

Integration of Geological Data as a Knowledge Graph

Evgeny Cherkashin, Oksana Lunina, Tatiana Cherkashina,
Vadim Pellinen, Anton Gladkov

Matrosov Institute for System Dynamics and Control Theory,
SB RAS, Irkutsk, Russia
Institute of the Earth's Crust, SB RAS, Irkutsk, Russia

eugeneai@icc.ru

IWCI-2021, Baikalsk, Russia

Problem statement

Integration the collected geological data, so they can be

- ❑ easily accessed
- ❑ of a standardized representation
- ❑ integrated with the data sources of the similar structure (access way and data representation paradigm)
- ❑ thus, reuse the existing data sources for the domain modeling
- ❑ processed within Big Data paradigm
- ❑ represented as cartographic work (map) dynamically with/without need of GIS software installation

The requested requirements are fulfilled with the present Semantic Web technologies.

Semantic web technologies & Knowledge graphs

Semantic Web (WEB 3.0) is characterized with

- ❑ Technological basis, oriented to the web
- ❑ Standardized data formats, storage, and processing
- ❑ Open principles of data publishing
- ❑ Services for data storage and access provision
- ❑ Generalized and special user interfaces are used for data presentation

For the Knowledge Graphs (KG), the following is of interest.

- ❑ Converged notions **data** and **knowledge** as something is **known**
- ❑ Contain data, relations, and metadata (vocabularies)
- ❑ Distinguished **node filling in** and **processing** graph triples, e.g., within SPARQL queries with UPDATEs
- ❑ Allow **postpone** the formal definition of a schema
- ❑ Three types of graph schemata: **semantic** (aimed at generalization), **validating** (e.g. semantics, **completeness** w.r.t. sets of relations), and **emergent** (infer a set of generalized structures and **reconstruct** the KG).

Knowledge graph: Validating semantic example

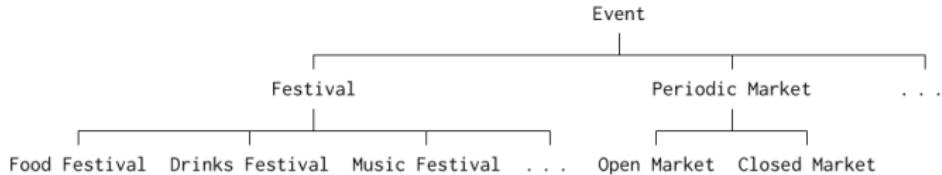


Fig. 10. Example class hierarchy for Event

Table 2. Definitions for sub-class, sub-property, domain and range features in semantic schemata

Feature	Definition	Condition	Example
SUBCLASS	$c \xrightarrow{\text{subc. of}} d$	$x \xrightarrow{\text{type}} c \text{ implies } x \xrightarrow{\text{type}} d$	$\text{City} \xrightarrow{\text{subc. of}} \text{Place}$
SUBPROPERTY	$p \xrightarrow{\text{subp. of}} q$	$x \xrightarrow{p} y \text{ implies } x \xrightarrow{q} y$	$\text{venue} \xrightarrow{\text{subp. of}} \text{location}$
DOMAIN	$p \xrightarrow{\text{domain}} c$	$x \xrightarrow{p} y \text{ implies } x \xrightarrow{\text{type}} c$	$\text{venue} \xrightarrow{\text{domain}} \text{Event}$
RANGE	$p \xrightarrow{\text{range}} c$	$x \xrightarrow{p} y \text{ implies } y \xrightarrow{\text{type}} c$	$\text{venue} \xrightarrow{\text{range}} \text{Venue}$

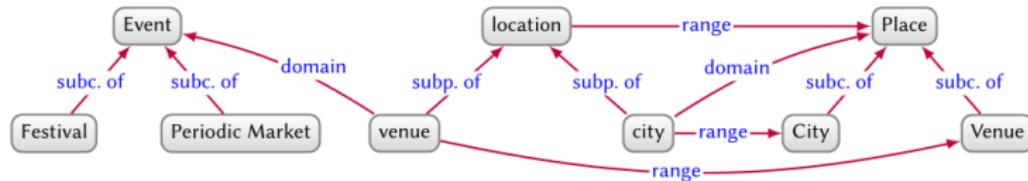
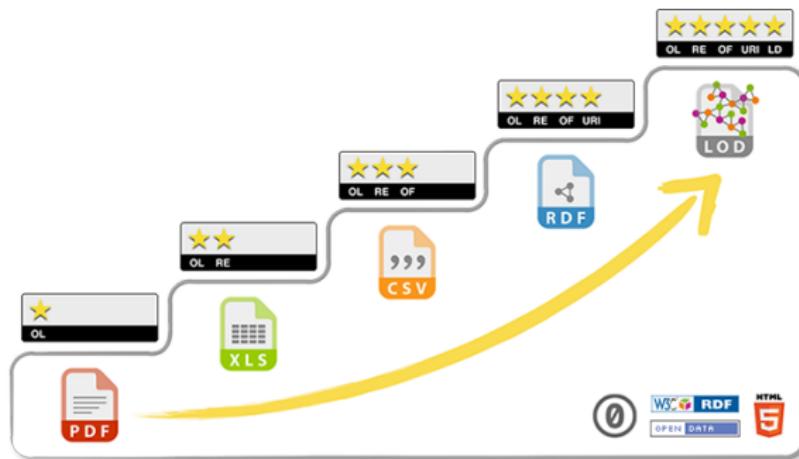


Fig. 11. Example schema graph describing sub-classes, sub-properties, domains, and ranges

Linked Open Data (LOD) star evaluation

Data are available in

- 1* any format **openly**
- 2* a **structured format**, such as Microsoft Excel file format (.xls)
- 3* a **non-proprietary structured format**, such as .csv
- 4* **W3C standards**, like using RDF and employing URIs
- 5* a hypercontent form **having links to other Linked Open Data sources**



The Aim and the plan

The **aim of the project** is to represent geological data accumulated at IEC SB RAS and other institutes into the Semantic Web infrastructure.

The **main problems** to be solved are

1. Design a web-based GIS system, representing data from SPARQL endpoints
2. Convert the existing data into RDF adhering LOD
3. Implement natural language query interface using GeoBase (© Borland) with a conversion into a SPARQL query
4. Bidirectional versioned data transfer between user GIS (QGIS, OSM Mapnik) and the SW storage
5. Implement various analytical functionality for domain problem solving

Related works and techniques review

LinkedGeoData project

The project¹ was to represent OpenStreetMap (OSM) data as a KG,

- ❑ Resembles the DBpedia project formalizing Wikipedia data but over the OSM database
- ❑ Converts SM data into RDF adhering LOD
- ❑ Designed an ontology for object georeferencing (nodes)
- ❑ Related the object to DBpedia, GeoNames, icon sets
- ❑ Developed a taxonomy of the objects on a various levels (Road → Way (list of nodes))
- ❑ Stated the relations between nodes and ways defining complex objects
- ❑ Implemented REST and SPARQL (does not work now) endpoints for actual data
- ❑ Had a live updates services from OSM changesets.

¹Claus Stadler, Jens Lehmann, Konrad Höffner, Sören Auer. LinkedGeoData: A core for a web of spatial open data. Semantic Web 3 (2012) 333–354. DOI:[10.3233/SW-2011-0052](https://doi.org/10.3233/SW-2011-0052)

LinkedGeoData project

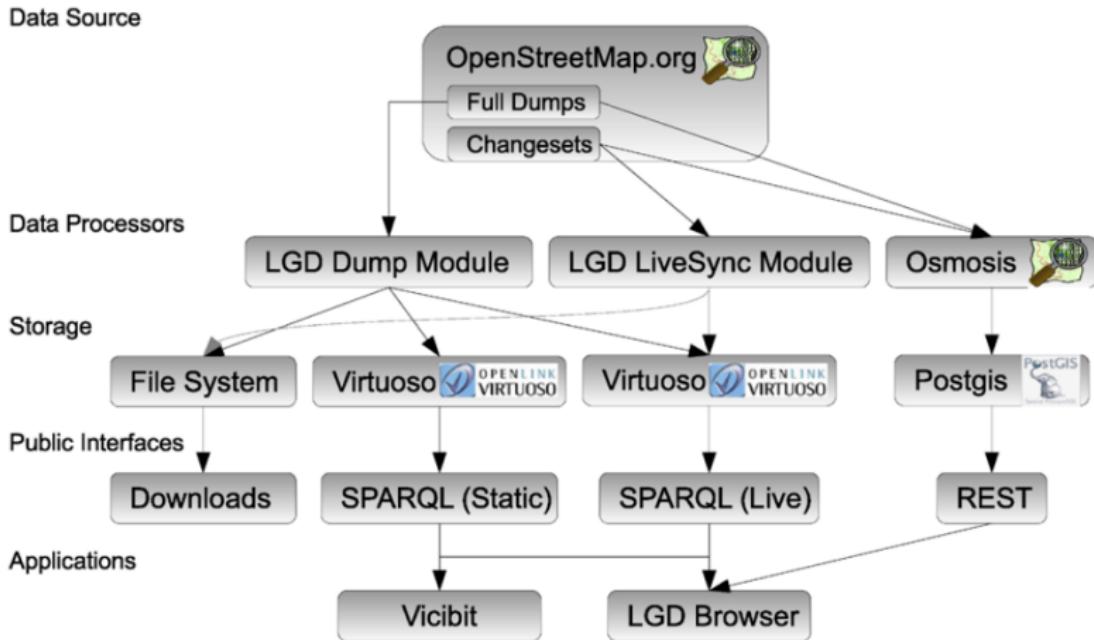


Fig. 2. Overview of LinkedGeoData's architecture.

LinkedGeoData project

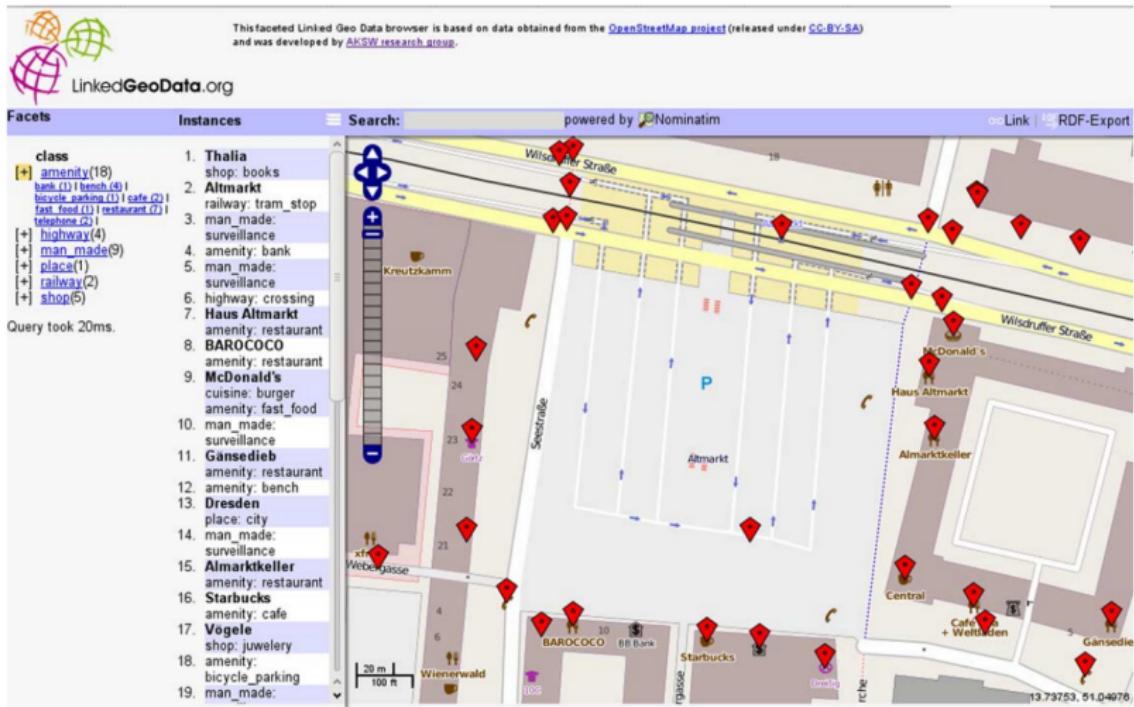


Fig. 6. LinkedGeoData Browser.

GeoLink knowledge graph

GeoLink KG²

- ❑ Includes diverse information as port calls made by oceanographic cruises, physical sample metadata, research project funding and staffing, and authorship of technical reports
- ❑ Implements LOD (4 of 5 stars) and federated SPARQL integration
- ❑ Contains 45 millions RDF triples with ontologies and geo-visualization tools
- ❑ Describes interlinked **R2R**, expeditions, **BCO-DMO**, oceanography, **IODP**, ocean floor microbiome, **MBLWHOI**, marine life papers, **SESAR**, rock samples, **DataONE**, metadata of external research, **AGU-NSF**, projects & conferences, **NGDB**, sediment geochemistry, **USAP**, Antarctica ice.

In the project, an update procedure (harvesting) is implemented to ensure the consistence of the KG w.r.t. the geo-base ontology (GBO).

²Michelle Cheatham, Adila Krisnadhi, Reihaneh Amini, Pascal Hitzler, et al (2018) The GeoLink knowledge graph, Big Earth Data, 2:2, 131-143, DOI:[10.1080/20964471.2018.1469291](https://doi.org/10.1080/20964471.2018.1469291)

GeoLink knowledge graph

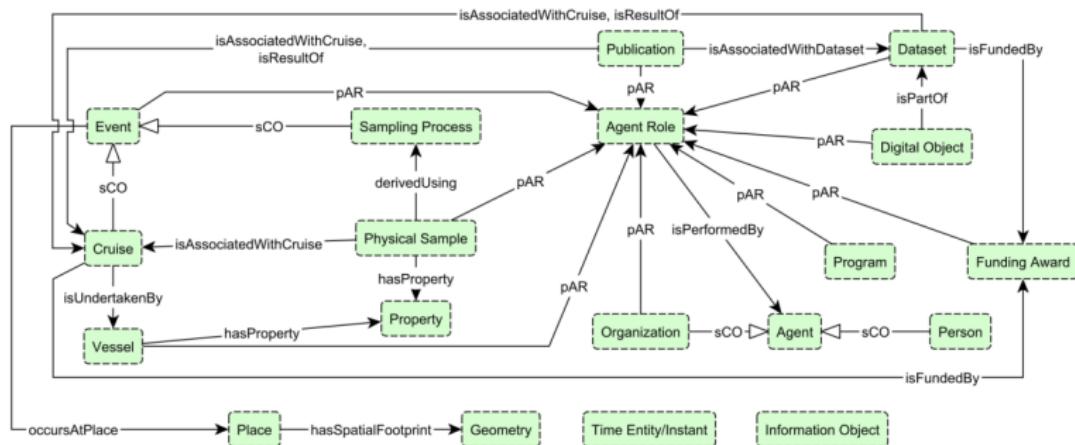


Figure 1. Schema diagram containing (almost) all patterns in the GeoLink ontology and their main links. All patterns have links to Time Entity/Instant and Information Object, but they have been omitted for clarity. **sCO**=subClassOf; **pAR**=providesAgentRole; each box is a pattern, represented by its main class.

The GeoLink knowledge graph is deployed at
<http://data.geolink.org>.

Integrating LOD into GIS

The project³ deals with developing a web GIS automatically publishing DBpedia data.

- ❑ LOD resembles Open Government Data principles
- ❑ Modules are
 - ▶ GIS is Google Map API v3
 - ▶ SPARQL used to query DBpedia
 - ▶ Viewing DBpedia data with Data Table plug-in of JQuery
- ❑ Test application allows user querying celebrities by their home town/city pointed by mouse on the Google Map.

³Tarek Abid, Hafed Zarzour. Integrating Linked Open Data in Geographical Information System. International Conference on Information Technology for Organization Development. 2014.

Integrating LOD into GIS

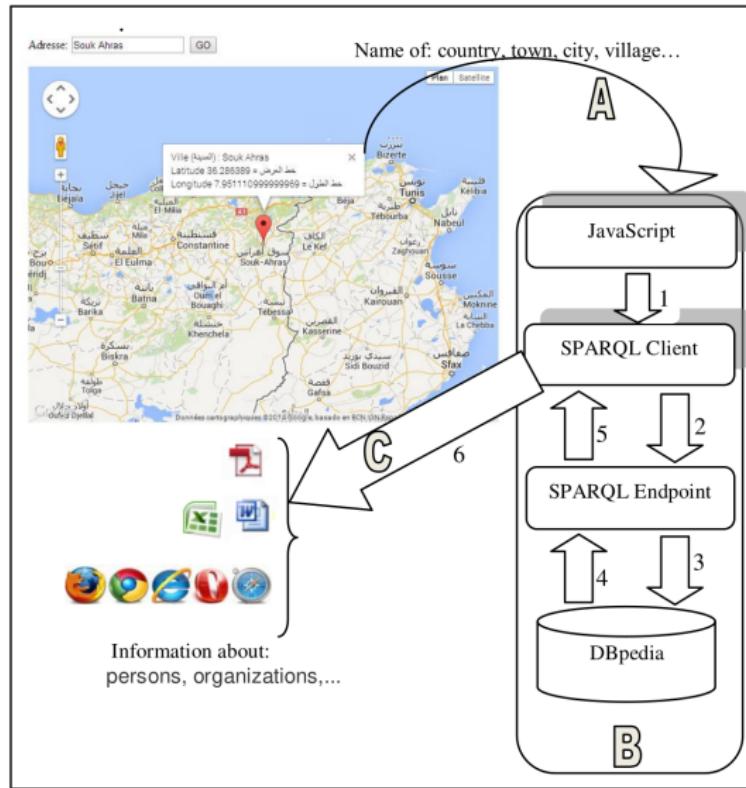


Fig. 1. System architecture.

Integrating LOD into GIS

```
PREFIX dbo: http://dbpedia.org/ontology/
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX : http://dbpedia.org/resource/
SELECT ?name ?birth ?death ?person
WHERE {
?person dbo: birthPlace :Souk Ahras .
?person dbo: birthDate ?birth .
?person foaf: name ?name .
?person dbo: deathDate ?death .
}
ORDER BY ?name
```

Fig. 3. Example of SPARQL code enabling to extract Celebrity person in Souk Ahras.

Publishing Geodata LOD in context dependent pages

The project⁴ goal is to convert existing GIS data into explicit knowledge, thus, forming a Spatial Data Infrastructure (SDI).

- ❑ Integrate existing geoportal data into a KB, including dynamic data
- ❑ Geoportal data must be LOD, e.g., HTML is enriched with RDFa
- ❑ Relation interpreters implemented as Expert systems (“building near forest”)
- ❑ Semantic enrichment of raw data to make it more usable/discoverable
- ❑ Targeting to GeoSPARQL (sfIntersects, sfOverlaps, sfTouches, sfWithin, sfContains))
- ❑ Metadata inference from the data source properties
- ❑ Test application is to integrate public services data in Mazowieckie Voivodeship of Poland, queries are realized by a limited set of keywords

⁴Adam Iwaniak, Marta Leszczuk, Marek Strzelecki, Francis Harvey, Iwona Kaczmarek. A Novel Approach for Publishing Linked Open Geodata from National Registries with the Use of Semantically Annotated Context Dependent Web Pages. International Journal of Geo-Information. 6, 252, 2017. DOI:[10.3390/ijgi6080252](https://doi.org/10.3390/ijgi6080252)

Publishing Geodata LOD in context dependent pages

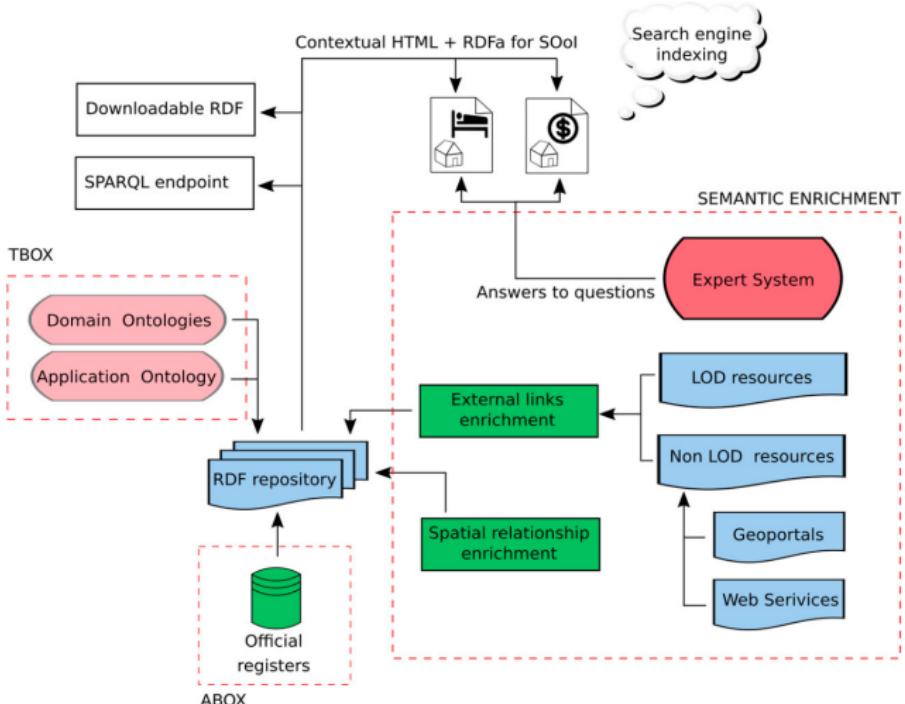


Figure 1. Concept of system architecture.

Enriching and improving the quality of linked data with GIS

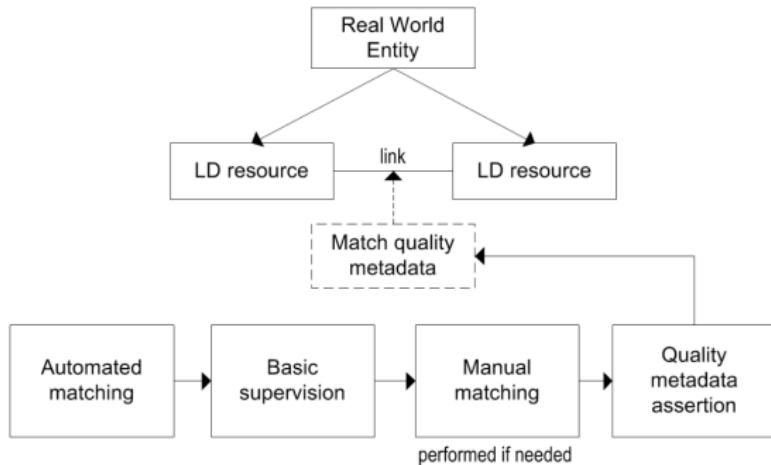


Figure 9: Basic supervision of linked data resources location quality with the use of Desktop GIS map visualization.

From an early work⁵.

⁵Adam Iwaniak, Iwona Kaczmarek, Marek Strzelecki, Jaromar Lukowicz, Piotr Jankowski. Enriching and improving the quality of linked data with GIS. DOI:[10.1515/geo-2016-0c020](https://doi.org/10.1515/geo-2016-0c020)

Enriching and improving the quality of linked data...

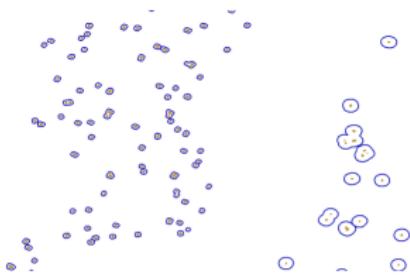


Figure 3: Analysis with different buffer distances dependent on the spatial distribution of features.

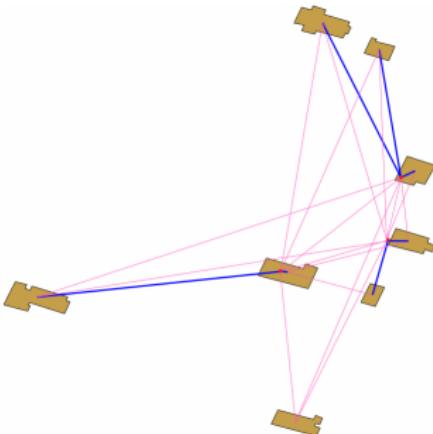


Figure 5: The most promising links between resources are highlighted.

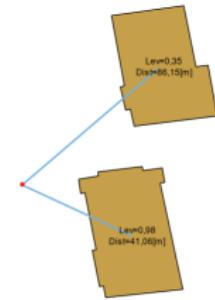


Figure 4: Two polygons related to single point object with the values of Euclidean and Levenshtein distances (Levenshtein and Euclidean distances values from first polygon are respectively 0.98 and 41.06 m; from the second polygon are 0,35 and 86,15 m).

Publishing Geodata LOD in context dependent pages

(a)

Mazowsze, since Polish
www.mazova.pl

Local spatial development plan of the city of Bialobrzegi no. XV / 81/2004
SEMANTIC METADATA

Abstract

Local Spatial Development Plan is the basis for spatial planning and is a planning document prepared for the municipality area in accordance with the Planning and Land Use Planning Act of 27 March 2003 (Journal of Laws 2003 no. 80 item 717, as amended). It establishes rules commonly applicable in the area, which are the basis for issuing administrative decisions. Legal basis - Resolution No. XV/ 81/2004 of the Council of the Town and Commune of Bialobrzegi.

Selected locations in the area

Bialobrzegi , Pierzchnia , Bialobrzegi

(b)

TOPOGRAPHIC OBJECTS INCLUDED IN THE RESOURCE AREA (GEONAMES):

Place / place / topographic object:

Place / location identifier:	7533955
Name of the place / location:	Bialobrzegi
Resource URI:	http://sws.geonames.org/7533955/

Place / place / topographic object:

Site / location identifier:	762277
Name of location / locator:	Pierzchnia
Resource URI:	http://sws.geonames.org/762277/

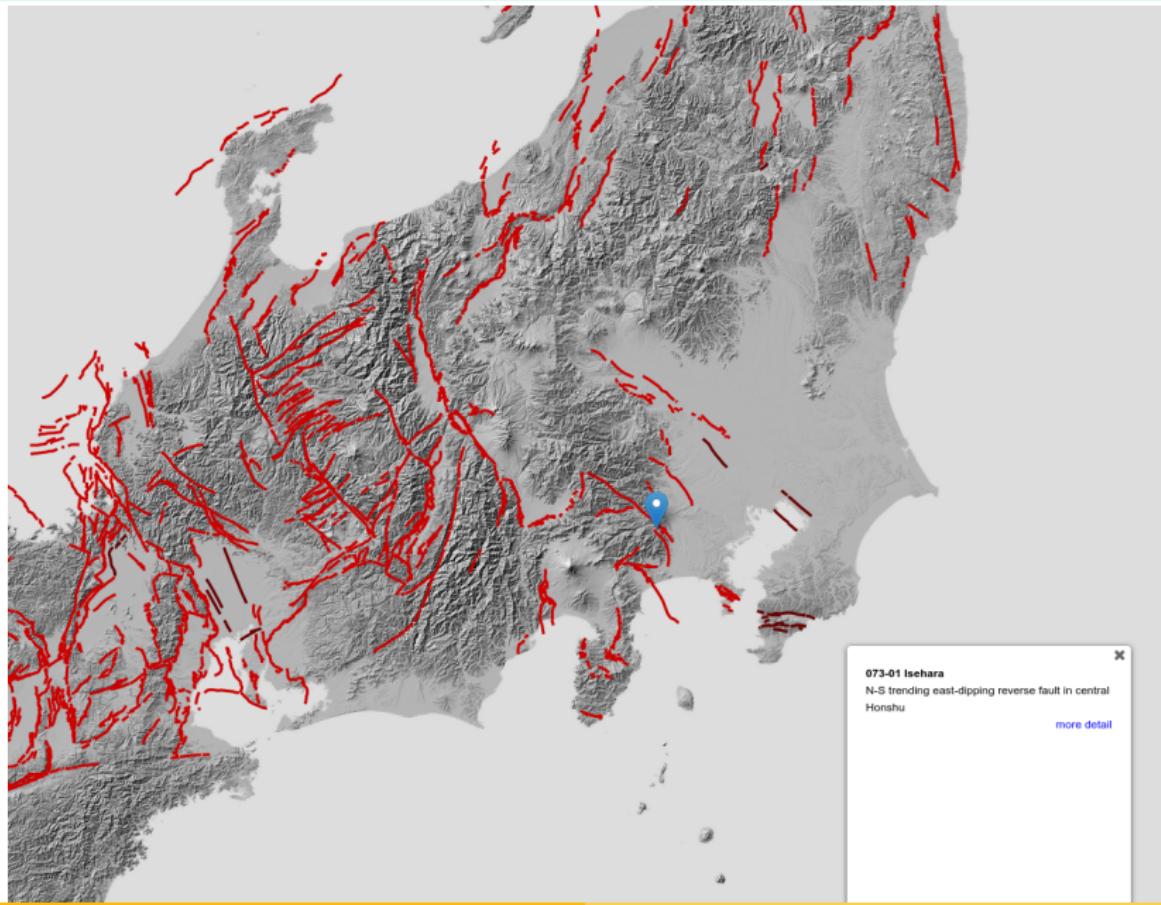
Place / place / topographic object:

Site / location identifier:	776114
Name of the place / location:	Bialobrzegi
Resource URI:	http://sws.geonames.org/776114/

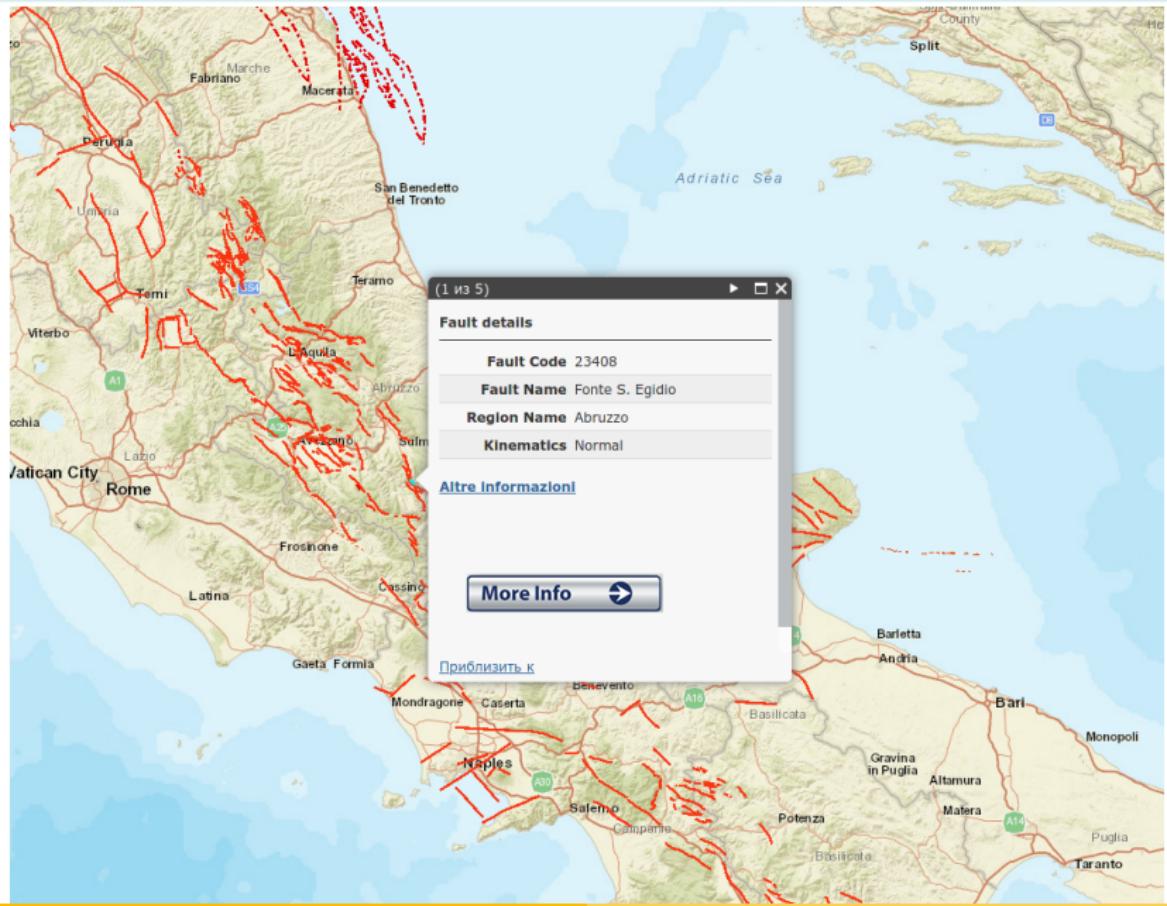
Figure 6. Fragments of HTML document, containing metadata for local spatial development plan: (a) title, basic information and bounding box of data; (b) links to Geonames features, included in the resource area.

Existing geological GIS resources

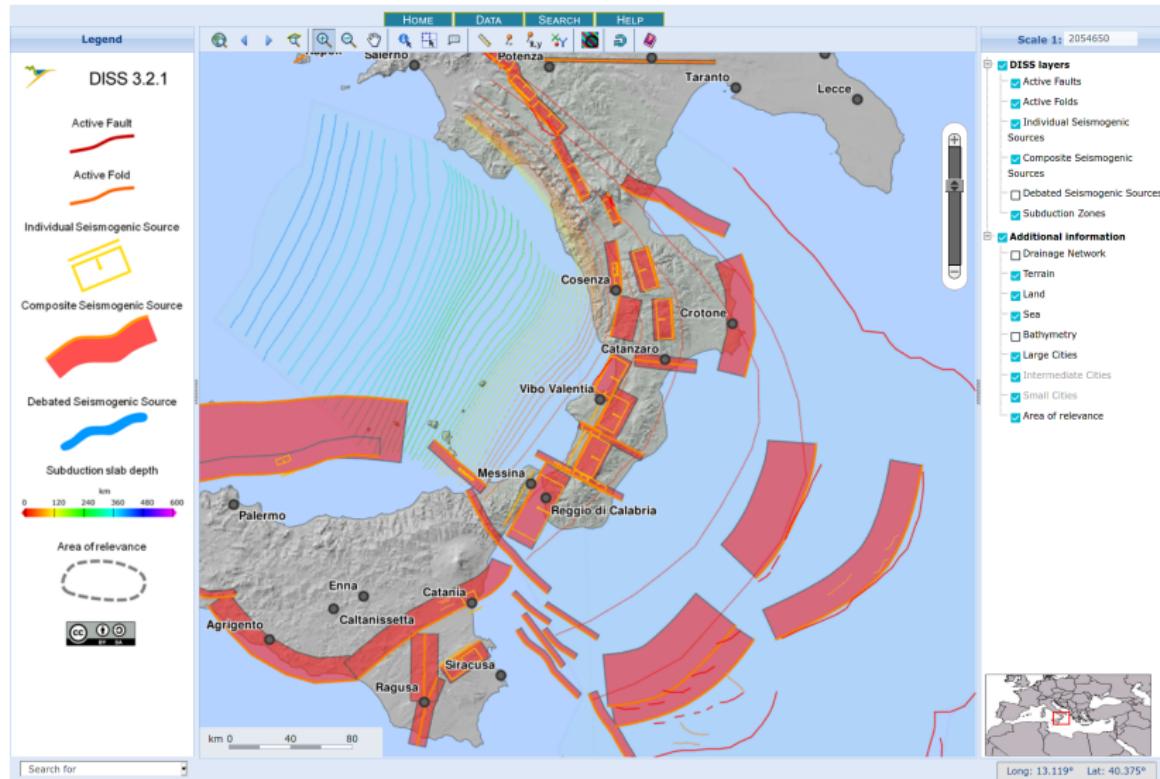
Active faults of Japan islands



Active fault of Italian Peninsula

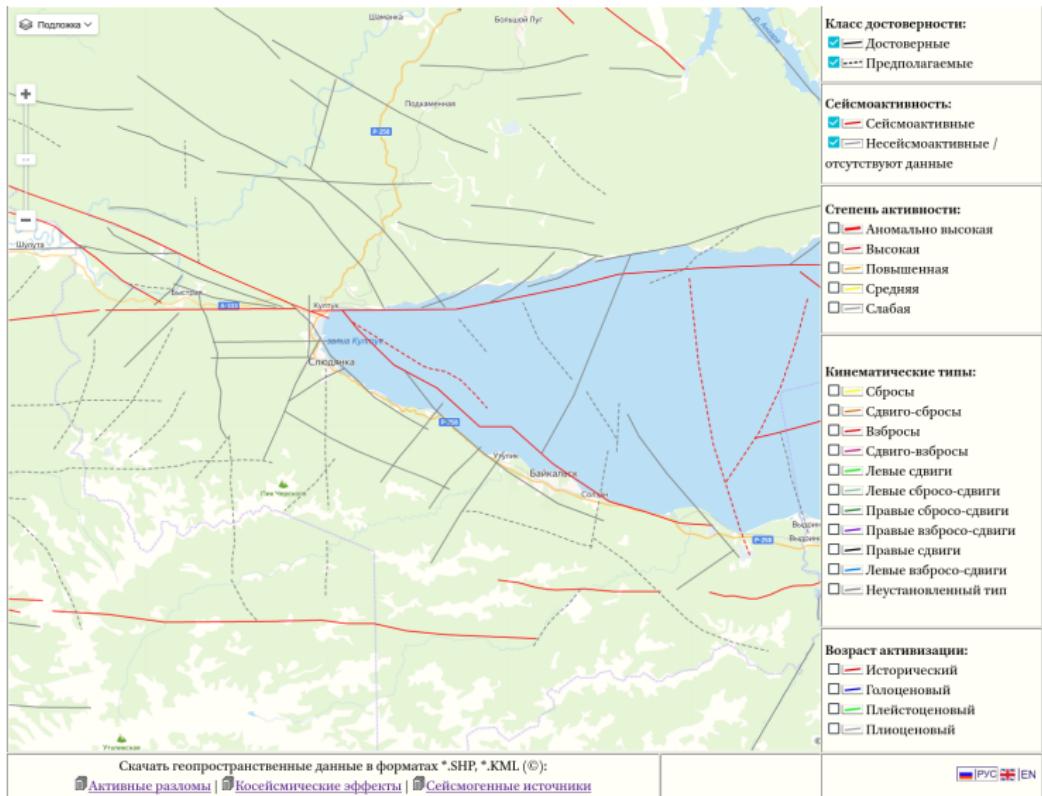


Active fault of Italian Peninsula



Resources to be presented

Active faults of the South of East Siberia



<http://activetectonics.ru/>

Resources to be presented: Active faults of the South of East Siberia

The fault data⁶ have the following properties:

- ❑ Faults are the objects having linear projection on the surface, and a depth and a slope in the lithosphere
- ❑ Various geological events are also subjects of presentation
- ❑ Any object accompanied by various valued characteristics having also characteristics, *e.g.*, unit name, measurement precision and related publications
- ❑ Images, tables, expedition trails are also used as auxiliary materials.

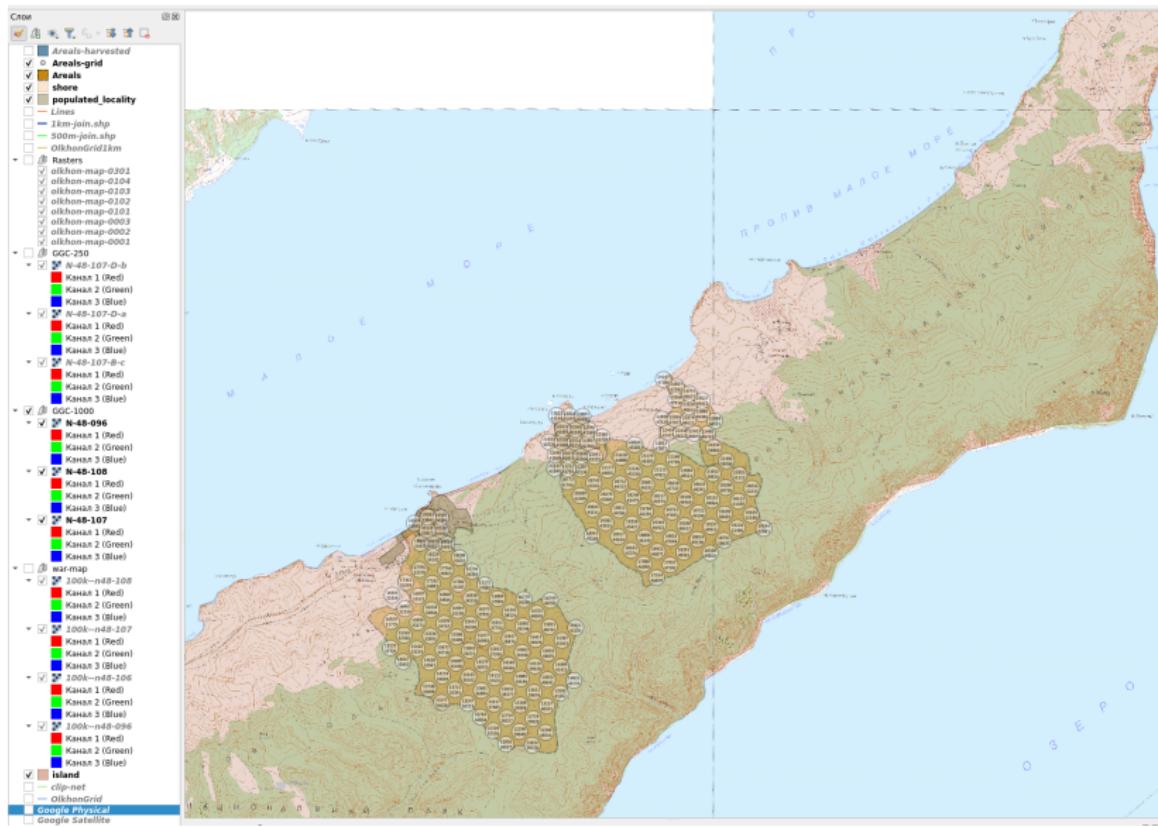
⁶Oksana V. Lunina. The digital map of the pliocene quaternary crustal faults in the southern east siberia and the adjacent northern Mongolia. Geodynamics & Tectonophysics. 2016. 7(3):407-434. DOI:[10.5800/GT-2016-7-3-0215](https://doi.org/10.5800/GT-2016-7-3-0215)

Database Structure

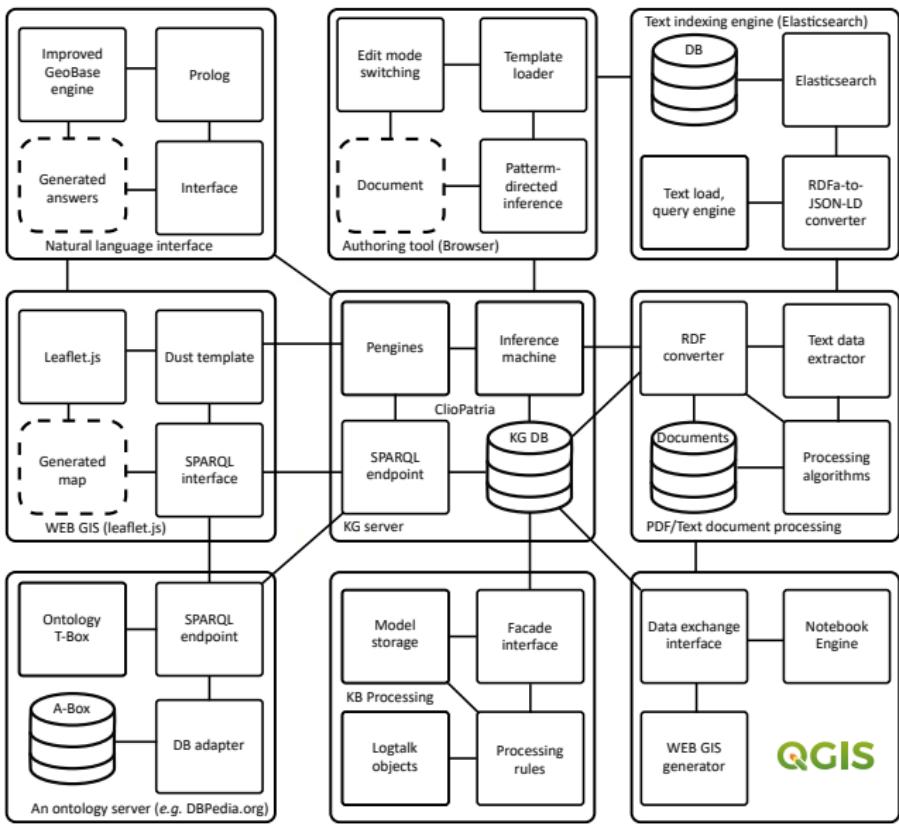
```
!table
!version 300
!charset WindowsCyrillic

Definition Table
Type NATIVE Charset "WindowsCyrillic"
Fields 72
ID Char (15) Index 1 ;
Name Char (40) Index 2 ;
Location Char (250) Index 3 ;
Strike Char (10) Index 4 ;
Strike_Q Char (2) ;
Dip_azimuth Char (10) Index 5 ;
Dip_azimuth_Q Char (2) ;
Dip_angle Char (10) Index 6 ;
Dip_angle_Q Char (2) ;
Length_km Char (10) Index 7 ;
Length_Q Char (2) ;
Depth_km Char (10) Index 8 ;
Depth_Q Char (2) ;
Width_damage_zone_km Char (10) Index 9 ;
Width_damage_zone_Q Char (2) ;
Slip_sense Char (30) Index 10 ;
Slip_sense_Q Char (2) ;
Slip_sense_Index Decimal (2, 0) ;
Total_Cenozoic_lateral_slip_m Char (20) ;
Total_Cenozoic_lateral_slip_Q Char (2) ;
Total_Cenozoic_vertical_slip_m Char (20) ;
Total_Cenozoic_vertical_slip_Q Char (2) ;
Reliability_class Decimal (1, 0) Index 11 ;
Geomorphological_features Char (250) ;
Geomorphological_grade Integer Index 12 ;
Geophysical_features Char (254) ;
Geophysical_grade Integer Index 13 ;
Engineering_geological_features Char (254) ;
Engineering_geological_grade Integer Index 14 ;
Gydrogeological_features Char (200) ;
Gydrogeological_grade Integer Index 15 ;
Structural_geological_features Char (254) ;
Structural_geological_grade Integer Index 17 ;
Paleoseismological_features Char (254) ;
Paleoseismological_grade Integer Index 18 ;
Seismological_features Char (100) ;
Seismological_grade Integer Index 19 ;
Slip_rate_features Char (10) ;
Slip_rate_grade Integer Index 20 ;
Total_activity_grade Decimal (4, 0) Index 21 ;
Activity_degree Char (25) Index 22 ;
Last_activation_age Char (30) Index 23 ;
Last_activation_age_index Decimal (1, 0) Index 24 ;
Last_historical_earthquake Char (150) Index 25 ;
Last_historical_earthquake_Q Char (2) ;
Absolute_deformation_age Char (254) ;
Absolute_deformation_age_Q Char (2) ;
Vertical_max_slip_per_event_m Float ;
Vertical_max_slip_per_event_Q Char (2) ;
Lateral_max_slip_per_event_m Float ;
Lateral_max_slip_per_event_Q Char (2) ;
Total_max_slip_per_event_m Float ;
Total_max_slip_per_event_Q Char (2) ;
Slip_rate_mm_year Char (30) ;
Slip_rate_Q Char (2) ;
Averaged_slip_rate_mm_year Float ;
Averaged_slip_rate_Q Char (2) ;
Recurrence_interval Char (10) Index 26 ;
Recurrence_interval_Q Char (2) ;
Potential_Ms_max Float ;
Potential_Ms_max_Q Char (2) ;
Potential_Mw_max Float ;
Potential_Mw_max_Q Char (2) ;
Elapsed_time_years Char (10) ;
Elapsed_time_Q Char (2) ;
Associated_CSS Char (25) ;
Associated_IGGSS Char (50) ;
Seismic_activity_of_fault Char (3) ;
```

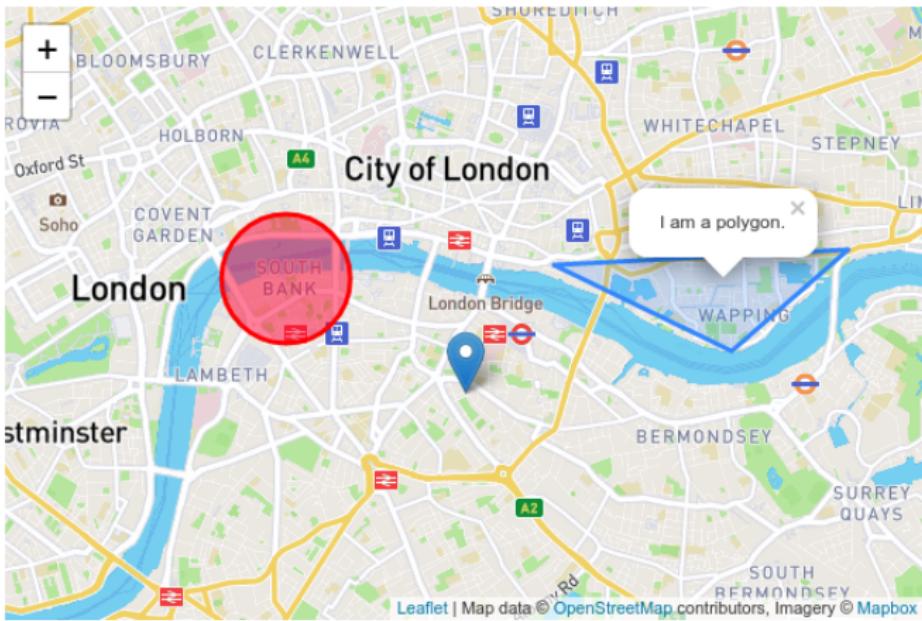
The pollutions of Olkhon Island



System architecture



Leaflet



Conclusion

The following results have been obtained as for today:

- ❑ A short review of similar& basic projects has been carried out
- ❑ An experience has been obtained in usage of standardized ontologies and SW techniques
- ❑ The project is relevant in sense of application of standards in data publishing
- ❑ There are technologies allowing to implement such services, they're mostly web-oriented
- ❑ Application of the reviewed technologies will extend the research possibilities

The results were obtained within the framework of the State Assignment of the Ministry of Education and Science of the Russian Federation for the project “Methods and technologies of cloud-based service-oriented platform for collecting, storing and processing large volumes of multi-format interdisciplinary data and knowledge based upon the use of artificial intelligence, model-guided approach and machine learning”. Some results were obtained using the facilities of the Centre of collective usage “Integrated information network of Irkutsk scientific educational complex”.

External links

The presentation URL



[https://raw.githubusercontent.com/
eugeneai/paper-2021-iwci/main/
talk-IWCI-2021-03-30.pdf](https://raw.githubusercontent.com/eugeneai/paper-2021-iwci/main/talk-IWCI-2021-03-30.pdf)

Thank You!

Integration of Geological Data as a Knowledge Graph

Evgeny Cherkashin, Oksana Lunina, Tatiana Cherkashina,
Vadim Pellinen, Anton Gladkov

Matrosov Institute for System Dynamics and Control Theory,
SB RAS, Irkutsk, Russia
Institute of the Earth's Crust, SB RAS, Irkutsk, Russia

eugeneai@icc.ru

IWCI-2021, Baikalsk, Russia

External links

-  Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'Amato *et al.* Knowledge Graphs. <https://arxiv.org/abs/2003.02320v5>
-  Claus Stadler, Jens Lehmann, Konrad Höffner, Sören Auer. LinkedGeoData: A core for a web of spatial open data. *Semantic Web* 3 (2012) 333–354. DOI:[10.3233/SW-2011-0052](https://doi.org/10.3233/SW-2011-0052)
-  Adam Iwaniak, Iwona Kaczmarek, Marek Strzelecki, Jaromar Lukowicz, Piotr Jankowski. Enriching and improving the quality of linked data with GIS. DOI:[10.1515/geo-2016-0c020](https://doi.org/10.1515/geo-2016-0c020)
-  Adam Iwaniak, Marta Leszczuk, Marek Strzelecki, Francis Harvey, Iwona Kaczmarek. A Novel Approach for Publishing Linked Open Geodata from National Registries with the Use of Semantically Annotated Context Dependent Web Pages. *International Journal of Geo-Information*. 6, 252, 2017. DOI:[10.3390/ijgi6080252](https://doi.org/10.3390/ijgi6080252)
-  Tarek Abid, Hafed Zarzour. Integrating Linked Open Data in Geographical Information System. *International Conference on Information Technology for Organization Development*. 2014.
-  Michelle Cheatham, Adila Krishnadi, Reihaneh Amini, Pascal Hitzler, *et al* (2018) The GeoLink knowledge graph, *Big Earth Data*, 2:2, 131–143, DOI:[10.1080/20964471.2018.1469291](https://doi.org/10.1080/20964471.2018.1469291)
-  Oksana V. Lunina. The digital map of the pliocene quaternary crustal faults in the southern east siberia and the adjacent northern Mongolia. *Geodynamics & Tectonophysics*. 2016. 7(3):407-434. DOI:[10.5800/GT-2016-7-3-0215](https://doi.org/10.5800/GT-2016-7-3-0215)

Auxiliary materials

Technologies used (open source)

Python-3.x.x (<http://python.org>)

ZCA (<https://muthukadan.net/docs/zca.html>)

SWIG (<http://swig.org/>)

SWI-Prolog (<https://www.swi-prolog.org/>)

Logtalk (<https://logtalk.org/>)

ClioPatria (<https://cliopatria.swi-prolog.org/home>)

Virtuoso Open Source Edition (<http://vos.openlinksw.com/owiki/wiki/VOS>)

Pengines (<https://pengines.swi-prolog.org/docs/index.html>)

LOV (<https://lov.linkeddata.es/dataset/lov/>)

Elastic Search (<https://www.elastic.co/>)

Kyotocabinet (<https://fallabs.com/kyotocabinet/>)

DBPedia (<https://wiki.dbpedia.org/>)

Dust.js (<https://akdubya.github.io/dustjs/>)

QGIS (<https://qgis.org/ru/site/>)

TabbyDOC (<http://td.icc.ru/>)

GeoBase (<https://github.com/eugeneai/geobase>)

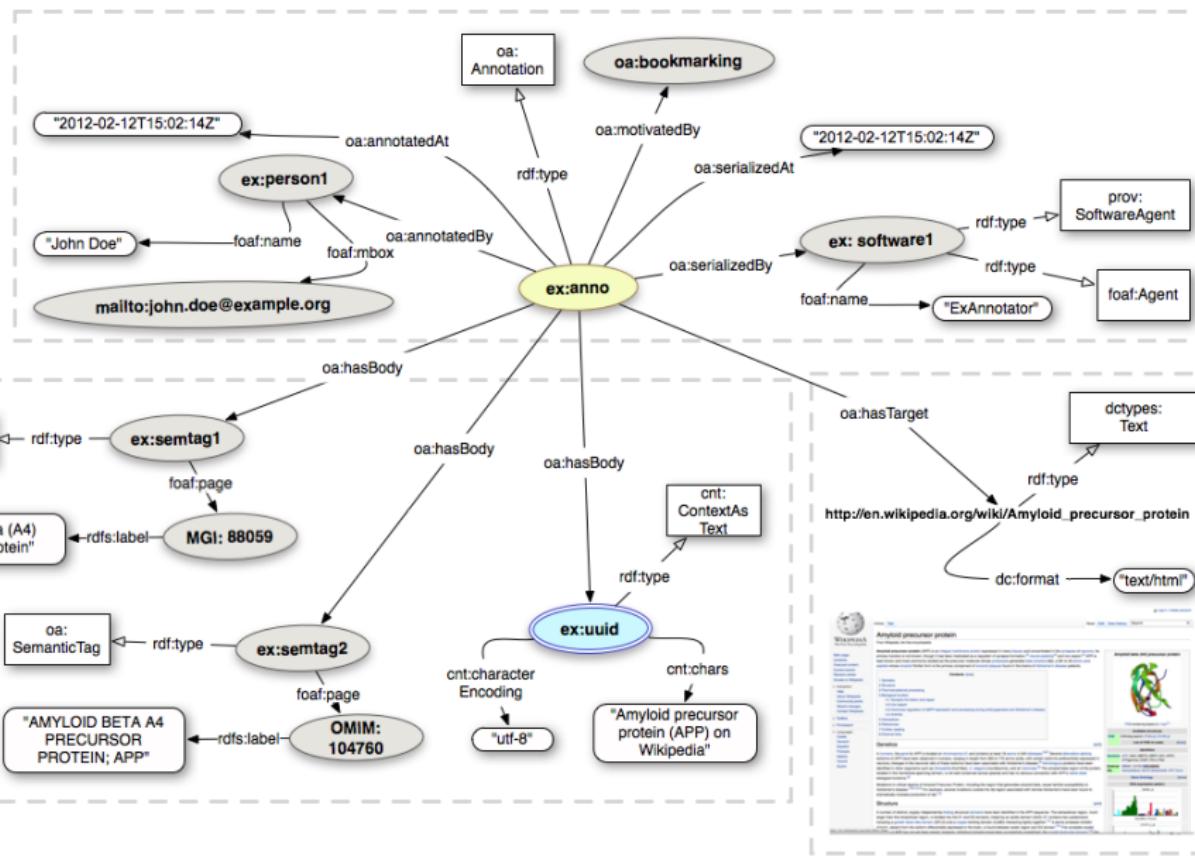
Authoring Tool (<https://github.com/isu-enterprise/isu.college>)

Used ontologies

Standardized ontologies

- ❑ Friend-of-a-friend (**foaf**) for agent information: individuals, legal entities, program agents
- ❑ Provenance (**prov**) for making references between documents
- ❑ Dublin Core (**dc**) for published resource metadata mark up
- ❑ DBpedia resource (**dbr**) to refer external classes and instance objects
- ❑ Schema.org (**schema**) for Google, Yandex, Yahoo, *etc.* searchable objects, structural elements
- ❑ The Bibliographic Ontology (**bibo**) used for literature reference
- ❑ Open annotation (**oa**) as an “bookmark” ontology
- ❑ LinkedGeoData A-Box (**lgd**)
- ❑ LinkedGeoData T-Box (**lgdo**)
- ❑ Coordinate system (**wgs84**)
- ❑ the Ontology WEB Language (**owl**)
- ❑ XML Schema (**xsd**)

Open Annotation (oa)



Document authoring and storage

In most cases documents are created as a result of

- ❑ creative activity of a person with text processors (authoring);
- ❑ printing a digital copy or a data record in a database;
- ❑ aggregation operation over database records (report).

Then it is stored either as a physical paper and/or a digital document (PDF, DOCX, HTML).

Since 2000-th, Semantic Web and Linked Open Data (LOD) is being developed, allowing

- ❑ structural storage of data within published documents;
- ❑ processing stored data computationally;
- ❑ integration of data structures and data objects globally.

The **aim of this research** is to develop technologies, software and services allowing construction of digital archives supporting document data inclusion and inference from existing documents.

Logtalk as transformation definition language

We have chosen Logtalk as it

- ❑ inherits widely known Prolog language syntax and runtime
- ❑ implemented as macro package, performance penalties are about 1.5%
- ❑ has flexible semantics: we can define transformations and constraints within the same syntax
- ❑ implement object-oriented knowledge (rules) structuring, encapsulation and replacement
- ❑ compositional way of transformation implementation
- ❑ powerful engine to post constraints on object-to-object messages (events)
- ❑ has implementation for many Prolog engines.

The «regular» language allow us to use its libraries not directly related to MDA transformations.