# GeoCubes Finland – A Unified Approach for Managing Multi-resolution Raster Geodata in a National Geospatial Research Infrastructure

Lassi Lehto, Jaakko Kähkönen, Juha Oksanen and Tapani Sarjakoski
Finnish Geospatial Research Institute (FGI)
National Land Survey of Finland
Finland
e-mail: lassi.lehto@nls.fi, jaakko.kahkonen@nls.fi, juha.oksanen@nls.fi, tapani.sarjakoski@nls.fi

*Abstract*—**Providers of geospatial data are facing the challenge of diverse user needs when delivering their products to different user groups. Academic researchers represent a user group with quite specific requirements, like good support for analysis and high-performance computing. A national infrastructure providing both geospatial data and powerful geocomputing facilities for research use is being developed in Finland. The part of the infrastructure described in this paper focuses on the management, storage and efficient delivery of raster-formatted geospatial data by applying the concept of datacube.**

***Keywords-research infrastructure; raster data; datacube; GeoTIFF; GDAL.***

## I. Introduction

National Spatial Data Infrastructures (SDIs) are mostly developed as general-purpose data delivery platforms. The main driving force is usually the availability of various data sets that providers have initially built for their own use. As data sharing principles gain momentum in society, existing data sets are being made available without any specific adaptation. An example of development aiming at a customised, user-oriented SDI is the Finnish Open Geospatial Information Infrastructure for Research (oGIIR) initiative [1]. The oGIIR is a part of a major national programme developing research infrastructures (Finnish Research Infrastructure [FIRI]). The building phase of oGIIR is funded by the Academy of Finland in the context of Finland's Roadmap for Research Infrastructures [2].

The oGIIR is an open-access virtual infrastructure supporting the broad multidisciplinary scientific research community by offering geospatial data services, scalable geocomputing services and a knowledge-sharing network. The oGIIR is jointly developed by the Finnish Geospatial Research Institute (FGI) in the National Land Survey of Finland (NLS), the University of Turku, Aalto University, the University of Eastern Finland, the Finnish Environment Institute (SYKE), the Geological Survey of Finland (GTK), the Natural Resources Institute Finland (LUKE) and CSC – IT Center for Science (the provider of high-performance computing facilities for Finnish universities). The oGIIR will make the Finnish geospatial research infrastructure internationally unique in two ways: 1) by providing a strong network of cooperation, open access infrastructure and researcher knowledge sharing in order to support scientific research with geospatial information and 2) by facilitating access to high-performance geocomputing resources for research organisations.

An initiative called GeoCubes Finland (hereafter also referred to as 'GeoCubes') has been launched in the context of the oGIIR to develop a cached storage of geospatial data for supporting the needs of the Finnish research community. GeoCubes is a unified, multi-resolution repository of raster-formatted geospatial data. The main use case for this data storage is a research task involving spatial components and requiring geospatial raster source data sets. The substantial effort involved in acquiring and combining disparate spatial data sets is often seen as a major impediment for wider utilization of spatial methods in research. GeoCubes aims at facilitating spatial analysis processes by providing interoperable data sets that have been pre-processed for easy access and integration.

GeoCubes Finland contains a representative selection of Finnish geospatial data sets with national coverage. The contained data sets are transformed into a common two-dimensional grid and into a unified set of resolution levels. Standardised mechanisms are applied for the storage and provision of essential metadata. A wide set of access protocols are supported for accessing the contents of GeoCubes in order to facilitate utilisation in various client applications. In particular, mechanisms are provided for easy access to GeoCubes data sets from the high-performance geocomputing platform of CSC. The GeoCubes Finland platform is currently in its early stages of development. Thus, detailed information on performance, adaptability for a particular purpose, or user acceptance of the platform, is not yet available.

The rest of the paper is organised as follows. Section II describes the concept of a datacube and its application in the geospatial domain. Section III describes the main aspects of the GeoCubes Finland data repository. Section IV deals with the implementation details of GeoCubes Finland. Section V contains discussion and Section VI presents conclusions and possible future developments of the platform.

## II. Datacubes for Geodata

In general computing technology, a *datacube* is understood as a multi-dimensional array of data (the term *OLAP cube* is also used; OLAP: Online Analytical Processing). The dimensions of a datacube represent the points of view from which a certain value (called a *measure*)

ISBN: 978-1-61208-617-0

is looked at. If a datacube contains more than three dimensions, the term *hypercube* is also used [3]. The concept of a datacube has recently raised interest also in the geospatial domain. In this context, datacubes are defined as multi-dimensional arrays containing spatially referenced data. Examples of datacubes include one-dimensional arrays of geolocated sensor observation time series, two-dimensional arrays containing range values of geospatial coverage, three-dimensional arrays of volumetric data sets (like voxel representations of data sets in geoscience) and four-dimensional arrays representing the time series of volumetric data sets [4].

In particular, satellite images can be seen as a promising application area for datacube-based storage and data management [5]. Earth observation (EO) missions have been carried out regularly since the sixties, and the images captured thus form an extensive time series. This allows for natural treatment of EO data as a three-dimensional datacube [6].

Open Data Cube (ODC) is a large international initiative aimed at improving access to EO imagery through a unified pre-processing, harmonisation and indexing procedure [7]. An open source Python-based implementation is available to help communities in organising and analysing vast amounts of EO data and in creating useful end-user applications based on those data resources [8].

An important example of an operational national-level ODC implementation is the Australian Geoscience Data Cube (AGDC) [9]. The main three components of the AGDC include a) data preparation for improved comparability and better time-series analysis, b) a software platform that supports better data access and management, and c) the provision of a high-performance computing platform for data analysis tasks. In the data ingestion process, source imagery is processed in order to achieve comparable spatial, spectral and quality properties, and then it is tiled and stored as netCDF files. The AGDC can also deal with data sets that are only indexed and processed into the common form in an on-the-fly manner, when needed.

In standardisation, the concept of a datacube has also been raised as a possible organising principle for storing massive amounts of raster-formatted geodata. A working group, called Datacube Domain Working Group (Datacube.DWG), is planned to start working on this topic in the Open Geospatial Consortium (OGC) [10].

Recently, Baumann has made an attempt to formalise the properties of a geospatial datacube in the Datacube Manifesto [11]. According to Baumann, geospatial datacubes are supposed to express the following properties: a) they must support at least one through to four dimensions, b) datacubes must treat all axes equally, in particular they must yield good performance in selecting subsets along all axes c) datacubes must support adaptive partitioning to improve query and processing efficiency, d) datacube service implementations must support a well-defined query language for accomplishing various tasks (like data extraction, filtering, processing and integration).

The most important existing specifications unifying datacube access methods include the following OGC standards: Coverage Implementation Schema (CIS) [12], Web Coverage Service (WCS) [13] and Web Coverage Processing Service (WCPS) [14].

## III. GeoCubes Finland's specifications

### A. Content

The contents of GeoCubes Finland include a representative selection of spatial data sets maintained by governmental research organisations in Finland (like SYKE, LUKE and GTK). As reference data, some general-purpose data sets provided by the NLS are also included. Data sets are organised as individual layers of information with common representational properties for easy integration and analysis.

Examples of data sets to be stored in GeoCubes in the first phase include high-resolution elevation models and surface models (from the NLS), land-use layers (from the SYKE), soil map layers (from the GTK) and national forest inventory layers (from the LUKE).

### B. Metadata

Metadata concerning the data sets stored in GeoCubes Finland are provided as a centralised resource. Because of the particular nature of the data sets, special attention is put on providing descriptive information about the classifications applied in raster layers. This information will be made available either as Raster Attribute Tables (RATs), as internal metadata fields of the raster data file or as online code list files. As GeoCubes provides multi-resolution data storage, the applied nomenclature in most cases form hierarchical classification structures.

### C. Encoding

The encoding of GeoCubes Finland cell values depends on the nature of the data set being represented. Both classified data sets (like land use or soil maps) and data sets with continuous value ranges (like Digital Elevation Models (DEMs) or orthophotos) are included. No-data areas are represented as zero-valued cells in classified data sets and by a separate mask channel in data sets with continuous value ranges. Where practicable, the data capture date is presented as a separate time layer.

### D. Grid

The standardised grid applied in GeoCubes is based on the Finnish national Coordinate Reference System (CRS) ETRS-TM35FIN (EPSG code 3067). This projected CRS is compatible with the pan-European ETRS89 system. ETRS-TM35FIN covers the whole country in one projection zone and has the false easting value of 500 000 m on its central meridian at 27°E longitude. The origin of the GeoCubes Finland's grid (top-left corner) is located at the coordinate point (0, 7800000). The easting value of the origin is selected to avoid negative coordinates. The northing coordinate value is selected as a round 100 km value, allowing for good coverage of the country.

### E. Resolution Levels

GeoCubes Finland applies the following resolution levels: 1, 2, 5, 10, 20, 50, 100, 200, 500, and 1000 m. The resolution levels applicable for a given source data essentially depend on

ISBN: 978-1-61208-617-0

the properties, like spatial accuracy, of the data set. Round resolution values, rather than the traditionally used exponents of two in image pyramids, are selected to facilitate integration with external sources (like statistical data sets) and to follow the values commonly used in spatial analysis reporting.

### *F. Spatial Subdivision*

For easy transfer and processing of the GeoCubes Finland data sets, the content is subdivided in 100 km * 100 km blocks with a round 100 km origin (top-left corner) coordinate values. The territory of Finland can be covered with 60 such blocks (see Figure 1). The so-called virtual raster mechanism is used to treat the 60 individual files as a one continuous data set.

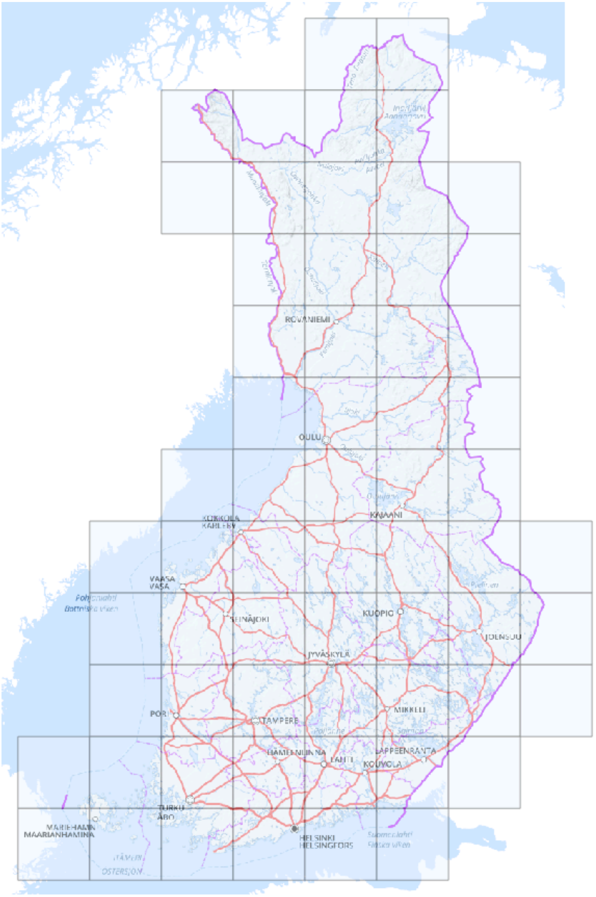

Figure 1. The block-wise spatial subdivision of GeoCubes Finland's data storage.

### *G. Access Methods*

Several access methods are supported in GeoCubes Finland in order to enable smooth usage in various user environments. Block-wise raster files are available for easy http access. A custom-made download service supports the selection of an arbitrary bounding box at a given resolution level. The efficient partial downloading of an individual block-wise raster file is also supported using an http 'GET range' request. A Web Coverage Service (WCS) interface is available for downloading GeoCubes content, supporting the definition of spatial extents both in ground and raster coordinates.

Visualisation of the GeoCubes content is provided via a Web Map Service (WMS) instance (MapServer) that can use the virtual raster representation of an individual GeoCubes data theme as its source data.

## IV. Implementation

### *A. Platform*

GeoCubes Finland – like most of the other research infrastructure components being developed in the oGIIR project – will be running on the high-performance cloud computing platform provided by the IT services provider CSC. The platform consists of two different computing environments: supercluster Taito, which is destined for massive parallel computing tasks, and cPouta, a traditional Infrastructure as a Service (IaaS) cloud computing platform. Large-scale geocomputing tasks will be run on the Taito platform, whereas interactive applications and the open data access interfaces of GeoCubes will be located in the cPouta environment. Data storage will be organised in the CSC's fast storage units. Investigation into the best possible manner to share the data storage between Taito and cPouta usage is underway.

### *B. Processing*

Data sets available in raster form are transformed into the standardised grid and into all applicable resolution levels. Vector-formatted source data sets are rasterised with selected attribute values and added to the raster storage. The needed generalisation processes are carried out in order to fill in the required resolution levels. If source data sets are available in generalised forms, those layers are used as input data.

### *C. Storage format*

In the first implementation, GeoCubes Finland data sets are stored as GeoTIFF files [15]. Each file covers one block. The different resolution levels are maintained as internal GeoTIFF overview layers. The internal structure of the GeoTIFF file is organised in the so-called cloud-optimised form for efficient extraction of the various overview levels using standard http transmission mechanisms. The BigTIFF mode is used to support vast spatial raster files. Raster content is organised into internal 256*256 cell GeoTIFF tiles to facilitate the fast extraction of sub-regions. The architecture of the initial GeoCubes implementation is illustrated in Figure 2.

### *D. Tools*

Processing of GeoCubes Finland data sets is mainly performed by the open source spatial data processing platform called the Geospatial Data Abstraction Library (GDAL) [16]. Automated processes for importing different data sources into the GeoCubes are based on the use of GDAL functionalities via Python scripting.

ISBN: 978-1-61208-617-0

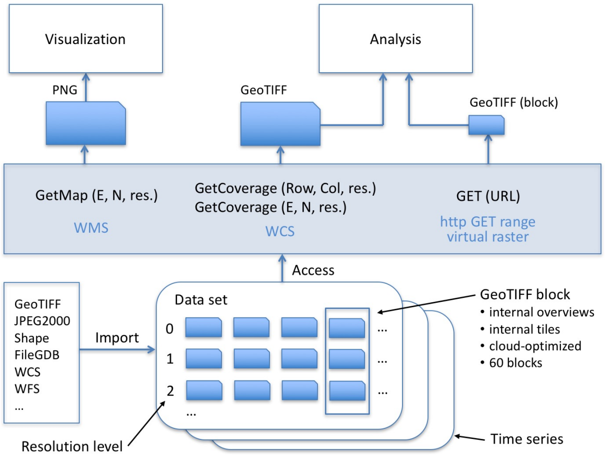

Figure 2. The system architecture of GeoCubes Finland and planned data access mechanisms.

## V. Discussion

The development of GeoCubes Finland arose from the need to have a unified approach for handling raster geodata in a multi-resolution fashion, without direct relation to the advancement of datacube approaches in the geospatial domain that we reviewed in Section II. Referring to Section III, the very central aspects of GeoCubes specifications are the following:

1. It supports handling of raster geodata sets that are mutually heterogeneous with respect to their content.
2. It uses a single coordinate reference system and one specified location for the origin of the raster data sets.
3. It uses multiple resolutions to store and represent geodata, like image pyramids are used in remote sensing imagery.
4. It uses unified principles to encode the data in raster cells.

*Northing* and *Easting* are the only natural dimensions of our datacube. Height and time could be considered to be other natural dimensions in the datacube. Looking at the datasets that are to be primarily used in the oGIIR infrastructure, there seems to be little (if any) need for voxel data with height as the third dimension. For time to be a natural dimension in a datacube, we should have data representing a phenomenon at rather regular intervals, thus forming time series data over longer periods of time. In our case we have data representing current phenomena, or only some snapshots representing the phenomenon at certain specified time instances. To summarise, on a logical level our way of modelling the data is, in many aspects, the traditional layer- and raster-based approach. Related to datacubes, GeoCubes Finland seems to mainly resemble the AGDC and ODC. However, it is useful to analyse how GeoCubes Finland fits with a datacube as defined by Baumann in his manifesto [11].

First of all, it is rather natural to consider each resolution level to form its own datacube. That is why the approach is named in the plural form – GeoCubes Finland. Secondly, the enumeration of the layers or themes can be considered to form the first dimension within each of these datacubes. Thirdly, Northing and Easting would form the second and the third dimension. GeoCubes Finland's formalism can be extended to handle volumetric and/or time series data by replacing a layer with three- or four-dimensional datacubes. Baumann's Datacube Manifesto assumes that the data values within a cube are of the same data type. This is not the case in our approach; the data type is only constrained to be the same within each layer.

Baumann's Datacube Manifesto implicitly defines or describes a datacube as a database management system and data processing environment for multi-dimensional raster data. In GeoCubes Finland, on the contrary, the focus is on representation (i.e., how the data is modelled and represented on a logical level). On the implementation level, GeoCubes uses a file-based approach and utilises the features available in the GDAL library, for example virtual rasters and overviews. As such GeoCubes' implementation is not restricted to GDAL – other realisations may be made using any software that suits the purpose. GeoCubes is not a processing and analysis environment for geodata; processing is assumed to take place in GIS or other software that has the capability to process raster geodata. The plan is to use WCS as the primary mechanism for accessing selected parts of the data. These observations make it evident that all the issues related to optimisation and performance will remain highly dependent on the specific solutions made in each implementation.

## VI. Conclusions and Outlook

The oGIIR project aims at improving access to geospatial data sets and geocomputing resources for academic and governmental research organisations. One of the aims of the project is to set up a datacube, called GeoCubes Finland, to facilitate researchers' work in cases where spatial data in raster form can contribute to the problem resolution.

GeoCubes Finland is currently in its early development phase. The first version of the specification has been developed, and the first tests with real data sets are ongoing. The future work includes further testing to refine the specifications, develop the service modules and user interfaces, and better integrate the data storage with high-performance computing facilities for spatial analysis.

## Acknowledgment

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 'Finnish Grid and Cloud Infrastructure', urn:nbn:fi:research-infras-2016072533, grant number 283818.

## References

[1] oGIIR, Open Geospatial Information Infrastructure. http://ogiir.fi [retrieved: Jan, 2018]

[2] Academy of Finland, “Finland’s Strategy and Roadmap for Research Infrastructures 2014-2020”. http://www.aka.fi/globalassets/awanhat/documents/firi/tutkimusinfrastruktuurien_strategia_ja_tiekartta_2014_en.pdf [retrieved: Jan, 2018]

[3] Wikipedia, Data cube. https://en.wikipedia.org/wiki/Data_cube [retrieved: Jan, 2018]

[4] U. Pyysalo and T. Sarjakoski, ”Voxel approach to landscape modelling”. *The International Archives of the Photogrammetry and Remote Sensing*, July 2–11, 2008, Beijing, China, XXXVII(B4/1), pp. 563–568

[5] A. Lewis et al., “Rapid, high-resolution detection of environmental change over continental scales from satellite data – the Earth Observation Data Cube”. *International Journal of Digital Earth*, Vol. 9 , Iss. 1, 2016, pp: 106-111

[6] ESA, Earth Observation Datacube. https://eodatacube.eu/ [retrieved: Jan, 2018]

[7] CEOS, Open Data Cube. https://www.opendatacube.org [retrieved: Jan, 2018]

[8] ODC, Open Data Cube Source Code Repository. https://github.com/opendatacube [retrieved: Jan, 2018]

[9] A. Lewis et al., “The Australian Geoscience Data Cube – Foundations and lessons learned”. *Remote Sensing of Environment*, 2017, ISSN 0034-4257, https://doi.org/10.1016/j.rse.2017.03.015, pp. 276-292

[10] OGC, Datacube Domain Working Group Charter. https://external.opengeospatial.org/twiki_public/pub/CoveragesDWG/Datacubes/17-071_Datacube-DWG_Charter.pdf [retrieved: Jan, 2018]

[11] P. Baumann, The Datacube Manifesto. http://earthserver.eu/sites/default/files/upload_by_users/The-Datacube-Manifesto.pdf [retrieved: Jan, 2018]

[12] OGC, Coverage Implementation Schema. http://docs.opengeospatial.org/is/09-146r6/09-146r6.html [retrieved: Jan, 2018]

[13] OGC, Web Coverage Service. http://www.opengeospatial.org/standards/wcs [retrieved: Jan, 2018]

[14] OGC, Web Coverage Processing Service (WCPS) Standard. http://www.opengeospatial.org/standards/ wcps [retrieved: Jan, 2018]

[15] GeoTIFF, GeoTIFF home page. http://trac.osgeo.org/geotiff/ [retrieved: Jan, 2018]

[16] GDAL, Geospatial Data Abstraction Library. http://gdal.org [retrieved: Jan, 2018]