Towards Linked Data Conventions for Delivery of Environmental Data Using netCDF

Jonathan Yu, Nicholas Car, Adam Leadbetter, Bruce Simons, Simon Cox

Towards Linked Data Conventions for Delivery of Environmental Data Using netCDF

Jonathan Yu[1], Nicholas Car[1], Adam Leadbetter[2], Bruce A. Simons[1], and Simon J.D. Cox [1]

[1]Land & Water Flagship: CSIRO, Highett and Dutton Park Labs, Australia
[2]British Oceanographic Data Centre, Liverpool, United Kingdom

{jonathan.yu, nicholas.car, bruce.simons,
simon.cox}@csiro.au
alead@bodc.ac.uk

Abstract. netCDF is a well-known and widely used format to exchange array-oriented scientific data such as grids and time-series. We describe a new convention for encoding netCDF based on Linked Data principles called netCDF-LD. netCDF-LD allows metadata elements, given as string values in current netCDF files, to be given as Linked Data objects. netCDF-LD allows precise semantics to be used for elements and expands the type options beyond lists of controlled terms. Using Uniform Resource Identifiers (URIs) for elements allows them to refer to other Linked Data resources for their type and descriptions. This enables improved data discovery through a generic mechanism for element type identification and adds element type expandability to new Linked Data resources as they become available. By following patterns already established for extending existing formats, netCDF-LD applications can take advantage of existing software for processing Linked Data and supporting more effective data discovery and integration across systems.

Keywords: netCDF, linked data, data discovery, environmental data

1 Introduction

Scientific data is increasingly being made available from environmental agencies, universities, research organisations and government departments via web services. There is, however, a need for better integration and discoverability across datasets from the various services. The Network Common Data Form (netCDF) is a suite of software libraries and a data format for producing array-oriented scientific data, which is commonly used to exchange environmental data organized as images, grids and time-series[1]. netCDF has been developed and maintained by Unidata, which is part of the University Corporation for Atmospheric Research (UCAR) funded by the

[1] http://www.unidata.ucar.edu/software/netcdf

adfa, p. 1, 2011.

United States National Science Foundation. The Open Geospatial Consortium (OGC) has adopted netCDF and formalized it as an implementation standard to support encoding of geospatial information to communicate and store multi-dimensional data[2]. netCDF datasets are typically published through service interfaces using Thematic Real-time Environmental Distributed Data Services (THREDDS), which allows users to find and access the data, and use them without necessarily downloading the entire file. The THREDDS Data Server (TDS) is a web service to access catalogues, metadata and data, and allows sub-sampling of the data using OpenDAP, OGC Web Coverage Service (WCS), OGC Web Map Service (WMS), and netCDF subset service interfaces. The netCDF format is not proprietary and THREDDS and TDS have been made available online at the Unidata website and via their GitHub repository (https://github.com/Unidata/thredds). As such, this technology stack has been widely used to publish large datasets and provide the ability to efficiently query subsets of the data.

netCDF files are designed to be 'self-contained', through headers that describe the structure and content of the dataset. Standardization is based on a community 'convention' for a set of vocabulary names, with the Climate and Forecasting (CF) conventions [1] most widely known. However, these names are only text values, and often conflate various concerns. These pose a key challenge in providing automatic semantic mediation between datasets to compare equivalent parameters by its conceptual meaning. A number of controlled vocabularies and ontologies currently enable the definition, publication and access of these semantics on the web [2–6], however, there is no facility in the netCDF headers to link to and reference externally defined terms and semantics. Although, the CF community and other communities publish standard names, the publishing of these names can be a lengthy process involving review of new proposed names by the respective committee, and some names may be duplicated and have different meaning or be incomplete [7]. Figure 1 gives an example of the current practice where applications request data using multiple names that may refer to the same concept or set of vocabulary terms.

The World Wide Web Consortium (W3C) has developed principles for publishing data on the Web. Known as "Linked Data" [8], the primary feature is the use of HTTP Uniform Resource Identifiers (URIs) to link between any information resources (data or metadata), including non-hypertext formats, so that more information can be discovered. The Linked Data principles have stimulated the design of a lightweight extension to JSON [9] which is a commonly-used data format in browser applications. JSON-LD (for "Linked Data") annotates JSON data with URIs, allowing JSON data to be interpreted as Linked Data with minimal additions to the original structure. The JSON-LD pattern has now been proposed for annotation of tabular data formatted as Comma-Separated-Values or similar in the W3C activity 'CSV on the Web'[10].

Thus, extending netCDF by borrowing and adopting the *-LD pattern from JSON-LD and 'CSV on the Web' is a natural extension in order to link and reference existing controlled vocabularies and terms via the Semantic Web. In this paper, we describe initial work in defining netCDF-LD conventions, which allow netCDF to be

[2] http://www.opengeospatial.org/standards/netcdf

interpreted as Linked Data with non-intrusive annotations. The netCDF-LD conventions allow the use of URIs to define the identity of a netCDF file, attribute names and attribute values. The aim is to enhance existing netCDF datasets with metadata that links to standardized vocabulary terms and to express precise semantics about the data so that it can be better discovered, integrated and used. By following patterns already established for extending existing formats, netCDF-LD applications will be able to take advantage of existing software for processing Linked Data; existing controlled vocabularies describing environmental domains published as Linked Data; and allow data, such as environmental data, to be more easily discovered, interpreted and integrated into applications, such as environmental models (see Fig. 2 for an example).

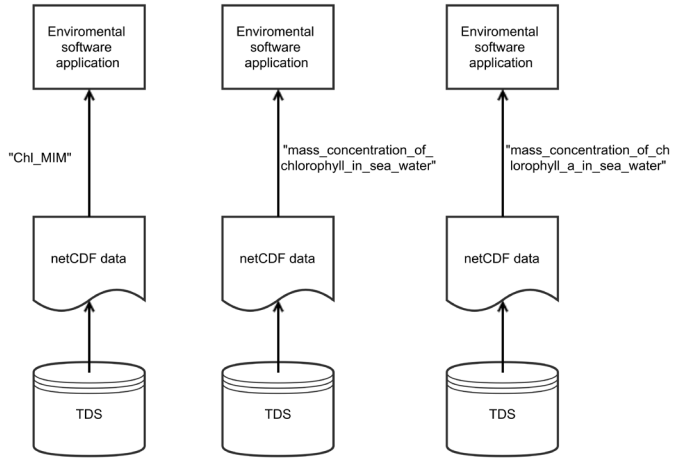

Fig. 1. Current practice of applications consuming netCDF data via TDS services

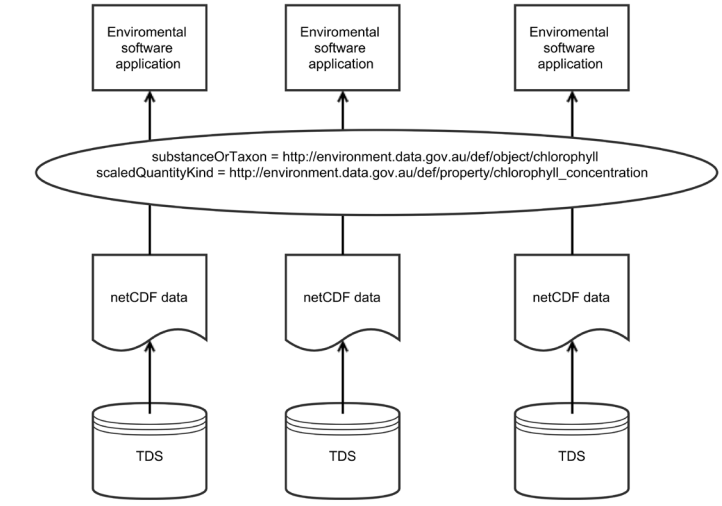

Fig. 2. Proposed data integration example consuming netCDF-LD

2 Linked data and JSON-LD

Linked Data is a methodology for publishing data and metadata in a structured format so that links may be created and exploited between objects. The key enabling components are URIs, HTTP, the Resource Description Framework (RDF) and the SPARQL Protocol and RDF Query Language (SPARQL) [8]. RDF is the component which provides a data model for describing subject and their relationships to other objects via predicates thus forming 'triples'. RDF data is encoded in a number of serializations including RDF/XML, N-Triples, Turtle, and RDF/JSON. RDF data can be loaded into repositories called 'triple stores' and queried using the SPARQL language. Tooling has also been developed to exploit Linked Data as RDF such as Linked Data browsers and mashups, semantic search engines, and data extraction tools. To date, there are around 700 million triples across about 1,800 vocabularies discoverable in publicly available linked data services [11].

2.1 JSON-LD

JSON-LD is a W3C recommendation and is a format that adds URIs to JSON data, allowing JSON data to be interpreted as Linked Data by supporting links from JSON data to other data [12]. JSON-LD provides an interchange language for libraries utilizing JSON, such as client applications, web services, and NoSQL or unstructured databases (e.g. *CouchDB* and *MongoDB*).

```
{
    "@id": "http://foo.bar/linked_netCDF_example",
    "@type": [
        "http://www.w3.org/ns/prov#Entity",
        "http://www.w3.org/ns/dcat#Dataset"
    ],
    "http://data.ba.gov.au/def/ba#dataOwner": [
        {"@id": "http://data.ba.gov.au/id/person/car587"}
    ],
    "http://purl.org/dc/elements/1.1/created": [
   {"@type": "http://www.w3.org/2001/XMLSchema#dateTime",
         "@value": "2014-10-15T23:23:25+11:00"}
    ],
    "http://purl.org/dc/elements/1.1/title": [
        {"@value": "Victorian Biodiversity Atlas"}
    ]
}
```

Fig. 3. A JSON-LD object with ID ("@id") of http://data.ba.gov.au/dataset/f4107720 and type ("@type") http://www.w3.org/ns/prov#Entity & http://www.w3.org/ns/dcat#Dataset, each defined at their URIs. Other properties ("created", "title", "dataOwner") defined in well-known vocabularies such as Dublin Core or in a project-specific vocabulary.

JSON-LD introduces:

- URI identifiers for JSON objects
- contexts or namespaces to disambiguate keywords in JSON documents
- a linking mechanism
- a way to associate datatypes with values such as dates and times
- a way to express one or more graphs in a single document

Tooling has been developed for reading JSON-LD data and creating an RDF graph, as well as serializing RDF data into JSON-LD, for example the *rdfLib-jsonld* plugin for Python, *JSONLD-JAVA*, and *PHP-JSON-LD*. JSON-LD contributes to the design of other microformats, such as the extension of GeoJSON with JSON-LD, called *GeoJSON-LD*.

The precedent of JSON-LD is being followed in the W3C 'CSV on the Web' [10] project, another effort to enrich a well-known format with more rigorous definitions and semantics. 'CSV on the Web' directly reuses key features of JSON-LD to allow terms to be assigned URIs and values bound to datatypes.

3 netCDF-LD

Given this uptake and prior work, we propose to apply the same approach for netCDF, to be called netCDF-LD. In this section we outline the conventions for netCDF-LD, based on the general approach from JSON-LD, and present examples of netCDF-LD encodings and the interpreted Linked Data content. The intention is that interpreted Linked Data content encoded as RDF could be used to support richer queries for data aggregation and query.

3.1 netCDF Conventions

netCDF-LD uses the "context" array concept from JSON-LD and maps it onto global variables within the NetCDF file. This allows variable names within the netCDF file to be used as shorthand for the URI of an external (Linked Data) resource, which should provide details of the variable definition. A reserved global variable, "context-id", assigns a URI to the netCDF file itself. The suffixes "_a" and "_ref" are reserved for the assignment of an RDF class and a URI to any variable level attribute within a netCDF file. The RDF datatype may be inferred from the declaration of the datatype in the netCDF file. The "_lang" variable level suffix is reserved for occasions when a human readable language is required to be assigned to an attribute or data value.

In the following sections, the reserved global attribute names and variable level suffixes are defined, and example encodings of a generic netCDF file and a Climate and Forecast (CF) metadata conventions compliant netCDF file are shown.

The "context-" Global Attributes.

The JSON-LD "@context" array is modelled in netCDF-LD as a series of global attributes with the "context-" prefix. netCDF does not allow the use of the "@" symbol as the opening character of an attribute name, and the "_" character is reserved for use by system attributes only.

Boilerplate code of three (with an optional fourth) "context-x" global attributes is recommended, and as many more as necessary may be added (see Fig. 4). Optionally, the rdf:datatype property may be declared as shown below in Fig. 5. Finally, a generic vocabulary from which all attribute URIs in the netCDF-LD file are used may be assigned as shown below in Fig. 6. Any given attribute name to be used on variables may be assigned a URI as shown below in Fig. 7.

```
:context-id = "http://foo.bar/baz";
:context-a = "http://www.w3.org/1999/02
               /22-rdf-syntax-ns#type";
:context-ref = "http://www.w3.org/1999/02
                 /22-rdf-syntax-ns#resource";
```

Fig. 4. Boilerplate code for netCDF-LD using Common Data form Language (CDL) syntax. CDL provides a human readable text respresentation of netCDF data.

```
:context-datatype = "http://www.w3.org/1999/02
                      /22-rdf-syntax-ns#datatype";
```

Fig. 5. Declaration of the datatype in the boilerplate code section for netCDF-LD

```
:context-vocab = "http://def.seegrid.csiro.au
            /isotc211/iso19156/2011/observation#";
```

Fig. 6. Declaration of default vocabulary in the context block for netCDF-LD

```
:context-attribute_ref = "http://bar.foo/baz";
```

Fig. 7. Declaration of attribute URI reference in netCDF-LD

Assigning URIs to Variable Level Attributes

A netCDF variable may be defined by a URI in netCDF-LD by adding a "ref" attribute with a value equal to the URI.

```
variable:ref= "http://vocab.nerc.ac.uk
              /collection/P07/current/CFSN0600/";
```

The values of variable level attributes may be defined by URIs thus:

```
variable:unit = "Meters";
variable:unit_ref = " http://qudt.org/vocab
                                        /unit#Meter";
```

3.2 Example Encodings and Resulting RDF Graphs

The following example encodes the netCDF file example from http://www.unidata.ucar.edu/software/netcdf/docs/CDL.html using the netCDF-LD conventions. In the example, the dimensions and variables blocks near the start of the file define the respective dimensions and variables contained in the metadata headers. The next block provides the global attributes to describe boilerplate content for the contexts of the attributes as described in the previous section. LD attributes are specified next for the listed variables, e.g. lat, lon, time, and z, each having the appropriate URI references for its value. For example, 'lat:ref' allows the value to be bound to the URI 'http://vocab.nerc.ac.uk/collection/P07/current/CFSN0600/'. This is then used as the subject for the statements bound to the variable 'lat', so that the other attributes are related objects, e.g. its units, its type, and the data values. Figure 9 shows the RDF encoding from the netCDF-LD example where the URI values and data values are mapped.

```
netcdf foo {// Example netCDF specification in CDL
  dimensions:
     lat = 10, lon = 5, time = unlimited;
  variables:
     int   lat(lat), lon(lon), time(time);
     float   z(time,lat,lon), t(time,lat,lon);
     double  p(time,lat,lon);
     int     rh(time,lat,lon);

  // Global attributes
  :context-id = "http://foo.bar/linked_netCDF_example";
  :context-units = "http://qudt.org/1.1/schema/qudt#unit";
  :context-ref = "http://www.w3.org/1999/02/22-rdf-syntax-ns
                                                  #resource";
  :context-quantityKind = "http://environment.data.gov.au/def
                                  /op#ScaledQuantityKind";
  :context-dcPartOf = "http://purl.org/dc/terms/isPartOf";
  :context-a = "http://www.w3.org/1999/02/22-rdf-syntax-ns#type";
  :Conventions = "LD-1.0";

  lat:ref = "http://vocab.nerc.ac.uk/collection/P07/current
                                                  /CFSN0600/";
  lat:units = "degrees_north";
  lat:units_ref = "http://qudt.org/vocab/unit#DegreeAngle";
```

```
  lat:a = "http://environment.data.gov.au/def/op#ScaledQuantityKind";
  lat:dcPartOf = "http://foo.bar/linked_netCDF_example";
  lat:datalink = "http://www.w3.org/1999/02/22-rdf-syntax-ns#value";


  lon:ref = "http://vocab.nerc.ac.uk/collection/P07/current
                                                  /CFSN0554/";
  lon:units = "degrees_east";
  lon:units_ref = "http://qudt.org/vocab/unit#DegreeAngle";
  lon:a = "http://environment.data.gov.au/def/op#ScaledQuantityKind";
  lon:dcPartOf = "http://foo.bar/linked_netCDF_example";
  lon:datalink = "http://www.w3.org/1999/02/22-rdf-syntax-ns#value";

  time:units = "seconds";
  time:units_ref = "http://qudt.org/vocab/unit#SecondTime";
  time:a = "http://environment.data.gov.au/def/op#ScaledQuantityKind";
  time:dcPartOf = "http://foo.bar/linked_netCDF_example";

  z:units = "meters";
  z:units_ref = "http://qudt.org/vocab/unit#Meter";
  z:a = "http://environment.data.gov.au/def/op#quantityKind";
  z:dcPartOf = "http://foo.bar/linked_netCDF_example";
  z:valid_range = 0., 5000.;

  p:_FillValue = -9999.;

  rh:_FillValue = -1;

  data:
     lat   = 0, 10, 20, 30, 40, 50, 60, 70, 80, 90;
     lon   = -140, -118, -96, -84, -52; }
```

Fig. 8. Example netCDF-LD encoding

```
@prefix unit: <http://qudt.org/vocab/unit#> .
@prefix qudt: <http://qudt.org/1.1/schema/qudt#> .
@prefix op: <http://environment.data.gov.au/def/op#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<http://vocab.nerc.ac.uk/collection/P07/current/CFSN0600/>
   a op:ScaledQuantityKind;
   qudt:unit unit:DegreeAngle;
   dcterms:isPartOf <http://foo.bar/linked_netCDF_example>;
```

```
   rdf:value
      "{0, 10, 20, 30, 40, 50, 60, 70, 80, 90}"^^xsd:integer .

<http://vocab.nerc.ac.uk/collection/P07/current/CFSN0554/>
   qudt:unit unit:DegreeAngle;
   a op:ScaledQuantityKind;
   dcterms:isPartOf <http://foo.bar/linked_netCDF_example>;
   rdf:value "{-140, -118, -96, -84, -52}"^^xsd:integer.

_:z qudt:unit unit:Meter;
   a op:ScaledQuantityKind;
   dcterms:isPartOf <http://foo.bar/linked_netCDF_example>.

_:time qudt:unit unit:SecondTime;
   a op:ScaledQuantityKind;
   dcterms:isPartOf    <http://foo.bar/linked_netCDF_example> .
```

Fig. 9. RDF description based on example netCDF-LD metadata

4 Discussion and Related Work

The use of Linked Data within netCDF headers makes the headers less readable by humans but far more powerful for machine interpretation. Tools able to understand and process Linked Data can "follow their nose" in order to collect more information about resources as needed. When encountering a header or other term not understood directly, such as a new variable type, a Linked Data tool will be able to dereference its type URI and obtain further metadata about that variable. With current string-based controlled terms, if a variable is not understood it is unusable. The issue of reduced human readability can, we believe, be bypassed with client tooling. Commonly used netCDF tools such as Python's *ncdump* program can easily be extended to render data from a netCDF-LD file as readable to humans as the headers from a regular netCDF file.

The inverse may not be true: it may not be possible to convert existing netCDF files into Linked Data files with as rich semantics as a netCDF-LD file would offer with new tooling. Such tools would need to contain all of the metadata about string-based terms used in current netCDF files locally in order to reference it when needed – there is no equivalent to the "follow-your-nose" procedure for string values. This would be very hard to implement as instances of such a tool would need to be kept up-to-date with all possible netCDF terms metadata in order to be relevant. This precludes tools such as netCDF Markup Language (NcML) generators.[3]

netCDF-LD adapts related work on decorating netCDF variable names with URIs. Metadata linkages in netCDF were presented by [13]. In netCDF-U [14], the

[3] See http://www.unidata.ucar.edu/software/thredds/current/netcdf-java/ncml/ for a description of the NcML and links to generators.

:ref and *:rel* suffixes are used to link netCDF variable declarations with URIs, but only relate specifically to concepts of uncertainty. In [15], a similar pattern is used to append variable declarations to link to ontology classes in the Observable Property ontology and concepts defined in the Water Quality vocabulary. However, these are limited to *scaledQuantityKind_id*, *unit_id*, *substanceOrTaxon_id*, *procedure_id*, and *medium_id*. netCDF-LD adapts these ideas for a generic mechanism for variable annotation and aligns it with current Linked Data best practice from JSON-LD. The approach used in netCDF-LD allows the whole netCDF metadata header to be translated into RDF as presented in this paper, compared with the other approaches which only map partial aspects of the netCDF metadata header.

Translating netCDF headers into RDF allows tooling to be developed to support data discovery and dataset aggregation. Figure 10 shows how the RDF descriptions from the netCDF-LD metadata headers can be harvested into a data brokering component. The data brokering component essentially would provide automated semantic mediation between netCDF datasets. This would allow applications or end users to browse or search over the available datasets based on the identifiers of the variables. An example query could be search for all datasets with 'chlorophyll concentration'.

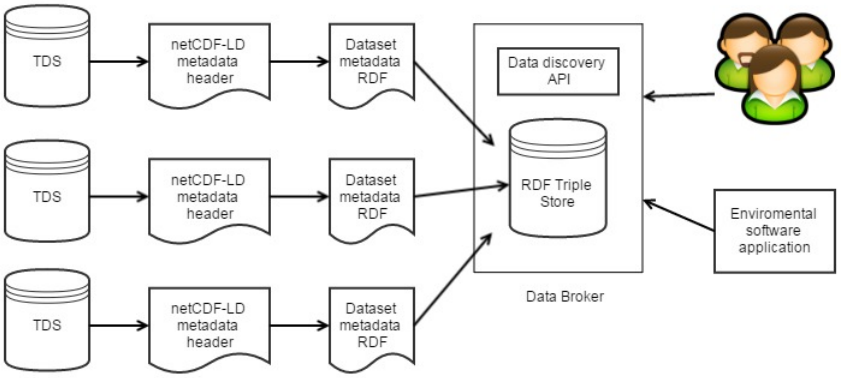

Fig. 10. Using netCDF-LD to support data discovery and dataset aggregation via a data broker

5 Conclusion and Future Work

In this paper, we have proposed initial work towards a netCDF-LD specification which applies principles from the *-LD pattern to netCDF data. The intention of netCDF-LD is to allow it to be interpreted as Linked Data with non-intrusive annotations to standard netCDF encodings. netCDF-LD conventions were presented which aim to be consistent with Linked Data design principles, allowing the use of URIs to define the identity of a netCDF file, its attribute names and the attribute values. These conventions allow existing netCDF datasets to be enhanced with structured metadata conventions for linking to standardized vocabulary terms. This allows binding to precise semantic descriptions and vocabularies which enables netCDF data exposed via

THREDDS to be better discovered, integrated and used. Also by following patterns already established for extending existing formats, netCDF-LD applications can take advantage of existing semantic web tooling for handling querying and reasoning over RDF triples translated from netCDF-LD data.

How large volumes of data is to be encoded in netCDF-LD files has not been explored. The direct serialization of array data as JSON arrays, as given in **Fig. 6**, is unlikely to be attractive to netCDF with large data volumes (potentially much greater than 2GB for netCDF4) given its verbosity compared with binary data arrays. Future work will see us test various methods for data encoding while still retaining the same netCDF-LD approach to header metadata encoding. Semantic data formats such as WaterML present a number of options that may be used.

References

1. Eaton, B., Gregory, J., Drach, B., Taylor, K., Hankin, S.: NetCDF Climate and Forecast (CF) Metadata Conventions, (2011).
2. Cox, S.J.D., Simons, B.A., Yu, J.: A harmonized vocabulary for water quality. 11th International Conference on Hydroinformatics (HIC). IWA Publishing, New York, NY, USA (2014).
3. Cox, S., Yu, J., Rankine, T.: SISSVoc: A Linked Data API for access to SKOS vocabularies. Semant. Web. (2014).
4. Caracciolo, C., Stellato, A., Morshed, A.: The agrovoc linked dataset. Semanit. Web. (2013).
5. Leadbetter, A., Lowry, R.K., Clements, O.: The NERC Vocabulary Server: Version 2.0. EGU General Assembly. p. 2943. Copernicus, Vienna, Austria (2012).
6. Summers, E., Isaac, A., Redding, C., Krech, D.: LCSH, SKOS and Linked Data. CoRR. abs/0805.2, (2008).
7. Peckham, S.D.: The CSDMS Standard Names: Cross-Domain Naming Conventions for Describing Process Models, Data Sets and Their Associated Variables. In: Ames, D.P., Quinn, N.W.T., and Rizzol, A.E. (eds.) International Environmental Modelling and Software Society (iEMSs). , San Diego, California, USA (2014).
8. Berners-Lee, T.: Linked Data - Design Issues.
9. Crockford, D.: The application/json Media Type for JavaScript Object Notation (JSON), (2006).
10. Herman, I., Archer, P.: CSV on the Web Working Group Charter, http://www.w3.org/2013/05/lcsv-charter, (2013).
11. LODStats, http://stats.lod2.eu.
12. Sporny, M., Kellogg, G., Lanthaler, M.: JSON-LD 1.0 -A JSON-based Serialization for Linked Data, http://www.w3.org/TR/2013/CR-json-ld-20130910/.
13. Palmer, D.: WaterML 2.0 – Timeseries – NetCDF Discussion Paper, (2012).
14. Bigagli, L., Nativi, S.: NetCDF Uncertainty Conventions (NetCDF-U) 1.0. (2011).
15. Yu, J., Simons, B.A., Car, N., Cox, S.J.D.: Enhancing water quality data service discovery and access using standard vocabularies. 11th International Conference on Hydroinformatics (HIC). IWA Publishing, New York, NY, USA (2014).