

GeoE3_A4_1_2 Report on technical aspects for including tabular data in the Geospatially Enabled Ecosystem



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

Publishing Information/Date



CONFIDENTIAL DRAFT – NOT FOR FURTHER CIRCULATION

f

Revision History

Version	Status	Author	Modifications
0.2	Updated	Timo Aarnio	Updated version
0.1	Updated	Timo Aarnio, Pekka Latvala	Initial version

Distribution

Version	Status	Recipient
1.0	Final	Hadea

Table of Contents

Revision History.....2

Distribution2

Table of Contents.....3

1. Introduction4

1.1. OGC API Joins.....4

2. Technical requirements for statistical data interfaces4

2.1. Hyperlinks5

1. Introduction

Tabular (or, statistical) and spatial data are mostly distributed via various data APIs and this document aims describe how statistical data can be integrated to the Geospatially Enabled Ecosystem for Europe (GeoE3): which formats and APIs can be used and how.

1.1. OGC API Joins

The standardization organisation Open Geospatial Consortium (OGC) has previously released a standard called Table Joining Service (TJS). The standard describes an API to combine tabular and spatial data. Like most OGC standards released in the previous decade or earlier, TJS utilizes XML-based formats for data interchange.

Some years ago, OGC started an effort to modernize its specifications and to make them more inviting for web developers at large. As part of this effort the OGC API -family of specifications was introduced. The OGC API -family specifications rely on the use of JSON-based data interchange instead of using XML and the specifications are very stripped-down trying to hide the various complexities of spatial data while still being usable by application developers.

The original Table Joining Service standard had some shortcomings that needed to be addressed and National Land Survey of Finland took on the task of revisioning the standard by entering the most critical Change Requests to the OGC. At the same time the idea of OGC API -family was cooking and in June 2021 it was decided to change from the revision work to specifying a new standard altogether. This new specification is called OGC API Joins and will be used in GeoE3 for achieving data integration.

2. Technical requirements for statistical data interfaces

Statistical data interfaces are the APIs through which statistical data or metadata can be acquired over the Internet. Statistical data behind the interfaces can be multi-dimensional or readily processed into fewer dimensions for application use. In the use-case of integrating statistical data with spatial data the data must match the geographical features i.e., the statistical values must match directly to the spatial features or some form of aggregation must take place. In the context of GeoE3 we will look solely at use-cases where the data can be directly matched without performing an additional step of aggregation.

The draft OGC API Joins will use comma separated values (CSV) for tabular (statistical) data, other formats may be used but will not be mandatory in the core specification. In GeoE3 the statistical data interfaces must therefore be able to output statistical data in CSV-format. It might be possible to implement adapters that are capable of transforming data in a different format to CSV. Such an adapter would read in data in one format and then output it in a suitable CSV format for use with OGC API Joins.

As for metadata, the core of OGC API Joins will support only a title-field describing the tabular data and if other needs for metadata are recognized they can be handled by extensions to the specification. As a join between spatial and statistical data is done a report of the join operation can be generated. The report will outline how many spatial units matched with the statistical units and vice-versa. Also, the number of unmatched units (rows) is included.

To carry out the join operation matching identifiers are needed in both datasets. The core will be able to handle alphanumeric identifiers and more advanced functionality can be offered again via the development of extensions

to the specification. One recognized possible need is to offer mappings between different identifiers (i.e., the spatial and statistical units are the same, but identifiers used are different), but this will not be in the scope of GeoE3.

The aim of this document is not to describe the technical aspects on spatial data but in short, the needs are like those for tabular data. Matching identifiers are of course needed and as the OGC API Joins will expose its resources in a similar way as other APIs in the same family, it will provide the following metadata for spatial datasets:

- attribution / source
- description
- spatial and temporal extent
- collection identifier

2.1. Identifiers in spatial and statistical data

To carry out data joining using OGC API Joins, matching identifiers need to be present in both spatial and statistical data. The most optimal identifiers are **global**, **persistent**, and **unique** so that there can be no question as to what entity they refer to. The OGC API Joins specification itself does not define any requirements for the identifiers, but they must match between the used spatial and statistical data. For example, the identifiers defined by the INSPIRE directive (INSPIRE ID's) fulfill all the criteria listed above, but they are not commonplace in spatial or statistical data.

When identifiers with such properties are not available, we must make sure that they are at least unique in the used context (i.e., do not collide with identifiers in other kind of data used).

2.2. Hyperlinks

The current draft for OGC API Joins can be found from GitHub (it is likely to update several times during the project): <https://github.com/opengeospatial/tjs/tree/master/docs/drafts>