BORDERLESS GEOSPATIAL WEB (BOLEGWEB)

V. Cetl [a], T. Kliment [a], M. Kliment [b]

[a] Faculty of Geodesy, University of Zagreb, Zagreb, Croatia - (vcetl, tkliment)@geof.hr
[b] Horticulture and Landscape Engineering Faculty, Slovak University of Agriculture in Nitra, Nitra, Slovakia - marcel.kliment@uniag.sk

WG IV/FIG - FIG's contributions to the Geo-Spatial Society

KEY WORDS: Mainstream Web, Geoinformation, SDI, OGC resources, Search, Access and Use

ABSTRACT:

The effective access and use of geospatial information (GI) resources acquires a critical value of importance in modern knowledge based society. Standard web services defined by Open Geospatial Consortium (OGC) are frequently used within the implementations of spatial data infrastructures (SDIs) to facilitate discovery and use of geospatial data. This data is stored in databases located in a layer, called the invisible web, thus are ignored by search engines. SDI uses a catalogue (discovery) service for the web as a gateway to the GI world through the metadata defined by ISO standards, which are structurally diverse to OGC metadata. Therefore, a crosswalk needs to be implemented to bridge the OGC resources discovered on mainstream web with those documented by metadata in an SDI to enrich its information extent. A public global wide and user friendly portal of OGC resources available on the web ensures and enhances the use of GI within a multidisciplinary context and bridges the geospatial web from the end-user perspective, thus opens its borders to everybody.

Project "Crosswalking the layers of geospatial information resources to enable a borderless geospatial web" with the acronym BOLEGWEB is ongoing as a postdoctoral research project at the Faculty of Geodesy, University of Zagreb in Croatia (http://bolegweb.geof.unizg.hr/). The research leading to the results of the project has received funding from the European Union Seventh Framework Programme (FP7 2007-2013) under Marie Curie FP7-PEOPLE-2011-COFUND. The project started in the November 2014 and is planned to be finished by the end of 2016. This paper provides an overview of the project, research questions and methodology, so far achieved results and future steps.

1. INTRODUCTION

1.1 Background

Rapid development of Spatial Data Infrastructures (SDIs) around the world triggered by INSPIRE (INSPIRE, 2007) and other similar initiatives make more and more geospatial information (GI) resources (data and services) available on the web. The main objective of SDI is to promote and enable data sharing and access. Crucial SDI component which enable users to search and find GI resources are metadata. Metadata are first visible component of each SDI for users. Figure 1 shows general usage scenario of SDI and role of metadata for search and evaluation.

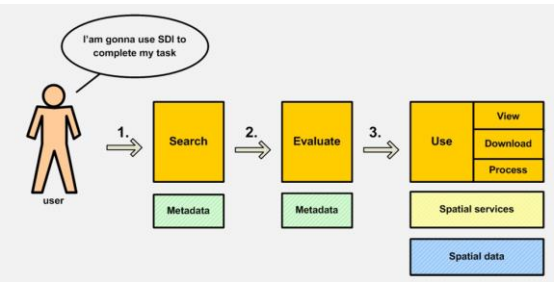

Figure 1 Generic use scenario of an SDI

Metadata should enable search for GI resources but also evaluation of fit for purpose. On the service level also harvesting of metadata should be enabled.

In the frame of SDI, the metadata are usually divided as metadata for geospatial data and metadata for geospatial services. Both are served in the standardised way by using discovery services. Most popular one is the Open Geospatial Consortium (OGC) catalogue service which is implemented in many existing either commercial or open source software solutions. Geospatial data users search for GI resources within an SDI using discovery clients of a Geoportal application (i.e. INSPIRE Geoportal[1]) (Kliment et al., 2013a).

All mentioned before works fine in an SDI environment where both producers and users are aware of SDI, corresponding services and how to use them. On the other hand, there are many ordinary users who are not aware of SDIs. They usually search for GI resources through web search engines such as Google, Yahoo, Bing etc. Also there are still a lot of GI data producers making their resources available on the Web without documenting and publishing in a standardised way. They would need to create and publish metadata in predefined structure describing GI resources in order to make them discoverable through an SDI. This approach allows for either distributed searches or harvesting metadata from different SDI nodes. The problem is that this data is stored in databases located in a layer, called the invisible web (Figure 2), thus are ignored by conventional web search engines. Therefore, a crosswalk needs to be implemented to bridge the mainstream web with SDI to enrich its information extent.

[1] http://inspire-geoportal.ec.europa.eu/

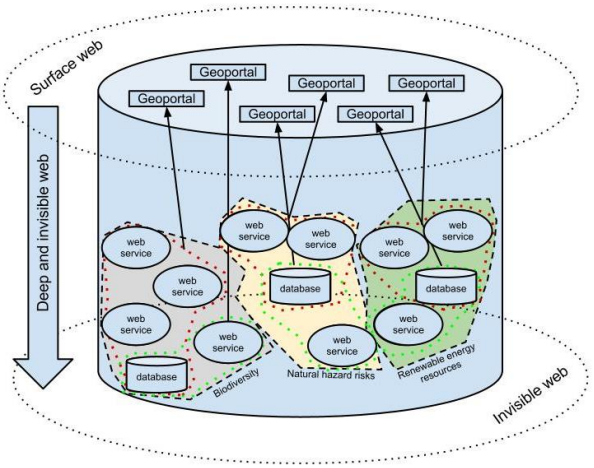

Figure 2 Geospatial-Web components and their positions in the web (Florczyk, 2012)

More over nowadays SDIs provide a gateway to vast amounts of GI resources in the web. SDIs are not only sharing platform for traditional data producers and their services but also for others e.g. semantic web, linked data, mobile applications etc. However, if someone searches for those services he/she may not find them since not all OGC services have online available and accessible metadata.

This paper provides preliminary results of the main activities performed in the research project Bolegweb[2]. The Bolegweb project aims at the development of a geospatial meta-search crawler to collect online accessible GI resources published on the Web using OGC services, harvest the geospatial metadata and deploy Graphic User and Application Programming Interfaces (GUI and API) facilitating access for different user groups (Kliment et al., 2015).

1.2 Previous research

Discovery of Geospatial Resources: Methodologies, Technologies, and Emergent Applications book (Díaz et al., 2012) presents a collection of attempts to push forward the automated discovery of GI resources. Contributions in the book from different authors provide a wide spectrum of perspectives and possible methods.

For the producers an added value would be to provide an automated way of metadata creation and update in order to decrease their work load and let them focus on actual GI resources (Kliment, 2012; Florczyk et al., 2012).

Many research activities (Abargues et al., 2009; López-Pellicer et al., 2011; Kliment et al., 2013) have reported that so called mainstream web provides many valuable GI resources of several types (e.g. OGC services, KML data, etc.). They can be discovered using web search engines. The benefit in comparison with SDI engines (geocatalogues) is that the former automatically crawls the web in order to discover GI resources of several types.

GI resources discovered within web search engines may extend significantly the information richness of the current SDI portals and may be used to develop specific domain oriented SDI portal also with combination of social media, voluntary geographic information (VGI) etc. (Kliment et al., 2013b).

Current research in this area is oriented towards discovering GI resources produced by the crowd i.e. VGI (Poorazizi, 2015). Besides the different methods that are used to store data, different technologies for the construction of GeoWeb2.0 applications are implemented. This wide variety in data storage and use of technologies makes the discovery by the crowd more and more difficult. On the other hand, the huge potential of VGI is still not explored especially impact on many human activities e.g. disaster risk management, migrations etc.

The importance and relevance of the discovery of GI resources can be seen also through the establishment of Spatial Data on the Web OGC/W3C working group in 2014. The main objective is to make easier to publish and use spatial data on the web. Improving discovery of spatial data on the Web is recognized as a one of the use cases that demand a combination of geospatial and non-geospatial data sources and techniques[3]. There are two important requirements defined as:

- Crawlability. Spatial data on the Web should be crawlable, allowing data to be found and indexed by external agents
- Discoverability. It should be easy to find spatial data on the Web, e.g. by means of metadata aimed at discovery. When spatial data are published on the Web, both humans and machines should be able to discover those data.

2. BOLEGWEB PROJECT

2.1 Research idea

The main objective of the project is to design, develop and implement a complex solution for the discovery of GI resources from available web services defined by OGC standards on the Internet (later called as “OGC Metasearch Enhanced Crawler”); provide geospatial metadata according to standards of current communities that potentially might be seeking for geospatial resources (SDI, Semantic and Open Data communities). It is based on the following sub-objectives: *i)* To design and develop OGC metasearch enhanced crawler for automatic collection of OGC services end points with all related resources available on the Internet; *ii)* To publish metadata describing both OGC-services and the content published (geographic layers, datasets, coverages, observations etc.) through implementation of an SDI catalogue, thus contribute to the information coverage provided by current SDI’s implementations at any level; and *iii)* To design and develop a user-friendly web based graphic user and application programming interfaces implementing search facilities on GI resources provided by OGC-services discovered on the Internet. The following seven OGC services are currently collected within the Bolegweb project:

1. Web Map Service (WMS) – is a web service operating on both raster and vector geospatial data and provides their image representation in a map form.
2. Web Feature Service (WFS) – is a web service operating on vector geospatial data and provides access to actual datasets features (geometries) and related data types (attributes).

[2] http://bolegweb.geof.unizg.hr/

[3] http://www.w3.org/TR/sdw-ucr/#ImprovingDiscoveryOfSpatialDataOnTheWeb

3. Web Coverage Service (WCS) – is a web service operating on raster multidimensional gridded data, extending WMS by formats used for complex modelling and analysis.
4. Web Processing Service (WPS) – is a web service operating on raster/vector data in either directions as process inputs and outputs as results of analysis executed by the service.
5. Sensor Observation Service (SOS) – is a web service provisioning observational and measurement data collected on a spatial feature.
6. Catalogue Service for Web (CSW) – is a web service operating on collections of descriptive information (metadata) being created and maintained for geospatial data and services.
7. Web Map Tile Service (WMTS) – is a web service operating on map tiles of spatially referenced data using tile images with predefined content, extent, and resolution.

2.2 Methodology

System concept definition of the Meta-Search Enhanced Crawler System (MSECS) high-level architecture was designed and documented with the architecture schema depicted in Figure 3.

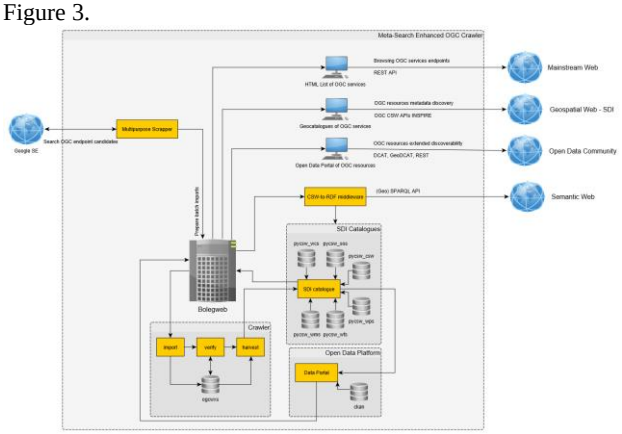

Figure 3 Bolegweb system architecture overview

The overall infrastructure of the Bolegweb system is quite complex and consists of several components. The services collection flow starts on *Google search engine* where the potential candidates of OGC services endpoints are collected by *Multipurpose Scrapper* component using advanced search parameters.

In the next step the results enter the *Crawler* component to process the data as follows: (i) Gathered OGC services' endpoints and related information are imported into the crawler database (import);
(ii) Crawler's verification script checks availability of collected services and extracts the service type, version, basic quality parameters, server location and other parameters (verify);
(iii) Harvesting script collects the metadata for services and resources they operate on in an *SDI catalogue* (harvest).

Another component of the system is represented by set of *SDI catalogues* implemented as Catalogue Service for Web (CSW). Software pycsw[4] is a python based implementation of the OGC's CSW server, which allows the discovery and publishing of metadata for geospatial resources. It can be deployed as a standalone server or can also be embedded in other applications (Sibolla et al., 2014). We use pycsw as the main metadata management system. Individual service type has its own pycsw instance (e.g. pycsw_wms), where the metadata from the service endpoint are harvested for both services and the content they operate on.

For the open data community, we designed component *Open Data Platform* based on available data portal implementation based on CKAN out-of-the-box software solution[5], the world's leading open-source data portal platform that makes data accessible – by providing tools to streamline publishing, sharing, finding and using data. It has native capability to harvest several types of open resources, including SDI catalogue CSWs. We use this functionality to collect the metadata about discovered OGC resources managed by SDI catalogues communicating to CSWs and storing into the CKAN internal data model. Metadata can be exposed to communities using mainstream REST or DCAT as well as geospatial community specific GeoDCAT standards.

The last component of the system provides an extension over the SDI catalogues and serves as a gateway between SDI and Semantic Web communities. *CSW-to-RDF middleware* is implemented using the TripleGeo-CSW[6] software developed by the GeoKnow project team that as a proof of concept set up an instance of this middleware against CSWs from public authorities across Europe, which involved datasets complying with the EU INSPIRE Directive. Experience gained testifies that TripleGeo-CSW can assist stakeholders to repurpose existing CSWs with minimal overhead and readily expose spatial metadata on the Semantic Web (Athanasiou et al., 2015). The metadata we collect into SDI catalogues from available OGC services are yeah I exposed using GeoSPARQL endpoint in a way where SPARQL queries are parsed to identify filter criteria and generate respective CSW GetRecords request, which is then send the remote CSW service via HTTP/POST. The service responds with ISO/XML metadata files that are transformed to the RDF/XML form and provided as RDF triples in the result set.

2.3 Results and discussions

The collection process of OGC services endpoints has been launched in October 2013 (Figure 4).

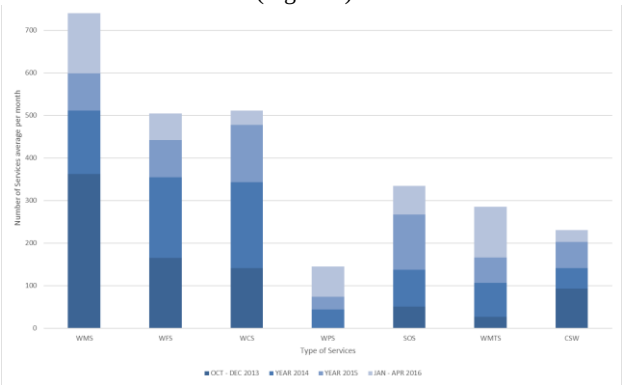

Figure 4. Number of URLs discovered within individual years for OGC services type.

[4] http://pycsw.org/

[5] http://ckan.org/

[6] https://github.com/GeoKnow/TripleGeo-CSW

Process have been repeated one per month until today. The last metasearch and crawling was performed on 30 April and the Crawler database contains in total 21368 records. The number of URLs collected within the whole period aggregated into individual calendar years and OGC service type is represented by Figure 4.

It is important to stress that this data characterise also those URLs that might not provide direct links to OGC services, instead to a web page, which obviously may after further crawling of URLs provide an access point. The current verification step in the crawler workflow, however does not investigate further those URLs, which does not provide an XML type response with a root attribute identifying the OGC Service version. (e.g. <sos:Capabilities *version="2.0.0"* ….. >). In addition, some of the discovered URLs might no longer be available, or even invalid. After filtering those potential OGC service candidates, the current database underneath the crawler system provides interesting numbers of records representing functioning services (Figure 5).

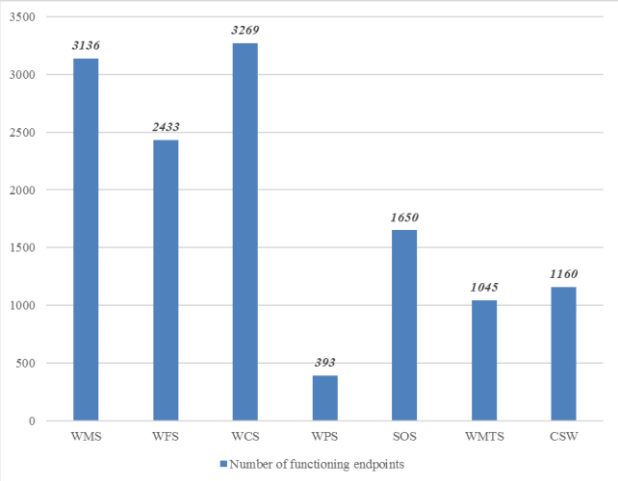

Figure 5. Number of functioning endpoints of OGC services types (valid to date 10 May 2016).

Obviously the figures are changing frequently due to the fact of dynamicity of the Web environment and geospatial web specifically. Some services available today might be offline or removed tomorrow and new ones may become publicly available.

The geospatial coverage of the results is world-wide, covering the majority of continents (Figure 6). Naturally the lesser amount of services is covered by Africa continent, however the situation has noticeably improved within the last period of the Bolegweb crawler system operation. In addition, countries of middle east have no records of OGC services operating from servers located there. United States are covered by 6022 functioning OGC services (915 WMS, 891 WFS, 2181 WCS, 1631 SOS, 102 WPS, 156 WMTS and 149 CSW) endpoints and second Germany by 1053 (444 WMS, 270 WFS, 101 WCS, 16 SOS, 40 WPS, 64 WMTS and 120 CSW). The second group is represented by more than 20 countries that publish their data with OGC services in amounts between one hundred and one thousand. Remaining more than 85 countries have less than hundred services available on the Web.

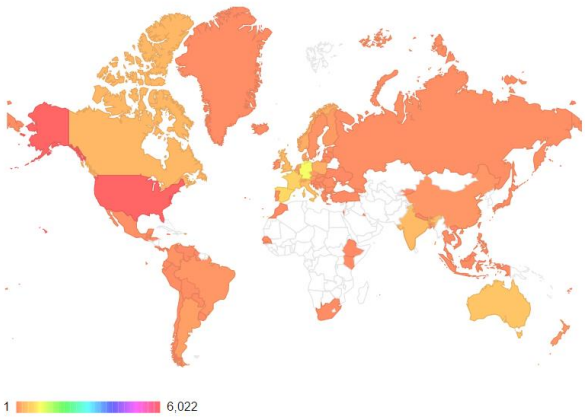

Figure 6. Geographic distribution of available OGC services endpoints.

Regarding the service types versions, the following results can be reported: 65% of WMS services has version 1.3.0 and 31,1% still the older 1.1.1. Situation with WFS services is more balanced where the majority is represented by older versions of the standard WFS 1.1.0 (40.7%) and WFS 1.0.0 (36,7%), whereas the last version 2.0 covers 21,5% but with increasing trend in time. The numbers of WCS services are increased due spatial and thematic scale of the data WCS can serve. The vast majority of service use the standard version WCS 1.0.0 (72.8%) and then version WCS 1.1.2 (13.8%). The last version of the standard WCS 2.0 is represented by only 2% of services. The youngest from the standards, SOS has also increasing number of available services in the last period. The 96.7% of available services report the version 1.0.0 and only 2.4% the last version 2.0.0. Very similar situation exists for WPS and CSW services where WPS version 1.0.0 is covered by 89.8% and CSW 2.0.2 by 90.8%. The WMTS services are covered by WMTS 1.0.0 on 97.7%.

The achieved results are in the line with current developments in GI resources distribution. They also show clear trend of more and more available GI resources as a result of rapid SDI developments around the world. Very strong push for it in Europe is driven by INSPIRE Directive and fulfilment of its roadmap which is legally binding obligation for European member states.

2.4 Products

The system architecture represented by the Figure 3 identifies at its end graphic user and application programming interfaces (GUIs and APIs) serving the data resources collected from within the Bolegweb infrastructure to various communities using the Web.

For the mainstream developers who have interest in discovering OGC service in rather easy and straightforward way both GUI and API based on REST are available. The GUI developed for the mainstream Web community and easy access to OGC services is entitled *HTML List of OGC Services* (Figure 7).

Figure 7. GUI of HTML List of OGC Services[7].

This simple page provides a list of OGC Web Services represented in a simple tabular view on the database implemented underneath Crawler. Information as OGC Service type, version, the location of the server where the service is operated from, the date when the service was discovered on Google SE, date when it's availability has been checked lastly and finally its status after the last check are available. User can filter by "Free text" way each column. In example, a user can filter OGC service type typing WFS in the service type column, to restrict version, the country name (e.g. "Czech") in the Server location column and value one for functioning service in the column status as depicted in figure 7. To access the same information from the remote application, REST service is made available and provides the same filtering capabilities as the described GUI using HTTP GET protocol[8].

The second community where Bolegweb would like to contribute is represented by users facilitating geospatial web technologies mostly used in SDI implementation. The CSW APIs are made available individually for OGC Services Types. In addition, in order to provide intuitive and user friendly GUI, the portal of OGC resources metadata is being developed on current opensource geospatial libraries (Figure 8). The portal represents a think web client application, which communicates to the CSW services managing the metadata collected from the crawler.

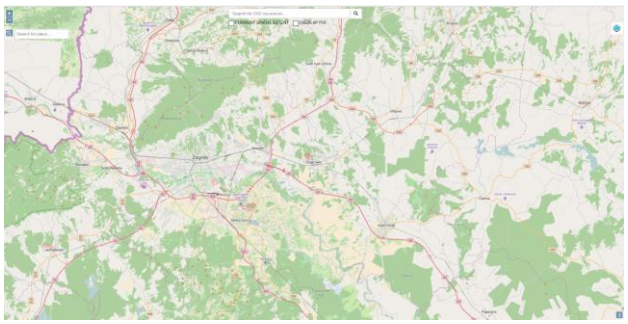

Figure 8. GUI of Bolegweb Geospatial Portal of OGC Resources[9].

The Geospatial portal provides GUI to search metadata of OGC resources discovered MSECS. Simple full text and spatial queries are supported by defining the point of interest location on the map defining by double click) and map extend (bounding box) are provided. This pilot is in the development phase and foresees to provide more advanced functionalities in the future for consuming the geospatial data discovered on the Web.

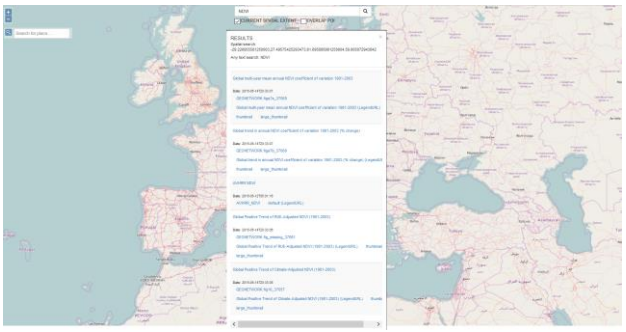

Figure 9. GUI of Bolegweb Geospatial Portal of OGC Resources – sample user query searching for "NDVI" data overlapping spatially overlapping the bounding box defined by the map window

Sample user scenario combining thematic (e.g. Normalized Difference Vegetation Index - NDVI) and spatial query is represented by results presented in Figure 9. Another way of using spatial queries is by enabling the option "OVERLAY POI", which ensures the user will get results overlapping the POI identified by clicking the map.

Third group of users identified by the Bolegweb project as community that might be attracted by the resources made available is the Open Data Community. For that reason, one instance of CKAN[10] data portal was deployed and being populated by the metadata collected from the Bolegweb CSWs. Collected information are categorized in groups corresponding the OGC service type and related resource type. Free text, spatial (BBOX definition) and faceted search is available (Figure 10).

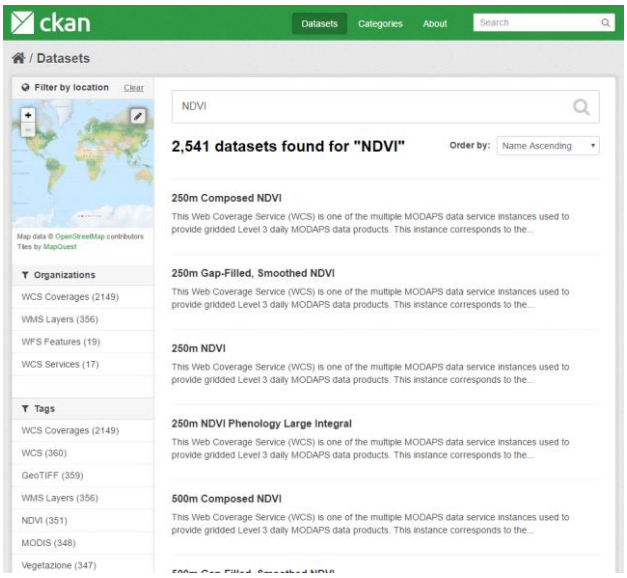

Figure 10. GUI of CKAN instance deployed as pilot data portal providing access to Bolegweb resources for Open Data Community.

[7] http://bolegweb.geof.unizg.hr/ogcwxs/html/

[8] http://bolegweb.geof.unizg.hr/ogcwxs/rest/json.php?type=wfs&location=Czech&status=1

[9] http://bolegweb.geof.unizg.hr/ogcwxs/portal/

[10] http://bolegweb.geof.unizg.hr:5000/

2.5 Future steps

The Bolegweb project will end up in October 2016. As reported in the paper, most of the system component are in the development and pilot phase. The major future achievements are to develop further Geospatial Portal of OGC Resources with visualization, download and processing functionalities, widgets, trying to use available OGC Services as much as possible.

3. CONCLUSION

The availability of GI resources on the Web is increasing on a day by day basis. The progress could be clearly seen on many geoportals around the world (e.g. on INSPIRE geoportal). This is mainly driven through overall SDI developments. On the other hand, also the number of interested users is following it. It brings us new ideas and new added value applications. However, for a layman it is not easy to understand, search, find and efficiently use GI resources. The most convenient way for them to search for something on the web is through the common search engines. The results of it will be that most of GI resources published in standardised SDI environments will not be found. This fact triggered many researchers to try to find best solution of how to solve this issue.

This paper describes the idea and products of the Bolegweb project entitled "Metasearch Enhanced OGC Crawler". It provides currently a tabular view of the various services available online via OGC standards like the Web Feature Service, Web Map Tile Service etc. collectively known as WxS services. Results achieved so far already provide a solid basis of information about available GI resources around the world. On the Bolegweb page it is possible to get information about currently available OGC Web Services that are discovered on Google Search Engine and their geographic distribution together with basic statistics. It is possible to filter results by different attributes (e.g. title, type of service, version, server location etc.). In order to provide intuitive and user friendly GUI, the portal of OGC resources metadata is being developed. Simple full text and spatial queries are supported by defining the point of interest location and map extend. This is already operating but still in the development phase and foresees to provide more advanced functionalities in the future. The contribution to open data community is also ongoing. CKAN instance has been deployed as a pilot data portal providing access to Bolegweb resources.

A likely outcome of the OGC/W3C Spatial Data on the Web working group is the definition of how to wrap those services in software that automatically creates human and machine readable Web pages for each of the items behind the service. This would mean that not just the service itself, but also the data behind would be discoverable and thus enabling tools such as the Metasearch Enhanced OGC Crawler to provide a much richer information coverage and user experience.

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7 2007-2013) under grant agreement n° 291823 Marie Curie FP7-PEOPLE-2011-COFUND (The new International Fellowship Mobility Programme for Experienced Researchers in Croatia - NEWFELPRO). This article has been written as a part of a project " Borderless Geospatial Web (Bolegweb) which has received funding through NEWFELPRO project under grant agreement n° 3.

REFERENCES

Abargues, C., Granell, C., Díaz, L., Huerta, J., Beltran, A., 2009. Discovery of User-Generated Geographic Data Using Web Search Engines, Advances in Geoscience and Remote Sensing, Gary Jedlovec (Ed.), InTech.

Athanasiou, S., Georgomanolis, N., Patroumpas, K., Alexakis, M., & Stratiotis, T. 2015. TripleGeo-CSW: A Middleware for Exposing Geospatial Catalogue Services on the Semantic Web. In EDBT/ICDT Workshops (Vol. 1330, pp. 229-236).

Díaz, L., Granell, C., Huerta, J., Eds. 2012. Discovery of Geospatial Resources: Methodologies, Technologies and Emergent Applications. IGI Global. ISBN 978-1-4666-0945-7.

Florczyk, A. J., 2012. Search improvement within the geospatial web in the context of spatial data infrastructures Doctoral dissertation, Universidad de Zaragoza.

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, 2012, Vol.7, 151-172.

INSPIRE, 2007. EU Directive. Official Journal of the European Union, L 108/1, 50.

Kliment, T., Cetl, V., Kliment, M., Tuchyňa, M., 2015. Making more OGC services available on the web discoverable for the SDI community. In Proceedings of the 15th International Multidisciplinary Scientific GeoConference. 16-25 June 2015, Albena, Bulgaria.

Kliment, T., Granell, C., Cetl, V., Kliment, M., 2013a. Publishing OGC resources discovered on the mainstream web in an SDI catalogue. In *Proceedings of the 16th AGILE International Conference on Geographic Information Science*. 14-17 May 2013, Leuven, Belgium.

Kliment, T., Cetl, V., Tuchyňa, M., 2013b. Discovery of geospatial information resources on the web. In *Proceedings of SDI Days 2013*. 26-27 September 2013, Šibenik, Croatia.

Kliment, T., 2012. Geospatial information re- sources discovery on the Internet. PhD thesis, Bratislava: Slovak University of Technology in Bratislava, Faculty of Civil Engineering, Department of Theoretical Geodesy [in Slovak].

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.

Poorazizi, M. E., Hunter, A. J. S., Steiniger, S., 2015. A Volunteered Geographic Information Framework to Enable Bottom-Up Disaster Management Platforms. ISPRS International Journal of Geo-Information. 2015, 4, 1389-1422.

Sibolla, B., van Zyl, T., McFerren, G., Hohls, D., 2014. Adding temporal data enhancements to the advanced spatial data infrastructure platform. *Tenth International Conference of the African Association of Remote Sensing of the Environment (AARSE 2014)*, University of Johannesburg, South Africa, 27 - 31 October 2014.