Ritschel *et al. Earth, Planets and Space* (2016) 68:181
DOI 10.1186/s40623-016-0542-x

Earth, Planets and Space

TECHNICAL REPORT Open Access

Experiments using Semantic Web technologies to connect IUGONET, ESPAS and GFZ ISDC data portals

Bernd Ritschel[1*], Friederike Borchert[1], Gregor Kneitschel[1], Günther Neher[2], Susanne Schildbach[2], Toshihiko Iyemori[3], Yukinobu Koyama[3], Akiyo Yatagai[4], Tomoaki Hori[4], Mike Hapgood[5], Anna Belehaki[6], Ivan Galkin[7] and Todd King[8]

Abstract

E-science on the Web plays an important role and offers the most advanced technology for the integration of data systems. It also makes available data for the research of more and more complex aspects of the system earth and beyond. The great number of e-science projects founded by the European Union (EU), university-driven Japanese efforts in the field of data services and institutional anchored developments for the enhancement of a sustainable data management in Germany are proof of the relevance and acceptance of e-science or cyberspace-based applications as a significant tool for successful scientific work. The collaboration activities related to near-earth space science data systems and first results in the field of information science between the EU-funded project ESPAS, the Japanese IUGONET project and the GFZ ISDC-based research and development activities are the focus of this paper. The main objective of the collaboration is the use of a Semantic Web approach for the mashup of the project related and so far inoperable data systems. Both the development and use of mapped and/or merged geo and space science controlled vocabularies and the connection of entities in ontology-based domain data model are addressed. The developed controlled vocabularies for the description of geo and space science data and related context information as well as the domain ontologies itself with their domain and cross-domain relationships will be published in Linked Open Data.

Keywords: E-science, Semantic Web, Linked Open Data, Data system, System mashup, Controlled vocabulary, Domain ontology, Terminological ontology

Introduction

One of the main challenges of geo and space science activities is improving our understanding of the complex processes of the earth system including its interaction with solar-driven impacts, such as climate change or space weather. This requires an interdisciplinary approach which connects relevant and related data in the different geo and space science domains. Most of the geo and space domains have mature information models for describing available resources. Discovering available resources in multiple domains is a challenge which requires a level of expertise and knowledge of the individual data systems in each domain. This challenge can be met by the integration of the different geo and space science domains using Semantic Web-based mashup of the appropriate data and models (Allemang and Hendler 2008).

Scientific research has entered the fourth paradigm (Hey et al. 2009) and is more and more real data driven. There is an exponential growing of data (IDC White Paper 2011) with terabytes of data generated daily by sensors, digital models and social networks. This presents another type of challenge for the integration of systems because data are now Big Data (IDC White Paper 2011). This new paradigm has two contrary sides. On the one hand side, scientists are pleased about the potential of

*Correspondence: berndritschel@yahoo.de
[1] GFZ German Research Centre for Geosciences, Potsdam, Germany
Full list of author information is available at the end of the article

using more and more data from different domains, but on the other hand most data are not described and structured in a way for machine-based combination. Furthermore, the tools for finding, accessing and connecting such large amounts of data are not fully available. This challenge can be met by using the Resource Description Framework (RDF) (RDF Working Group 2004) standard as a metadata information model which is used by Semantic Web technology (Allemang and Hendler 2008) to automatically connect data systems and data. Another major reason for doing this research is the fact that standards implementation is at best patchy, and as a result, ontological mediation such as described here can be useful to address deficiencies and variations in quality of standards implementation.

A Sematic Web approach also addresses other related challenges. One is the development of a new culture of cooperative scientific work which is connected through the Web. With a Semantic Web, coherent research collaboratives can be formed that combine data, publications and social networks.[1,2] Also, while English is the main language in the field of science, scientific work is a personally organized effort, often discussed and reasoned in the researcher's primary language. This means that researchers use different vocabularies in different languages for the description of their research topics, results, applications and underlying data. Semantic Web technology (Allemang and Hendler 2008; Hebeler et al. 2009; Hitzler et al. 2008) can provide a solution by defining explicit expressions and connecting the different vocabularies using SKOS (W3C 1994–2012), RDFS (Brickley and Guha 2014) and OWL (OWL Working Group 2012).

In this paper, we mainly describe the GFZ ISDC efforts[3,4] to develop a Semantic Web-based data system using the ISDC ontology network.[5] This work is initial part for a planned Semantic Web technology-based connection of the ESPAS (European Commission, Research & Innovation, Research Infrastructures 2014),[6] IUGONET[7] (Abe et al. 2014; Yatagai et al. 2015) and GFZ ISDC[8] data portals (Hapgood and Iyemori 2013). A fruitful collaboration with the University of the Applied Sciences Potsdam, Department of Information Sciences, in research and education forms the basis for this project. The first activities involving the data modeling tasks for the ISDC ontology were started around 5 years ago. The first version of the ISDC ontology for mapping the information model of the ISDC repository was published in 2010 (Pfeiffer 2010). A Semantic Web-based data portal was developed using Virtuoso Universal Server[9] triple store and Drupal CMS in 2013.[10] The ISDC ontology and services were used to form connections to IUGONET, ESPAS and GFZ ISDC resources.

E-science projects—IUGONET, ESPAS and GFZ ISDC

To explore the use of Semantic Web technologies, a proof-of-concept project GFZ ISDC[11] was formed. The goal was to explore how to form a science collaborative using the IUGONET project[12] (Abe et al. 2014; Yatagai et al. 2015), the European Union ESPAS project[13] and the GFZ ISDC. This chapter describes the main requirements for scientific data systems and explains the background and main goals of the Japanese IUGONET project[14] (Abe et al. 2014; Yatagai et al. 2015), the European Union ESPAS project[15] and the GFZ ISDC (Ritschel et al. 2008a).

Requirements for e-science infrastructure

The main scientific and technical objectives for e-science or cyberspace projects are to improve the domain specific data management systems and make all resources available on the Web. Often data systems are responsible for sustainable ingestion, storage and provision of data. These systems usually have a specific data use policy. A basic service is to have data catalogs that describe repositories and data harvested from available metadata and context information. These catalogs can be searched for data and metadata and provide methods to access the data either anonymously or through authenticated channels. Some systems offer the publishing of data and the connection of data and publication as value-added services. Such systems are often based on common Content Management Systems (CMS) platforms like Typo3[16] or Drupal.[17] Additional value-added services such as moderated user forum or RSS-feed services may be offered. Interoperability between data systems is possible only if the systems are based on the same standards, for

[1] https://www.researchgate.net.

[2] https://www.linkedin.com/.

[3] http://isdc.gfz-potsdam.de.

[4] http://rz-vm125.gfz-potsdam.de/drupal/.

[5] http://rz-vm30.gfz-potsdam.de/ontology/isdc_1.4.owl.

[6] http://www.espas-fp7.eu/.

[7] http://www.iugonet.org/en/.

[8] http://isdc.gfz-potsdam.de.

[9] http://virtuoso.openlinksw.com/.

[10] https://drupal.org/.

[11] http://rz-vm125.gfz-potsdam.de/drupal/.

[12] http://www.iugonet.org/en/.

[13] http://www.espas-fp7.eu/.

[14] http://www.iugonet.org/en/.

[15] http://www.espas-fp7.eu/.

[16] http://typo3.org/.

[17] https://drupal.org/.

example, the same information model and a standardized service.

Additional motivations for open accessibility of data are reproducibility in science and better return on investment from tax-funded research.

IUGONET project

The Japanese Inter-university Upper Atmosphere Global Observation Network IUGONET[18] (Abe et al. 2014; Yatagai et al. 2015) project unifies the efforts of four Japanese universities from Kyoto, Nagoya, Tohuku and Kyushu and the National Institute for Polar Research. Its goal is to design, implement and operate a data system for the enhancement of the provision of mainly upper atmosphere and geomagnetic data. All project partners are responsible for the operation of specific ground-based observatories and instruments which are the basis for the geophysical data within the IUGONET data repository. The leading institution for the design and operation of the IUGONET data system called metadata database (MDB) is the WDC/WDS for Geomagnetism of the Kyoto University.[19] The 6-year research project IUGONET started in spring 2009. It is planned to continue the project with the addition of DOI[20]-based publishing of scientific data.

ESPAS project

The Near-Earth Space Data Infrastructure for e-Science ESPAS[21] project was founded by the European Union's Seventh Framework Program. The main objective is the design and implementation of an e-science infrastructure for distributed near-earth space data resources. The project started in November 2011 and will end in November 2015. There are more than 20 partners, mostly scientific institutions from all over Europe. The project is mainly driven by the RAL Space Department of the STFC's Rutherford Appleton Laboratory and the National and Kapodistrian University of Athens including the National Observatory of Athens. The tasks of the participants in the project vary from data provider and information modeler to software developer and system operator. More than 40 existing data repositories covering data from the atmosphere to outer radiation belts were measured by ground-based instruments and also satellites. The data providers mainly contribute metadata to a centralized ESPAS data system[22] which is still in development. Beside a catalog service and an access service to selected data, value-added services are part of the planned infrastructure ESPAS (2013).

GFZ ISDC project

The Information System and Data Center ISDC[23] of the Helmholtz Centre Potsdam—GFZ German Research Centre for Geosciences is an operational data portal for geoscientific data with corresponding metadata, scientific documentation and software tools (Ritschel et al. 2008a). The majority of the data and information are global geomonitoring products such as satellite orbit and earth gravity field data as well as geomagnetic and atmospheric data from GFZ-affiliated projects. It includes data from Challenging Minisatellite Payload (CHAMP) low earth orbit satellite,[24] the twin Gravity Recover And Climate Experiment (GRACE) low earth orbit satellites,[25] Global Navigation Satellite Systems (GNSS),[26] Global Geodynamic Project (GGP),[27] Global Geodetic Observing System (GGOS),[28] TerraSAR-X (TSX)[29] and other data associations.

Metadata for IUGONET, ESPAS and GFZ ISDC data portals

This chapter deals with information about the data portals of the IUGONET, ESPAS and GFZ ISDC projects. This includes the metadata, data models and the system architectures used in the ISDC Semantic Web framework.

Metadata formats and data models

Metadata or context data are used for the description of data. Such descriptions contain both information about the data itself, such as content information, start and stop time or spatial coverage of the measurement, and information about entities. It may also include descriptions of resources which are involved in the overall creation process, such as instruments and platforms, persons, institutions and projects. Metadata are also used to document parts of the data life cycle, such as the generation of knowledge in form of scientific publications. Data models, also known as information models, are the basis for system architectures of data systems. For the management of data repositories, underlying concepts and

[18] http://www.iugonet.org/en/.

[19] http://wdc.kugi.kyoto-u.ac.jp/.

[20] http://www.doi.org/.

[21] http://www.espas-fp7.eu/.

[22] https://www.espas-fp7.eu/portal/index.html.

[23] http://isdc.gfz-potsdam.de.

[24] http://www.gfz-potsdam.de/champ.

[25] http://www.gfz-potsdam.de/grace.

[26] http://www.gfz-potsdam.de/forschung/ueberblick/departments/department-1/.

[27] http://www.eas.slu.edu/GGP/ggphome.html.

[28] http://www.ggos.org/.

[29] http://terrasar-x.gfz-potsdam.de/.

relationships of appropriate entities are modeled. There are some standards for geoscience-related metadata and data models, such as DIF standard from NASA (DIF 2013), or ISO 19115 standard for metadata[30] and Observations and Measurements (O&M) data model standard from OGC/ISO.[31] In addition to structural standards for metadata and models, controlled terms or vocabularies are used for keyword-based tagging or indexing of entities. Examples of such vocabularies are the GCMD science keywords from NASA (Olsen et al. 2013) or the "allowed values" derived from the Space Physics Archive Search and Extract (SPASE) standard (King et al. 2010).

IUGONET common metadata format and model

The IUGONET data portal is based on the SPASE metadata and the SPASE data model (King et al. 2010). SPASE is a heliophysics community-based project for the design, implementation and operation of an e-science infrastructure in the heliophysics domain. The corresponding data model is used for the creation of data set descriptions for data collections. Main entities are data resources (numerical data, display data, catalog, granule and annotation), originating resources (observatory, instrument, person and document) and infrastructure resources (registry, repository and service). The SPASE data model specification[32] includes a conceptual ontology, shown in Fig. 1, with the primary implementation as an XML schema. Version 2.0.2 of the SPASE XML schema[33] was the basis for Version 1.0.0[34] of the IUGONET XML schema and the IUGONET common metadata format (Abe et al. 2014).[35] In the SPASE data model, all resource entities have a unique resource identifier URI and are described using the XML format. Recently, the IUGONET data model has been extended to include references to ORCID[36] and DOI[37] to enable connections between authors, publications and data. An important part of the metadata and the data model is the use of controlled vocabularies for classification and keyword-based search of entities. IUGONET uses both the SPASE keywords and GCMD science keywords.

ESPAS metadata and data model

The metadata used for the description of ESPAS entities are mainly based on the ISO 19115 standard for (Geographic Information—Metadata).[38] The ESPAS data model (ESPAS 2013) uses ISO standards, such as ISO 19101:2002 (Geographic information—Reference model)[39] and ISO 19109:2005 (Geographic information—Rules for creating and documenting application schemas).[40] The model is also partly based on the ISO 19156 Observations and Measurements (O&M) standard.[41] Core classes or entities of the O&M standard, which are also used for the ESPAS model, are feature of interest, observed property, observation result and designated procedure. In summary, the ESPAS data model version 2.0 consists of following concepts: organization, individual, project, instrument, platform, operation, acquisition process, computation process, composite process, collection and observation. The terminological ESPAS ontology[42] provides a controlled vocabulary for the near-earth space domain related to phenomena and observed properties. The terminological ESPAS ontology is modeled using the Semantic Web standard Simple Knowledge Organization System SKOS (W3C 1994–2012) for keyword collections, classifications and thesauri.

GFZ ISDC DIF standard and data model

The design of the operational GFZ ISDC data system[43] was based on NASA's DIF metadata standard (Directory Interchange Format (DIF) Writer's Guide 2013), mainly used for the GCMD and appropriate services. The DIF standard includes information about the data sets, such as title, temporal and spatial coverage, quality, access and use constraints, but also about instruments, platforms, projects, persons and data centers. An Entry ID is used for the identification of conforming DIF standard metadata documents. In former versions of the DIF standard, ASCII text was used. The recent version is available as DIF XML schema (Mende et al. 2008). The DIF standard is valid only for a collection of data or data sets called product types. In order to overcome the limitation, the GFZ ISDC derived an enhanced model to include information about granules or data products, such as a unique identifier, temporal and spatial coverage, revision and software version. Figure 2 shows the extension of the main DIF classes which form the ISDC DIF standard. The data model of the ISDC data portal is a relational data

[30] http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=53798.

[31] http://www.opengeospatial.org/standards/om.

[32] http://www.spase-group.org/data/dictionary/spase-2_2_2.pdf.

[33] http://www.spase-group.org/data/schema/.

[34] http://www.iugonet.org/data/schema/.

[35] http://www.iugonet.org/en/mdformat.html.

[36] http://orcid.org/.

[37] http://www.doi.org/.

[38] http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=53798.

[39] http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=26002.

[40] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=39891.

[41] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=32574.

[42] http://isdc.gfz-potsdam.de/ontology/spase_keywords.owl.

[43] http://isdc.gfz-potsdam.de.

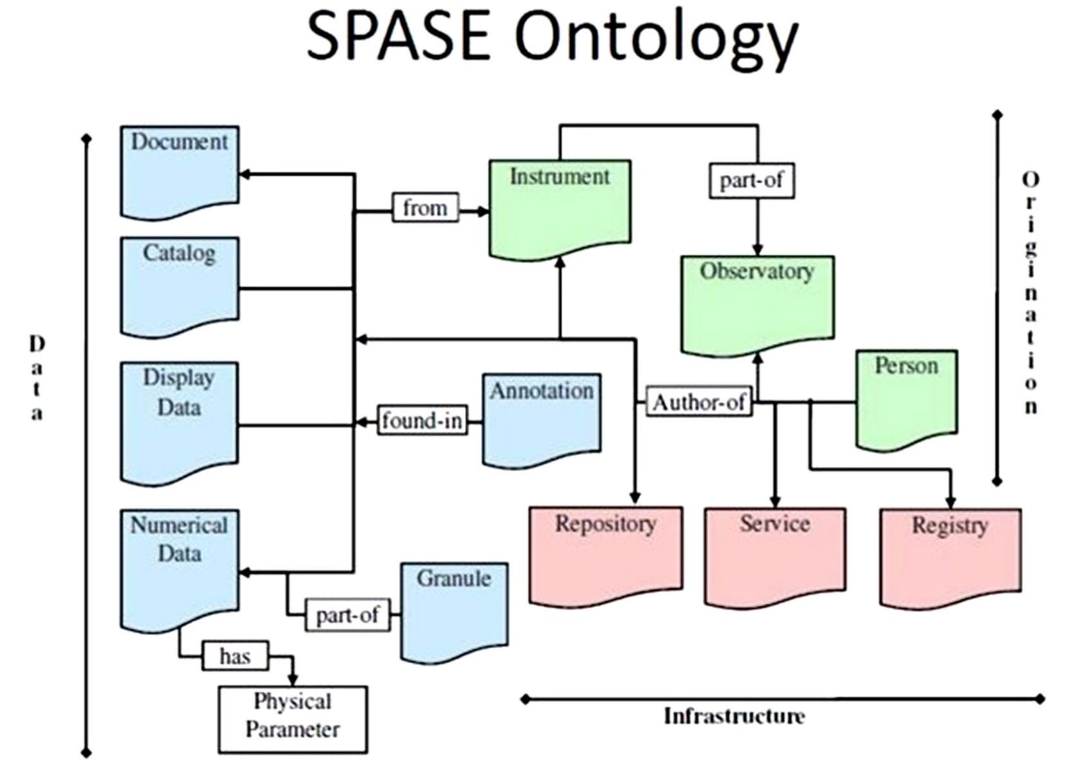


Fig. 1 Association map between resources in the SPASE information model. *Arrows* point in the direction of the association. This figure is taken from "A Space and Solar Physics Data Model" from the SPASE consortium (http://www.spase-group.org/data/dictionary/spase-2_2_2.pdf)

model and is implemented using a relational database management system (Ritschel et al. 2008a). The GFZ ISDC data catalog mainly consists of product type-related tables extended by aggregated tables for enhanced search capabilities. The ISDC metadata documents for product types benefit from the use of GCMD science keywords.

GFZ ISDC: Semantic Web proof of concept

Recognizing both the usefulness of each of the previously described data portals and the complementary nature of their content, we set out on the goal to interconnect the ESPAS, IUGONET and GFZ ISDC data portals. Our analysis showed that while each system used different metadata, conceptually there was a great deal of commonality. The ideal approach to achieving interoperability would be to form a Semantic Web.

Semantic Web stack and standards

From its inception in 1991, the WWW (Lee et al. 1992; Shadbolt et al. 2006) quickly became the standard infrastructure of the Internet. The World Wide Web Consortium (W3C),[44] with Tim Berners-Lee as its director, is the standardization body for the WWW specifications. An implementation of the WWW specifications is commonly referred to as a Web. One of the core WWW specifications is for Unique Resource Identifiers (URIs), or more specific Uniform Resource Locators (URLs), which are used to identify and address documents in the Web. The Hypertext Transfer Protocol HTTP[45] is responsible for the communication within the Web. This application layer protocol connects resources using hyperlinks in HTML documents. This allows HTML documents in the Web to be connected using links. This works exceptionally well, in part because the Web was created for human mind-based interaction. However, there are no explicit semantics of the elements and links of a Web page.

Adding semantics to the Web will allow data to be shared and reused across current boundaries. The technology stack to add semantics is referred to as the Semantic Web.[46]

The base technology is the Resource Description Framework (RDF) standard (RDF Working Group 2004). For data interchange, the RDF connects Web resources with specific properties which link to other resources or

[44] http://www.w3.org/.

[45] http://www.w3.org/Protocols/.

[46] http://www.w3.org/2004/Talks/1117-sb-gartnerWS/slide18-0.html.

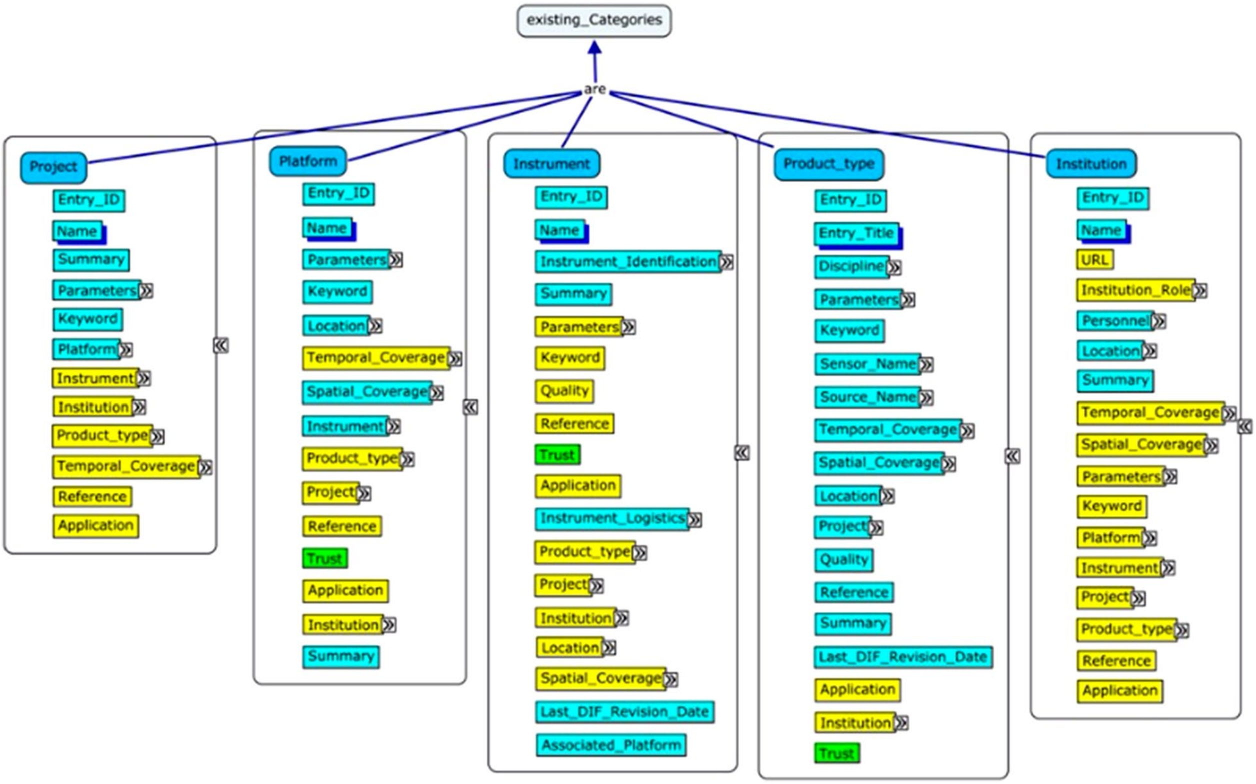


Fig. 2 Main classes and elements of the extended ISDC DIF data model. The cyan colored elements are taken from NASA's DIF standard; the yellow and green colored one are ISDC extensions to this standard. The figure is taken from Sabine Pfeiffer's Master of Engineering Thesis (Pfeiffer 2010)

just literals (strings or numbers). An example is the connection of an author and a book using a triple consisting of subject, predicate and object. Just like in natural language: The author (subject) is Creator (predicate) of the book (object). Each element of the triple may be resources and referenced with a URI. A formal representation or model of knowledge in a real world domain is called an ontology[47] (Gruber 1995). The design of an ontology may be described with RDF Schema (RDFS) (Brickley and Guha 2014) or Ontology Web Language OWL (OWL Working Group 2012). RDFS and OWL extend the features of RDF by the introduction of classes and subclasses, respectively. Subproperties and logical constructs, such as inverse, symmetric, transitive, disjunct and equivalent, provide inference capability based on the first-order predicate logic. Specific elements of OWL, such as "owl:sameAs," are used to connect entities from different ontologies. Populating an ontology with individuals creates a knowledge base. A knowledge base can be access and queries using the RDF Query Language SPARQL (2008). With SPARQL, individuals can be retrieved and manipulated according to rules defined in Rule Interchange Format RIF.[48] The highest layers in the Semantic Web stack, such as unifying logic, proof and trust, are still in an experimental status and not yet realized.

LOD: Semantic Web application

Linked Open Data LOD (Hebeler et al. 2009)[49] is the most known and a successful project and application in the Semantic Web and is based on the linked data principles defined by Tim Berners-Lee in 2007 (Hebeler et al. 2009; Christian et al. 2009; Berners-Lee 2006). These principles build on the Semantic Web standards and focus on the use and connection of URIs or Internationalized Resource Identifiers IRIs[50] as a way to make statements in RDF expressed as subject–predicate–object triples. Collections of statements can be evaluated and searched using query languages such as SPARQL (2008). When RDF expressions are defined for openly accessible resources, you can define a LOD cloud (Jentzsch et al. 2011). One of the first applications was DBpedia

[47] http://queksiewkhoon.tripod.com/ontology_01.pdf.

[48] http://www.w3.org/TR/rif-overview/.

[49] http://linkeddata.org/.

[50] http://www.w3.org/International/O-URL-and-ident.html.

(Lehmann et al. 2012).[51] DBpedia is the Semantic Web counterpart of Wikipedia in the Web. At present, DBpedia contains around 8.8 billion RDF transformed triple of about more than 6 million entities,[52] mainly referencing to the info boxes of Wikipedia. The DBpedia SPARQL endpoint[53] is used to connect DBpedia resources via SPARQL with other RDF resources in LOD. At present, LOD is composed of about 2200 data sets[54] mainly covering the domains of media, geographic, government, publication, cross-domain, life sciences and user-generated content (Jentzsch et al. 2011). In addition to GeoNames[55] and Linked GeoData[56] containing geographical information, there are also resources related to geo and space sciences, such as NASA Space Flight & Astronaut data in RDF[57,58] and related to e-infrastructure projects available, e.g., Linked Sensor Data (Kno.e.sis)[59] in LOD.

Methods for design and mashup of data in the Semantic Web

Structured resources in the RDF format (RDF Working Group 2004) managed by a triple store which include a SPARQL (2008) endpoint are necessary for an efficient mashup of different entities. RDF data reflect the use of entities, such as classes or properties of one or more appropriate ontologies. For enhanced interoperability, it is best to adopt existing ontologies when available. Domain ontologies such as the Semantic Web for Earth and Environmental Terminology SWEET ontology[60] from NASA or the Semantic Sensor Net SSN ontology[61] from W3C[62] are good starting points for the creation of an ontology for a particular domain. There are also terminological ontologies containing controlled vocabularies for the tagging and indexing of resources of the geo and space science domain, such as GEMET (General Multilingual Environmental Thesaurus GEMET 2012).

Modeling the ISDC ontology network

The ISDC ontology (Pfeiffer 2010) was developed according to best practice process models (Noy and McGuinness 2001). The scope and domain of the ISDC ontology is the conceptual mapping of parts of the data life cycle valid for the objectives of the GFZ ISDC (Ritschel et al. 2008a). For the modeling of the ISDC ontology, both Protégé 3[63] and Protégé[64] 4 have been used.

Forming a Semantic Web

The ISDC ontology network is the basic model for the Semantic Web-based GFZ ISDC proof-of-concept[65] implementation. The main ISDC classes and properties are derived from the extended GCMD DIF standard used at the operational GFZ ISDC (Pfeiffer 2010; Ritschel et al. 2012; Ritschel et al. 2008b). This means the core metadata or context information describing the data—ISDC product types and data products—is still compliant to the DIF standard. The ISDC ontology was developed first with the intension to be a one-to-one translation of the ISDC DIF schema (Ritschel et al. 2008b). The main classes are ProductType and DataProduct describing the core context of the data itself. Instrument and Platform classes with information about the sensors and carriers of the sensors, such as observatories or satellites, provide contextual information. Additional classes for Person, Institution and Project are included to provide information of the roles of people, institutions and projects who are involved in the data life cycle. Finally, Publication and Phenomenon classes were added. An important aspect of the ISDC ontology network (Ritschel and Neher 2013) is the ability to connect ISDC ontology classes and properties with ontology entities available in Linked Data (Hebeler et al. 2009) or Linked Open Data.[66] Classes and properties from such ontologies, such as FOAF (Brickley and Miller 2014), Bibo (D'Arcus and Giasson 2009) or Geonames,[67] have been linked to the appropriate ISDC ontology entities. For example, "isdc:person owl:equivalentClass foaf:person" connects the ISDC class Person with the appropriate FOAF class. In this process, the core GCMD ontology was taken out of the ISDC ontology and the GCMD classes and properties also have been linked to the appropriate ISDC entities. Figure 3 shows the main entities and relationships of the ISDC ontology network. Most metadata elements of the schema could be transformed into object properties modeling the relationship between classes. For example, "isdc:isCreatedBy" connects individuals of

[51] http://dbpedia.org/About.

[52] http://wiki.dbpedia.org/Datasets.

[53] http://dbpedia.org/sparql.

[54] http://datahub.io/en/dataset?q=linked+open+data.

[55] http://www.geonames.org/.

[56] http://linkedgeodata.org/About.

[57] https://earthdata.nasa.gov/esdswg.

[58] http://datahub.io/dataset/data-incubator-nasa.

[59] http://datahub.io/dataset?q=Kno.e.sis.

[60] http://sweet.jpl.nasa.gov/.

[61] http://www.w3.org/2005/Incubator/ssn/wiki/Semantic_Sensor_Net_Ontology.

[62] http://www.w3.org/.

[63] http://protegewiki.stanford.edu/wiki/Protege_Desktop_Old_Versions#Protege_3.

[64] http://protegewiki.stanford.edu/wiki/Protege_Desktop_Old_Versions#Protege_4.

[65] http://rz-vm125.gfz-potsdam.de/drupal/.

[66] http://linkeddata.org/.

[67] http://www.geonames.org/.

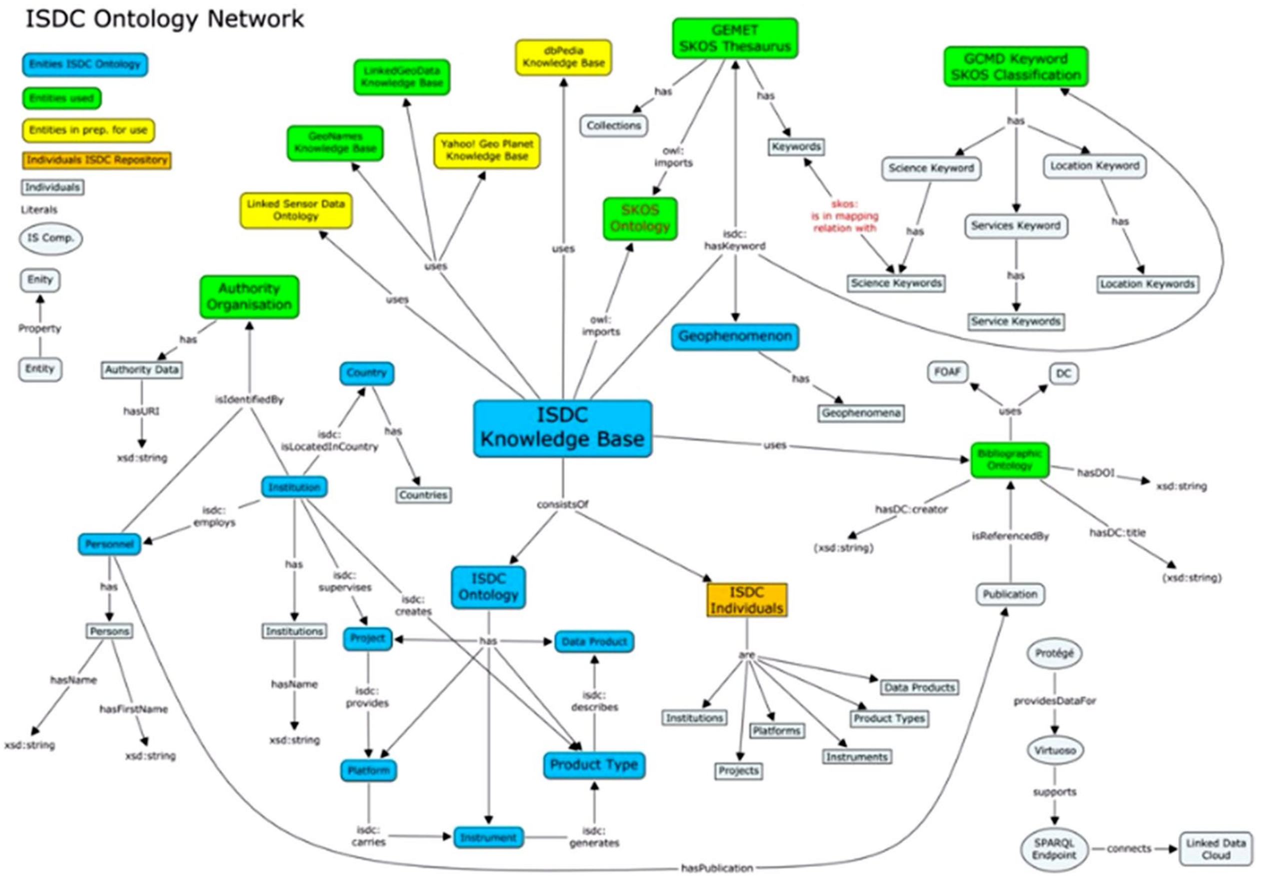


Fig. 3 ISDC ontology network. The network is composed of the ISDC core ontology and appropriate individuals, connected with further domain and terminological ontologies

ProductType with Institution (Fig. 4, relationship or property 4) and “isdc:isMeasuredBy” connects ProductType with Instrument (Fig. 4, relationship or property 10). Because the ISDC ontology is modeled in OWL (OWL Working Group 2012), powerful OWL constructs such as “owl:inverseOf” to define inverse features or “owl:transitiveProperty” for the expression transitive features of a property are used. For example, “isdc:isMeasuredBy owl:inverseOf isdc:measuresDataFor” expresses that the property isMeasuredBy is the inverse of the property measuresDataFor. When used to describe that a Product Type “is measured by” the Instrument, there is a corresponding inverse relationship that asserts that the Instrument “measures data for” the Product Type.

In addition to the data life cycle concepts, terminological ontologies have been modeled and included into the ISDC ontology network[68] (Ritschel and Neher 2013). Again the DIF standard plays an important role. SPASE and other organizations which are providing controlled vocabularies for the indexing of entities are also included. Similar to the Parameters field of the ISDC DIF metadata documents containing controlled terms from the GCMD earth science keywords document (Olsen et al. 2013), these keywords are used as a controlled index in the ISDC ontology network. For the use of the GCMD keywords at the ISDC ontology network, the hierarchically structured science keywords have been modeled as concepts with appropriate relationships (properties) and translated into SKOS.[69] In a similar process, the SPASE “allowed values” have been classified and the hierarchically related concepts assigned to the appropriate SKOS concept schemas.[70] In addition to GCMD and SPASE keywords, the SKOS version of the GEMET (2012) (General Multilingual Environmental Thesaurus GEMET 2012) vocabulary designed and controlled by the participants of the European Environment Agency was added to the ISDC ontology network.

[68] http://rz-vm30.gfz-potsdam.de/ontology/isdc_1.4.owl.

[69] http://isdc.gfz-potsdam.de/ontology/gcmd_science.skos.rdf.

[70] http://isdc.gfz-potsdam.de/ontology/spase_keywords.owl.

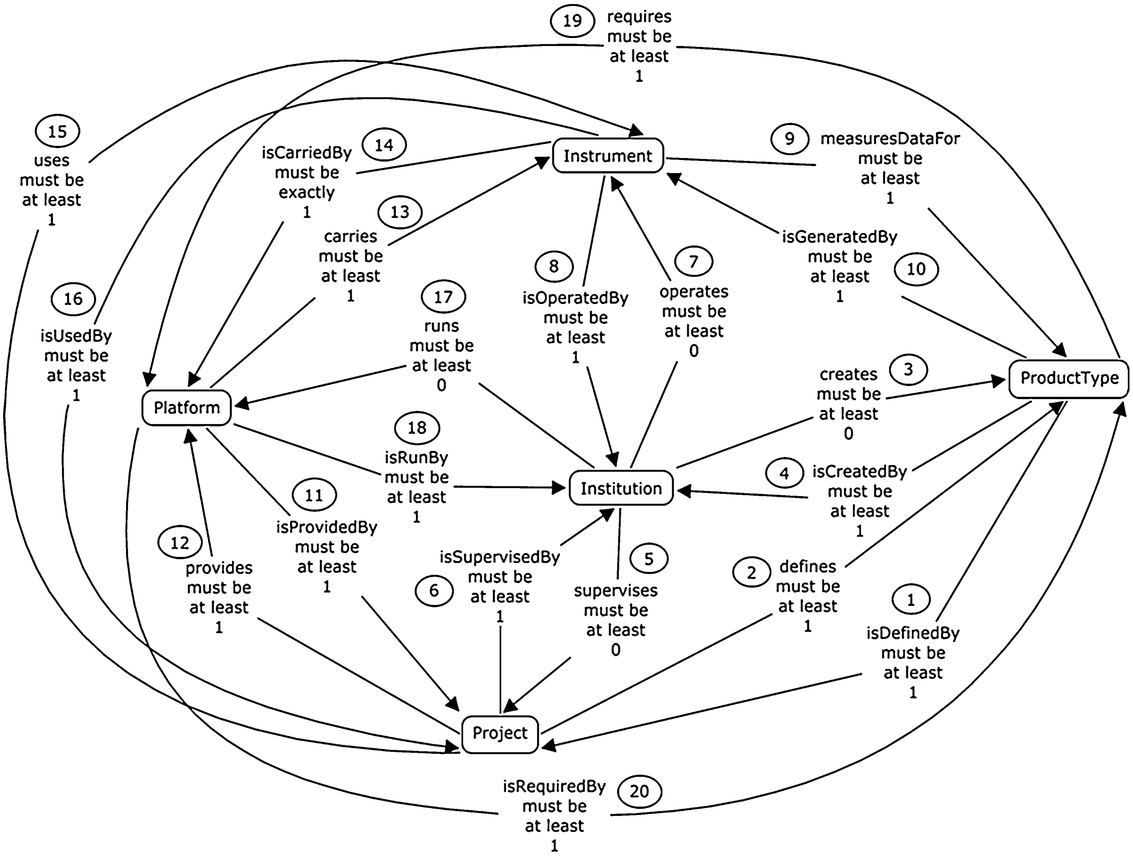

19 requires must be at least 1
15 uses must be at least 1
isCarriedBy must be exactly 1 14
Instrument
9 measuresDataFor must be at least 1
13 carries must be at least 1
8 isOperatedBy must be at least 1
7 operates must be at least 0
isGeneratedBy must be at least 1 10
16 isUsedBy must be at least 1
17 runs must be at least 0
creates must be at least 0 3
ProductType
Platform
18 isRunBy must be at least 1
Institution
4 isCreatedBy must be at least 1
11 isProvidedBy must be at least 1
12 provides must be at least 1
isSupervisedBy must be at least 1 6
5 supervises must be at least 0
2 defines must be at least 1
1 isDefinedBy must be at least 1
Project
isRequiredBy must be at least 1 20

Fig. 4 Object properties reflecting the relationships between main classes of the ISDC core ontology. Also shown are the corresponding inverse properties. The small numbers below the property names are cardinalities

Transforming GCMD's science keywords and SPASE "allowed values"

The team of the Global Change Master Directory from NASA has developed different controlled vocabularies covering the geo and space science domain, as well as geographical and specific data parameters aspects (Olsen et al. 2013). For the use within the Semantic Web approach, these vocabularies have been transformed into RDF data using the SKOS standard (W3C 1994–2012). Hierarchical relationships between keywords (SKOS concepts) have been translated into transitive semantic relations such as "...skos/core:broader" and "...skos/core:narrower." For example, "concept#Atmosphere skos/core:narrower concept#Atmospheric Chemistry" expresses that "Atmospheric Chemistry" is a narrower concept of an "Atmosphere." To become independent from the notation of terms, and for future multilingualism, an independent decimal classification system has been introduced to link to the terms of the vocabulary. The English notation of the term is kept in the annotation property field "prefLabel," whereas the definition or explanation of the terms related to the specific domain of the vocabulary is documented in the annotation property field definition (Ritschel and Neher 2013).[71]

The SPASE schema (King et al. 2010)[72] provides various enumeration lists and appropriate concepts for different elements. These elements are related to a specific domain, such as instrument type and measurement type or observatory region and observed region. Some enumeration lists are even hierarchically structured, such as observatory region and observed region, as demonstrated in Fig. 5. The idea to transform these lists as part of a controlled SPASE vocabulary into the SKOS format was realized by mapping such schema elements which are related to an enumeration list to an appropriate SKOS concept schema. For example, SPASE schema element "instrument type" was mapped to the SKOS concept schema Instrument Type. The list of values then became SKOS concepts of the appropriate SKOS concept schema. Again SKOS object properties reflecting broader or narrower relationships are used for the mapping of the

[71] http://isdc.gfz-potsdam.de/ontology/gcmd_science.skos.rdf.

[72] http://www.spase-group.org/data/dictionary/spase-2_2_2.pdf.

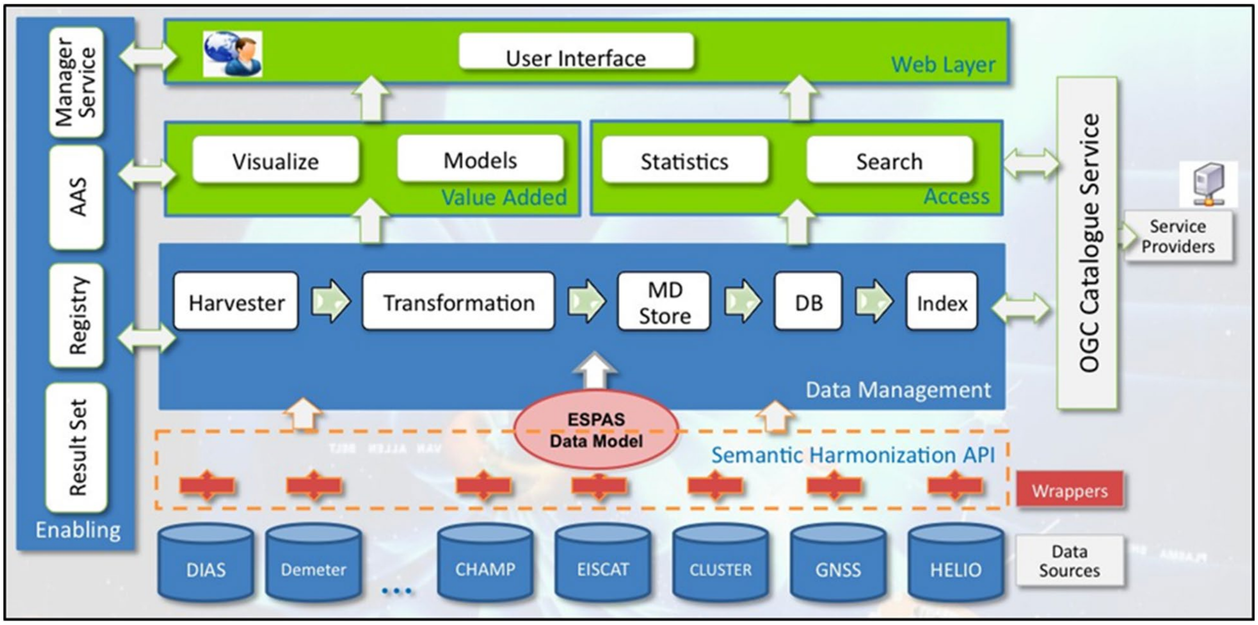


Fig. 5 Transformation of the SPASE "allowed values" as controlled vocabulary into the SKOS standard. Shown is the example of the concept schema "Observatory Region" and appropriate concepts

hierarchical structure between some values and related concepts of the enumeration lists.[73]

Mapping and merging of domain and terminological ontologies with the example of SPASE/IUGONET, ESPAS and GFZ ISDC ontologies

Mapping and merging are techniques for the semantic integration of different domain and terminological ontologies (Allemang and Hendler 2008; Hebeler et al. 2009; Hitzler et al. 2008). Specific OWL constructs provide the capability for the mapping or merging of entities, such as classes or properties. Such OWL properties are sameAs, equivalentClass or equivalentProperty. The semantic similarity or the semantic distance of classes, properties or individuals of different ontologies is the key to semantic integration. The estimation of the semantic similarity of entities was done for the SPASE/IUGONET and GFZ ISDC domain ontologies (Schildbach 2013). If you compare the object properties for the relationship between data and instrument in the SPASE and GFZ ISDC ontology, the value of the semantic similarity is 0.81, as shown in Fig. 6. In this case, you can reason the object property "spase:isDataOf" is very similar to the appropriate property "isdc:isMeasuredBy." The connection of these properties can be done using the OWL constuct "owl:equivalentProperty" (Schildbach 2013).

A similar approach can be used for the connection of concepts of terminological ontologies. Using a lexical analysis, the comparison of the similarity of strings or substrings of concepts can help to estimate the semantic similarity of the concepts. Stemming and the extraction of term signatures of concepts before the string comparison increase the equivalence assumptions. A structural analysis of the terminological ontology comparing parent and child concepts also improves the process of the ontology mapping/merging. Figure 7 shows a simplified process model of the merging of two vocabularies. The terminological ontology derived from the SPASE/IUGOENT schema[74,75] and the GCMD science keywords ontology[76] developed for the GFZ ISDC Semantic Web have been mapped and merged[77] (Kneitschel 2013). In this case, an automatic procedure for performing a lexical analysis, adapted for use with ontology mapping, detected 23 "equal" concepts. But only 14 concepts of the different ontologies had a real semantic similarity for the use of the SKOS construct "closeMatch." Examples are the concepts Atmosphere, Corona and Electric Field (Kneitschel 2013). The small number of semantic equal concepts comes from the small overlap or intersection of the terminological ontologies or controlled vocabularies SPASE/IUGONET and GCMD science keywords.

[73] http://isdc.gfz-potsdam.de/ontology/spase_keywords.owl.

[74] http://www.spase-group.org/data/dictionary/spase-2_2_2.pdf.

[75] http://www.spase-group.org/data/schema/.

[76] http://isdc.gfz-potsdam.de/ontology/gcmd_science.skos.rdf.

[77] http://isdc.gfz-potsdam.de/ontology/spase_keywords.owl.

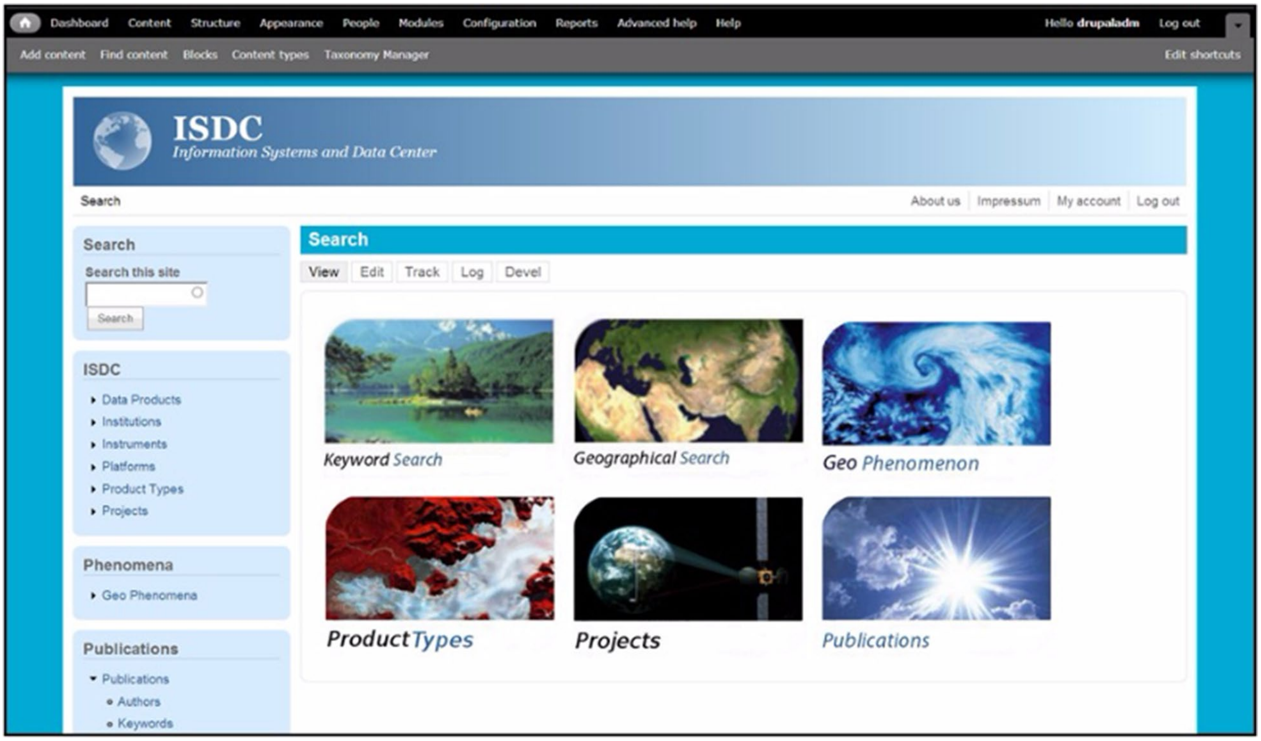


Fig. 6 Particular result of the estimation of semantic similarities of the SPASE and ISDC domain ontologies. Shown are the similarities of the properties of the "spase:Data" and "isdc:ProductType" classes. This figure is taken from Susanne Schildbach's Bachelor of Art thesis (Schildbach 2013)

The reason is quite simple. The domain of the SPASE/IUGONET is specific to near-earth space science, whereas the vocabulary of the GCMD science keywords covers all geo and space science domains.

System architecture, frameworks and services

The next step was to use the ISDC ontology in an operational system. In a complete system, the system architecture describes the components and relationships between the components and subcomponents as well as the interfaces between components and the available API. This process begins with a functional view of the system architecture which is defined by use cases that describe each workflow. This leads to a logical view of the system architecture which is the basis for design decisions related to software implementation and hardware platforms. With a logical view of the system, it is possible to define or select a framework as the software development environment.

To determine an appropriate ISDC Semantic Web system architecture, we looked at the system architecture for our selected data portals. The overall system architecture—seen from a global scope—is very similar for the IUGONET, ESPAS or GFZ ISDC data systems. Each system architecture is layered and service oriented, consisting of the following main components: data sources, data registration, data access, harvesting and transformation, indexing and catalog ingestion, catalog search and data download. Some portals also have value-added services, such as visualization or statistics.

IUGONET platform

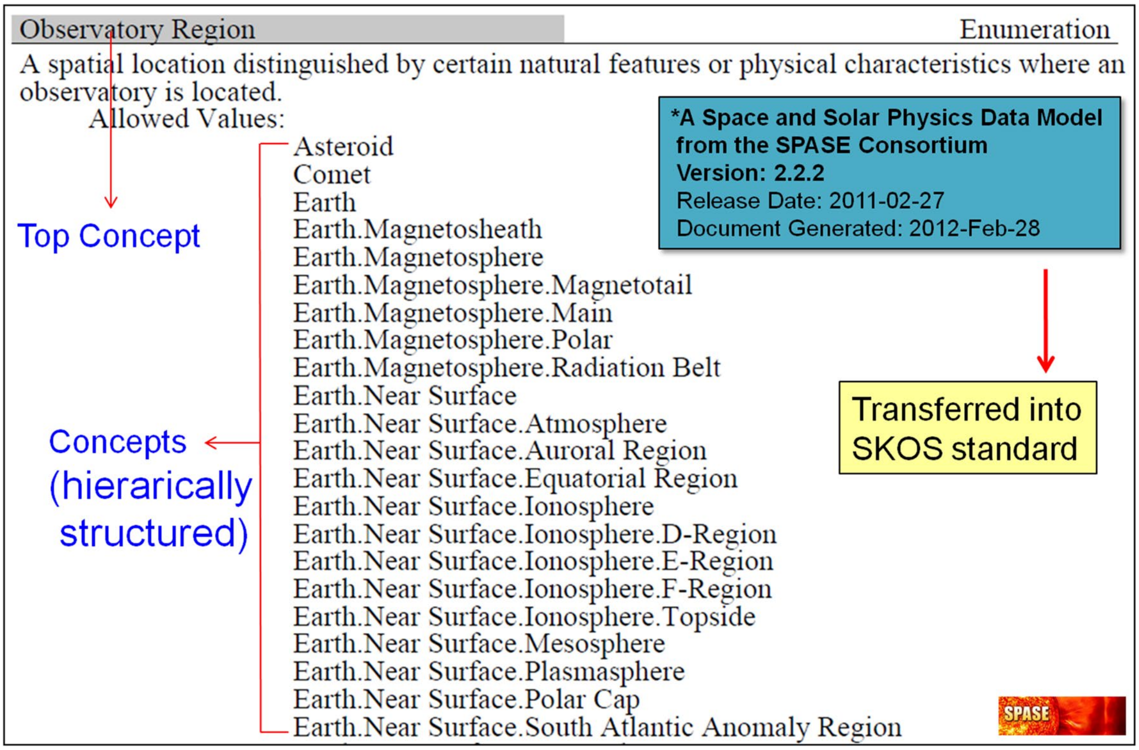
The IUGONET data system is built upon the open source platform DSpace[78] for the creation and management of digital repositories. Resources are described using the IUGONET/SPASE data model,[79] expressed in XML with the XML documents managed by DSpace.[80] New resources and documents can be registered, and every single resource entity is referenced by a unique identifier. Data search and access capabilities are implemented and reflected in the GUI of the data portal.[81]

[78] http://www.dspace.org/.

[79] http://www.iugonet.org/data/schema/.

[80] http://www.dspace.org/.

[81] http://search.iugonet.org/iugonet.



Observatory Region Enumeration

A spatial location distinguished by certain natural features or physical characteristics where an observatory is located.

Allowed Values:

Asteroid
Comet
Earth
Earth.Magnetosheath
Earth.Magnetosphere
Earth.Magnetosphere.Magnetotail
Earth.Magnetosphere.Main
Earth.Magnetosphere.Polar
Earth.Magnetosphere.Radiation Belt
Earth.Near Surface
Earth.Near Surface.Atmosphere
Earth.Near Surface.Auroral Region
Earth.Near Surface.Equatorial Region
Earth.Near Surface.Ionosphere
Earth.Near Surface.Ionosphere.D-Region
Earth.Near Surface.Ionosphere.E-Region
Earth.Near Surface.Ionosphere.F-Region
Earth.Near Surface.Ionosphere.Topside
Earth.Near Surface.Mesosphere
Earth.Near Surface.Plasmasphere
Earth.Near Surface.Polar Cap
Earth.Near Surface.South Atlantic Anomaly Region

Top Concept

Concepts (hierarically structured)

*A Space and Solar Physics Data Model from the SPASE Consortium
Version: 2.2.2
Release Date: 2011-02-27
Document Generated: 2012-Feb-28

Transferred into SKOS standard

Fig. 7 Merging workflow of two thesauri with selected process steps. Such steps are pre-integration, analysis, disambiguation, restructuring and integration and finally evaluation. This figure is taken from the Gregor Kneitschel's Bachelor of Art thesis (Kneitschel 2013)

ESPAS platform

The system architecture of the ESPAS data system[82] is service-oriented architecture (SOA), as shown in Fig. 8. For the integration of distributed resources and applications, XML, SOAP, REST, UDDI and WSDL technology is used (ESPAS 2013). The ESPAS data system is based on the D-NET framework[83] for the construction of digital data infrastructures. The D-Net framework provides services for data mediation, data mapping, data storage and indexing, data curation and enrichment, and data provision. After an authorized registration of distributed ESPAS resources, appropriate XML metadata documents are harvested using OAI-PMH[84] mechanism. The implemented OGC Catalog Service OGC CSW[85] connects ESPAS data provider and the centralized catalog of the ESPAS data repository over the Web. The OGC CSW catalog service also provides search capabilities. A new version of the ESPAS data system,[86] demonstrating the main features, is available on the Web.

GFZ ISDC platform

The operational GFZ ISDC[87] was developed using the open source PostNuke CMS and portal framework.[88] In order to adapt the functionality of the PostNuke framework to the requirements of a data system, unnecessary components were removed and others were added (Ritschel et al. 2008a). ISDC/DIF metadata extracted from the ASCII and/or XML documents and stored in relational database which is the foundation for the GFZ ISDC data catalog (Mende et al. 2008). Unique identifiers also stored at the catalog are used to reference all granules in the data archive of the ISDC system. Main components of the current GFZ ISDC data system are proprietary and therefore not ready for interoperability.

GFZ ISDC: Semantic Web-based proof-of-concept platform

After evaluating the selected data portals, we selected the open source CMS Drupal 7[89] and the Virtuoso Universal Server[90] for the backbone of the Semantic Web-based

[82] https://www.espas-fp7.eu/portal/index.html.

[83] http://www.d-net.research-infrastructures.eu/.

[84] http://www.openarchives.org/pmh/.

[85] http://www.opengeospatial.org/standards/cat.

[86] https://www.espas-fp7.eu/portal/.

[87] http://isdc.gfz-potsdam.de.

[88] http://www.pn-cms.de/.

[89] https://drupal.org/.

[90] http://virtuoso.openlinksw.com/.

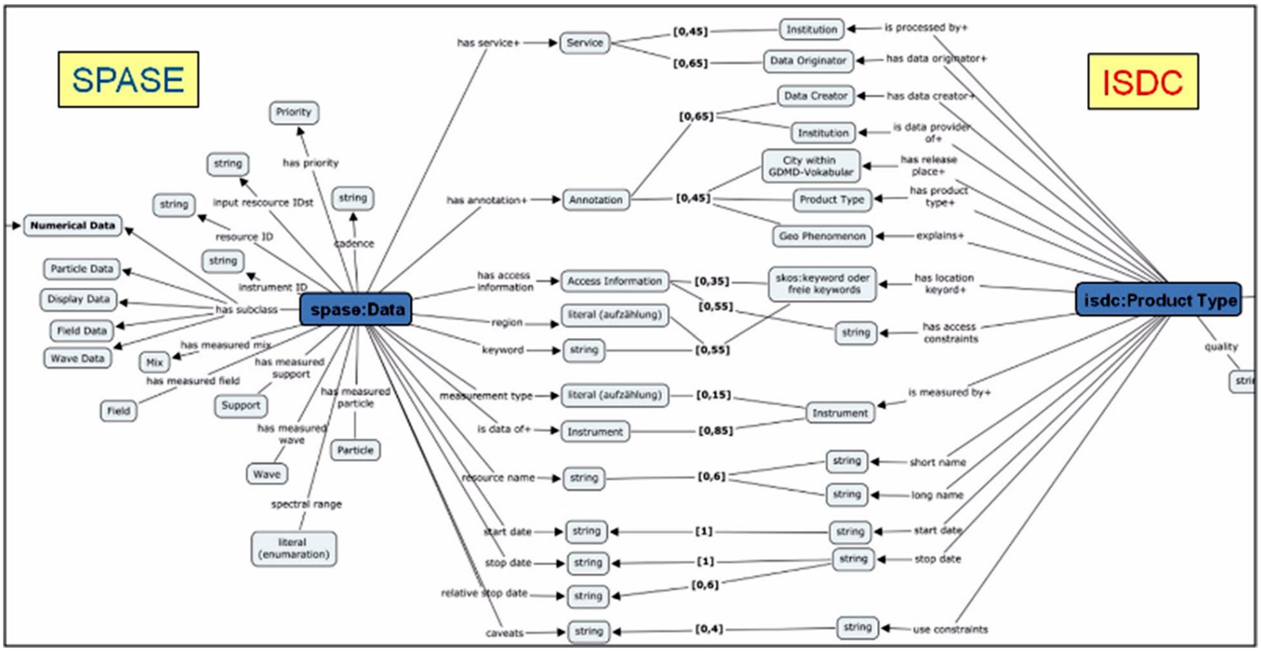


Fig. 8 ESPAS architecture overview based on service-oriented architecture principles. This figure is taken from the "ESPAS, the near-Earth space data infrastructure for e-Science" (ESPAS 2013)

GFZ ISDC data server.[91] Virtuoso is used for the RDF data management providing a triple store and SPARQL endpoint, in our case the management of the GFZ ISDC knowledge base consisting of the ISDC ontology network (OWL file)[92] and appropriate individuals (RDF data). The complete business logic of the Semantic Web-based ISDC data server is implemented in Drupal 7. The RDF triples of the GFZ ISDC knowledge base are imported from Virtuoso and indexed by an Apache Solr index server.[93] The individuals and appropriate relationships of the ISDC ontology network including the terminological ontologies are visualized in the GUI of the Drupal system. Drupal also provides a SPARQL interface (SPARQL 2008) for the connection of ISDC entities with external resources in Linked Open Data (LOD). In order to answer the question why we made the choices and how Drupal[94] and Virtuoso[95] compare to other alternatives, such as Apache Jena framework (Apache Software Foundation 2011–2014), we refer to Christoph Seelus's Bachelor of Art thesis about Sementic Web CMS for scientific data management (Seelus 2014). The thesis focuses on the development of an evaluation procedure for the comparison of Semantic Web CMS including appropriate data storage management systems and the subsequent use of this procedure for the features of well-known Semantic Web CMS. Beside Drupal,[96] DSpace,[97] Semantic MediaWiki,[98] OntoWiki[99] und Ximdex[100] were evaluated. In addition, the Semantic Web Frameworks Apache Stanbol,[101] Erfurt SWF[102] and OpenRDF Sesame[103] were proofed. Without going into details, the procedure focuses on requirements and performance indicators, such as technology and system requirements, content and user management, security and software ecosystem, and especially Semantic Web features including knowledge representation, queries and rules. The results of the evaluation clearly show that none of the currently available and tested systems really can meet professional user's requirements regarding functionality and ecosystem. Only Drupal and with a lower degree DSpace[104] and Semantic MediaWiki achieve satisfactory results.

[91] http://rz-vm125.gfz-potsdam.de/drupal/.

[92] http://rz-vm30.gfz-potsdam.de/ontology/isdc_1.4.owl.

[93] https://lucene.apache.org/solr/.

[94] https://drupal.org/.

[95] http://virtuoso.openlinksw.com/.

[96] https://drupal.org/.

[97] http://www.dspace.org/.

[98] https://www.semantic-mediawiki.org/wiki/Semantic_MediaWiki.

[99] http://aksw.org/Projects/OntoWiki.html.

[100] http://www.ximdex.com/.

[101] https://stanbol.apache.org/.

[102] http://aksw.org/Projects/Erfurt.html.

[103] http://www.openrdf.org/.

[104] http://www.dspace.org/.

User interfaces and services

Graphical user interfaces and APIs for inter-machine communication are necessary for the interaction with the data systems. Such interactions include data search and catalog browsing but also data access and data download. System interoperability mainly depends on the underlying data model and also depends on API functionality. A survey of the user interfaces and APIs for the selected data portals helped to inform the selection for the ISDC Semantic Web portal.

IUGONET system interfaces and services

The IUGONET data system provides a simple but efficient GUI to the end users.[105] Correspondent to the data model, metadata are searchable related to resource types but also using temporal and spatial coverage data or keywords from the controlled SPASE and GCMD science keyword vocabulary. Value-added services, such as data analysis, are realized using IUGONET Data Analysis Software (UDAS).[106]

ESPAS system interfaces and services

The ESPAS data system offers GUI-based services and APIs for data providers and end users.[107] New data resources can be registered entering the metadata according to the data model. Web-based harvesting mechanism automatically ingests metadata of observations and measurements from the different distributed data providers. A qualified search for data is realized using the GUI of the ESPAS data system.

GFZ ISDC system interfaces and services

The operational GFZ ISDC provides not only the search for data but also the access and download of data files. The system also manages the documents necessary for the use of the data. The portal GUI only provides a search for data products of a specific product type for end users.[108] There is no search across all product types which may be available in the ISDC data repository. A proprietary API provides a machine-based request for data. All requested data are delivered from the ISDC archive to end user-specific directories.

GFZ ISDC: Semantic Web-based proof of concept

Ideally the user interface and capabilities of the ISDC Semantic Web should encompass all the capabilities of the selected data portals. We found that the RDF capabilities of Drupal 7 provide a GUI for the interaction with the Semantic Web-based GFZ ISDC data system.[109] Search for data-related context information is ontology class based and enhanced by the use of controlled vocabulary terms. Context-dependent DBpedia data (Lehmann et al. 2012) from LOD are automatically requested and visualized, such as DBpedia information about institutions. Open street map data are used for the geographical referencing and visualization of search results. The graphical user interface of the ISDC GFZ is shown in Fig. 9.

At present, the Virtuoso Universal Server[110] and the Drupal 7 CMS[111]-based GFZ ISDC—Semantic Web-based proof-of-concept data server[112] only contain a limited number of entities of the GFZ ISDC repository. The knowledge base consists of the ISDC ontology network, version 1.4[113] and appropriate individuals. Most RDF data are related to the gravity field of the earth measured by superconducting gravimeter but also related to the atmosphere and ionosphere derived from GPS measurements, and related to the geomagnetic field from CHAMP satellite magnetometers. These data are linked with RDF data about instruments and platforms, and also persons, institutions, projects and geophenomena. SPARQL queries are used for the connection of known resources with DBpedia[114] information for institutions, instruments, platforms and geophenomena. In addition, Linked GeoData[115] from LOD is used for a visual representation of geographical information for institution and platforms. The SKOS ontology of the GCMD science keywords[116] uses concepts for the tagging of product types and geophenomena. A substantial retrievable publication collection mainly about earth gravity research is also included of the GFZ ISDC Semantic Web.[117]

Conclusion and future work

By combining and integrating Semantic Web approaches, appropriate Web standards and LOD data, the resulting approach has the potential to play an important role in meeting the challenges of interoperability and sharing in the geo and space science domains.

Prior to the development of the GFZ ISDC Semantic Web, there was no common and unique interoperable

[105] http://search.iugonet.org/iugonet.

[106] http://www.iugonet.org/en/software.html.

[107] https://www.espas-fp7.eu/portal/index.html.

[108] http://isdc.gfz-potsdam.de.

[109] http://rz-vm125.gfz-potsdam.de/drupal/.

[110] http://virtuoso.openlinksw.com/.

[111] https://drupal.org/.

[112] http://rz-vm125.gfz-potsdam.de/drupal/.

[113] http://rz-vm30.gfz-potsdam.de/ontology/isdc_1.4.owl.

[114] http://dbpedia.org/sparql.

[115] http://linkedgeodata.org/About.

[116] http://isdc.gfz-potsdam.de/ontology/gcmd_science.skos.rdf.

[117] http://rz-vm125.gfz-potsdam.de/drupal/.

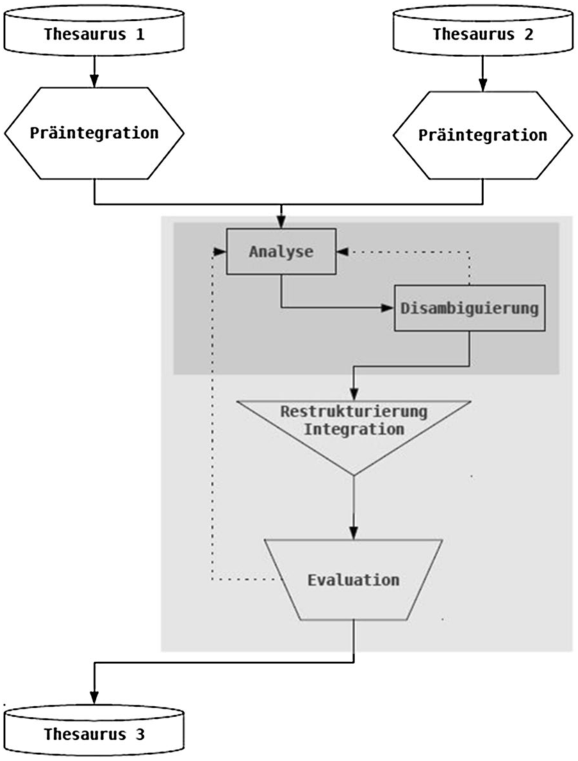


Fig. 9 Graphical user interface of the ISDC GFZ—Semantic Web-based proof of concept. The data system is realized using Virtuoso Universal Server (http://virtuoso.openlinksw.com/) and Drupal 7 (https://drupal.org/)

e-science infrastructure available to connect the Japanese IUGONET,[118] European Union ESPAS[119] and GFZ ISDC[120] data portals. We found that while each of the data portals had different data models, there were similarities of concepts. Also each system was built on a different software framework making interoperability difficult at the API level. We found that the most promising approach to achieving interoperability was to use Semantic Web-based technology. A transformation of XML schema into OWL models is possible,[121] and with lexical analysis of definitions for terms, the semantic similarity can be quantified. By storing the metadata transformed into RDF triples in appropriate databases,[122] we were able to achieve cross-system queries and reasoning. This enables the integration of multiple domain ontologies and, through references, access to the appropriate data servers. This was fully demonstrated using the SPASE/IUGONET and GFZ ISDC ontologies.[123,124]

The next important step in the realization of a Semantic Web-based e-infrastructure is the real integration of mapped or merged terminological ontologies into the data server of the involved projects. The installation of triple stores and SPARQL endpoints provides a query-based connection to the distributed and different data resources. It is planned to publish the terminological ontologies and the mapped parts in LOD. In order to overcome the limitations of Drupal 7,[125] especially to avoid the broken links which can occur between the CMS and the triple store Virtuoso Universal Server,[126] other CMS supporting Semantic Web technology, such as Ontowiki[127] and Semantic MediaWiki,[128] was validated for the use as a possible framework for the GFZ ISDC—Semantic Web data server (Seelus 2014), as shown in Sect. "GFZ ISDC: Semantic Web-based proof-of-concept platform". There is also a collaboration project with the University of Applied Sciences, Department of Information Sciences, based on the GFZ ISDC[129] for the integration of unstructured data in the Web, such as publications derived from data of the GFZ ISDC repository using entity recognition and text of speech tagging methods. Further planed activities including the validation and usage of the recently published Open Semantic Framework OSF[130] for the management of the IUGONET data repository will also focus on the efficiency of the ontological approach and a performance comparison between appropriate relational database management systems and triple stores.

The main result from this work shows that the Semantic Web, with multilingual terminological ontologies, can establish a new collaborative science culture in the Web age.

Abbreviations

EU: European Union; ESPAS: Near-Earth Space Data Infrastructure for e-Science (project); IUGONET: Inter-University Upper Atmosphere Global Observation Network (project); GFZ ISDC: German Research Centre for Geosciences Information System and Data Center; RDF: Resource Description Framework; SKOS: Simple Knowledge Organization System; RDFS: Resource Description Framework Schema; OWL: Ontology Web Language; WDC: World Data Center; WDS: World Data System; DIF: Directory Interchange Format; ISO: International Organization for Standardization; GCMD: Global Change Master Directory; NASA: National Aeronautics and Space Administration; SPASE: Space Physics Archive Search and Extract; XML: Extensible Markup Language; URI:

[118] http://search.iugonet.org/iugonet.

[119] https://www.espas-fp7.eu/portal/.

[120] http://isdc.gfz-potsdam.de.

[121] http://rhizomik.net/html/redefer/#XML2RDF.

[122] http://d2rq.org/.

[123] http://isdc.gfz-potsdam.de/ontology/spase_keywords.owl.

[124] http://isdc.gfz-potsdam.de/ontology/gcmd_science.skos.rdf.

[125] https://drupal.org/.

[126] http://virtuoso.openlinksw.com/.

[127] http://aksw.org/Projects/OntoWiki.html.

[128] https://www.semantic-mediawiki.org/wiki/Semantic_MediaWiki.

[129] http://isdc.gfz-potsdam.de.

[130] http://opensemanticframework.org/

Uniform Resource Identifier; ORCID: Open Researcher and Contributor ID; DOI: Digital Object Identifier; LOD: Linked Open Data; GEMET: General Multilingual Environmental Thesaurus; SOAP: Simple Object Access Protocol; REST: Representational State Transfer; UDDI: Universal Description, Discovery and Integration; OAI-PMH: Open Archive Initiative-Protocol for Metadata Harvesting; OGC: Open Geospatial Consortium; CSW: Web Catalogue Service; SPARQL: SPARQL Protocol and RDF Query Language; API: Application Programming interface; WWW: World Wide Web; W3C: WWW Consortium; HTML: Hypertext Markup Language; URL: Unique Resource Identifier; HTTP: Hypertext Transfer Protocol; RDFS: RDF Schema; RIF: Rule Interchange Format; IRS: Internationalized Resource Identifier; SWEET: Semantic Web for Earth and Environmental Terminology; SSN: Semantic Sensor Net ontology; FOAF: Friend-Of-A-Friend ontology; CHAMP: Challenging Minisatellite Payload; GRACE: Gravity Recover and Climate Experiment; GNSS: Global Navigation Satellite Systems; GGP: Global Geodynamic Project; GGOS: Global Geodetic Observing System; OSF: Open Semantic Framework.

Authors' contributions

BR, FB and GN designed and implemented the GFZ ISDC—Semantic Web-based proof-of-concept platform and carried out the experiments. FB installed and configured the different components of the Semantic Web platform. BR created the ISDC data model and developed the ISDC ontology network. BR and FB transferred the ISDC metadata into OWL entities. GK developed the software for the merging of terminological ontologies. SS developed a method for the mapping of domain ontologies and created the mapping of the ESPAS and ISDC ontology. BR, GN, TY, YK and IG carried out the interpretation of the ontology mapping and merging results. TY, YK, AY and TH participated in the development of the IUGONET data model, the adoption of the SPASE framework and the implementation of the metadata server. TY, TH, MH and AB participated in the conceptual work for the mashup experiments of IUGONET, ESPAS and ISDC platforms. IG participated in the modeling of the ESPAS metadata and created the ESPAS ontology. TK, TY and GN conceived of the study, participated in its design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

Author details

[1] GFZ German Research Centre for Geosciences, Potsdam, Germany. [2] University of Applied Sciences Potsdam, Potsdam, Germany. [3] Kyoto University, Kyoto, Japan. [4] Nagoya University, Nagoya, Japan. [5] Rutherford Appleton Laboratory, Chilton, Oxfordshire, UK. [6] National Observatory of Athens, Athens, Greece. [7] University of Massachusetts, Lowell, MA, USA. [8] University of California, Los Angeles, CA, USA.

Acknowledgements

We wish to acknowledge the work of the Bachelor students of the University of Applied Sciences Potsdam, Department of Information Sciences, for their contribution to the success of the GFZ ISDC—Semantic Web-based proof-of-concept project. Especially, we are grateful for the work of Juliane Pilgrim for the translation of the GCMD science keywords into the SKOS format. Furthermore, we would like to thank Christoph Seelus for the evaluation of Semantic Web frameworks for scientific data management and his support in system administration and maintenance work. This study was supported partly by JSPS Invitation Fellowship grant L15524, JSPS grant 25287128 and 15H05815, and 2016 Kyoto University "So-cho Sairyo Keihi".

Competing interests

The authors declare that they have no competing interests.

Received: 22 April 2016 Accepted: 6 October 2016
Published online: 14 November 2016

References

Abe S, Umemura N, Koyama Y, Tanaka Y, Yagi M, Yatagai A, Shinbori A, UeNo S, Sato Y, Kaneda N (2014) Progress of the IUGONET system—metadata database for upper atmosphere ground-based observation data. Earth Planets Space 66:133. https://earth-planets-space.springeropen.com/articles/10.1186/1880-5981-66-133

Allemang D, Hendler J (2008) Semantic web for the working ontologist, modeling in RDF, RDFS and OWL. Elsevier. ISBN-13: 978-0-12-373556-0

Apache Software Foundation, Apache Jena (2011–2014). http://jena.apache.org/

Apache Software Foundation, Apache Solr. https://lucene.apache.org/solr/

Apache Stanbol. https://stanbol.apache.org/

Berners-Lee TJ et al (1992) "World-wide web: information universe". Electronic Publishing: Research, Applications and Policy, p 4. http://pdfcollector.com/World-Wide-Web:-The-Information-Universe.html

Berners-Lee T (2006-07-27) Linked data—design issues. W3C. Retrieved 2010-12-18. http://www.w3.org/DesignIssues/LinkedData.html

Brickley D, Guha RV (2014) RDF schema 1.1, W3C recommendation. http://www.w3.org/TR/rdf-schema/

Brickley D, Miller L (2014) FOAF vocabulary specification 0.99. http://xmlns.com/foaf/spec/

CHAMP (Challenging Mini-satellite Payload). http://www.gfz-potsdam.de/champ

Christian B, Tom H, Tim B-L (2009) Linked data—the story so far. Int J Seman Web Inf Syst 5 (3):1–22. doi:10.4018/jswis.2009081901. ISSN 1552-6283. http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf

D'Arcus B, Giasson F (2009) Bibliographic ontology specification. http://bibliontology.com/

DBpedia data sets (web site). http://wiki.dbpedia.org/Datasets

DBpedia (web site). http://dbpedia.org/About

Directory Interchange Format (DIF) Writer's Guide (2013) Global change master directory. National Aeronautics and Space Administration. http://gcmd.gsfc.nasa.gov/add/difguide/index.html

D-Net software toolkit. http://www.d-net.research-infrastructures.eu/

Drupal (Content Management System). https://drupal.org/

DUARASPACE DSpace. http://www.dspace.org/

D2RQ, Accessing Relational Databases as Virtual RDF Graphs. http://d2rq.org/

Earth Science Data System Working Groups (ESDSWG). https://earthdata.nasa.gov/esdswg

Erfurt SWF. http://aksw.org/Projects/Erfurt.html

ESPAS data system. https://www.espas-fp7.eu/portal/

ESPAS data system. https://www.espas-fp7.eu/portal/index.html

ESPAS, the near-Earth space data infrastructure for e-Science: architecture, data model and first release (2013). http://www.espas-fp7.eu/index.php/publications/papers-manuals/file/52-espas-manual-for-espas-release-2

European Commission, Research & Innovation, Research Infrastructures (2014) http://ec.europa.eu/research/infrastructures/index_en.cfm

General Multilingual Environmental Thesaurus GEMET (2012). http://www.eionet.europa.eu/gemet/

GRACE (Gravity Recover And Climate Experiment). http://www.gfz-potsdam.de/grace

GeoNames. http://www.geonames.org/

GFZ Information System and Data Center—Semantic web based proof of concept. http://rz-vm125.gfz-potsdam.de/drupal/

GGP (Global Geodynamic Project). http://www.eas.slu.edu/GGP/ggphome.html

GGOS (Global Geodetic Observing System). http://www.ggos.org/

GNSS (Global Navigation Satellite Systems). http://www.gfz-potsdam.de/forschung/ueberblick/departments/department-1/

Gruber T (1995) Toward principles for the design of ontologies used for knowledge sharing. Int J Hum Comput Stud 43 (5-6): 907–928. doi:10.1006/ijhc.1995.1081. http://tomgruber.org/writing/onto-design.pdf

Hapgood M, Iyemori T (2013) Memorandum of understanding between near Earth Space Data Infrastructure or e-Sciences (ESPAS) and Inter-university Upper atmosphere Global Observation Network (IUGONET), , http://www.iugonet.org/doc/MOU-ESPAS-IUGONET.pdf

Hebeler J, Fisher M, Blace R, Perez-Lopez A, Dean M (2009) Semantic web programming. Wiley, Indianapolis. ISBN: 978-0-470-41801-7

Hey T, Tansley S, Tolle K (2009) The fourth paradigm: data-intensive science discovery. Microsoft Corporation. ISBN 978-0-9825442-0-4. http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf

Hitzler P, Krötzsch M, Rudolph S, Sure Y (2008) Semantic Web Grundlagen. Springer, Berlin. ISBN 978-3-540-33993-9, doi:10.1007/978-3-540-33994-6

HTTP—hypertext transfer protocol. http://www.w3.org/Protocols/

IDC White Paper (2011) The diverse and exploding digital universe. http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf

Information system and data center of the Helmholtz Centre Potsdam—GFZ German Research Centre for Geosciences. http://isdc.gfz-potsdam.de

International DOI Foundation. The DOI system ISO 26324. http://www.doi.org/

Internationalized Resource Identifiers (IRIs). http://www.w3.org/International/O-URL-and-ident.html

ISDC ontology network, version 1.4. http://rz-vm30.gfz-potsdam.de/ontology/isdc_1.4.owl

ISO 19101:2002. Geographic information—reference model. http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=26002

ISO 19109:2005. Geographic information—rules for application schema. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=39891

ISO 19115-1:2014. Geographic information—metadata—part 1: fundamentals. http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=53798

ISO 19156:2011. Geographic information—observations and measurements. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=32574

IUGONET data analysis software (UDAS). http://www.iugonet.org/en/software.html

IUGONET—Inter-university Upper atmosphere Global Observation Network. http://www.iugonet.org/en/

IUGONET metadata DB for upper atmosphere data system. http://search.iugonet.org/iugonet

IUGONET metadata format. http://www.iugonet.org/en/mdformat.html

Jentzsch A, Cyganiak R, Bizer C (2011) Linked data—state of the cloud. http://lod-cloud.net/state/

King T, Thieman J, Aaron Roberts D (2010) Earth science informatics, vol 3, no 1–2, pp 67–73. SPASE 2.0: a standard data model for space physics. http://link.springer.com/article/10.1007%2Fs12145-010-0053-4

Kneitschel G (2013) Methodenentwicklung und Implementierung zum Thesaurusmerging am Beispiel naturwissenschaftlicher Keywordontologien, Bachelor of Arts Thesis. University of Applied Sciences Potsdam, Department of Information Sciences. https://www.fh-potsdam.de/studieren/informationswissenschaften/forschung-und-entwicklung/publikationen/bachelorarbeiten-von-2013/

Lehmann J, Isele R, Jakob M, Jentzsch A, Kontokostas D, Mendes PN, Hellmann S, Morsey M, van Kleef P, Auer S, Bizer C (2012) DBpedia—a large-scale, multilingual, knowledge base extracted from Wikipedia, semantic web 1. 1–5 1, IOS Press. ISSN 1570-0844. http://svn.aksw.org/papers/2013/SWJ_DBpedia/public.pdf

Linked data community, linked data—connect distributed data across the web. http://linkeddata.org/

LinkedGeoData.org (web site). http://linkedgeodata.org/About

Linked Open Data—data hub. http://datahub.io/en/dataset?q=linked+open+data

Linked sensor data (Kno.e.sis). http://datahub.io/dataset?q=Kno.e.sis

LinkedIn (Social Network). https://www.linkedin.com/

Mende V, Ritschel B, Freiberg S, Palm H, Gericke L (2008) Directory interchange format (DIF) Metadata and Handling at the German Research Center for geosciences' Information System and Data Center, Geoinformatics 2008-data to knowledge, Proceedings, June 11–13, Potsdam, Germany. Scientific investigation report 2008-5172, US Department of the Interior, US Geological Survey, S. 43-46. ISBN 978-1-4113-2279-0. http://pubs.usgs.gov/sir/2008/5172/sir2008-5172.pdf

NASA space flight & astronaut data in RDF. http://datahub.io/dataset/data-incubator-nasa

Near earth space data infrastructure for e-science ESPAS project funded by European Union's Seventh Framework Program. http://www.espas-fp7.eu/

Noy NF, McGuinness DL (2001) Ontology development 101: a guide to creating your first ontology. Stanford University, Stanford, CA, 94305. http://www.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness.pdf

OGC observations and measurements. http://www.opengeospatial.org/standards/om

OGC Catalogue Service. http://www.opengeospatial.org/standards/cat

Olsen LM, Major G, Shein K, Scialdone J, Ritz S, Stevens T, Morahan M, Aleman A, Vogel R, Leicester S, Weir H, Meaux M, Grebas S, Solomon C, Holland M, Northcutt T, Restrepo RA, Bilodeau R (2013) NASA/Global Change Master Directory (GCMD) earth science keywords. Version 8.0.0.0.0. http://gcmd.nasa.gov/learn/keyword_list.html

OntoWiki. http://aksw.org/Projects/OntoWiki.html

Open archives initiative protocol for metadata Harvesting http://www.openarchives.org/pmh/

OPENLINK virtuoso universal server. http://virtuoso.openlinksw.com/

OpenRDF Sesame. http://www.openrdf.org/

Open researcher and contributor ID ORCID. http://orcid.org/

Open Semantic Framework. http://opensemanticframework.org/

OWL Working Group (2012) OWL web ontology language. http://www.w3.org/2001/sw/wiki/OWL

Paulquek, Ontology. http://queksiewkhoon.tripod.com/ontology_01.pdf

Pfeiffer S (2010) Entwicklung einer Ontologie für die wissensbasierte Erschließung des ISDC-Repository und die Visualisierung kontextrelevanter semantischer Zusammenhänge, Master of Engineering Thesis. urn:nbn:de:gbv:519-thesis2010-0139-4. http://digibib.hs-nb.de/file/dbhsnb_derivate_0000000780/Masterarbeit-Pfeiffer-2010.pdf

Postnuke open source application framework. http://www.pn-cms.de/

Protégé 3. http://protegewiki.stanford.edu/wiki/Protege_Desktop_Old_Versions#Protege_3

Protégé 4. http://protegewiki.stanford.edu/wiki/Protege_Desktop_Old_Versions#Protege_4

RDF Working Group (2004) (2014) Resource description framework (RDF). http://www.w3.org/RDF/

ResearchGate (Social Network). https://www.researchgate.net

Rhizomik framework for the transformation XML2RDF. http://rhizomik.net/html/redefer/#XML2RDF

Ritschel B, Neher G (2013) Enhancing interoperability in the GFZ ISDC ontology by embedding SKOS transformed NASA's GCMD keywords for describing and connecting entities and VIAF authority data for referencing personal and corporate names, AGU fall meeting, San Francisco, 5–9 December 2013. http://gfzpublic.gfz-potsdam.de/pubman/faces/viewItemOverviewPage.jsp?itemId=escidoc:385322:1

Ritschel B, Mende V, Palm H, Gericke. L, Freiberg S, Kopischke R, Bruhns C (2008) The German Research Center for Geosciences' Information System and Data Center—portal to geoscientific data, information and knowledge, geoinformatics—data to knowledge, Proceedings, June 11–13, Potsdam, Germany, Scientific investigation report 2008-5172. US Department of the Interior, US Geological Survey, S. 33-35. ISBN 978-1-4113-2279-0. http://pubs.usgs.gov/sir/2008/5172/sir2008-5172.pdf

Ritschel B, Pfeifer S, Mende V, Freiberg S (2008) Semantic web technologies for value added services at the German Research Center for Geosciences' Information System and Data Center, Geoinformatics 2008-data to knowledge, Proceedings, June 11–13, Potsdam, Germany, Scientific Investigation Report 2008-5172, US Department of the Interior, US Geological Survey, S. 66-69. ISBN 978-1-4113-2279-0. http://www.pubs.usgs.gov/sir/2008/5172/sir2008-5172.pdf

Ritschel B, Mende V, Gericke L, Kornmesser R, Pfeiffer S (2012) Web approach for ontology-based classification, integration, and interdisciplinary usage of geoscience metadata. Data Sci J. doi:10.2481/dsj.IGY-014. https://www.jstage.jst.go.jp/article/dsj/11/0/11_IGY-014/_pdf

Schildbach S (2013) Ontology Merging in den Geowissenschaften, Bachelor of Arts Thesis. University of Applied Sciences Potsdam, Department of Information Sciences. https://www.fh-potsdam.de/studieren/informationswissenschaften/forschung-und-entwicklung/publikationen/bachelorarbeiten-von-2013/

Seelus C (2014) Semantic web CMS für wissenschaftliches Datenmangement, Bachelor of Arts Thesis. University of Applied Sciences Potsdam, Department of Information Sciences. https://www.fh-potsdam.de/studieren/informationswissenschaften/forschung-und-entwicklung/publikationen/bachelorarbeiten-von-2014/

Semantic MediaWiki. https://www.semantic-mediawiki.org/wiki/Semantic_MediaWiki

Semantic web stack. http://www.w3.org/2004/Talks/1117-sb-gartnerWS/slide18-0.html

Semantic Sensor Network (SSN) ontology. http://www.w3.org/2005/Incubator/ssn/wiki/Semantic_Sensor_Net_Ontology

Semantic Web for Earth and Environmental Terminology (SWEET) ontology. http://sweet.jpl.nasa.gov/
Shadbolt N, Hall W, Berners-Lee T (2006) The semantic web revisited. IEEE Intell Syst J 96–101. http://eprints.soton.ac.uk/262614/1/Semantic_Web_Revisted.pdf
SPARQL query language for RDF (2008) (2013). http://www.w3.org/TR/rdf-sparql-query/
SPASE consortium, a space and solar physics data model, version: 2.2.2, release date: 2011-02-27. http://www.spase-group.org/data/dictionary/spase-2_2_2.pdf
SPASE metadata model. http://www.spase-group.org/data/schema/
Terminological GCMD science keyword ontology. http://isdc.gfz-potsdam.de/ontology/gcmd_science.skos.rdf
Terminological SPASE ontology. http://isdc.gfz-potsdam.de/ontology/spase_keywords.owl
TerraSAR-X (TSX). http://terrasar-x.gfz-potsdam.de/
Typo3 (Content Management System). http://typo3.org/
Virtuoso SPARQL Query Editor, DBpedia SPARQL endpoint. http://dbpedia.org/sparql
W3C (1994–2012) SKOS simple knowledge organization system. http://www.w3.org/2004/02/skos/
W3C Working Group Note 5 February 2013, RIF overview, 2nd edn. http://www.w3.org/TR/rif-overview/
World Data Center for Geomagnetism, Kyoto. http://wdc.kugi.kyoto-u.ac.jp/
WWW consortium, W3C. http://www.w3.org/
Ximdex. http://www.ximdex.com/
XML schema for the IUGONET common metadata format. http://www.iugonet.org/data/schema/
Yatagai A, Sato Y, Shinbori A, Abe S, UeNo S (2015) The capacity-building and science-enabling activities of the IUGONET for the solar-terrestrial research community. Earth Planets Space 67:2. https://earth-planets-space.springeropen.com/articles/10.1186/s40623-014-0170-2