CSW2LD: a Linked Data frontend for CSW

Francisco J. Lopez-Pellicer, Aneta J. Florczyk, Walter Rentería-Aguaviva,
Javier Nogueras-Iso, and Pedro R. Muro-Medrano.

Computer Science and Systems Engineering Department
Universidad de Zaragoza
{fjlopez,florczyk,walterra,jnog,prmuro}@unizar.es

Abstract

Catalogues are at the core of Spatial Data Infrastructures. The most noticeable catalogue specification in the geospatial domain is the *Catalogue Services* specification issued by the Open Geospatial Consortium. That specification defines a set of abstract interfaces for the discovery, access, maintenance and organization of metadata repositories of geospatial information in distributed computing scenarios, such as the Web. In particular, the specification defines a HTTP protocol binding named *Catalogue Service for the Web* or CSW. The standard is complex and requires specialized clients. The Web of Data is an alternative approach to the publication of data and metadata on the Web that enables the use of generic clients and the mix of information from different domains. This paper proposes the CSW2LD Server as a solution for publishing on the Web as Linked Data the content of metadata repositories offered through the CSW protocol.

Keywords: CSW, Linked Data, Wrapper, Metadata.

1 Introduction

Geospatial data and services, such as satellite imagery, transit networks, census data, and mapping services, developed for internal use can often be reused in external applications once they are published on the Web with open licenses. Geospatial catalogues are a usual means to publish descriptions of geospatial holdings for enabling their discovery and access by potential users [1]. These systems use indexed and searchable metadata against which users can query on geospatial resources. Metadata are information and documentation about data intended to be understood, shared and exploited effectively by all users over time. The *Open Geospatial Consortium* (OGC) *Catalogue Service* specification [2]

proposes an abstract model for the discovery and retrieval of metadata, and standardizes its bindings to different platforms. The OGC has led the development of open and standardized Web service interface specifications for accessing geospatial information since 1994. The OGC is an international voluntary consensus standards organization with more than 400 members including companies, government agencies and universities. In this regard, the OGC is similar to other consensus standards organizations, such as W3C and OASIS.

The *Catalogue Service for the Web* (CSW) [2] is the HTTP protocol binding of the abstract model defined in the *Catalogue Service* specification. The CSW protocol is considered a key element for the development of *Spatial Data Infrastructures* (SDI) [3]. The term SDI often denotes a consensus initiative in a community materialized as a collection of technologies, policies and institutional arrangements that facilitate the access to spatial data [1]. Lopez-Pellicer et al. [4, 5] provides an estimation of the spread of the CSW protocol: 15 CSW servers in Spain and 45 CSW servers in Europe related with SDI initiatives were found during a Web crawl for OGC Web services.

This paper describes the CSW2LD Server, a Linked Data server able to publish the content of CSW-based geospatial catalogues. Setting up this server is not trivial due to the characteristics of the CSW protocol. In addition, the RDF data published by the CSW2LD Server provide only a glimpse of the information stored in the catalogue. In order to clarify the semantics of published resources, the CSW2LD Server uses the metadata provided by wrapped CSW servers in its original format as one of the representations available for the Linked Data best practice *303 URIs forwarding to different documents* [6].

The structure of this paper is as follows. Section 2 identifies related work. Section 3 introduces the CSW protocol. Section 4 presents the design of the CSW2LD Server and its prototype. Finally, the conclusion reviews the ideas presented and sets the next research goals.

2 Related work

The Linked Data community has shown interest in easing the access to large amounts of data available on the Web through catalogues. The approaches found in the literature can be classified as *curated data*, *third-party data conversion*, *consensus standard vocabularies* and *protocol wrappers*. *Curated data* identifies the publication of catalogues interwoven with the Web of Linked Data by their

owners or curators. An example is the linked open data service provided by German National Library [7]. *Third-party data conversion* retrieves raw data found in catalogues, and then applies a conversion into the RDF data model. A case is the Data-gov Wiki site [8], which host linked government data retrieved mostly from the US government's data.gov catalogue. *Consensus standard vocabularies* are vocabularies developed with the participation of stakeholders whose aim is to get catalogue operators to publish as Linked Data their metadata. The Data Catalog Vocabulary, which is intended for Government Data catalogues [9], illustrates this strategy. Finally, a *protocol wrapper* is a component that enables the integration of a standard catalogue service with the Linked Data cloud. The OAI2LOD Server [10], which wraps the protocol OAI-PMH used in digital libraries, is a renowned example.

The geospatial community is an early adopter of Semantic Web technologies. Egenhofer [11] proposed in 2002 the use of the Semantic Web to face problems of semantic heterogeneity in geo-resource discovery and interoperability. Nowadays, geographical information producers, such as Ordnance Survey [12], are investigating how Linked Data technologies can improve the diffusion of geographic data. The Linked Data community has also interest in publishing geospatial datasets. For instance, *GeoLinked Data (.es)* publishes official geospatial datasets linked with official statistical data [13]. Schade et al. [14] analysed the use of Linked Data in SDIs using different technological approaches. In particular, these authors suggest that CSW servers may return RDF data and, even, behave as Linked Data servers. Their work conclude that now it is at the time to develop prototypes for being tested in a Linked Data augmented SDI scenario. In this regard, the CSW2LD server could be considered as one of these prototypes.

3 The CSW protocol

This section gives an introduction to the abstract model defined in the OGC *Catalogue Service* specification, presents the main features of its HTTP binding, the *Catalogue Service for the Web* (CSW), and shows some issues hindering its use for the publication of metadata on the Web that a Linked Data protocol wrapper, such as the CSW2LD Server, could improve.

3.1 Abstract catalogue model

The OGC *Catalogue Service* specification [2] defines the interaction between a catalogue client and a catalogue server using concepts that can be assumed as shared by the geospatial community. The specification defines a set of abstract interfaces that includes interfaces for retrieving the metadata of a catalogue server (*OGC_Service interface*) and client discovery of resources registered in a catalogue (*Discovery interface*). In addition, this specification defines a set of core queryable and returnable attributes (names, definitions, conceptual datatypes) derived from the *Dublin Core Metadata Initiative* (DCMI) Element Set (as defined in February 2003 [15]) and a query language called CQL similar to SQL.

3.2 HTTP binding

The OGC *Catalogue Service* specification also defines the CSW protocol. This protocol standardizes the HTTP binding of the abstract interfaces *OGC_Service* and *Discovery* among others, accepts GET and POST requests, and may return metadata documents using different output schemas, mainly based in XML. The CSW operations *GetRecords* and *GetRecordsById* implement respectively the abstract operations *query* and *present* defined in the *Discovery* interface. The *query* operation searches the catalogued metadata and produces a result set containing URIs that reference to all the resources that satisfy the query. This result set may expire immediately. The query operation returns an estimate of the records matched and allows start position and max records parameters. This operation optionally returns metadata for some or all of the found result set. The *present* operation returns selected metadata for resources identified by their URIs. In addition, the operation *GetCapabilities* defined in the abstract *OGC_Service* interface returns a description of the server capabilities. Figure 1 shows an example of a CSW conversation between a server and a client.

3.3 Issues

The CSW protocol provides a solution for the dissemination of metadata within the geospatial community via the HTTP protocol. That is, the CSW protocol is not intended for the publication of geospatial metadata on the Web. Indeed, the design of the CSW protocol is not consistent with Web semantics (as described in Fielding and Taylor [16]). For example, the support of HTTP POST *GetRecords* requests is mandatory meanwhile the support of HTTP GET *GetRecords* requests is optional. In Web semantics, GET should be mandatory and the use of POST is discouraged in such a query. In addition, some URIs found in the responses are URN, URI

fragments, or local to the catalogue. It is not mandatory the use of dereferenceable URIs for the identification of metadata records in CSW responses. Hence, only domain clients may know how to use these URIs to compose a request to the appropriate CSW server for additional information. Finally, the CSW server only returns representations conforming to the CSW XML schema and representation standards well known in the geographic domain (e.g. ISO 19139 geographic metadata formats [17]). Therefore, only domain clients can understand these responses.

Request:

```
GET/csw/servlet/cswservlet?request=GetRecordById&id=
ESIGNMAPASRELIEVESERIE200701180000&elementSetName
full&outputSchema=http://www.opengis.net/cat/csw/
2.0.2 HTTP/1.1
Host: www.idee.es
```

Response:

```
HTTP/1.x 200 OK
Content-Type: application/xml;charset=ISO-8859-1
...

<?xml version = '1.0' encoding = 'ISO-8859-1'?>
<GetRecordByIdResponse
 xmlns="http://www.opengis.net/cat/csw/2.0.2"
 xmlns:dc="http://purl.org/dc/elements/1.1/"
 xmlns:dcterms="http://purl.org/dc/terms/"
 ...>
 <SummaryRecord>
  <dc:title>Mapas en Relieve</dc:title>
  <dc:identifier>
   ESIGNMAPASRELIEVESERIE200701180000
  </dc:identifier>
  ...
  <dc:format>PVC</dc:format>
  <dc:subject>elevation</dc:subject>
  <dc:subject>imageryBaseMapsEarthCover</dc:subject>
  <dc:type>dataset</dc:type>
  ...
  <dcterms:spatial>COUNTRIES.SPAIN</dcterms:spatial>
  ...
  <dcterms:spatial>
   northlimit=43.8;
   southlimit=37.83;
   westlimit=-9.32;
   eastlimit=0.72;
  </dcterms:spatial>
 </SummaryRecord>
</GetRecordByIdResponse>
```

Figure 1. Sample CSW GetRecordById request and response.

4 The CSW2LD server

The CSW2LD Server is a protocol wrapper that gives access to CSW-based servers as if they were Linked Data sources. It allows geospatial catalogue curators to publish the descriptions of their assets in the Web of Data, and Linked Data practitioners to wrap third-party geospatial catalogues.

4.1 Published concepts

The concepts *CatalogueService*, *RecordList*, *MetadataRecord* and *Resource* are the things published by the CSW2LD server that should have URIs. The concept *CatalogueService* describes a wrapped CSW server. The description about a *CatalogueService* is grounded in the *GetCapabilities* response. Each *RecordList* describes a query result from a *GetRecords* request. The concepts *MetadataRecord* and *Resource* allow decoupling *GetRecordById* responses in information about the metadata record and information about the resource described.

The current version of the CSW2LD server only assigns HTTP URIs to *CatalogueService*, *RecordList* and *MetadataRecord* instances. As design decision, the representation of *MetadataRecord* instances piggybacks the description of the associated *Resource* instances. Table 1 shows the default wrapping patterns. The variable *{service}* should match with to the identifier given to a wrapped CSW server in the configuration. The variable *{?keys+}* represents a map of parameters which is expanded in the URI as a query. The current version of the CSW2LD server supports the parameters start and max. That is, a *RecordList* represents in the current version an ordered subset of the available metadata records. The variable *{id}* is interpreted as the identifier of a metadata record in a wrapped CSW server.

Concept	***URI Template***	***Wraps***	***Example***
Catalogue Service	/resource/{service}	*GetCapabilities*	/resource/idee
Record List	/resource/{service}/ records{?keys+}	*GetRecords*	/resource/idee/records? start=50&max=15
Metadata Record	/resource/{service}/ record/{id}	*GetRecordById*	/resource/idee/record/ 100000_full1038_es

Table 1. The CSW2LD Server wrapping patterns.

4.2 Vocabularies

The publishing strategy of the CSW2LD Server requires the introduction of a specific RDFS vocabulary that provides a definition for the concepts

CatalogueService, *RecordList*, *MetadataRecord* and their relations. Figure 2 depicts part of this vocabulary: the classes *CatalogueService*, *RecordList*, *MetadataRecord* and the properties *query*, *present* and *about*. The class *RecordList* represents an ordered collection and is a subclass of the RDFS class *rdf:List*. The properties *query* and *present* relates record lists and metadata records respectively with a catalogue service. Instances of the class *CatalogueService* should have at least a query pointing to a *RecordList* resource that represents a *GetRecords* request with the default values start position 1 and max values 10. The property *about* ties a metadata record with the resource described.

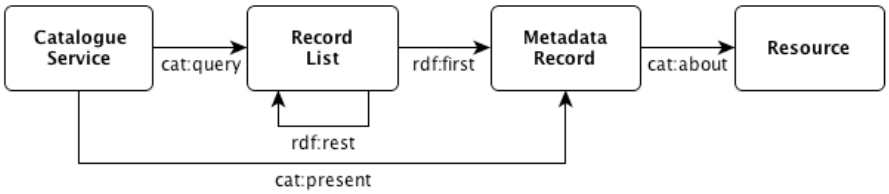

Figure 2. CSW2LD Server companion vocabulary.

The class *RecordList* requires additional caveats. The first element of the list should identify a *MetadataRecord* resource. The rest-of-list element may point to a *RecordList* resource that begins with the next record or to the resource *rdf:nil* if all records have been returned. The information about the next record can be found in the *GetRecords* response. In addition, two consecutive identical *GetRecords* requests may return different result sets because the result set of a *GetRecords* request may expire immediately. Therefore, the contents of two consecutives dereferences of a URI that identifies a *RecordList* resource may differ. The CSW2LD Server solves this issue by associating the dereferenced *RecordList* with a blank node of type *RecordList* that contains the results using the DCMI property *hasVersion* (see Figure 3). The concepts of the above vocabulary are currently published within the dereferenceable namespace http://iaaa.cps.unizar.es/vocs/cat# and identified with the prefix *cat*.

The use of the DCMI vocabulary is pervasive in the descriptions returned. Usually, the CSW2LD Server asks the CSW server for representations of the metadata records conforming to the core returnable attributes of the *Catalogue Services* specification, which are based in the DCMI vocabulary. The CSW2LD Server asks for richer representations in other representation schemas if a crosswalk of that representation schema to the DCMI vocabulary has been registered. Additionally, each description includes information, such as its provenance, when was retrieved

from the CSW server, and, if the CSW server accepts HTTP GET *GetRecordById* requests, a link to a raw version of the metadata record.

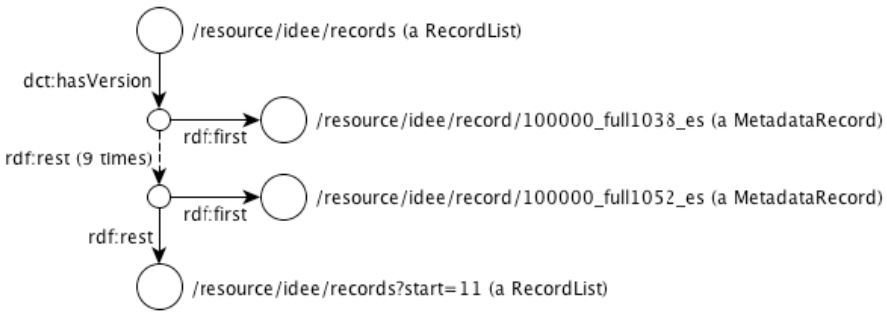

Figure 3. Description of a RecordList.

4.3 Content negotiation

Given a resource part of the underlying information model of a catalogue, it is possible to identify using HTTP URIs: the resource itself, a HTML representation describing the resource that can be consumed by humans and crawlers employed by search engines, a RDF representation describing the resource that can be consumed by semantic clients [20] and crawlers employed by semantic search engines [21], and a domain representation, usually in XML, describing the resource that can be consumed by CSW-aware clients. As the XML representation of the resource may differ substantially from the RDF and HTML representations, the approach for the content negotiation is the Linked Data best practice *303 URIs forwarding to different documents* [6] (see Figure 4): a 303 status code and a *Location* HTTP header that points to the appropriate representation.

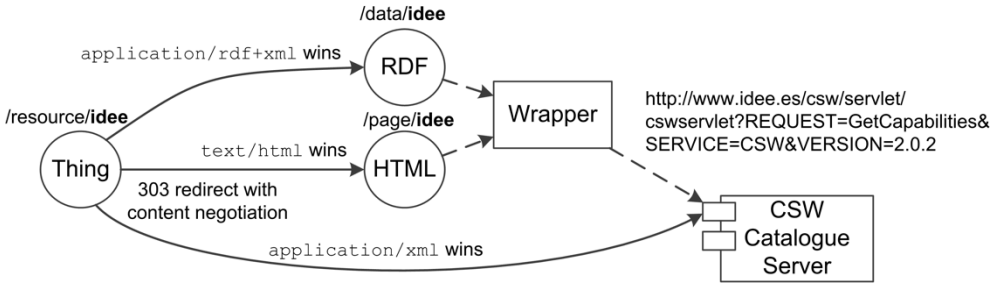

Figure 4. The 303 URI solution extended with forwarding to CSW requests.

4.4 Prototype implementation

The CSW2LD Server (figure 5) is a Web application implemented in Java compatible with Tomcat and Jetty servlet containers. The publishing front-end and the 303 URIs forwarding with content negotiation is based on the Linked Data frontend Pubby. The current version uses a Jena TDB triple store as storage. The requested resources can be considered as *cached*, *missed*, *expired* and *non-cacheable*. A resource is *cached* if the storage contains information derived from a CSW request and the data is not stale. The resource is considered *missed* if the storage does not contain information data about the resource, and then, a CSW client makes a request to the CSW server for details about the resource. This is the default behaviour for metadata record data (*GetRecordById*). Cached data may become stale or *expired* after a given date, triggering a CSW request for fresh data. This is the default behaviour for catalogue service data (*GetCapabilities*). Finally, some resources can be considered as *non-cacheable*. This is the default behaviour for record list data (*GetRecords*). The output schema requested by default to the CSW server is the defined in the Catalogue Services specification. However, if the CSW server supports ISO 19139 representations, these are requested instead, and mapped to DCMI using the CWA 14857 [22] crosswalk as reference.

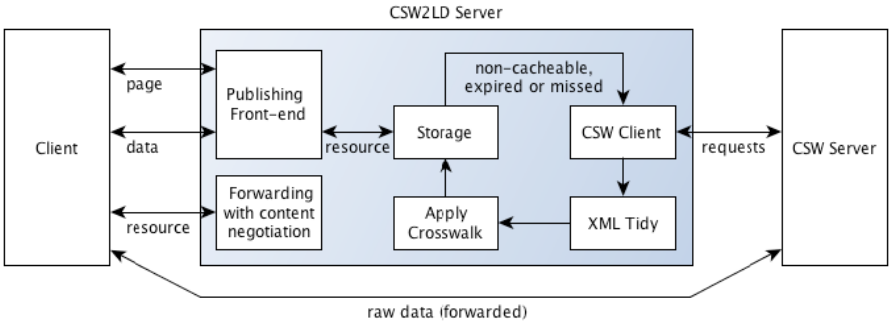

Figure 5. The CSW2LD Server architecture.

5 Conclusions

This paper presents the CSW2LD Server: a software component that republishes according to the Linked Data metadata from repositories accessible through CSW. Applied to SDI metadata catalogues, the CSW2LD Server exposes the description

of SDI assets as dereferenceable Web resources, and allows search engines to index them. Future versions of the CSW2LD Server should include additional technical features, such as additional crosswalks, and functional features, such as the generation of links between the metadata and existing thesauri and ontologies, and augment the meta-metadata available about the provenance and quality of the exposed information.

Acknowledgements. This work has been partially supported by Spanish Government (projects "España Virtual" ref. CENIT 2008-1030 and TIN2009-10971), the National Geographic Institute (IGN) of Spain and GeoSpatiumLab S.L. The work of Aneta Jadwiga Florczyk has been partially supported by a grant (ref. AP2007-03275) from the Spanish Government. The work of Walter Rentería-Agualimpia has been partially supported by a grant (ref. B181/11) from the Aragon Government.

References

[1] Douglas D. Nebert (Ed.): Developing Spatial Data Infrastructures: The SDI Cookbook v.2.0. Technical report, Global Spatial Data Infrastructure (2004).
[2] D. Nebert, A. Whiteside, and P. Vretanos: Open GIS Catalogue Services Specification, version 2.0.2.. OpenGIS Publicy Available Standard OGC-07-006r1, Open GIS Consortium Inc., (2007).
[3] J. Nogueras-Iso, F.J. Zarazaga-Soria, R. Béjar, P.J. Álvarez, and P.R. Muro-Medrano: OGC Catalog Services: a key element for the development of SDIs. Computers & Geosciences, 31(2):199-209, (2005).
[4] F. J. Lopez-Pellicer, R. Béjar, A. J. Florczyk, P. R. Muro-Medrano, and F. J. Zarazaga-Soria: State of Play of OGC Web Services across the Web. In INSPIRE Conference 2010: INSPIRE as a framework for cooperation. Krakow, Poland, 22-25-June 2010, (2010).
[5] F. J. Lopez-Pellicer, A. J. Florczyk, R. Béjar, J. Nogueras-Iso, F. J. Zarazaga-Soria, and P. R. Muro-Medrano: State of Play: Spain and Portugal. SDI services? state of play in autumn 2010. In I Jornadas Ibéricas de Infraestructuras de Datos Espaciales (JIIDE'2010), Lisboa, (2010).
[6] L. Sauermann and R. Cyganiak: Cool URIs for the Semantic Web. W3C Interest Group Note, December (2008).

[7] Jan Hannemann and Jürgen Kett: Linked Data for Libraries. In World Library and Information Congress: 76th IFLA general conference and assembly, 10-15 August 2010, Gothenburg, Sweden. IFLA, (2010).
[8] Li Ding, Dominic DiFranzo, Alvaro Graves, James R. Michaelis, Xian Li, Deborah L. McGuinness, and Jim Hendler: Data-gov Wiki: Towards Linking Government Data. In Proceedings of the 2010 AAAI Spring Symposium on Linked Data Meets Artificial Intelligence, (2010).
[9] Fadi Maali, Richard Cyganiak, and Vassilios Peristeras: Enabling interoperability of government data catalogues. In Maria Wimmer, Jean-Loup Chappelet, Marijn Janssen, and Hans Scholl, editors, Electronic Government, LNCS 6228, pp. 339-350, (2010).
[10] Bernhard Haslhofer and Bernhard Schandl: Interweaving OAI-PMH data sources with the linked data cloud. International Journal of Metadata, Semantics and Ontologies, 5(1):17-31, (2010).
[11] M. J. Egenhofer: Toward the semantic geospatial web. In GIS '02: Proceedings of the 10th ACM international symposium on Advances in geographic information systems, pages 1-4, McLean, Virginia, USA, (2002).
[12] John Goodwin, Catherine Dolbear, and Glen Hart: Geographical Linked Data: The Administrative Geography of Great Britain on the Semantic Web. Transactions in GIS, 12(s1):19-30, (2008).
[13]Luis M. Vilches-Blázquez, Boris Villazón-Terrazas, Victor Saquicela, Alexander de León, Oscar Corcho, and Asunción Gómez-Pérez: Geolinked data and inspire through an application case. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS '10, pg. 446-449, New York, NY, USA, (2010).
[14] Sven Schade, Carlos Granell, and Laura Diaz: Augmenting SDI with Linked Data. In Workshop On Linked Spatiotemporal Data 2010 (LSTD-2010), CEUR workshop proceedings, 691(2010).
[15] ISO 15836:2003(E) - Information and documentation - The Dublin Core metadata element set. International Standard, International Organization for Standardization (ISO), (2003). Replaced by ISO 15836:2009.
[16] R.T. Fielding and R.N. Taylor: Principled design of the modern web architecture. ACM Trans. on Internet Tech. (TOIT), 2(2):115-150, (2002).
[17] W. Kresse and K. Fadaie: ISO Standards for Geographic Information. Springer, Berlin, (2004).
[18] F. J. Lopez-Pellicer, A. J. Florczyk, J. Nogueras-Iso, P. R. Muro-Medrano, and F. J. Zarazaga-Soria: Exposing CSW Catalogues as Linked Data. In Geospatial Thinking, LNGC, pp. 183-200. (2010).
[19] T. Berners-Lee: Linked Data — Design Issues, (2006).

[20] T. Berners-Lee, Y. Chen, L. Chilton, D. Connolly, R. Dhanaraj, J. Hollenbach, A. Lerer, and D. Sheets: Tabulator: Exploring and Analyzing linked data on the Semantic Web. In Proc. of the 3rd Int. Sem. Web User Interaction, (2006).
[21] G. Tummarello, E. Oren, and R. Delbru: Sindice.com: Weaving the open linked data. In Proc. of the 6th Int. Sem. Web Conf. and 2nd Asian Sem. Web Conf. (ISWC/ASWC2007), LNCS 4825, pp 547-560, (2007).
[22] WS/MMI-DC. CWA 14857:2003: Mapping between Dublin Core and ISO 19115, Geographic Information — Metadata. CEN/ISSS Workshop on Metadata for Multimedia Inf. - Dublin Core, (2003).