SUPPORTING WIDE USER-BASE IN RASTER ANALYSIS – GEOCUBES FINLAND

L. Lehto *, J. Kähkönen, J. Oksanen, T. Sarjakoski

Finnish Geospatial Research Institute in the National Land Survey of Finland, Geodeetinrinne 2, FI-02430 Masala, Finland -
(lassi.lehto, jaakko.kahkonen, juha.oksanen, tapani.sarjakoski)@nls.fi

Commission IV, WG IV/7

KEY WORDS: Geodata harmonization, Raster analysis, Multi-resolution, Cloud storage, GeoTIFF

ABSTRACT:

An initiative has been started in Finland to build a harmonized raster geodata repository, called GeoCubes Finland, to help academic users in carrying out geospatial analysis operations in research projects of diverse disciplines. The aim of the initiative is to expand the adoption of GI methods and tools to new sectors of academic research. GeoCubes Finland is being built in the context of a large-scale program developing national geodata infrastructure for academic studies. The main building blocks of the GeoCubes Finland initiative are: a harmonized, multi-resolution raster geodata storage, a set of standardized mechanisms for accessing the raster contents of the storage, and support facilities for cloud-based geocomputing to help users carry out spatial analysis with the contained datasets.

1. INTRODUCTION

National Spatial Data Infrastructures (NSDIs) have widely reached a rather mature level of development (Hvingel et al., 2014). In various NSDIs an ample selection of high quality geospatial data sets has been produced and is readily available from standardized content access services. However, real practical use of those data sets is still limited, particularly in some tightly focused sectors, like academic research. Treatment of issues in geospatial context could yield valuable insight in many research problems. It is therefore essential that geospatial community is willing and able to lower the bar for outsiders for making use of these valuable geodata collections.

Reasons for this situation are manifold. Geospatial data sets cannot be easily integrated with other research data, because of their peculiar geometry data types (de Man, 2007). In most cases processing of geospatial data sets require experience with specialized software and understanding of awkward algorithms. Organizations have varying capabilities in dealing with these issues (Mäkelä, 2012). Furthermore, the level of harmonization across data sets produced by different national data providers is still far from being perfect (Tegtmeier et al., 2008).

One feasible approach for improving the situation would be to build a specialized SDI, targeted for a particular user sector, taking their specific needs into account. This kind of SDI could be built on top of the existing NSDI services and data sets. If the sector-specific SDI is targeted for users having little experience with geospatial data sets, it would be recommendable to start with a simple data format, raster data. To ease the processing of geodata on general purpose computers, it would be advisable to develop the solution on a cloud computing and Web browser - based architecture.

An initiative has been started in Finland to build a harmonized raster geodata repository, called GeoCubes Finland, to help academic users in carrying out geospatial analysis in research projects of diverse disciplines (Lehto et al., 2018). This work has been initiated in the framework of general research infrastructure development that is guided by the Roadmap for Finnish Research Infrastructures (FIRI) (Academy of Finland, 2018). This national infrastructure development plan is maintained by the national research fund, Academy of Finland.

The aim of the GeoCubes Finland initiative is to expand adoption of GI methods and tools to new sectors of academic research. GeoCubes Finland is being built as a part of a large-scale FIRI program developing a national geodata infrastructure to support academic studies. The program is called Open Geospatial Information Infrastructure for Research (oGIIR, 2018).

The three main building blocks of the GeoCubes Finland initiative are: a) setting up of a harmonized, multi-resolution raster geodata storage, b) provision of a set of standardized mechanisms for accessing the raster contents of the storage, and c) developing support facilities for cloud-based geocomputing to help users carry out spatial analysis with the contained datasets.

The rest of the paper is structured as follows. In Section 2 the harmonized raster data storage of the GeoCubes Finland initiative is described. The available data access mechanisms are detailed in Section 3. Section 4 continues to discuss the cloud computing facilities developed for processing of the available geodata resources. Section 5 deals with the implementation issues encountered in the GeoCubes Finland development and Section 6 concludes the paper, together with some outlook for further work.

2. GEOCUBES FINLAND DATA STORAGE

2.1 Data sets

The raster data storage of GeoCubes Finland can be seen as a harmonized cache of national datasets provided by various governmental institutions. The first datasets to be added to the GeoCubes platform include high-resolution Digital Elevation Model (DEM) provided by National Land Survey of Finland, CORINE Land Cover (EEA, 1995) datasets provided by Finnish

* Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-4-329-2018 | © Authors 2018. CC BY 4.0 License.

Environment Institute, Superficial Deposits dataset of Geological Survey of Finland, and National Forest Inventory dataset of National Resources Institute Finland. As auxiliary layers of content topographic base maps, ortho imagery and administrative units of the country are also available. The data sets include thus both data sets with discrete classification values, like land cover or surface deposits maps, and data sets with continuous value ranges, like DEMs or ortophotos. All data layers included into the GeoCubes Finland data storage are open data and thus available without restrictions and free of charge.

2.2 Content harmonization

The GeoCubes data storage is standardized in the sense of georeferencing, the set of spatial resolutions available, spatial subdivision of the country's area, and data encoding mechanisms. The national ETRS-based grid, ETRS89/TM35FIN (EPSG:3067), is used as the basis of georeferencing, with fixed point of grid origin. By using a detailly agreed grid, all data sets can be accurately processed together with exact cell-level correspondence.

Depending on the accuracy of the original dataset, GeoCubes contains up to ten different resolution levels in the range from 1 m to 1000 m. The selected resolutions are [m]: 1, 2, 5, 10, 20, 50, 100, 200, 500 and 1000. Thus, the series of resolutions applied does not follow the traditional image pyramid approach with resolutions in two's exponents, but rather offers resolution levels that are typically used in statistical analysis.

The area of the country is divided into 100 km * 100 km blocks that are stored as individual GeoTIFF files, with internal overviews storing the generalized resolution levels (GeoTIFF, 2018). The size of the blocks is selected appropriately for achieving reasonable file sizes, even in the case of the most detailed 1 m resolution data sets. The territory of Finland can be covered with 62 such blocks (see Figure 1). The block structure also facilitates efficient parallelization of the computations related for instance to the ingestion of data sets into the GeoCubes data storage, extraction of subareas from the storage, and raster operations in various analysis tasks. The recently introduced cloud optimized form of GeoTIFF imagery (COG) is used for efficient extraction of subareas from individual image files (COG, 2018).

2.3 Ingestion processing

The datasets are ingested into the GeoCubes Finland data storage by automated procedures, developed in Python with help of routines provided by the open source software package Geospatial Data Abstraction Library (GDAL, 2018). Data sets available in raster form are transformed into the standardized grid and generalized to all applicable resolution levels. Vector-formatted source data sets are first rasterized with selected attribute value and then added to the raster storage. Needed generalization processes are carried out to fill in the required resolution levels. If source data sets are available in generalized forms, those data sets are used as input data for ingestion. In case of data sets with time series, classification and source resolution transformations have also been carried out. After these preprocessing steps the data sets are easy to take into use and combine in various analysis scenarios, thus revealing end users from the tedious data preparation work.

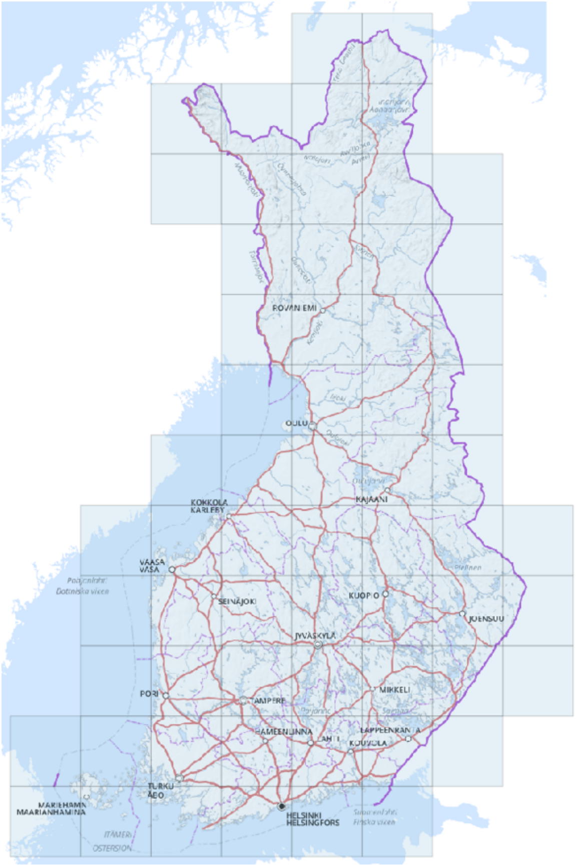


Figure 1. The block structure used in GeoCubes Finland

3. STANDARDIZED ACCESS

The contents of GeoCubes Finland are made available in various standardized ways (see Figure 2). Complete block-wise GeoTIFF files can be readily downloaded by their published URL. Subareas of the files can be efficiently accessed via HTTP GET Range operation, making use of the cloud optimized structure of the file. A complete country-wide data layer can be conveniently treated as a single raster data set by applying the so-called GDAL virtual raster (VRT) mechanism. The end user can be simply provided with a single textual VRT file that refers to the whole set of 62 block files on the server. This file can then be opened for instance in QGIS application, or any other GDAL VRT supporting client software, instantly getting access to the complete data layer.

The Open Geospatial Consortium (OGC) -standardized Web Coverage Service (WCS) raster data interface is also provided as an access mechanism to the GeoCubes Finland data storage (OGC, 2012). The WCS interface enables selection of spatial subsets from the storage, based on both ground and raster coordinates. The uniformly applied cell-level georeferencing ensures that all data layers can be consistently accessed using nationwide column and row values. Source data sets for the WCS server can be easily configured by using VRT-based sources. MapServer is currently used as the WCS engine for the GeoCubes Finland platform (Open Source Geospatial Foundation, 2018).

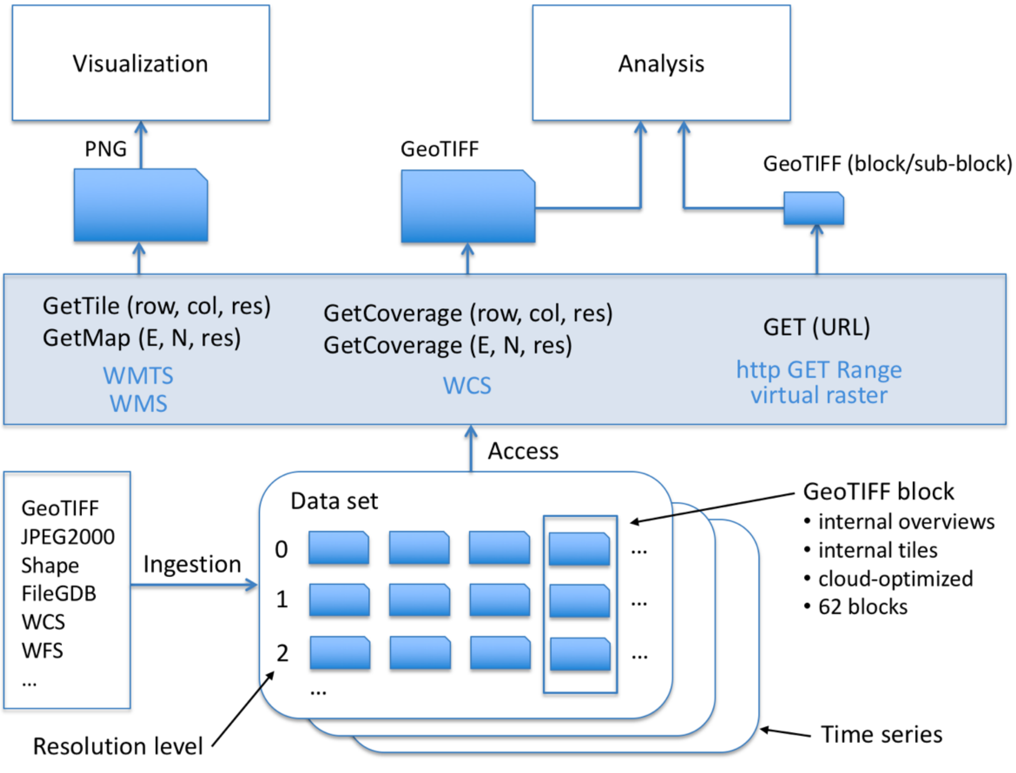


Figure 2. GeoCubes Finland data storage architecture

Although raster data processing is the main use case for the GeoCubes Finland, also visual representation of its contents is important. For visualization purposes an OGC Web Map Service (WMS) instance is available, supported by a Web Map Tile Service (WMTS) cache for efficient rendering of GeoCubes layers. The WMS service of GeoCubes Finland is realized with MapServer and the WMTS service with MapProxy (Omniscale, 2018).

4. CLOUD COMPUTING FACILITIES

4.1 Computing platform

GeoCubes Finland is implemented as a cloud service with tight connection to High Performance Computing (HPC) environment. This facilitates efficient processing of spatial analysis operations involving several GeoCubes data layers. The cloud computing platform used in GeoCubes Finland is provided by CSC – IT Center for Science (the provider of high-performance computing facilities for Finnish universities). At the moment the GeoCubes services are running on a CSC computing node with 6 virtual cores and 16 GB of RAM memory. The data storage is maintained on a 10 TB cloud volume.

4.2 Analysis in multiple resolutions

One of the important principles of the GeoCubes Finland is the provision of data in multiple resolutions. One might justifiably argue that it is always desirable to use the most detailed resolution available for best analysis results. However, there are exceptions. When the analysis result is shown as visual map representation, the analysis can be run on the resolution level that corresponds to the visualization scale. In this case the number of pixels involved, and the number of data cells needed for resolving the values of those pixels, remains constant over various zooming levels.

First tests performed on the GeoCubes Finland data sets and the cloud-based computing platform indicate that visualization-oriented analysis processes can be run in a nearly constant time over a range of scales of the visualized analysis results. For example, while performing change detection analysis between two editions of the CORINE land cover data set, processing time only doubled, when the scale, resolution level and, consequently, the number of data blocks involved in the computing was increased ten-fold.

Certain phenomena can also arguably be best analyzed on the level of detail, and thus on the data set resolution level, that corresponds to the spatial frequency that the phenomenon occurs in reality. In these cases, the multi-resolution nature of the GeoCubes platform would serve the user in an optimal way.

4.3 Processing in cloud computing environment

A cloud-based service platform and an OpenLayers-based Web browser client application was developed to demonstrate the GeoCubes content and the analysis possibilities it provides (OpenLayers, 2018). The service platform is based on the Django framework (Django Software Foundation, 2018) and developed making use of the GDAL Python API. The service platform is located on the same host with the data storage, thus making data access operations as efficient as possible.

The Web browser-based client application is designed to demonstrate advantages that can be achieved, when data storage and geocomputing facilities are tightly integrated. The GeoCubes service platform follows RESTful communication principles and is supported by a database containing basic metadata of the data sets served. A critical piece of metadata in case of raster data sets is the list of classification code values used and their corresponding textual labels. In the following is an example of a request to get the text label for a classification value of a data set (‘maapera’ = ‘superficial deposits’) on a given location. This request is used for displaying a tooltip containing the label, when user explores the data layer on the map view.

```
/legend/maapera/200/298100,6741200
```

operation | data set | resolution | coordinates

Another example is a request for calculating the change between two editions of the CORINE data set on a given land cover class (‘Pellot’ = ‘fields’). The computing area is limited by a BBOX (the viewport on the client application). The result of the analysis is a new map layer, returned as a PNG image, showing the changes detected between the selected editions of the data set.

```
/analyze/changedetect/200/corine/2000,2012/
Pellot/bbox/225700,6660000,494300,6740000
```

operation | method | resolution | data set | editions | class | spatial extent (by BBOX)

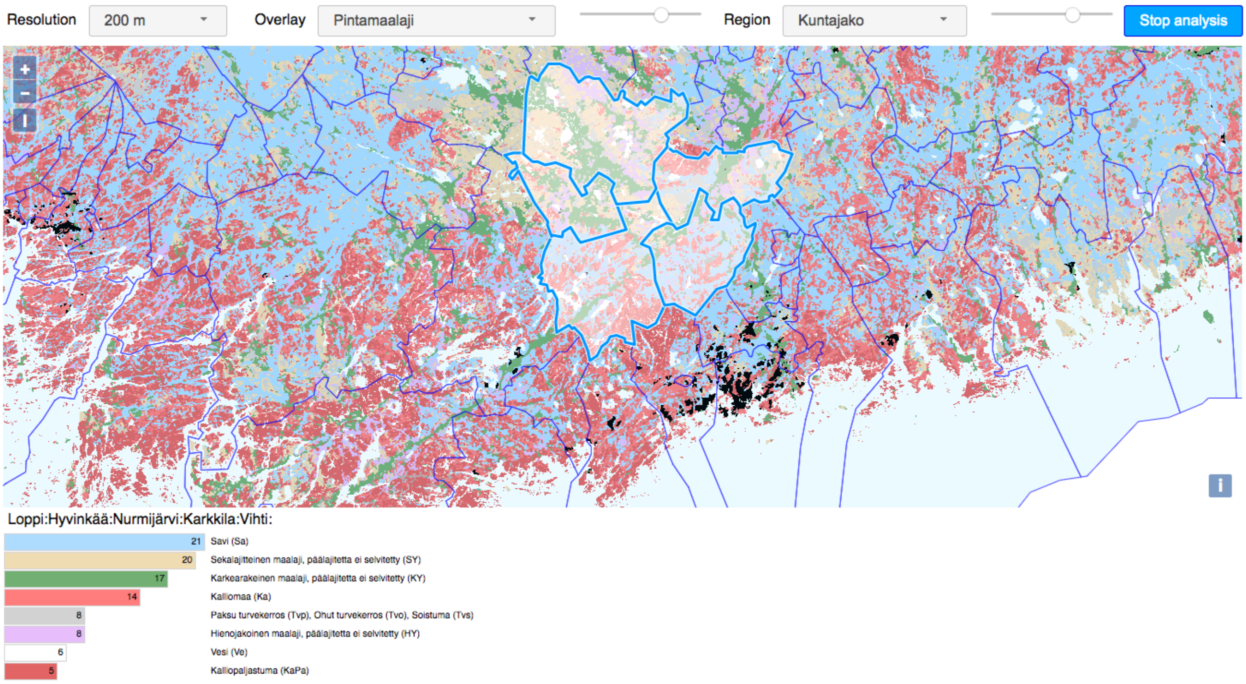
The third example shows a request for analysing the distribution of different classification values (class of superficial deposits) inside the given administrative area (‘kuntajako’ = ‘municipalities’), indicated by a list of area codes.

```
/analyze/distribution/100/maapera/kuntajako
/433,106,543,224,927
```

operation | method | resolution | data set | admin level | admin area codes

The result of the analysis is a list of classification code/percentage value pairs. These values are visualised in the client application using the D3.js library (Bostock, 2018). The target region is selected using vector-formatted administrative units requested from the GeoCubes auxiliary data storage. The GeoCubes Web client with the bar chart resulting from the analysis is shown in Figure 3.

It is worth to note that in all requests the resolution level is indicated as one of the parameters. This gives the user full control for selecting, on which level of detail the task in question should be resolved. Even the classification value of a data set can change depending on the resolution level, as the classification systems in many cases are hierarchical.



Resolution 200 m

Overlay Pintamaalaji

Region Kuntajako

Stop analysis
+
−
i
i
Loppi:Hyvinkää:Nurmijärvi:Karkkila:Vihti:
21 Savi (Sa)
20 Sekalajitteinen maalaji, päälajitetta ei selvitetty (SY)
17 Karkearakeinen maalaji, päälajitetta ei selvitetty (KY)
14 Kalliomaa (Ka)
8 Paksu turvekerros (Tvp), Ohut turvekerros (Tvo), Soistuma (Tvs)
8 Hienojakoinen maalaji, päälajitetta ei selvitetty (HY)
6 Vesi (Ve)
5 Kalliopaljastuma (KaPa)

Figure 3. The GeoCubes Finland Web client with analysis results (distribution of classes of superficial deposits inside the selected administrative areas)

4.4 QGIS as a client platform

In addition to the developed OpenLayers-based Web client also QGIS application has been used as the client application in accessing the GeoCubes data sets and for performing analysis processes on them (QGIS, 2018). The GeoCubes data sets can easily be handled in QGIS by the VRT mechanism. The local VRT file is just configured to refer through HTTP connection to the GeoTIFF files on the GeoCubes server. Use of the HTTP GET Range queries speeds up the content requests considerably. Analysis processes can be run in similar way as with the GeoCubes service platform, as GDAL functions are also available in QGIS application.

5. IMPLEMENTATION ISSUES

At the moment the multiple generalized resolution levels of the GeoCubes Finland raster storage are maintained as internal GeoTIFF overviews. The arrangement makes it easy to manage the files, as each block can be represented as a single file. The mechanism is GDAL-specific though, and poorly supported in other software. In the further development this issue will be considered. Most probably both internal and external overviews will be supported in future releases of the storage.

GDAL's 'BuildOverviews' method is applied in the data set generalization process. 'BuildOverviews' is a flexible processing tool and supports many resampling methods. However, one specific restriction for the selection of resolution levels is that 'BuildOverviews' only supports integer-valued factors between the base level and the computed overview levels. Thus, base level resolution of a data set cannot be 2m, 20 m or 200 m, as the first level generalized overview would then require use of a non-integer factor.

A fundamental issue in overview creation is that the generalization process tends to be application-dependent. The resampling method to be selected ultimately depends on the planned use of the data set. In many cases the generalization process also seems to require expert level understanding of the data set being generalized. In the case of GeoCubes Finland, the generalization method generally deemed as most appropriate is selected, honouring the opinion of the original data provider in the process. For specific use cases the base level resolution is always available and can be generalized at will in the end user's processing environment.

The generalization process applied in the GeoCubes ingestion mechanism also uses coarser classification system for generalized resolutions, if available. Class integration procedures had to be implemented as part of the data ingestion process for that end. A specific problem encountered while designing the data ingestion methods is that consistent classification mappings between data set editions could not always be found.

A particular practical problem for Web client-based visualization is that some of the data sets require 16-bit cell values that cannot easily be translated to standard Web image formats.

6. CONCLUSIONS

In the paper the initial experiences gained while building the GeoCubes Finland data storage, and running first tests on it, have been detailed. The storage is aimed at helping users in the academic sector to make use of spatial analysis in raster domain. The harmonized, cloud service-based raster data storage, together with the attached geocomputing facilities, makes it simple to take the spatial dimension into account in a research project.

In particular, an approach for accessing spatial analysis functions in a cloud-based environment has been discussed. A RESTful access API has been designed for this purpose. The multiresolution nature of the GeoCubes data storage enables visualization-oriented analysis operations to be run in nearly constant time over the whole available scale range. This way a visual map image, which actually represents the result of an analysis, can be configured as an ordinary map layer in the client application. This opens wholly new prospects for cloud-based analysis and processing of spatial data.

The spatial subdivision in blocks of the GeoCubes raster content naturally facilitates parallel processing of analysis functions. In the development of the GeoCubes data storage only standard Python subprocess mechanisms have so far been used. The future work on the GeoCubes Finland data storage will consider better integration with the HPC facilities of the CSC's computing platform. These include technologies like Apache Spark (Apache, 2018) and GeoTrellis (Eclipse Foundation, 2018).

Generally, the GeoCubes Finland platform will be continuously developed by adding more content data sets and by further improving the level of data harmonization. External GeoTIFF overviews will be taken into use and content download mechanisms further improved. First promotional user workshops will be organized in near future.

ACKNOWLEDGEMENTS

The work described in this paper has been carried out in the context of the project 'Open Geospatial Information Infrastructure for Research' (oGIIR, urn:nbn:fi:research-infras-2016072513), a part of Finland's Roadmap for Research Infrastructures. The project is funded by the Academy of Finland, grant number 306536. The computing infrastructure used in the work is funded by the Academy of Finland through the development program 'Finnish Grid and Cloud Infrastructure', urn:nbn:fi:research-infras-2016072533, grant number 283818.

REFERENCES

Academy of Finland, “Finland’s Strategy and Roadmap for Research Infrastructures 2014-2020, Interim review report 2018”. http://www.aka.fi/globalassets/tiedostot/aka_infra_tiekartta_raportti_en_030518.pdf (9 July 2018)

Apache, 2018. Spark Home Page. https://spark.apache.org (12 July 2018)

Bostock M., 2018. Data-Driven Documents, D3.js Home Page. https://d3js.org (12 July 2018)

COG, 2018. Cloud Optimized GeoTIFF Home Page. http://www.cogeo.org (12 July 2018)

de Man, W. H. E., 2007. Are Spatial Data Infrastructures Special? In: Onsrud, H. J. (Ed.), *Research and Theory in Advancing Spatial Data Infrastructure Concepts*, ESRI Press, Redlands, USA, pp. 33-54.

Django Software Foundation, 2018. Django Home Page. https://www.djangoproject.com (13 July 2018)

Eclipse Foundation, 2018. GeoTrellis Home Page. https://geotrellis.io (12 July 2018)

EEA, 1995. CORINE land cover. https://www.eea.europa.eu/publications/COR0-landcover/at_download/file (13 July 2018)

GDAL, 2018. Geospatial Data Abstraction Library. http://gdal.org (11 July 2018)

GeoTIFF, GeoTIFF home page. http://trac.osgeo.org/geotiff/ (11 July 2018)

Hvingel, L., Baaner, L. and Schrøder, L., 2014. Mature e-Government based on spatial data - legal implications. In: *International Journal of Spatial Data Infrastructures Research*, 2014, Vol.9, pp. 131-149.

Lehto, L., Kähkönen, J., Oksanen, J. and Sarjakoski, T., 2018. GeoCubes Finland - A Unified Approach for Managing Multi-resolution Raster Geodata in a National Geospatial Research Infrastructure. In: *GEOProcessing 2018, the Tenth International Conference on Advanced Geographic Information Systems, Applications and Services*, March 25-29, 2018, Rome, Italy. ISBN: 978-1-61208-617-0.

Open Source Geospatial Foundation, 2018. MapServer Home Page. http://mapserver.org (13 July 2018)

Mäkelä, J., 2012. Model for Assessing GIS Maturity of an Organization. In: A. Rajabifard and D. Coleman, (Eds.), *Spatially Enabling Government, Industry and Citizens: Research and Development Perspectives*, GSDI Association Press, Needham, MA, pp. 143-165.

OGC, 2012. Web Coverage Service. http://www.opengeospatial.org /standards/wcs (11 July 2018)

oGIIR, Open Geospatial Information Infrastructure for Research Home Page, http://ogiir.fi (9 July 2018)

Omniscale, 2018. MapProxy Home Page. https://mapproxy.org (13 July 2018)

OpenLayers, 2018. OpenLayers Home Page. https://openlayers.org (11 July 2018)

QGIS, 2018. QGIS Home Page, https://qgis.org/en/site/ (11 July 2018)

Tegtmeier, W., Hack, R., Zlatanova S. and van Oosterom, P. J. M., 2008. The problem of incertainty integration and geo-information harmonization. In: Coors V., Rumor M., Fendel E. M., Zlatanova S., (Eds.), *Urban and regional data management: UDMS annual 2007*, Leiden: Taylor & Francis, 2008, pp. 11-184.