Publishing OGC resources discovered on the mainstream web in an SDI catalogue

Tomas Kliment
Slovak University of Technology
Faculty of Civil Engineering
Radlinského 11
Bratislava, Slovakia
tomas.kliment@gmail.com

Carlos Granell
Joint Research Centre
Institute for Environment and Sustainability
Via E. Fermi 2749
I-21027 Ispra, Italy
carlos.granell@jrc.ec.europa.eu

Vlado Cetl
Joint Research Centre
Institute for Environment and Sustainability
Via E. Fermi 2749
I-21027 Ispra, Italy
vlado.cetl@jrc.ec.europa.eu

Marcel Kliment
Slovak University of Agriculture
Horticulture and Landscape Engineering Faculty
Tulipánová 7
Nitra, Slovakia
marcel.kliment@uniag.sk

Abstract

Nowadays geospatial data users search for geospatial information within an SDI using discovery clients of a Geoportal application (i.e. INSPIRE Geoportal). If data producers want to promote related resources and make them available in the SDI, then they need to create metadata according to the predefined rules (i.e. INSPIRE metadata regulation) and publish them using a CSW standard. This approach allows for either distributed searches or harvesting metadata from different SDI nodes. Nevertheless, there are still a lot of data producers making their resources available on the Web without documenting and publishing in a standardised way. The paper describes a workflow to provide a tool to make OGC-based geospatial services found on the Internet discoverable through CSW-compatible service catalogues and, hence, more visible to a wider SDI community.

Keywords: mainstream web, OGC services, metadata, harvesting, geospatial catalogue.

1 Introduction

"I need some datasets for my project but I do not find anything". Most users registered in GIS-related mailing lists have encountered this kind of message. Unfortunately, these desperate messages are not an exception but still the common manner to ask colleagues and peers for missing information. In the digital era, „a word of mouth" via social networking services and thematic mailing lists may work well.

Paradoxically, Spatial Data Infrastructures (SDIs) are particularly aimed to address this issue by promoting and enabling data sharing and access [1]. Apart from organizational arrangements and suitable policies in place, from a technical perspective, standardization is crucial to support interoperability and access to resources from distinct SDIs, i.e., a network of interrelated SDI nodes. Standards-based Services, such as the Open Geospatial Consortium (OGC) Catalog Service for Web (CSW), [2] have been long established to let the SDI community search and discover relevant datasets in a uniform manner [3]. The CSW catalogue service specification along with a series of profiles enables discovery, retrieval, and access to large volumes of distributed resources such as geospatial datasets and services [4]. Our desperate user might have used the CSW services. Probably he did but found no results.

Nowadays, collaboration, interrelation, multidisciplinary projects and social networks are trendy terms that all in all pursue a common objective: sharing, discovery and processing of distributed web resources. In essence there is no difference between the objective of these terms and that of the SDIs. Despite an increasing number of OGC-based services deployed on the Web [5], we can estimate, as we later argue in this paper, that not the entire group of the available OGC-based geospatial services is in reality discoverable through the standards-based geospatial-oriented catalogues (e.g. CSW). This fact undoubtedly limits sharing and access to geospatial datasets and services as those remain hidden from the SDI communities, but still could enrich them [6]. This implies for instance that the sought datasets searched by our user is likely available on the Internet through the OGC-based services which in turn are not published (i.e., registered) in the CSW-compatible geospatial service catalogues. Therefore a tool capable of an automatic transformation and extension of metadata describing the discovered geospatial resources with compliance to the requirements defined in an SDI framework (i.e. INSPIRE metadata regulation [7]) would bridge them with the SDI world. In this paper, we describe a workflow to provide a tool to make OGC-based geospatial services found on the Internet discoverable through CSW-compatible service catalogues by collecting and publishing metadata and, hence, more visible to a wider SDI community.

The paper is organised as follows: Section 2 provides an overview of the related works; Section 3 describes a methodology of the workflow applied to discover OGC services using the Google Search engine (SE), extract relevant metadata into the required structure and provide CSW services for further usage in SDIs; Section 4 summarizes the results and provides figures reporting the quantitative analyses performed; Section 5 concludes the paper and outlines future work.

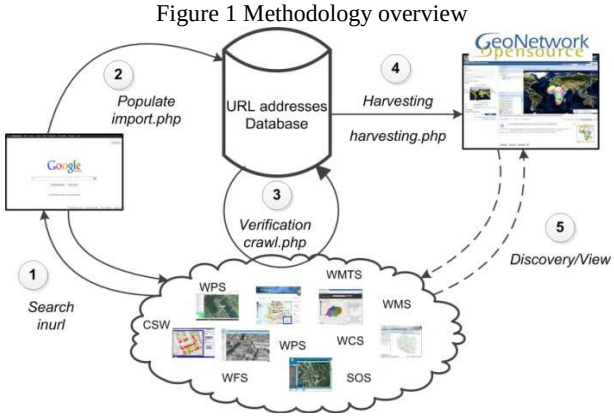
2 Related work

The EuroGEOSS Discovery Broker --an outcome of the European project EuroGEOSS [23] and an overall contribution to GEOSS (Global Earth Observation System of Systems) and its Common GEOSS Infrastructure (CGI) [8], allows users to access resources from both SDI and non-SDI infrastructures with a uniform search interface. As its name indicates, the Discovery Broker follows a brokering approach, which extends the basic broker pattern [9] by transferring the required business logic, such as coordinating requests and responses and handling with specific standards for data encoding to a brokering middleware. In such a context, the EuroGEOSS Discovery Broker provides a unique access point to services and data sources from biodiversity, forestry, and drought domains. Indeed, the potential of the brokering approach is the ability of aggregating various brokers to work cooperatively to provide an integrated functionality that augments that of any single broker [10]. Although the Discovery broker works well in connecting multidisciplinary catalogues, it is still constrained to a fixed set of back-end services, repositories and catalogues. As new geospatial services emerge on the Web, we suggest to use the entire web and the high potential of general-purpose search engines that have been proved to be valid mechanisms to discover and retrieve geospatial datasets [11] and services [5, 12], to bear in mind the dynamics of the Web and provide an up-to-date view of the huge amounts of geospatial services available on the Web. In a similar spirit, the recent Spatineo Directory service [24] embeds a search engine to look up over 8500 geospatial web services discovered on the Web. It also offers a kind of "service health" functionality of each discovered service in compliance with the INSPIRE monitoring guidelines [13, 14, 15]. This functionality is very important, because the testing results [16, 17] achieved applying standardized methodology [18] may provide an overview of the interoperability within an SDI ensured by implementing discovered OGC services.
As we describe in the following section, the tool presented is similar to that of Spatineo but based on the well-known open source/free geospatial components.

3 Methodology

In this section we describe the methodology for publishing OGC services discovered on the Web in a CSW-compliant geospatial catalogue. Figure 1 illustrates five main steps of a workflow that allow us to keep a wide range of geospatial services updated and accessible through a CSW catalogue. In order to support the steps proposed in the workflow, we make use of the available tools and applications such as Google Search Client [25], OutWit Hub [26], GeoNetwork opensource [27], MySQL DBMS (database management system) [28] and Apache HTTP Server [29] used to run developed PHP (Hypertext Preprocessor) [30] scripts.

Figure 1 Methodology overview



In particular, the individual steps shown in figure 1 are summarized in the following subchapters [19].

3.1 Search in Google SE to discover URL addresses pointing to OGC services

Google search client was used as a search interface to discover available OGC services endpoints. An advanced search operator inurl was used for a query definition in order to restrict the search to predefined query strings in the URLs of records stored in the Google database. An example of a query definition used to discover WMS services GetCapabilities URLs was as follows:

inurl:service=WMS inurl:request=GetCapabilities

The same pattern was used for all seven types of OGC services we were searching for, thus the part of the string after *service=* (in the example above WMS) has been changed with the standardized acronyms as follows: WFS (Web Feature Service), WCS (Web Coverage Service), SOS (Sensor Observation Service), WPS (Web Processing Service), CSW (Catalogue Service for the Web) and WMTS (Web Map Tile Service). Useful information for each type of the OGC service such as Title, Description and URL were extracted from the HTML representation of a Google result list and converted into comma separated value (csv) files per service type (i.e., wms.csv, wfs.csv, sos.csv etc.). These csv files were used as inputs for the next step in the workflow.

3.2 Populate a database with discovered URL addresses

A MySQL database was created as a central storage of the discovered OGC services endpoints (GetCapabilities URL addresses). It is a simple database, which contains currently only one table *services*, used for storage and further processing. In order to populate the database with the results retrieved from Google, a PHP script *import.php* was developed. The script retrieves URL addresses from the csv files created in step 1 and populates the predefined columns in the database with the parsed URL addresses and related information as title and description.

3.3 Verify OGC services endpoints

The objective of this step was to verify whether the services found are available to be used. In order to implement this step, another PHP script *crawl.php* was developed. The script fetches a GetCapabilities URL stored in the database (Step 2), triggers it (sends a GetCapabilities request) and verifies service availability. If the service is available, the script extracts relevant information (e.g. version, etc.) to update the corresponding table columns and sets the *status* column as "available". Otherwise, the *status* column is flagged as "unavailable" and *version* remains empty. The script tests those records that have the *status* column empty. Once the validation script is finished, each record of the resulting database contains an URL pointing to a GetCapabilities method of the OGC service together with additional information (version and status) necessary for the next step.

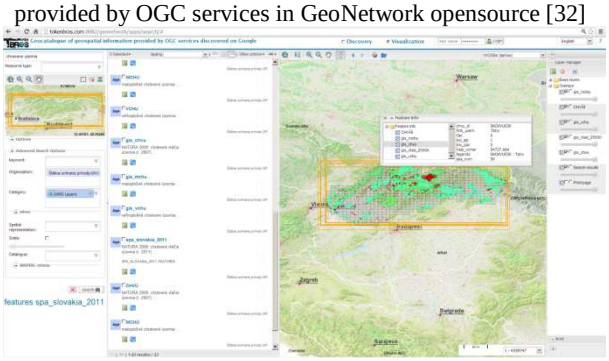
3.4 Create and run metadata harvesting tasks in GeoNetwork for OGC services endpoints

GeoNetwork opensource catalogue was used as a central repository for ISO 19115/19119 metadata created for the discovered OGC services (WMS, WFS, WCS, WFS, SOS and CSW) and their content (WMS Layers, WFS Features, WCS Coverages, SOS Observations and CSW datasets, series and services). Harvesting tasks were created and run for all the OGC services endpoints stored and verified in the previous step 3. Harvesting XML services (xml.harvesting.get, xml.harvesting.add and xml.harvesting.run [31]) provided by the GeoNetwork API were used to manage harvesting task in a programmatic way with a third developed PHP script *harvesting.php* through HTTP POST messages. This script looks for services for which the version was detected (they have the value "available" in the column *status*) and creates harvesting tasks in GeoNetwork. Two types of harvesting tasks have been defined: the first one for catalogue services - csw (for the records that have stored the value *csw* in the column *type*) and the second one for other OGC services such as WMS, WFS, WCS, WPS and SOS - ogcwxs (other acronyms stored in the column *type*). In order to avoid memory performance issues on the server running GeoNetwork, the script executes only one harvesting task at the time. The following harvesting task is created and executed only if the previous one has finished. WMTS services have not been used in this step, because GeoNetwork currently does not provide the corresponding type of a harvesting task. Associated categories for each OGC service type and its content have been created in the GeoNetwork database in order to enable distinct searches as described below.

3.5 Provide discovery/viewer web interface

GeoNetwork opensource has also been used as a client application to provide a web-based search interface on the metadata harvested in the previous step, as well as a map client to portray layers from the discovered WMS services. The figure below shows a customized GUI of GeoNetwork opensource, version 2.8.0 RC0, deployed in September 2012.

Figure 2 Discovery and view of geospatial information provided by OGC services in GeoNetwork opensource [32]



The discovery component (left side of figure 2) provides a simple (full text) and an advanced search query definition, which is matched with the information contained in the metadata harvested in the previous step and subsequently provided to the user. The view client (right side) communicates directly with remote services in order to portray layers on a map and retrieve additional information about the features if available. Finally, another GeoNetwork feature called Virtual CSW has been used in order to provide several CSW interfaces for individual OGC services and their content to be discoverable with remote SDI clients. For instance, any other GeoNetwork node or an INSPIRE Geoportal can use this endpoint to query individual types of services and geospatial content they provide. The endpoints of individual CSW endpoints are listed in the next section.

4 Results

This section first summarizes some interesting statistics of individual OGC service types after conducting five searches on Google and the total amount of OGC services GetCapabilities URL addresses stored in the database. Figure 3 below shows the numbers of URL addresses discovered for individual OGC services in different periods and a summary of all imported into the database. Figure 4 refers to the same figures, but after the verification procedures performed on both URL addresses parsing and OGC services availability tests.

Figure 3 Discovered and imported URL addresses

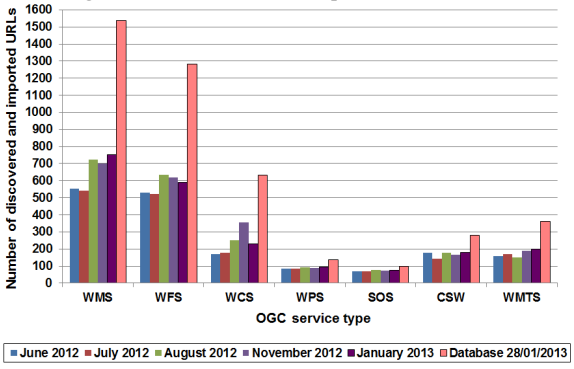
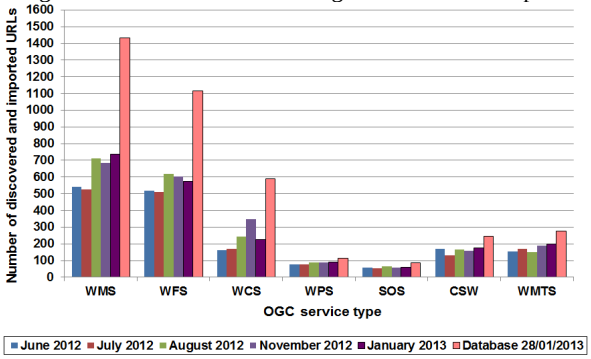


Figure 4 Verified and functioning OGC services endpoints



The total number of functioning OGC services stored in the database after a verification procedure performed on the 28th of January was 3855 out of 4328 total amount of the discovered URL addresses (89.07%). A percentage increase of an amount of the stored functioning services was for WMS 123.70%, WFS 97.35%, WCS 155.86%, WPS 33.02%, SOS 45.55%, CSW 52.12% and WMTS 60.86%. For instance, the average number of WMS services discovered after 5 searches was 640.6 and the number of the functioning stored in the database was 1433. Therefore an increase of WMS services stored was 792.42, which means that on average 158.48 new WMS services were found in each search. The following figures (figure 5 and 6) show the quantity of metadata records created in GeoNetwork after the individual harvesting tasks for all the OGC services summarized in figure 4 have been finished. Figure 5 summarizes the results gained from the harvesting tasks type ogcwxs, thus for all OGC services and their content except CSW (which has a particular harvesting task) and WMTS (which is not supported yet). Figure 6 provides an overview of the amounts of metadata records generated after the harvesting tasks of type ogccsw that were created and executed for the discovered CSW services. A metadata element hierarchyLevel [21] was used to distinguish types of geospatial information described by the metadata harvested into the catalogue.

Figure 5 Numbers and categories of metadata collected from *ogcwxs* harvesting tasks

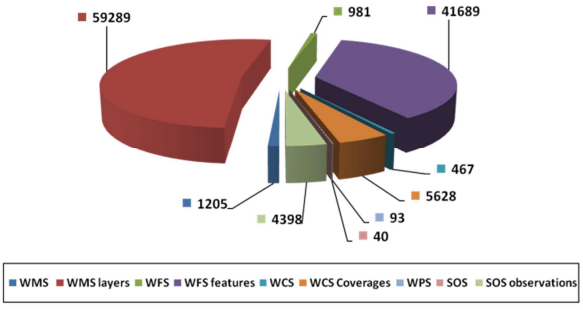
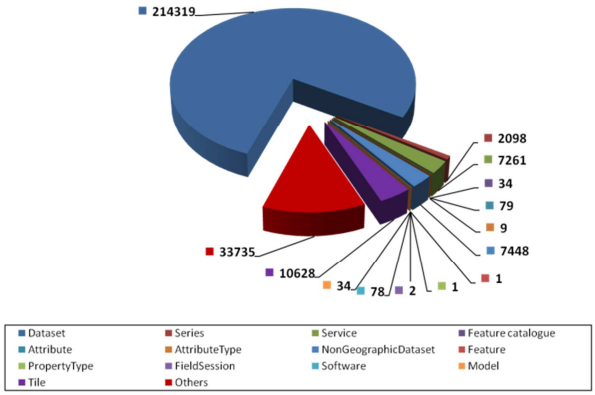


Figure 6 Numbers and categories of metadata collected from *ogccsw* harvesting tasks



The results shown in figure 5 provide an overview of the metadata records created within the harvesting tasks executed for the functional OGC services (figure 4) by GeoNetwork. However, the provided figures show that the numbers of functioning services and corresponding metadata records do not match (i.e. 1433 functioning WMS services stored in the database and 1205 metadata records created in GeoNetwork). This might reflect compatibility issues between the GeoNetwork application and the interfaces of remote services being requested due to the versions currently supported by GeoNetwork and the incorrectly provided values in capabilities documents (i.e. value 1.3.2 was provided for 10 discovered WMS services). Figure 6 shows the number of metadata records created after the successful harvesting tasks for the csw type. A large majority of the created metadata describes resources of the type dataset (214319 records). Another interesting number is 7261 metadata records describing network services, of which: 4016 for view services, 239 download, 24 discovery, 6 invoke, 3 transformation and 36 other types of services defined by the INSPIRE classification to assist in the search of the available spatial data services [7]. Not all resource types defined in [21] were presented in the created metadata records and some had either not any values defined or not standardized values (Others - 33735 records). The harvested metadata records are available also through a CSW interface on the following CSW virtual endpoints:

- CSW services (275727 records): http://tokenbros.com:8082/geonetwork/srv/csw-csw
- WMS services (1205 records): http://tokenbros.com:8082/geonetwork/srv/csw-wms
- WMS layers (59289 records): http://tokenbros.com:8082/geonetwork/srv/csw-wms-layers
- WFS services (981 records): http://tokenbros.com:8082/geonetwork/srv/csw-wfs
- WFS features (41689 records): http://tokenbros.com:8082/geonetwork/srv/csw-wfs-features
- WCS services (467 records): http://tokenbros.com:8082/geonetwork/srv/csw-wcs
- WCS coverages (5628 records):

http://tokenbros.com:8082/geonetwork/srv/csw-wcs-coverages

- WPS services (93 records): http://tokenbros.com:8082/geonetwork/srv/csw-wps
- SOS services (40 records): http://tokenbros.com:8082/geonetwork/srv/csw-sos
- SOS observations (4398 records) - http://tokenbros.com:8082/geonetwork/srv/csw-sos-observations.

5 Conclusions and future work

The approach presented here enhances discoverability (or visibility) of OGC services through SDI mechanisms on a global scale. By visibility we mean the ability to discover OGC services through CSW services, as a standard discovery mechanism to SDI users. However, not all OGC services are registered and hereby reachable via CSW services, or some are registered in individual CSW nodes that are not interlinked. For that reason, many provided information may not be discoverable through the common discovery interface (e.g. 1537 view services metadata available on the INSPIRE Geoportal, whereas 4016 available on the Geocatalogue). On the positive side, the results of the paper indicate that service providers and organizations use extensively OGC services, as already identified in recent surveys [12]. However, on the negative side, it seems that service providers do not pay attention to enhance their discoverability, i.e., such services remain hidden to the SDI community, either due to a lack of skilled personnel or of easy-to-use publishing tools that would automate the registration of access, download and view services into catalogues [20]. In essence, the approach allows us to reflect continuously the current state of play of the OGC services available on the Web and accessible to the SDI community, since those services and the geospatial content they provide are now discoverable through a standards-based discovery tool described earlier.

Further analysis will be carried out on the metadata harvested from the discovered OGC service in order to provide a more accurate overview of the number of OGC services that are not registered in the available CSWs as a next step. Another extension to the present work shall be driven by promoting Linked Data principles to semantically connect the related geospatial information resources through the implemented CSW [22]. Additionally, the range of geospatial resources discovered through general-purpose search engines may be extended, from OGC services to actual geospatial datasets published in several ways as well as to geospatial applications, research projects, etc. This would leverage CSW services coverage to their full potential as a key SDI service for discovery, retrieval, and access to any kind of geospatial resources available on the Web.

References

[1] INSPIRE EU Directive: Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). Official Journal of the European Union, L 108/1, 50 (2007).

[2] Nebert, D., Whiteside, A., Vretanos, P., 2007. OpenGIS Catalogue Services Specification. OGC 07-006r1.

[3] Nogueras-Iso, J., Zarazaga-Soria, F.J., Béjar, R., Álvarez, P. J., Muro-Medrano, P.R., 2005. OGC Catalog Services: a key element for the development of Spatial Data Infrastructures. Computers & Geosciences, 31(2): 199–209.

[4] Aijun Chen, Liping Di, Yuqi Bai, Yaxing Wei, Yang Liu, 2010. Grid computing enhances standards compatible geospatial catalogue service. Computers & Geosciences, 36(4): 411–421.

[5] Lopez-Pellicer, F.J., Béjar, R., Florczyk, A.J., Pedro R., Muro-Medrano, P. R., Zarazaga-Soria, F. J., 2011. A review of the implementation of OGC Web Services across Europe. International Journal of Spatial Data Infrastructures Research, Vol 6 (2011), 168-186.

[6] Florczyk, A. J., López-Pellicer, F. J., Nogueras-Iso, J., Zarazaga-Soria, F.J., 2012. Automatic Generation of Geospatial Metadata for Web resources. International Journal of Spatial Data Infrastructures Research, Vol 7 (2012), 151-172.

[7] Commission Regulation (EC) No 1205/2008 of 3 December 2008 implementing Directive 2007/2/EC of the European Parliament and of the Council as regards metadata.

[8] Nativi, S., Craglia, M., Pearlman, J., 2012. The Brokering Approach for Multidisciplinary Interoperability: A Position Paper. International Journal of Spatial Data Infrastructures Research, Vol 7 (2012), 1-15.

[9] Buschmann, F., Meunier, R., Rohnert, H., Sommerland, P., Stal. M., 1996. Pattern-oriented software architecture volume 1: A system of patterns John Wiley & Sons Ltd.

[10] Vaccari, L., Craglia, M., Fugazza, C., Nativi, S., Santoro, M., 2012. Integrative Research: The EuroGEOSS Experience. IEEE Journal of Selected Topics in Applied Earth Observations and remote Sensing, 5(6): 1603-1611.

[11] Abargues, C., Granell, C., Díaz, L., Huerta, J., Beltran, A, 2009. Discovery of user-generated geographic data using web search engines. In: G. Jedlovec (Ed). Advances in Geoscience and Remote Sensing. Vukovar: In-Tech, 2009, pp. 207-228.

[12] López-Pellicer, F. J., Florczyk, A. J., Béjar, R., Muro-Medrano, P. R., Zarazaga-Soria, F. J., 2011. Discovering geographic web services in search engines. Online Information Review 35 (6), 909–927.

[13] Technical Guidance for the Implementation of INSPIRE discovery services. IOC Task Force for Network Services, 07/11/2011. Online: http://inspire.jrc.ec.europa.eu/documents/Network_Services/TechnicalGuidance_DiscoveryServices_v3.1.pdf

[14] Technical Guidance for the Implementation of INSPIRE view services. IOC Task Force for Network Services, 07/11/2011. Online: http://inspire.jrc.ec.europa.eu/documents/Network_Services/TechnicalGuidance_ViewServices_v3.1.pdf

[15] Technical Guidance for the Implementation of INSPIRE download services. Initial Operating Capability Task Force, 12/06/2012. Online: http://inspire.jrc.ec.europa.eu/documents/Network_Services/Technical_Guidance_Download_Services_3.0.pdf

[16] Ardielli J., Horak J, Ruzicka J. 2012. View Service Quality Testing according to INSPIRE Implementing Rules. In Elektronika ir Elektrotechnika. 2012. No. 3(119), Išleido Kauno technologijos universitetas, Kaunas, Latvia. ISSN 1392 – 1215. PP 69-74.

[17] Horak J, Ruzicka J, Ardielli J. 2013. Výkonové a zátěžové testy stahovacích služeb ČUZK dle požadavků INSPIRE. Zborník sympózia. Ostrava, ČR, 23.-26.1.2011. Ostrava: VŠB-Technická univerzita Ostrava, 2013. ISBN 978-80-248-2951-7. p.9.

[18] Kliment, T., Tuchyňa, M., Kliment, M. 2012. Methodology for conformance testing of spatial data infrastructure components including an example of its implementation in Slovakia Slovak Journal of Civil Engineering. Volume XX, Issue 1, Pages 10–20, ISSN (Online) 1210-3896, ISSN (Print) 1338-3973.

[19] Kliment, T., 2012. Searching for geospatial information resources on the Internet. Dissertation thesis. Slovak University of Technology in Bratislava, Faculty of Civil Engineering, Department of Theoretical Geodesy, (2012).

[20] Díaz, L., Granell, C., Gould, M., Huerta J., 2011. Managing user generated information in geospatial cyberinfrastructures. Future Generation Computer Systems, 27(3): 304-314, 2011.

[21] ISO, 2003. ISO19115:2003, Geographic information - Metadata. Tech. Rep. 19115:2003, ISO/TC 211.

[22] Lopez-Pellicer, F. J., Florczyk, A., Rentería-Aguaviva, W., Nogueras-Iso, J., Muro-Medrano, P., 2011. CSW2LD: a Linked Data frontend for CSW. In: II Iberian Conference on Spatial Data Infrastructures (JIIDE 2011). Institut Cartogràfic de Catalunya.

Online resources

[23] EuroGEOSS - www.eurogeoss.eu

[24] Spatineo Directory - http://www.spatineo.com/

[25] Google Search Client - https://www.google.com

[26] OutWit Hub - http://www.outwit.com/

[27] GeoNetwork opensource - http://geonetworkopensource.org/

[28] MySQL DBMS - http://www.mysql.com/

[29] Apache HTTP Server - http://httpd.apache.org/

[30] PHP - http://php.net/

[31] Geonetwork harvesting services - http://www.geonetworkopensource.org/manuals/trunk/developer/xml_services/services_harvesting.html

[32] Geocatalogue of geospatial information provided by OGC services discovered on Google - http://tokenbros.com:8082