solid foundation for further data joining service implementations. The PoC implementation serves as one of the references implementations required for approving the standard.

As for further development of the present data joining service concept, some ideas as file upload of CSV data and joining several statistical variables onto a single geography were presented. In addition, an even more interesting concept to be tested could be a business intelligence dashboard-like functionality based on the present implementation. In a dashboard, statistical or more generally any attribute data, provided by an API, would be joined with corresponding geographical features such as buildings in a smart city case. The result of this data integration would be analysed and visualized in a user interface.

5. References

[1]: The URL of the data joining service: https://geoe3platform.eu/djs/

[2]: The draft for the OGC API - Joins standard in GitHub: https://github.com/opengeospatial/ogcapi-joins

[3]: The URL of the PxWeb API of Statistics Finland; https://pxweb2.stat.fi/PxWeb/api/v1/en/

[4]: Statistics Finland’s PxWeb API Help document: https://www.stat.fi/static/media/uploads/org_en/avoindata/px-web_api-help.pdf

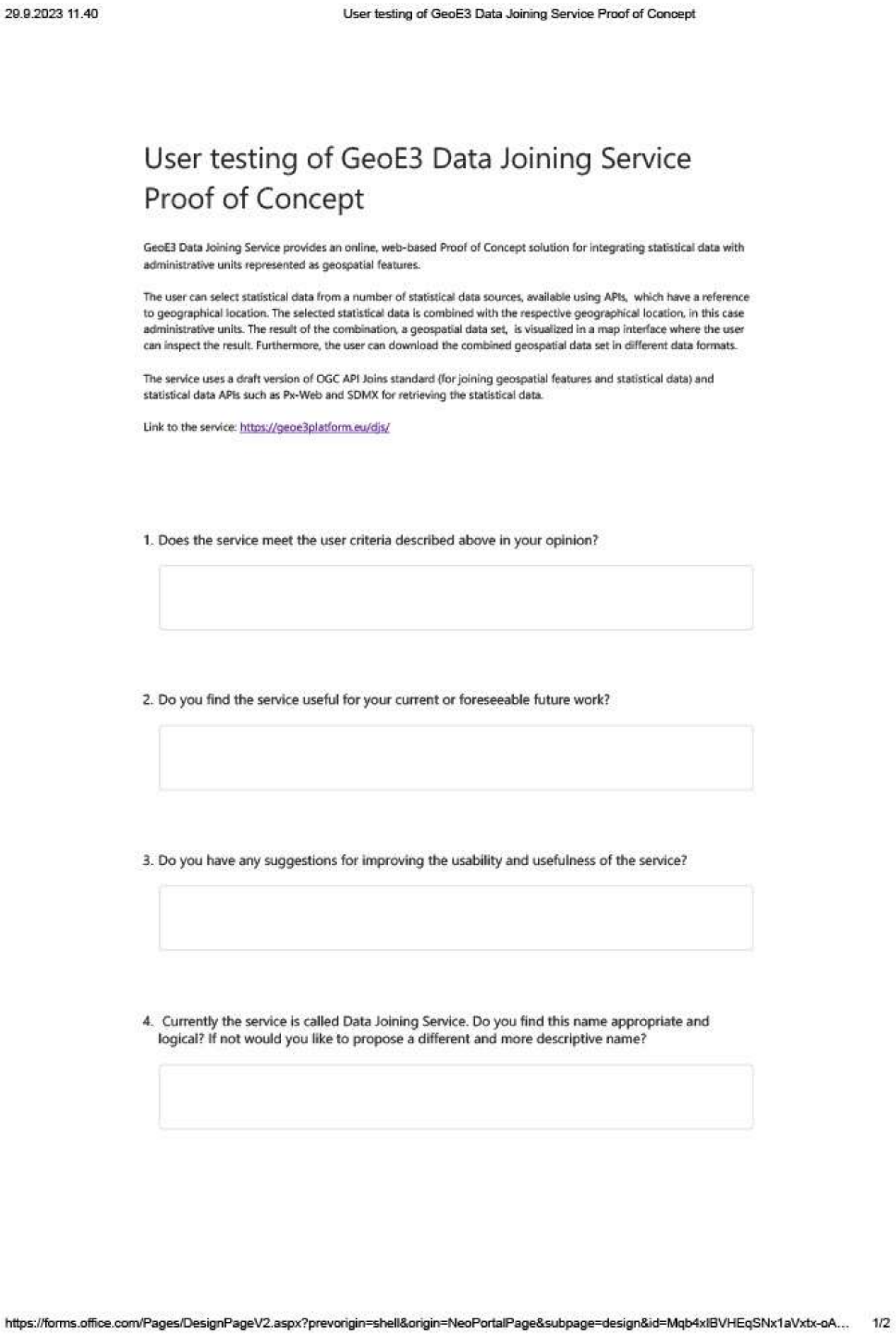
[5]: The SDMX 2.1 API of the Eurostat: https://ec.europa.eu/eurostat/api/dissemination/sdmx/2.1/dataflow/ESTAT/all?detail=allstubs&format=json

[6]: Eurostat API SDMX 2.1 documentation: https://wikis.ec.europa.eu/pages/viewpage.action?pageId=44165555

[7]: Oskari.org web page: https://oskari.org

6. Appendix A

Questionnaire for user feedback of the data joining service PoC.



29.9.2023 11.40 User testing of GeoE3 Data Joining Service Proof of Concept

User testing of GeoE3 Data Joining Service Proof of Concept

GeoE3 Data Joining Service provides an online, web-based Proof of Concept solution for integrating statistical data with administrative units represented as geospatial features.

The user can select statistical data from a number of statistical data sources, available using APIs, which have a reference to geographical location. The selected statistical data is combined with the respective geographical location, in this case administrative units. The result of the combination, a geospatial data set, is visualized in a map interface where the user can inspect the result. Furthermore, the user can download the combined geospatial data set in different data formats.

The service uses a draft version of OGC API Joins standard (for joining geospatial features and statistical data) and statistical data APIs such as Px-Web and SDMX for retrieving the statistical data.

Link to the service: https://geoe3platform.eu/djs/

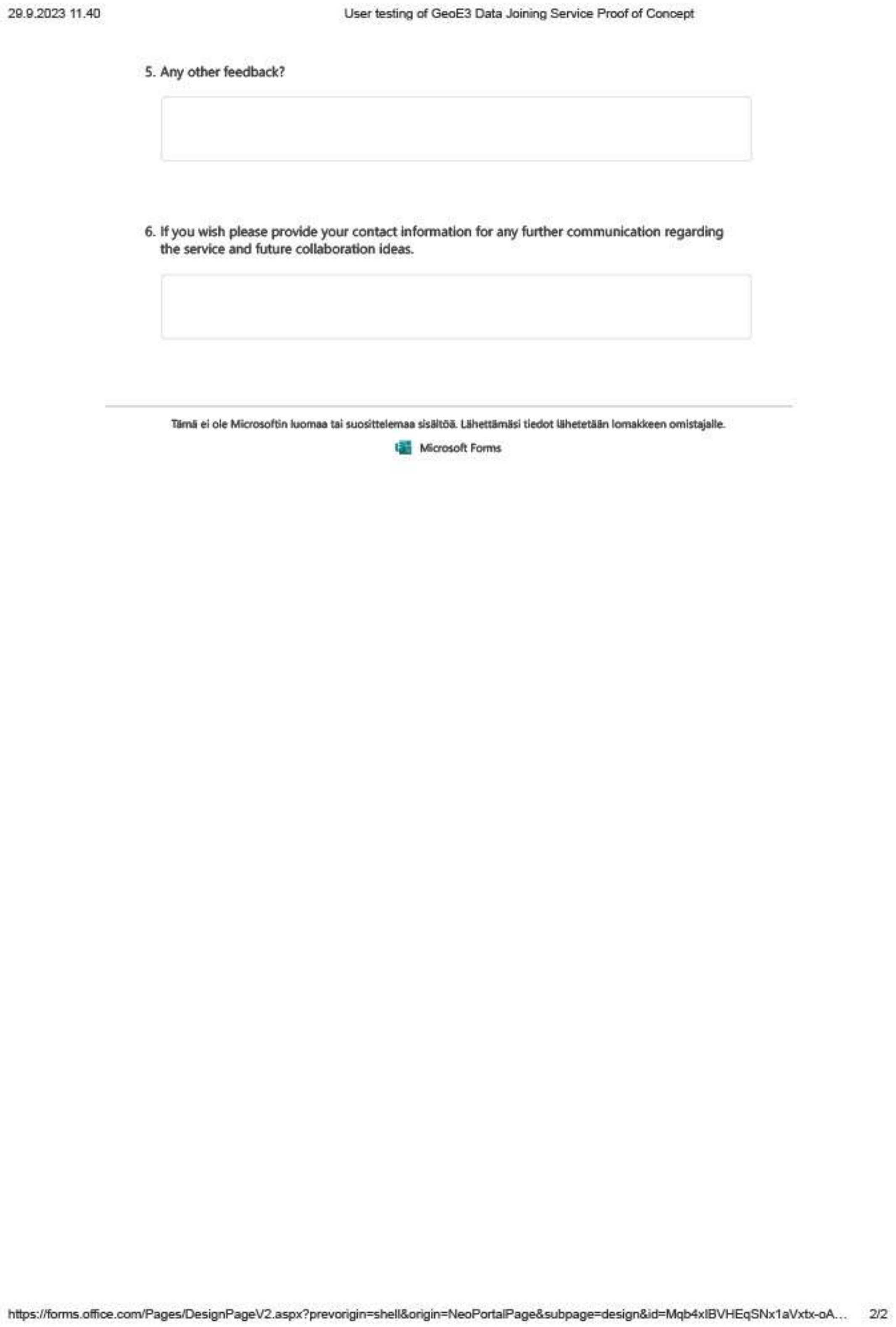
1. Does the service meet the user criteria described above in your opinion?

2. Do you find the service useful for your current or foreseeable future work?

3. Do you have any suggestions for improving the usability and usefulness of the service?

4. Currently the service is called Data Joining Service. Do you find this name appropriate and logical? If not would you like to propose a different and more descriptive name?

https://forms.office.com/Pages/DesignPageV2.aspx?prevorigin=shell&origin=NeoPortalPage&subpage=design&id=Mqb4xIBVHEqSNx1aVxtx-oA... 1/2



29.9.2023 11.40 User testing of GeoE3 Data Joining Service Proof of Concept

5. Any other feedback?

6. If you wish please provide your contact information for any further communication regarding the service and future collaboration ideas.

Tämä ei ole Microsoftin luomaa tai suosittelemaa sisältöä. Lähettämäsi tiedot lähetetään lomakkeen omistajalle.

Microsoft Forms

https://forms.office.com/Pages/DesignPageV2.aspx?prevorigin=shell&origin=NeoPortalPage&subpage=design&id=Mqb4xIBVHEqSNx1aVxtx-oA... 2/2