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 cartographical 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.*, withing 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 re-represent the KG).

Knowledge graph: Validating semantic example

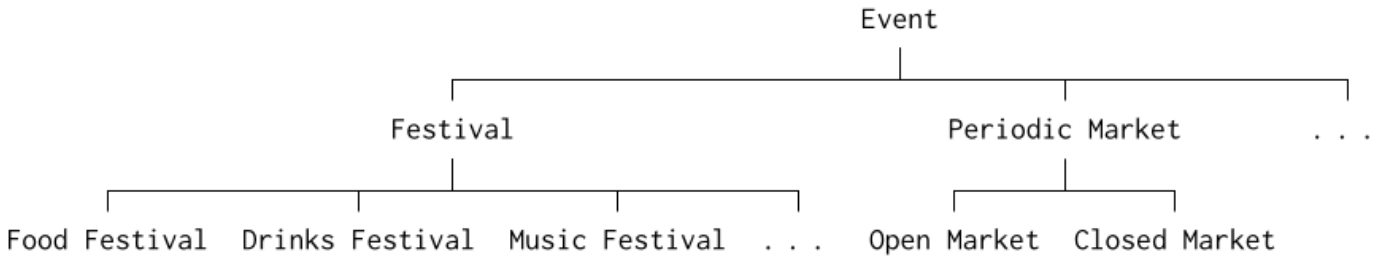


Fig. 10. Example class hierarchy for Event

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

Feature	Definition	Condition	Example
SUBCLASS	(c) –subc. of→ (d)	(x) –type→ (c) implies (x) –type→ (d)	(City) –subc. of→ (Place)
SUBPROPERTY	(p) –subp. of→ (q)	(x) –p→ (y) implies (x) –q→ (y)	(venue) –subp. of→ (location)
DOMAIN	(p) –domain→ (c)	(x) –p→ (y) implies (x) –type→ (c)	(venue) –domain→ (Event)
RANGE	(p) –range→ (c)	(x) –p→ (y) implies (y) –type→ (c)	(venue) –range→ (Venue)

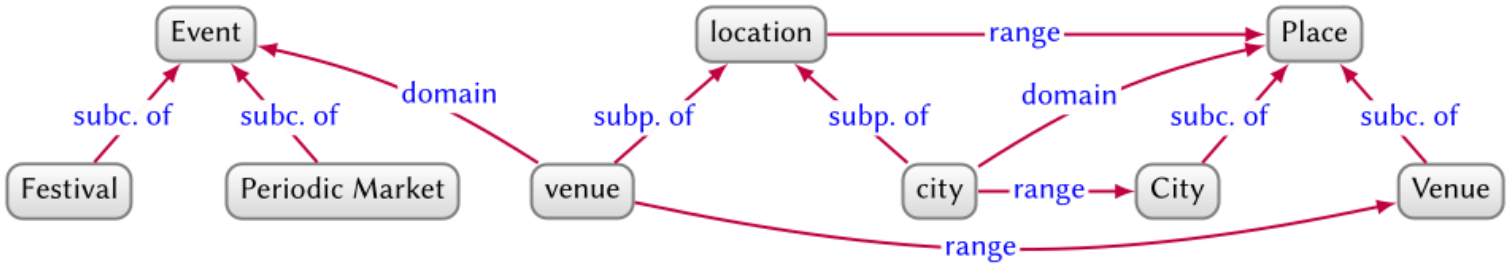


Fig. 11. Example schema graph describing sub-classes, sub-properties, domains, and ranges

Linked Open Data (LOD) star evaluation

Data are available in

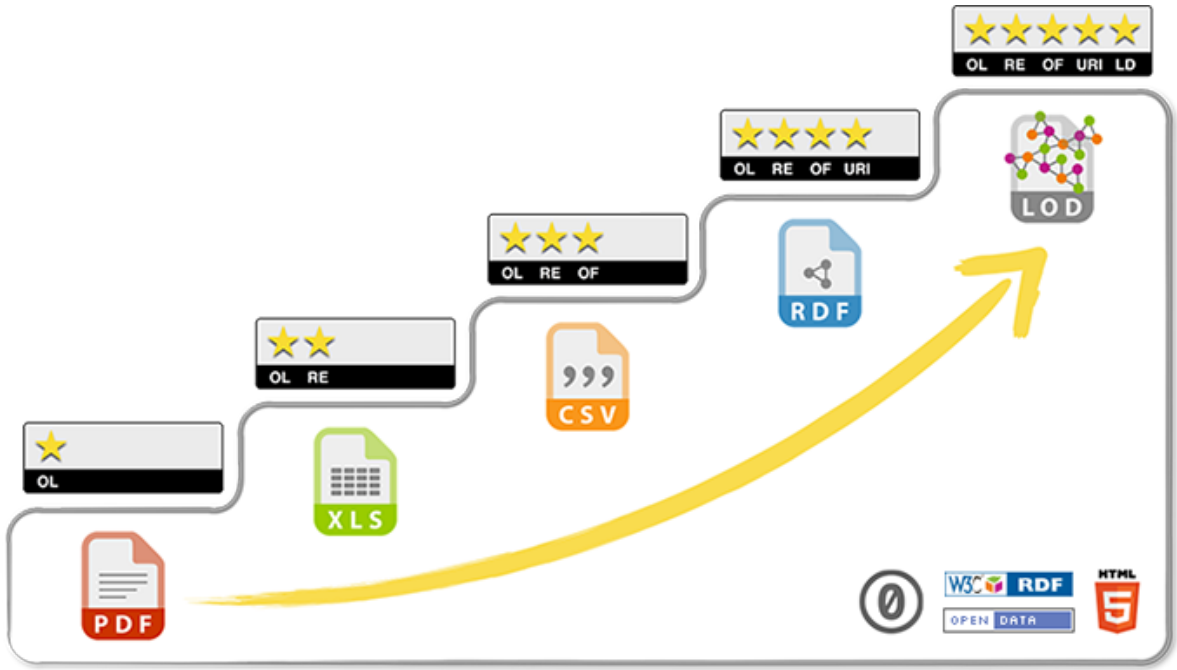
1* any format **openly**

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: **LinkedGeoData** project

The project [?] was to represent OpenSteetMap (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 $\longrightarrow$ 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.

Related works: LinkedGeoData project

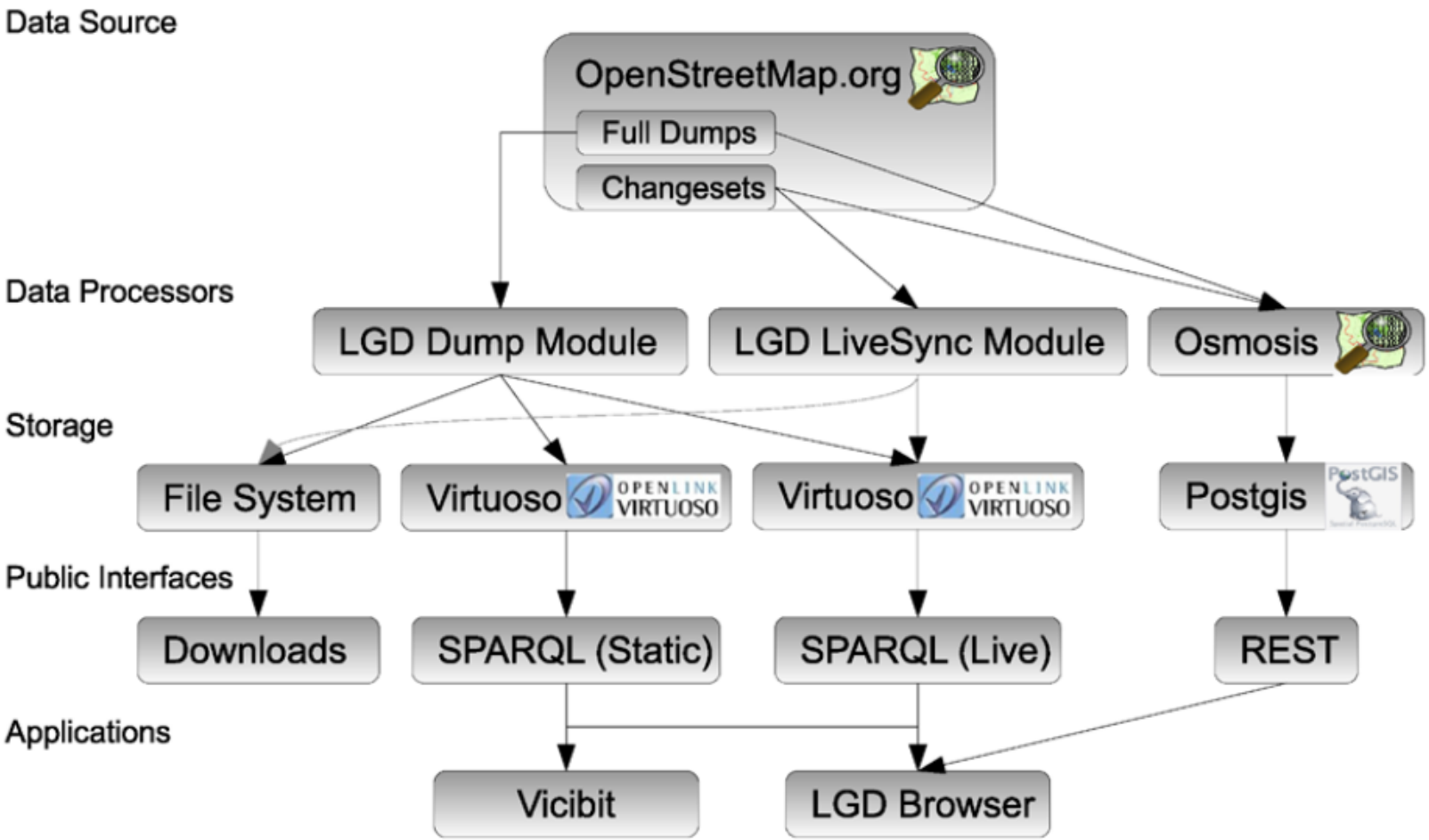


Fig. 2. Overview of LinkedGeoData's architecture.

Related works: LinkedGeoData project

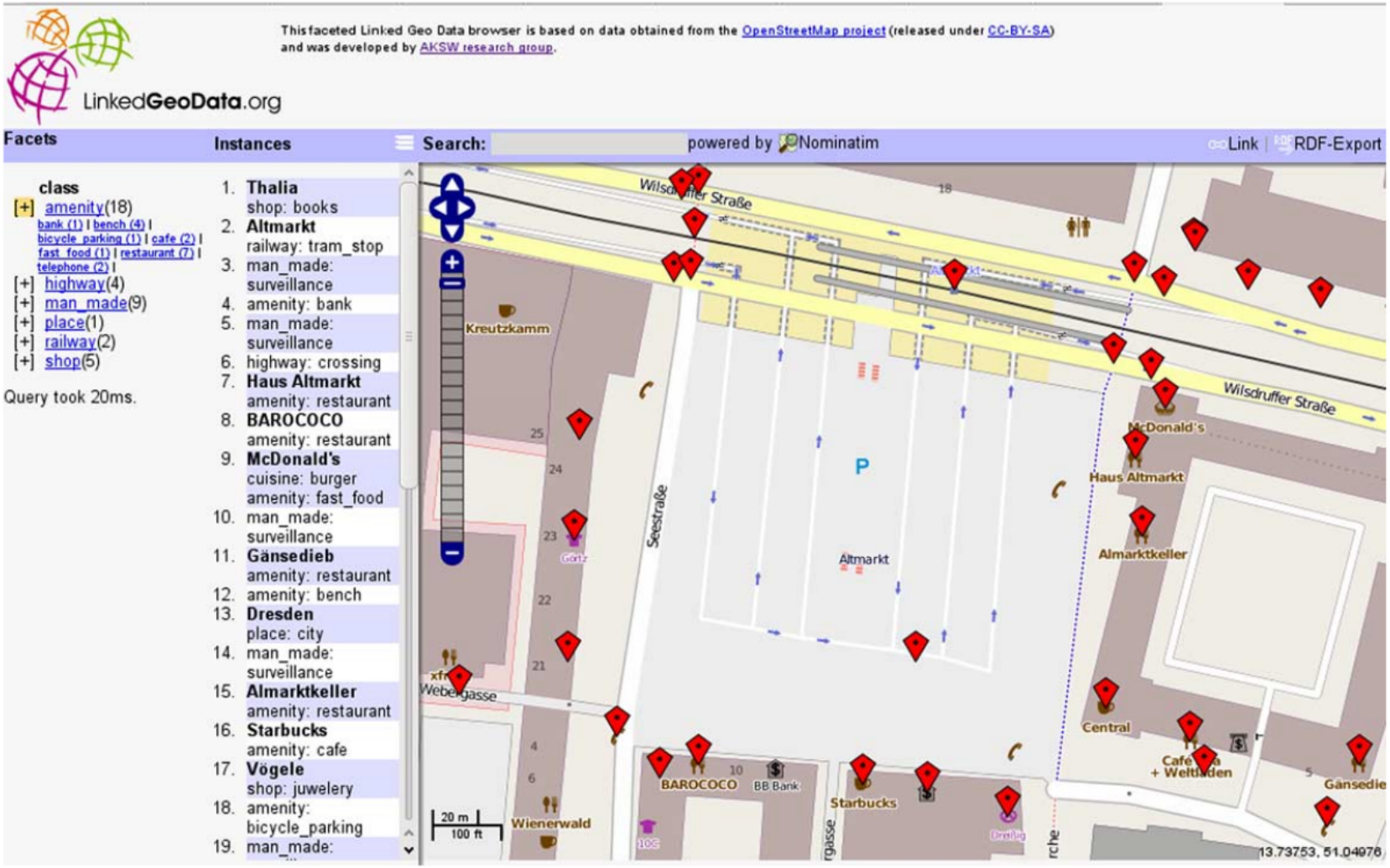


Fig. 6. LinkedGeoData Browser.

Related Works: GeoLink Knowledge graph

GeolLink KG [?]

- 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 geochemy, **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).

Related Works: GeoLink Knowledge graph

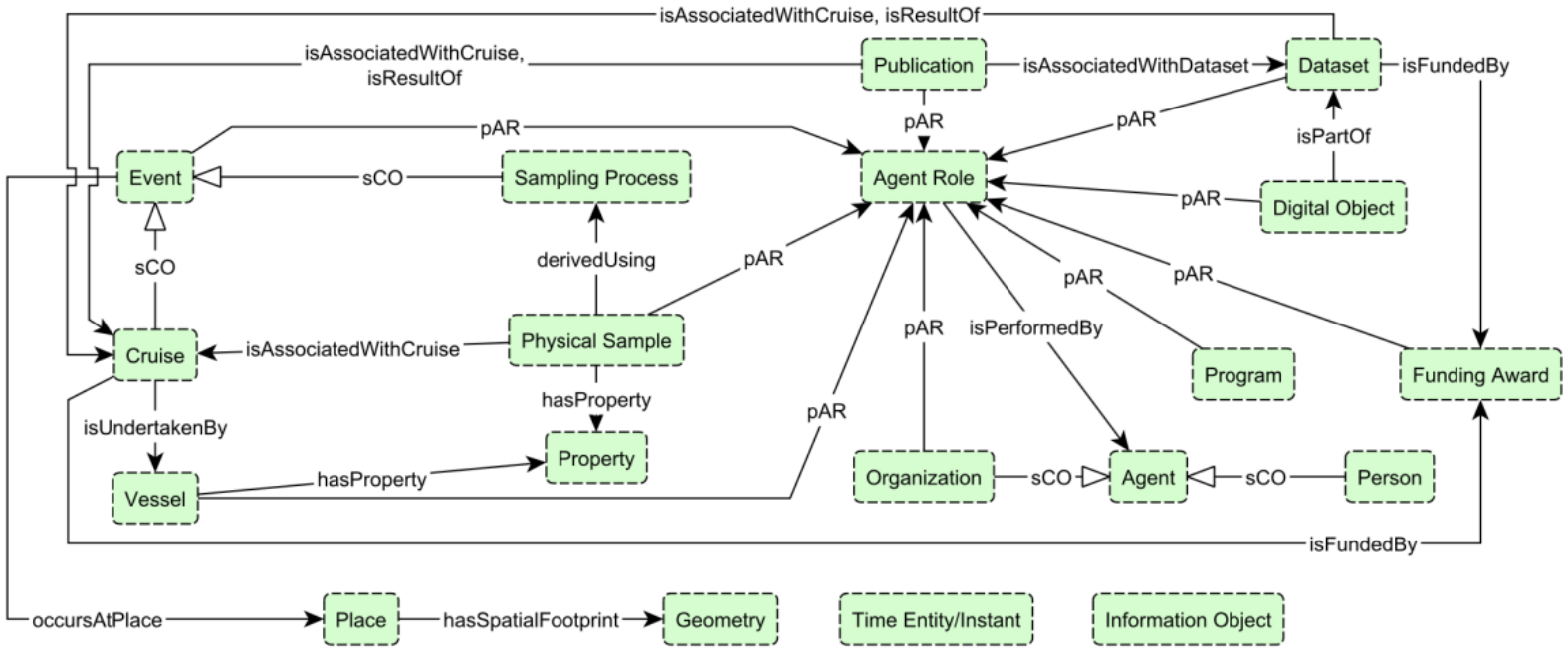


Figure 1. Schema diagram containing (almost) all patterns in the GeoLink ontology and their main links. All patterns have links to Time Entity/Instant and Information Object, but they have been omitted for clarity. sCO=subClassOf; pAR=providesAgentRole; each box is a pattern, represented by its main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

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

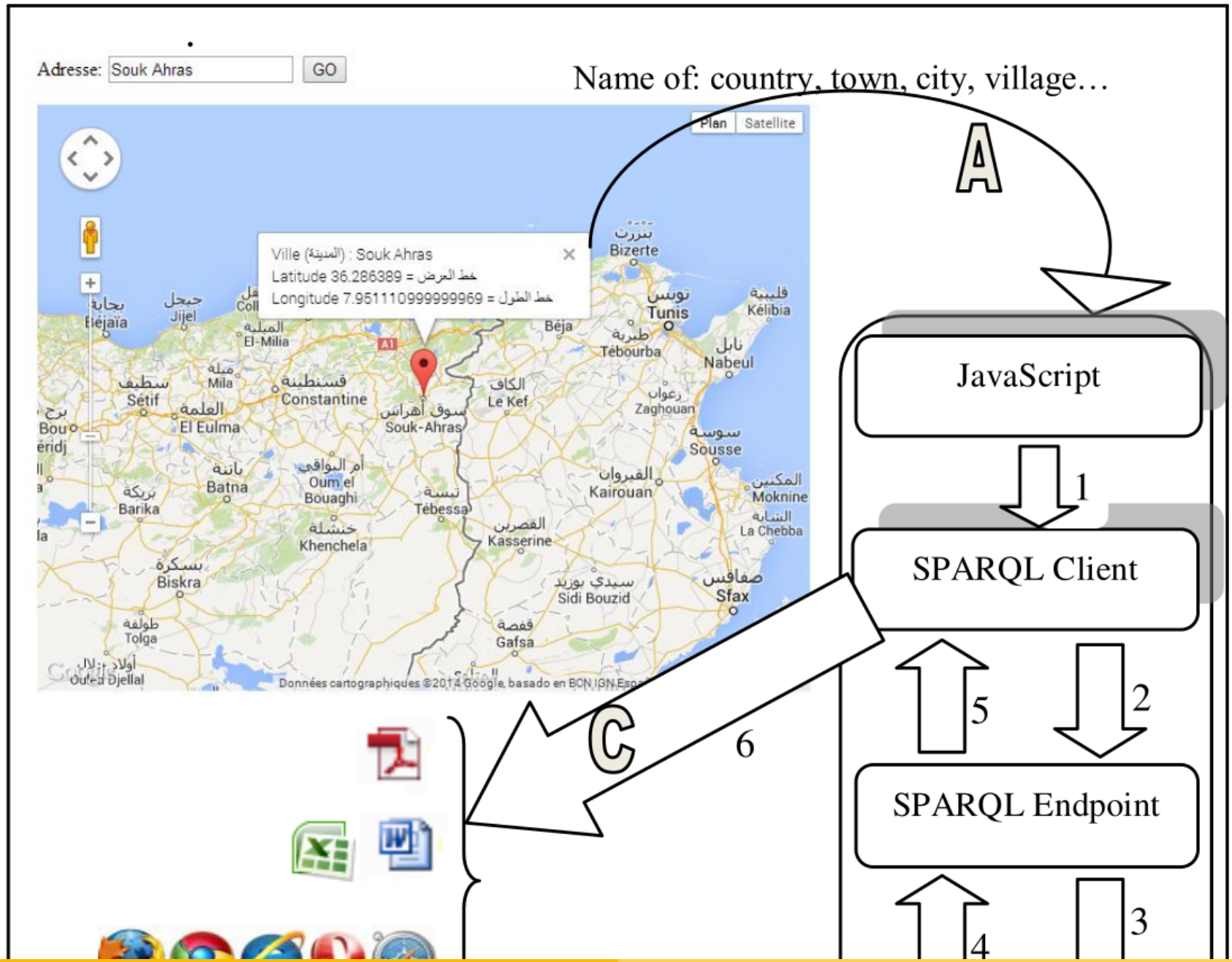
Related Works: Integrating LOD into GIS

The project[1] 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.

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

Related Works: Integrating LOD into GIS



Related Works: 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
```

Related Works: Publishing Geodata LOD in context dependent pages

The project[2] 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

[2] 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

Publishing Geodata LOD in context dependent pages

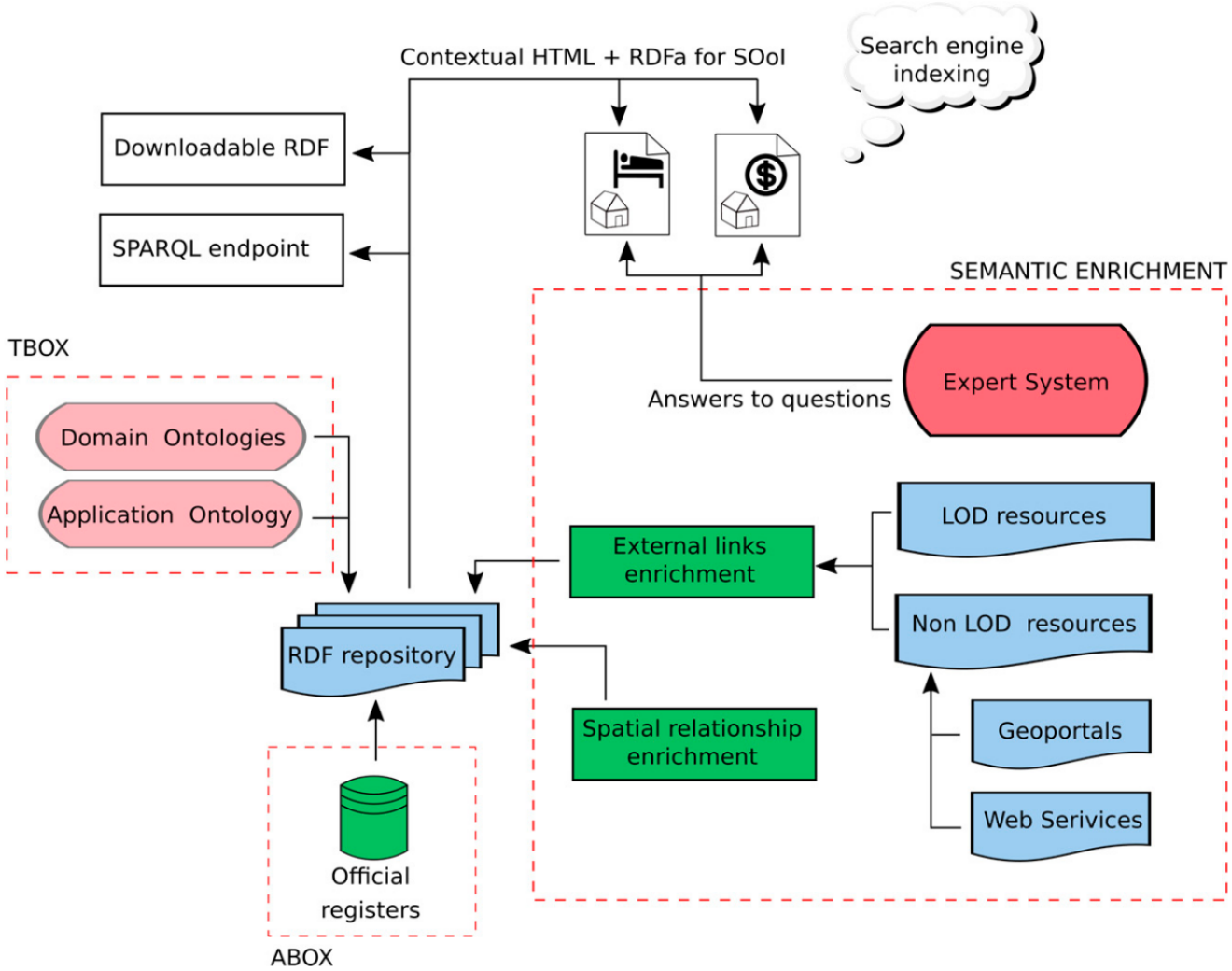


Figure 1. Concept of system architecture.

Enriching and improving the quality of linked data with GIS

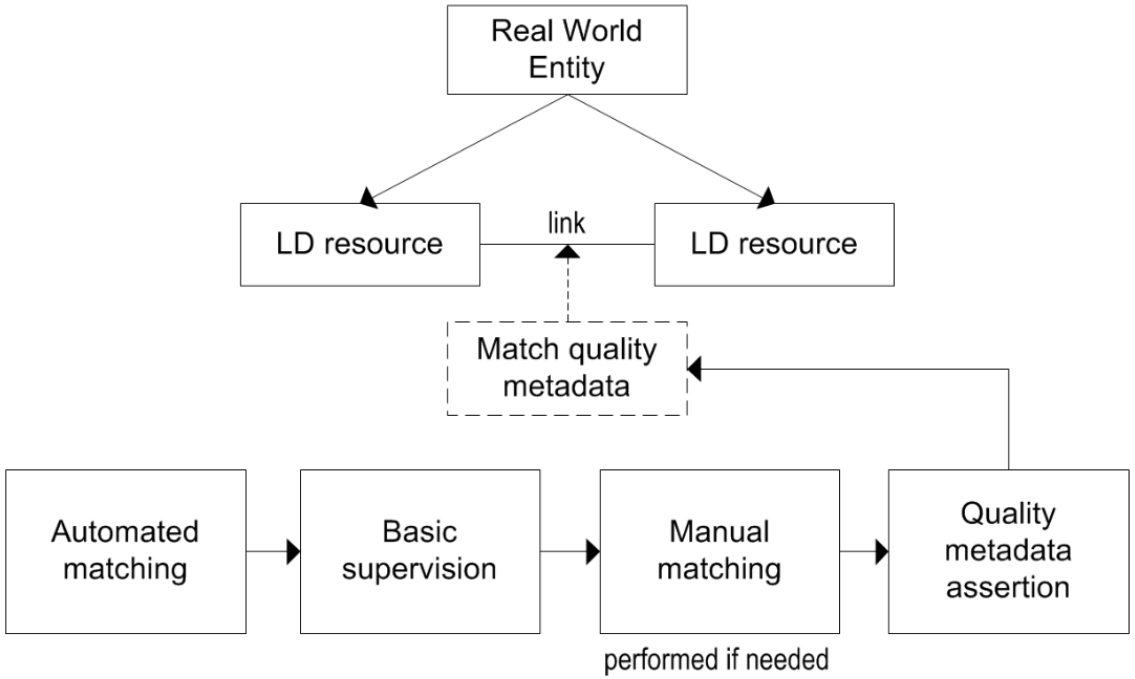


Figure 9: Basic supervision of linked data resources location quality with the use of Desktop GIS map visualization.

From an early work[3].

[3] Adam Iwaniak, Iwona Kaczmarek, Marek Strzelecki, Jaromar Lukowicz, Piotr Jankowski. Enriching and improving the quality of linked data with GIS. doi:10.1515/geo-2016-0020

Publishing Geodata LOD in context dependent pages

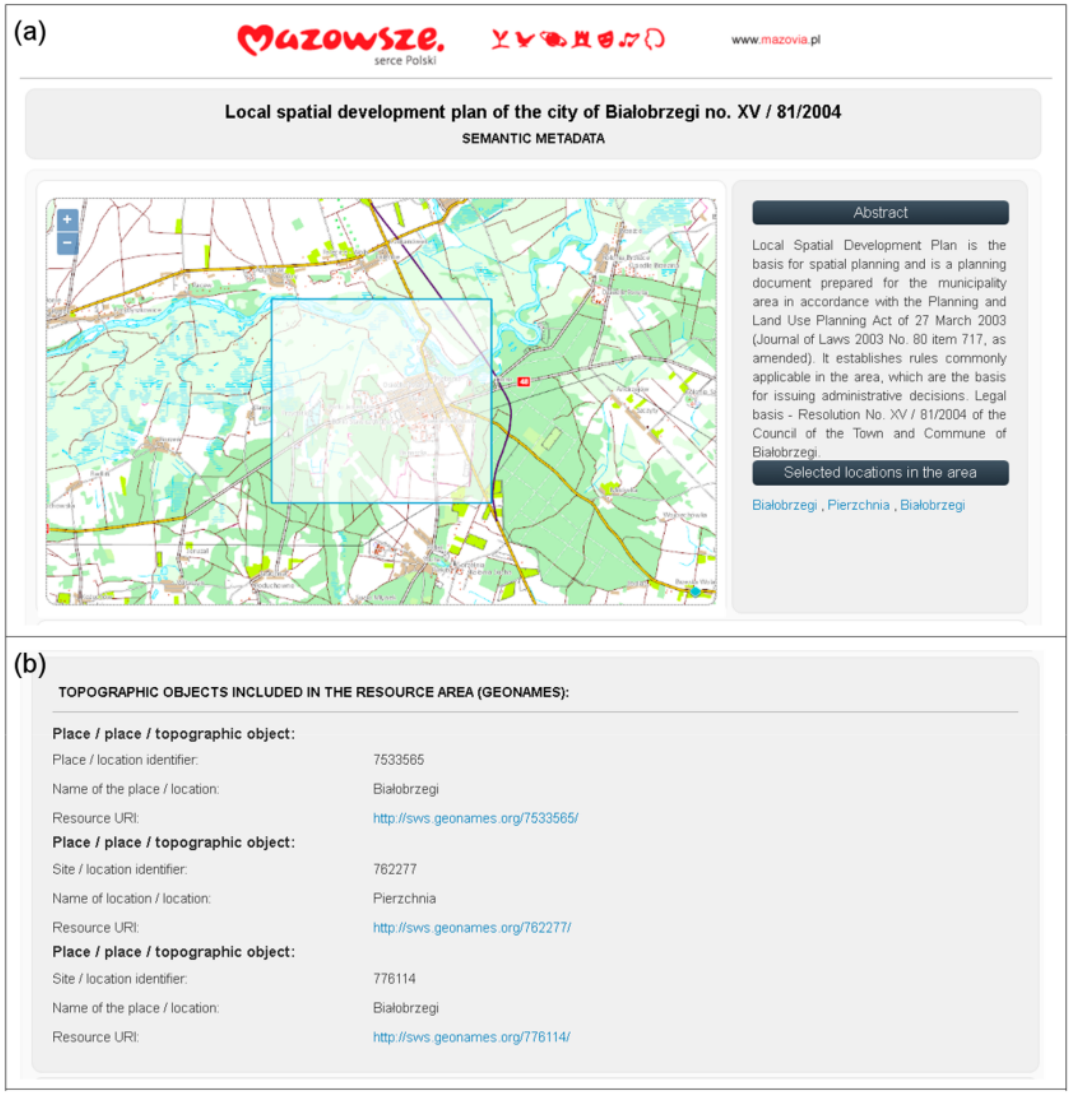
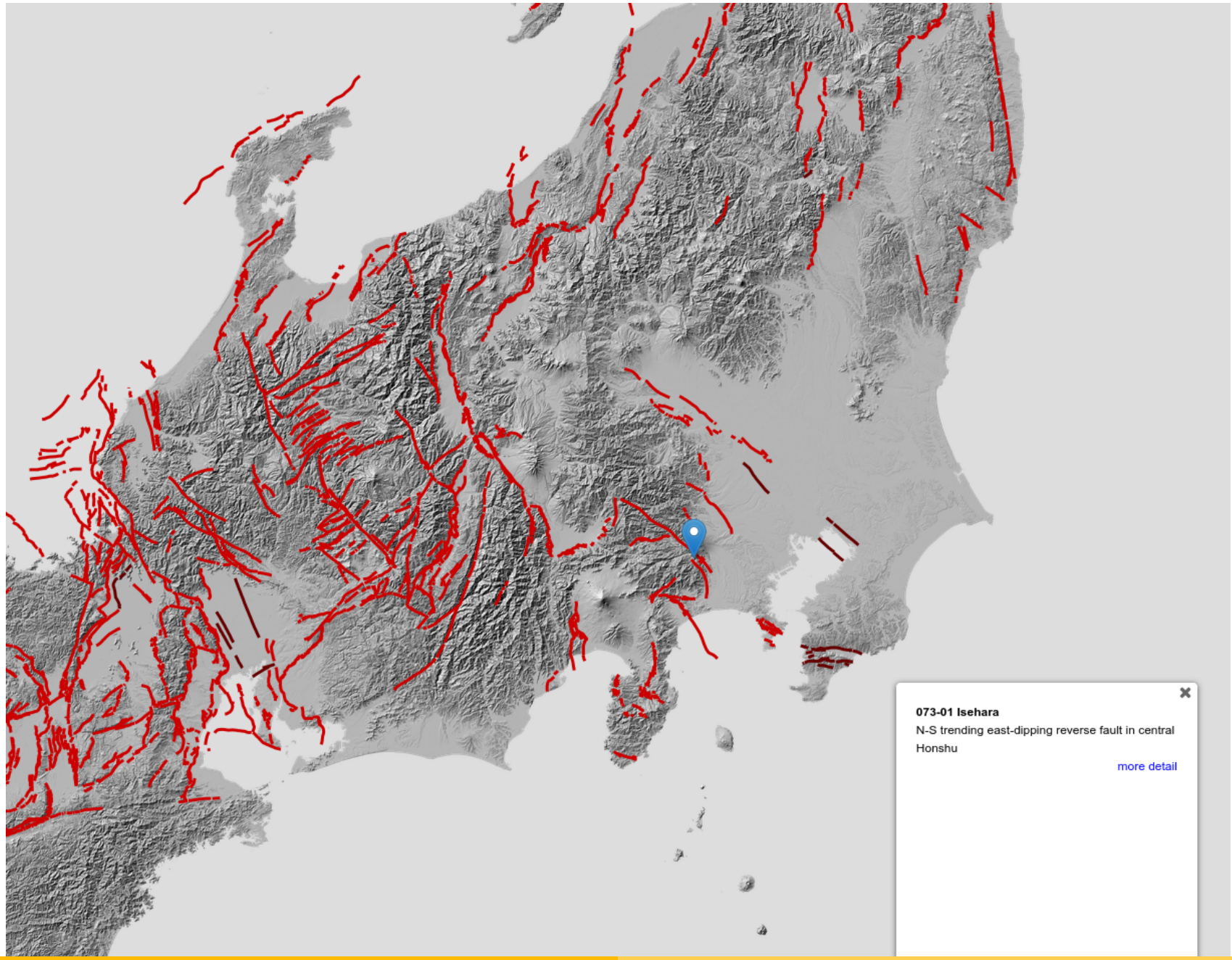


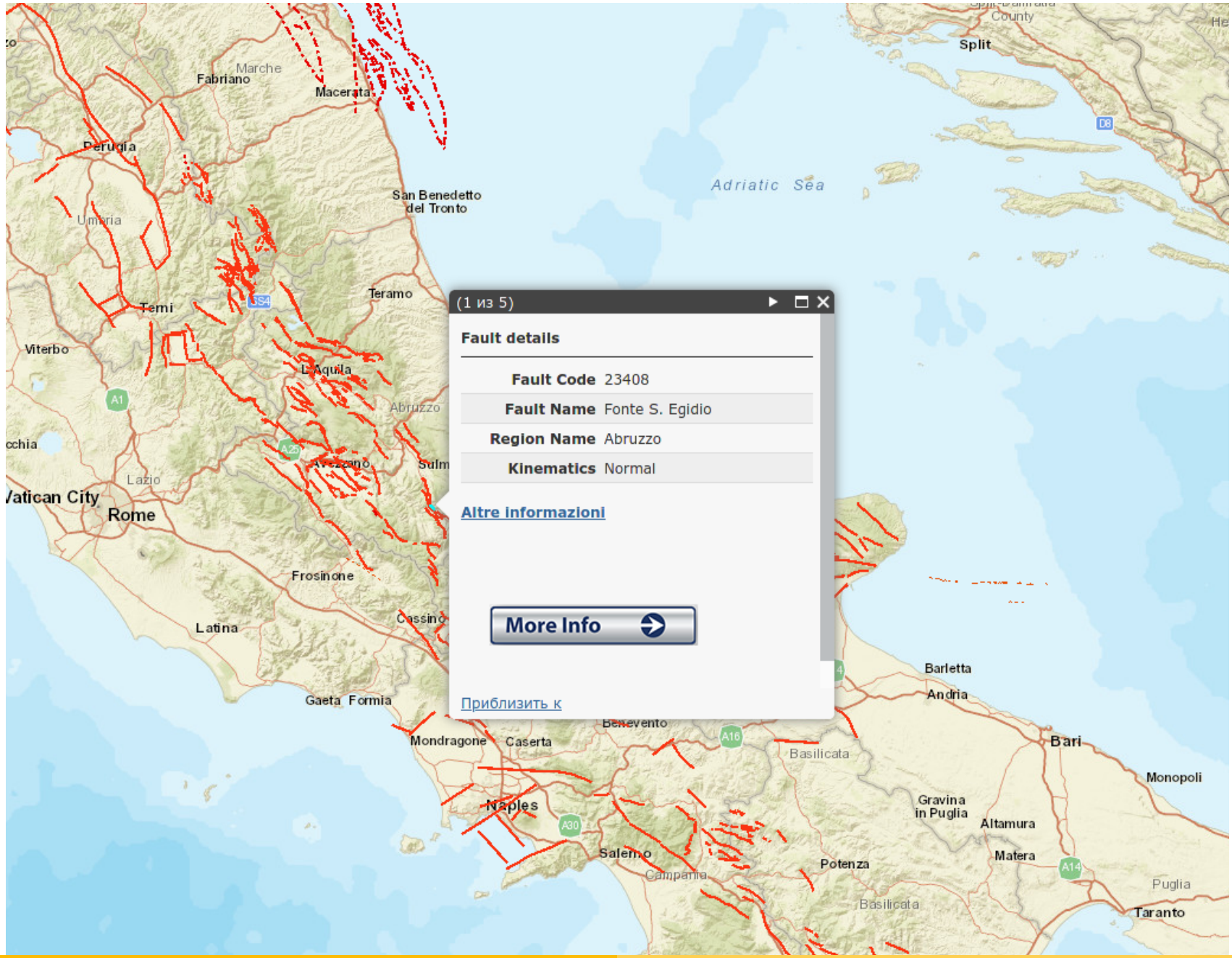
Figure 6. Fragments of HTML document, containing metadata for local spatial development plan: (**a**) title, basic information and bounding box of data; (**b**) links to Geonames features, included in the resource area.

Related Works: Publishing Geodata LOD in context dependent pages

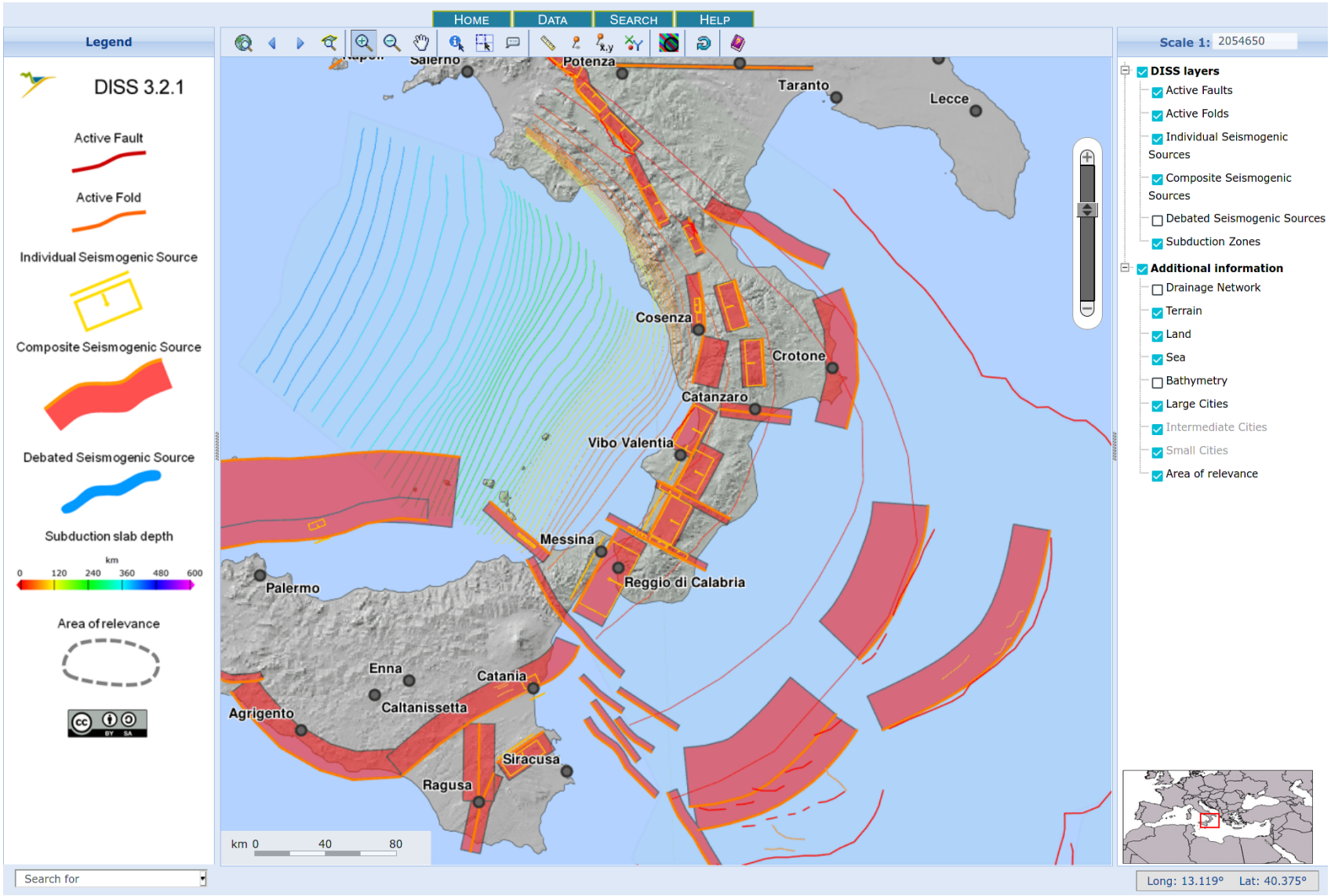
Existing resources



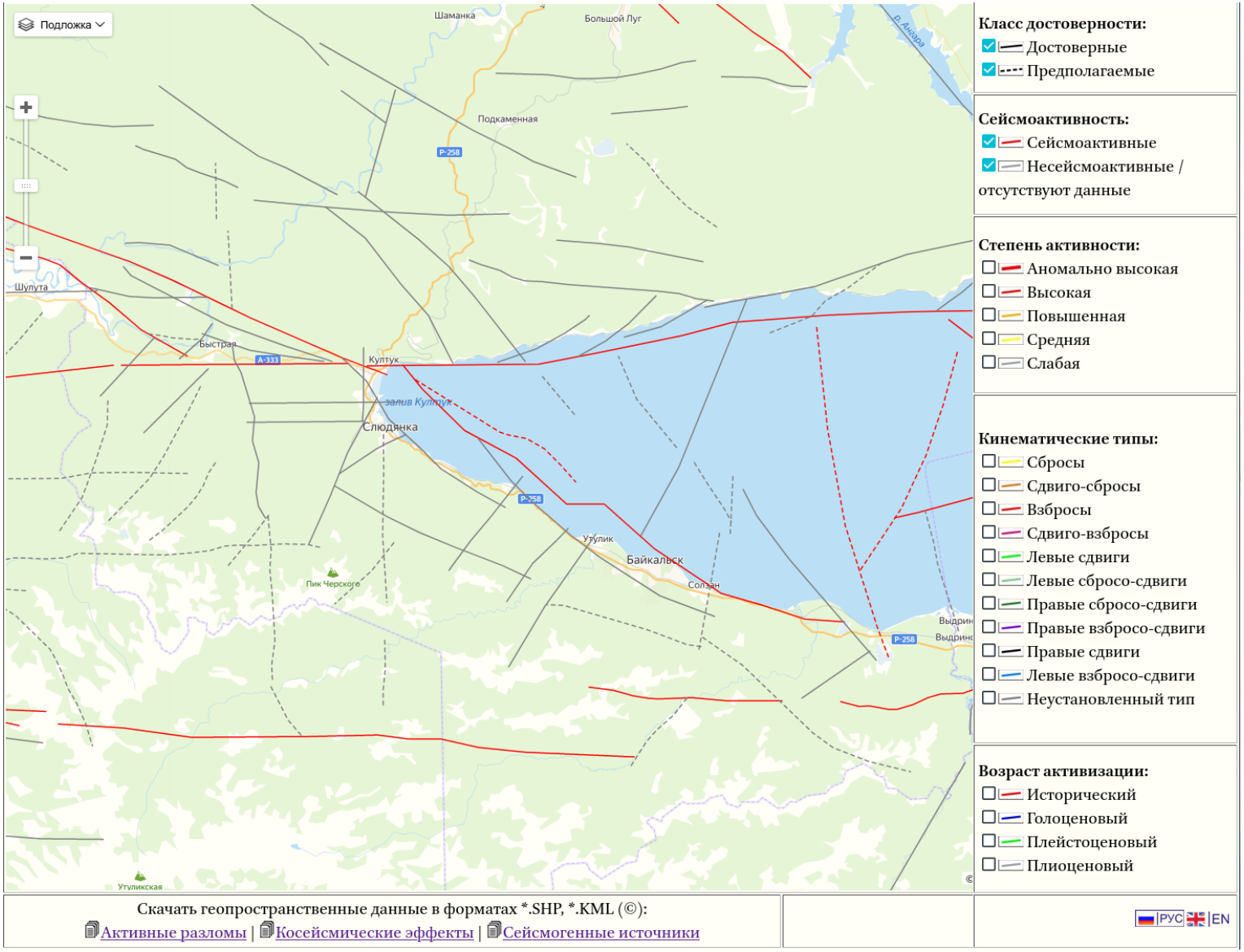
Existing resources



Existing resources



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



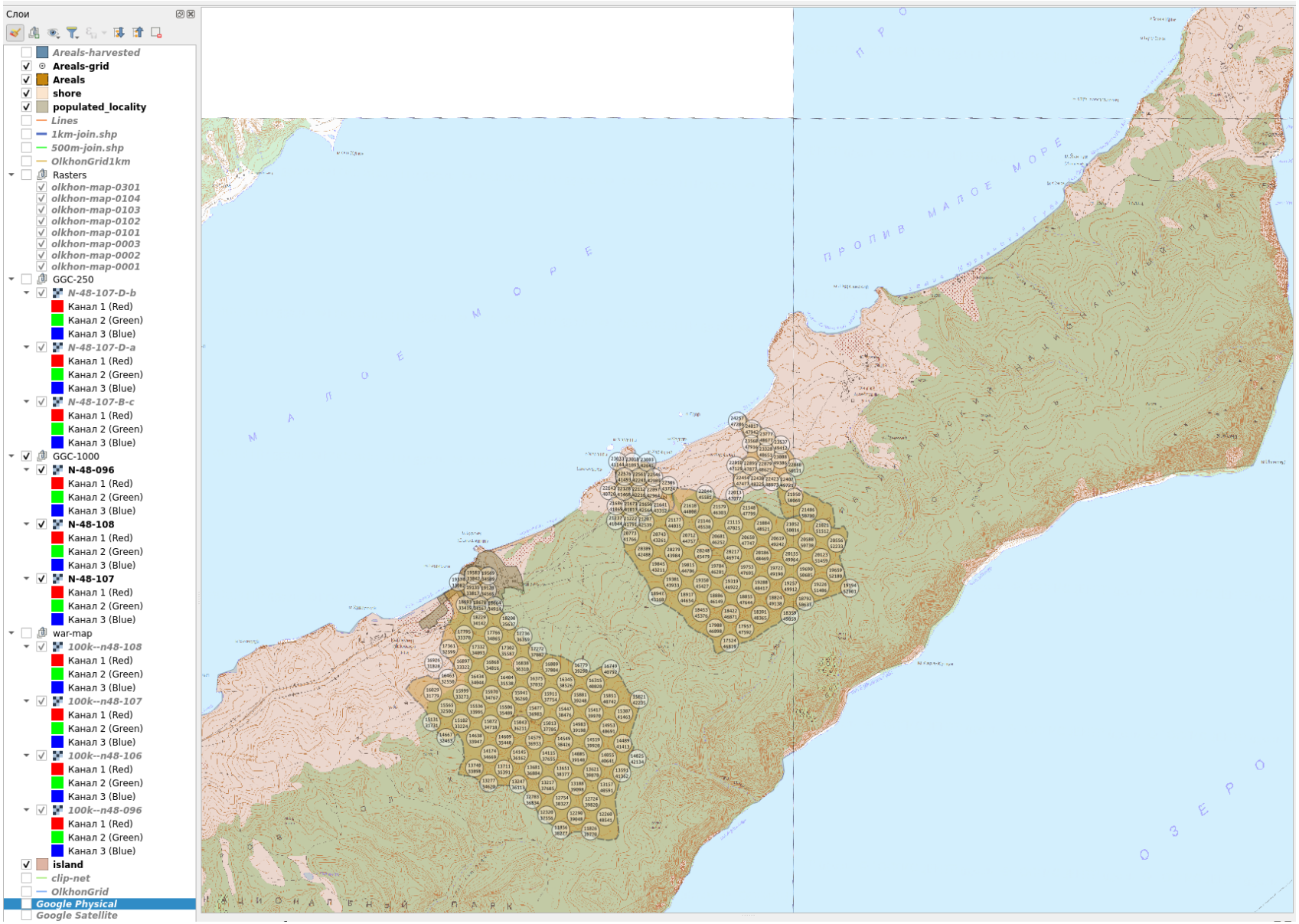
Resources to be presented: Active faults of the South of East Siberia

The fault data[4] have the following properties:

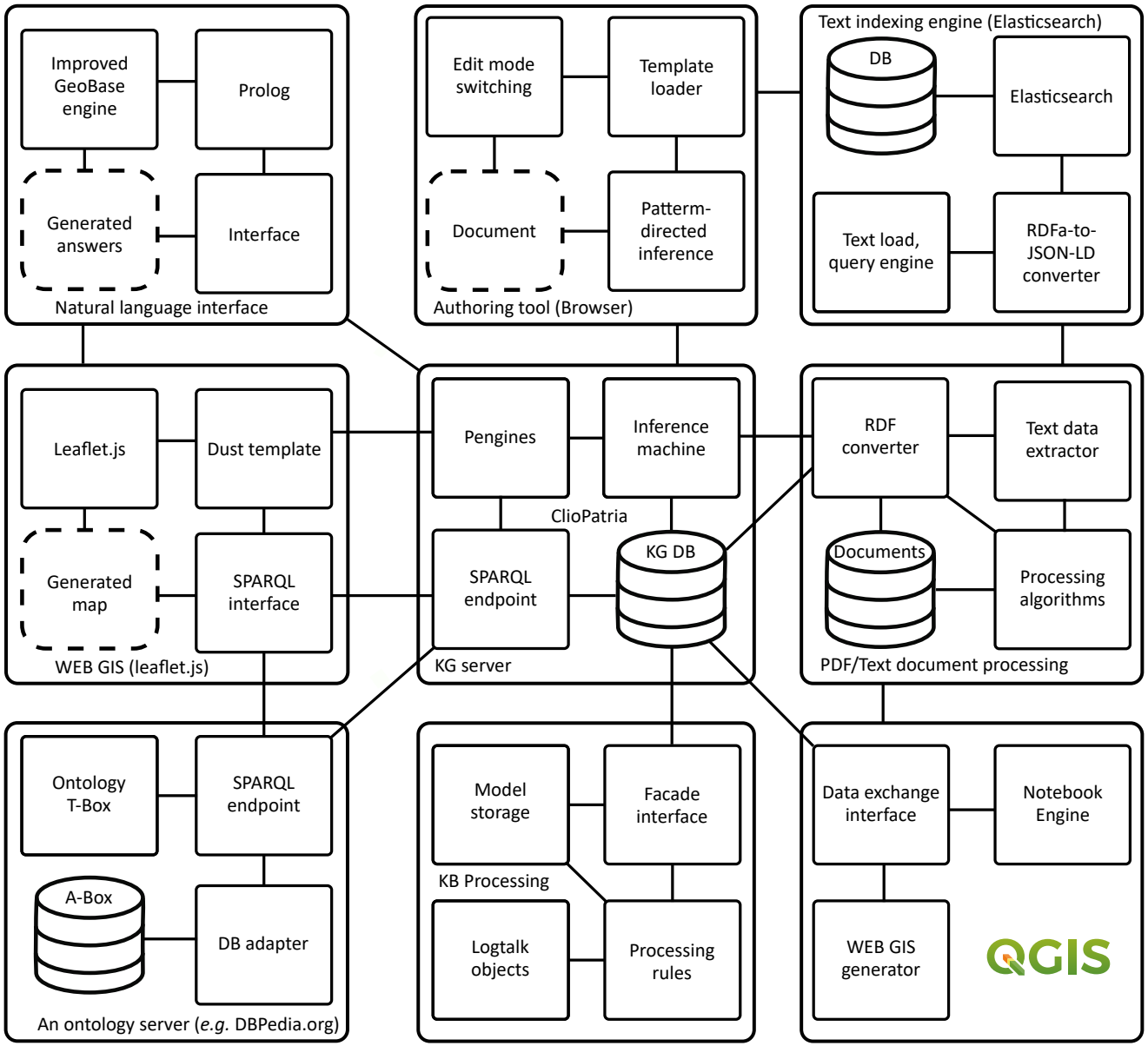
- Faults are the objects having linear projection on the surface, and a depth and a slope in the lithosphere
- Various geological events are also subjects of presentation
- Any object accompanied with various valued characteristics having also characteristics, *e.g.*, unit name, measurement precision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

[4] 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

Resources to be presented: the Olkhon Island



System architecture



Conclusion

[[[]]]The following results have been obtained as for today:

- Biologists' activities are mastered, investigated and regular patterns are described.
- A technique for interpretation of Mothur interfaces has been developed, implemented, refined and extended.
- Transformation tools are tested in application areas and no significant technical problems were detected.
- A technique of document authoring is being developed and adapted to the domain.
- Integration with Galaxy project is the primary aim of the future development.

The source codes are available at
https://github.com/isu-enterprise/icc.xmitransform,
https://github.com/eugeneai/icc.mothurpim.
This research is supported by Irkutsk scientific center of SB RAS, project No 4.2;

External links

The presentation URL

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

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

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'Amato *et al.* Knowledge Graphs. https://arxiv.org/abs/2003.02320v5

Thanks for Your interest in our project!

Auxiliary materials

Technologies used (open source)

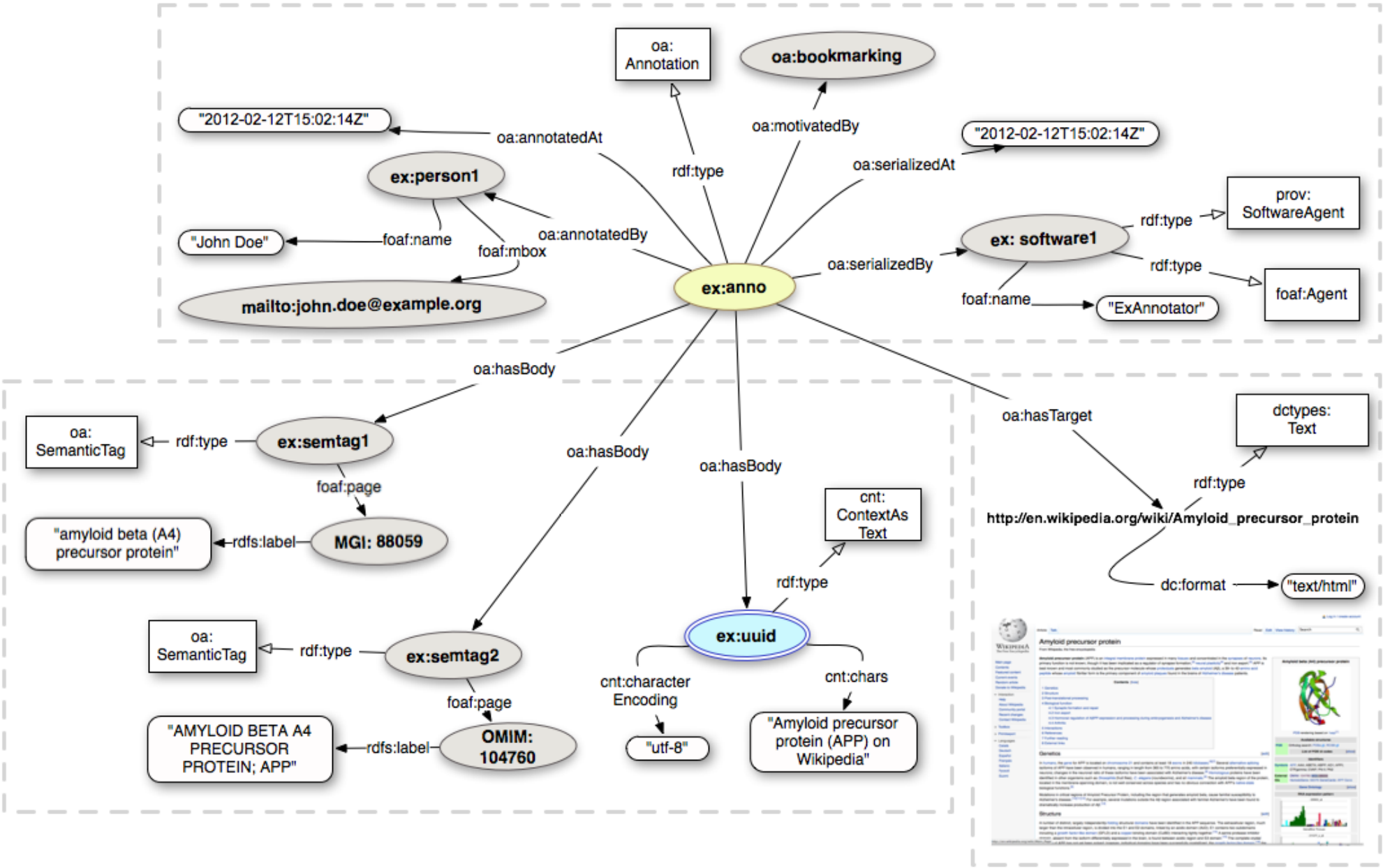
Python-3.x.x (http://python.org)
ZCA (https://muthukadan.net/docs/zca.html)
SWIG (http://swig.org/)
SWI-Prolog (https://www.swi-prolog.org/)
Logtalk (https://logtalk.org/)
ClioPatria (https://cliopatria.swi-prolog.org/home)
Virtuoso Open Source Edition (http://vos.openlinksw.com/owiki/wiki/VOS)
Pengines (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 a 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.