HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

Technische Universität Berlin

Artificial Intelligence Techniques for Earth Observation Data

Dimitris Bilidas, Begüm Demir, Theofilos Ioannidis, Manolis Koubarakis, Despina-Athanasia Pantazi, George Papadakis, Dharmen Punjani, George Stamoulis and Eleni Tsalapati

ISWC 2022 Tutorial, Virtual
October 23, 2022

Outline

- Introduction and motivation
- The data science pipeline for big linked Earth observation data
  - Discovering Earth Observation data
  - Deep Earth Query: Information Discovery from Big Earth Observation Data Archives
  - RDF and SPARQL extensions for geospatial data
  - Geospatial RDF stores
  - Transformation of geospatial data
  - Interlinking geospatial RDF data
  - Geospatial knowledge graphs
  - Question answering over geospatial knowledge graphs
  - Searching, browsing, exploring and visualizing linked geospatial data
- Open questions for future work

Introduction and motivation

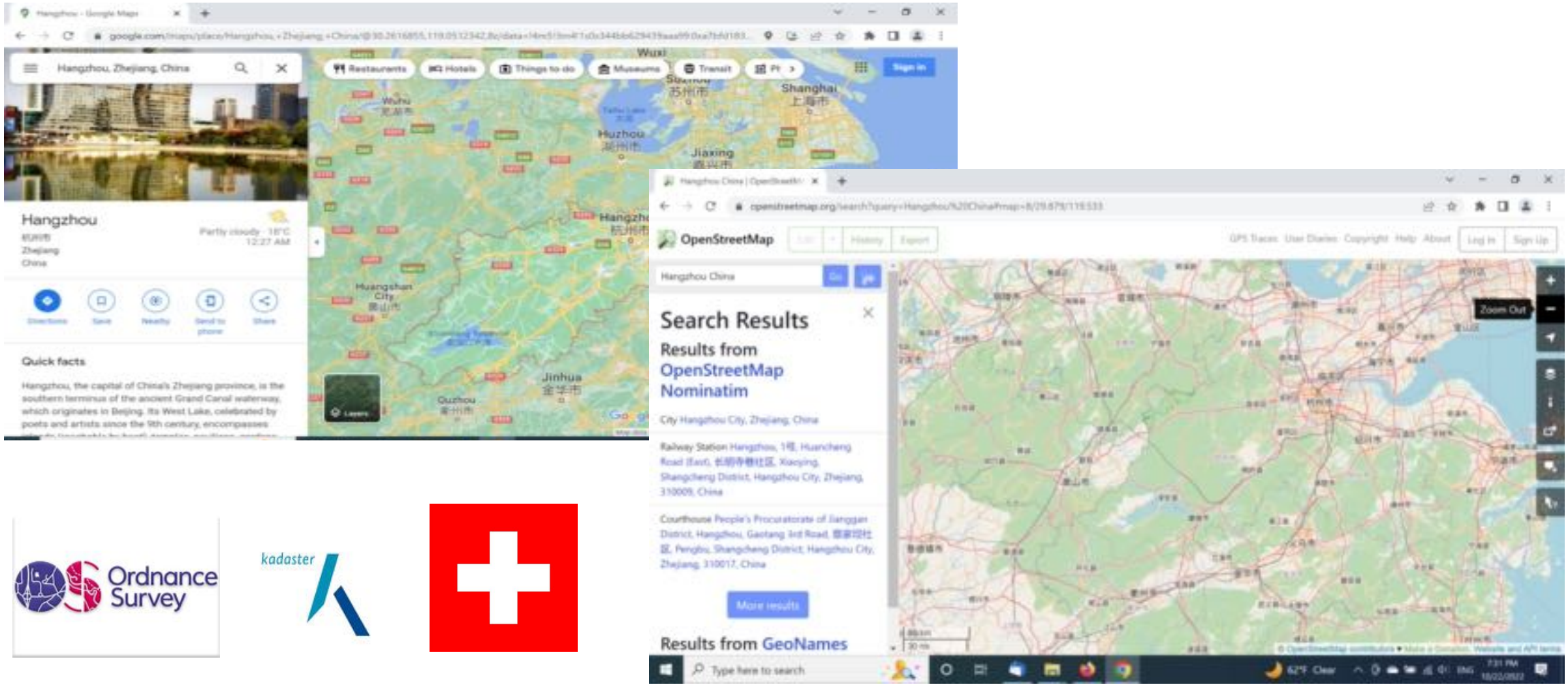
Open Government Data

- Lots of **public sector data** has been made open and freely available recently through various government portals.

DATA.GOV
EMPOWERING PEOPLE

European Union Open Data Portal

(Open) Geospatial Data



Ordnance Survey

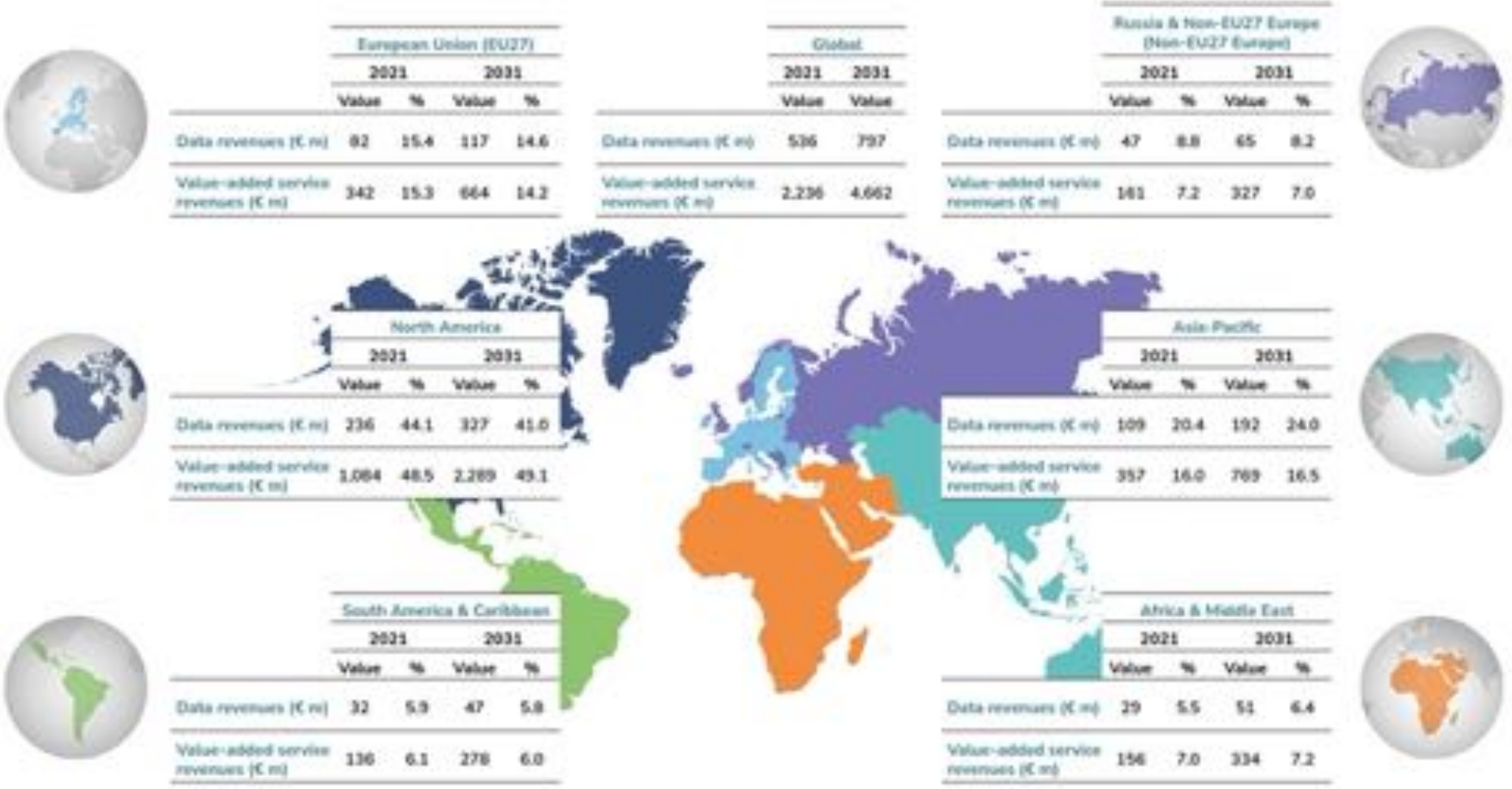
kadaster

Open Earth Observation Data

- Lots of **Earth Observation (EO) data** has also been made freely available recently in Europe and the United States.

- Europe is a pioneer in this area with its flagship Earth Observation Programme **Copernicus**.

Economic Impact



European Union (EU27)

	2021 Value	2021 %	2031 Value	2031 %
Data revenues (€ m)	82	15.4	117	14.6
Value-added service revenues (€ m)	342	15.3	664	14.2

Global

	2021 Value	2031 Value
Data revenues (€ m)	536	797
Value-added service revenues (€ m)	2.236	4.662

Russia & Non-EU27 Europe (Non-EU27 Europe)

	2021 Value	2021 %	2031 Value	2031 %
Data revenues (€ m)	47	8.8	65	8.2
Value-added service revenues (€ m)	161	7.2	327	7.0

North America

	2021 Value	2021 %	2031 Value	2031 %
Data revenues (€ m)	236	44.1	327	41.0
Value-added service revenues (€ m)	1.084	48.5	2.289	49.1

Asia-Pacific

	2021 Value	2021 %	2031 Value	2031 %
Data revenues (€ m)	109	20.4	192	24.0
Value-added service revenues (€ m)	357	16.0	769	16.5

South America & Caribbean

	2021 Value	2021 %	2031 Value	2031 %
Data revenues (€ m)	32	5.9	47	5.8
Value-added service revenues (€ m)	136	6.1	278	6.0

Africa & Middle East

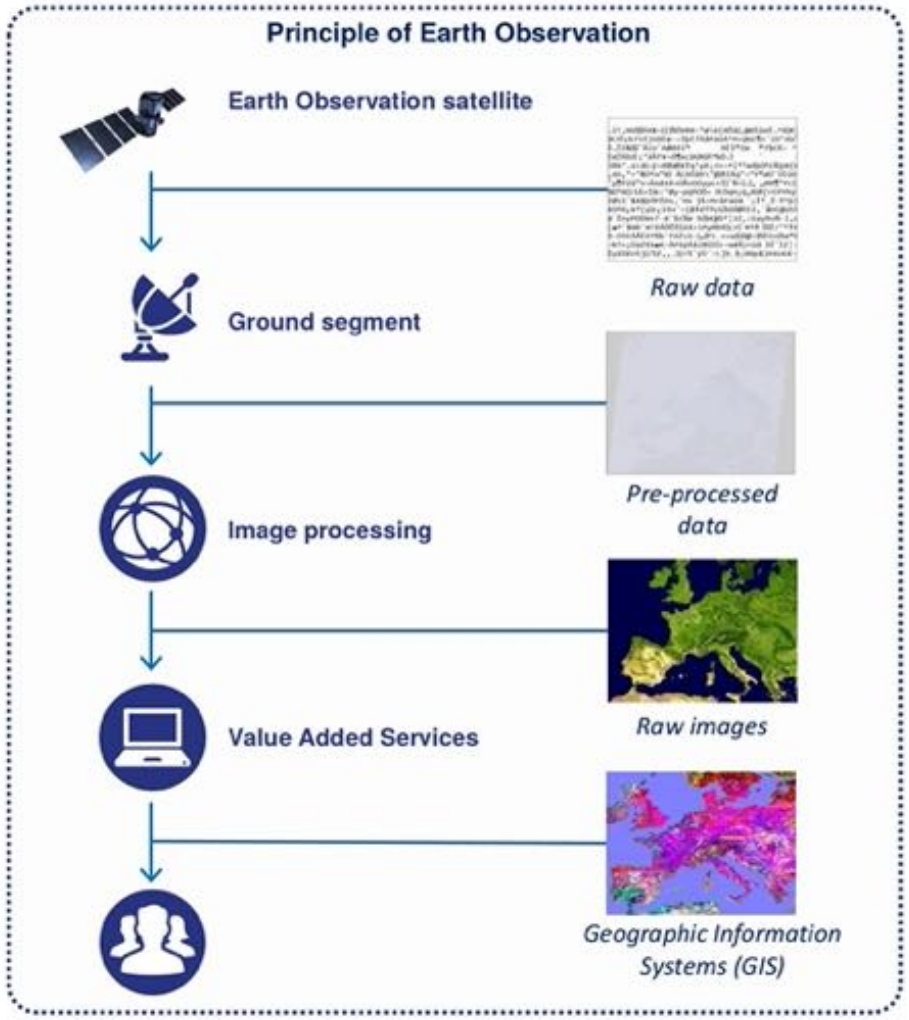
	2021 Value	2021 %	2031 Value	2031 %
Data revenues (€ m)	29	5.5	51	6.4
Value-added service revenues (€ m)	156	7.0	334	7.2

- See https://www.euspa.europa.eu/sites/default/files/uploads/euspa_market_report_2022.pdf .

Earth Observation Applications (from the same EUSPA report)

- Agriculture
- Aviation and drones
- Biodiversity, ecosystems and natural capital
- Climate services
- Consumer solutions, tourism and health
- Emergency management and humanitarian aid
- Energy and raw materials
- Environmental monitoring
- Fisheries and aquaculture
- Forestry
- Infrastructure
- Insurance and finance
- Maritime and inland waterways
- Rail
- Road and automotive
- Urban development and cultural heritage

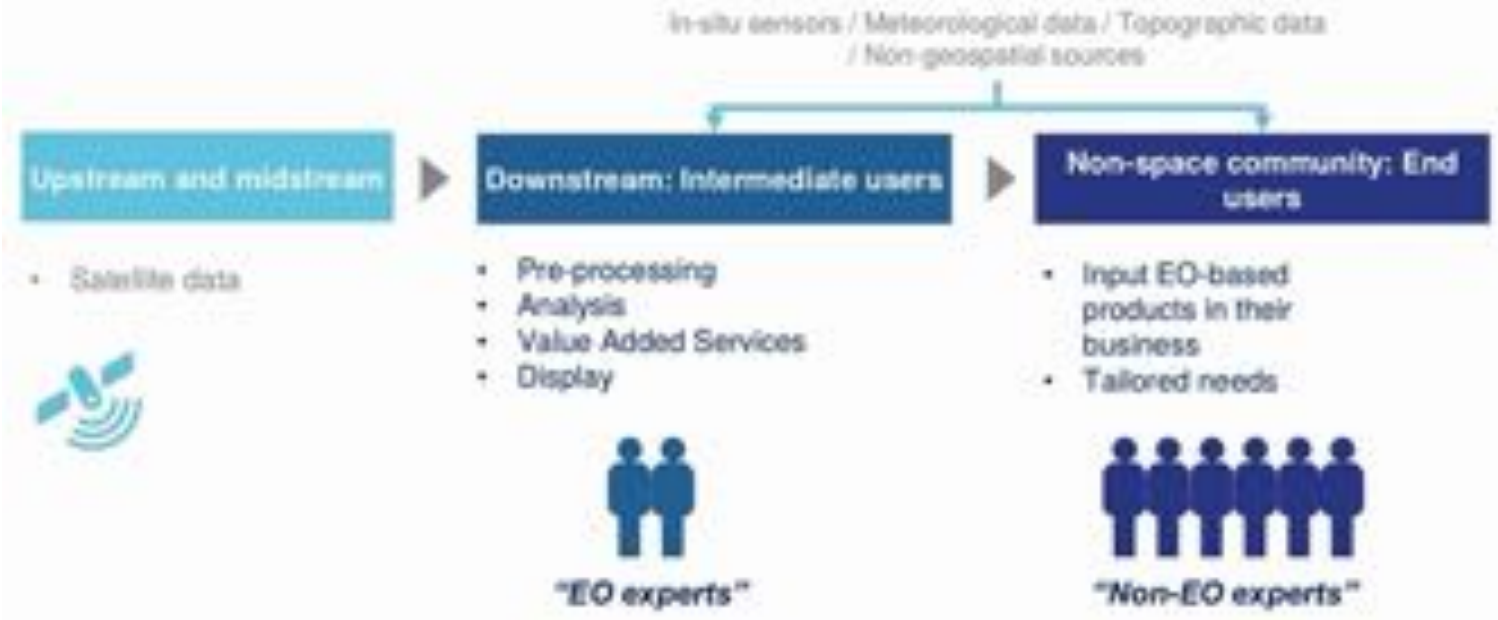
Earth Observation in Three Slides



From the report https://www.copernicus.eu/sites/default/files/2019-02/PwC_Copernicus_Market_Report_2019_PDF_version.pdf.

Earth Observation (cont'd)

Earth Observation (cont'd)



Some Information about Copernicus (http://www.copernicus.eu/)

- Copernicus is the European programme for Earth Observation.
- Copernicus collects data about our planet using a set of **dedicated satellites** (the **Sentinel** families) and **contributing missions** (existing commercial and public satellites).
- The first satellite (**Sentinel-1A**) was launched in **2014**. **Almost 20 satellites will be deployed by 2030**.
- Copernicus also collects information from **in-situ systems** such as ground stations, which deliver data acquired by a multitude of sensors on the ground, at sea or in the air.

The Five Vs of Copernicus Big Data

- **Volume** in the Copernicus Open Access Hub (https://scihub.copernicus.eu/)

Mission	No. of user-level data published in Y2021	No. of user-level data published since start of Ops	Y2021 No. as % of total published per mission since start of Ops	Volume of user-level data published in Y2021 (PiB)	Volume of user-level data published since start of Ops (PiB)	Y2021 volume as % of total published per mission since start of Ops
S1	1,439,646	7,451,432	19%	2.18	11.65	19%
S2	8,147,340	31,798,019	26%	4.18	16.20	26%
S3	3,302,695	12,350,106	27%	0.82	3.85	21%
S5P	624,541	2,016,670	31%	0.15	0.50	31%
ALL	13,514,222	53,616,227	25%	7.34	32.21	23%

Table 1: Overall number and volume of published user-level data on the Open Hub both in Y2021 and since the start of operations, per Sentinel mission

The Five Vs of Copernicus Big Data (cont'd)

- **Velocity** in the Copernicus Open Access Hub

Mission	Daily Average Vol (TiB) published in November 2021	Nov 2021 Volume as % of overall daily average	Daily Average Vol (TiB) published in November 2020	Nov 2020 Volume as % of overall daily average
S1	6.51	35%	6.23	31%
S2	9.67	51%	9.56	47%
S3	2.19	12%	4.03	20%
S5P	0.42	2%	0.40	2%
All	18.79		20.22	

Table 2: Average volume of user-level data published per day in the last month of Y2021 and Y2020, with percentage splits per Sentinel mission

The Five Vs of Copernicus Big Data (cont'd)

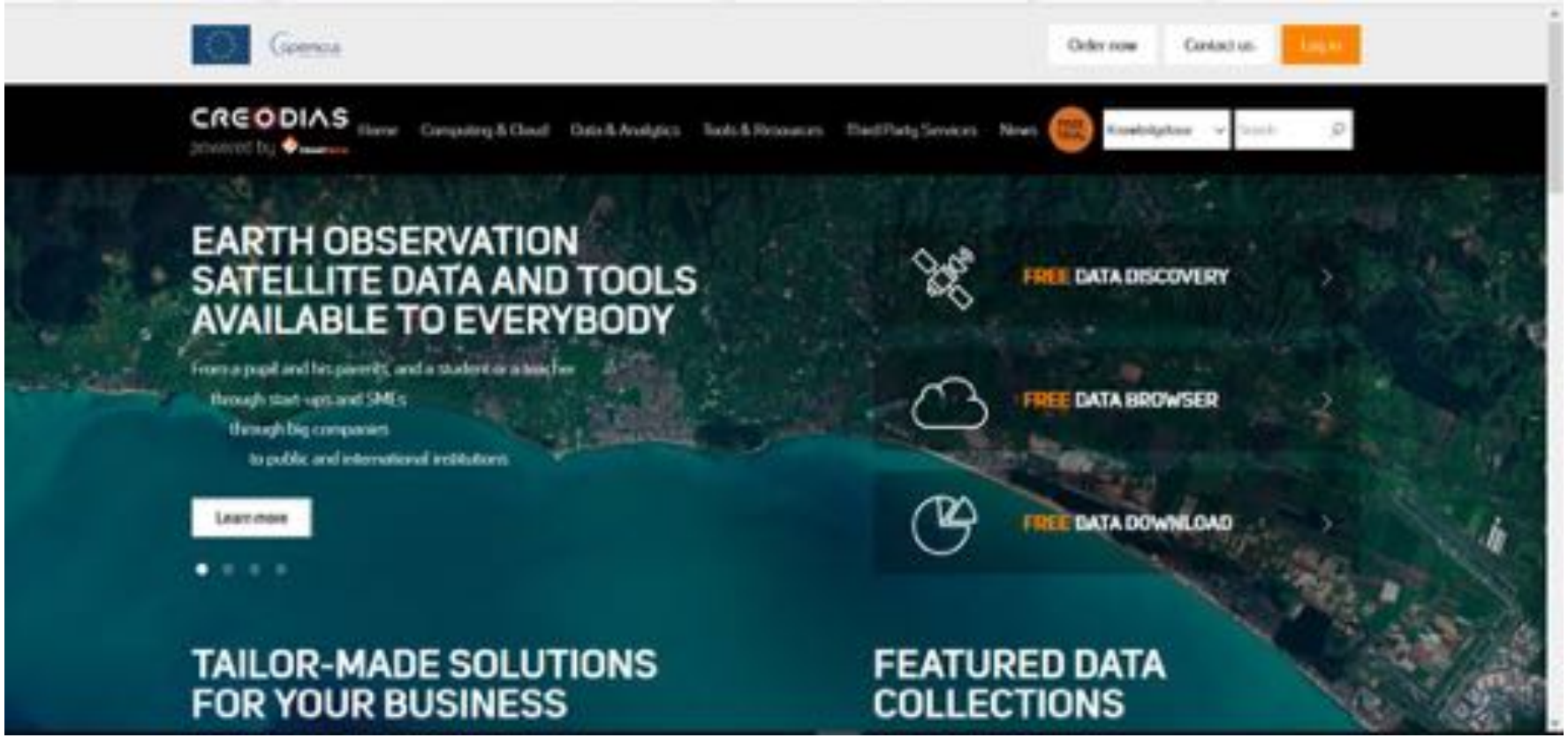
- **Variety**:
  - The Sentinel satellites comprise **different type of sensors** (e.g. optical, radar and thermal) and **different levels of processing** (from raw to advanced products).
  - Datasets used for geospatial applications can be composed not only by satellite data but also by **aerial imagery, in-situ data** and **other collateral information** (e.g. media data, public government data, etc.).
  - This wealth of data is processed by EO actors to extract **information and knowledge.** This information and knowledge is also **Big** and **similar Big Data challenges apply**. For example, 1PB of Sentinel data may consist of about 750.000 datasets which, when processed, about **450TB of content information and knowledge** (e.g. classes of objects detected) can be generated.

The Five Vs of Copernicus Big Data (cont'd)

- **Veracity:** Decision-making and operations require **reliable sources**. Thus, **assessing the quality of the data is important** for whole information extraction chain.
- **Value:** The Copernicus programme has big economic impact as we discussed earlier.

Copernicus Data and Information Access Services (DIAS)

- **Five DIAS** now in operation
- One of them used linked data for their catalogue: **https://creodias.eu/**



Copernicus Services

- **Copernicus Services** (https://www.copernicus.eu/en/services) transform the wealth of satellite and in-situ Copernicus data into **value-added products** by processing and analysing the data.

- There are six Copernicus services covering the following thematic areas: **Atmosphere, Marine, Land, Climate, Emergency** and **Security.**






Two Examples of Copernicus Services Products

- The CORINE land cover dataset (available at http://land.copernicus.eu/pan-european/corine-land-cover).

- Global solar UV index forecast (available at http://atmosphere.copernicus.eu/catalogue#/).

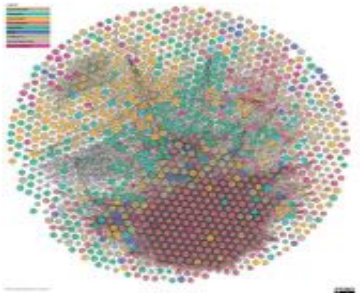
The CORINE Land Cover Dataset of 2012 (most recent version 2018)

- It covers **39 European countries**.
- Land cover is characterized using a **3-level hierarchy of classes** (e.g., olive groves or vineyards) with 44 classes in total at the 3rd level.
- The **minimum mapping unit** is 25 hectares for areal phenomena and 100 meters for linear phenomena.
- It is made available in **raster (GeoTIFF)** and **vector (ESRI/SQLite geodatabase)** formats.

Land Monitoring Service

Main Objective of Our Work Since 2010

- **Open up EO data silos** by publishing their metadata, data and the information and knowledge extracted from this data on the Web using **Semantic Web, Linked Data and Knowledge Graph technologies**.



Semantic Web

RDF

W3C SPARQL

W3C OWL2

Why Linked Data?

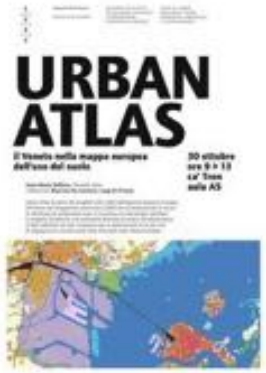
The vision of **linked data** is to go from a Web of documents to a Web of data:

- Unlock open data dormant in their silos
- Make it available on the Web using Semantic Web technologies (HTTP, URIs, RDF, SPARQL)
- Interlink it with other data (e.g., from the European data portal)

Examples of Linked Open EO Data

- CORINE land cover of the year 2012
- Urban Atlas of the year 2012

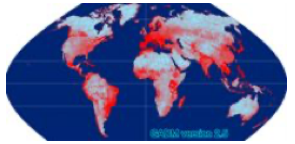
https://ai.di.uoa.gr/#datasets



Examples of Interesting Linkages

- The CORINE land cover dataset can be usefully linked with the following datasets:
  - GeoNames
  - Global Administrative Areas
  - DBpedia
  - OpenStreetMap

GeoNames

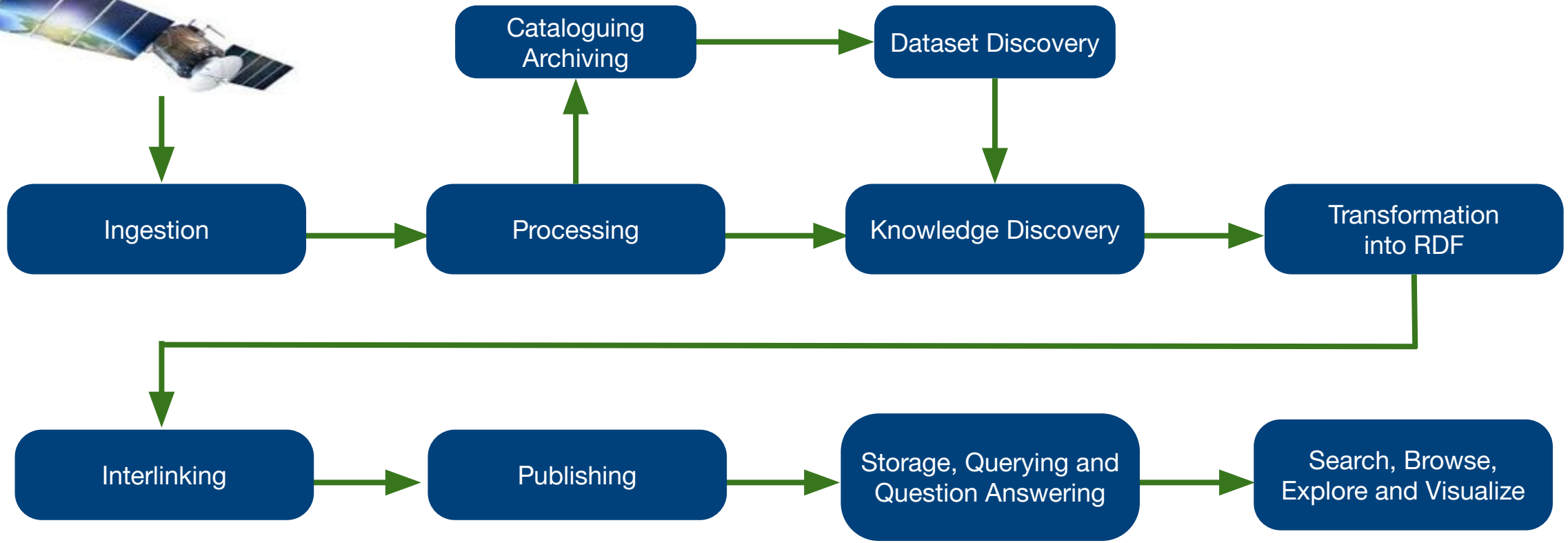


Copernicus Data, Information and Knowledge as Open Linked Data: Benefits

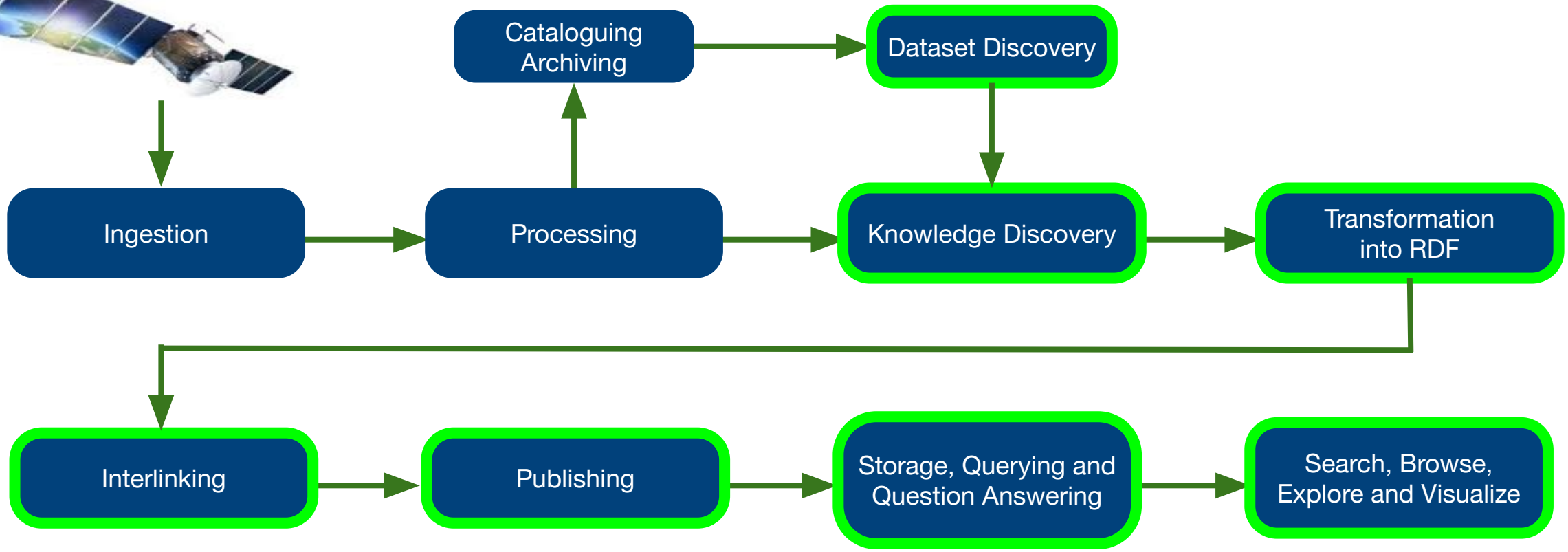
- Make Copernicus data more **easily discoverable by search engines and new services like Google Dataset Search** by using technologies such as **schema.org** for encoding the metadata. schema.org is now used by all major search engines.
- Once datasets are transformed into linked data (e.g., the CORINE land cover dataset), we can **interlink** them with other open linked data sources (e.g., GADM, OpenStreetMap or DBpedia data) to build **geospatial knowledge graphs.**
- Enable **semantics-based querying** and **visualization** of these graphs.
- Enable **question answering** using natural language questions.
- **Therefore: enable easier utilization** e.g., by software developers who may not be specialists in Earth Observation.

The data science pipeline for big linked Earth observation data

The Data Science Pipeline



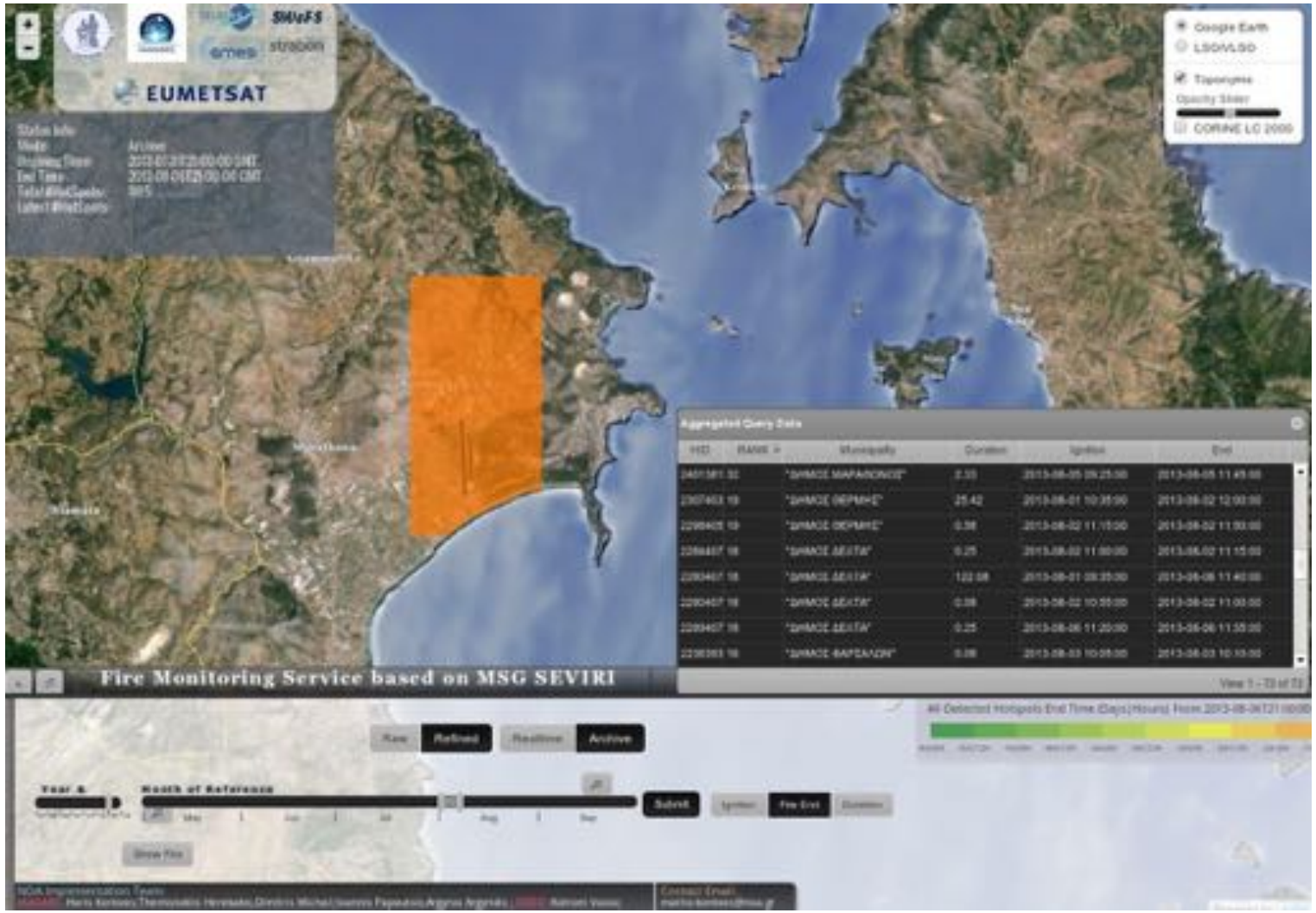
The Data Science Pipeline



M. Koubarakis et al. Managing Big, Linked, and Open Earth-Observation Data: Using the TELEIOS/LEO software stack. In: IEEE GRSM (2016).
M. Koubarakis et al. Big, Linked Geospatial Data and Its Applications in Earth Observation. IEEE Internet Computing 21(4), pages 87-91, 2017.

Applications

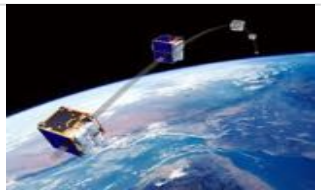
The FIREHUB service of the National Observatory of Athens (http://195.251.203.238/seviri/)



M. Koubarakis et al. Real-time wildfire monitoring using scientific database and linked data technologies. EDBT 2013.

Precision Farming

RapidEye, Landsat, Sentinel 2 images

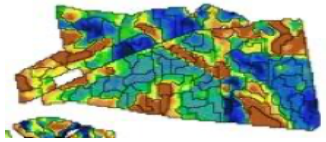


Processing

Vista

Water bodies
Protected areas
Legal regulations

Biomass Map
Fertilization Map



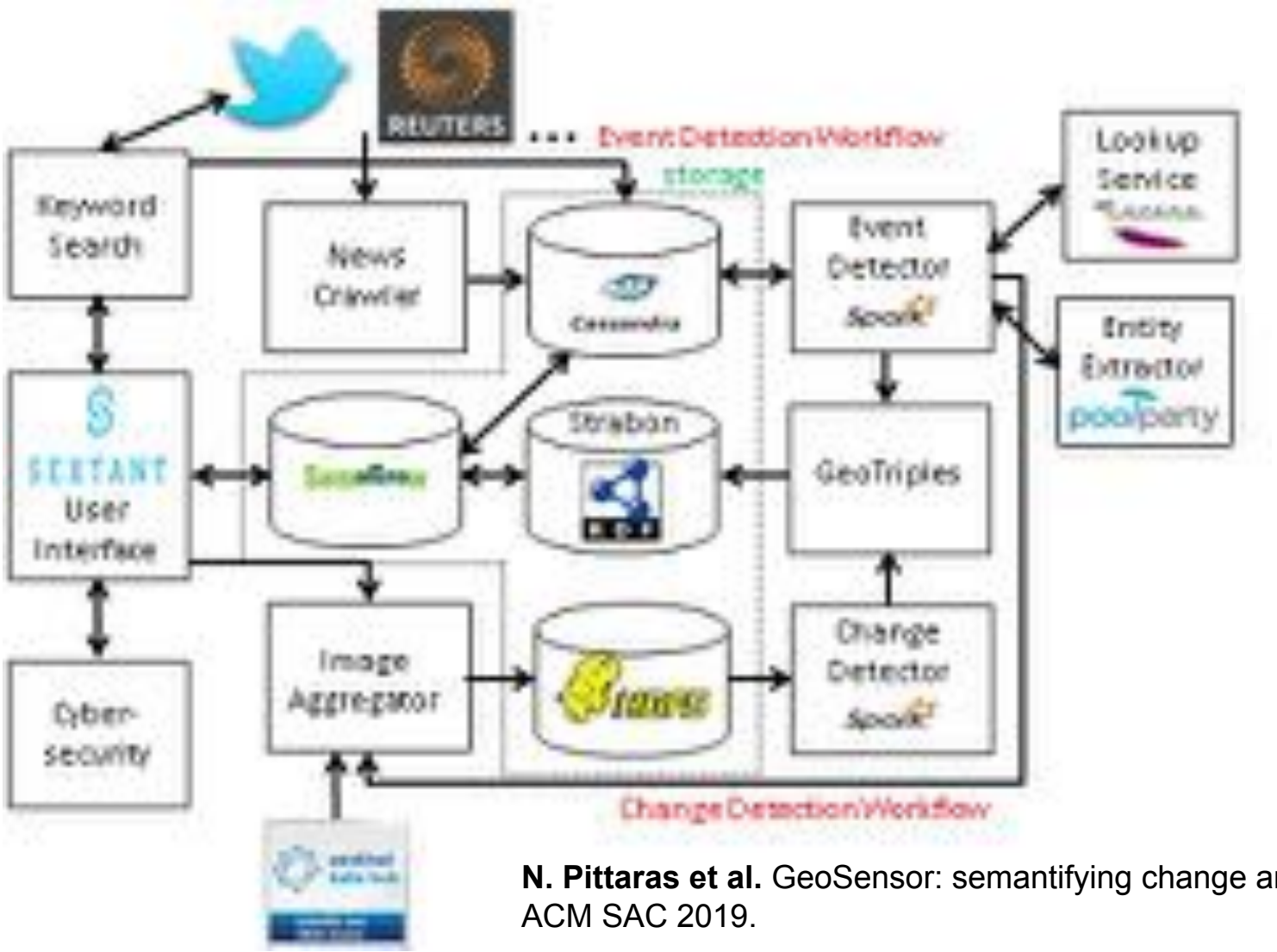
Precision Farming Application

pcagrar

LEOpatra

S. Burgstaller et al. LEOpatra: A Mobile Application for Smart Fertilization Based on Linked Data. HAICTA 2017.

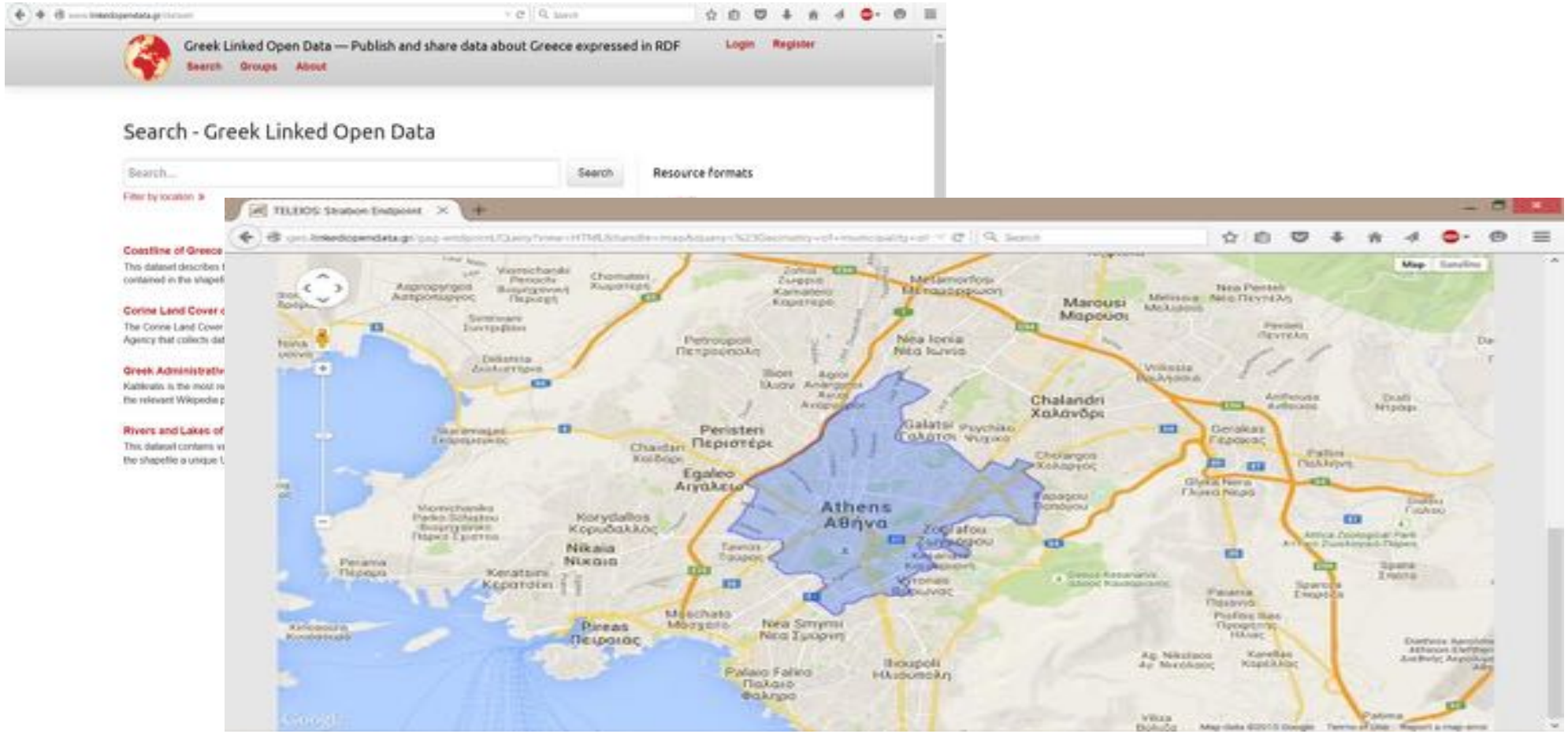
Change Detection Pilot in BigDataEurope



SatCen
European Union Satellite Centre

N. Pittaras et al. GeoSensor: semantifying change and event detection over big data. ACM SAC 2019.

Education (http://linkedopendata.gr/)



Coming up …

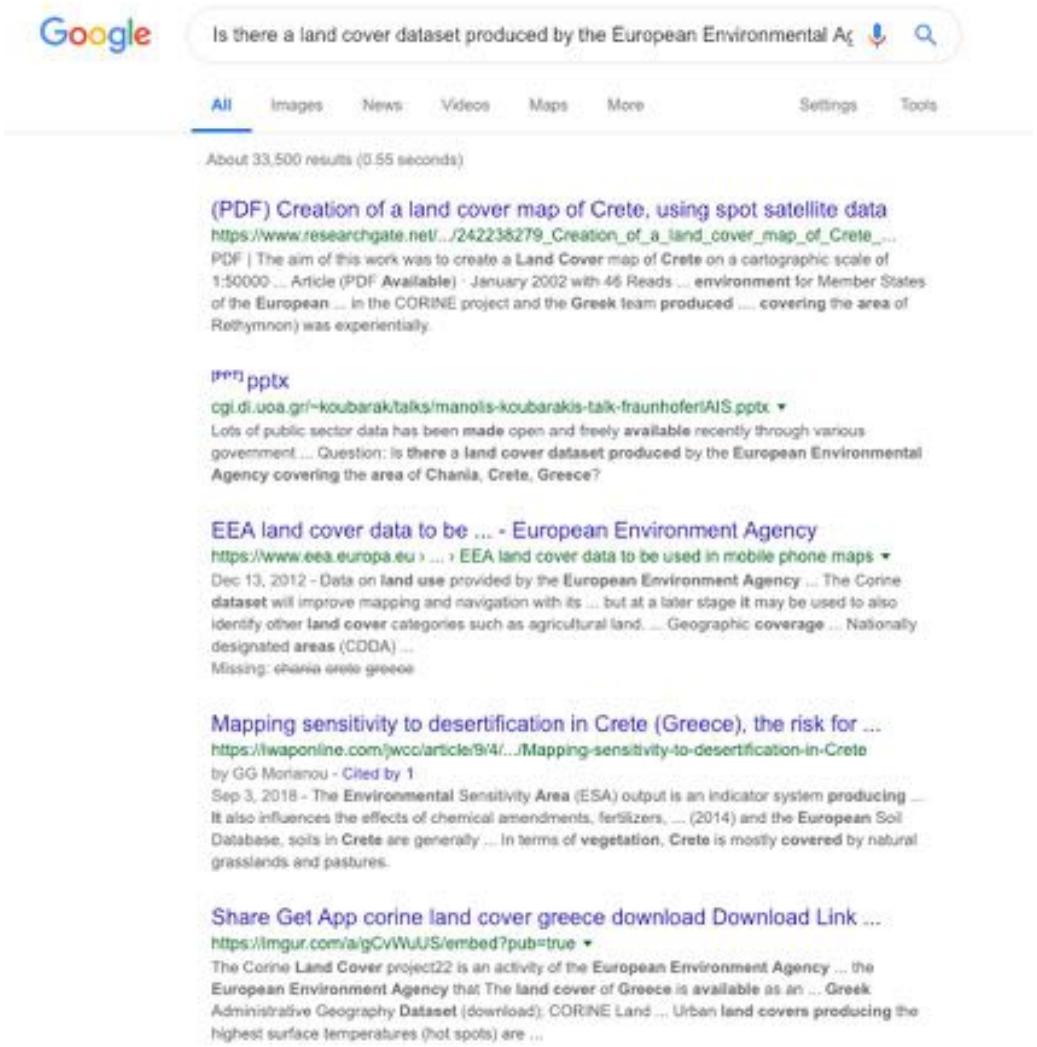
- Book on “Geospatial Data Science: A hands-on approach for building geospatial applications using linked data technologies” (to be published by ACM Books).

Discovering Earth Observation data

Using Google for dataset discovery

- Is there a land cover dataset produced by the European Environmental Agency covering the area of Chania, Crete, Greece?
- Google it!

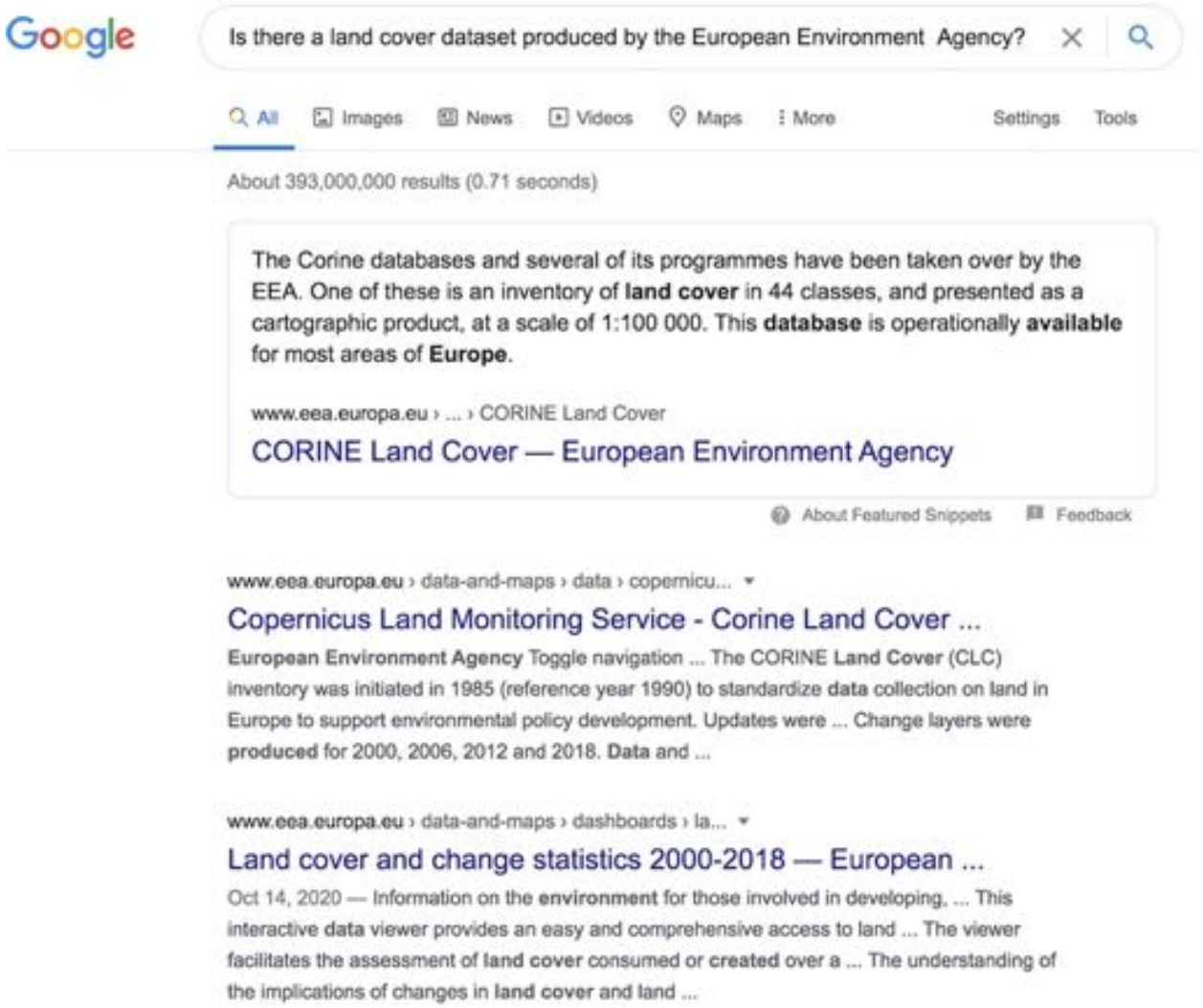
Results


Google
Is there a land cover dataset produced by the European Environmental Ag
All
Images
News
Videos
Maps
More
Settings
Tools
About 33,500 results (0.55 seconds)
(PDF) Creation of a land cover map of Crete, using spot satellite data
https://www.researchgate.net/.../242238279_Creation_of_a_land_cover_map_of_Crete_...
PDF | The aim of this work was to create a Land Cover map of Crete on a cartographic scale of 1:50000 ... Article (PDF Available) · January 2002 with 46 Reads ... environment for Member States of the European ... in the CORINE project and the Greek team produced .... covering the area of Rethymnon) was experientially.
[PPT] pptx
cgi.di.uoa.gr/~koubarak/talks/manolis-koubarakis-talk-fraunhoferIAIS.pptx
Lots of public sector data has been made open and freely available recently through various government ... Question: Is there a land cover dataset produced by the European Environmental Agency covering the area of Chania, Crete, Greece?
EEA land cover data to be ... - European Environment Agency
https://www.eea.europa.eu › ... › EEA land cover data to be used in mobile phone maps
Dec 13, 2012 - Data on land use provided by the European Environment Agency ... The Corine dataset will improve mapping and navigation with its ... but at a later stage it may be used to also identify other land cover categories such as agricultural land. ... Geographic coverage ... Nationally designated areas (CDDA) ...
Missing: chania crete greece
Mapping sensitivity to desertification in Crete (Greece), the risk for ...
https://iwaponline.com/jwcc/article/9/4/.../Mapping-sensitivity-to-desertification-in-Crete
by GG Morianou - Cited by 1
Sep 3, 2018 - The Environmental Sensitivity Area (ESA) output is an indicator system producing ... It also influences the effects of chemical amendments, fertilizers, ... (2014) and the European Soil Database, soils in Crete are generally ... In terms of vegetation, Crete is mostly covered by natural grasslands and pastures.
Share Get App corine land cover greece download Download Link ...
https://imgur.com/a/gCvWuUS/embed?pub=true
The Corine Land Cover project22 is an activity of the European Environment Agency ... the European Environment Agency that The land cover of Greece is available as an ... Greek Administrative Geography Dataset (download): CORINE Land ... Urban land covers producing the highest surface temperatures (hot spots) are ...

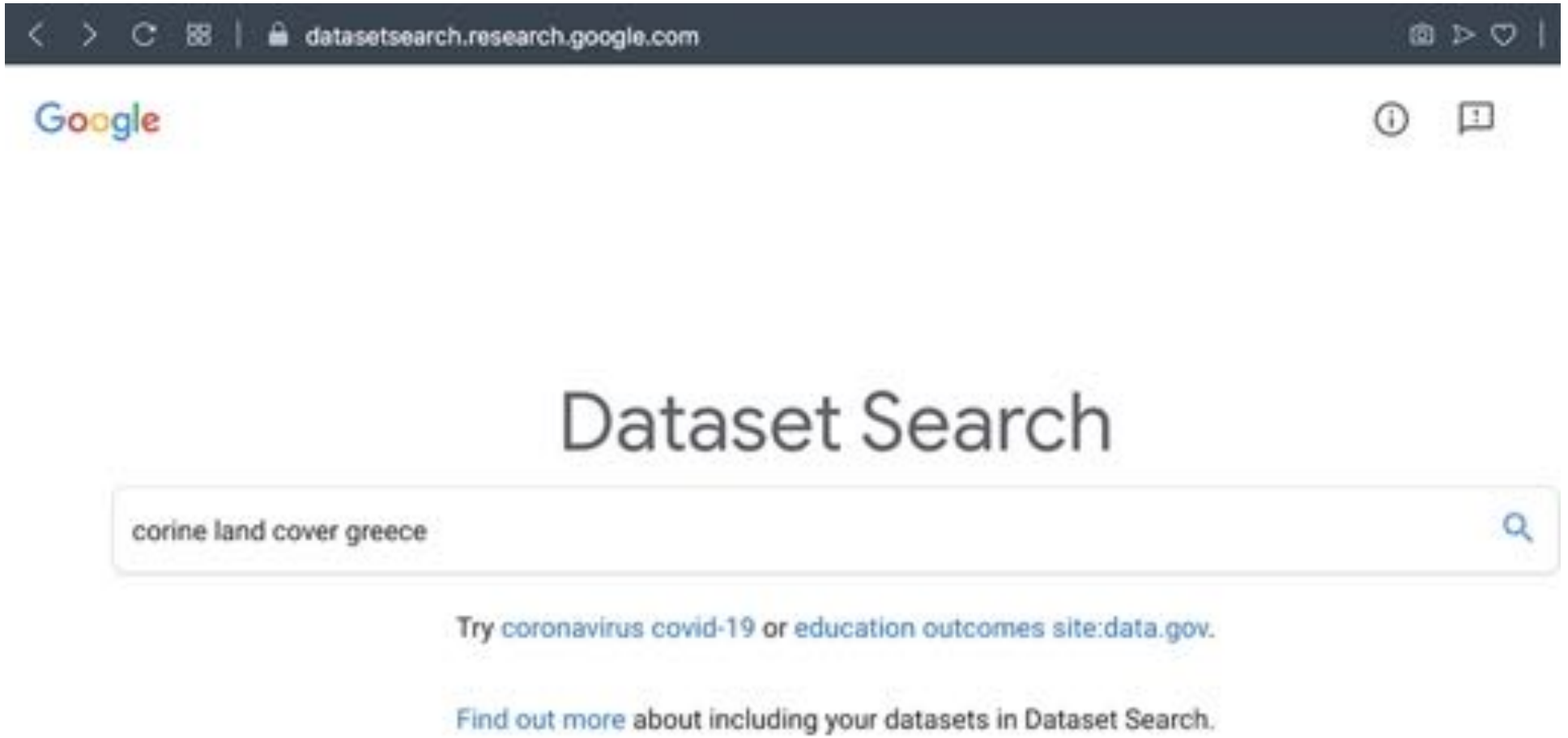
Let us pose a different query

- Is there a land cover dataset produced by the European Environment Agency?

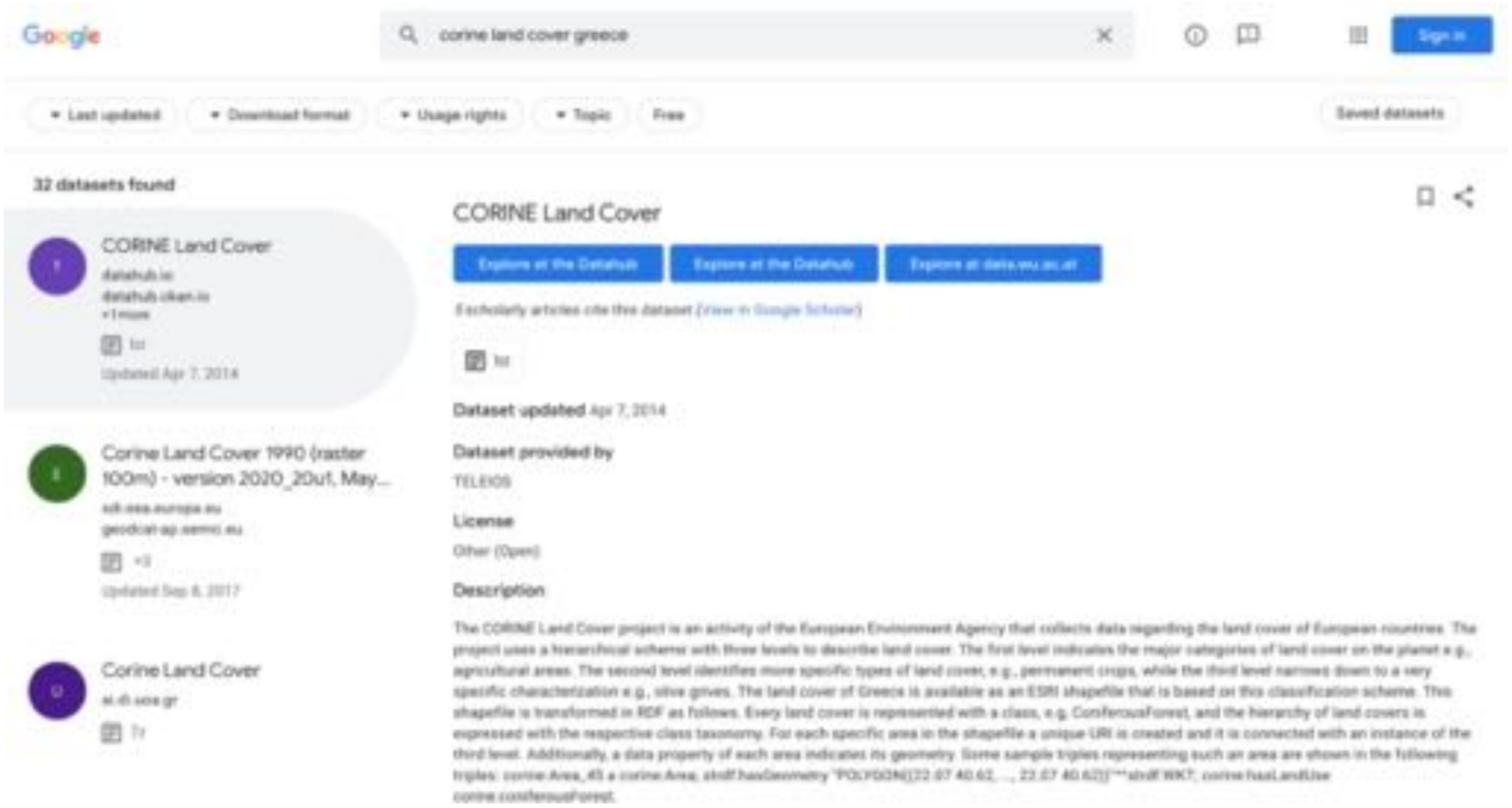
Results


Google
Is there a land cover dataset produced by the European Environment Agency?
All
Images
News
Videos
Maps
More
Settings
Tools
About 393,000,000 results (0.71 seconds)
The Corine databases and several of its programmes have been taken over by the EEA. One of these is an inventory of land cover in 44 classes, and presented as a cartographic product, at a scale of 1:100 000. This database is operationally available for most areas of Europe.
www.eea.europa.eu › ... › CORINE Land Cover
CORINE Land Cover — European Environment Agency
About Featured Snippets
Feedback
www.eea.europa.eu › data-and-maps › data › copernicu...
Copernicus Land Monitoring Service - Corine Land Cover ...
European Environment Agency Toggle navigation ... The CORINE Land Cover (CLC) inventory was initiated in 1985 (reference year 1990) to standardize data collection on land in Europe to support environmental policy development. Updates were ... Change layers were produced for 2000, 2006, 2012 and 2018. Data and ...
www.eea.europa.eu › data-and-maps › dashboards › la...
Land cover and change statistics 2000-2018 — European ...
Oct 14, 2020 — Information on the environment for those involved in developing, ... This interactive data viewer provides an easy and comprehensive access to land ... The viewer facilitates the assessment of land cover consumed or created over a ... The understanding of the implications of changes in land cover and land ...

Google Dataset Search



Results


Google
corine land cover greece
Sign in
Last updated
Download format
Usage rights
Topic
Free
Saved datasets
32 datasets found
CORINE Land Cover
Corine Land Cover 1990 (raster 100m) - version 2020_20u1, May...
Corine Land Cover
CORINE Land Cover
Dataset updated
Dataset provided by
TELEIOS
License
Other (Open)
Description

Google Dataset Search

- Datasets that are indexed using **schema.org**, as proposed by Google, show up.
- Enables users to find datasets stored across the Web by doing a simple keyword search.
- Uncovers information about datasets hosted in thousands of repositories across the web, making these datasets universally accessible and useful.

How does Dataset Search find datasets?

Authors need to add metadata in **schema.org** to each page that describes a dataset.

Schema.org:

- Founded by Google, Microsoft, Yahoo!, Yandex. Currently schema.org vocabularies are developed by an open community process.

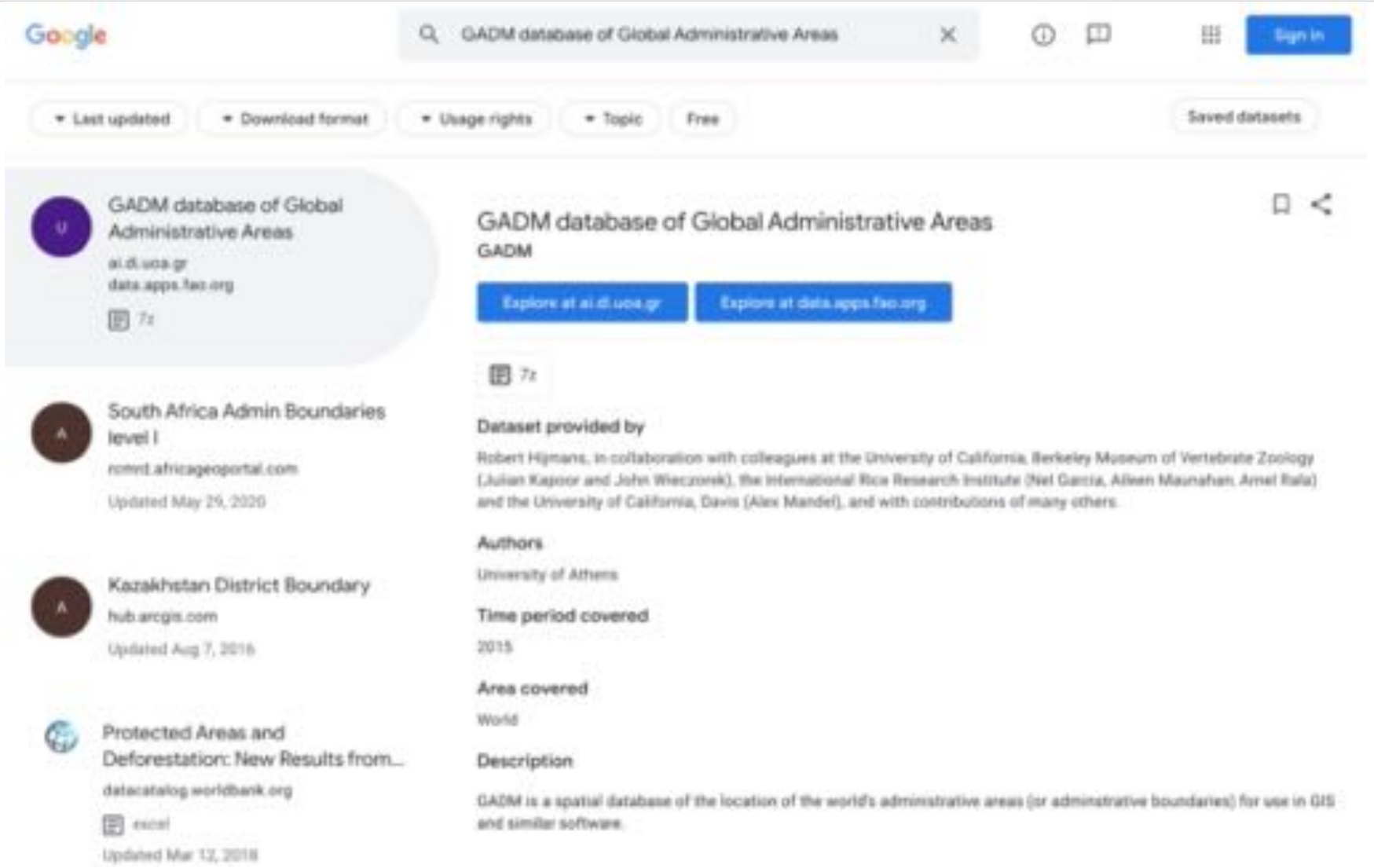
- Provides a unique structured data markup schema to annotate a variety of topics.
  - Tags added to HTML as JSON-LD, Microdata, or RDFa.

- On-page markup allows search engines to understand information included in web pages.

Schema.org

```
<script type="application/ld+json">{
    "@context": "http://schema.org",
    "@type": ["ItemList", "Dataset"],
    "itemListOrder": "http://schema.org/ItemListOrderAscending",
    "numberOfItems": "7",
    "itemListElement": [{
        "@type": "ListItem", "position": 1,
        "item":{
            "@type" : "Dataset",
            "name": "GADM database of Global Administrative Areas",
            "alternateName": "GADM",
            "description": "GADM is a spatial database of the location ...",
            "author": "University of Athens",
            "sourceOrganization": "Robert Hijmans, in collaboration with ...",
            "copyrightYear": "2018",
            "keywords":["GADM", "Global Administrative Areas", "GADM 2015"],
            "spatialCoverage": "World", "temporalCoverage": "2015", "fileFormat": "7z",
            "isBasedOn": "GADM database of Global Administrative Areas 2015, Version 2.8",
            "isAccessibleForFree": true,
            "distribution": {
                "@type":"DataDownload", "encodingFormat":"7z",
                "contentUrl":"https://datahub.ckan.io/dataset/gadm"},
            "url": "https://datahub.ckan.io/dataset/gadm"}}, ... ]} </script>
```

Google Dataset Search



Google Dataset Search by the Numbers

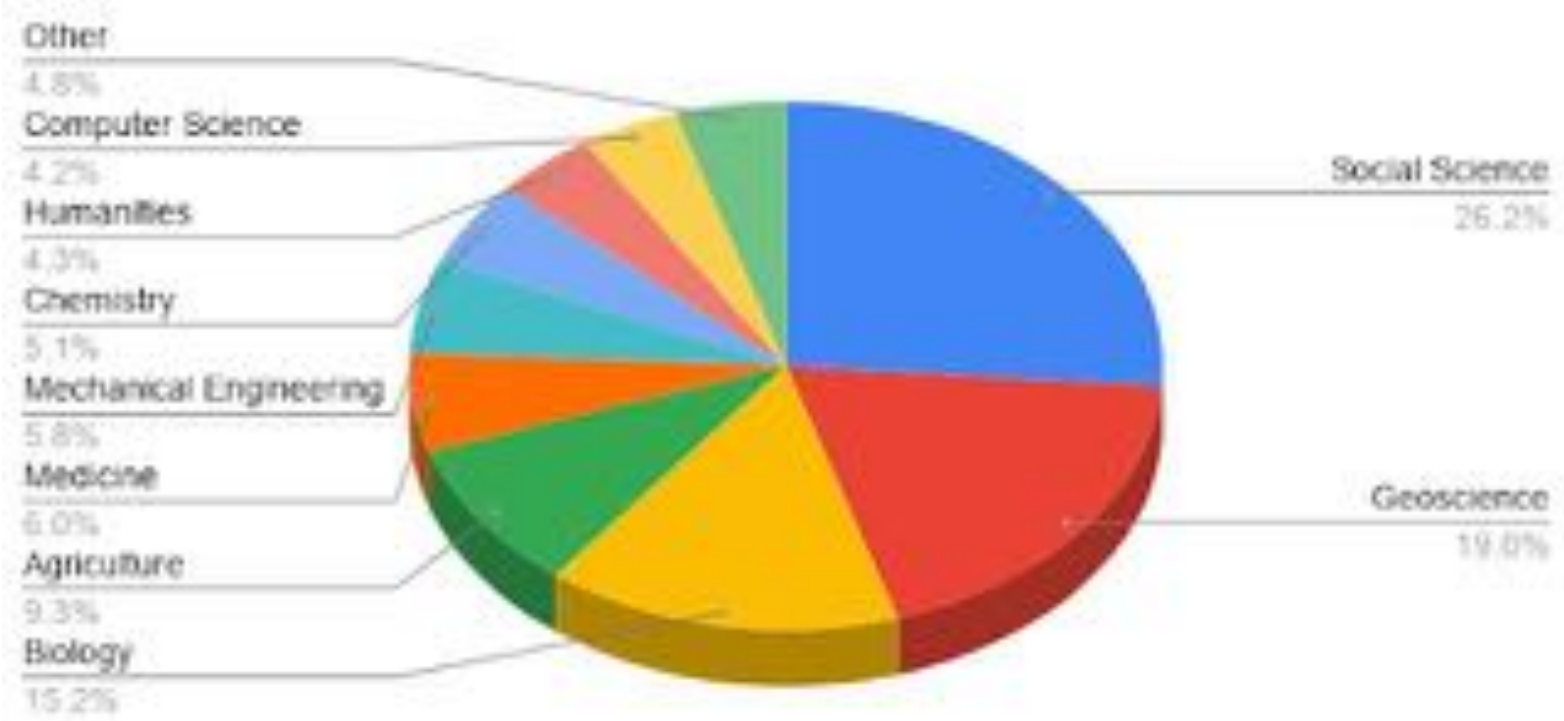
- O. Benjelloun, S. Chen, N. Noy: **Google Dataset Search by the Numbers** (Jun 2020) https://arxiv.org/abs/2006.06894
  - As of March 2020, the corpus contained 28 million datasets from more than 3,700 sites
  - The corpus is a reasonably representative snapshot of the datasets published on the Web, but there is no way of measuring how well the corpus covers all the datasets available on the Web

Google Dataset Search by the Numbers

- O. Benjelloun, S. Chen, N. Noy: **Google Dataset Search by the Numbers** (Jun 2020) https://arxiv.org/abs/2006.06894
  - **Licenses and access:** Only 34% of the datasets provide any licensing information
    - most of datasets available for free, almost always allowed reuse for both commercial and non-commercial purposes
  - **Linked Data**: Fewer than 1% of datasets in the corpus are in linked data formats
    - there is plenty of shared data that the Semantic Web community produces, but the final step of describing it appears to be less common

Google Dataset Search - An analysis of online datasets

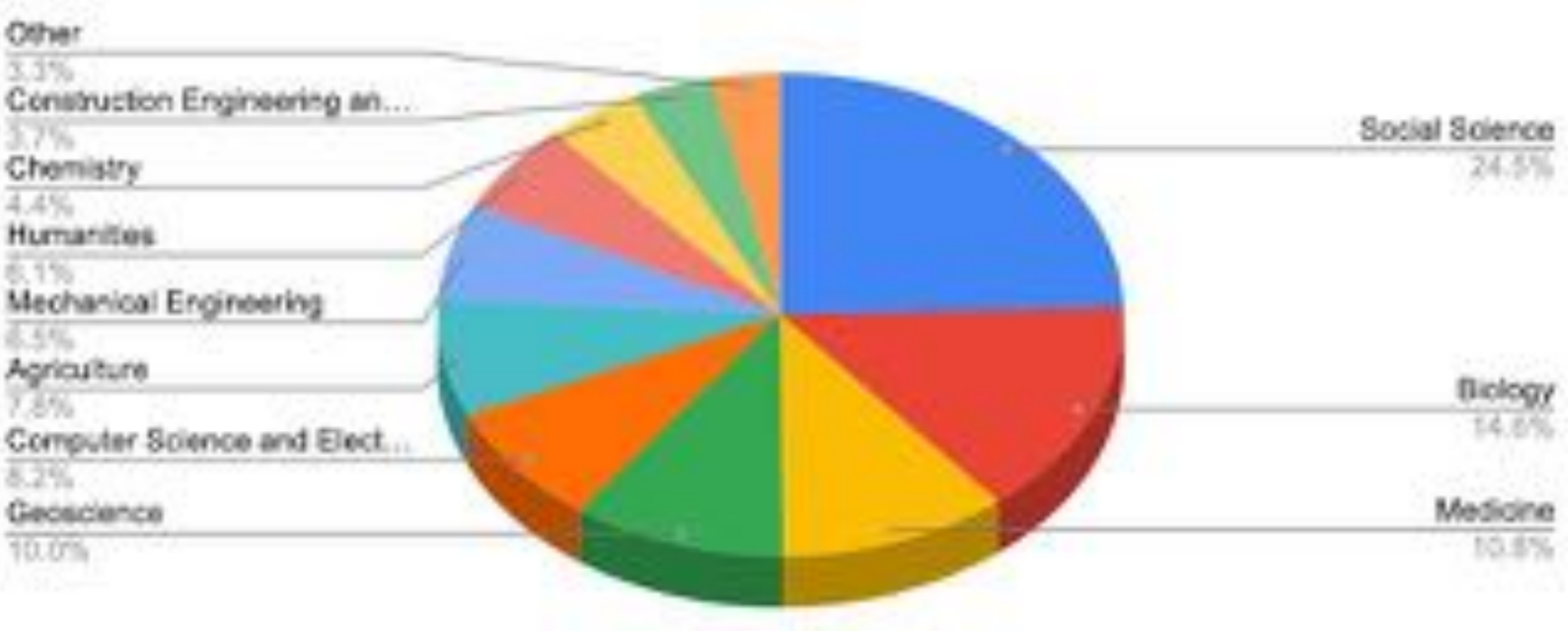
- Distribution of dataset topics



Reference: https://ai.googleblog.com/2020/08/an-analysis-of-online-datasets-using.html

Google Dataset Search - An analysis of online datasets

- What do users access?

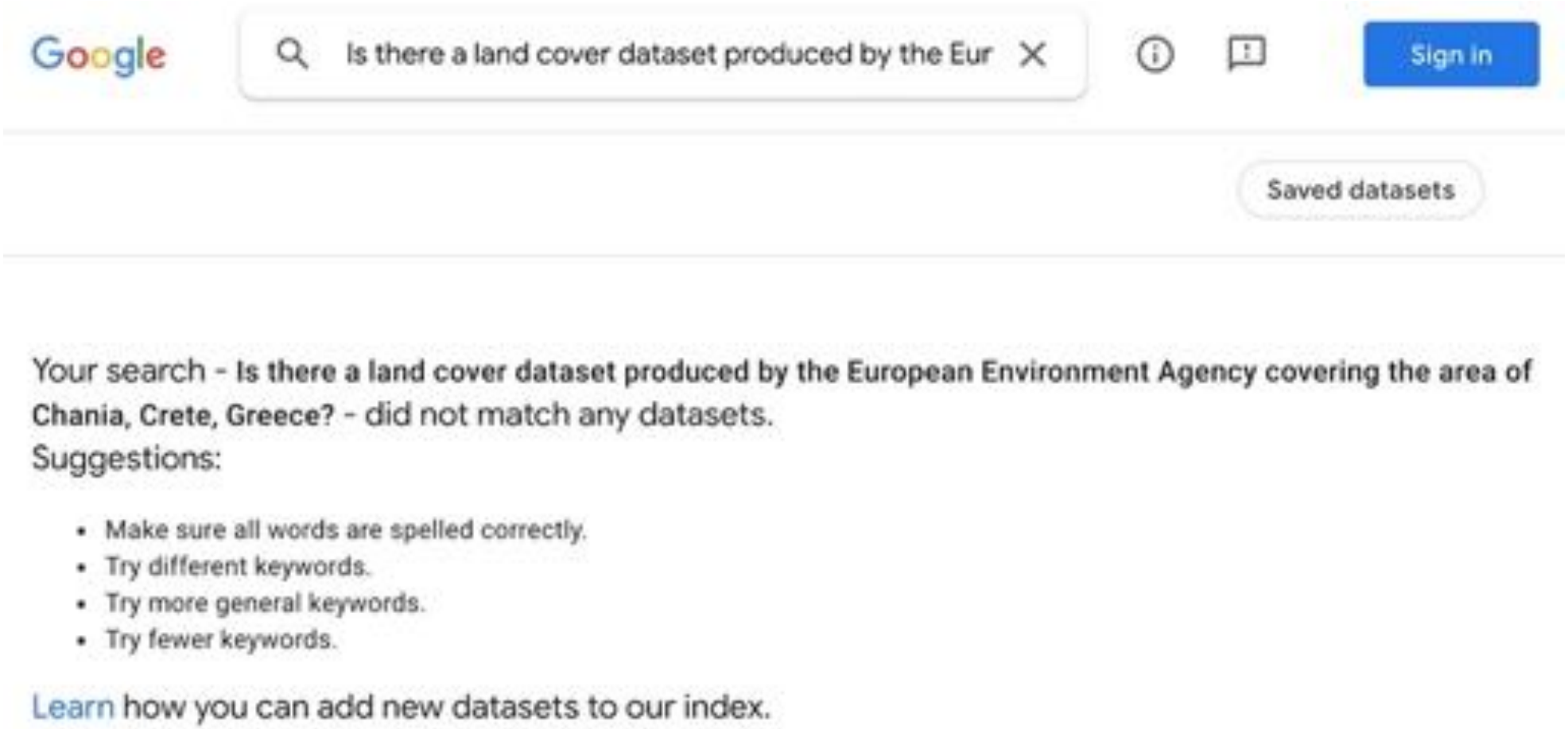


Reference: https://ai.googleblog.com/2020/08/an-analysis-of-online-datasets-using.html

Google Dataset Search

- T. Alrashed, D. Paparas , O. Benjelloun, Y. Sheng, and N. Noy: **Dataset or Not? A Study on the Veracity of Semantic Markup for Dataset Pages** (October 2021) https://research.google/pubs/pub50547/
  - **Schema.org** has become **prevalent** on the Web as a way to express the semantics of Web page content
    - it is present on more than **30%** of Web pages
  - We cannot always take **Schema.org/Dataset** markup at face value
    - pages may include this markup erroneously or for the purposes of search-engine optimization

Let us pose our original query to Dataset Search



EO dataset search - Copernicus Open Access Hub

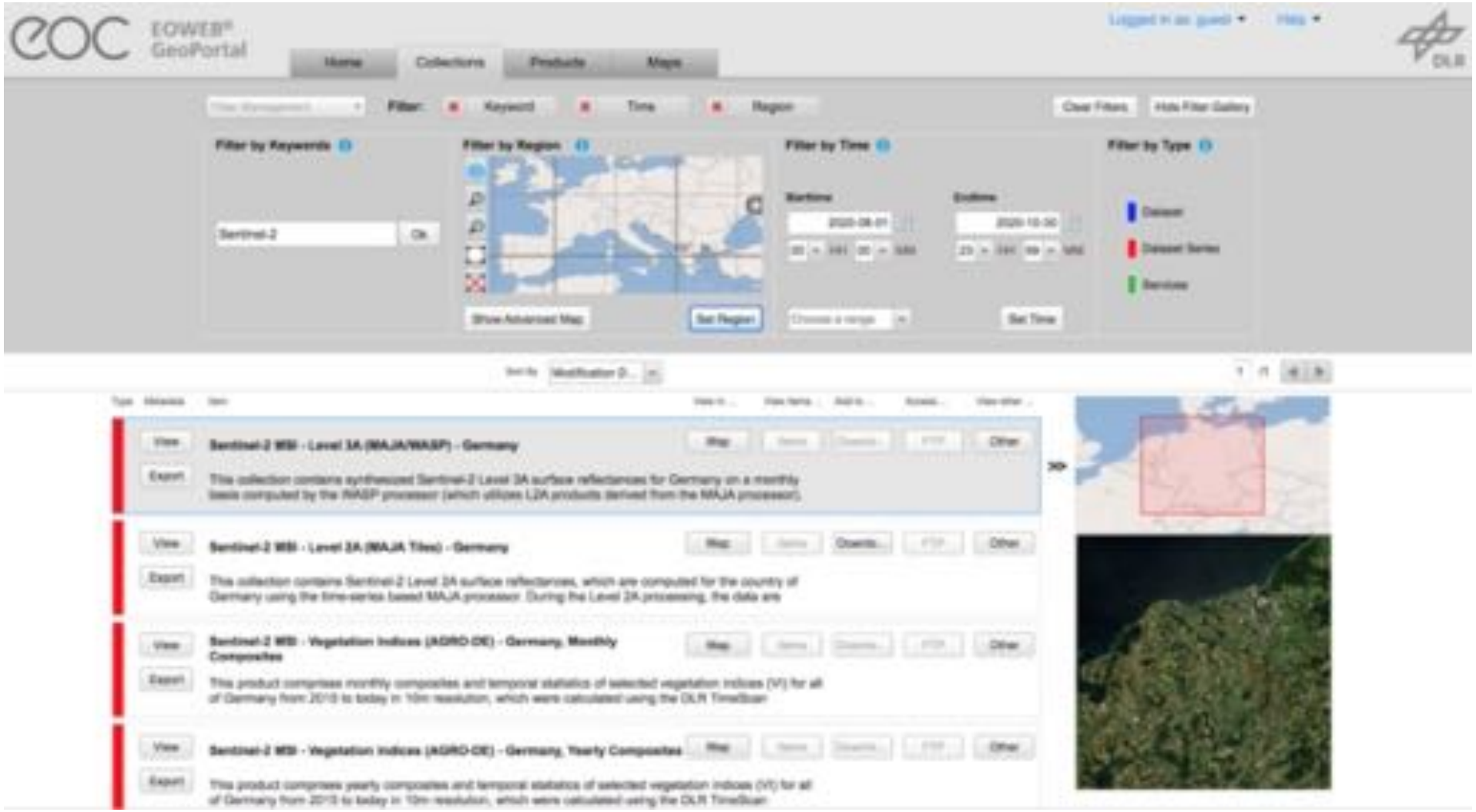
- Provides complete, free and open access to Sentinel-1, Sentinel-2, Sentinel-3 and Sentinel-5P user products (https://scihub.copernicus.eu)

EO dataset search - Copernicus Open Access Hub

EO dataset search - EOWEB® GeoPortal

- A multi-mission web portal for interactive access to the German Aerospace Center (DLR) Earth observation data holdings
  - Combines classic discovery and order services for data held in the German Satellite Data Archive (D-SDA) with browse and download features via interoperable, OGC-compliant visualization and download services (https://eoweb.dlr.de/egp/)

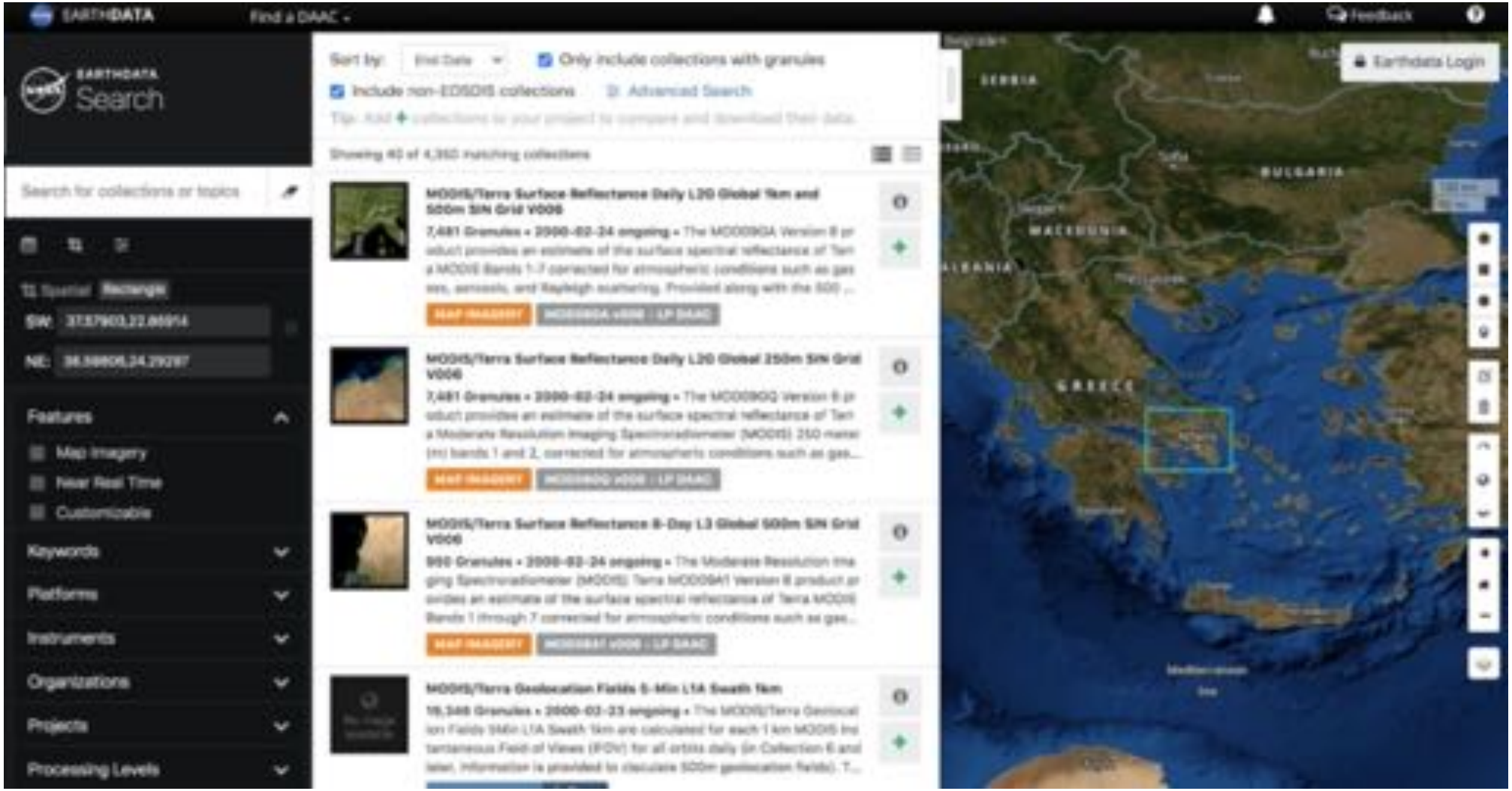
EO dataset search - EOWEB® GeoPortal



EO dataset search - Earthdata Search

- Earthdata Search enables data discovery and access to more than 33,000 EO data collections from NASA's EOSDIS, U.S. and international partner agencies (search.earthdata.nasa.gov/)
  - Users can:
    - **Search for EO data:** Earthdata Search uses the Common Metadata Repository (CMR) for sub-second search
    - **Preview EO data:** Using Global Imagery Browse Services (GIBS), enables high-performance data visualization
    - **Download and access EO data:** In addition to direct download, surfaces OPeNDAP services for simpler spatial and parameter subsetting

EO dataset search - Earthdata Search

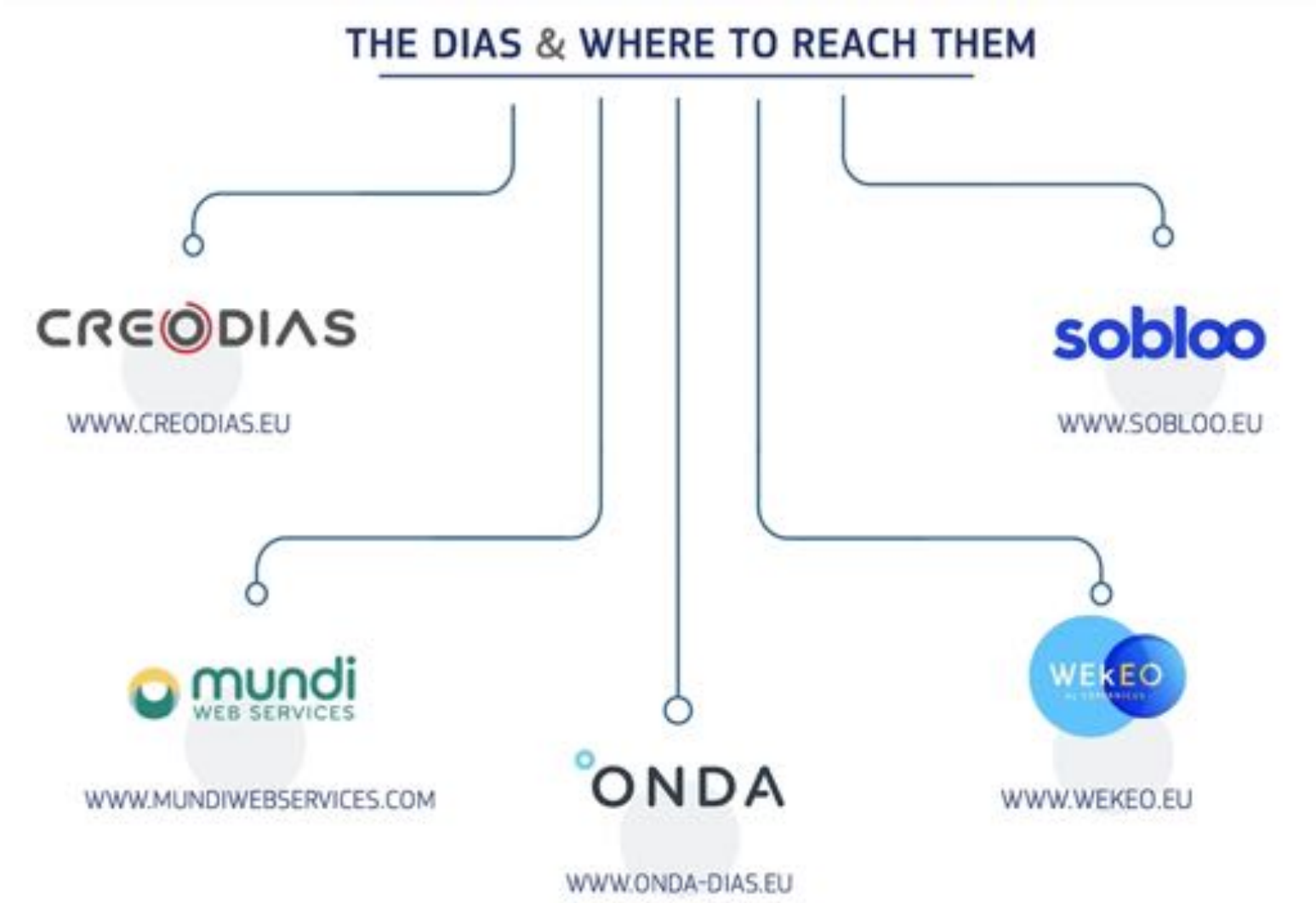


Copernicus Data and DIAS

Data and Information Access Services (DIAS):

- Five competitive platforms for quick access to a huge resource of Earth Observation data (satellite imagery) and Copernicus themed services.
- Easy and quick search, processing and sharing of satellite data.
- Quick access to satellite imagery via virtual machines.
- Data is free but computing power on a DIAS platform needs to be paid for.

Copernicus Data and DIAS



CREODIAS - Uses linked data in its catalogue!

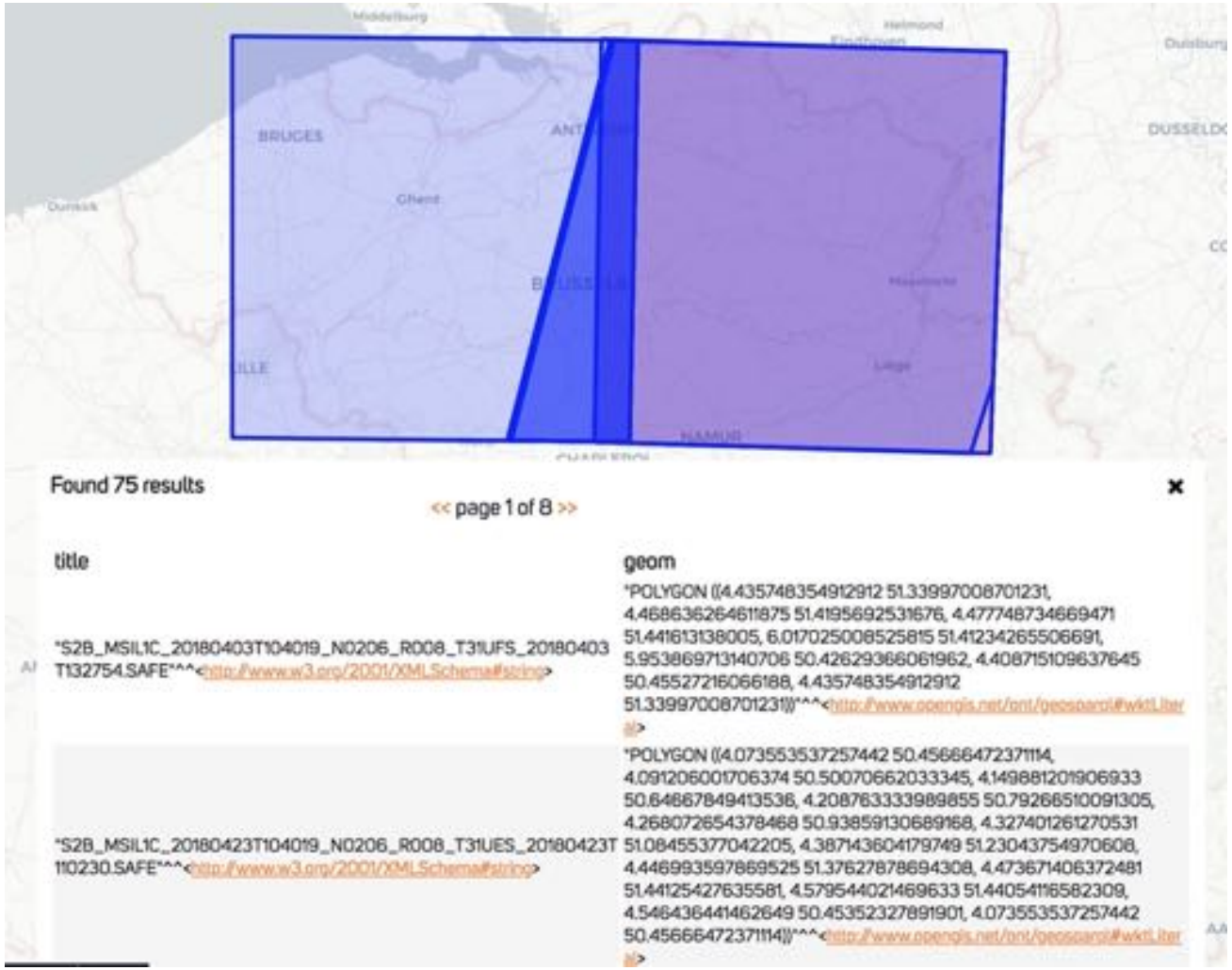
SPARQL interface: provides extended search capabilities for linking metadata of all products stored in the repository with various information from the Internet

Example Query: Find all Sentinel-2 images in the area of Brussels

SPARQL Query:

```
SELECT DISTINCT ?title ?geom {
?feature <http://ws.eodias.eu/metadata/attribute#title> ?title.
?feature <http://ws.eodias.eu/metadata/attribute#mission>
                    <http://ws.eodias.eu/metadata/mission/Sentinel-2> .
?hex      <http://ws.eodias.eu/metadata/attribute#feature> ?feature.
?hex      <http://ws.creodias.eu/metadata/object/settlement>
                    <http://dbpedia.org/resource/Brussels> .
?feature <http://ws.eodias.eu/metadata/attribute#geometry> ?geom .
} LIMIT 100
```

CREODIAS


Found 75 results
<< page 1 of 8 >>
title
geom
"S2B_MSIL1C_20180403T104019_N0206_R008_T31UFS_20180403T132754.SAFE"^^<http://www.w3.org/2001/XMLSchema#string>
"POLYGON ((4.435748354912912 51.33997008701231, 4.468636264611875 51.4195692531676, 4.477748734669471 51.441613138005, 6.017025008525815 51.41234265506691, 5.953869713140706 50.42629366061962, 4.408715109637645 50.45527216066188, 4.435748354912912 51.33997008701231))"^^<http://www.opengis.net/ont/geosparql#wktLiteral>
"S2B_MSIL1C_20180423T104019_N0206_R008_T31UES_20180423T110230.SAFE"^^<http://www.w3.org/2001/XMLSchema#string>
"POLYGON ((4.073553537257442 50.45666472371114, 4.091206001706374 50.50070662033345, 4.149881201906933 50.64667849413536, 4.208763333989855 50.79266510091305, 4.268072654378468 50.93859130689168, 4.327401261270531 51.08455377042205, 4.387143604179749 51.23043754970608, 4.446993597869525 51.37627878694308, 4.473671406372481 51.44125427635581, 4.579544021469633 51.44054116582309, 4.546436441462649 50.45352327891901, 4.073553537257442 50.45666472371114))"^^<http://www.opengis.net/ont/geosparql#wktLiteral>

Current work

- We are developing an **extension to schema.org for Earth Observation data (eo.schema.org).**
- Main idea:
  - Use OGC 17-003: EO Dataset Metadata GeoJSON(-LD) Encoding Standard
- We are also developing an **annotation tool** that can be used to annotation Earth Observation datasets using this extension.

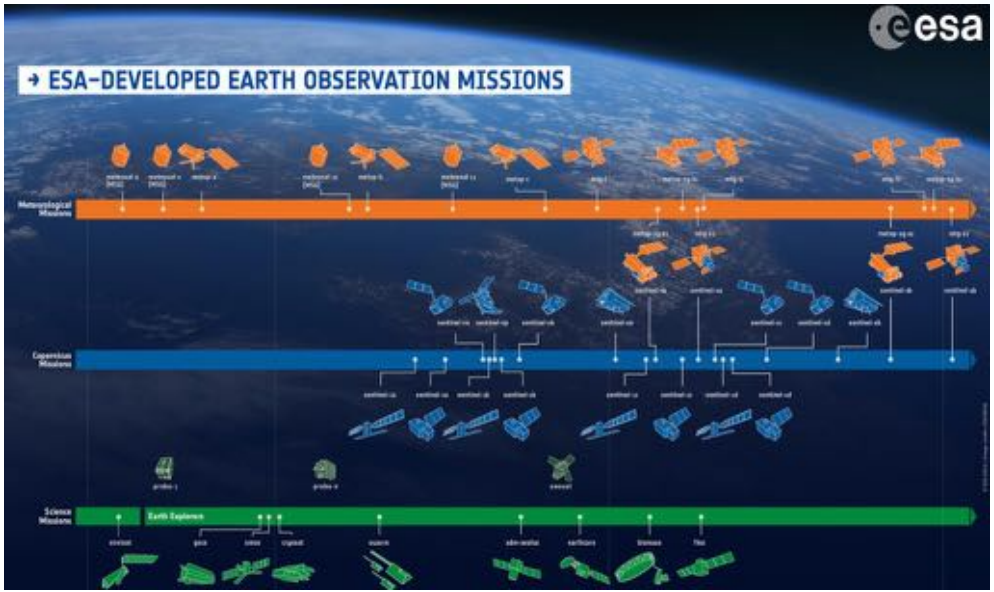
Deep Earth Query: Information Discovery from Big Earth Observation Data Archives

Deep Earth Query: Information Discovery from Big Earth Observation Data Archives

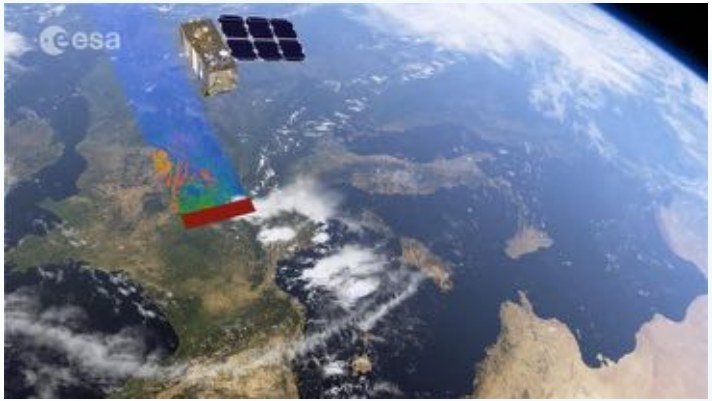
Prof. Dr. Begüm Demir
Big Data Analytics for Earth Observation (BigEarth) Group, BIFOLD
Remote Sensing Image Analysis (RSiM) Group, Faculty of EECS, TU Berlin

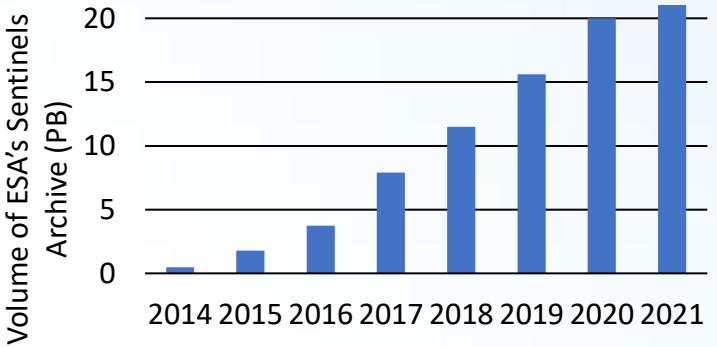
Space Renaissance

- ✓ Recent Earth Observation (EO) satellite missions have led to a significant growth of EO image archives.



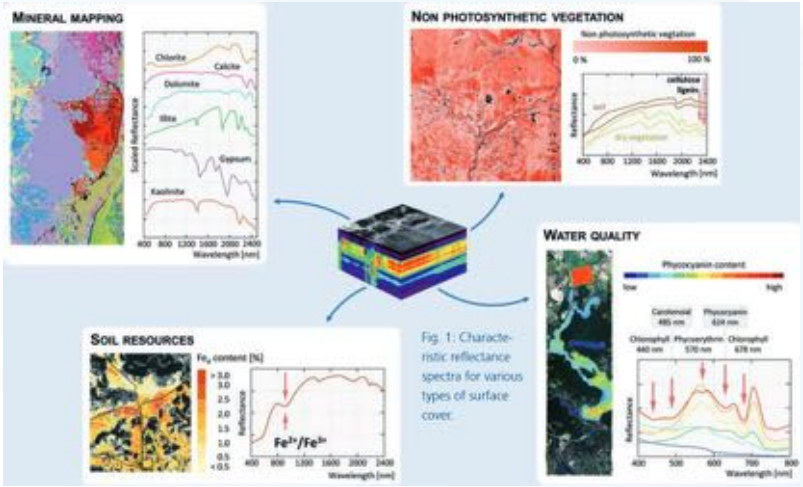
©ESA





EnMAP: Germany's Hyperspectral Satellite for Earth Observation

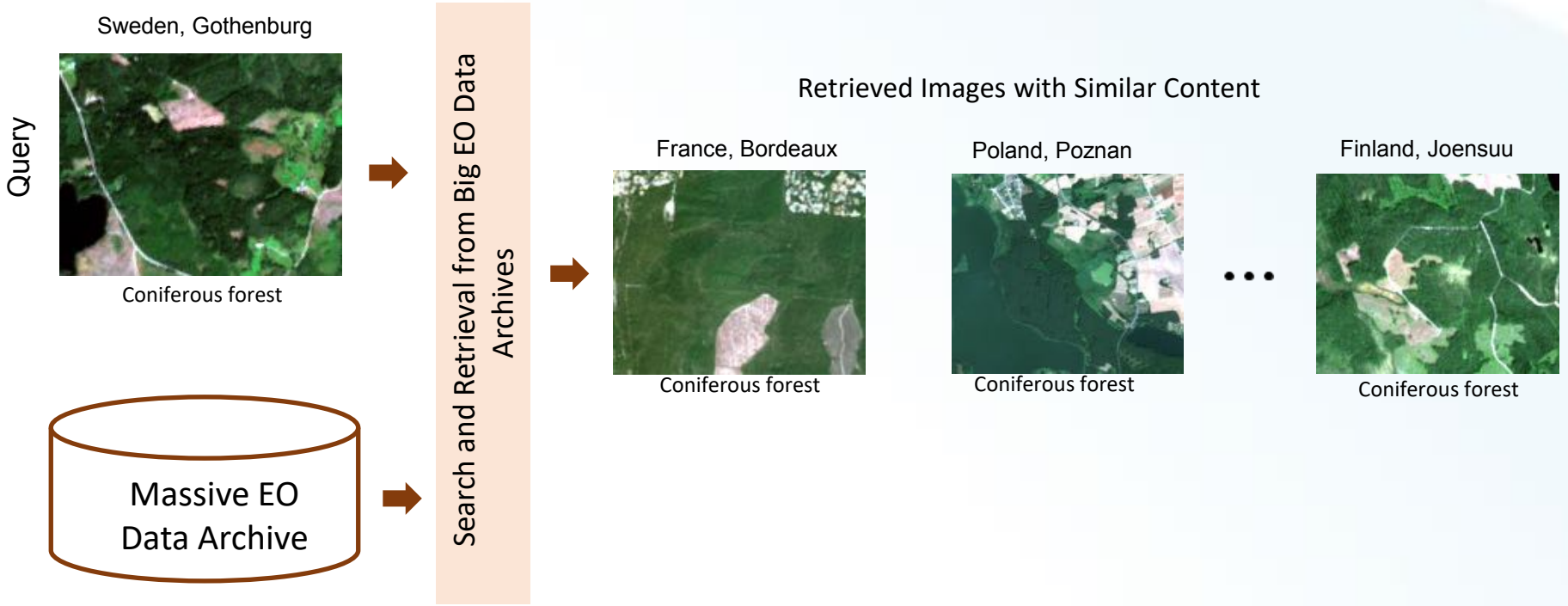
©DLR



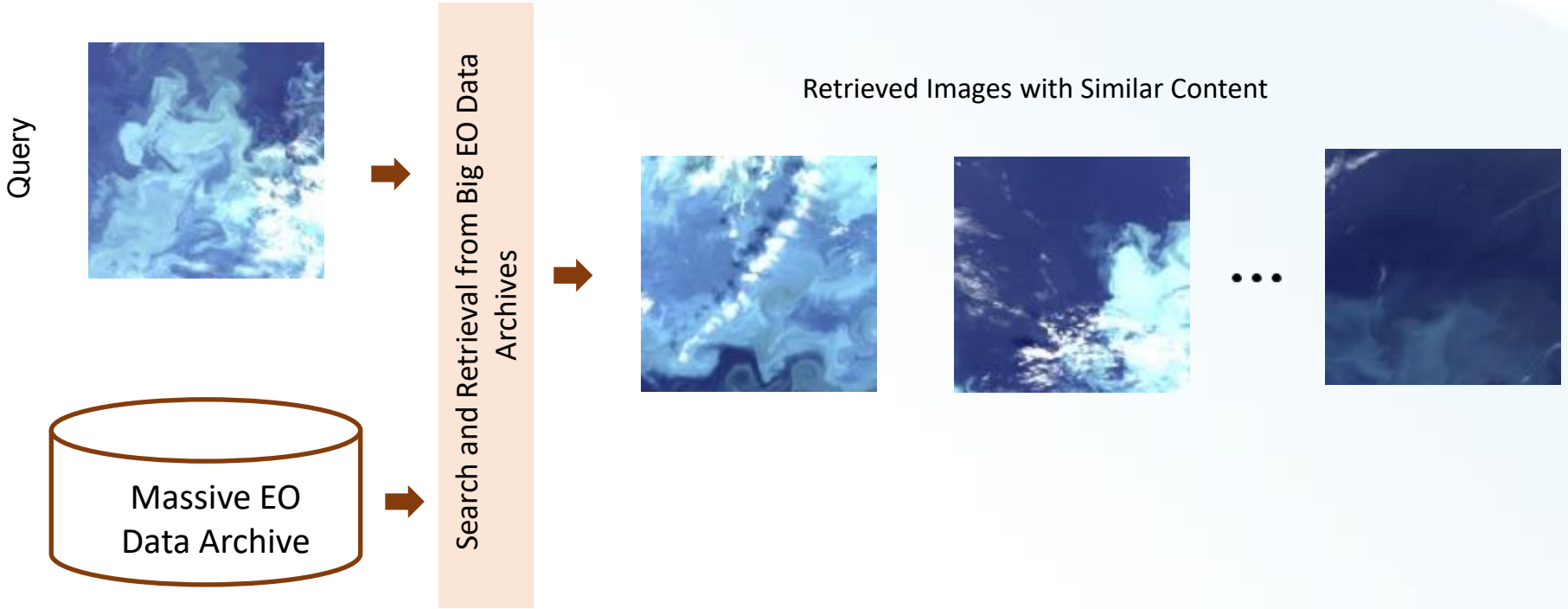
- ✓ #bands: 242
- ✓ spatial resolution: 30m.
- ✓ revisit time: 27 days
- ✓ radiometric resolution: 14 bits

First EnMAP image

Information Discovery by Querying Archives

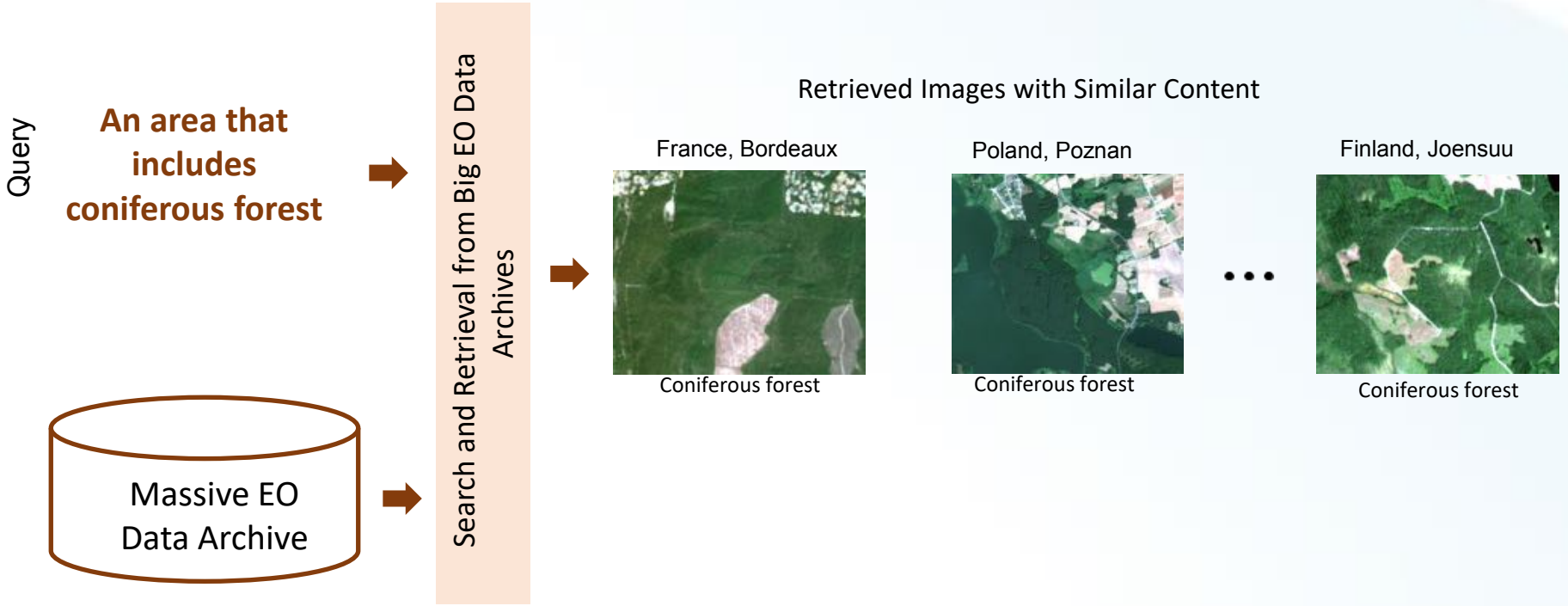


Information Discovery by Querying Archives

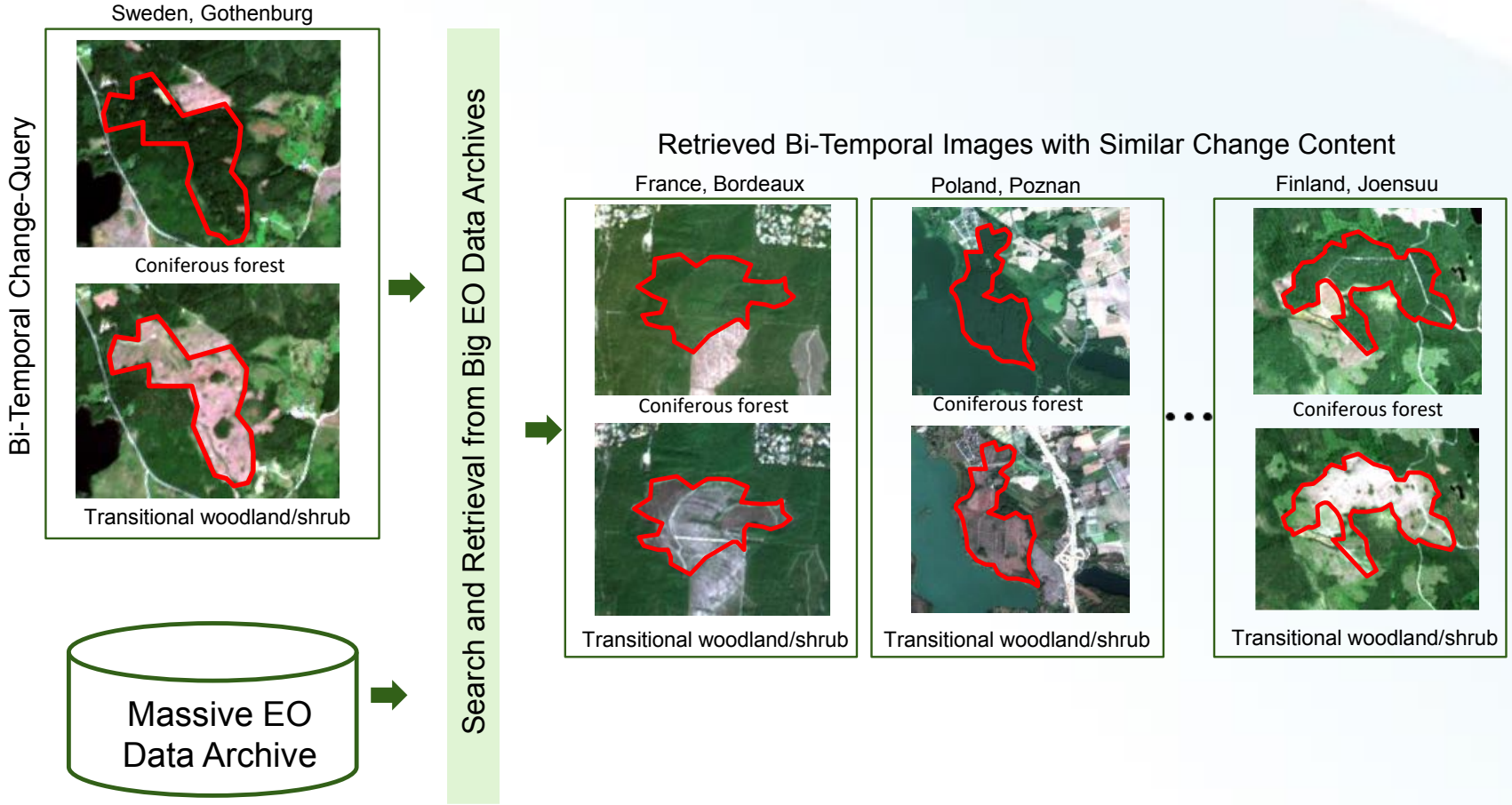


Information Discovery by Querying Archives

Information Discovery by Querying Archives



Information Discovery by Querying Archives



Information Discovery by Querying Archives

Q: Are there any *coastal lagoons* present?
A: No.

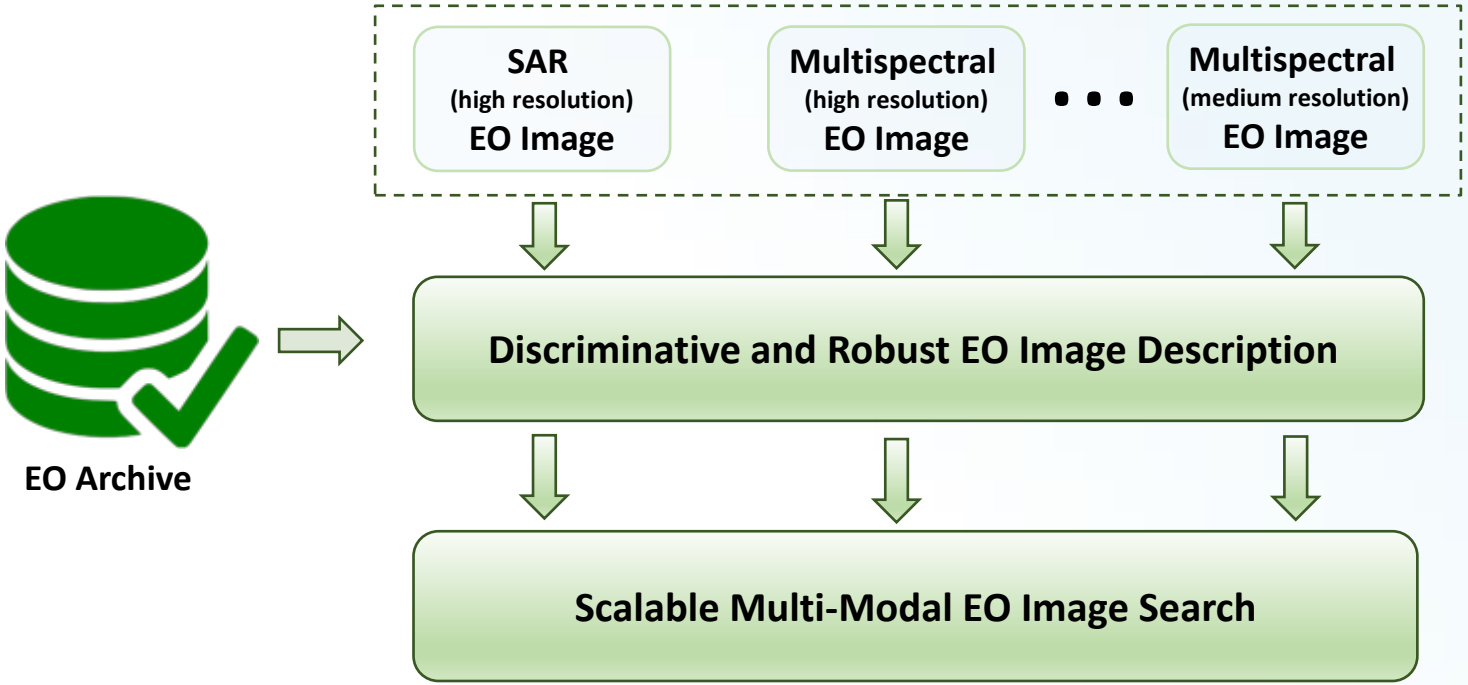
Q: Besides *agricultural areas*, what classes are in the image?
A: Water bodies.

Q: Are there *artificial areas* and *water bodies*?
A: Yes.

Q: Besides *artificial areas* and *water bodies*, what classes are in the image?
A: None.

Visit Our Group Webpage
https://rsim.berlin

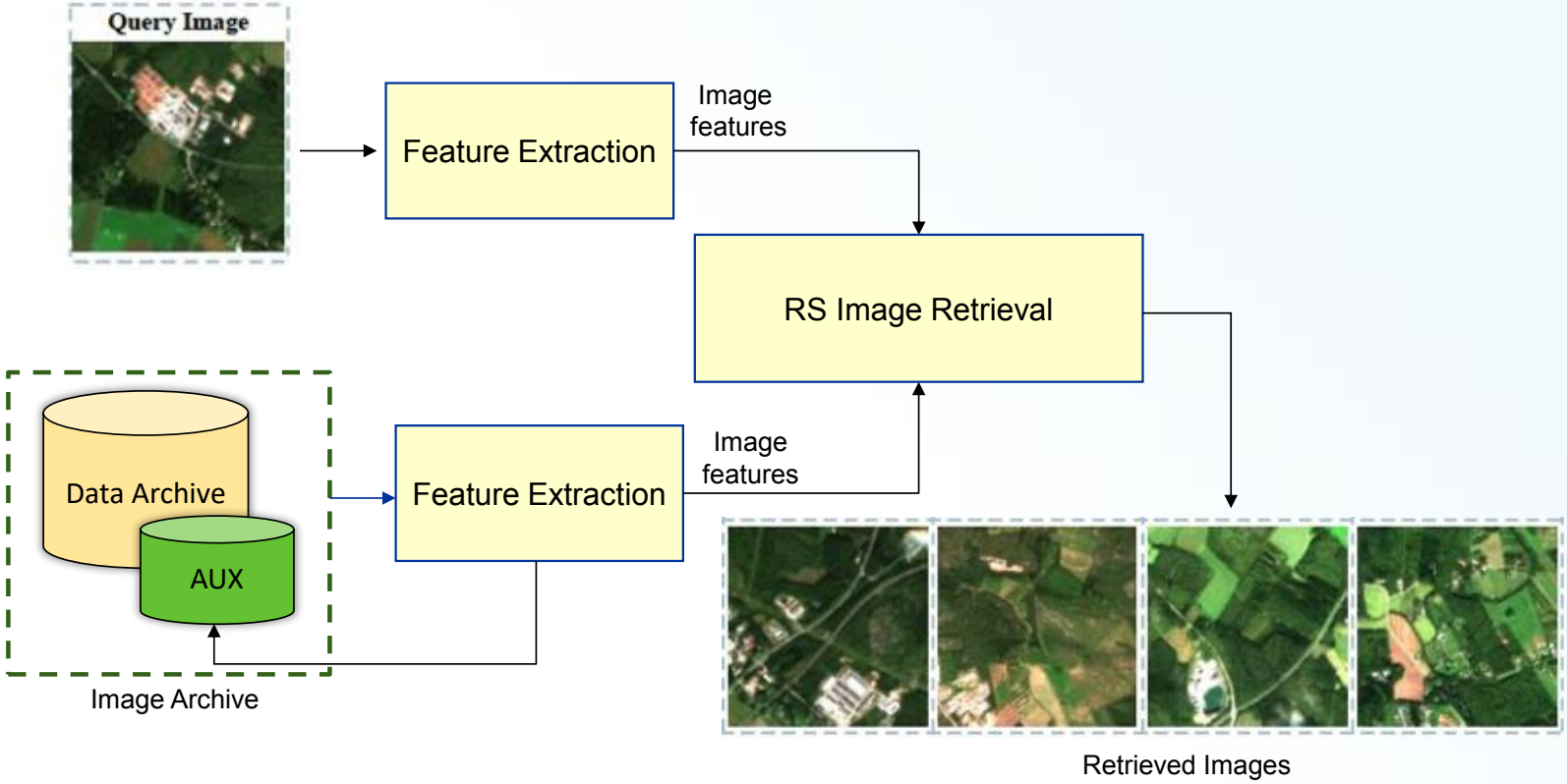
Information Discovery by Querying Archives



G. Sumbul et al. “DL for Image Search and Retrieval in Large Remote Sensing Archives", in "Deep learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science and Geosciences", John Wiley & Sons, 2021.

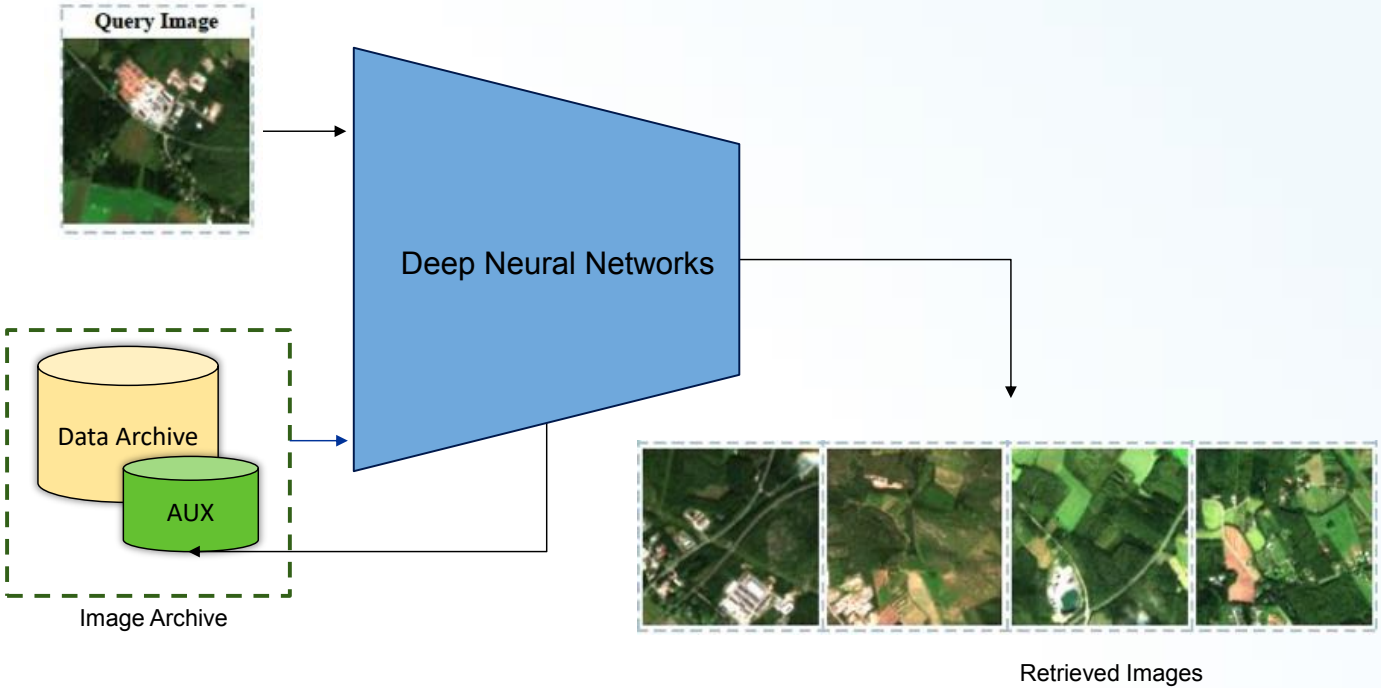
Information Discovery by Querying Archives

Traditional systems

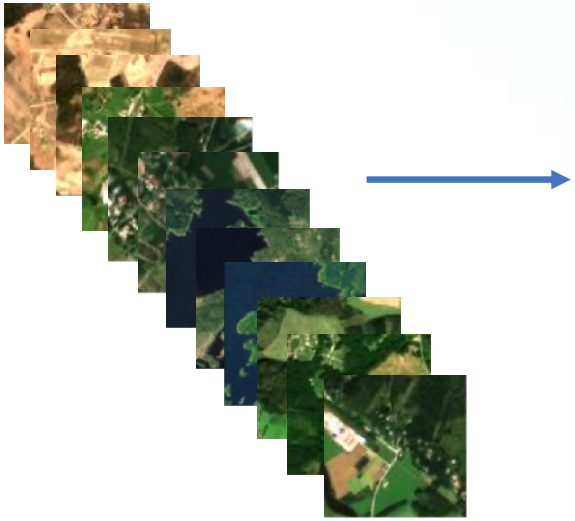


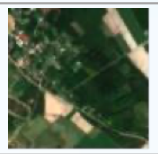
Information Discovery by Querying Archives

Recent systems



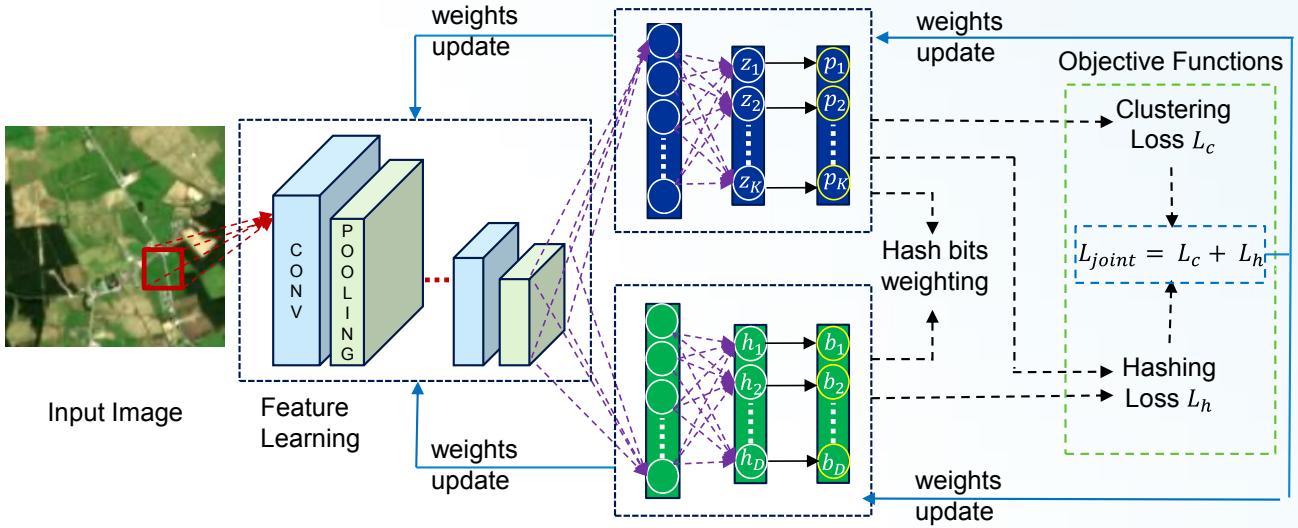
Query by Image through Hashing



Hash Bucket			Code
		...		00
		...		01
		...		10
		...		11

S. Roy, E. Sangineto, B. Demir, N. Sebe, "Metric-Learning based Deep Hashing Network for Content Based Retrieval of Remote Sensing Images", IEEE Geoscience and Remote Sensing Letters, vol. 18, no. 2, pp. 226-230, 2021.

Query by Image through Hashing



Deep class-wise hashing

Graph-based hashing

Zero-shot hashing

Adversarial hashing

Multi-modal hashing

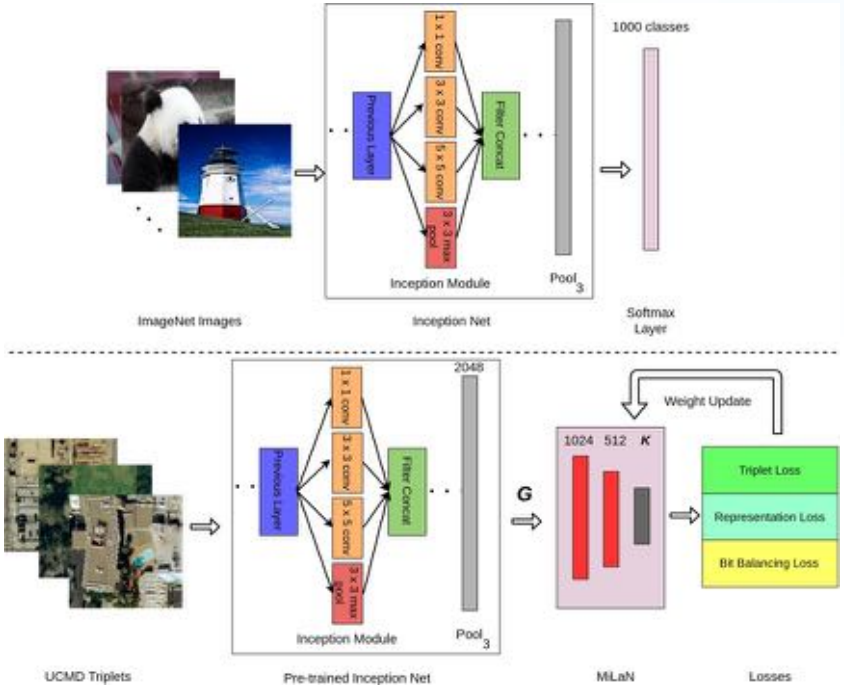
Unsupervised hashing

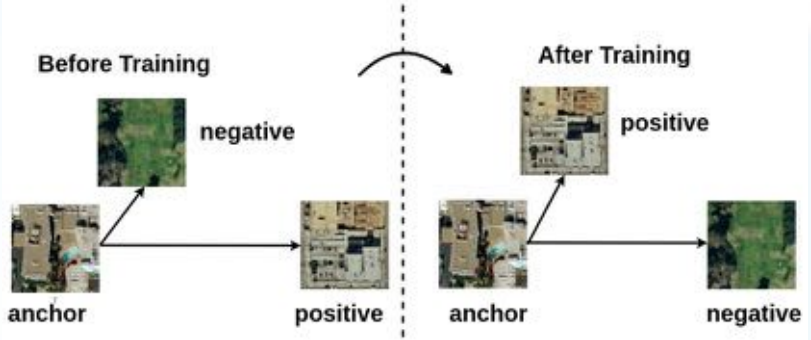
Semantic-preserving hashing

Attention guided hashing

Weakly-supervised hashing

Metric Learning based Deep Hashing (MiLAN)





- ✓ Triplet loss: after training, a positive sample is "moved" closer to the anchor sample than the negative samples;
- ✓ Representation penalty loss: pushes the activations of the last layer of the network to be binary;
- ✓ Bit-balancing loss: encourages the network to produce hash codes having an equivalent number of 0s and 1s.

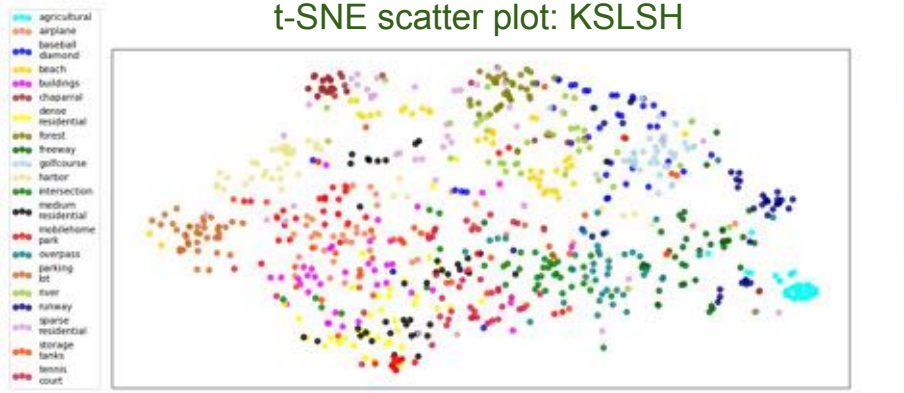
S. Roy, E. Sangineto, B. Demir, N. Sebe, "Deep Metric and Hash-Code Learning for Content-Based Retrieval of Remote Sensing Images", International Geoscience and Remote Sensing Symposium, 2018.

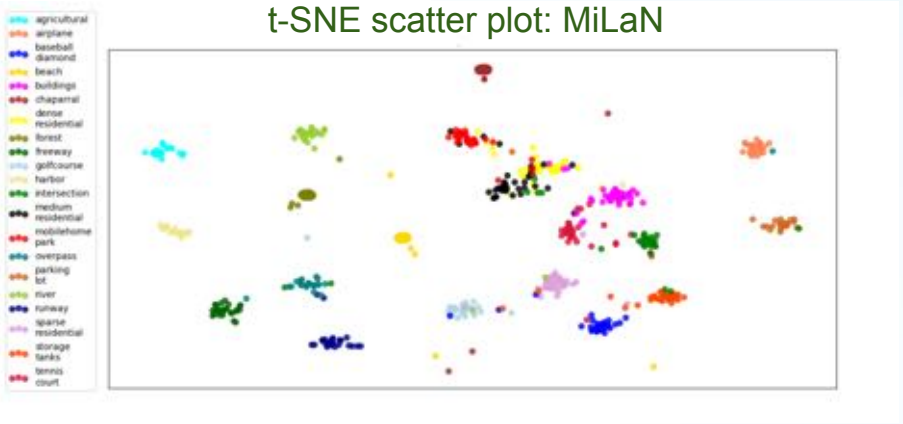
Metric Learning based Deep Hashing (MiLAN)

- ✓ **Data set:** archive that consists of annotated images associated with 21 categories selected from aerial orthophotos with a pixel resolution of 30 cm. Each class includes 100 images that were downloaded from the USGS National Map of several US regions.

Y. Yang, and S. Newsam, “Geographic image retrieval using local invariant features”, IEEE Transactions on Geoscience and Remote Sensing, vol. 51, no. 2, pp. 818-832, Feb. 2013.

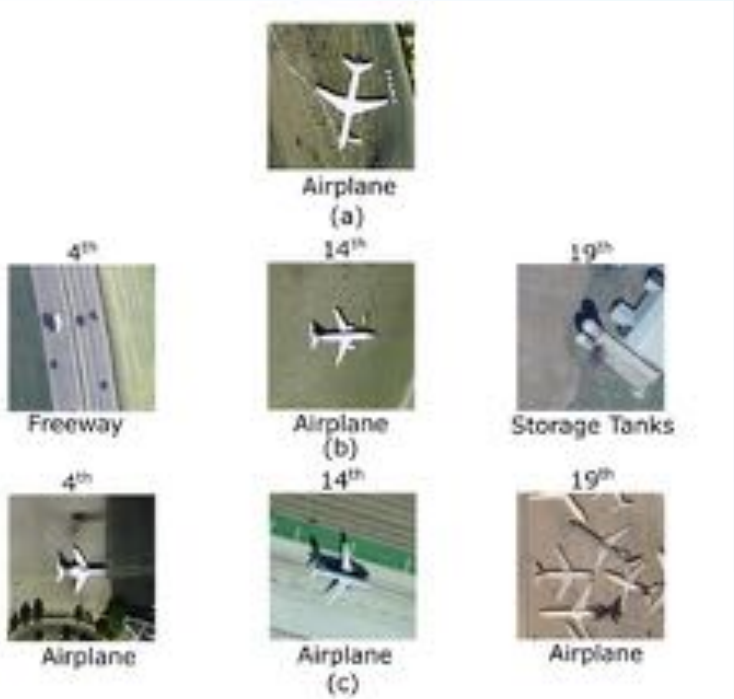
Metric Learning based Deep Hashing (MiLAN)





T-SNE: t-distributed stochastic neighbor embedding

Metric Learning based Deep Hashing (MiLAN)



(a) The query image
(b) Images retrieved by kernel-based hashing
(c) Images retrieved by the MiLaN

On the Way to BigEarthNet

Problem: Differences on the characteristics of images between computer vision and RS.

BigEarthNet: A Benchmark Archive for EO

- ✓ To support the studies on search and retrieval, we developed **BigEarthNet** that:
  - consists of 590,326 Sentinel-1&2 images.
  - opens up promising directions to advance studies for the analysis of large-scale EO data archives.

Impact:

Available also at:

TensorFlow

Radiant MLHub
EARTH IMAGERY FOR IMPACT

BigEarthNet: A Large-Scale Benchmark Archive for Remote Sensing Image Understanding. IEEE IGARSS, Yokohama, Japan, 2019.

BigEarthNet-MM: A Large Scale Multi-Modal Multi-Label Benchmark Archive for Remote Sensing Image Classification and Retrieval. IEEE GRSM, 2021.

BigEarthNet: A Benchmark Archive for EO

- ✓ Each image patch is associated with one or more land-cover class labels provided from the CORINE Land Cover database of the year 2018 (CLC 2018).
- ✓ CLC 2018 has been produced by the European Environment Information and Observation Network of the European Environment Agency.

Urban fabric,
Arable land,
Mixed forest, Land principally occupied by agriculture.

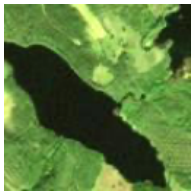
Arable land,
Mixed forest,
Urban fabric.

Urban fabric,
Arable land, Coniferous forest,
Transitional woodland/shrub,
Land principally occupied by agriculture.

Urban fabric,
Arable land,
Land principally occupied by agriculture.

Urban fabric,
Arable land,
Coniferous forest,
Mixed forest, Transitional woodland/shrub.

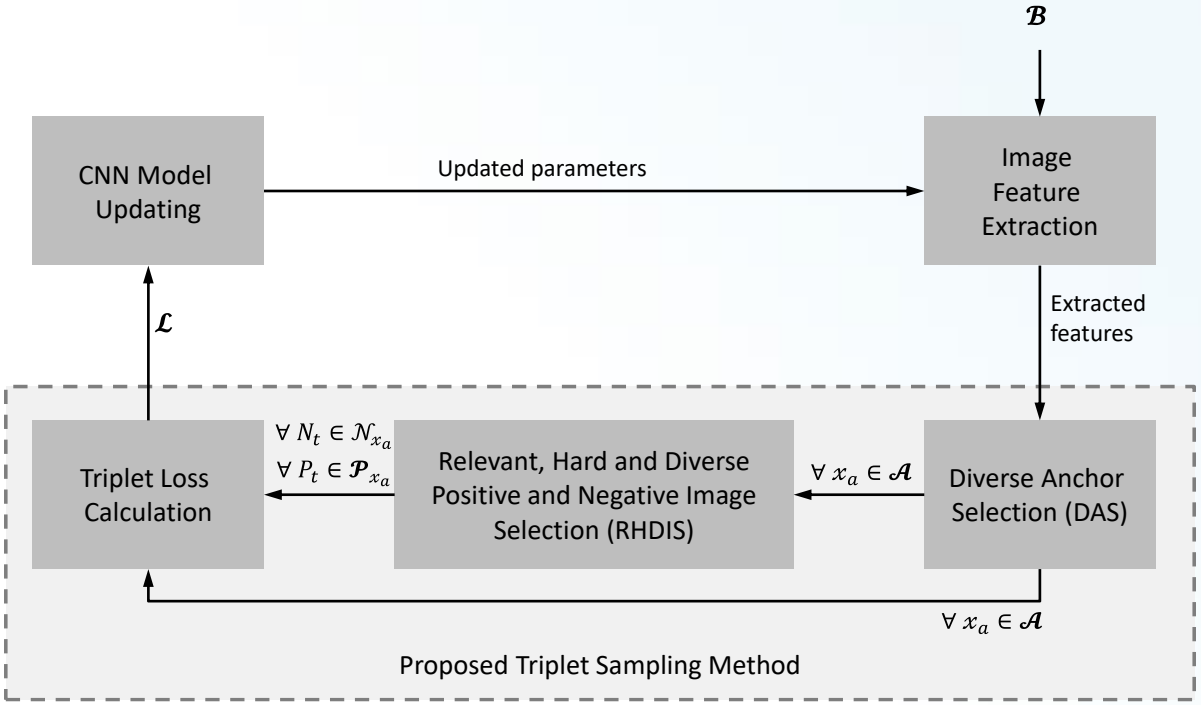
Urban fabric,
Arable land,
Pastures,
Complex cultivation patterns



Coniferous forest,
Mixed forest,
Inland waters,
Transitional woodland/shrub.

Urban fabric,
Arable land,
Coniferous forest,
Mixed forest,
Transitional woodland/shrub.

Informative and Representative Triplet Selection



G. Sumbul, M. Ravanbakhsh, B. Demir, Informative and Representative Triplet Selection for Multi-Label Remote Sensing Image Retrieval, IEEE TGRS, in press, 2022.

Informative and Representative Triplet Selection

- ✓ Diverse anchor selection (DAS) step aims to find a small set of the most representative anchors.
- ✓ An iterative algorithm is introduced to evaluate the diversity in the feature space among the images from a training mini-batch $\mathcal{B}$.
- ✓ At each iteration, it selects a new anchor, which is associated with the highest distance from already selected anchors.

$$x_H = \underset{x_b \in \mathcal{B} \setminus \mathcal{A}}{\operatorname{argmax}} \max_{x_a \in \mathcal{A}} D(x_b, x_a)$$

- ✓ To select relevant, hard and diverse positive-negative image selection, we develop a strategy that evaluates:
  - relevancy based on class label similarity;
  - hardness based on the distance to an anchor;
  - diversity based on the distance among candidate positive/negative images.
- ✓ For each $x_a \in \mathcal{A}$ it initially calculates the informativeness of the images to select the positive and negative image candidates.

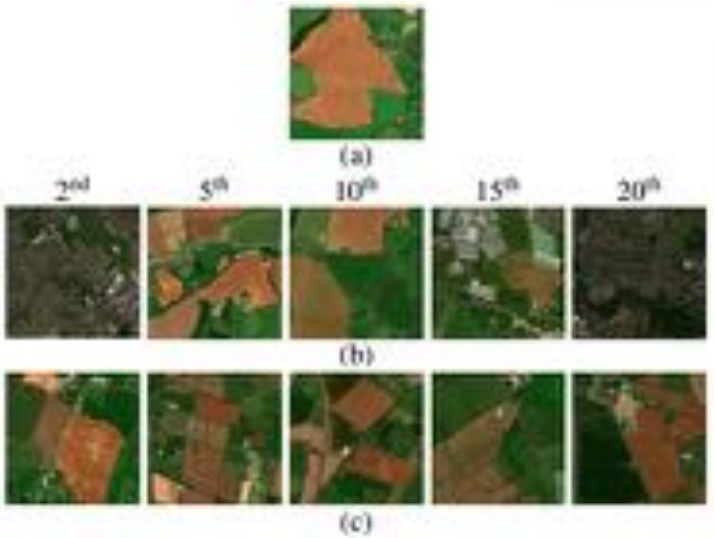
$$I_P(x_a, x_b) = \beta \times S(x_a, x_b) + (1 - \beta) \times D(x_a, x_b)$$

Soft pair-wise similarity

Euclidean distance in feature space

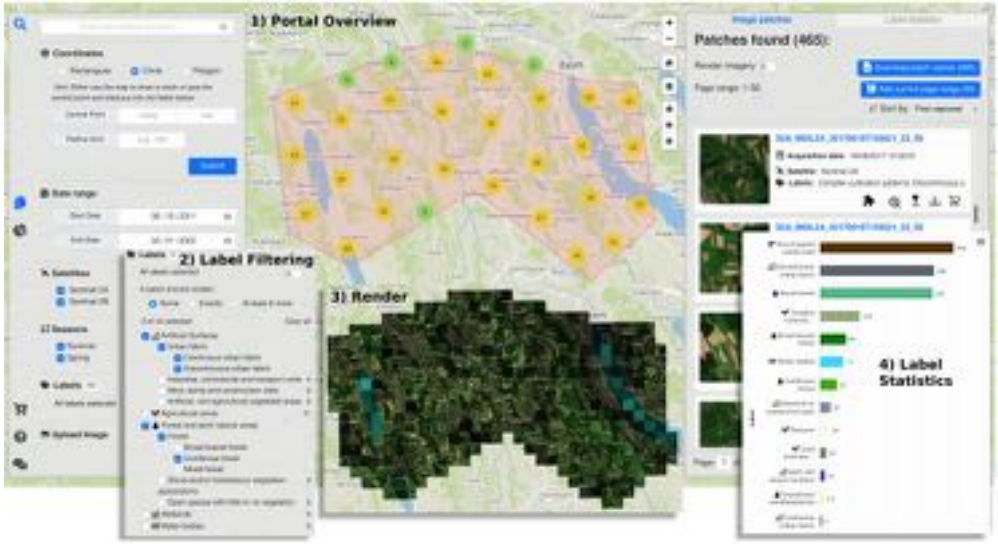
$$I_N(x_a, x_b) = \beta \times [1 - S(x_a, x_b)] + (1 - \beta) \times [1 - D(x_a, x_b)]$$

Results and EarthQube Portal



a) query image; images retrieved by b) standard triplet selection; and c) our triplet sampling approach. Query Classes: arable land, pastures, and complex cultivation patterns.

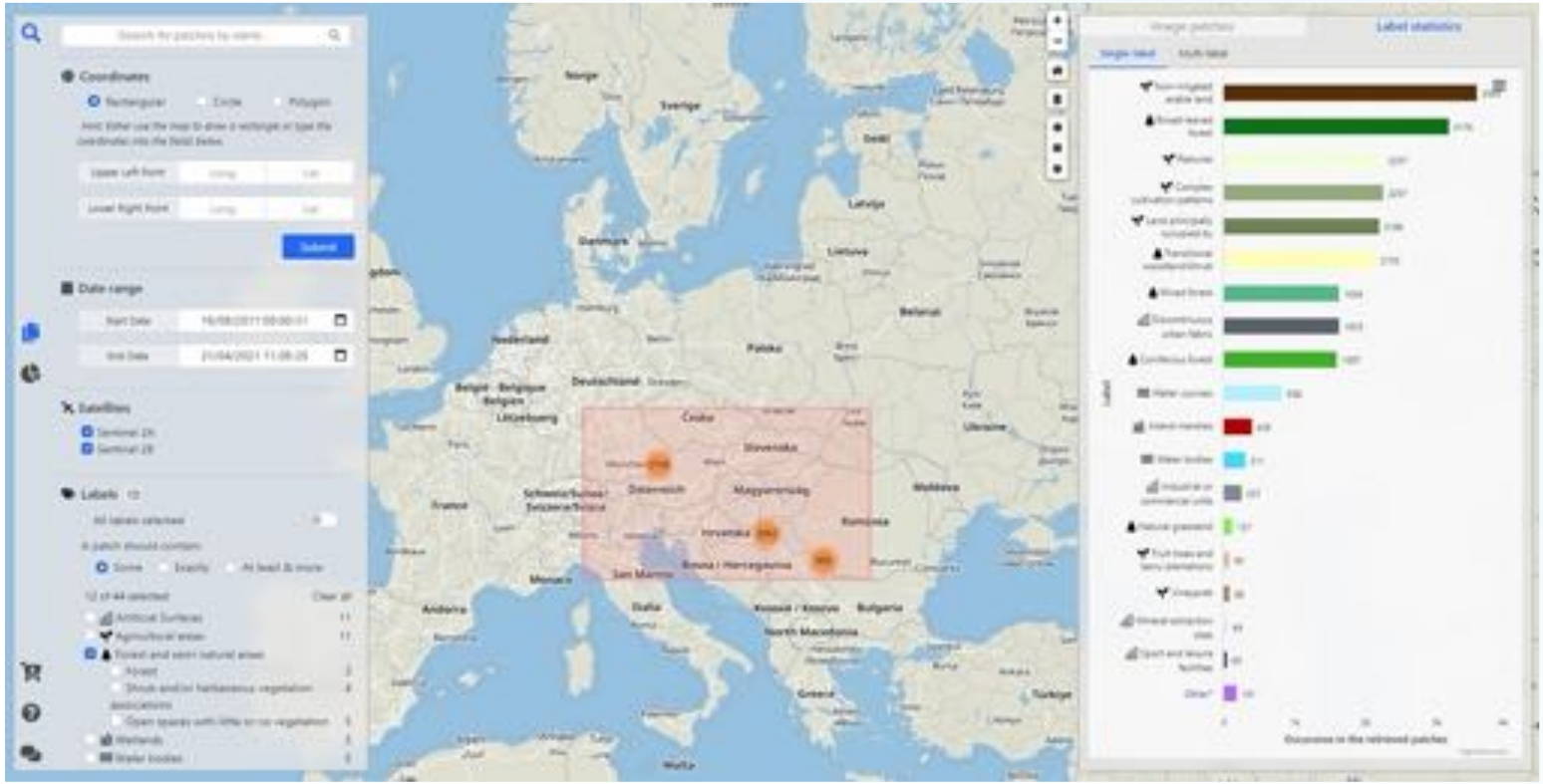
Method	Precision	Recall	F_1 Score
Standard	73.7%	73.8%	73.8%
Our	77.7%	75%	76.7%



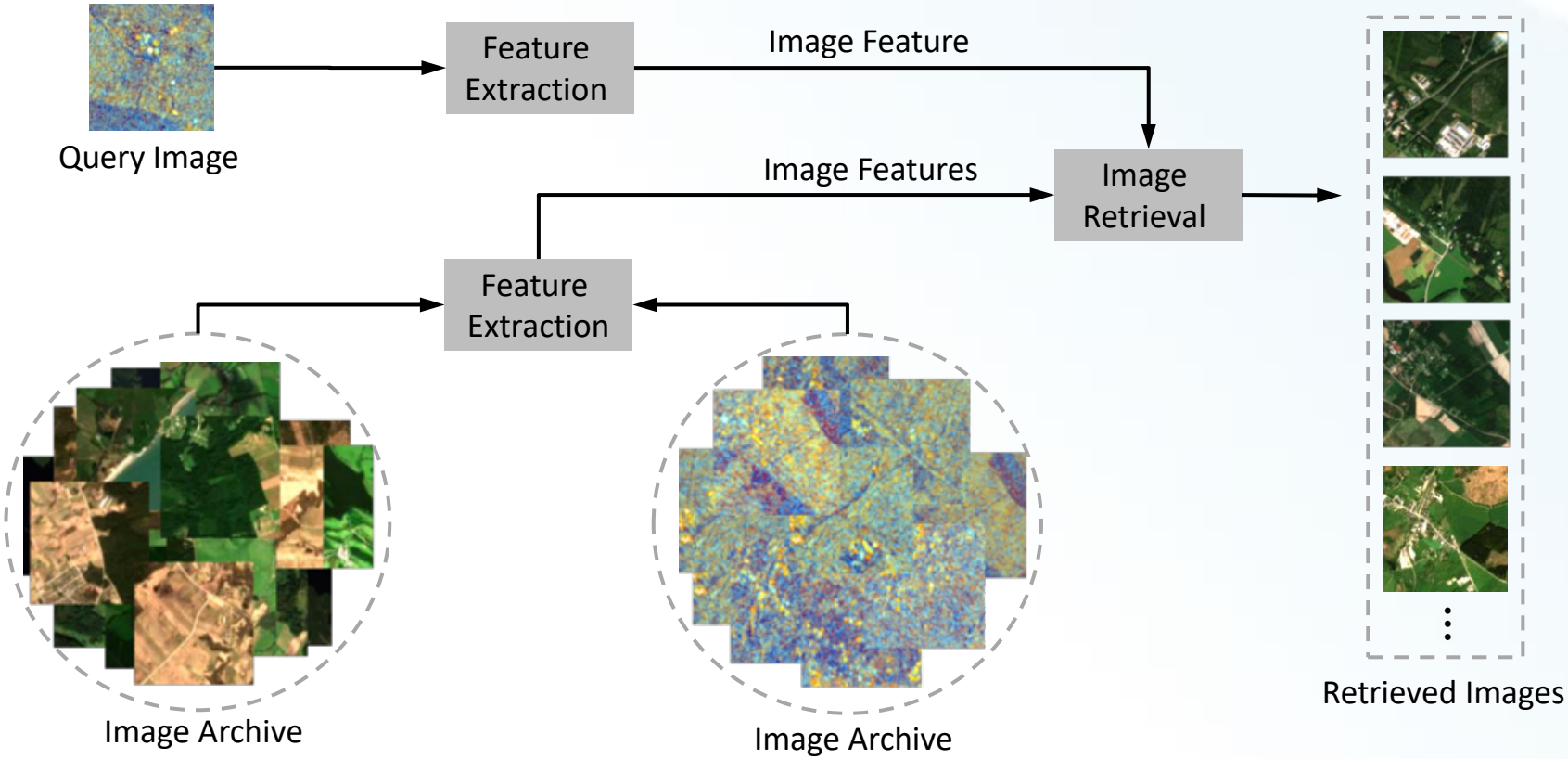
A. K. Aksoy et al., “Satellite Image Search in AgoraEO", International Conference on Very Large Databases, Sydney, Australia, Sept. 2022.

EarthQube Portal

EarthQube Portal

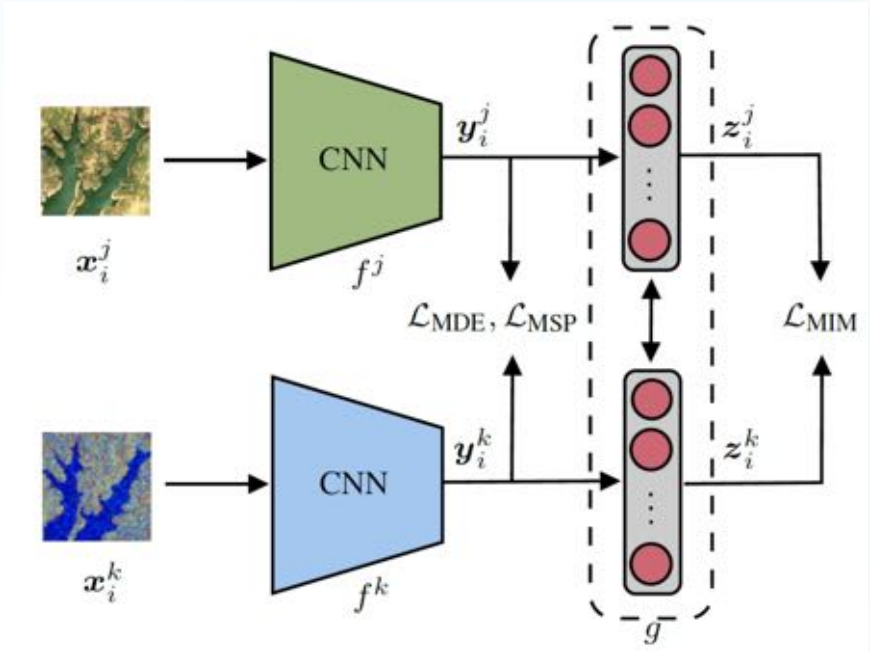


Cross Modal Retrieval



Cross Modal Self-Supervised Retrieval

- ✓ Inter-modal similarity preservation: Mutual information maximization loss function $\mathcal{L}_{MIM}$ is defined based on normalized temperature-scaled cross entropy in a self-supervised manner.
- ✓ Inter-Modal Discrepancy Elimination: $\mathcal{L}_{\mathrm{MDE}}$ is defined aiming to minimize the angular distance of multi-modal image pairs.
- ✓ Intra-Modal Similarity Preservation: $\mathcal{L}_{MSP}$ is defined aiming to maximize the cosine similarity of the most similar image pairs within each modality.



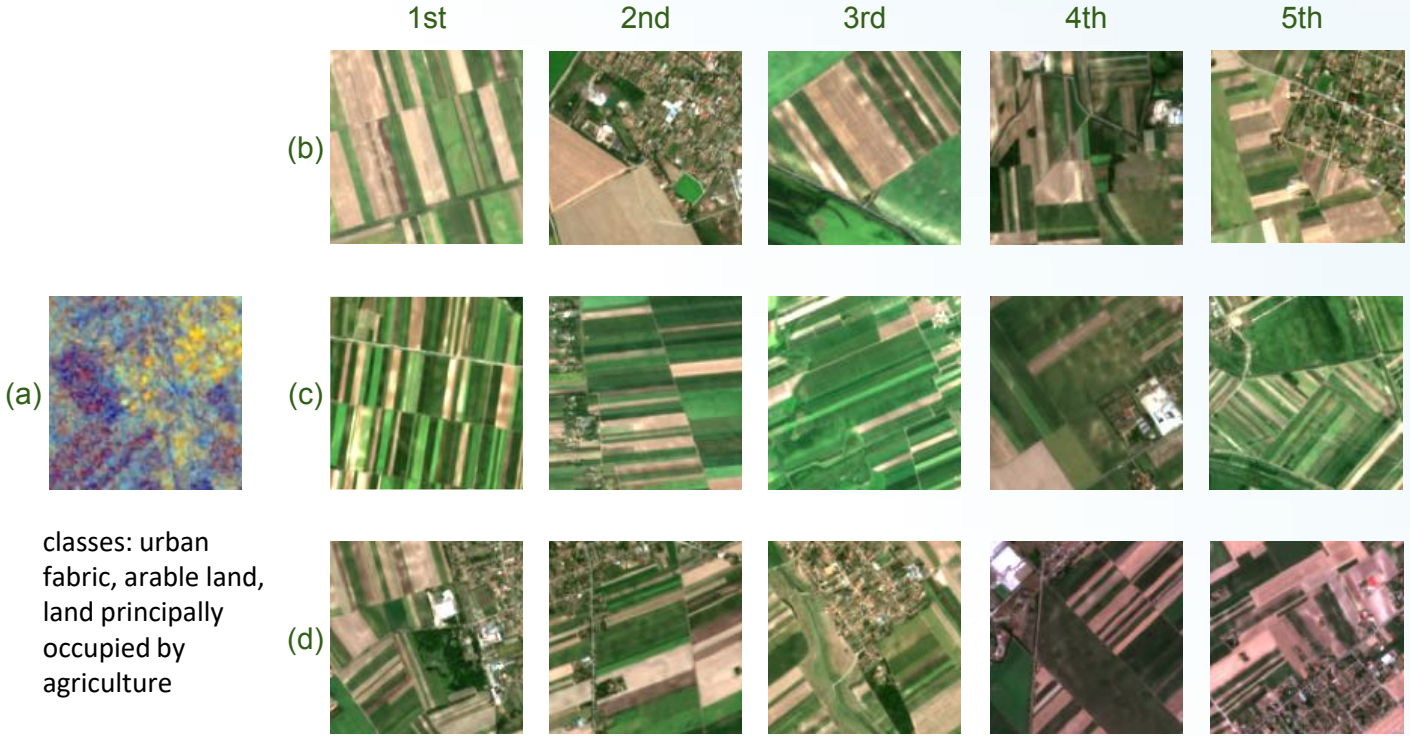
Gencer Sumbul, Markus Müller, Begüm Demir, 'A Novel Self-Supervised Cross-Modal Image Retrieval Method in Remote Sensing', IEEE ICIP, Bordeaux, France, Oct. 2022.

Results: Cross Modal Self-Supervised Retrieval

Method	Requirement of Labels	S1 → S2	S2 → S1	Average
S2MC	✓	41.7	45.6	43.7
deep-SM	✓	65.0	68.7	66.8
DSCMR	✓	<u>73.9</u>	<u>71.3</u>	<u>72.6</u>
DCCA	✗	49.1	40.3	44.7
SimCLR	✗	47.3	53.5	50.4
DUCH	✗	66.6	67.8	67.2
Ours	✗	**71.5**	**71.3**	**71.4**

[1] Li et al, “Semantically supervised maximal correlation for cross-modal retrieval,” in IEEE ICIP, pp. 2291–2295, 2020.
[2] Wei et al. “Cross-modal retrieval with CNN visual features: A new baseline,” IEEE Transactions on Cybernetics, vol. 47, no. 2, 2017.
[3] Zhen et al, “Deep supervised cross-modal retrieval,” in IEEE CVPR, pp. 10386–10395, 2019.
[4] Andrew et al. “Deep canonical correlation analysis,” ICML, vol. 28, no. 3, pp. 1247–1255, 2013.
[5] Chen et al, “A simple framework for contrastive learning of visual representations”, ICML, pp. 1597–1607, 2020.
[6] G. Mikriukov, M. Ravanbakhsh, and B. Demir, “Unsupervised contrastive hashing for cross-modal retrieval in remote sensing,” IEEE ICASSP, 2022.

Results: Cross Modal Self-Supervised Retrieval

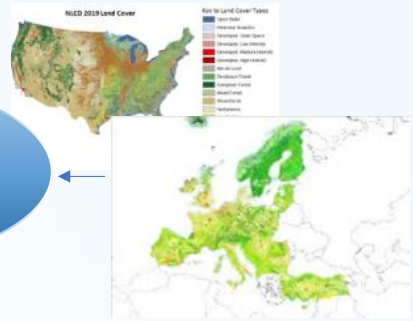


(a) S1 query image; and S2 images retrieved by (b) self-supervised DUCH; (c) fully supervised DSCMR; and (d) our method.

Outlook

- ✓ Hashing-based methods are promising for RS retrieval problems due to their capability on fast and scalable image search and retrieval.
- ✓ Development of methods and tools is needed for global-scale scalable RS CBIR with high accuracy and zero-annotation cost.
- ✓ Thematic maps (which may contain noisy labels) can be used. Methods that are robust to the label noise are required.

Discontinues urban fabric, Complex cultivation patterns, non-irrigated arable land



Noise-Robust DL/ML Models

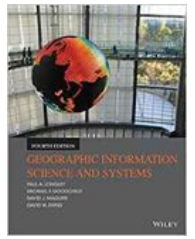
Correct caption:
A red church with a white cross in top is near a river with boats.
Noisy caption (typos):
A rad churhc with a white cros in top is neer a rver with boatss.
Noisy caption (wrong caption):
A rectangular playground and many tall buildings around.

Discontinuous urban fabric
Coniferous forest
Mixed forest
Industrial or commercial units
missing label

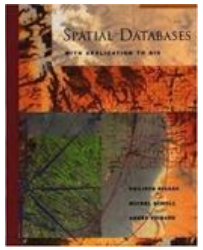
Discontinuous urban fabric
Industrial or commercial units
Non-irrigated arable land
Coniferous forest
wrong label

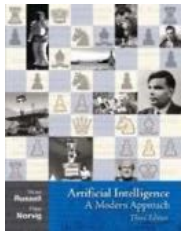
RDF and SPARQL extensions for geospatial data

Background



- Geographic information systems
- Spatial database research
- Spatial logics and reasoning
- Industry standards and implemented systems





OGC®
Making location count.
www.opengeospatial.org

ArcGIS ESRI

QGIS

PostGIS

Overview

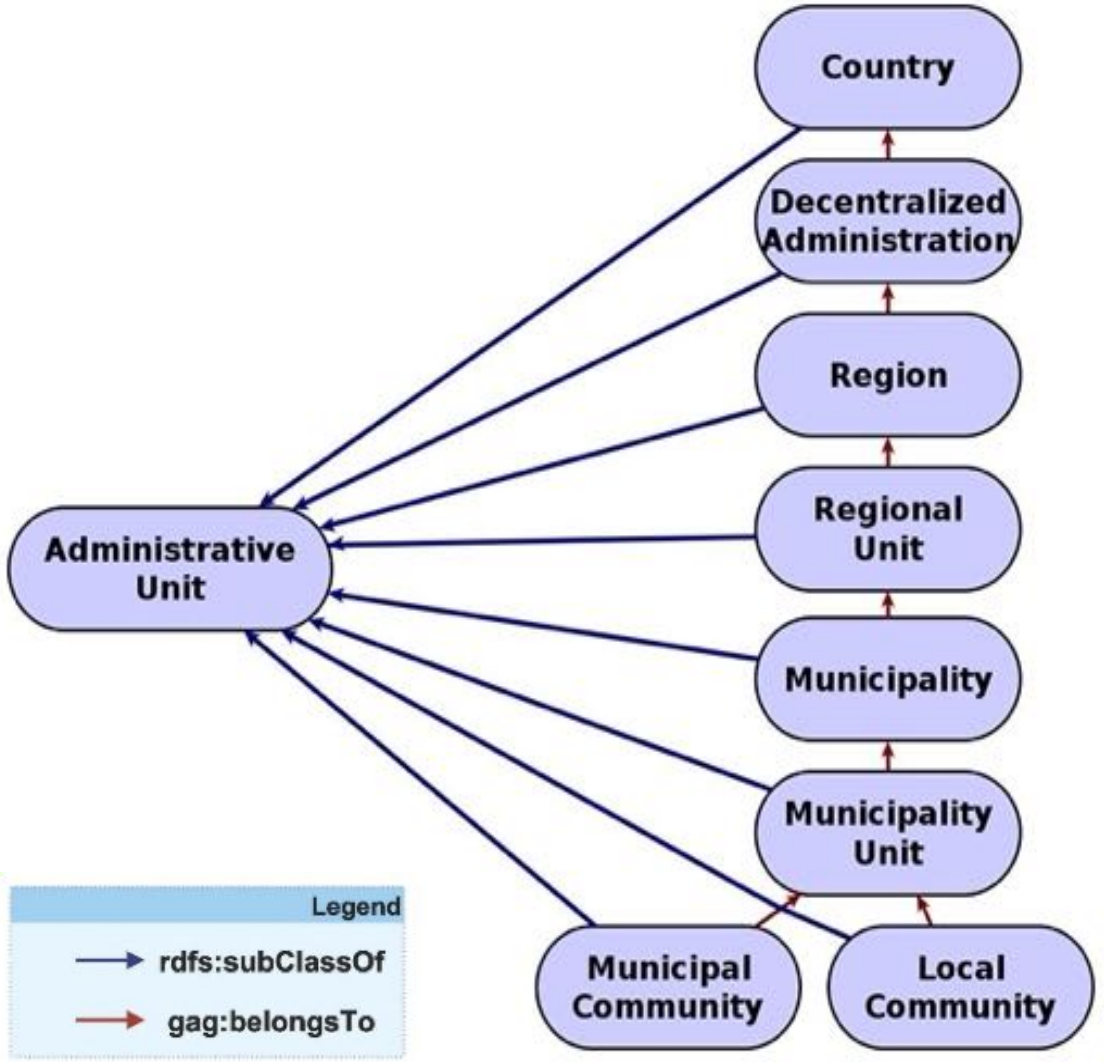
- The data model stRDF/stSPARQL (2012)
- The OGC standard GeoSPARQL (2012)
- The framework RDFi (2013)
- GeoSPARQL+ (2020)
- The proposed language GeoSPARQL 1.1 (2019-2022)

The Model stRDF

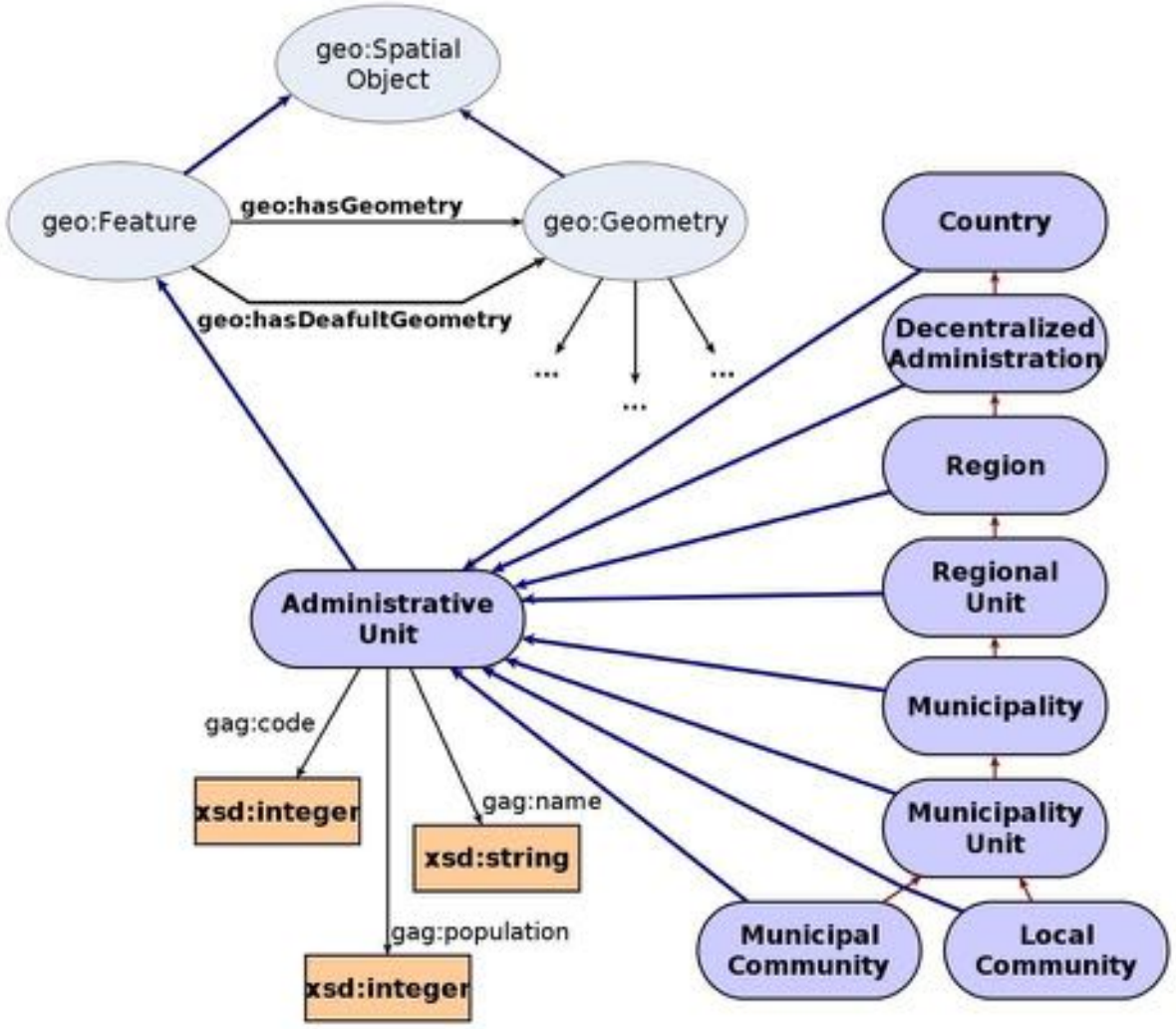
- An extension of RDF for the representation of **geospatial information that changes over time**.
- Geospatial dimension:
  - Two **spatial data type**s are introduced.
  - Geospatial information is represented using **spatial literals** of these datatypes.
  - OGC standards **WKT** and **GML** are used for the serialization of spatial literals.
- **Temporal dimension** (not covered in this tutorial)

Example: Greek Administrative Geography

Domain Ontology

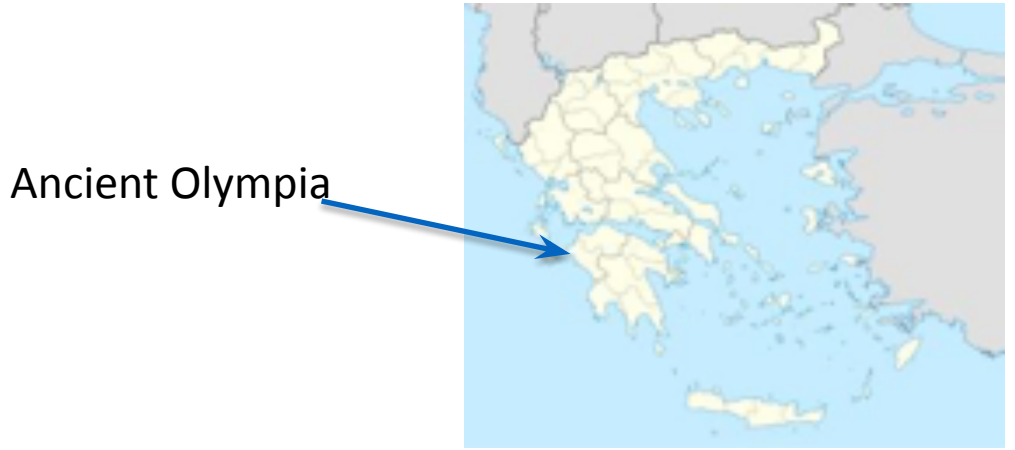


Connect to a Top Level Geospatial Ontology + Introduce Some Properties



Example of stRDF (geospatial dimension)

```
gag:Olympia
   rdf:type gag:MunicipalCommunity;
   gag:hasName "Ancient Olympia";
   gag:hasPopulation "184"^^xsd:int;
   strdf:hasGeometry "MULTIPOLYGON(((308511.906249999
   4201042,308763.8125 4200714,
   308840.09375 4200629,308939.3125 4200545,.....308390.000000001
   4201276,308451.593749999 4201167,308467 4201125,308511.906249999
   4201042)));<http://www.opengis.net/def/crs/EPSG/0/2100>"^^strdf:WKT.
```



The Query Language stSPARQL (geospatial dimension)

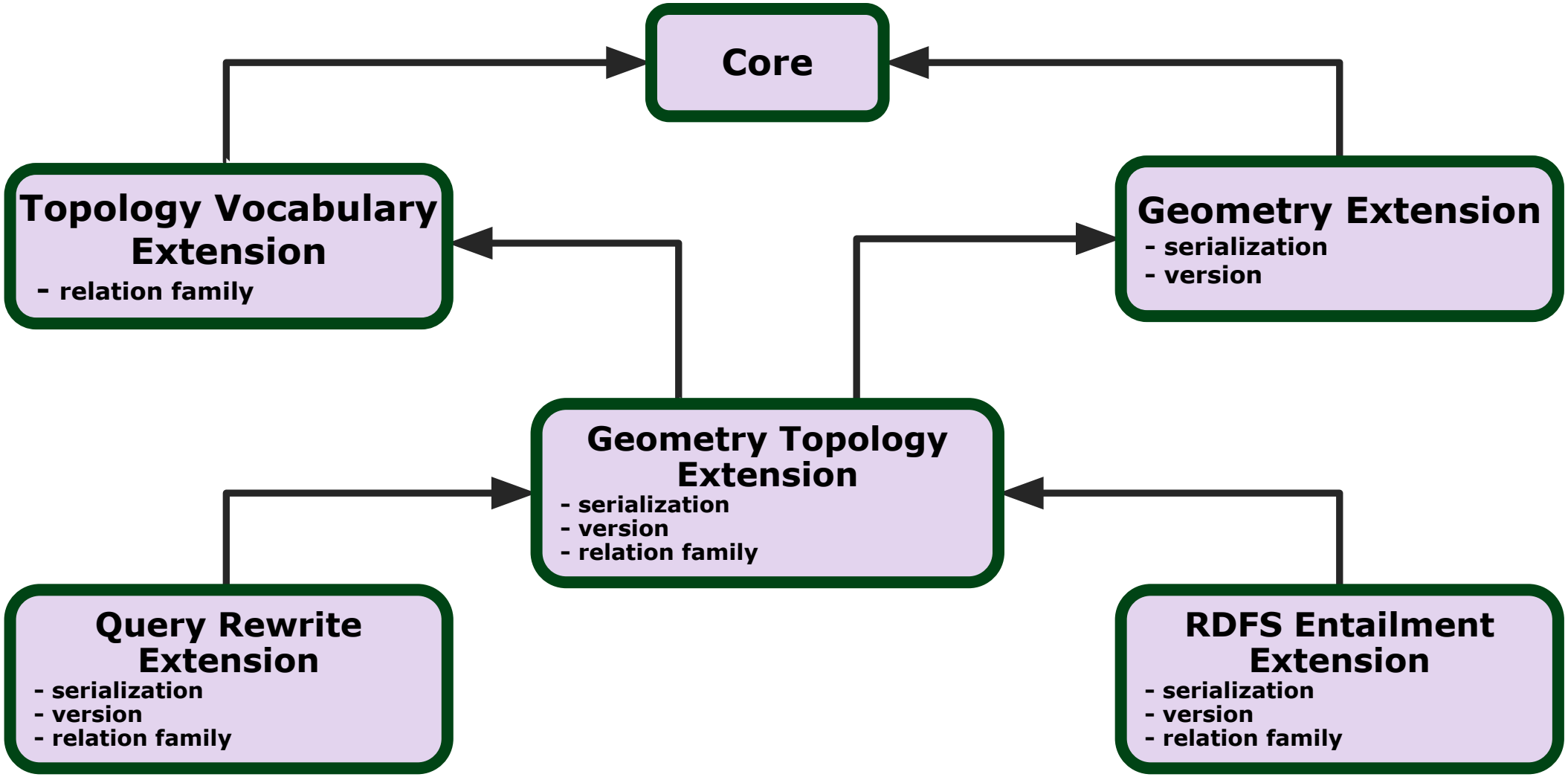
- It is an **extension of SPARQL 1.1**
- It offers **families of functions for querying geometries**.
- The functions are taken mostly from the **OGC standard** “OpenGIS Simple Feature Access - Part 2: SQL Option”.
- They are similar to the ones offered by **spatially-enabled relational database management systems** (e.g., PostGIS).

Example of stSPARQL (geospatial dimension)

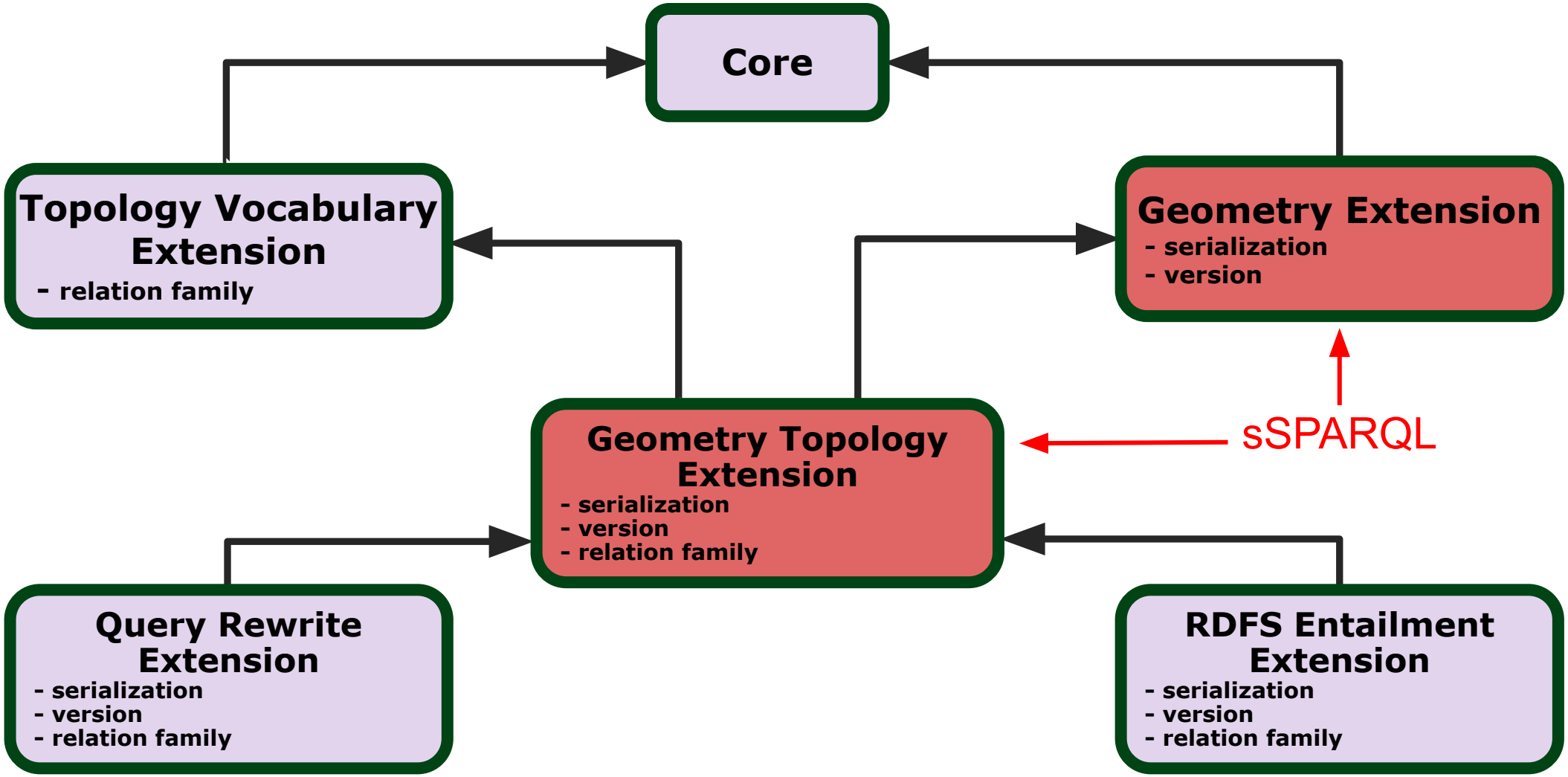
Query: Compute the parts of burnt areas that lie in coniferous forests


```
SELECT ?burntArea (strdf:intersection(?baGeom,
                   strdf:union(?fGeom)) AS ?burntForest)
WHERE {
       ?burntArea  rdf:type noa:BurntArea;
                   strdf:hasGeometry ?baGeom.
        ?forest    rdf:type clc:Region;
                  clc:hasLandCover clc:ConiferousForest;
                   strdf:hasGeometry ?fGeom.

FILTER  (strdf:intersects(?baGeom,?fGeom)) }
GROUP BY ?burntArea ?baGeom
```

The OGC Standard GeoSPARQL (2012)



GeoSPARQL vs. stSPARQL



Example of the Topology Vocabulary Extension

Triples:

```
gag:Olympia rdf:type gag:MunicipalCommunity .
gag:OlympiaMunicipality rdf:type gag:Municipality .
gag:WesternGreece rdf:type gag:Region .

gag:Olympia   geo:sfWithin  gag:OlympiaMunicipality .
gag:OlympiaMunicipality   geo:sfWithin   gag:WesternGreece .
```

Query: Find the **region** in which Ancient Olympia is located.

Answer: `gag:WesternGreece`

Example of the Topology Vocabulary Extension

Triples:

```
gag:Olympia rdf:type gag:MunicipalCommunity .
gag:OlympiaMunicipality rdf:type gag:Municipality .
gag:WesternGreece rdf:type gag:Region .

gag:Olympia   geo:sfWithin  gag:OlympiaMunicipality .
gag:OlympiaMunicipality   geo:sfWithin   gag:WesternGreece .
```

Query: Find the **region** in which Ancient Olympia is located.

Answer: `gag:WesternGreece`

Method: By transitivity of geo:sfWithin.

Not supported by GeoSPARQL!

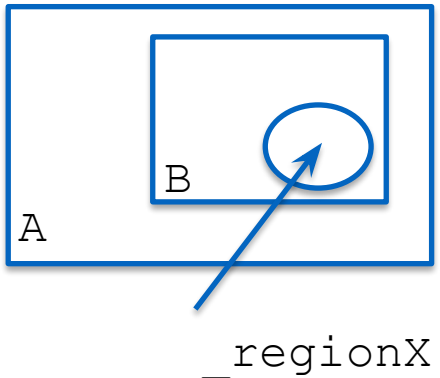
The Query Rewrite Extension

- Enables the **translation of qualitative topological information appearing in a query to quantitative.**
- This is done by rewriting of queries with triple patterns involving topological relations into queries with topological functions on geometries.
- The rewriting is based on **RIF rules.**

Beyond the Topology Vocabulary Extension

Triples:

```
ex:regionA strdf:hasGeometry "POLYGON(A)"^^strdf:WKT .
ex:regionB strdf:hasGeometry "POLYGON(B)"^^strdf:WKT .

_:regionX geo:sfWithin ex:regionB
```

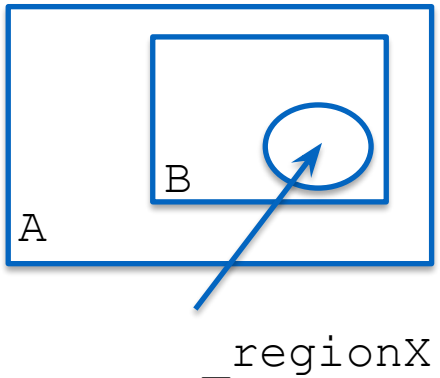


Query: Is regionX contained in regionA?

Beyond the Topology Vocabulary Extension (cont'd)

Triples:

```
ex:regionA strdf:hasGeometry "POLYGON(A)"^^strdf:WKT .
ex:regionB strdf:hasGeometry "POLYGON(B)"^^strdf:WKT .

_:regionX geo:sfWithin ex:regionB
```



Query: Is regionX contained in regionA?

Answer: **Yes**

Not supported by GeoSPARQL.

The Framework RDFi (RR2013, AIJ 2016)

- Extension of RDF with incomplete information.
- New kind of literals (**e-literals**) for each datatype.
  - Property values that exist but are unknown or partially known.

- **Partial knowledge**: captured by constraints (appropriate constraint language *L)*.

- RDF graphs extended to **RDF^i databases**: pair (G, φ)

  G: RDF graph with e-literals

  φ: quantifier-free formula of *L*

The Framework RDFi (cont'd)

- **Formal semantics** for RDF^{i} and SPARQL query evaluation.
- **Representation Systems**:
  - CONSTRUCT queries
    - Without blank nodes in their templates
    - With monotone graph patterns (using only operators AND, UNION and FILTER).
  - CONSTRUCT queries
    - Without blank nodes in their templates
    - With well-designed graph patterns (graph patterns using only AND, FILTER and OPT plus some intuitive conditions).
- **Computational Complexity**:
  - Query answering is **coNP-complete (data complexity)** for certainty queries and various interesting classes of spatial constraints. Compare this with LOGSPACE complexity for the standard SPARQL case.

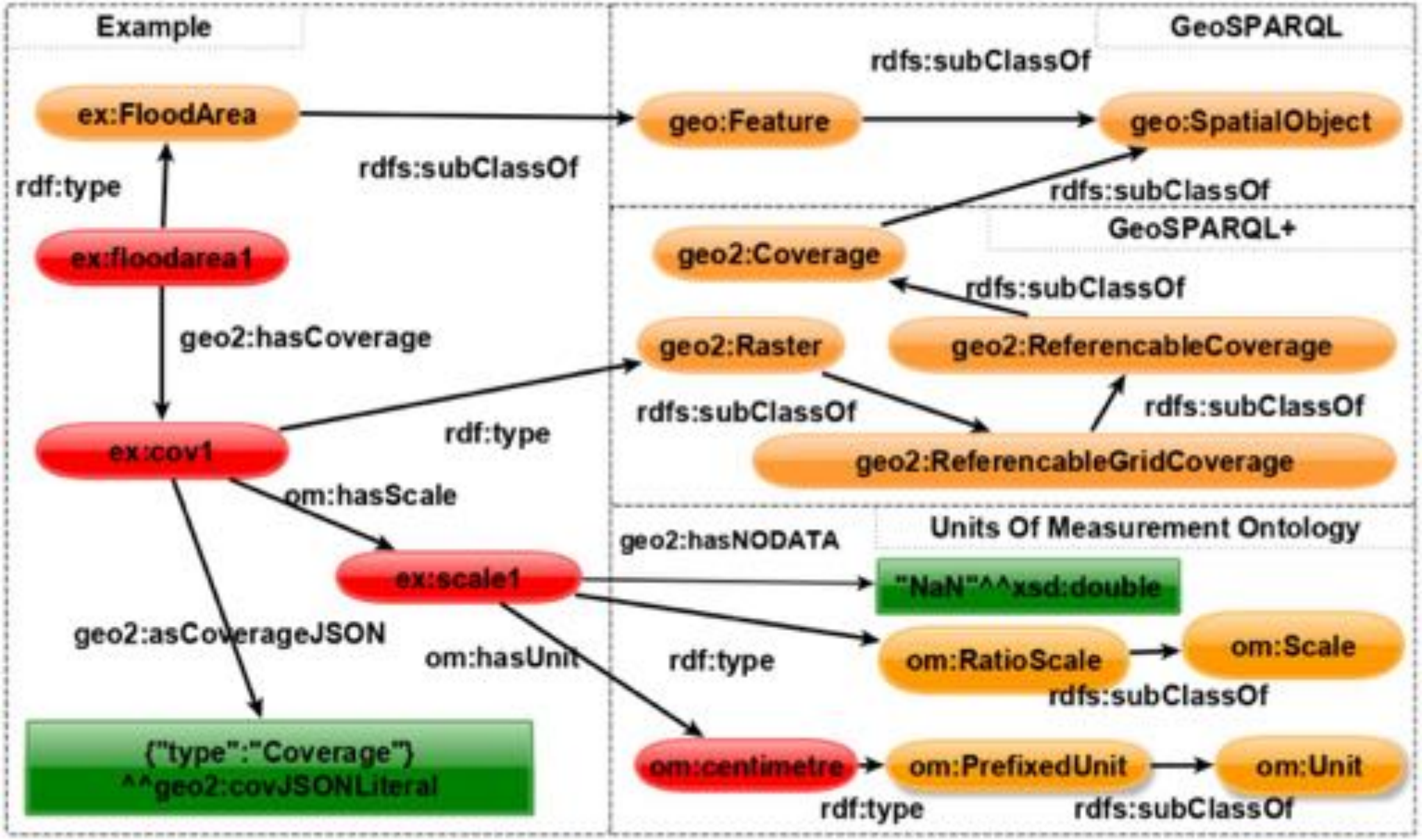
GeoSPARQL+ (Homburg et al. ISWC 2020)

- GeoSPARQL+ ontology in order to integrate geospatial **raster data** into the Semantic Web
  - New type of geospatial data for raster

- Extension of GeoSPARQL query language
  - **New filter functions** based on raster algebra operations, e.g. **rasterPlus**, **rasterSmaller**

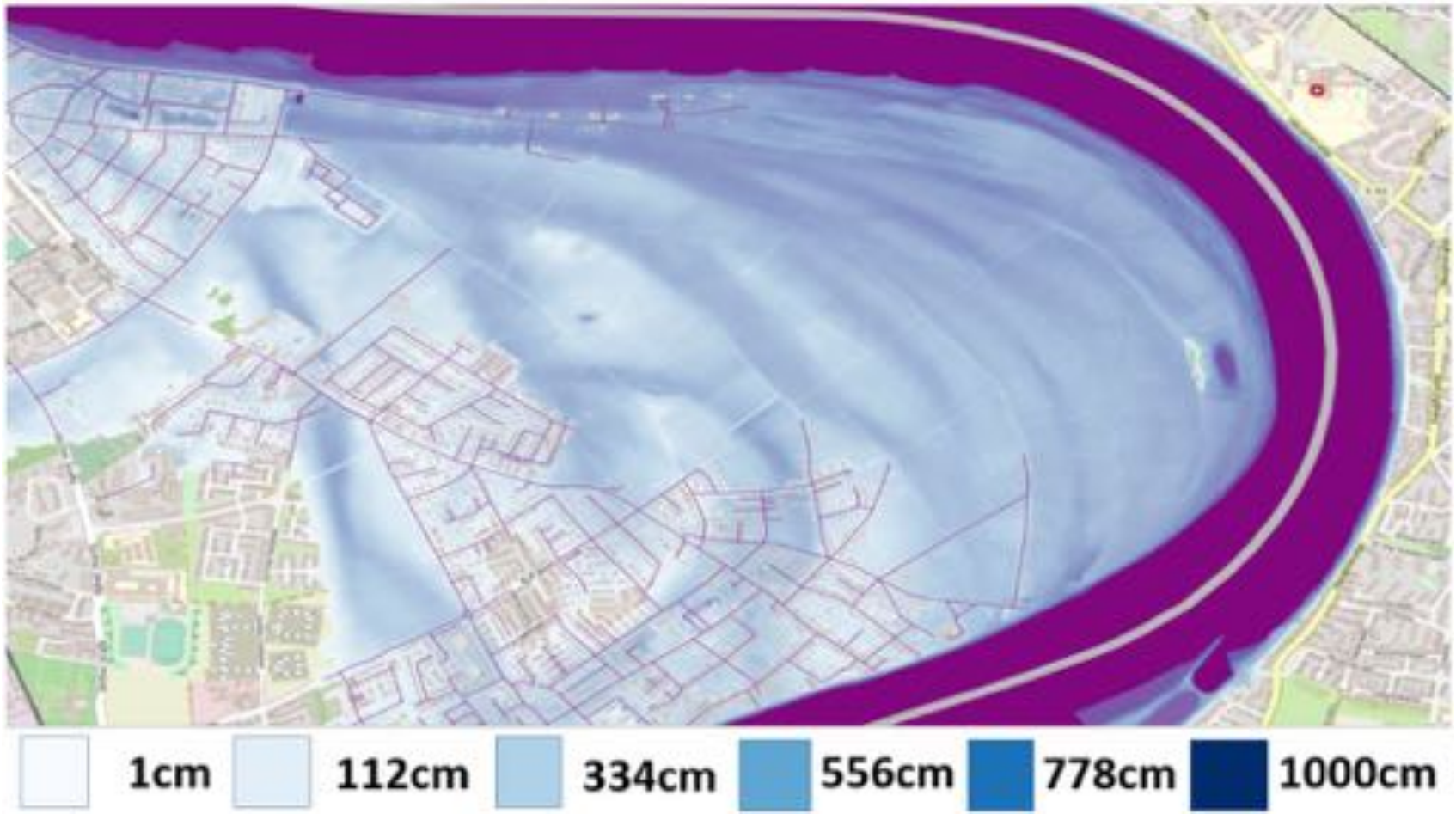
- **Combined vector and raster data analysis** can be achieved from a single query

GeoSPARQL+ Ontology (Homburg et al. ISWC 2020)



GeoSPARQL+ Example (Homburg et al. ISWC 2020)

- Give me all roads which are not flooded by more than 10cm
  - Road network vector dataset
  - Flood altitude raster dataset



GeoSPARQL+ Example (Homburg et al. ISWC 2020)

- Give me all roads which are not flooded by more than 10cm
  - Road network vector dataset
  - Flood altitude raster dataset

```
SELECT ?road
WHERE {
        ?road a ex:Road ; geo:hasGeometry ?roadseg .
        ?roadseg geo:asWKT ?roadseg_wkt .
        ?floodarea a ex:FloodRiskArea ;
                   geo2:asCoverage ?floodarea_cov .
        ?floodarea_cov geo2:asCoverageJSON ?floodarea_covjson.
        BIND(geo2:rasterSmaller(?floodarea_covjson,10) AS? relfloodarea)
        FILTER(geo2:intersects(?roadseg_wkt,?relfloodarea))
        }
```

GeoSPARQL 1.1

- In 2019 OGC reactivated the **GeoSPARQL Standards Working Group (SWG**) in order to publish a new version of GeoSPARQL
  - GeoSPARQL 1.1 will contain **non-breaking changes** with respect to the 2012 GeoSPARQL version (version 1.0)
  - SWG foresees another future version that will no longer be backward compatible with version 1.0
  - Currently (as of Oct 2022), SWG is **finalizing** the proposal regarding version 1.1
- GeoSPARQL 1.1 will incorporate **modifications and improvements** resulted from users' requests after a **decade of wide adoption** of GeoSPARQL 1.0
- **Living document** with the draft version available at: https://opengeospatial.github.io/ogc-geosparql/

GeoSPARQL 1.1

- Extended ontology
  - **Classes:** Spatial Object Collection, Feature Collection, Geometry Collection
    - These **collections** are used to represent groups of spatial objects, features and geometries respectively, in order to offer to the users **interoperability with popular GIS tools**, which usually use the notion of collections of such entities
  - **Spatial Object Properties** used to define **scalar** properties: geo:hasLength, geo:hasPerimeterLength, geo:hasArea and geo:hasVolume are subproperties of **geo:hasSize**, whereas geo:hasMetricLength, geo:hasMetricPerimeterLength, geo:hasMetricArea and geo:hasMetricVolume are subproperties of **geo:hasMetricSize**.
  - **Feature Properties:** geo:hasBoundingBox and geo:hasCentroid
  - **Geometry Serializations:** New geometry serializations are now defined in GeoSPARQL 1.1 in order to support the following formats: **GeoJSON, KML** and **DGGS,** and also new functions are also defined for **transforming between different serialization** formats.

GeoSPARQL 1.1

- **Non-topological Query Functions** which **add functionality** in GeoSPARQL that was **missing in comparison to capabilities commonly offered by GIS tools or spatial databases**.
  - geof:area, geof:coordinateDimension, geof:dimension, geof:geometryN, geof:geometryType, geof:is3D, geof:isEmpty, geof:isMeasured, geof:isSimple, geof:length, geof:maxX, geof:maxY, geof:maxZ, geof:minX, geof:minY, geof:minZ, geof:numGeometries, geof:spatialDimension and geof:transform.
- **Spatial Aggregate Functions** that accept as input a set of geometries:
  - geof:aggBoundingBox, geof:aggBoundingCircle, geof:aggCentroid, geof:aggConcaveHull and geof:aggUnion.
- **Profile definition** using the RDF the Profiles Vocabulary that facilitates the formal definition of profiles of specifications in a machine-readable way.
- **RDF validation file** expressed in the Shapes Constraint Language (SHACL) used to express specific conditions in order to perform validity check on input RDF graphs.
  - e.g. Ensure that
    - each entity that corresponds to a geometry has at least one declared serialization
    - an entity cannot act as both subject and object in triples whose properties are geo:hasGeometry or a subproperty of it
    - RDF literals that correspond to WKT should be well-formed according to the official WKT specification

Thank you!

Questions ?

Geospatial RDF stores

Theofilos Ioannidis
@tioannid1
tioannid@di.uoa.gr
tioannid@yahoo.com

Outline

- Part A - Geospatial RDF stores (25 min)
- Part B – Q&A or Benchmarking geospatial RDF stores (5 min)

Part A - Geospatial RDF stores

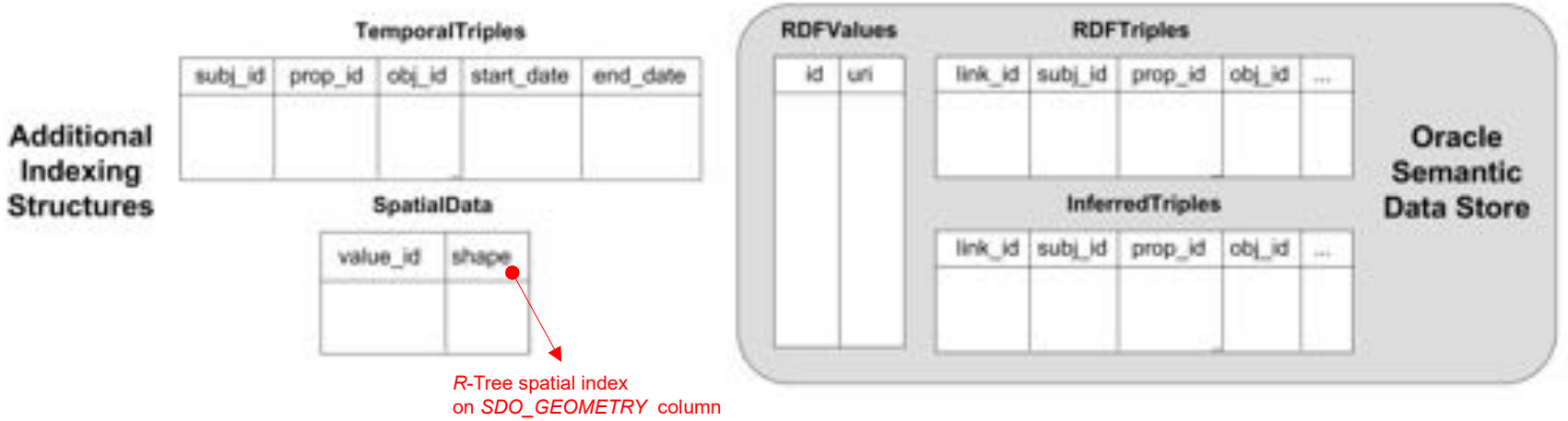
Part A.1 - Early Experimental Systems

Perry's Ph.D dissertation

The system of Brodt et al.

Perry PhD dissertation

- Implementation on top of Oracle 10g by Wright State University
- Support for SPARQL-ST
- GeoRSS GML serialization of geometries
- Spatial and temporal variables
- Spatial and temporal filters (RCC8, Allen)
- R-tree spatial index



Perry, Matthew Steven. "A framework to support spatial, temporal and thematic analytics over semantic web data." (2008)

Brodt et al.

- Built on top of RDF-3X by University of Stuttgart
- No GeoSPARQL support
- Geometries represented as typed WKT literals
- WGS84 supported
- OGC-SFA spatial operations as SPARQL filter functions
- R-tree supported (but only used for spatial selections)

Brodt, Andreas, Daniela Nicklas, and Bernhard Mitschang. "Deep integration of spatial query processing into native RDF triple stores." Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems. 2010

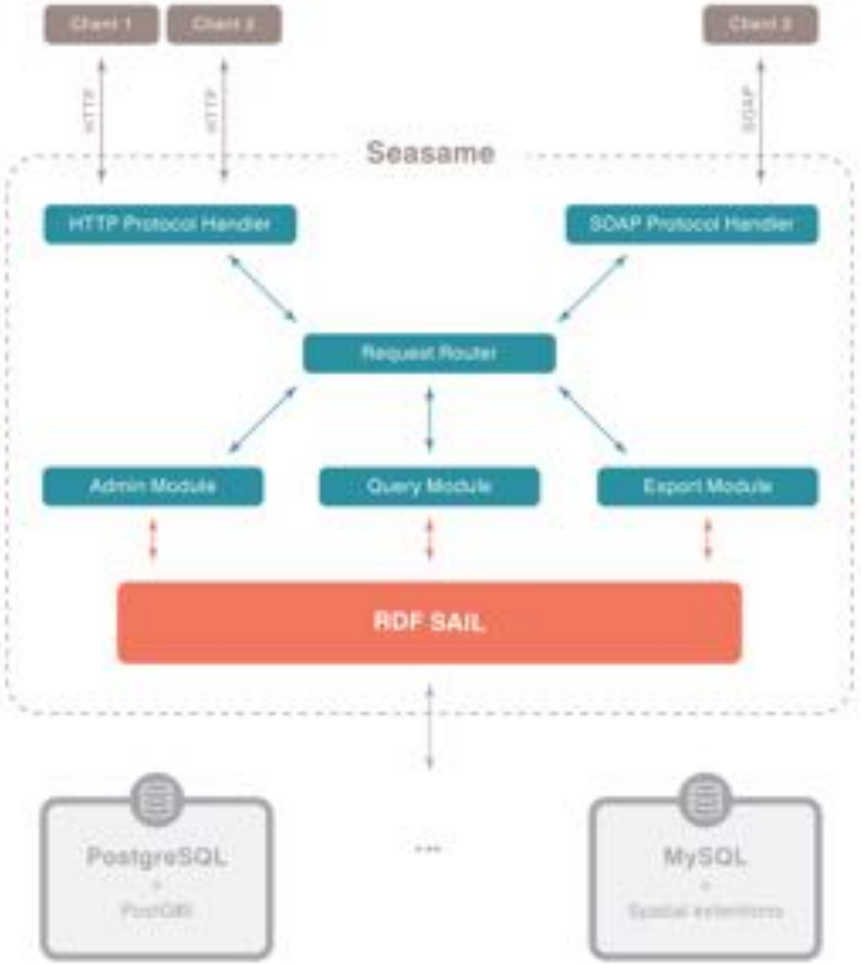
Part A.2 - Outdated Systems

OpenRDF Sesame

uSeekM

OpenRDF Sesame

- Java framework for processing and handling RDF data by Aduna.
- SAIL (Storage and Inference Layer), stackable architecture
- Spatial extensions: extending RDBMS SAIL with spatial databases
- Major releases available: v2.x and v4.x
- Open source available at: https://sourceforge.net/projects/sesame/



uSeekM

- Spatial plugin for Sesame by OpenSahara. Supports:
- GeoSPARQL support
  - Core
  - Topology Vocabulary Extension (Simple Features, Egenhofer, RCC8)
  - Geometry Extension (WKT)
  - Geometry Topology Extension (Simple Features, Egenhofer, RCC8, WKT)
  - RDFS Entailment Extension (Simple Features, Egenhofer, RCC8, WKT)
- Only WGS84 CRS
- Available spatial indexes: R-Tree-over-GiST, Quadtree, Geohash (PostGIS or ElasticSearch as providers)
- Open source (Apache v2.0). Currently, no available official source code location.
- https://web.archive.org/web/20140415085418/https://dev.opensahara.com/projects/useekm

Part A.3 - Systems with Limited Geospatial/GeoSPARQL Functionality

AllegroGraph

OpenLink Virtuoso

Blazegraph DB

MarkLogic Server

Allegrograph

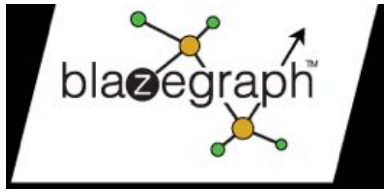
FRANZ INC.

- Quad store developed by Franz Inc. Supports:
- No GeoSPARQL
- Point geometries (N-dimensional Geospatial)
- Non standard CRSs
- Only a few spatial operations supported (Buffer, Bounding Box, Distance)
- Includes temporal functionalities: datetimes, time points, and time intervals and relations (properties) between these
- Is horizontally scalable through federation and sharding (FedShard)
- Latest v7.3.0 (May 2022)
- Closed source. Free edition available at https://allegrograph.com/downloads/

OpenLink Virtuoso

- Developed by OpenLink. Virtuoso Open Source (VOS)
- 3 Java APIs: Jena (v2.x, v3.0.x, v4.3.x), Sesame (v1, v2, v4), RDF4J v2.x.
- No GeoSPARQL (stable/7 branch), GeoSPARQL* (develop/7 branch)
- Points only
- Serialized as typed literals (datatype `virtrdf:Geometry`)
- Spatial operations (subset of SQL/MM)
- Multiple CRS
- WKT for geometry types
- R-tree spatial index
- Latest v7.2.7 (May 2022)
- Latest from develop/7 branch v7.2.6-rc1 (Oct 2018)
- VOS available at http://vos.openlinksw.com/owiki/wiki/VOS

Blazegraph DB



- Previously known as Bigdata. Developed by SYSTAP, LLC. Supports:
- Written in Java, uses Sesame, no GeoSPARQL yet
- Dedicated, typed literals
  - http://www.bigdata.com/rdf/geospatial/literals/v1#lat-lon
  - http://www.bigdata.com/rdf/geospatial/literals/v1#lat-lon-time
- Rectangle and distance queries over geospatial coordinates, possibly combined with range scans over the temporal coordinate
- Geospatial queries are done through a custom SERVICE extension http://www.bigdata.com/rdf/geospatial#search
- Generic z-order index implementation for multi-dimensional range scans
- Latest v2.1.6 RC (Feb 2020)
- Available at https://blazegraph.com/

MarkLogic Server

- NoSQL database that supports multi-model data representation
- JSON/XML documents, RDF triples and relational data
- Data access languages: JavaScript/XQuery, SPARQL and SQL
- Triple query languages: SPARQL, XQuery or JavaScript
- No GeoSPARQL
- Geospatial serialization vocabularies: GML, KML, GeoRSS and GeoJSON
- Geospatial types : Point, Box, Circle, Polygon, Complex Polygon (WKT), Linestring (WKT)
- Supports Cartesian, WGS84 coordinate systems and ETRS89 (European specific)
- Spatial indexes: (i) Point index, fast, in memory, point-only, (ii) Region path index, Geohash or R-tree, allows DE-9IM operations
- Latest v10.0-9.5 (Sep 2022)
- Available at https://www.marklogic.com/

Part A.4 - Systems with Partial GeoSPARQL Support

Stardog

Strabon

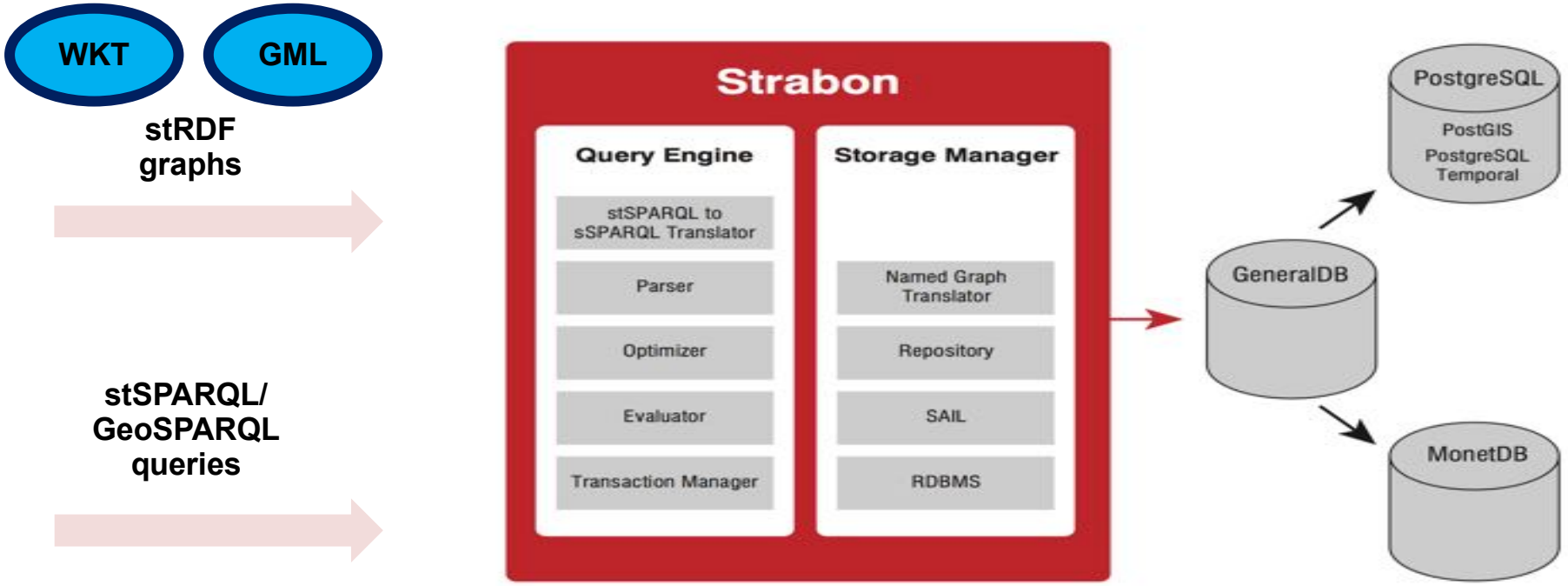
Eclipse RDF4J

Stardog

- Developed by Stardog. Supports:
- partial GeoSPARQL support (after 2018)
  - WKT serialization
    - points and rectangles, by default
    - All shapes, by installing JTS (Java Topology Suite https://locationtech.github.io/jts/)
  - Functions geof:relate, geof:distance, geof:within, geof:nearby and geof:area
- Spatial Indexing
  - spatial indexing, based on Lucene Spatial, geohash prefix tree (precision 11 for sub-meter accuracy)
  - approximate matching controlled with a precision parameter
  - precision specified upon database creation
- Supports WGS84 Geo Positioning RDF vocabulary (World Geodetic System 1984) https://www.w3.org/2003/01/geo/wgs84_pos
- Provides support for Jena and Sesame APIs
- Latest v8.1.0 (Sep 2022)
- Closed source. Available at: https://www.stardog.com

Strabon

Find more at: http://www.strabon.di.uoa.gr/



[1] Kyzirakos, Kostis, Manos Karpathiotakis, and Manolis Koubarakis. "Strabon: a semantic geospatial DBMS." International Semantic Web Conference. Springer, Berlin, Heidelberg, 2012
[2] Kyzirakos, Kostis, et al. "The Spatiotemporal RDF Store Strabon." SSTD 2013 Proceedings of the 13th International Symposium on Advances in Spatial and Temporal Databases - Volume 8098, 2013, pp. 496–500.

Strabon - Geospatial features

Support for:

- stRDF and stSPARQL
- GeoSPARQL support
  - Core
  - Geometry Extension (WKT, GML)
  - Geometry Topology Extension (Simple Features, Egenhofer, RCC8, WKT, GML)
- Multiple Coordinate Reference Systems (CRS)
- Builds on Sesame RDBMS Sail
- Geospatial relational database as back-end (PostGIS, MonetDB)
- R-tree over GiST index (PostGIS)
- Latest v3.2.9. Available at http://www.strabon.di.uoa.gr/

Eclipse RDF4J

- Java framework for processing and handling RDF data (former OpenRDF Sesame by Aduna).
- Based on Spatial4J and JTS libraries for geospatial reasoning.
- SAIL (Storage and Inference Layer), stackable architecture
- Memory store, NativeStore, Lucene SAIL and its derivates (the SolrSail and the ElasticSearchSail), (Sesame's RDBMS Sail removed), LMDB Store fast embeddable key-value based on the Symas Lightning Memory-Mapped Database (https://www.symas.com/lmdb)
- GeoSPARQL :
  - By default supported on any type of store (memory, native, etc)
  - The Lucene SAIL and its derivates (the SolrSail and the ElasticSearchSail) have built-in optimizations for geospatial querying.
  - Default spatially indexed property is geo:asWKT
  - LuceneSail configuration allows for customization of spatially indexed properties
    - Core
    - Geometry (partial) (WKT)
    - Geometry Topology Extension (Simple Feature, Egenhofer, RCC8, WKT)
    - Only WKT serializations
    - RDFS entailment
- Latest v4.2.0 (Sep 2022)
- Open source available at: https://rdf4j.org

Part A.4 - Systems with Extensive GeoSPARQL Support

Parliament

Oracle RDF Spatial and Graph

AnzoGraph DB

GraphDB

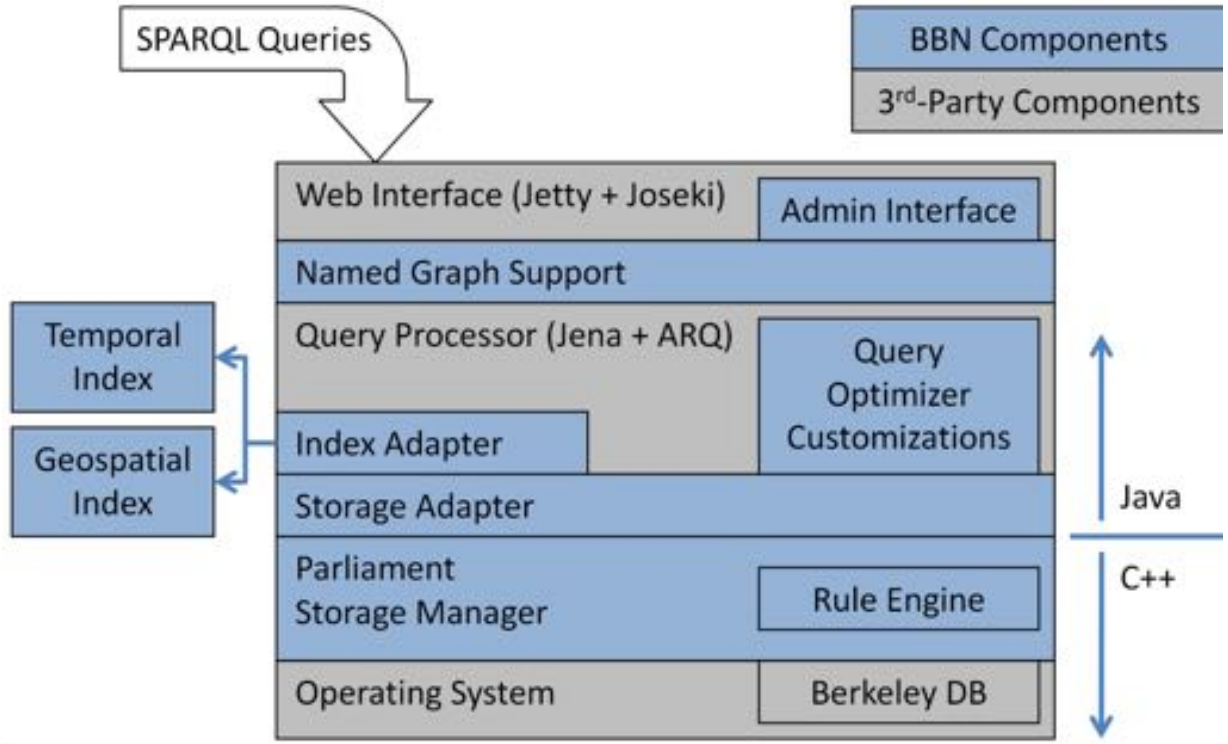
Apache Jena GeoSPARQL

Parliament

- Developed (2009) by Raytheon BBN Technologies (Dave Kolas).
- First GeoSPARQL implementation. Supports:
  - Core
  - Topology vocabulary
  - Geometry
  - Geometry Topology
  - Multiple CRSs
  - Both WKT and GML serializations
  - RDF entailment required for (geo:asWKT, geo:asGML)
- It uses BerkeleyDB as backend and Jena+ARQ as SPARQL processor
- Standard R-tree index

Latest v2.8.1 (June 2022) features: separation between Parliament's software and data files, Java 8 and 11 support, offered as Docker image

- Open source available (since 2018) at : https://github.com/SemWebCentral/parliament



Oracle Spatial and RDF Graph

- Developed by Oracle
- GeoSPARQL support
  - Core
  - Topology Vocabulary Extension (Simple Features)
  - Geometry Extension (WKT 1.2.0, GML 3.1.1)
  - Geometry Topology Extension (Simple Features, WKT 1.2.0, GML 3.1.1)
  - RDFS Entailment Extension (Simple Features, WKT 1.2.0, GML 3.1.1)
- CRS support
- R-Tree as spatial index for up to 4 dimensions or composite B-tree index on point data for non spatial join operations
- Virtual RDF graphs (as of Oracle Spatial and Graph 12c R2)
- Latest version 21c (April 2022)

AnzoGraph DB

- Massively parallel processing native graph database built for data harmonization and analytics
- Scales from single server to multiple servers in a cluster and cloud environments
- It holds the record for the fastest execution of the LUBM 1 Trillion Triple Benchmark (Google Cloud Platform, October 2016)
- GeoSPARQL support
  - Core
  - Topology Vocabulary Extension (Simple Features, Egenhofer, RCC8)
  - Geometry Extension (WKT 1.x, GML 2.x)
  - Geometry Topology Extension (Simple Features, Egenhofer, RCC8, WKT 1.x, GML 2.x)
  - RDFS Entailment Extension (Simple Features, Egenhofer, RCC8, WKT 1.x, GML 2.x)
- CRS support
- Geospatial aggregate functions provided through user-defined aggregate extensions
- Latest v2.5.10
- Closed source. Available at: https://cambridgesemantics.com/

GraphDB Free

ontotext

- Former OWLIM. Based on RDF4J framework, developed by Ontotext.
- GeoSPARQL
  - Core
  - Topology Vocabulary Extension (Simple Features, Egenhofer, RCC8)
  - Geometry Extension (WKT 1.x, GML 3.x)
  - Geometry Topology Extension (Simple Features, Egenhofer, RCC8, WKT 1.x, GML 2.x)
  - RDFS Entailment Extension (Simple Features, Egenhofer, RCC8, WKT 1.x, GML 2.x)
  - WKT serialization
- Multiple CRSs (v9.x onward)
- GeoSPARQL support through the GeoSPARQL plugin
  - Indexing algorithms, based on Lucene Spatial, quad or geohash prefix tree
  - Approximate matching controlled with a precision parameter
  - Reconfiguration of algorithm and precision allowed through SPARQL update statements
  - Rebuilding GeoSPARQL index on demand
- Offers useful GeoSPARQL extensions based on the USeekMSail, i.e., ext:isValid(geometry), ext:area(geometry)
- Supports WGS84 Geo Positioning RDF vocabulary (World Geodetic System 1984) https://www.w3.org/2003/01/geo/wgs84_pos
- Features 2 very fast bulk importers (LoadRDF, Preload)
- Latest v10.0.2 (Aug 2022)
- Closed source. Available at: https://graphdb.ontotext.com/

Apache Jena GeoSPARQL

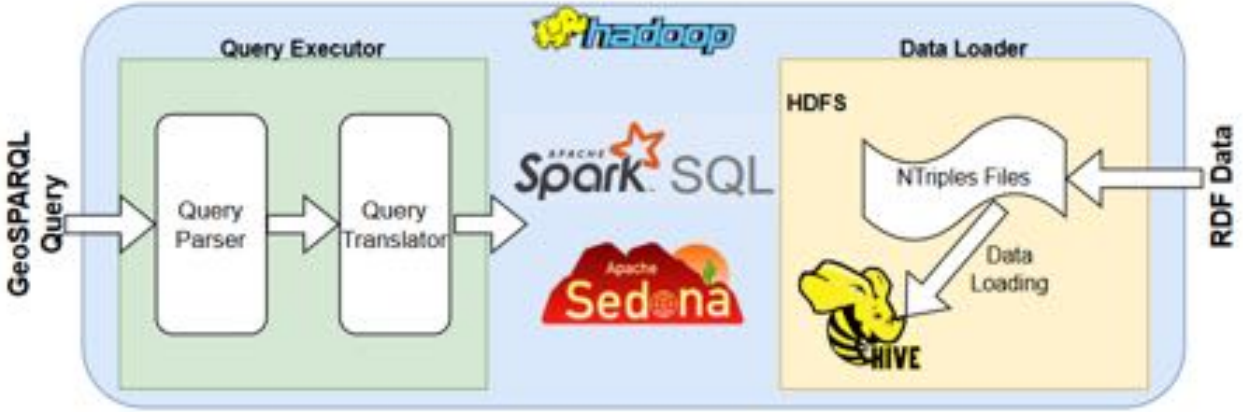
- Pure Java implementation which does not require any setup or configuration of any third party relational databases and geospatial extensions
- Full GeoSPARQL support
  - Core
  - Topology Vocabulary Extension (Simple Feature, Egenhofer and RCC8)
  - Geometry Extension (WKT, GML 2.0)
  - Geometry Topology Extension (Simple Feature, Egenhofer, RCC8, WKT, GML 2.0)
  - RDFS Entailment Extension (Simple Feature, Egenhofer, RCC8, WKT, GML 2.0)
  - Query Rewrite (Simple Feature, Egenhofer, RCC8, WKT, GML 2.0)
- All indexing and caching is performed during query execution
- 3 indexes (Geometry Literal, Geometry Transform, Query Rewrite)
- Also supports WGS84 Geo Positioning RDF vocabulary (World Geodetic System 1984)
- Latest v3.16.0 (Jul 2020) as a maven artifact org.apache.jena:jena-geosparql:3.16.0
- Open source available at: https://github.com/galbiston/geosparql-jena

Part A.5 – Distributed RDF Stores with GeoSPARQL Support

Strabo2

Strabo2

- Apache Spark Java implementation for Hadoop clusters
- Extends Ontop-spatial for GeoSPARQL to Spark-SQL translation
- Geospatial capabilities provided by Apache Sedona (former GeoSpark)
- Apache Hive as persistence layer with Parquet storage format
- Vertical Partitioning used as the logical partitioning strategy
- GeoSPARQL support: Simple Feature family, WKT serialization
- Open source available at: https://github.com/db-ee/Strabo-2



ISWC 2022: Main Track 6b: Querying, 26th October 2022
Strabo 2: Distributed Management of Massive Geospatial RDF Datasets
Dimitris Bilidas, Theofilos Ioannidis, Nikos Mamoulis and Manolis Koubarakis

Part B - Q&A or Geographica 2 benchmark

Useful links

Source code, Datasets, Query sets and most of the results are available on the following 2 web sites:

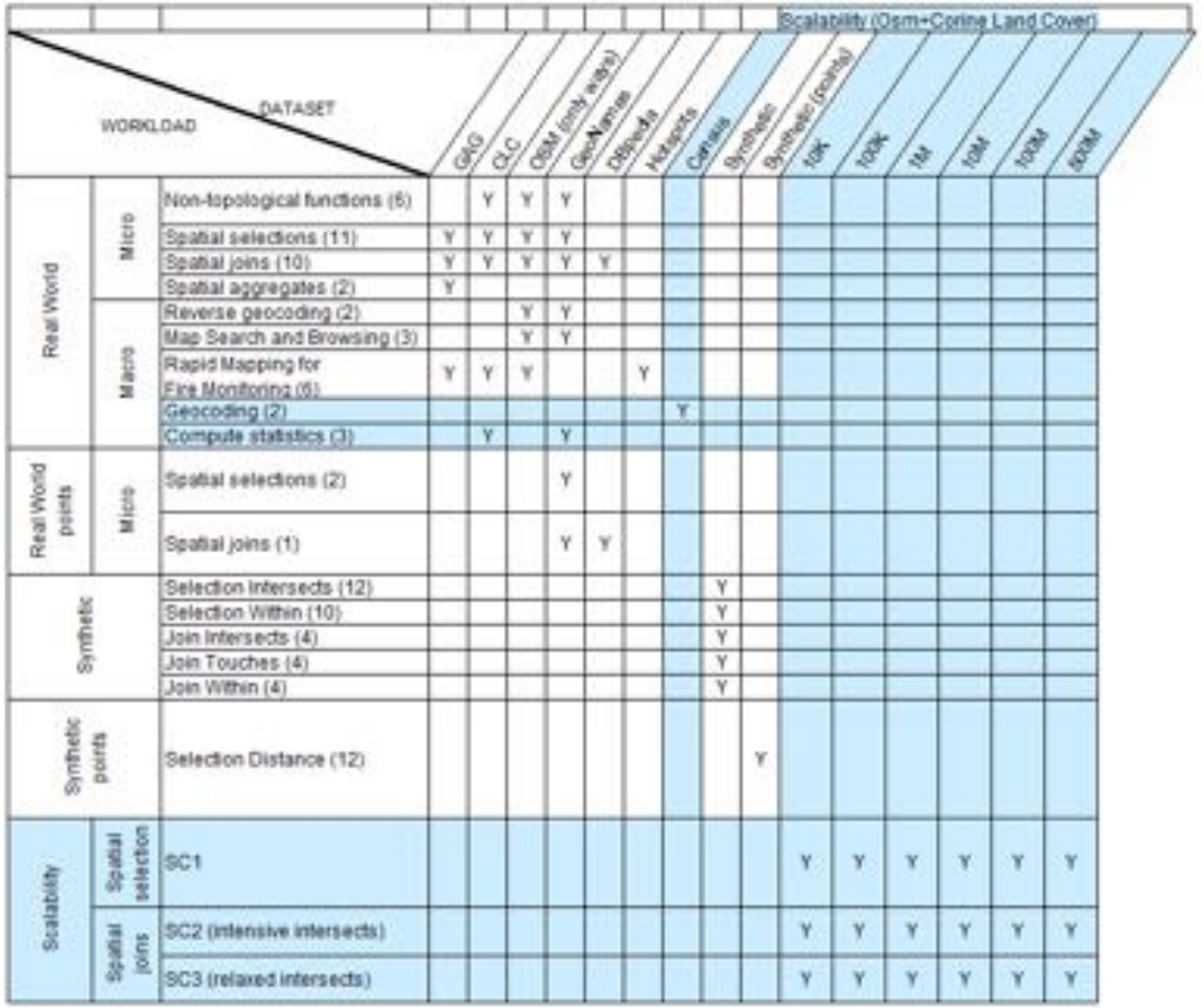
- http://geographica.di.uoa.gr/
- http://geographica2.di.uoa.gr/

Geographica 2

- **Purpose**: evaluate RDF stores supporting GeoSPARQL and stSPARQL
- **Workloads**:
  - Real world (micro and macro benchmarks)
  - Synthetic
  - Scalability: OSM+Corine Land Cover Reference Dataset 1 (RD-1)
- **Datasets**: GAG, CLC, LGD (OSM), GeoNames, DBpedia, Hotspots, Synthetic-512, Synthetic-Points-1024, Census, Scalability (OSM+CLC RD-1: 10K, 100K, 1M, 10M, 100M, 500M triples)
- **Scenarios**: Micro, Macro (Reverse geocoding, Map search & browsing, Rapid mapping for fire monitoring, Geocoding, Computation of Statistics), Scalability
- **Systems**: uSeekM, Parliament, Strabon, System X (Par, Ser), GraphDB, RDF4J (NativeSail, LuceneSail+NativeSail)
- **Limited functionality**: VOS (Virtuoso OpenLink Server), System Y in addition to the 6 (8 variants) other systems against points only versions of Real World and Synthetic datasets

Ioannidis Theofilos, Garbis George, Kyzirakos Kostis, Bereta Konstantina, & Koubarakis Manolis (2021). "Evaluating geospatial RDF stores using the benchmark Geographica 2". Journal on Data Semantics, 10(3), 189-228.

Geographica 2 – Workload/Dataset matrix



WORKLOAD			DATASET: GAG	CLC	OSM (only edges)	GeoNames	DBpedia	Hotspots	Corine	Synthetic	Synthetic (points)	Scalability (Osm+Corine Land Cover): 10K	100K	1M	10M	100M	800M
Real World	Micro	Non-topological functions (6)		Y	Y	Y											
		Spatial selections (11)	Y	Y	Y	Y											
		Spatial joins (10)	Y	Y	Y	Y	Y										
		Spatial aggregates (2)	Y														
	Macro	Reverse geocoding (2)			Y	Y											
		Map Search and Browsing (3)			Y	Y											
		Rapid Mapping for Fire Monitoring (6)	Y	Y	Y			Y									
		Geocoding (2)							Y								
		Compute statistics (3)		Y		Y											
Real World points	Micro	Spatial selections (2)				Y											
		Spatial joins (1)				Y	Y										
Synthetic		Selection Intersects (12)								Y							
		Selection Within (10)								Y							
		Join Intersects (4)								Y							
		Join Touches (4)								Y							
		Join Within (4)								Y							
Synthetic points		Selection Distance (12)									Y						
Scalability	Spatial selection	SC1										Y	Y	Y	Y	Y	Y
	Spatial joins	SC2 (intensive intersects)										Y	Y	Y	Y	Y	Y
		SC3 (relaxed intersects)										Y	Y	Y	Y	Y	Y

Geographica 2 – Workload/System matrix

WORKLOAD \ SUT			uStarMM v1.2.1	Parliament v2.7.4	Strabon v3.2.9	System X parallel	System X Serial	Virtuoso	System Y	GraphDB Free v8.5.1	RDF4J v2.4.3	RDF4J + Lucene v2.4.3
Real World	Micro	Non-topological functions (6)	Y	Y	Y	Y	Y			Y	Y	Y
		Spatial selections (11)	Y	Y	Y	Y	Y			Y	Y	Y
		Spatial joins (10)	Y	Y	Y	Y	Y			Y	Y	Y
		Spatial aggregates (2)	Y	Y	Y	Y	Y			Y	Y	Y
	Macro	Reverse geocoding (2)	Y	Y	Y	Y	Y			Y	Y	Y
		Map Search and Browsing (3)	Y	Y	Y	Y	Y			Y	Y	Y
		Rapid Mapping for Fire Monitoring (6)	Y	Y	Y	Y	Y			Y	Y	Y
		Geocoding (2)	Y	Y	Y	Y	Y			Y	Y	Y
		Compute statistics (3)	Y	Y	Y	Y	Y			Y	Y	Y
Real World points	Micro	Spatial selections (2)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
		Spatial joins (1)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Synthetic		Selection Intersects (12)	Y	Y	Y	Y	Y			Y	Y	Y
		Selection Within (10)	Y	Y	Y	Y	Y			Y	Y	Y
		Join Intersects (4)	Y	Y	Y	Y	Y			Y	Y	Y
		Join Touches (4)	Y	Y	Y	Y	Y			Y	Y	Y
		Join Within (4)	Y	Y	Y	Y	Y			Y	Y	Y
Synthetic points		Selection Distance (12)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Scalability	Spatial selection	SC1			Y					Y	Y	Y
	Spatial joins	SC2 (intensive intersects)			Y					Y	Y	Y
		SC3 (relaxed intersects)			Y					Y	Y	Y

Geographica 2 - Scalability Workload

- **Purpose**: the scalability experiment aims at discovering the limits of the systems under test as the number of triples in the dataset increase
- **Method**: Each selected system is tested against six increasingly bigger, proper subsets of a big real world reference geospatial dataset. For each system-dataset combination we measure:
  - ✔ the repository size on disk,
  - ✔ the bulk loading time taking into consideration the limitations of loading methods of each system and
  - ✔ the response time in three queries which represent a spatial selection, a heavy spatial join with higher spatial selectivity and a lighter spatial join with lower spatial selectivity
- **Systems**: Strabon, GraphDB, RDF4J

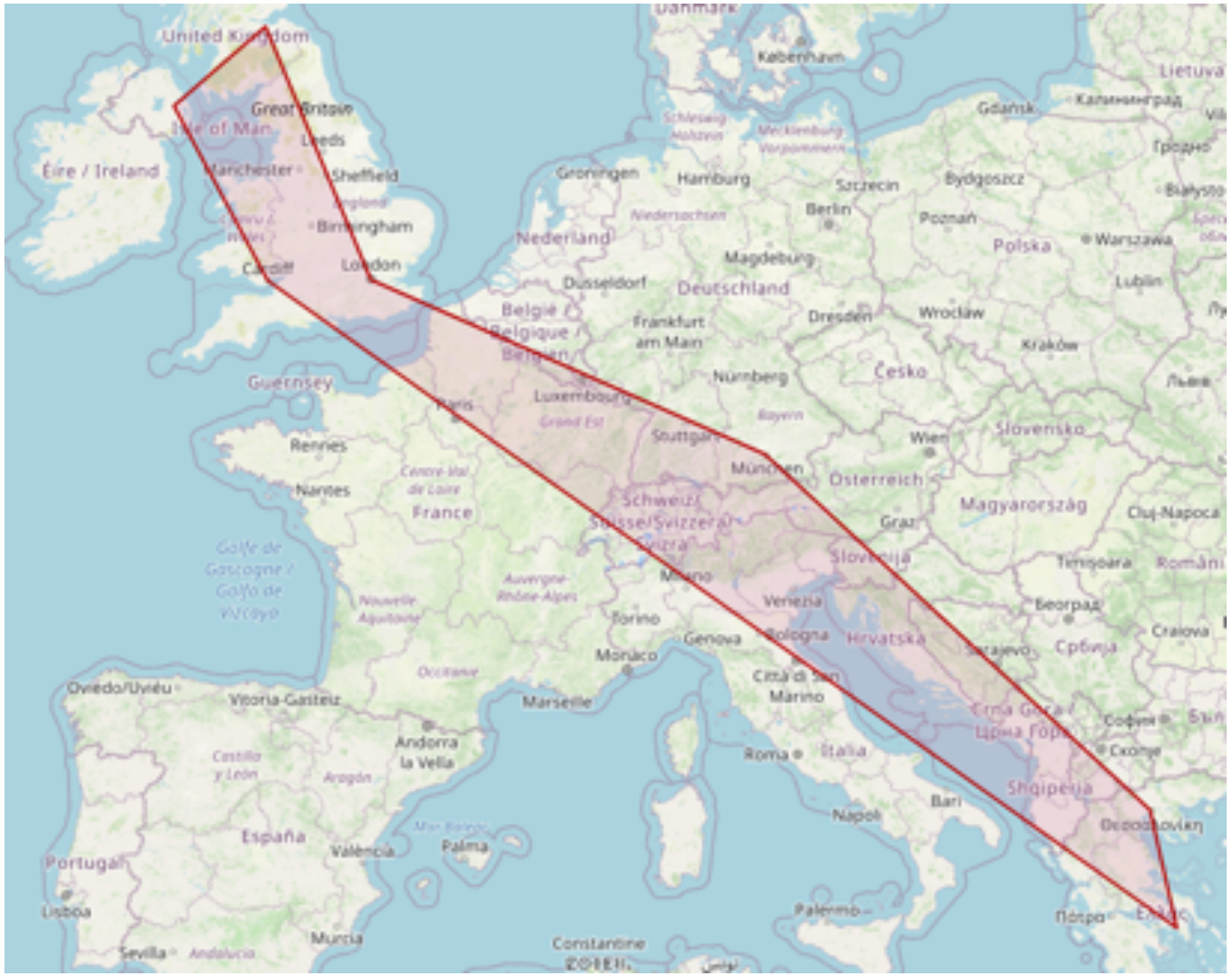
Geographica 2 - Scalability Reference Dataset characteristics

Datasets	Country	Triples (M)	Size (MB)
OSM	Wales	6.56	1,206
	Scotland	15.78	2,913
	Greece	15.22	2,877
	N. Ireland	15.27	3,240
	England	104.21	18,965
	Germany	326.48	59,002
CLC-2012	*39 countries*	16.60	11,283
	Totals	483.52	99,486

Geographica 2 - Scalability Datasets basic characteristics

Dataset	# of Features	# of Points	# of Lines	# of Polygons
10K	1,135	587	0	900
100K	12,166	6,623	4,239	2,531
1M	118,161	46,781	45,238	29,200
10M	1,038,739	317,865	328,630	427,842
100M	10,259,959	904,677	2,058,386	7,553,440
500M	48,623,878	5,520,767	15,771,932	23,390,220

Geographica 2 - Scalability Query SC1 (Spatial Selection)



Geographica 2 - Scalability Query SC1 (Spatial Selection)

SC1_Geometries_Intersects_GivenPolygon: Find all geometries that intersect with the given polygon

```
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX lgd: <http://data.linkedeodata.eu/ontology#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT ?s1 ?o1
WHERE {
 ?s1 geo:asWKT ?o1 .
 FILTER(geof:sfIntersects(?o1, "POLYGON((23.708496093749996 37.95719224376526,22.906494140625
40.659805938378526,11.524658203125002 48.16425348854739,-0.1181030273437499 51.49506473014367,-
3.2189941406250004 55.92766341247031,-5.940856933593749 54.59116279530599,-3.1668090820312504
51.47967237816337,23.708496093749996
37.95719224376526))"^^<http://www.opengis.net/ont/geosparql#wktLiteral>)).
}
```

Geographica 2 - Scalability Value distribution of `lgo:has_code` property

`lgo:has_code ⇒ <http://data.linkedeodata.eu/ontology#has_code>`

lgo:has_code	lgo:has_fclass	10K	100K	1M	10M	100M	500M
1001 (used in SC2, SC3)	city	1	1	7	14	84	232
5601	railway_station	15	284	284	669	1,194	8,449
5621	bus_stop	4	4,416	4,416	22,337	35,555	503,455
5622	bus_station	36	46	46	98	425	2,647
5641	taxi	7	43	43	217	886	5,798
5661	ferry_terminal	4	18	18	153	583	1,508
5601,5621,5622,5641,5661 (used in SC3)		66	4,807	4,807	23,488	38,643	521,857
(5001-5999) - [5260] (used in SC2)	transportation except parking	66	4,875	11,412	264,199	1,978,632	16,151,652

Geographica 2 - Scalability Query SC2 (Spatial Join)

SC2_Intensive_Geometries_Intersect_Geometries: Find all transportation-related features (except parking) within cities

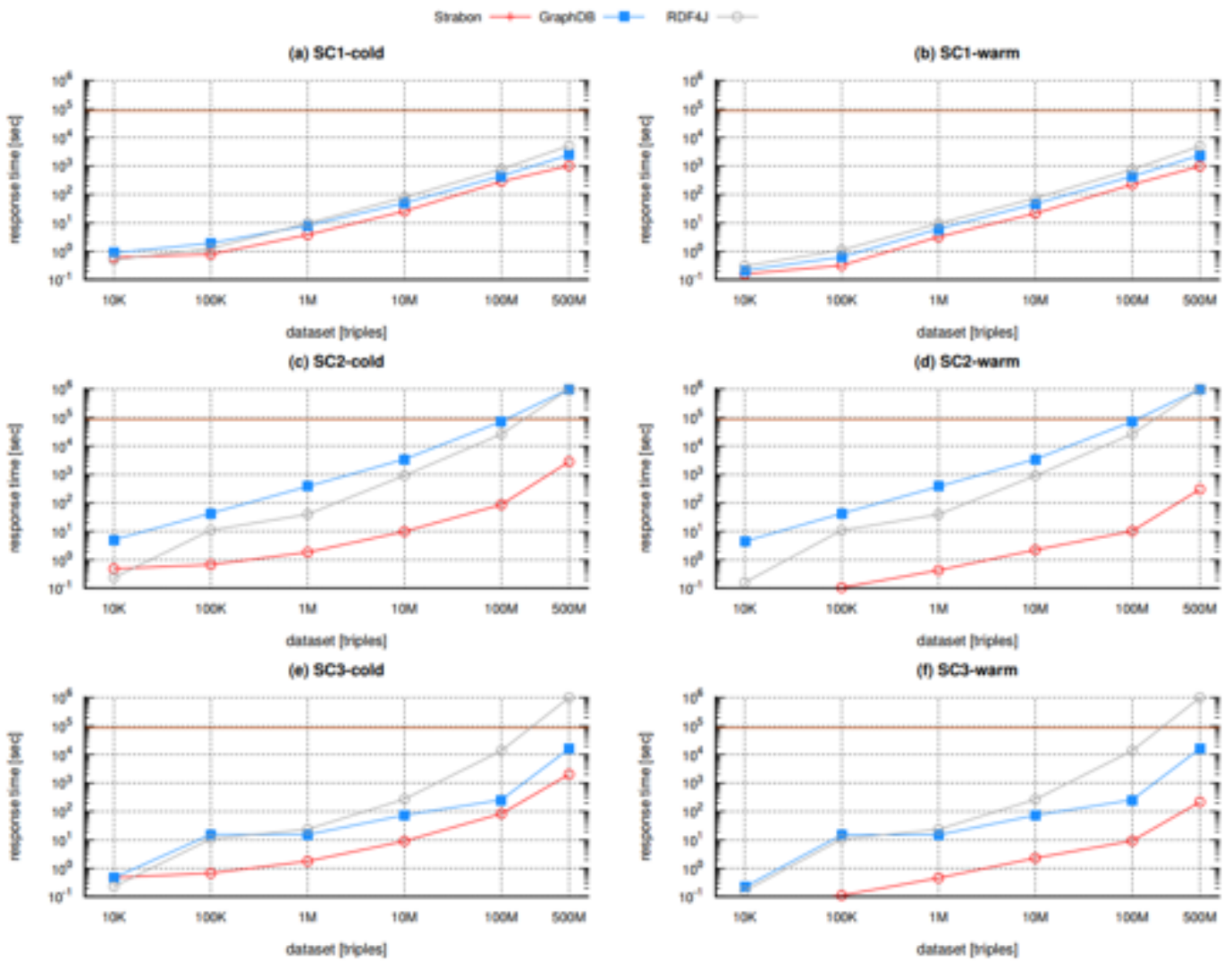
```
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX lgo: <http://data.linkedeodata.eu/ontology#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT ?s1 ?s2
WHERE {
 ?s1 geo:hasGeometry [ geo:asWKT ?o1 ] ;
    lgo:has_code "1001"^^xsd:integer .
 ?s2 geo:hasGeometry [ geo:asWKT ?o2 ] ;
    lgo:has_code ?code2 .
 FILTER(?code2>5000 && ?code2<6000 && ?code2 != 5260) .
 FILTER(geof:sfIntersects(?o1, ?o2)).
}
```

Geographica 2 - Scalability Query SC3 (Spatial Join)

SC3_Relaxed_Geometries_Intersect_Geometries: Find all bus stops, bus stations, railway stations, taxis and ferry terminals within cities

```
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX lgo: <http://data.linkedeodata.eu/ontology#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT ?s1 ?s2
WHERE {
 ?s1 geo:hasGeometry [ geo:asWKT ?o1 ] ;
    lgo:has_code "1001"^^xsd:integer .
 ?s2 geo:hasGeometry [ geo:asWKT ?o2 ] ;
    lgo:has_code ?code2 .
 FILTER(?code2 IN (5622, 5601, 5641, 5621, 5661)) .
 FILTER(geof:sfIntersects(?o1, ?o2)).
}
```

Geographica 2 - Scalability results



Transformation of geospatial data

Transforming to RDF

- **Direct mapping** approach became a W3C recommendation in 2012
  - tables becomes classes
  - tables' attributes are mapped to RDF properties that represent the relation between subject and object
  - SQL table and column identifiers compose RDF IRIs in the direct graph

A. Bertails, M. Arenas, E. Prudĥommeaux, J. Sequeda, A direct mapping of relational data to RDF. W3C Recommendation Available from: http://www.w3.org/TR/rdb-direct-mapping/, 2012.

Mapping Languages

R2RML

- A language for expressing customized mappings from relational databases to RDF graphs
- Became W3C recommendation in 2012.
- Express transformation of existing relational data into the RDF data model.

RML

- The RDF Mapping language (RML) is a generic mapping language.
- It can express rules that map data with heterogeneous structures to RDF graphs.
- RML is defined as a superset of R2RML and allows the expression of rules that map relational and semi-structured data (e.g., XML, JSON) into RDF graphs.

RML Example

Input Data

```
id,stop,latitude,longitude
6523,25,50.901389,4.484444
```

```
@prefix rr: <http://www.w3.org/ns/r2rml#>.
@prefix rml: <http://semweb.mmlab.be/ns/rml#>.
@prefix ql: <http://semweb.mmlab.be/ns/ql#>.
@prefix transit: <http://vocab.org/transit/terms/>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
@prefix wgs84_pos:
<http://www.w3.org/2003/01/geo/wgs84_pos#>.
@base <http://example.com/ns#>.

<#AirportMapping>
 rml:logicalSource [
   rml:source "Airport.csv" ;
   rml:referenceFormulation ql:CSV
 ];
 rr:subjectMap [
   rr:template "http://airport.example.com/{id}";
   rr:class transit:Stop
 ];

 rr:predicateObjectMap [
   rr:predicate transit:route;
   rr:objectMap [
     rml:reference "stop";
     rr:datatype xsd:int
     ]
   ];

 rr:predicateObjectMap [
   rr:predicate wgs84_pos:lat;
   rr:objectMap [
     rml:reference "latitude"
   ]
 ];

 rr:predicateObjectMap [
   rr:predicate wgs84_pos:long;
   rr:objectMap [
     rml:reference "longitude"
   ]
 ].
```

RML Mapping

Output Triples

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix transit: <http://vocab.org/transit/terms/>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
@prefix wgs84_pos:
<http://www.w3.org/2003/01/geo/wgs84_pos#>.

<http://airport.example.com/6523> rdf:type transit:Stop.
<http://airport.example.com/6523> transit:route "25"^^xsd:int.
<http://airport.example.com/6523> wgs84_pos:lat "50.901389".
<http://airport.example.com/6523> wgs84_pos:long "4.484444".
```

https://rml.io/specs/rml/#string-template

RML/R2RML Term Maps

- A **Term Map** is a function that generates an RDF term from a logical reference.
- Term maps are used to generate the subjects, predicates and objects of the RDF triples
  - **subject maps**
  - **predicate maps**
  - **object maps**

- A **Term Map** must be exactly one of the following:
  - a constant-valued term map,
  - a reference-valued term map,
  - a template-valued term map.

RML/R2RML Term Maps

```
rr:predicateObjectMap [
        rr:predicateMap [ rr:constant exp:hasType];
        rr:objectMap [
                rr:datatype xsd:string;
                rr:constant “Point”;
        ];
];
```

→ ?x exp:hasType “Point”^^<http://www.w3.org/2001/XMLSchema#string>

```
rr:predicateObjectMap [
        rr:predicateMap [ rr:constant ogc:hasName ];
        rr:objectMap [
                rr:datatype  xsd:string;
                rml:reference "name";
        ];
]
```

→ ?x exp:hasName “George”^^<http://www.w3.org/2001/XMLSchema#string>

```
 rr:subjectMap [
        rr:template "http://trans.example.com/stop/{@id}"
].
```

→ http://trans.example.com/stop/645

Tools/Methods

- The **Geometry2RDF** was one of the first tools that enabled users to transform spatially enabled RDB systems into RDF graphs.
- K. Chentout et al presented how R2RML can be combined with a spatially-enabled relational database in order to transform geospatial data into RDF
- **TripleGeo** is a tool for transforming geospatial features from various sources into RDF graphs.
  - Supports complex formats such as OpenStreetMap data and certain INSPIRE
  - Parallelized executions based on multi-threading and Apache Spark

K. Chentout, A.A. Vaisman, Adding spatial support to R2RML mappings, in: OTM Workshops, in: Lecture Notes in Computer Science, vol. 8186, Springer, 2013.

Kostas Patroumpas et al. "Exposing Points of Interest as Linked Geospatial Data". In:Proceedings of the 16th International Symposium on Spatial and Temporal Databases, SSTD2019

TripleGeo Repo: ***https://github.com/SLIPO-EU/TripleGeo***

The tool GeoTriples

- GeoTriples is a **semi-automated tool that enables the automatic transformation of geospatial data into RDF graphs**
- The transformation process comprises three steps.
    1. GeoTriples generates automatically extended R2RML or RML mappings
    2. As an optional second step, the user may revise these mappings according to her needs
    3. Finally, GeoTriples processes these mappings and produces an RDF graph.

- GeoTriples supports the transformation of Spatially-enabled relational databases, CSV, GeoJSON, ESRI Shapefiles, KML and XML documents

Kostis Kyzirakos et al. "GeoTriples: Transforming geospatial data into RDF graphs using R2RML and RML mappings". In:J. Web Semant.52-53 (2018)

Repo: ***https://github.com/LinkedEOData/GeoTriples***

RML/R2RML Extension

- When transforming geospatial data into RDF we may need to c**ompute on the fly values that are not explicitly present in the source data** (e.g. the dimension of a given geometry, the length of a line, the area of a polygon, etc.)
- We may also want to **compute on the fly which topological, directional or distance relations** hold between two spatial objects
- We have extended RML/R2RML with a **Transformation-valued** term map, that generates an RDF term by applying SPARQL extension function on one or more term maps.
- A transformation-valued term map has exactly one **rrx:function** property and one **rrx:argumentMap** property .
- The **rrx:argumentMap** property has as range an rdf:List of term maps that define the arguments to be passed to the transformation function.

```
rr:predicateObjectMap [
        rr:predicateMap [ rr:constant ogc:asWKT ];
        rr:objectMap [
                rr:datatype  ogc:wktLiteral;
                rrx:function geof:asWKT;
                rrx:argumentMap ( [ rml:reference "geometry"; ] );
        ];
];
```

```
rr:predicateObjectMap [
        rr:predicateMap [ rr:constant exp:hasBuffer];
        rr:objectMap [
                rrx:function geof:buffer;
                rrx:argumentMap (
                        [ rml:reference "geometry"; ]
                        [ rr:constant “10”; rr:datatype xsd:integer]
                        [rr:constant uom:metre]
                 );
        ];
];
```

RML/R2RML Extension

- We can assert topological relations using the topology vocabulary of GeoSPARQL
- A **referencing object map** is a map that allows a predicate–object map to generate as objects the subjects of another triples map.

```
rr:objectMap [
        rr:joinCondition [
                rrx:function geof:sfOverlaps;
                rrx:argumentMap (
                        [ rr:column "Geom" ]
                        [ rr:column "Geom" ; rr:triplesMap <MapB> ]
                )
        ]
];
```

```
rr:objectMap [
        rr:joinCondition [
                rrx:function ex:isClose;
                rrx:argumentMap(
                        [rr:column "Geom"]
                        [rr:column "Geom"; rr:triplesMap <MapB>]
                        [rr:constant "10"; rr:datatype xsd:int ]
                        [rr:constant uom:metre ]
                )
        ]
];
```

GeoTriples System Architecture

GeoTriples comprises three main components: the **Mapping Generator**, the **Mapping Processor** and the **stSPARQL/GeoSPARQL Evaluator.**

Mapping Generator

The mapping generator is given as input a data source and creates automatically an R2RML/RML mapping document, depending on the type of the input. The user may edit the generated mapping document to make it comply with her requirements.

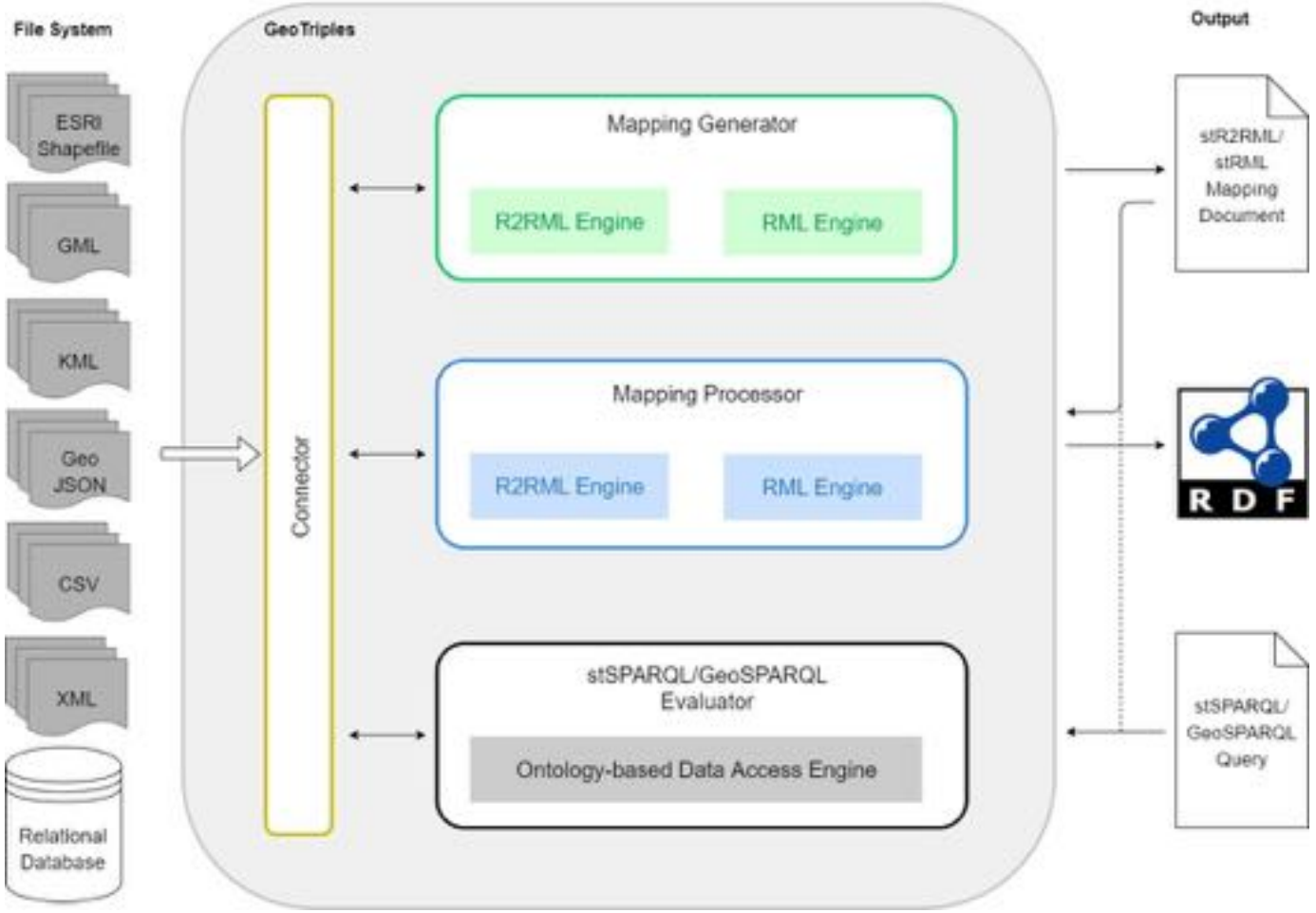
Mapping Processor

The mapping processor receives as input the mapping document. Based on the term maps, the Mapping Processor generates the final RDF graph, which can be manifested in any of the popular RDF syntaxes such as Turtle, RDF/XML, Notation3 or N-Triples.

stSPARQL/GeoSPARQL Evaluator

This component evaluates an stSPARQL/GeoSPARQL query over a relational database given an R2RML mapping.

GeoTriples System Architecture



Mappings Generation

- The mappings produced by GeoTriples consists of **two logical sources**:
  - non-geometric (thematic) data,
  - geospatial information
- The subjects are defined by combining a URI template with a unique identifier
- For each field of the input data source, GeoTriples generates an RDF predicate according to the name of the field and a predicate-object map
- The triples map that handles geospatial information contains a serialization of the geometric information according to the WKT format
- The generated RDF Graph will be compliant with the GeoSPARQL vocabulary

Mappings Generation

```
@prefix rr: <http://www.w3.org/ns/r2rml#>.
@prefix rml: <http://semweb.mmlab.be/ns/rml#> .
@prefix ql: <http://semweb.mmlab.be/ns/ql#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
@prefix rrx: <http://www.w3.org/ns/r2rml-ext#>.
@prefix rrxf: <http://www.w3.org/ns/r2rml-ext/functions/def/>.
@prefix ogc: <http://www.opengis.net/ont/geosparql#>.
@prefix onto: <http://example.di.uoa.gr/ontology#>.

<#shp_example_Geometry>
rml:logicalSource [
        rml:source "/path/to/shapefile/file.shp";
        rml:referenceFormulation ql:SHP;
        rml:iterator "shp_iterator";
];
rr:subjectMap [
        rr:template "http://example.di.uoa.gr/Geometry/{GeoTriplesID}";
        rr:class ogc:Geometry;
];
rr:predicateObjectMap [
        rr:predicateMap [ rr:constant ogc:asWKT ];
        rr:objectMap [
                rr:datatype  ogc:wktLiteral;
                rrx:function rrxf:asWKT;
                rrx:argumentMap ( [ rml:reference "geometry"; ] );
        ];
].
<#shp_example>
rml:logicalSource [
        rml:source "/path/to/shapefile/file.shp";
        rml:referenceFormulation ql:SHP;
        rml:iterator "gis_osm_natural_free_1";
];
rr:subjectMap [
        rr:template "http://example.di.uoa.gr/id/{GeoTriplesID}";
        rr:class onto:gis_osm_natural_free_1;
];
rr:predicateObjectMap [
        rr:predicateMap [ rr:constant onto:hasName ];
        rr:objectMap [
                rr:datatype  xsd:string;
                rml:reference "name";
        ];
];
rr:predicateObjectMap [
        rr:predicateMap [ rr:constant ogc:hasGeometry ];
        rr:objectMap [
                rr:template "http://example.di.uoa.gr/Geometry/{GeoTriplesID}";
        ];
].
```

Mapping Processor

- If the input mapping is an R2RML GeoTriples uses an extended version of D2RQ
- If the input mapping is an RML, GeoTriples uses an extended version of the iMinds processor.
- The transformation follows three main steps

1. Extracts the content of the logical source(e.g. using a SELECT query)
2. Defines Subject based on the Term map of the Subject map
3. Iterate the entities and forms the triples based on the Predicate object maps

GeoTriples-Spark

- We extended GeoTriples to run on top of Apache Spark
- It can run in standalone or in any cluster that supports Apache Spark
- GeoTriples-Spark is capable of transforming large amount of data into RDF graphs
- RML Mappings
- It supports the transformation of CSV, GeoJSON and ESRI Shapefiles

Apache Spark

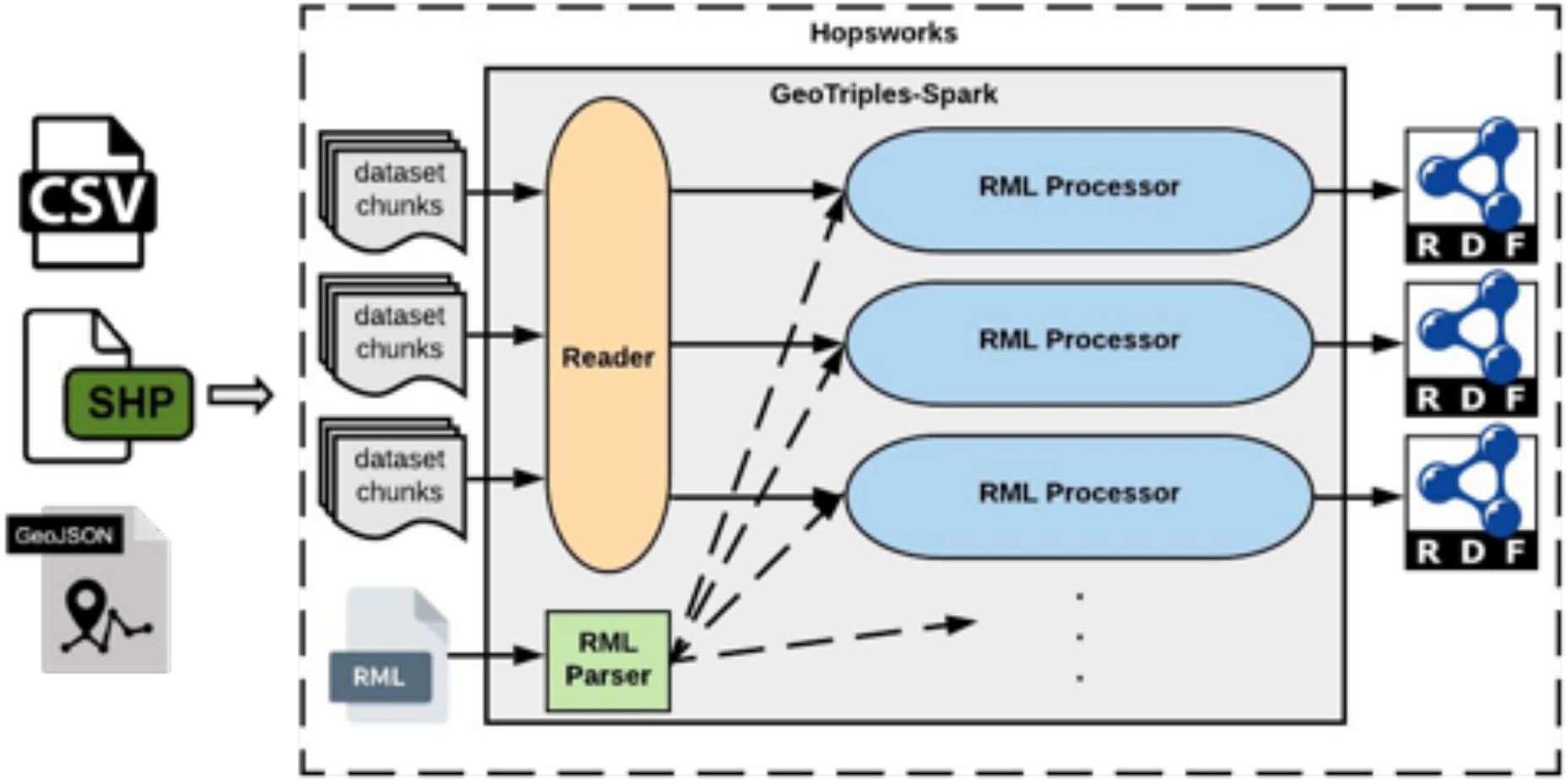
- Apache Spark is an open-source, distributed, general-purpose, cluster-computing framework.
- Spark uses a master/worker architecture.
- There is a Driver (JVM process) that talks to a single coordinator called Master which manages Workers in which Executors (JVM processes) run .

RDD

- **Resilient Distributed Dataset (RDD)** is an immutable distributed collection of elements of data.
- The data is partitioned across machines in the cluster, which can be operated in parallel using transformations and actions.
- **Dataframes** and **Datasets** are immutable collections of data in which is organized into named columns.

GeoTriples-Spark

Architecture



GeoTriples-Spark

- The Input data are loaded as multiple partitions, distributed across the cluster
- The RML mappings are extracted from the RML file and broadcasted to all servers
- Each partition is transformed into RDF triples by an RML processor
- The number of concurrent tasks is defined by the number of partitions and the number of the Executors (and their cores)
- There is no need for intermediate caching
- Except the broadcasting of the RML mappings, no further data shuffling occurs.
- Each partition is transformed independently from the rest.

GeoTriples-Spark

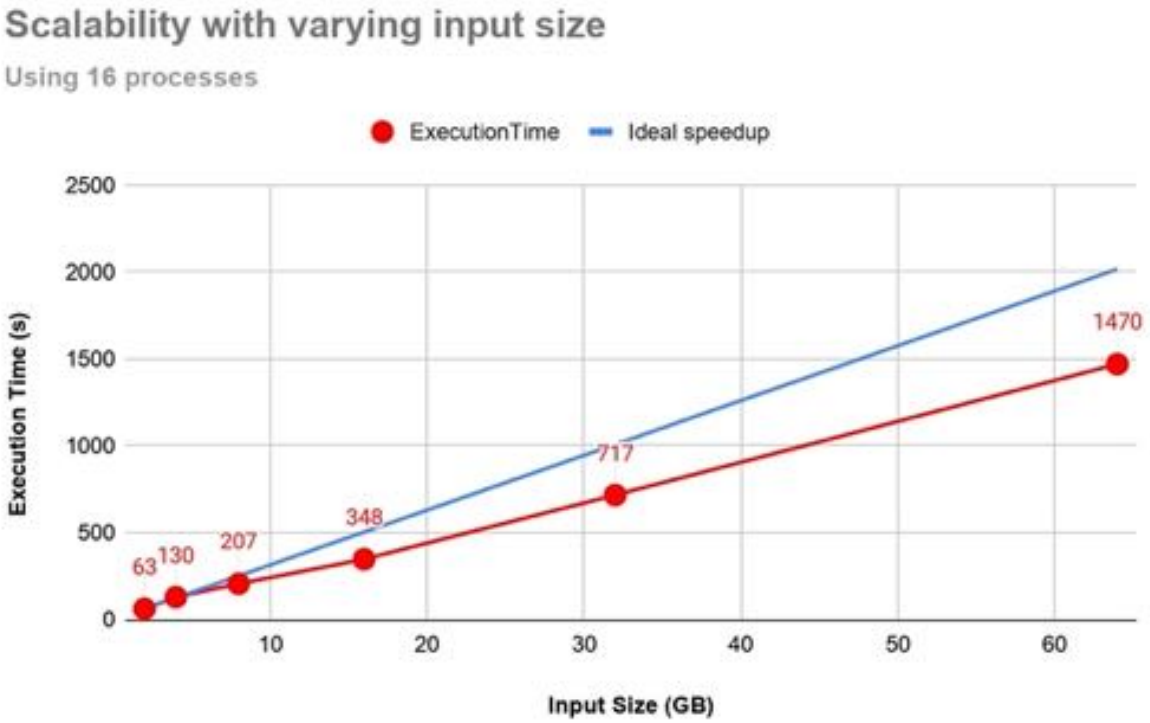
CSV: Transformation and Evaluation

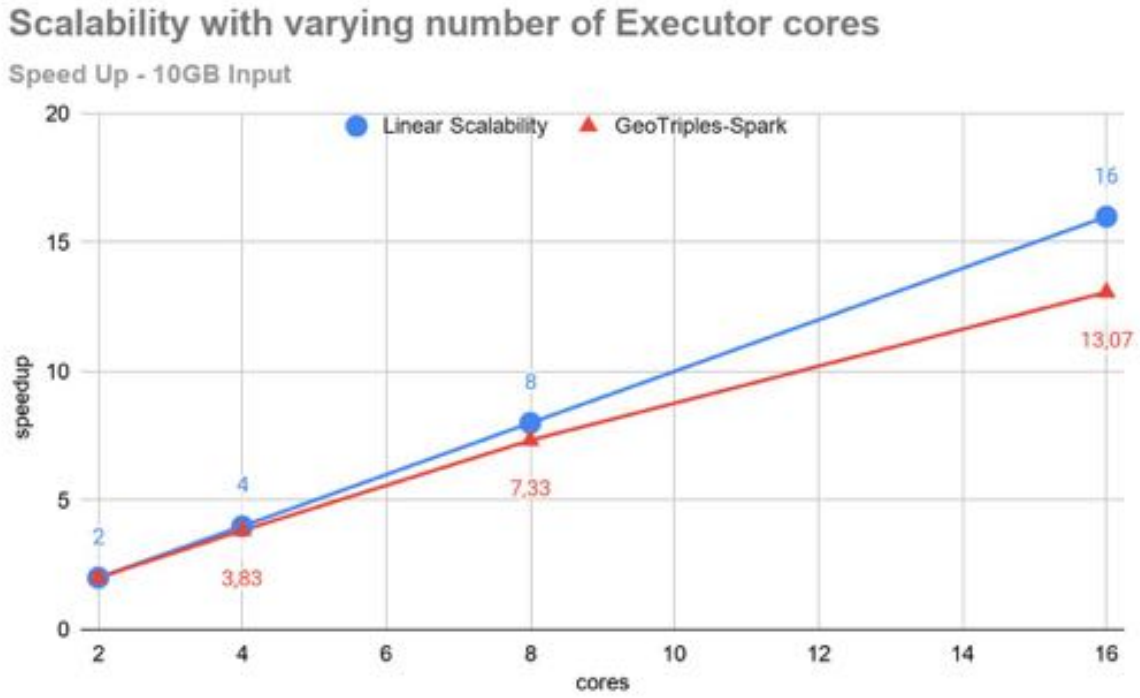
- CSV files are considered text files
- The geometry feature in CSV files is expected to be in Well Known Text (WKT)

Dataset	*Size*	*#Records*	*#Produced Triples*
1GB.csv	1GB	5M	58M
10GB.csv	10GB	52M	540M
100GB.csv	100GB	0.5B	5B
250GB.csv	250GB	1.3B	13B

Dataset	*Times loaded*	*Input Size*	*#Executors*	*Output Size*	*#Execution time (m)*
1GB.csv	1	1GB	2	7.7GB	1
10GB.csv	1	10GB	21	83.4GB	1.6
100GB.csv	1	100GB	41	840.1GB	3.3
250GB.csv	1	250GB	60	2.1**TB**	6.6
250GB.csv	2	500GB	65	4.1**TB**	13
250GB.csv	4	1**TB**	70	8.3**TB**	26
250GB.csv	8	2**TB**	80	16.6 **TB**	50

Scalability Experiments





GeoTriples-Spark

ESRI Shapefile: Transformation and Evaluation

- ESRI Shapefile is a file format for storing spatial data, and consist of multiple file (.shp, .shx, .dbf)
- **GeoSpark** is used for loading the input Shapefile into a Spark Dataset
- GeoSpark is an in-memory cluster computing framework developed by the Data Systems Lab, in order to support spatial data types, indexes, and processing of large-scale spatial data.
- It is important to mention that **GeoSpark always loads the input shapefile into a single partition** as it merges all the related component files of shapefile into one.
- Therefore, each Task have to transform a whole shapefile into RDF.

Jia Yu, Jinxuan Wu, Mohamed Sarwat: GeoSpark: a cluster computing framework for processing large-scale spatial data. SIGSPATIAL/GIS 2015

GeoTriples-Spark

ESRI Shapefile: Transformation and Evaluation

- There is a 2 GB size limit for any shapefile component file, which translates to a maximum of roughly 70 million point features.
- Therefore GeoTriples-Spark provides the option to load multiple shapefiles and transform them at once, by specifying a folder containing shapefile folders.
- GeoTriples-Spark loads each shapefile into an individual Spark Dataset and in the end it unites them into one.
- Otherwise, if the user want to transform a single big shapefile, it can repartition it into multiple partitions which will be transformed in parallel.

GeoTriples-Spark

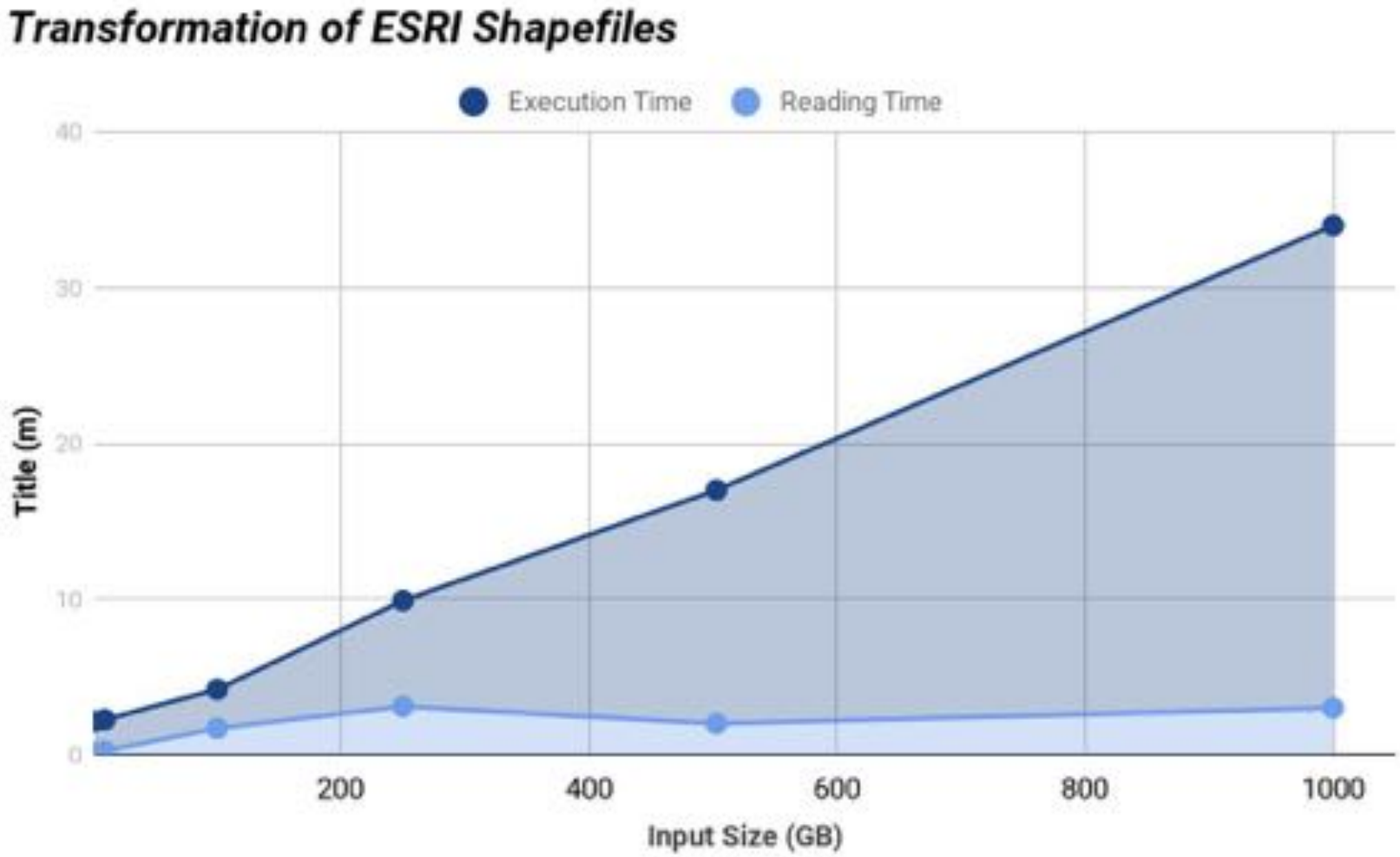
ESRI Shapefile: Transformation and Evaluation

Dataset	*.shp size*	*.dbf size*	*Total size*	*#Records*	*#Produced Triples*
RoadSystem_AUS	381.1MB	278.9MB	672.4 MB	1615868	17165376
RoadSystem_GER	1.7GB	1.9GB	3.7 GB	11107532	115146843

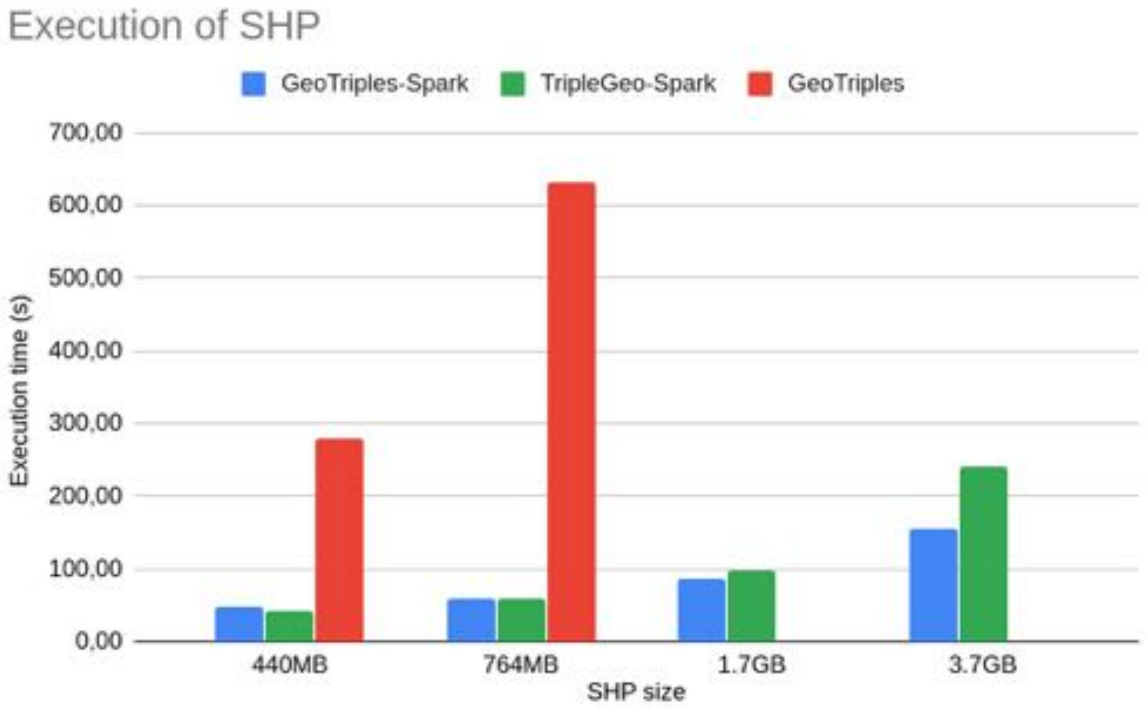
Dataset	Times	Input Size	#Executors	Output Size	Execution Time (m)
RoadSystem_AUS	2	1.3GB	1	5.5 GB	1.2
RoadSystem_AUS	16	9.8GB	3	41.6 GB	2.5
RoadSystem_AUS	153	100.5 GB	20	427.7 GB	4.3
RoadSystem_AUS	381	250.2 GB	30	1068.6 GB	9.9
RoadSystem_GER	136	502.9 GB	15	2.5 TB	17
RoadSystem_GER	258	1TB	27	5.1 TB	34

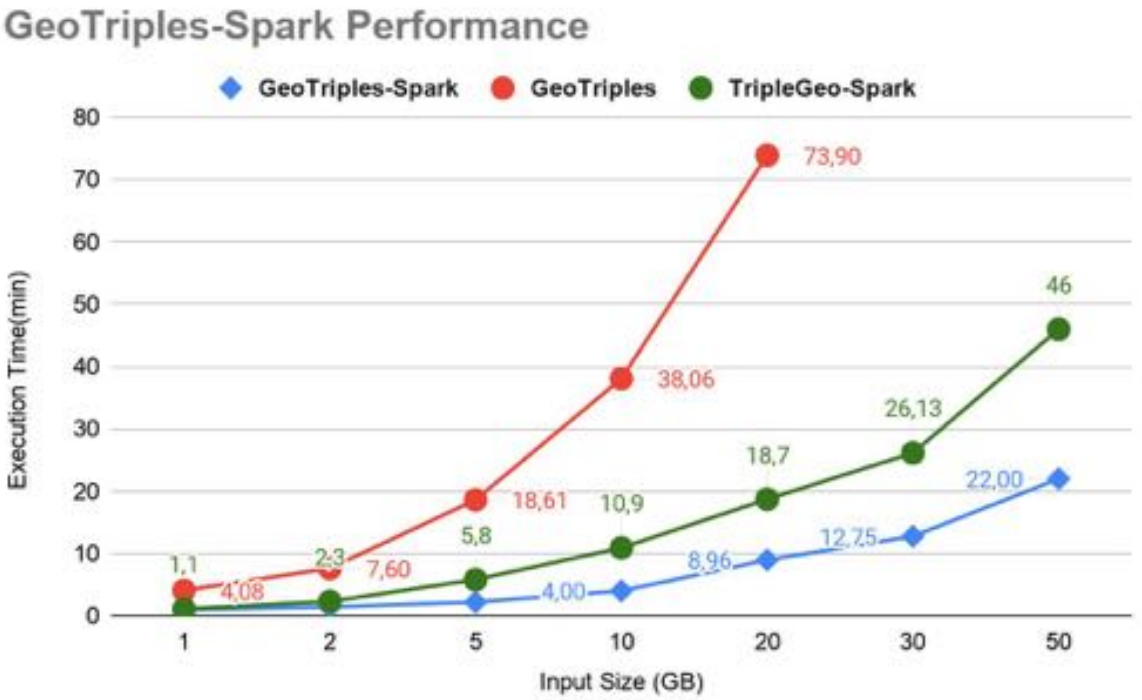
GeoTriples-Spark

ESRI Shapefile: Transformation and Evaluation



GeoTriples-Spark and TripleGeo-Spark





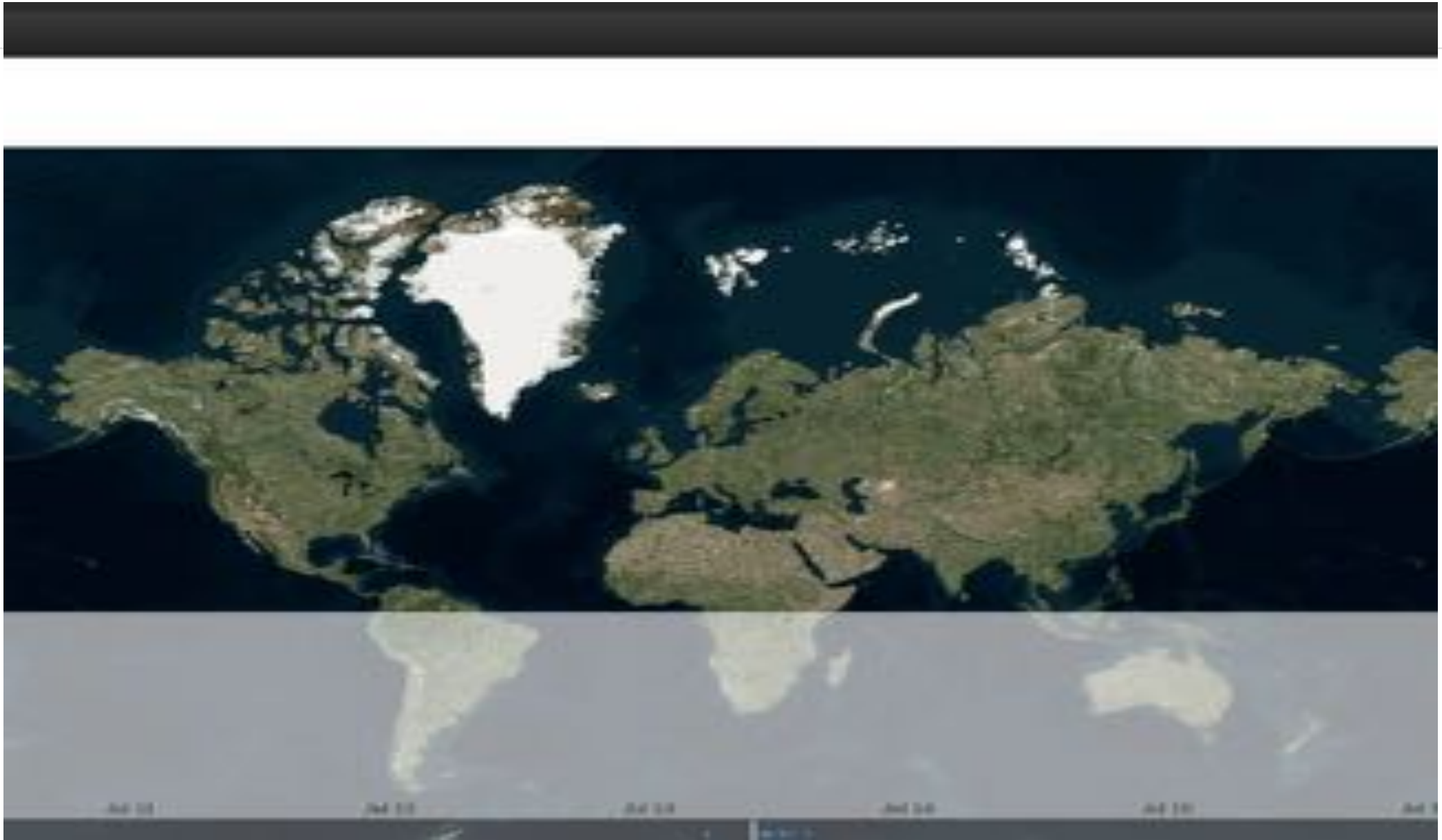
K Patroumpas, D Skoutas, et al., Exposing Points of Interest as Linked Geospatial Data. *SSTD 2019*

Thank you!

Questions ?

Interlinking geospatial RDF data

Geospatial Interlinking in action



Geospatial Interlinking

Input:

- A topological relation ***R***
- A source dataset of geometries ***S***
- A target dataset of geometries ***T***

  Types of Geometries:

  - LineStrings
  - Polygons

Output:

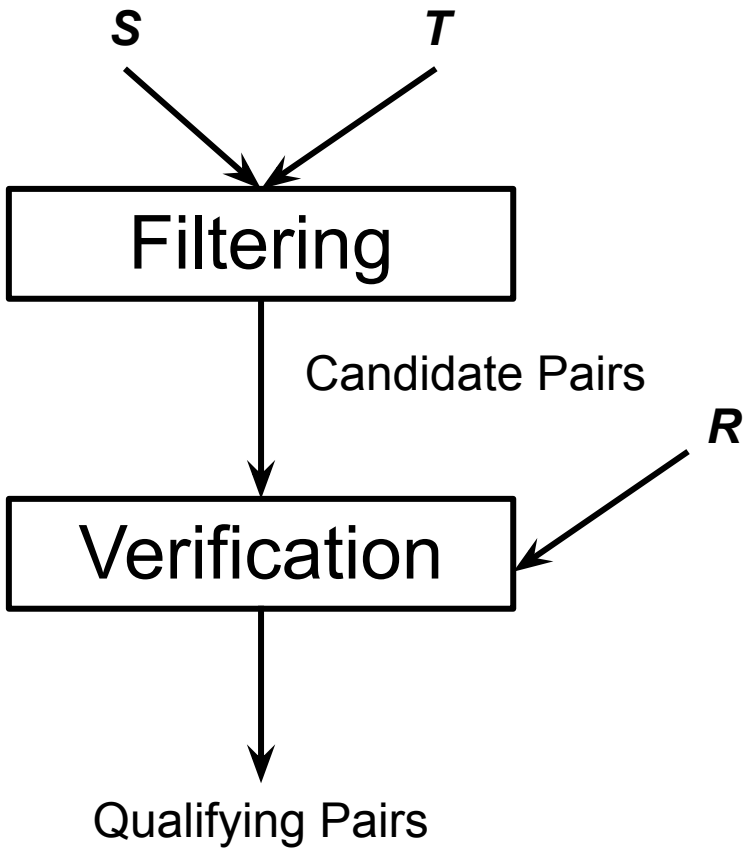
- All pairs (s,t) ∈ ***S*** x ***T*** such that ***R(s,t) = true***

Challenges:

- quadratic time complexity, $\mathbf{O(n^2)}$
- time-consuming topological relations over complex geometries

Filtering – Verification Framework

Two-step procedure to reduce the quadratic time complexity:



Filtering, a.k.a. Space Tiling

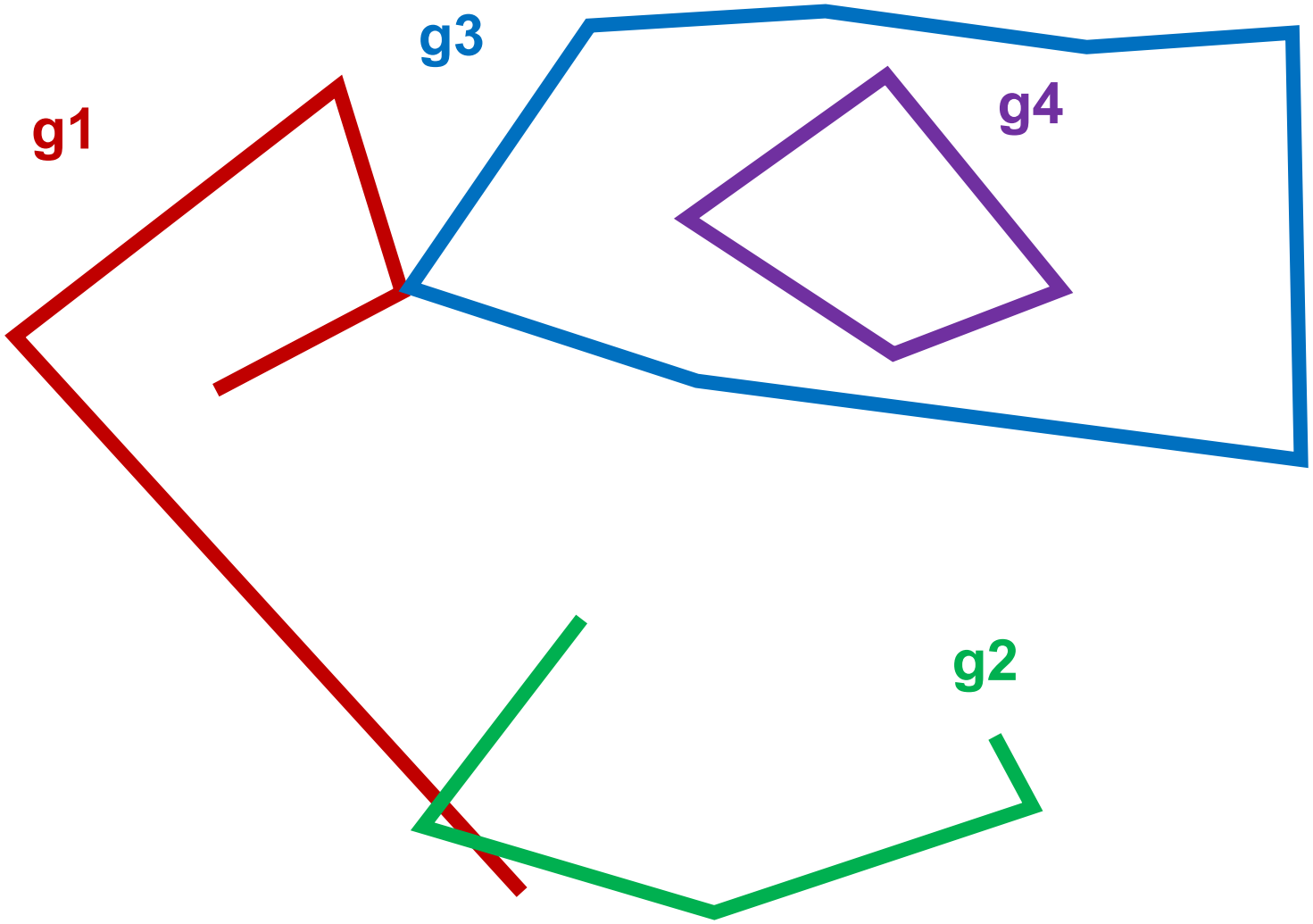
Involves three steps:

1. We define an ***Equigrid*** on Earth's surface
2. We index geometries according to their ***Minimum Bounding Rectangle***
3. We define as ***candidate pairs*** only the geometries that share at least one tile

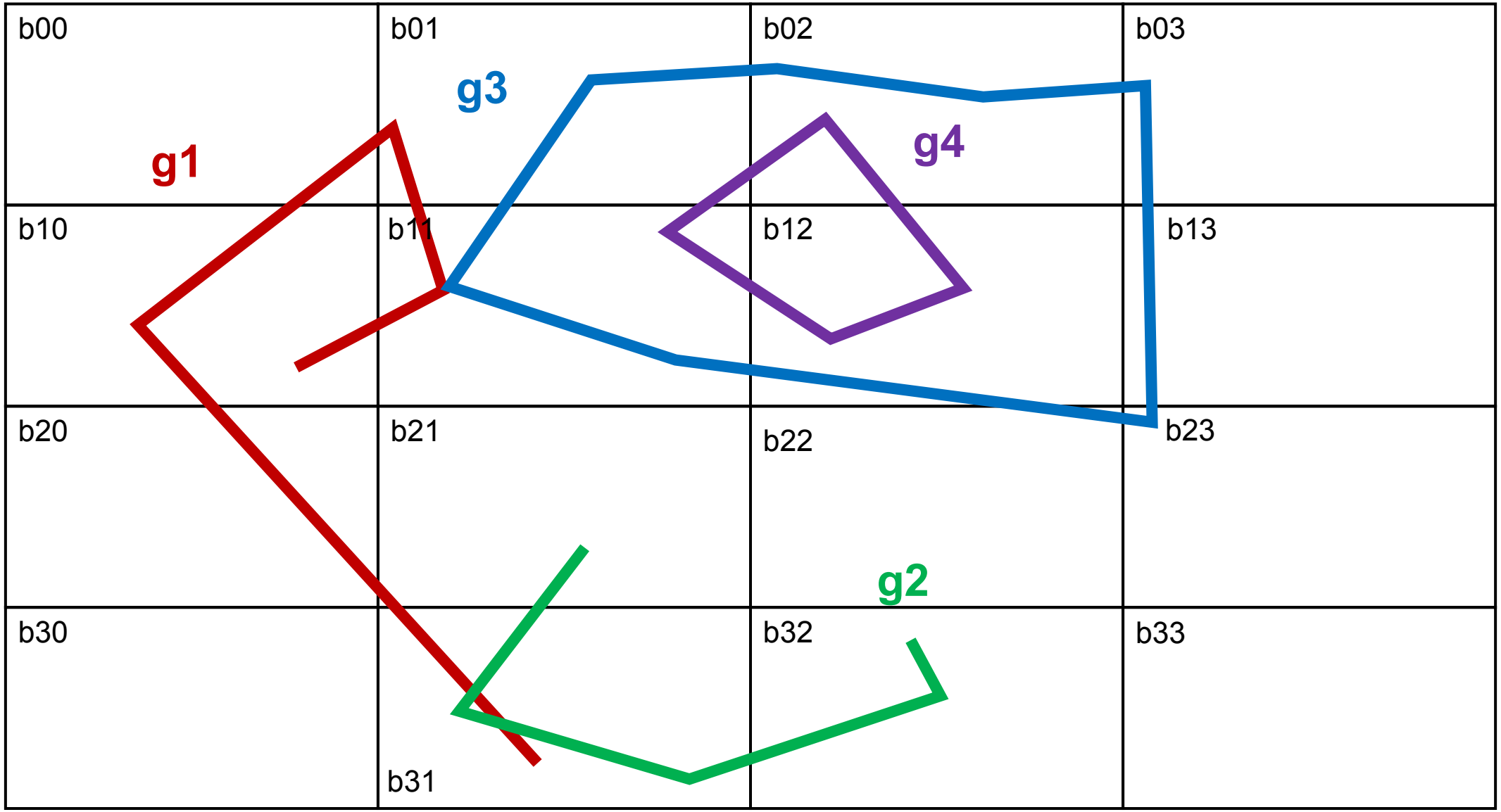
Advantages:

- Exact process
- Linear time complexity O(n)
- Significant gains in efficiency

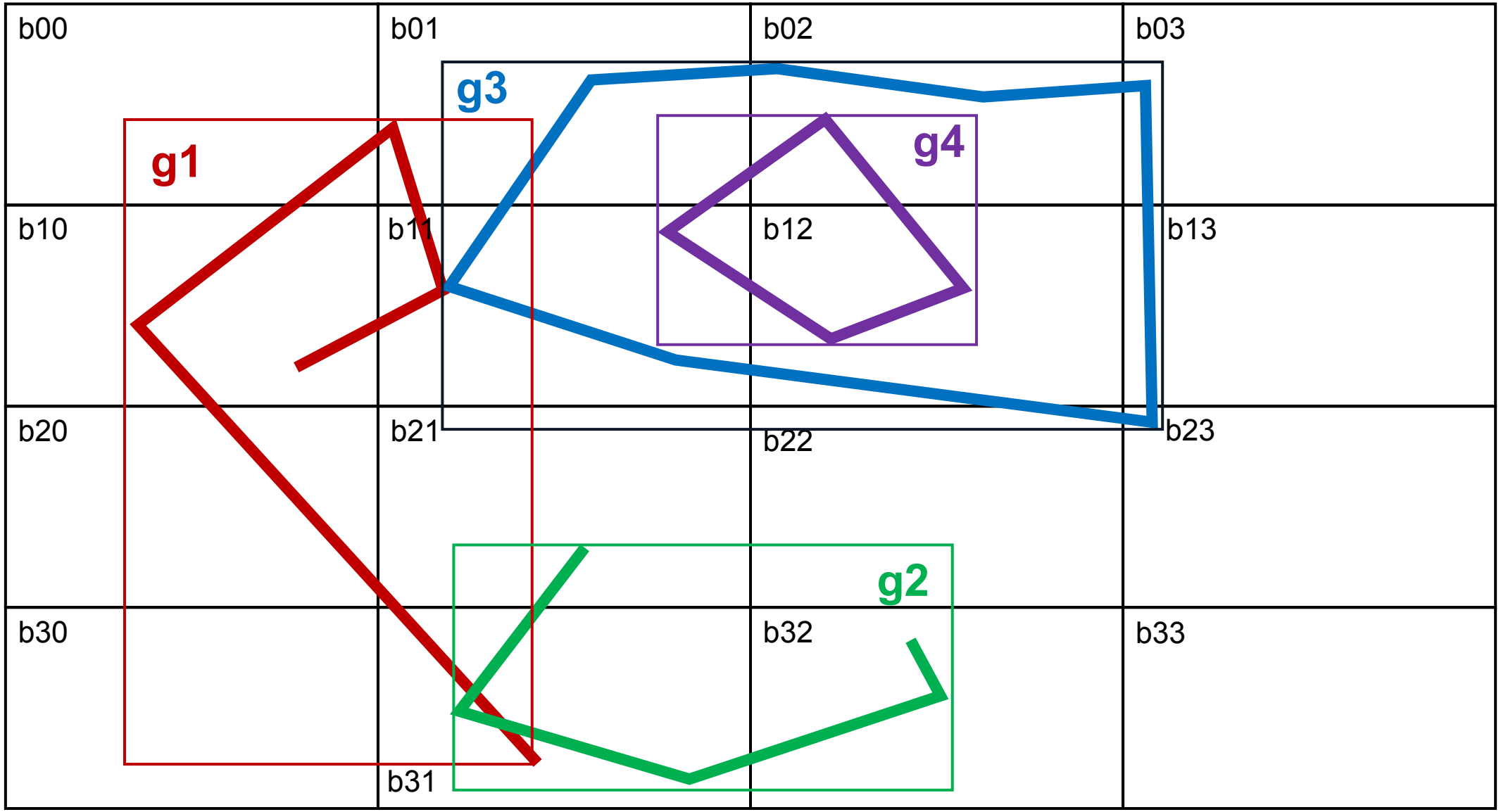
Space Tiling Example



Space Tiling Example - Equigrid



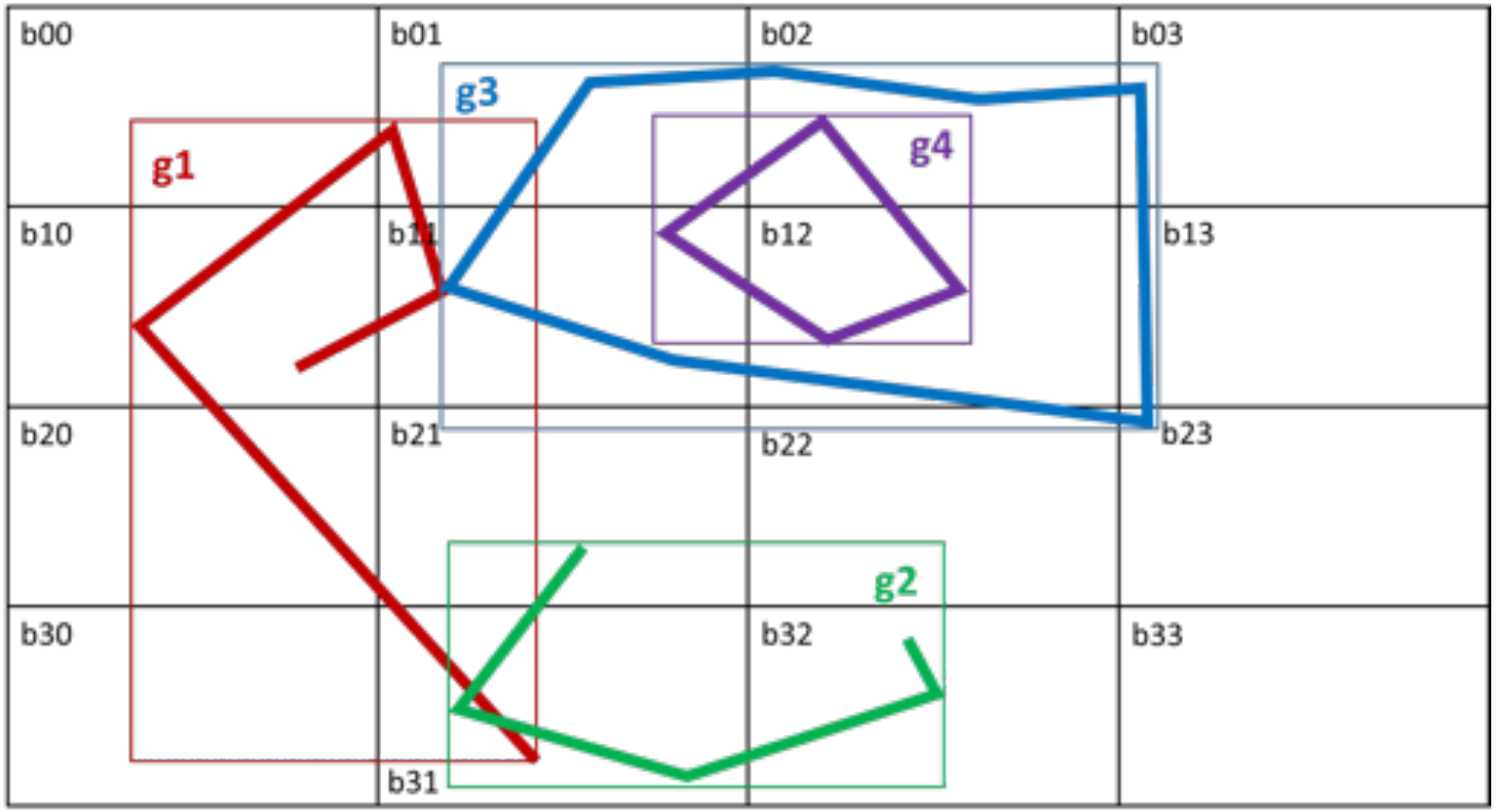
Space Tiling Example - MBR indexing



Space Tiling Example – Candidate Pairs

Just **3** pairs:
$g_1 - g_2$
$g_1 - g_3$
$g_3 - g_4$

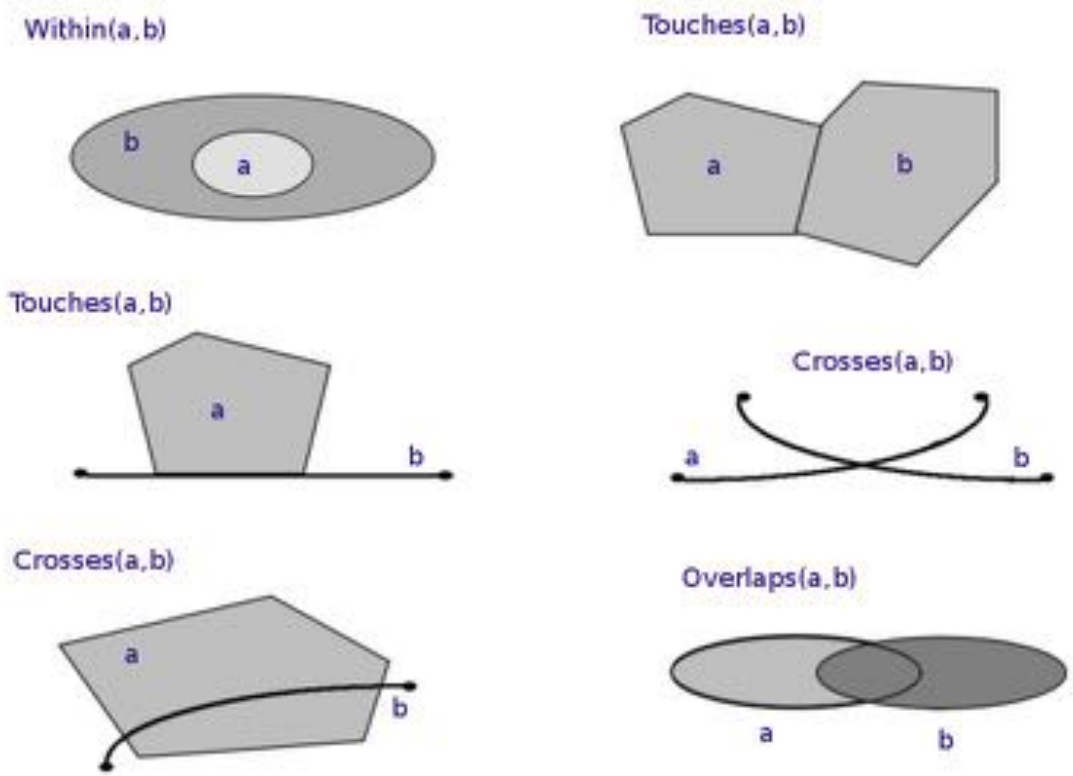
50% lower than the **6** pairs
of the brute-force approach.



Verification

Two different types:

1. Proximity relations (such as `dbp:near`) with a distance threshold
   - e.g., find all cities from ***S*** that are less than 1km away from any river in ***T***
2. Topological relations according to the Dimensionally Extended 9-Intersection Model (**DE9IM**)
   - Equals
   - Touches
   - Contains
   - Covers
   - Intersects
   - Within
   - CoveredBy
   - Crosses
   - Overlaps
   - Disjoint



ORCHID

Filtering:

- **Static** space tiling
- Granularity for width and height = **θ / R / a**
  - a = 1

Verification:

- Hausdorff distance $hd(s,t) = max_{si \in S}\{min_{ti \in T}\{\delta(s_i,t_j)\}\} <= \theta$
- Optimizations for efficient computation:
  - Bounding circles
  - Cauchy-Swarz Inequality for Distance Approximation

Open-source implementation (https://github.com/dice-group/LIMES)

Axel-Cyrille Ngonga Ngomo: ORCHID - Reduction-Ratio-Optimal Computation of Geo-spatial Distances for Link Discovery. International Semantic Web Conference, 2013: 395-410

Silk-spatial

Filtering:

- **Static** space tiling
- Granularity for width and height = $\mathbf{1/a^{o\ 2}}$
  - a = 10

Verification:

- DE9IM topological relations – single relation per run
- Massive parallelization (Apache Hadoop)

Open-source implementation (https://github.com/silk-framework/silk)

Panayiotis Smeros, Manolis Koubarakis: Discovering Spatial and Temporal Links among RDF Data. LDOW@WWW 2016

RADON

Filtering:

- Swapping strategy
- **Dynamic** space tiling
  - Width = ½ · (average$_{s \in S}$(s.width) + average$_{t \in T}$(t.width))
  - Length = ½ · (average$_{s \in S}$(s.length) + average$_{t \in T}$(t.length))

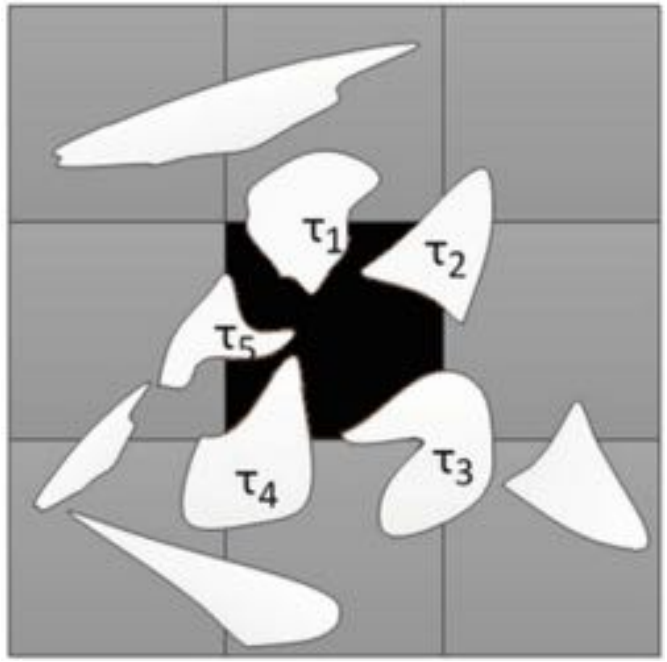
Verification:

- DE9IM topological relations – single relation per run
  - Relation-based optimizations
  - Hash-based redundancy elimination
- Multi-core parallelization

Open-source implementation (https://github.com/dice-group/LIMES)

Mohamed Ahmed Sherif, Kevin Dreßler, Panayiotis Smeros, Axel-Cyrille Ngonga Ngomo: Radon - Rapid Discovery of Topological Relations. AAAI 2017: 175-181

stLD



Filtering:

- **Static** Index
- Variety of approaches (e.g., R-Trees, Equigrid, Hierarchical Grid)
- Indexes exclusively the source dataset ***S***
- MaskLink algorithm

Verification:

- Both topological and proximity relations – single relation per run
- Massive parallelization (Apache Flink)
- Suitable for streams

Implementation not available.

Georgios M. Santipantakis, Apostolos Glenis, Christos Doulkeridis, Akrivi Vlachou, George A. Vouros: stLD: towards a spatio-temporal link discovery framework. SBD@SIGMOD 2019: 4:1-4:6
Georgios M. Santipantakis, Christos Doulkeridis, George A. Vouros, Akrivi Vlachou: MaskLink: Efficient Link Discovery for Spatial Relations via Masking Areas. CoRR abs/1803.01135 (2018)

GIA.nt: Geospatial Interlinking At large – Part A

Improving RADON's **Filtering**:

- **Dynamic** space tiling, based exclusively on the source dataset **S**
  - Width = average$_{s \in S}$(s.width)
  - Length = average$_{s \in S}$(s.length)
- No dataset swapping

Lower running time

- Target dataset (=largest input dataset) read one by one from the **disk**
- Inherent removal of redundant (i.e., repeated) geometry pairs

Memory requirements lower by >50%

- Easily parallelizable in MapReduce, due to its geometry-centric functionality

George Papadakis, Georgios M. Mandilaras, Nikos Mamoulis, Manolis Koubarakis. Progressive, Holistic Geospatial Interlinking. WWW 2021: 833-844

GIA.nt: Geospatial Interlinking At large – Part B

Improving RADON's **Verification**:

- **Holistic** Geospatial Interlinking:

  Simultaneous estimation of all DE9IM topological relations → **Intersection Matrix**

Run-time lower by **>80%**

$$\mathrm{DE9IM}(a,b) = \begin{bmatrix} \dim(I(a) \cap I(b)) & \dim(I(a) \cap B(b)) & \dim(I(a) \cap E(b)) \\ \dim(B(a) \cap I(b)) & \dim(B(a) \cap B(b)) & \dim(B(a) \cap E(b)) \\ \dim(E(a) \cap I(b)) & \dim(E(a) \cap B(b)) & \dim(E(a) \cap E(b)) \end{bmatrix}$$

Examples:

Equals	$\begin{bmatrix} T & * & F \\ * & * & F \\ F & F & * \end{bmatrix}$
	T*F**FFF*

Intersects	$\begin{bmatrix} T & * & * \\ * & * & * \\ * & * & * \end{bmatrix}$	$\begin{bmatrix} * & T & * \\ * & * & * \\ * & * & * \end{bmatrix}$	$\begin{bmatrix} * & * & * \\ T & * & * \\ * & * & * \end{bmatrix}$	$\begin{bmatrix} * & * & * \\ * & T & * \\ * & * & * \end{bmatrix}$
	T********	*T*******	***T*****	****T****

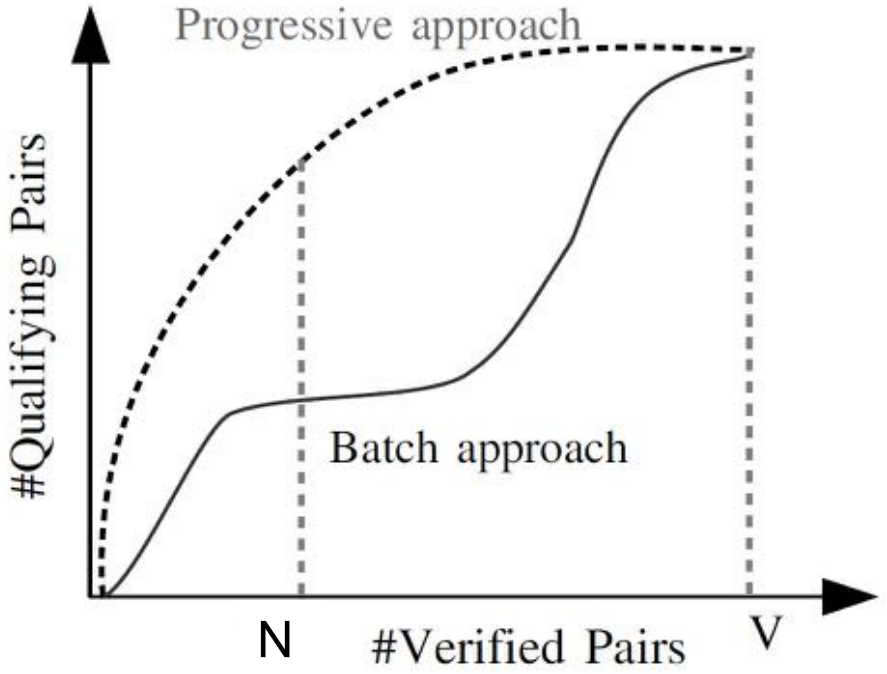
Progressive Geospatial Interlinking

Ideal for applications with limited resources:

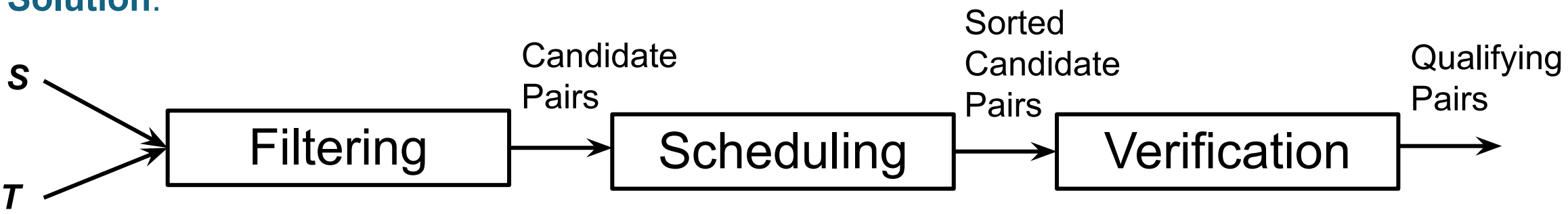
- Temporal or computational (e.g., Amazon Lambda functions)

Requirements with respect to batch approaches [1]:

1. Same Eventual Quality
2. Improved Early Quality
   - Measured through Progressive Geometry Recall (PGR)



Solution:



[1] *Steven Euijong Whang, David Marmaros, Hector Garcia-Molina: Pay-As-You-Go Entity Resolution. IEEE Trans. Knowl. Data Eng. 25(5): 1111-1124 (2013)*

Progressive GIA.nt

Input:

- Budget B + source dataset + target dataset

Filtering:

- Same as batch GIA.nt

Scheduling:

- Priority queue with top-B weighted candidate pairs based on either of the following functions:
  - Co-occurrence Frequency (CF): #common tiles
  - Jaccard Similarity (JS): normalized CF
  - Pearson's χ^2 test (χ^2): degree to which ***s*** and ***t*** occur independently in tiles

higher scores → more likely to satisfy at least one topological relation

Verification:

- Processes the pairs of the priority queue in decreasing weight

George Papadakis, Georgios M. Mandilaras, Nikos Mamoulis, Manolis Koubarakis. Progressive, Holistic Geospatial Interlinking. WWW 2021: 833-844

Dynamic Progressive Geospatial Interlinking

Improved Static Progressive Geospatial Interlinking in three ways:

1. New weighting schemes based on the complexity of geometries.
   - Normalized MBR overlap → higher effectiveness

$$MBR(s,t) = \frac{MBR(s \cap t)}{MBR(s \cup t)} = \frac{MBR(s \cap t)}{MBR(s) + MBR(t) - MBR(s \cap t)}$$

   - Inverse sum of points → higher time efficiency

$ISP(s,t) = \frac{1}{p(s)+p(t)}$, where p(g) denotes the number of boundary points

2. Composite weighting schemes → higher effectiveness, more deterministic behavior
   - the primary one is used for scheduling all pairs
   - the secondary one is used for resolving the ties

3. Dynamic Progressive GIA.nt

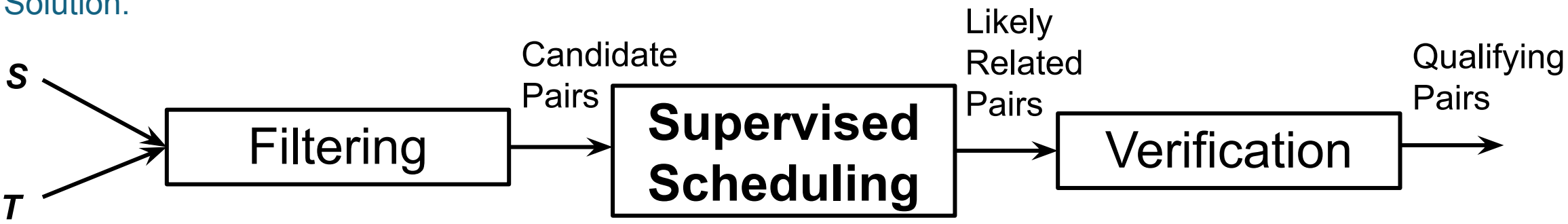
George Papadakis, Georgios M. Mandilaras, Nikos Mamoulis, Manolis Koubarakis. Static and Dynamic Progressive Geospatial Interlinking. ACM TSAS (to appear)

Supervised Progressive Geospatial Interlinking

Drawbacks of Progressive Geospatial Interlinking:

- Store the top-BU weighting pairs in main memory
- Might be hard to fine-tune BU
- Considers at most two sources of evidence, i.e., composite weighting schemes

Solution:



1. Filtering → as in (Progressive) GIA.nt
2. Supervised Filtering
   - Classify candidate pairs into "`likely related pairs`" & "`unlikely related pairs`" using a feature vector
3. Verification → as in Batch GIA.nt

Supervised Filtering

Challenges:

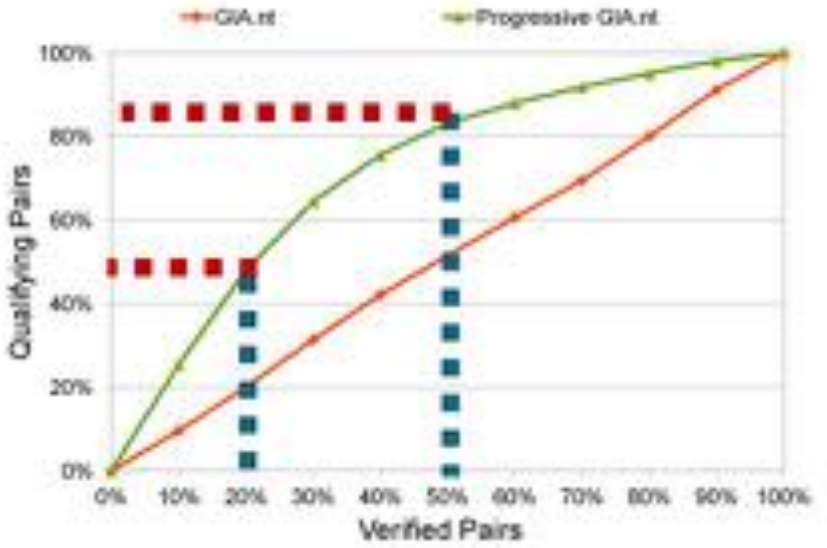
- Define generic, effective & efficient features
- Avoid any human intervention
- Address class imbalance
- Minimize the feature and the training set → simple & efficient classification models

Approach outline:

- **Self-supervised learning based on undersampling**
- 4 categories of features
    1. Area-based (source/target/intersection MBR area)
    2. Boundary-based (source/target #boundary points and boundary length)
    3. Grid-based (#common tiles, #tiles intersecting the target MBR)
    4. Candidate-based (total/distinct/real candidates per source/target geometry)
- 2 sub-categories in each case:
    - Atomic features
    - Composite features

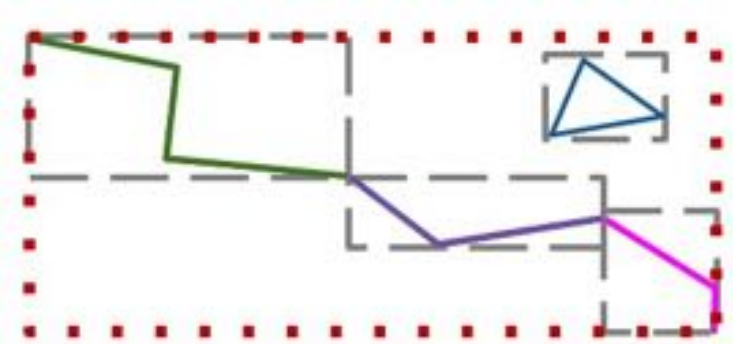
Future directions

- Proactive Geospatial Interlinking
  - Terminate Geospatial Interlinking automatically as soon as recall exceeds a desired level → minimize the time required for processing voluminous datasets



- Generalize to 3-dimensional data
  - Silk-spatial: 3rd dimension = time
  - stLD: 3rd dimension = height (e.g,. aviation data)

- Improve Intersection Matrix computation
  - O(n · logn) [1]
  - Fine-grained MBR



[1] *Edward P. F. Chan, Jimmy N. H. Ng: A General and Efficient Implementation of Geometric Operators and Predicates. SSD 1997: 69-93*

Thank you!

Questions?

Geospatial knowledge graphs

Knowledge Graphs

Knowledge graphs:

- Directed graphs
- Represent knowledge about:
  - World objects (nodes of the graph)
  - Relationships among world objects (the edges of the graph).

Geospatial Knowledge Graphs

- **Geospatial entities** are (abstractions of) real-world objects that have **spatial** and **non-spatial characteristics**.

- A **knowledge graph** will be called **geospatial** if it has *at least one* (but usually many) **geospatial entity**

- **Knowledge about the spatial attributes of a feature can be:**
  - Quantitative (distance between Athens and its airport)
    - Calculated using geometries
  - Qualitative (Athens is in the region of Attica)
    - Qualitative binary relations

Knowledge Graphs

- DBpedia
- Wikidata
- GeoNames
- YAGO2
- YAGO2geo
- YAGO4
- WorldKG
- KnowWhereGraph

Encyclopedic (+ geospatial knowledge)

DBpedia

- Open & Free KG (started in 2007!)
- Structured information from **Wikipedia articles** (infoboxes)
- **Multiple languages** (but different content)
- DBpedia (English): 1.6B RDF triples, 6.6M entities
- Automatically updated
- **Geospatial Information:**
  - 1.9 million **coordinate** pairs (centers of cities, towns, etc)
  - **Thematic attributes** (neighboring countries, cities of countries, etc)
  - Some **cardinal** directions (e.g., `dbp:north`)

S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives. 2007. DBpedia: A nucleus for a web of open data. In The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007, pp. 722–735.

DBpedia


```
dbr:Lamia_(city) dbp:name "Lamia";
                 rdf:type dbo:AdministrativeRegion ;
                 geo:lat "38.900002"^^xsd:float ;
                 geo:long "22.433332"^^xsd:float ;
                 …
                 dbo:country dbr:Greece ;
                 dbp:periph dbr:Central_Greece_(region) .
```

Wikidata

- Open & Free KG (started in 2012)
- Successor of Freebase
- >100M entities
- Crowdsourced (500 edits/min by community members)
- High quality structured data in real time
- Multilingual (same content)
- **Geospatial Information:**
  - 7M triples containing coordinate information
  - 20K geometries (polygons, multipolygons)
  - Topological information (e.g., neighboring countries, cities of countries)

D. Vrandecic and M. Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. Commun. ACM, 57(10): 78–85.

Wikidata

- **Geospatial Information:**
  - Properties:
    - coordinate location (`wd:P625`)
    - geoshape (`wd:P3896`)
  - Data types:
    - GlobeCoordinate (`wd:Q29934236`)
    - GeoShape (`wd:Q42742911`)

Wikidata


```
<Lamia_Municipality> <instance_of> <municipality_of_Greece>;
                <country>  <Greece> ;
                <population> "75315" ;
                <headquarters_location> geo:38.866389,22.367222 .
```

GeoNames Gazetteer

- A **freely** available geographical database.
- **7 M geographical names** in various languages.
- 9 feature classes + 645 feature codes:
  - Administrative Boundary Features, Hydrographic Features , Area Features, Populated Place Features, Road / Railroad Features, Spot Features, Hypsographic Features, Vegetation Features, Undersea Features
- All lat/long coordinates are in **WGS84**.
- **Users** may manually **edit**, correct and add new names.
- Topological Information:
  - **Partonomic** relations (e.g., Berlin is located in Germany is located in Europe)
  - **Neighboring countries** for each location.

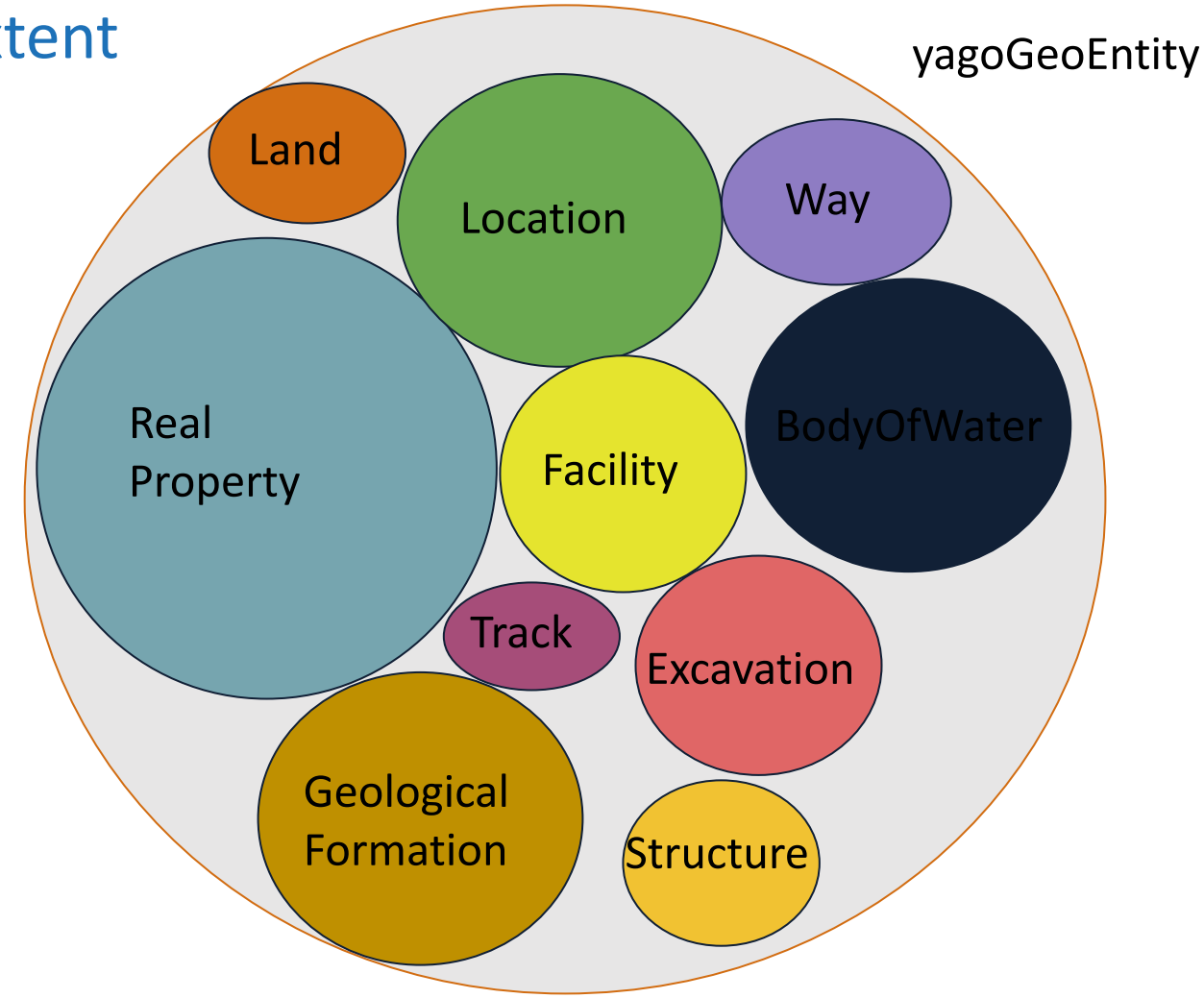
The YAGO 2 Knowledge Graph

- **YAGO:**
  - Wikipedia: Entities and their types
  - WordNet Hierarchy
  - **S**ubject, **P**redicate, **O**bject
- **YAGO2**:
  - Spatial Knowledge
    - Geonames
  - Temporal Knowledge
    - Wikipedia
  - 447M facts, 9.8M entities
    - Location existence: 30% of entities
    - Time existence: 47% of entities
  - SPOTL model:
    - **S**ubject, **P**redicate, **O**bject, **T**ime and **L**ocation

Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, Gerhard Weikum, YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia, Artificial Intelligence, Volume 194, 2013, p. 28-61.

The YAGO 2 Knowledge Graph

Entities that have a permanent spatial extent on Earth.



The YAGO 2 Knowledge Graph

```
@base <http://yago-knowledge.org/resource/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

<Auckland> rdfs:label "Auckland"@eng .
<Auckland> <isLocatedIn> <New_Zealand> .
<Auckland> <hasLatitude> "-36.840555555555556"^^<degrees>.
<Auckland> <hasLongitude> "174.73999999999998"^^<degrees>.
```

YAGO2geo: Motivation & Goal

- Motivation
  - Not enough geospatial knowledge in current KGs
  - Importance of geospatial data (e.g., GoogleMaps)
- Goal
  - Extend YAGO2 with precise geospatial knowledge (polygons, lines)
  - Combine multiple data sources
  - Follow OGC standards

Nikolaos Karalis, Georgios Mandilaras, and Manolis Koubarakis. Extending the YAGO2 Knowledge Graph with Precise Geospatial Knowledge, ISWC, Auckland, New Zealand, 26-30 October, 2019

Data Sources

- Official Administrative Data
  - Greek administrative data
    - Kallikratis: Greek Administrative Geography dataset
    - Kapodistrias (Source: geodata.gov.gr)
  - Ordnance Survey (England, Scotland, Wales)
  - Ordnance Survey Northern Ireland
  - Ordnance Survey Ireland
  - National Boundary Dataset: USGS Governmental Unit Boundaries dataset from the National Map represents major civil areas for the U.S

Data Sources (cont'd)

- Unofficial Administrative Data
  - Global Administrative Areas (GADM)
    - Administrative areas of every country
    - Six administrative layers
    - Over 386,000 administrative areas
    - May 2018

Data Sources (cont'd)

- Crowdsourced Knowledge: OpenStreetMap (OSM)
  - Volunteer, crowdsourced project
  - Free map of the world
  - Multiple geographical features
    - Natural (e.g., lakes)
    - Waterways (e.g., streams)
    - Places (e.g., cities)
    - Land Use (e.g., forests)
  - Data provided by Geofabrik

Geospatial Knowledge in YAGO2geo KG

	DBPedia	Wikidata	YAGO2	YAGO2geo
Coordinates	1M	7.2M	12M	12M
Lines and Polygons (Shapes)	-	23K	-	3,8M Linestrings 703K Polygons and Multipolygons

- OGC geometries

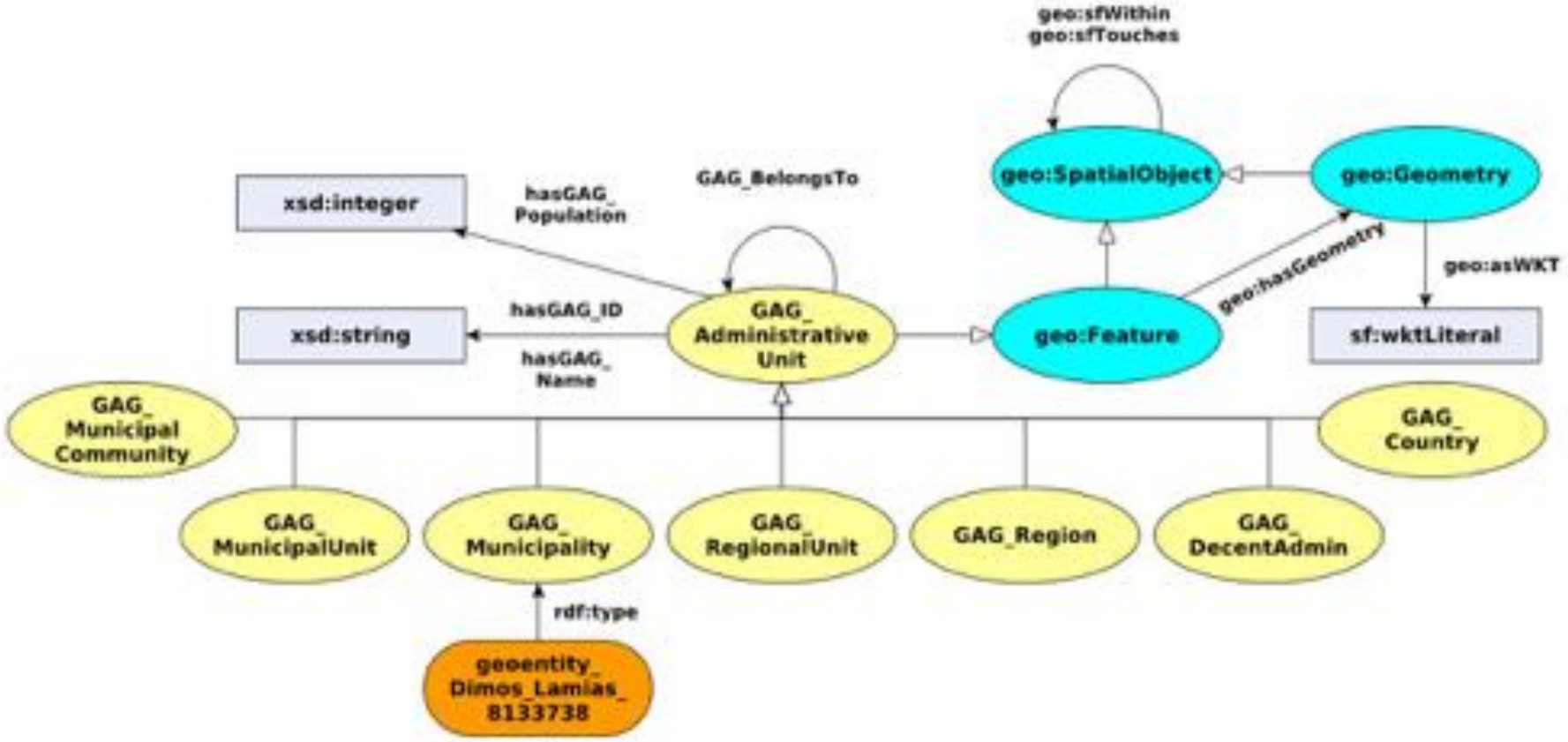
```
<Oxfordshire> geo:hasGeometry y2geo:Geometry_OS_8328.
y2geo:Geometry_OS_8328 geo:asWKT "POLYGON((...))" .
<geoentity_Dimos_Athens_8133876> geo:hasGeometry y2geo:Geometry_gag_9186.
y2geo:Geometry_gag_9186 geo:asWKT "MULTIPOLYGON(((...)))" .
```

- Topological relations, OGC vocabulary
  - Precise geometries allow us to extract additional information

```
<Oxford> geo:sfWithin <Oxfordshire> .
<geoentity_Dimos_Athens_8133876> geo:sfTouches <Kaisariani> .
```

YAGO2geo Ontology

- Extend YAGO2 ontology in order to support new geospatial knowledge
- Example: Greek Administrative Geography ontology in YAGO2geo



Geospatial Knowledge in YAGO2geo

- **Question**: Find the geometries of water bodies and forests that are within Saarland
- **Answer**:

Temporal Knowledge in YAGO2geo KG

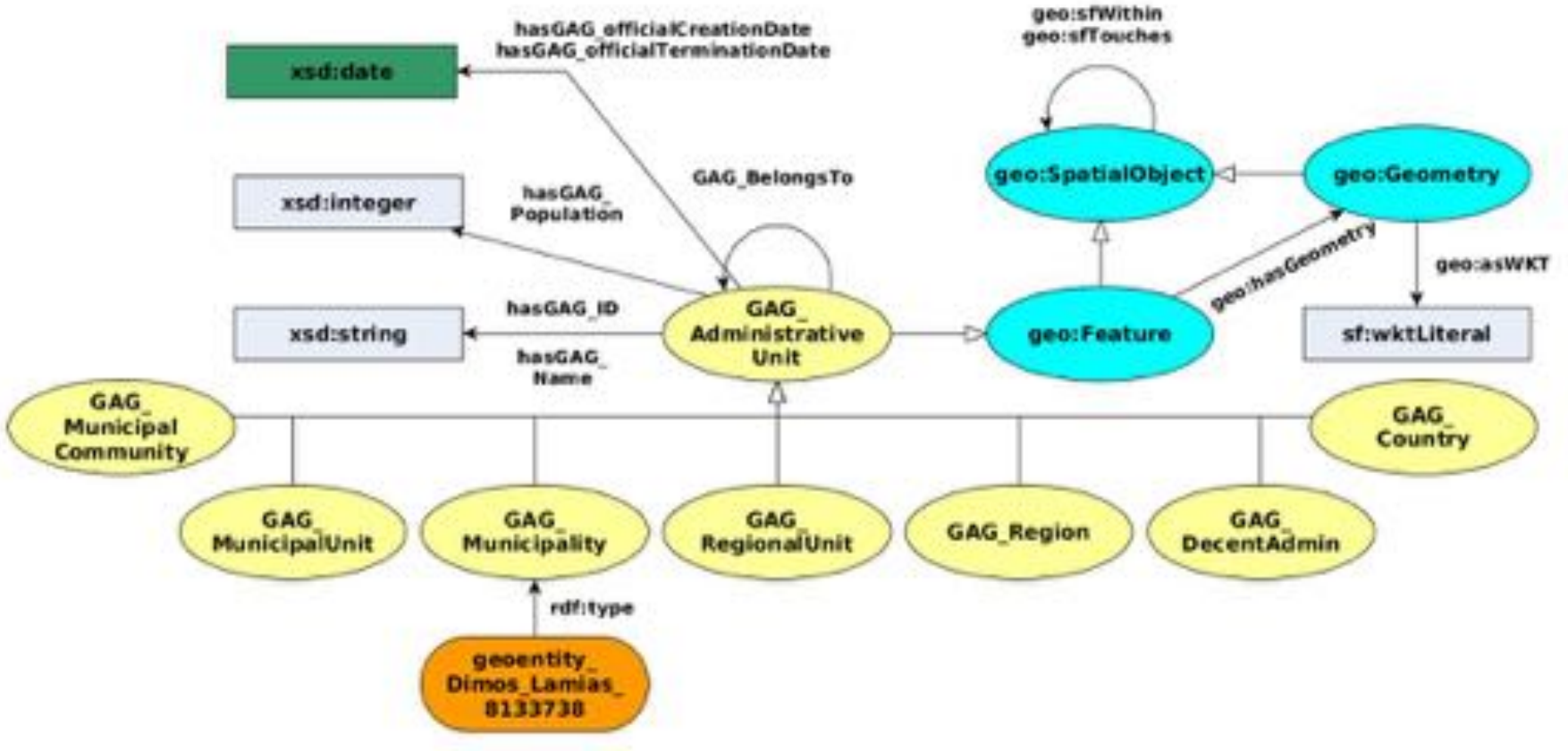
- Official temporal information for Greek administrative units
  - GAG dataset
  - Kapodistrias dataset

Temporal Knowledge in YAGO2geo KG

```
<geoentity_Nomarchía_Athínas_445408> yago2geo:hasKapo_OfficialCreationDate
                                                    "1997-##-##"^^xsd:date.
<geoentity_Nomarchía_Athínas_445408> yago2geo:hasKapo_OfficialTerminationDate
                                                    "2011-##-##"^^xsd:date.
<Magnesia_Prefecture> <wasCreatedOnDate> "1899-##-##"^^xsd:date  % existing fact
<Magnesia_Prefecture> yago2geo:hasKapo_OfficialCreationDate "1997-##-##"^^xsd:date.
<Magnesia_Prefecture> yago2geo:hasKapo_OfficialTerminationDate 2011-##-##"^^xsd:date.
```

Temporal Knowledge in YAGO2geo Ontology

- Extend YAGO2 ontology in order to support new temporal knowledge
- Exan



Methodology

- **Important:** Avoid duplicate information
- Match entities of YAGO2 with entities of the data sources (`owl:sameAs`)
- Matching Phase
  - Label Similarity Filter
    - Jaro-Winkler (>0.82 threshold)
    - Exploits multilingual labels
  - Geometry Distance Filter (over the label-based matched entities)
    - Disambiguation Step (many geographic entities that share the same name)
    - Euclidean distance in the WGS:84 coordinate system
- Manual evaluation of a subset of the total matches
- References: YAGO2, LinkedGeoData

Methodology

- The matching phase is applied on **manually created** pairs of classes
  - Municipalities of GAG and *third-order_administrative_division* of YAGO2
  - Level-1 of GADM and *first-order_administrative_division* of YAGO2
  - Forests of OSM and *forests* of YAGO2
  - …
- Matched entities of YAGO2: Extended with new information
- Unmatched entities data sources
  - Administrative data sources: New entities in YAGO2geo
  - OpenStreetMap: Noisy data → not included in YAGO2geo

Results, Greece

GAG	YAGO2	# Matches	Precision
decentralized administrations	administrative_division	6/7	1.000
regions	first-order	11/13	1.000
regional units	administrative_division	21/74	1.000
municipality	third-order	324/325	1.000
municipal unit and community	populated_place and locality	530/1037	0.907

- The second administrative level of YAGO2 contains former Greek administrative units

YAGO2geo OSM Updater

- YAGO2geo is automatically updated by OSM information.

YAGO4

- Latest version of YAGO
- Created from three sources:
  - Leaf nodes of the taxonomy and entities: **Wikidata** (with Wikipedia page titles)
  - Inner nodes of the taxonomy and properties: **schema.org** and **bioschemas.org**
- 10K classes
- SHACL & OWL constraints: consistent KG
- **Geospatial information:**
  - Schema defines GeoCoordinates or a GeoShape type (not integrated to YAGO4)

```
@base <http://yago-knowledge.org/resource/> .
@prefix schema: <http://schema.org/> .
<Athens> schema:geo geo:37.979444444444745,23.7161111111113 .
```

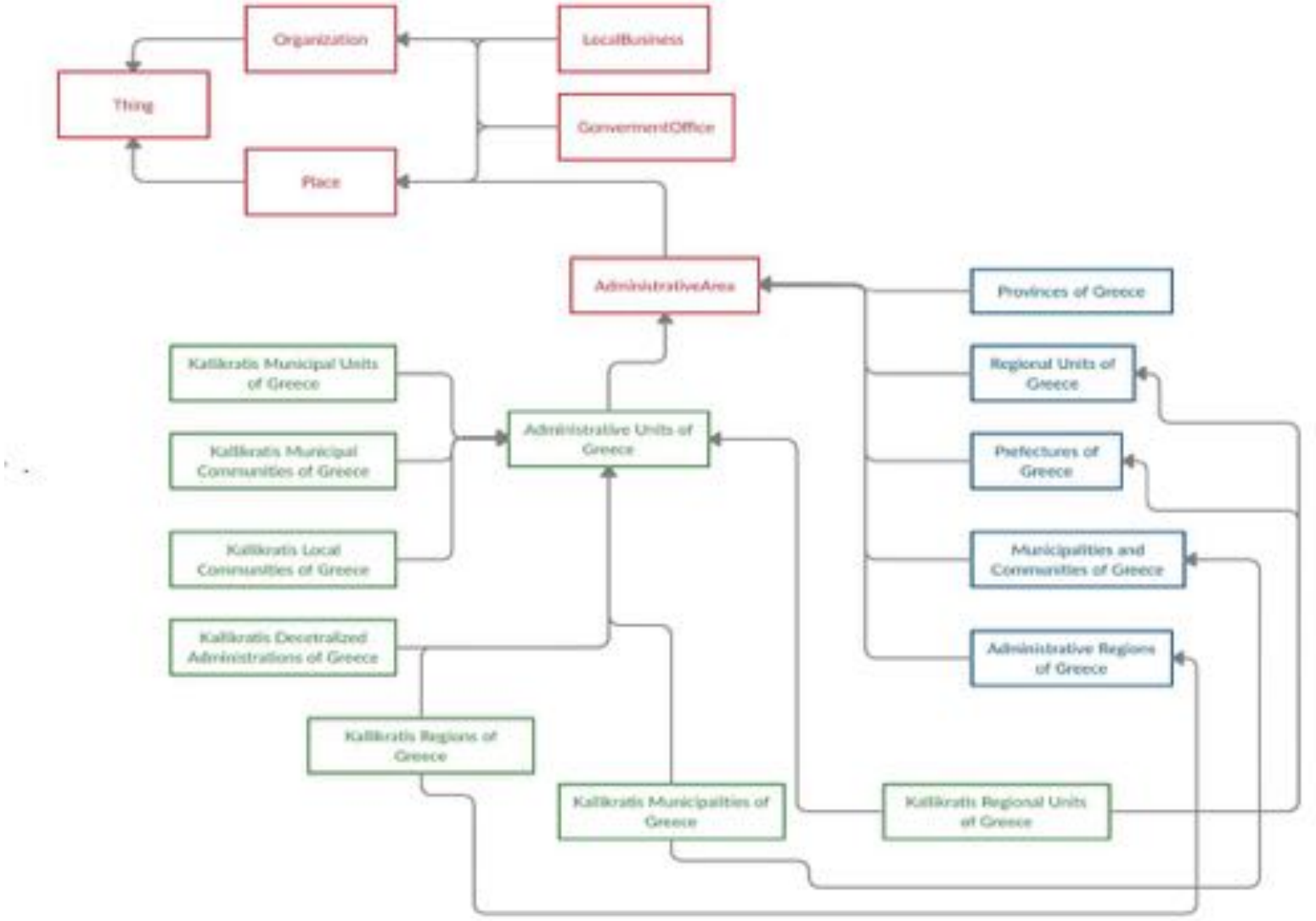

Pellissier Tanon, T., Weikum, G., Suchanek, F. (2020). YAGO 4: A Reasonable Knowledge Base. In: , et al. The Semantic Web. ESWC 2020. Lecture Notes in Computer Science(), vol 12123. Springer, Cham

YAGO4, Geospatial Extension

- Follows the methodology of YAGO2geo
  - GAG dataset
- Detailed geometries
  - Polygons
  - Multi-polygons
- Example:

```
@base <http://yago-knowledge.org/resource/> .
@prefix schema: <http://schema.org/> .
@prefix ext: <http://ai.di.uoa.gr/> .
<Athens> schema:geo ext:Geometry_Q1524 .
ext:Geometry_Q1524 schema:polygon "..." .
ext:Geometry_Q1524 schema:polygon "..." .
                   ...
ext:Geometry_Q1524 schema:polygon "..." .
```

YAGO4, Taxonomy Extension



Open Issues

- schema.org does not support:
  - Well-known spatial literal formats (e.g., WKT)
  - Complex geometries (e.g., multi-polygons, multiple coordinate reference systems)

WorldKG

- Geographic KG from OpenStreetMap (OSM) dataset
- OSM: rich, open, crowdsourced geographic dataset
  - >6.8B geographic entities in 188 countries
  - User-defined key-value pairs
- WorldKG:
  - Conversion of OSM schema to ontology
  - >100M geographic entities from 188 countries, >800M triples
  - Aligned with Wikidata & DBpedia
- WorldKG population:
  - Entities: OSM nodes with tags from most recent OSM dumps
  - Geographic objects represented as `sf:Point` objects
  - Property: `geo:asWKT`
  - Geographic coordinates: from OSM keys (`lat`, `long`) as `geo:WKTLiteral` literals

WorldKG

OSM map feature:

`natural = cave_entr`

Top-level class

Subclass

WorldKG Ontology

KnowWhereGraph

- For the integration of geospatial data silos to offer information for environmental intelligence:
  - Disaster relief
  - Food-related supply chains
  - Land valuation
- Information about:
  - Extreme events
  - Administrative boundaries
  - Soils
  - Crops
  - Climate
  - Transportation
- >12B triples

Janowicz, K. et al. 2022. " Know, Know Where, KnowWhereGraph: A densely connected, cross-domain knowledge graph and geo-enrichment service stack for applications in environmental intelligence." AI Magazine 43: 30– 39.

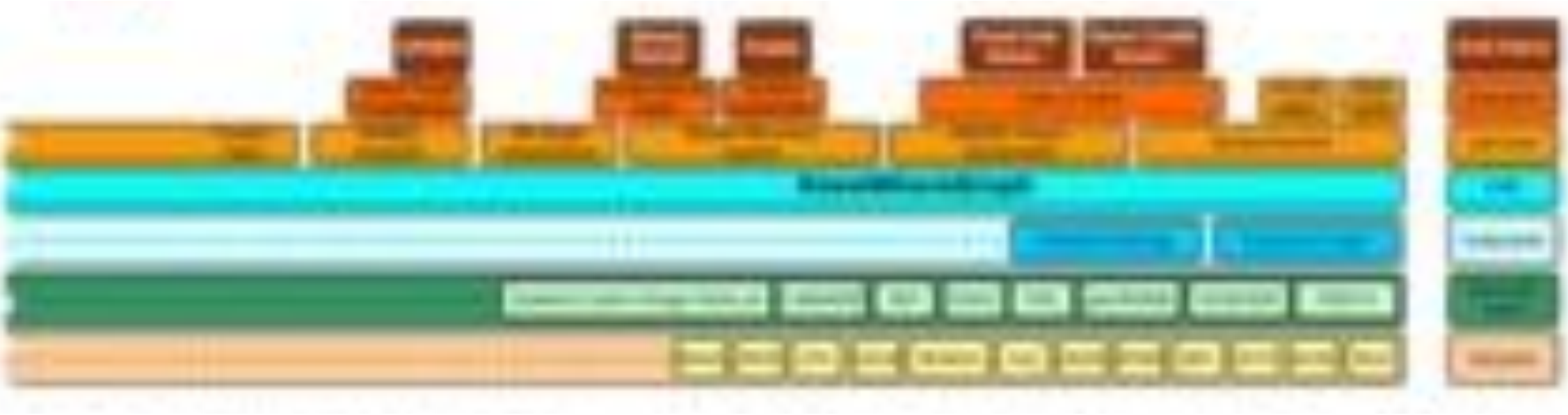
KnowWhereGraph

- **Geospatial Knowledge:** S2 Grid System
  - Hierarchical grid over Earth's surface
  - Each grid cell in a level: 4 subcells of increasing spatial resolution
  - US: at least 20km²/cell
- Ontologies:
  - GeoSPARQL
  - SOSA/SSN
  - QUDT
  - …
- Datasets:
  - 17 thematic raw datasets (e.g., Soil Properties dataset, Wildfire dataset, Climate Hazards dataset)
  - 11 place-centric raw datasets (e.g., S2 Cells dataset, ZIP dataset, GADM)

Datasets in KnowWhereGraph

Thematic Datasets					Place-Centric Datasets		
Dataset Name/ Theme	Source Agency	Key Attributes	Spatial Coverage	Temporal Coverage	Place-Centric Dataset	Defining Authority	Spatial Coverage
Soil Properties	USDA	soil type, farmland class	Targeted regions in US	Current	S2 Cells	Google	Lvl 9 (Global), Lvl 13 (US),
Wildfires	USGS, USDA, USFS, NIFC	wildfire type, burn severity, num. acres burned, contained date	US	1984–current	Global Administrative Regions	University of Berkeley, Museum of Vertebrate Zoology and the International Rice Research Institute	Global
Earthquakes	USGS	magnitude, length, width, geometry	Global (mag. over 4.5)	2011-01-01 to 2022-01-18			
Climate Hazards	NOAA	injuries, deaths, property damages	US	1950–2022			
Expert - Covid-19 Mobility	Direct Relief (DR)	name, affiliation, expertise	Global	2021	US Federal Judicial District	DoJ, ESRI	US
Expert - General	KWG, UC System, DR, Semantic Scholar	name, affiliation, expertise with spatiotemporal scopes	Global	unlimited	National Weather Zones	NOAA	US
Cropland Types	USDA	crop types (raster data)	US	2008-2021	FIPS Codes	NRCS	US
Air Qual. Obs.	U.S. EPA	AQI value, CO concentration	US	1980–2022	Designated Market Area	Nielen	US
Smoke Plumes	NOAA	daily smoke plumes extent	US	2010-2022	ZIP	ZCTA	US
Climate Observations	NOAA	temperature, precipitation, PDSI, PHSI	US	1950 - 2022	Climate Division	NOAA	US
Disaster Declaration	FEMA	designated area, program, amount approved, program designated date	US	1953 - 2022	Census Metropolitan Area	US Census	US
Smoke Plume Extents	NOAA	Smoke extent	US	2017 - 2022	Drought Zone	NDMC, USDA,NOAA	US
BlueSky Forecasts	Bluesky	PM10, PM5	US	2022-03-07	Geographic Name Information System	USGS	US
Transportation (highway network)	DOT	road type, road length, road sign	US	2014			
Public Health	CDC, US Census	below poverty level percent, diabetes age adjusted 20 plus percent, obesity age adjusted 20 plus percent	US	2017			
Social Vulnerability	CDC/ATSDR	social vulnerability index	US	2018			
Hurricane Tracks	NOAA	max wind speed, min pressure	US	1851-2020			

KnowWhereGraph



KnowWhereGraph - Interlinking

- The regions, such as climate divisions or counties, are linked to entities from Wikidata or the Geographic Names Information System
  - hence, access to population density, previous events, soil health, etc
- Topological relations among regions (e.g., RCC8)
- Interlinked places with events with causal relationships and provenance

E.g., Where a fire took place, which events it triggered,and which regions have been affected, for example, by heavy smoke

KnowWhereGraph

Thank you!

Questions?

Question answering over geospatial knowledge graphs

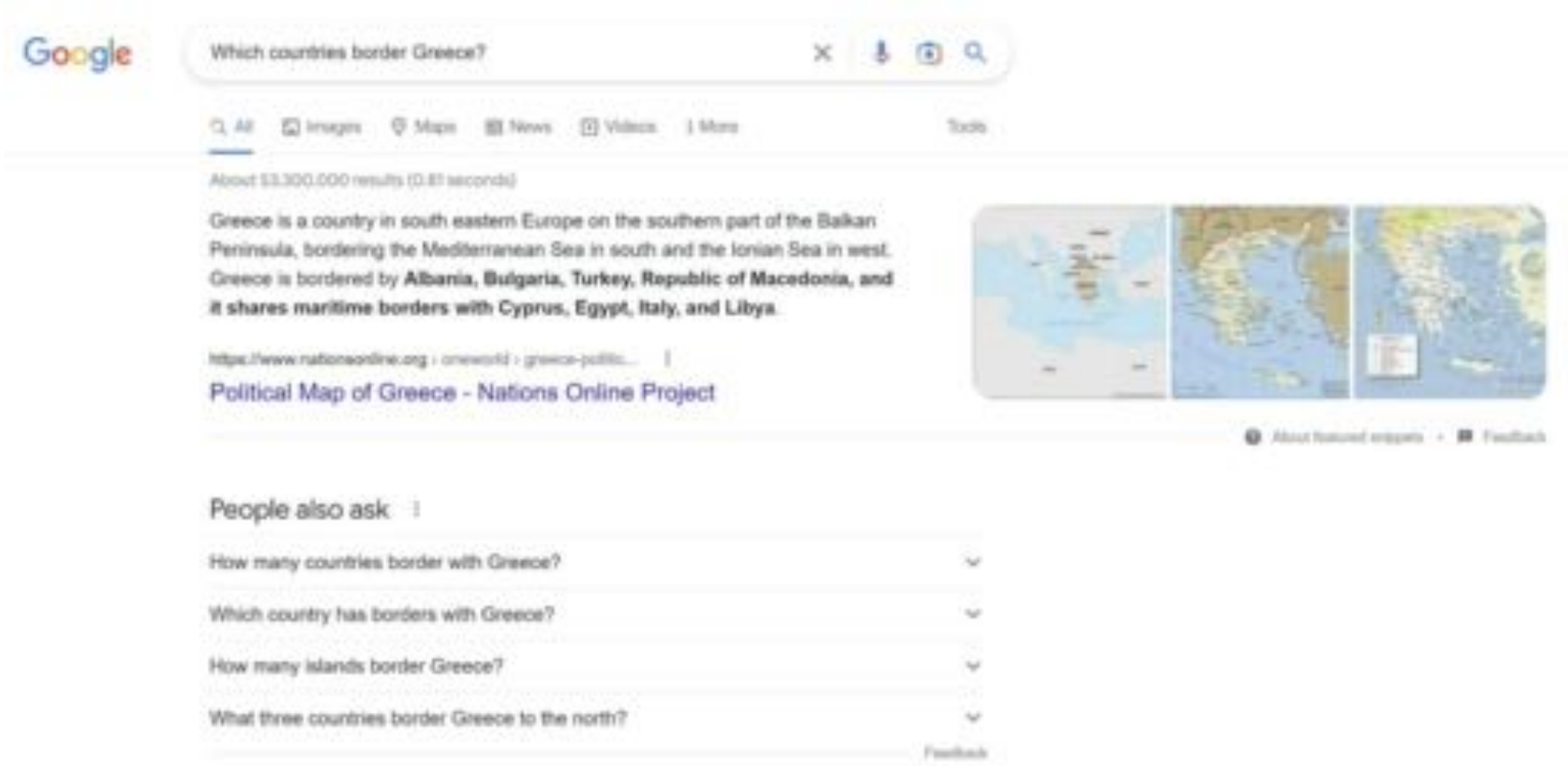
Motivation

- **Geospatial** or **geographic** knowledge
  - important in many applications (travel, tourism, education etc.)

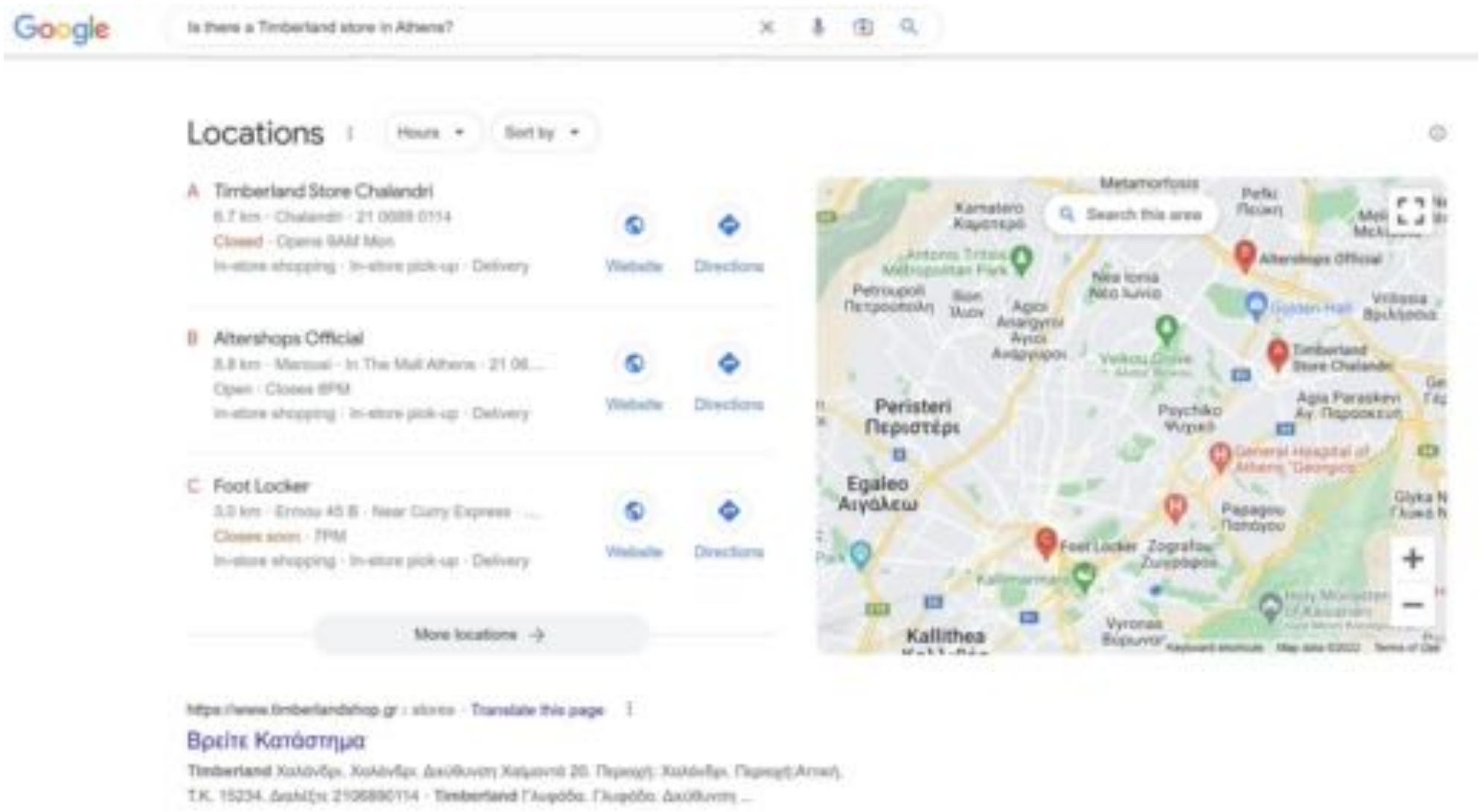
- **Increasing** use of **geospatial terms** in **search engine** queries. Examples:
  - Which countries border Greece?
  - Is there a Timberland store in Athens?

- **Can geospatial questions be answered by today's search engine ?**

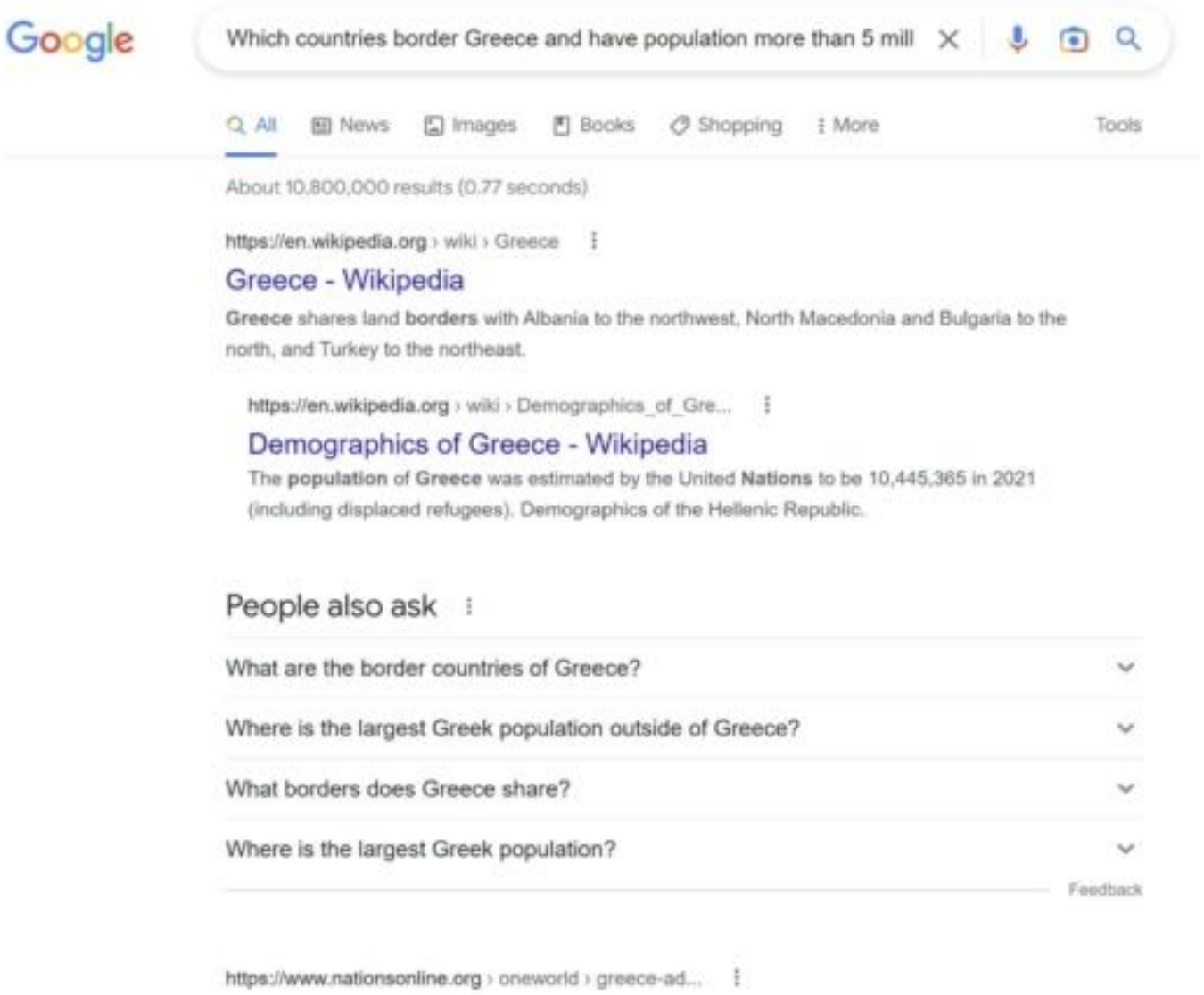
Which countries border Greece?



Is there a Timberland store in Athens?



Which countries border Greece and have population more than 5 million?



Question

- How can we go beyond what current search engines offer for geospatial questions?
- Answer: Develop **geospatial question answering systems!**

Existing Geospatial Question Answering Systems

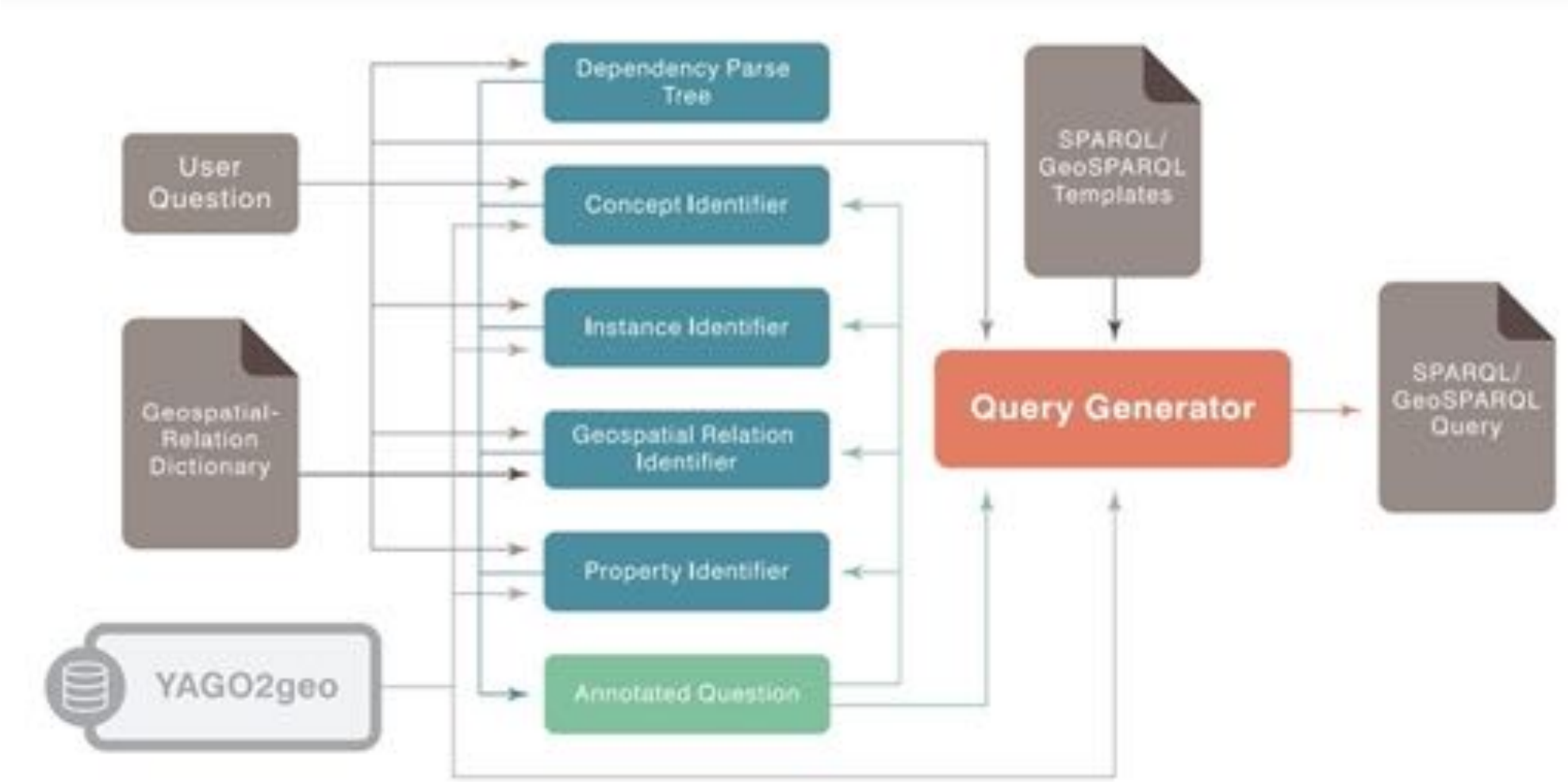
- GeoQA (2018) [1]
  - DBpedia + OpenStreetMap + GADM
- Revised version of GeoQA (2020) [2]
  - Same KGs
- Neural factoid geospatial question answering [3]
  - Does not generate executable queries over any KGs.
- **Question Answering System by Hamzei et al.** [4]
  - YAGO2geo (with more data added from OpenStreetMap)
- **GeoQA2** [5]
  - YAGO2geo

[1] Punjani, D., et al. "Template-based question answering over linked geospatial data." Proceedings of the 12th Workshop on Geographic Information Retrieval. 2018.
[2] Punjani, D., et al., 2020. Template-based question answering over linked geospatial data. CoRR, abs/2007.07060. Available from: https://arxiv.org/abs/2007.07060
[3] Li, H., et al. "Neural factoid geospatial question answering." Journal of Spatial Information Science 23 (2021): 65-90.
[4] Hamzei, E., et al. 2022. Translating place-related questions to GeoSPARQL queries. In: Proceedings of the Web Conference (WWW).
[5] Punjani, D., et al. (2023) The Question Answering System GeoQA2 for the Geospatial Knowledge Graph YAGO2geo and a New Benchmark For its Evaluation. Forthcoming.

The Question Answering Engine GeoQA2

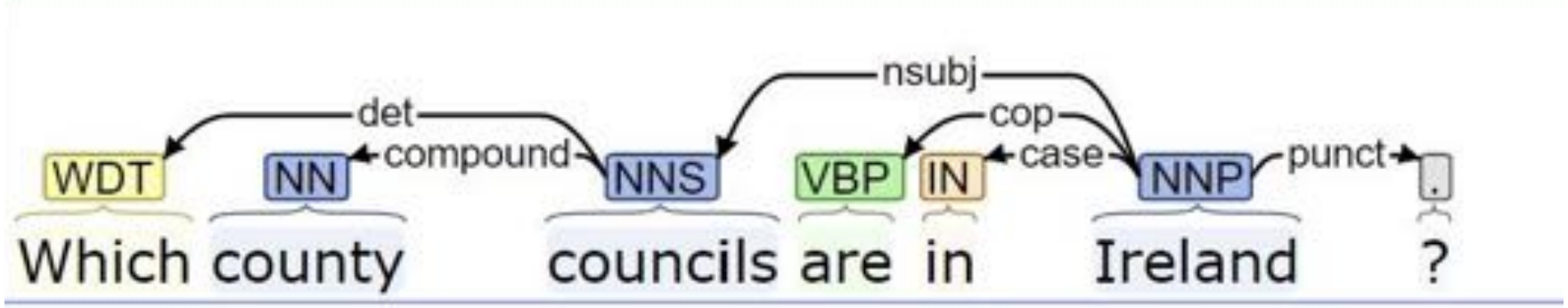
- GeoQA2 is developed using Qanary and Frankenstein.
  - Qanary
    - lightweight component-based QA methodology
  - Frankenstein
    - most recent implementation of the ideas of Qanary.
- In our work, we leverage the power of the Qanary framework to create six QA components which collectively implement the geospatial QA pipeline of GeoQA2.

The Conceptual Architecture Of The GeoQA2 System



Dependency Parse Tree Generator

- Part-of-speech (POS) tagging
- Dependency parse tree
- **Stanford CoreNLP API.**
- Example question: "**Which county councils are in Ireland?**"



Instance Identifier

- Identifies the **features**.
  - Example: **Country Ireland**, **city Dublin**, **river Shannon** etc.
- **Mapped** to **YAGO2geo resource**.
- **TagMeDisambiguate**.
  - It maps instances to Wikipedia (hence to YAGO2 as well)
- We also search for resources in the YAGO2geo(added instances) dataset but not in YAGO2, and add them to the list of identified instances.
- Example question: "**Which forests are contained in the Norfolk?** "
  - **yago2:Norfolk**
- The appropriate node of dependency parse tree is annotated with the results.

Concept Identifier

- Identifies the **class of features**
  - Maps them to the corresponding **classes** in the **YAGO2geo ontology**.
- Following process is followed in concept identifier:
  - Iterate through **classes of ontology**
  - generate **n-grams** from question based on number of words in class labels
  - n-grams with **string similarity** more than **99%** are mapped to the respective class
- E.g., "**Which county councils are in Ireland?**"
  - yago2geoo:OSI_County_Council
- The appropriate node of dependency parse tree is annotated with the identified concepts.

Geospatial Relation Identifier

- This module first identifies **geospatial relations** in the input question, and then maps them to a **spatial function** of the **GeoSPARQL** or **stSPARQL vocabulary**, or a data property with a spatial semantics in the YAGO2geo ontology.
- The appropriate node of dependency parse tree is annotated with the results of this module.

Supported Geospatial Relations

Category	Geospatial Relation
Topological relations	“within”,“crosses”,”borders”
Distance relations	“near”,“at most x units”, “at least x units”
Cardinal Direction relation	“north of”, “ south of”, “east of”, “west of”, “northwest of”, “northeast of”, “southwest of”, and “southeast of”

Geospatial relation	Synonyms in dictionary
within	In, inside, is located in, is included in
crosses	Cross, intersect
near	Nearby, close to, around
borders	is/are at the border of, is/are at the outskirts of, at the boundary of
North of	Above of
South of	below
East of	To the right
West of	To the left

Property Identifier

- The property identifier module
  - Identifies **attributes of types of features** and **attributes of features**
  - Maps them to the corresponding **properties** in **YAGO2geo**
- Consider question "**What is the population of villages in Rhodes ?**"
  - **yago2geoo:hasGAG_Population**

YAGO2geo Class	YAGO2geo Property	Property Label
yago2geoo:OSM_village	yago2geoo:hasGAG_Population	GAG Population
	yago2geoo:hasOSM_Area	OSM Area
	yago2geoo:hasGADM_UpperLevelUnit	GADM UpperLevelUnit
yago2:wordnet_river_109411430	yago2infobox:length	length
	yago2infobox:source	source
	yago2infobox:discharge	discharge

Property Identifier

- "Which is the largest lake in Greece?"
  - yagoinfobox:area

YAGO2geo Class	Keyword	Property
yago2:wordnet_lake_109328904	Largest, Smallest	yagoinfobox:area
yago2:wordnet_river_109411430 yago2:wordnet_bridge_102898711	Smallest, Largest, Biggest	yagoinfobox:length
yago2:wordnet_mountain_19359803	Highest, Tallest, Smallest	yagoinfobox:elevationm
yago2:wordnet_hospital_1025405959 yago2:wordnet_stadium_104295881 yago2:wordnet_hotel_103542333	Biggest	yagoinfobox:capacity
	Largest, Smallest	yagoinfobox:area
yago2:wordnet_dam_103160309	Tallest	yagoinfobox:damnheight
	Largest, Smallest	yagoinfobox:area
yago2:wordnet_building_102913152	Highest, Tallest, Smallest	yagoinfobox:height
yago2:wordnet_county_108546183 yago2:wordnet_city_108524735 yago2:wordnet_town_108524735 yago2:wordnet_village_108672738 yago2:wordnet_settlement_108672562	Smallest, Biggest, Largest	yagoinfobox:area

Property Identifier

- Supposingly YAGO2geo does not contain “**area**” property for the mentioned class
  - It will be calculated using spatial function
    - **strdf:area()**

Query Generator

- This module generates **GeoSPARQL/SPARQL** queries using handcrafted query templates.
- We have identified **question patterns** and **mapped** them to **GeoSPARQL/SPARQL** queries

Question Pattern	Example
CRI	Which rivers cross Limerick?
CRIRI	Which churches are close to the Shannon in Limerick?
CRC	Which restaurants are near hotels?
CRCRI	Which restaurants are near hotels in Limerick?
IRI	Is Hampshire north of Berkshire?
IP	What is total area of county Cheshire?
PCRI	What is the length of river that crosses London?
PCRIRI	Which Greek restaurant are near Wembley Stadium in London?
PCRCRI	Find the area of the parks that belong to counties north of Budgebury Forest.
NCRI	How many hospitals are there on Oxford?

Where

- C → Concept
- R → Geospatial relation
- I → Instance
- P → Property
- N → Count

Question Patterns : PCRI (Query Generator)

Pattern	Example natural language question	Templates	Query Generated

PCRI	What is the population of the villages in Rhodes?	**GeoSPARQL**:
		select ?property where {
		?x rdf:type _Concept;
		geo:hasGeometry ?xGeom;
		_Property ?property.
		?xGeom geo:asWKT ?xWKT.
		_Instance geo:hasGeometry ?iGeom.
		?iGeom geo:asWKT ?iWKT.
		FILTER(_Relation(?xWKT, ?iWKT))
		}
			select ?property where {
			?x rdf:type yago2geoo:OSM_village;
			geo:hasGeometry ?xGeom;
			yago2geoo:hasGAG_Population ?property.
			?xGeom geo:asWKT ?xWKT.
			yago2:Rhodes geo:hasGeometry ?iGeom.
			?iGeom geo:asWKT ?iWKT.
			FILTER(geof:sfWithin(?xWKT, ?iWKT))
			}

		SPARQL:
		Select ?property where {
		?x rdf:type _Concept.
		?x _Relation _Instance.
		?x _Property ?property.
		}
			select ?property where {
			?x rdf:type yago2:wordnet_village_108672738.
			?x yago2:isLocatedIn yago2:Rhodes.
			?x yagoinfobox:population ?property.
			}

Queries With Aggregates and Superlatives

- **"Which Civil Parishes in Ireland have more than 10 townlands?"**
  - Pattern present in the question is **CRIRI**
  - **Group By** and **Count**
- Use of Constituents and Dependencies from parse trees
  - **Quantifier phrase** (QP) from Constituency parse tree
  - Add **Count(distinct ?y) as ?total and Group By(?x) Having (?total > 10)**
- Generated Query :

```
select distinct ?x (count(distinct ?y) as ?total) where {
    ?x rdf:type yago2geoo:OSI_Civil_Parish;
        geo:hasGeometry ?cGeom1.
    ?cGeom1 geo:asWKT ?cWKT1.
    ?y rdf:type yago2geoo:OSI_Townland;
        geo:hasGeometry ?cGeom2. ?cGeom2 geo:asWKT ?cWKT2.
    yago2:Republic_of_Ireland geo:hasGeometry ?iGeom.
    ?iGeom geo:asWKT ?iWKT.
    FILTER(geof:sfWithin(?cWKT1,?iWKT) && geof:sfContains(?cWKT1,?cWKT2))
} GROUP BY(?x) HAVING(?total >  10 )
```

Query Ranking

- We use a very simple heuristic for the ranking of generated queries based on the estimated selectivity of the generated queries.
- We compute the selectivity of a SPARQL or GeoSPARQL query taking into account only the triple patterns present in the query and using the formulas of Stocker et al., 2008.
- The generated query with the lowest selectivity is selected to be executed.

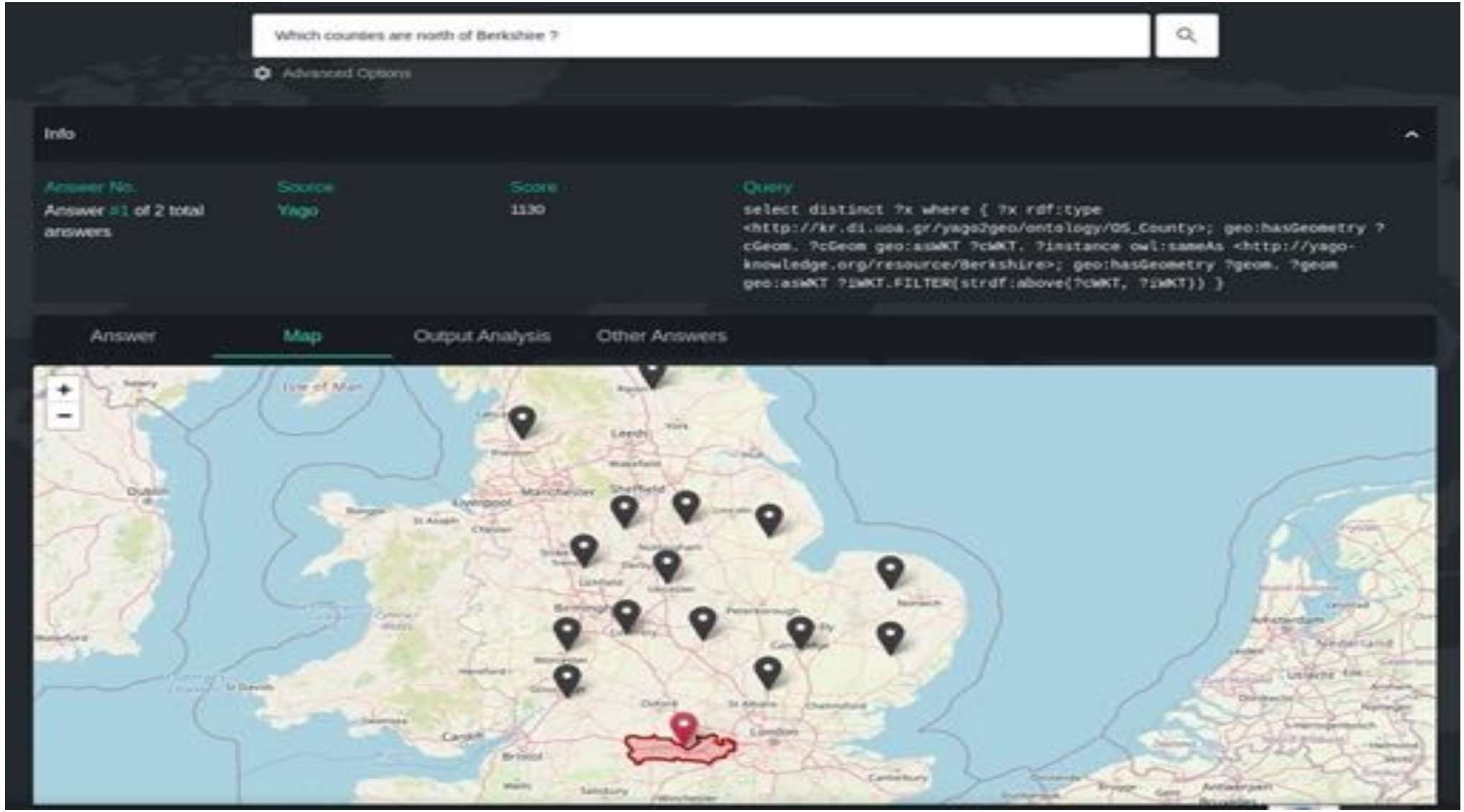
Stocker, M., et al., 2008. SPARQL basic graph pattern optimization using selectivity estimation. In: J. Huai, R. Chen, H. Hon, Y. Liu, W. Ma, A. Tomkins and X. Zhang, eds. Proceedings of the 17th International Conference on World Wide Web, WWW 2008, Beijing, China, April 21-25, 2008. ACM, 595–604.

Query Executor

- The last module executes the top-ranked GeoSPARQL/SPARQL query against a YAGO2geo Strabon endpoint at http://pyravlos2.di.uoa.gr:8080/yago2geo/.

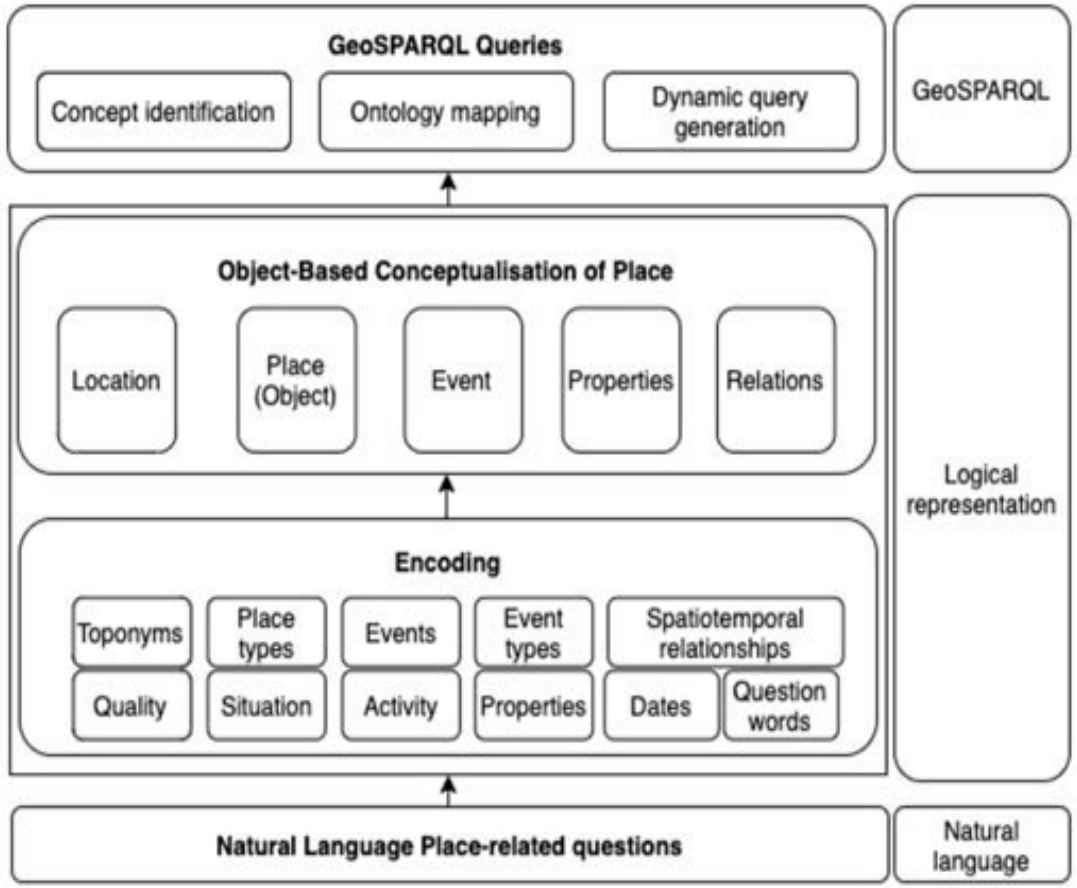
```
select ?property where {
?x rdf:type yago2geoo:OSM_village;
     geo:hasGeometry ?xGeom;
     yago2geoo:hasGAG_Population ?property.
?xGeom geo:asWKT ?xWKT.
yago2:Rhodes geo:hasGeometry ?iGeom.
?iGeom geo:asWKT ?iWKT.
FILTER(geof:sfWithin(?xWKT, ?iWKT))
}

Answer :

"2714"^^<http://www.w3.org/2001/XMLSchema#integer>
"1027"^^<http://www.w3.org/2001/XMLSchema#integer>
```

Which counties are north of Berkshire ?



QA Engine by Hamzei et al. (2022)

- Encoding Extraction
- Grammatical parsing
  - Identify relations among encodings
- First-Order Logic Statements
- GeoSPARQL query generation
  - Concept identification and ontology mapping
  - Query generation



Method workflow to translate place-related question to GeoSPARQL queries

Example Question

- How many pharmacies are in 200 meter radius of High Street in Oxford?

Encoding Extraction

- Pre-trained models for fine-grained
  - named entity recognition (NER) [1]
  - part-of-speech tagging [2]

Encoding class	Code	Encoding class	Code
where	1	*place name*	P
what	2	*place type*	p
which	3	*event*	E
when	4	*event type*	e
how	5	*properties*	o
how+adj	6	*activity*	a
why	7	*situation*	s
is/are	8	*spatial relation*	R
date	d	*temporal relation*	r
place quality	Q	*properties/events quality*	q
comparison	<, >, =	*and*	&
or	|	*negation*	!

[1] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, K. Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. ArXiv abs/1603.01360 (2016)
[2] V. Joshi, Matthew E. Peters, and Mark Hopkins. 2018. Extending a Parser to Distant Domains Using a Few Dozen Partially Annotated Examples. In ACL.

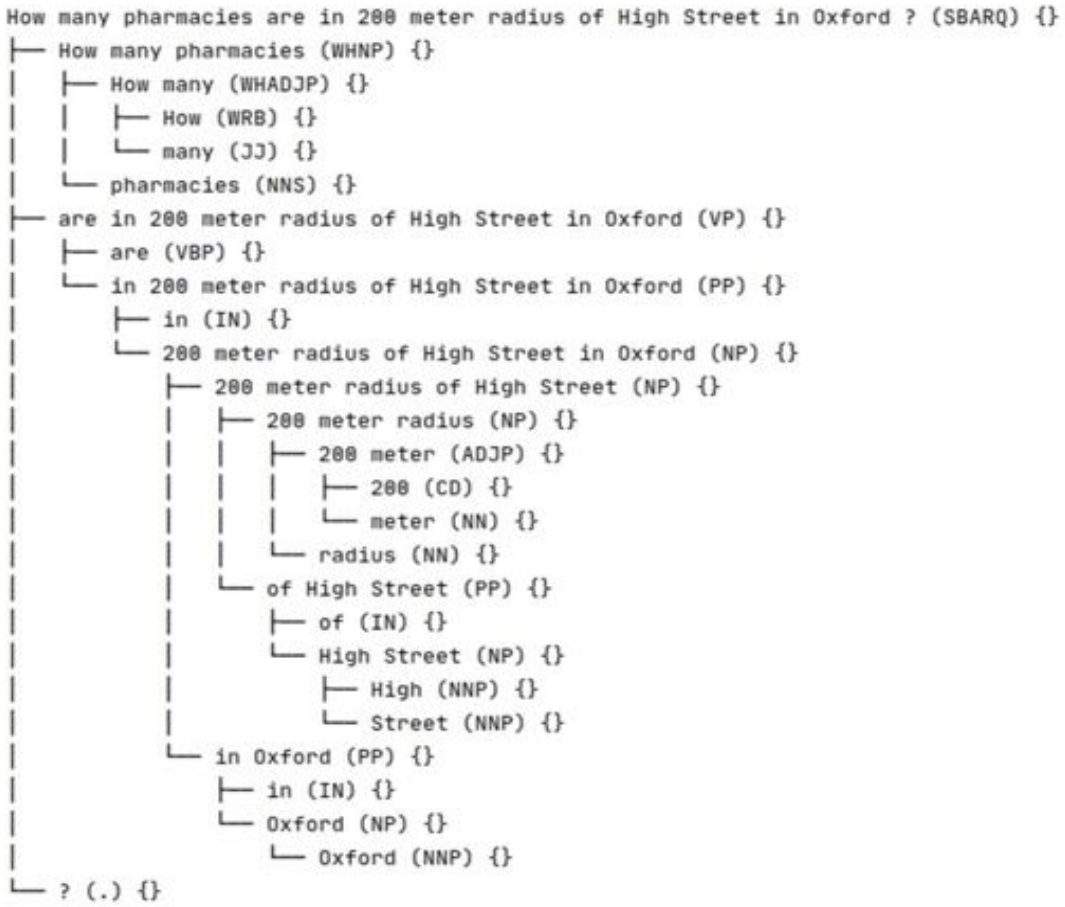
Encoding

name	part-of-speech	role
How many	ADV	6
High Street	NOUN	P
Oxford	NOUN	P
pharmacies	NOUN	p
in 200 meter radius of	ADP	R
in	ADP	R

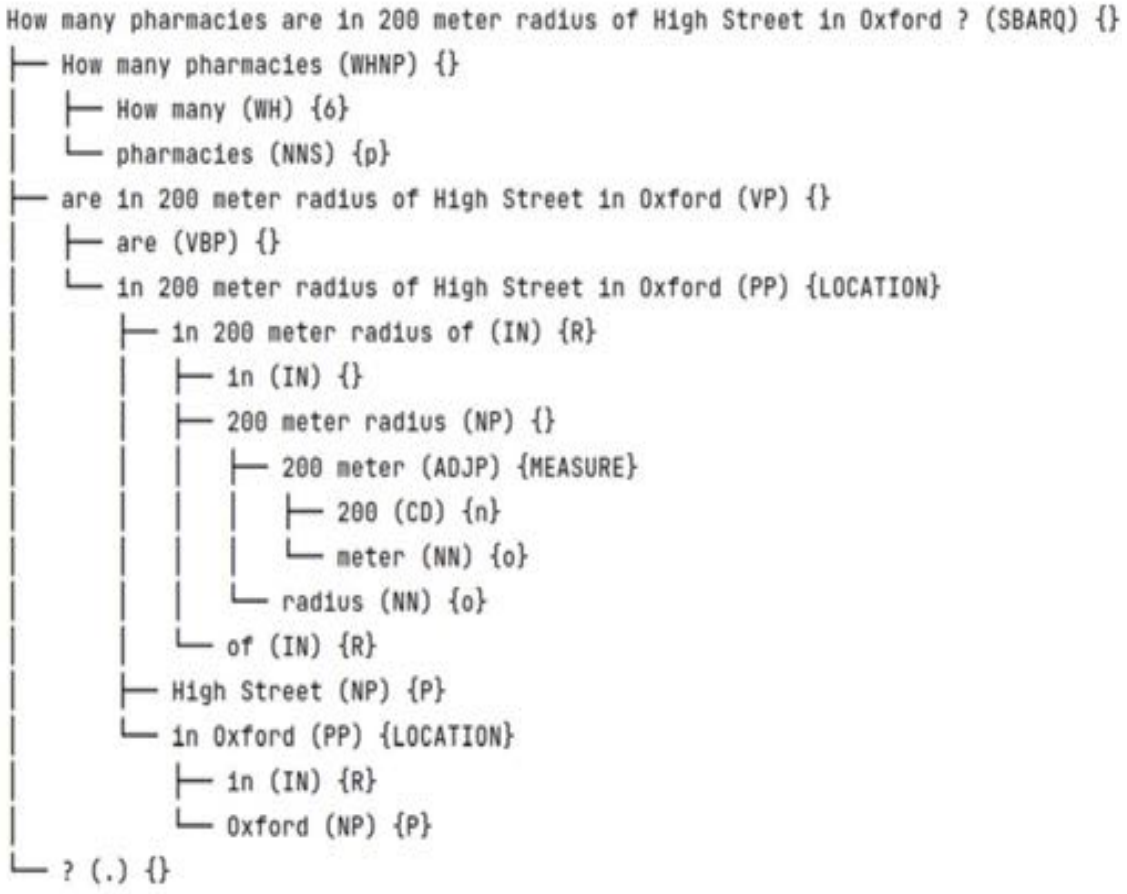
Grammatical Parsing

- Extract relationships between encodings
  - **Constituency Parsing**
    - Phrase-level information
      - Conjunctions phrases
      - Quality phrases
      - Location phrases
  - **Intent Detection**
    - Heuristic method derived from **constituency parsing** results
  - **Dependency Parsing**
    - Places/events with location phrases
    - Situation/activities with properties
    - Places with situations/activities
    - Comparison phrases and their source

Constituency Parsing



Constituency Parse Tree (before labelling)



Constituency Parse Tree (after labelling)

Intent Identification

- Word Phrase Rule
  - **How+adjective**
  - Intent could be **Count** or distance
- Specificity rule
  - More Specific Concepts
    - "How many **pharmacies** are in 200 meter radius of *High Street in Oxford*?"
- Identified Intent
  - **How many pharmacies**

Dependency Parsing

```
are (root) {['AUX']} []
├─ How (advmod) {['ADV']} []
├─ pharmacies (nsubj) {['NOUN']} []
│  └─ many (dep) {['ADJ']} []
├─ in (prep) {['ADP']} []
│  └─ radius (dep) {['NOUN']} []
│     ├─ meter (nn) {['NOUN']} []
│     │  └─ 200 (dep) {['NUM']} []
│     └─ of (prep) {['ADP']} []
│        └─ Street (pobj) {['PROPN']} []
│           └─ High (amod) {['PROPN']} []
├─ in (prep) {['ADP']} []
│  └─ Oxford (pobj) {['PROPN']} []
└─ ? (dep) {['PUNCT']} []
```

Dependency Parse Tree (before labelling)

```
are (root) {['AUX']} []
├─ How many (advmod) {['ADV']} [6]
├─ pharmacies (nsubj) {['NOUN']} [p]
├─ in 200 meter radius of (prep) {['ADP']} [R]
│  └─ High Street (pobj) {['NOUN']} [P]
├─ in (prep) {['ADP']} [R]
│  └─ Oxford (pobj) {['NOUN']} [P]
└─ ? (dep) {['PUNCT']} []
```

Dependency Parse Tree (after labelling)

Generating First-Order Logic Statements

- **Terms**
  - Either a **constant** or a **variable**
  - e.g., **High Street** or **x**
- **Predicates**
  - Symbols that either **declare terms** or **describe their relationships**
  - Every generic declaration is either a place or an event
    - e.g., $\forall$ x PHARMACY(x) $\rightarrow$ PLACE(x)
  - Binary or ternary predicates: spatiotemporal, situation/activity relationships, comparisons and qualities
    - e.g., declaration: PLACE(High Street) and relations: IN(High Street, Oxford) and IN_RADIUS_OF(x0, High Street, 200 meter)

- **Generated Statement**
  - **COUNT(x0) : PLACE(High Street) $\wedge$ PLACE(Oxford) $\wedge$ PHARMACY(x0) $\wedge$ IN_RADIUS_OF(x0, High Street, 200 meter) $\wedge$ IN(High Street, Oxford)**

GeoSPARQL Query Generation - Step 1

- **Concept identification and ontology mapping** over YAGO2geo (with more OSM data)
- Apache Solr to **index** the **names** and **identifier**s of **places** and **events**
  - **String matching using Solr search**
- **One-to-many mapping** to match **extracted place/event types** and **properties** to the **knowledge base ontology**
  - Exact matching
    - Knowledge graph ontology
  - Label matching using cosine similarity
    - The contextual BERT representations (Devlin et al., 2019)
    - Labels in the ontology
  - Glossary matching using cosine similarity
    - BERT representations of the definitions
      - WordNet and Wikipedia snippet search
    - Glossary in the ontology

Devlin et al. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of naacL-HLT (pp. 4171-4186).

Example

- **Names** and **identifier**s of **places** and **events**
  - **PLACE(High Street) :** yago2:High_Street_Lincoln yago2geor:OSM_HighStreet561 yago2geor:OSM_HIGHSTREET678 yago2geor:OSM_HighStreet541 yago2geor:OSM_HighStreet789 yago2geor:OSM_HighStreet302 yago2geor:OSM_HighStreet414 yago2geor:OSM_HighStreet985 yago2geor:OSM_HighStreet936 yago2geor:OSM_HIGHSTREET381…
  - **PLACE(Oxford) :** yago2:Oxford yago2geor:OSM_oxfordvapours470 yago2geor:osientity_2AE19629B1DE13A3E05500000000001 yago2geor:OSM_Oxford440 yago2geor:OSM_Oxford180 yago2geor:OSM_Oxford159 yago2geor:OSM_Oxford996 ……
- **Extracted place/event types**
  - **PHARMACY(x0) :** yago2geoo:OSM_amenity_veterinary yago2:wordnet_drugstore_103249342 yago2geoo:OSM_office_logistics yago2geoo:OSM_amenity_pharmacy yago2geoo:OSM_amenity_dentist

GeoSPARQL Query Generation - Step 2

- **Generation of an intermediate non-executable query**
  - **Predefined templates**
  - Determine structure of a query (i.e., ASK vs. SELECT query)
    - extracted intent
  - WHERE-clause is generated
    - concatenate individual concept and relation definition statements
  - Sorting and aggregation (ORDER-BY and GROUP-BY clauses) If needed.
- **Identified concept and their mappings** to the ontology are **replaced in query** to generate **final executable query**.

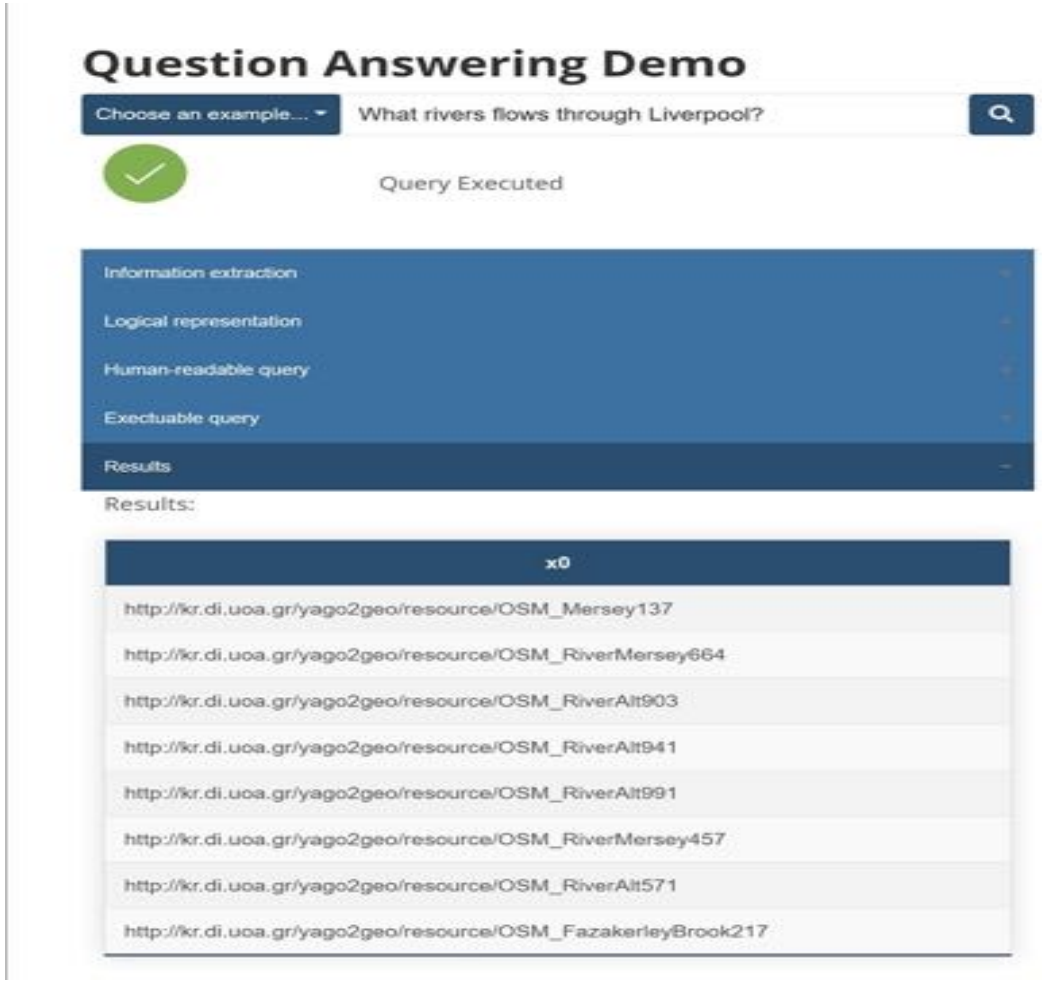
Example (Non-executable Query)

```
SELECT DISTINCT (COUNT(distinct ?x0) as ?countx0)
WHERE {
   VALUES ?c0  {< URIS >}.
   ?c0 geosparql:hasGeometry ?c0G .
   ?c0G geosparql:asWKT ?c0GEOM .
   VALUES ?c1  {< URIS >}.
   ?c1 geosparql:hasGeometry ?c1G .
   ?c1G geosparql:asWKT ?c1GEOM .
   ?x0 rdf:type ?x0TYPE;
      geosparql:hasGeometry ?x0G .
   ?x0G geosparql:asWKT ?x0GEOM .
   VALUES ?x0TYPE {< URIS >} .
   FILTER(geof:distance(?x0GEOM, ?c0GEOM, units:meter) < 200).
   FILTER (geof:sfContains(?c1GEOM, ?x0GEOM)).
}
```

Example (Final Generated Query)

```
SELECT DISTINCT (COUNT(distinct ?x0) as ?countx0)
WHERE {
	VALUES ?c0 {yago2:High_Street_Lincoln yago2geor:OSM_HighStreet561
yago2geor:OSM_HIGHSTREET678 yago2geor:OSM_HighStreet541 yago2geor:OSM_HighStreet789
yago2geor:OSM_HighStreet302 yago2geor:OSM_HighStreet414 yago2geor:OSM_HighStreet985
yago2geor:OSM_HighStreet936 yago2geor:OSM_HIGHSTREET381 …..}.
	?c0 geosparql:hasGeometry ?c0G .
	?c0G geosparql:asWKT ?c0GEOM .
	VALUES ?c1 {yago2:Oxford yago2geor:OSM_oxfordvapours470
yago2geor:osientity_2AE19629B1DE13A3E05500000000001 yago2geor:OSM_Oxford440
yago2geor:OSM_Oxford180 yago2geor:OSM_Oxford159 yago2geor:OSM_Oxford996 ……}.
	?c1 geosparql:hasGeometry ?c1G .
	?c1G geosparql:asWKT ?c1GEOM .
	?x0 rdf:type ?x0TYPE;
		geosparql:hasGeometry ?x0G .
	?x0G geosparql:asWKT ?x0GEOM .
	VALUES ?x0TYPE {yago2geoo:OSM_amenity_veterinary yago2:wordnet_drugstore_103249342
yago2geoo:OSM_office_logistics yago2geoo:OSM_amenity_pharmacy
yago2geoo:OSM_amenity_dentist} .
	FILTER(geof:distance(?x0GEOM, ?c0GEOM, units:meter) < 200).
	FILTER (geof:sfContains(?c1GEOM, ?x0GEOM)).
}
```

What rivers flow through Liverpool?



https://www.tomko.org/demo/

How can we evaluate geospatial QA systems?

- Benchmarks
  - GeoQuestions201 (Punjani et al., 2018)
    - GeoQA, revised version of GeoQA, Hamzei et al(2022).
  - GeoQuestion733 (forthcoming)
    - GeoQA2 (forthcoming)

Punjani, D., et al. "Template-based question answering over linked geospatial data." Proceedings of the 12th Workshop on Geographic Information Retrieval. 2018.

The GeoQuestions733 Benchmark (to be released soon)

- It consists of 733 geospatial questions, query and answers.
- The benchmark dataset was collected from students of the Department of Informatics and Telecommunications of the National and Kapodistrian University of Athens, in the context of an Artificial Intelligence undergraduate course and Knowledge Technologies graduate course taught by Prof. Manolis Koubarakis (academic year 2020-2021).
- The students were asked to include in their questions one or more features and various kinds of geospatial relations.
  - Distance relations
  - Topological relations
  - Cardinal direction relation
- The students were asked to provide a natural language question and write their GeoSPARQL queries and execute the queries over YAGO2geo knowledge graph.

Categories of Questions

- The questions in the benchmark GeoQuestions733 fall under the following categories:
- Asking for the attribute of a feature.
  - “Where is Loch Goil located?”
- Asking whether a feature is in a geospatial relation with another feature
  - “Is Liverpool east of Ireland?”
- Asking for features of a given class that are in a geospatial relation with another feature
  - “Which counties border county Lincolnshire?”
- Asking for features of a given class that are in a geospatial relation with any features of another class
  - “Which lakes are near streams?”

Categories of Questions

- Asking for features of a given class that are in a geospatial relation with an unspecified feature of another class which, in turn, is in another geospatial relation with a feature specified explicitly
  - “Which lakes are near streams in County Mayo?”
- As in category 3 to 5 plus more thematic and/or geospatial characteristics of the features that are expected as answers
  - “Which villages in Scotland have a population of less than 500 people?”
- Questions with quantities and aggregates and superlatives
  - “Which is the largest lake by area in Great Britain?”

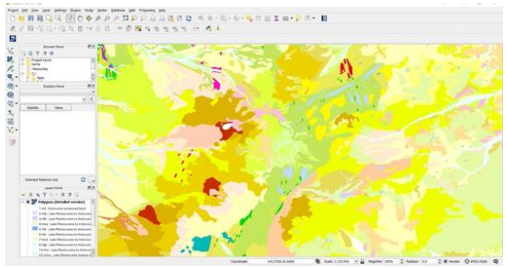
Performance evaluation (GeoQA2 vs. Hamzei et al.)

System	Generated Query (%)	Correctly Generated Query (%)
GeoQA2	92.6	**86.63**
Hamzei et al.	**96.2**	45.11

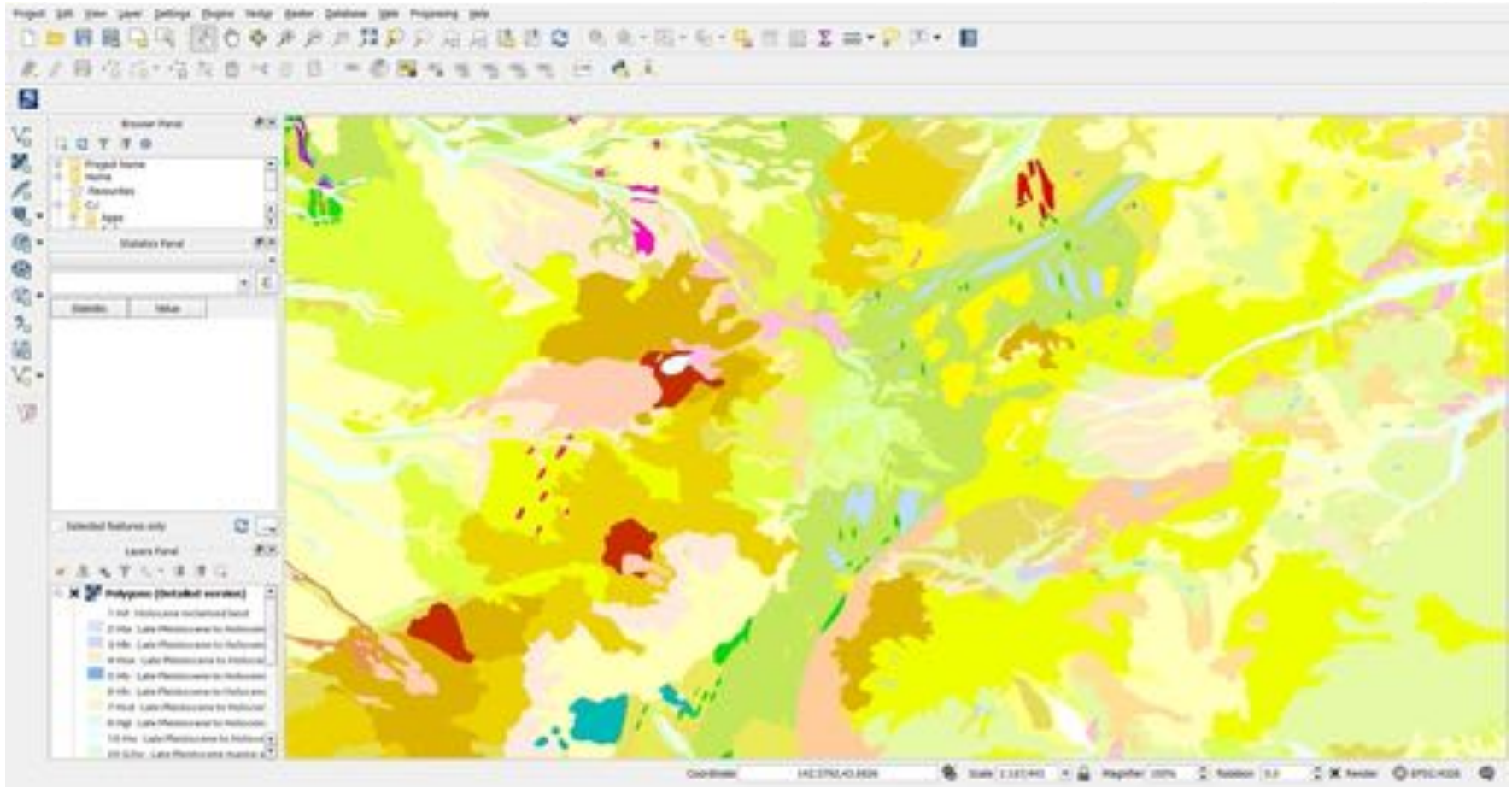
Searching, browsing, exploring and visualizing linked geospatial data

Geospatial Technologies

- Remote Sensing
- Geographic Information Systems (GIS)
- Global Positioning System (GPS)
- Digital Mapping Technologies



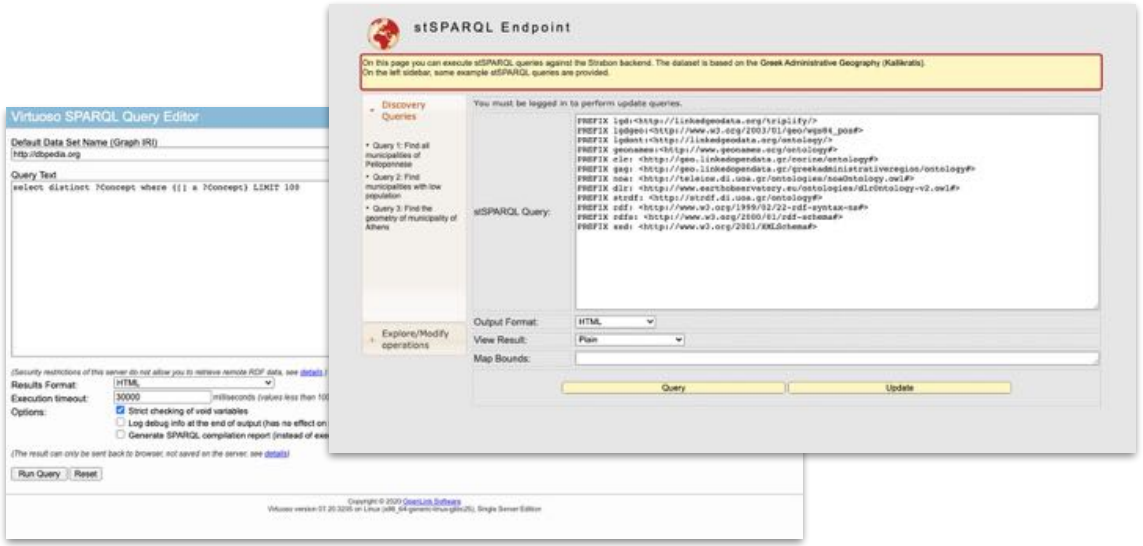
Visualizing geospatial data



Visualizing geospatial data

Visualizing linked geospatial data

- Explore and filter geospatial linked data sources
- Translate GeoSPARQL/stSPARQL queries into map layers
- Manipulate layers (e.g., zoom, filter, coloring)
- Combine layers into maps



Visualizing linked geospatial data

Map4RDF [Leon et al., 2012]

LinkedGeoData [Stadler et al., 2012]

SexTant [Bereta et al., 2013]

Spacetime [Ronchetti et al., 2014]

Facete [Stadler et al., 2014]

DBpedia Atlas [Valsecchi et al., 2015]

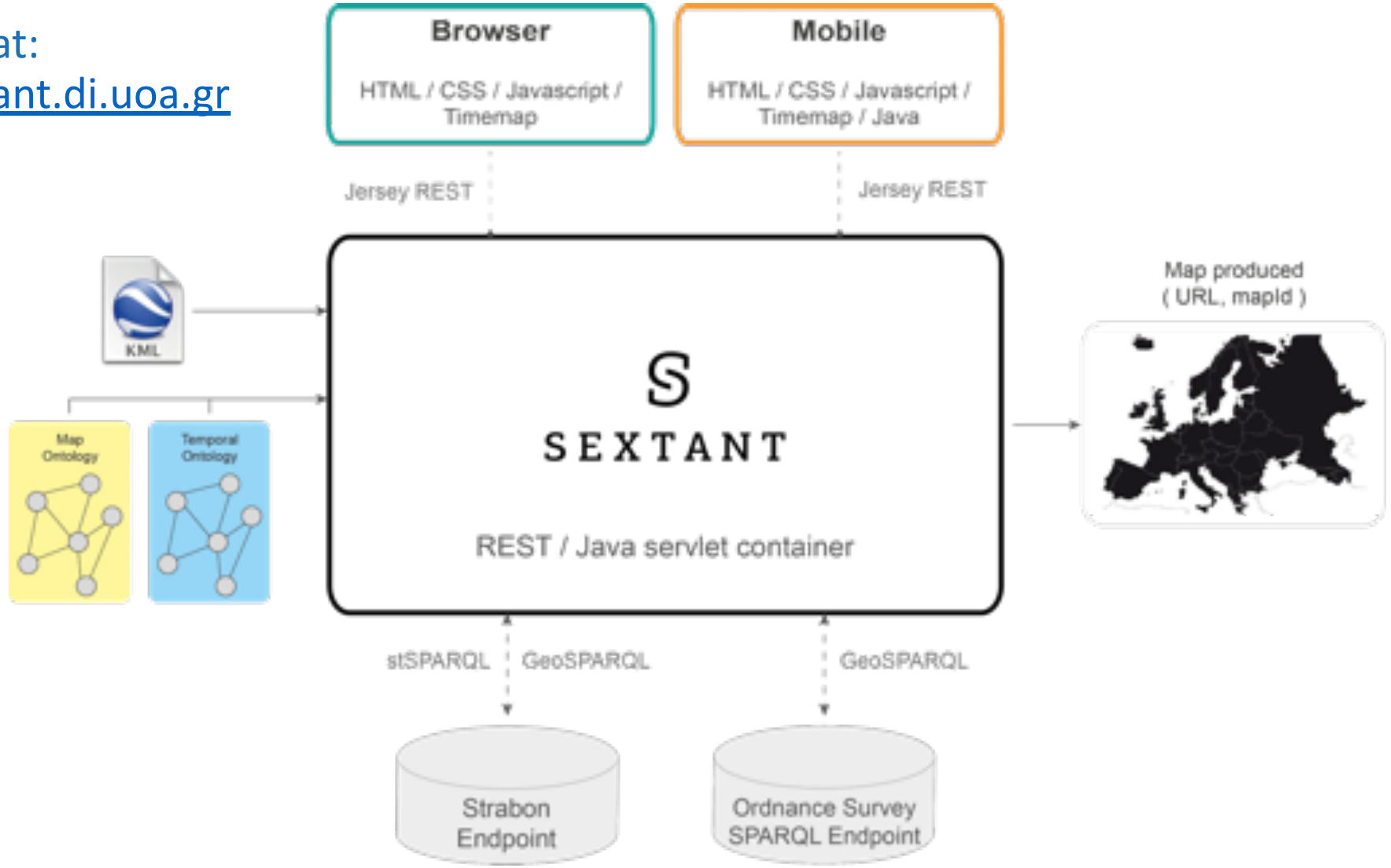
ESTA-LD [Mijovic et al., 2016]

GEOYASGUI [Beek et al., 2017]

GViz [McGlinn et al., 2019]

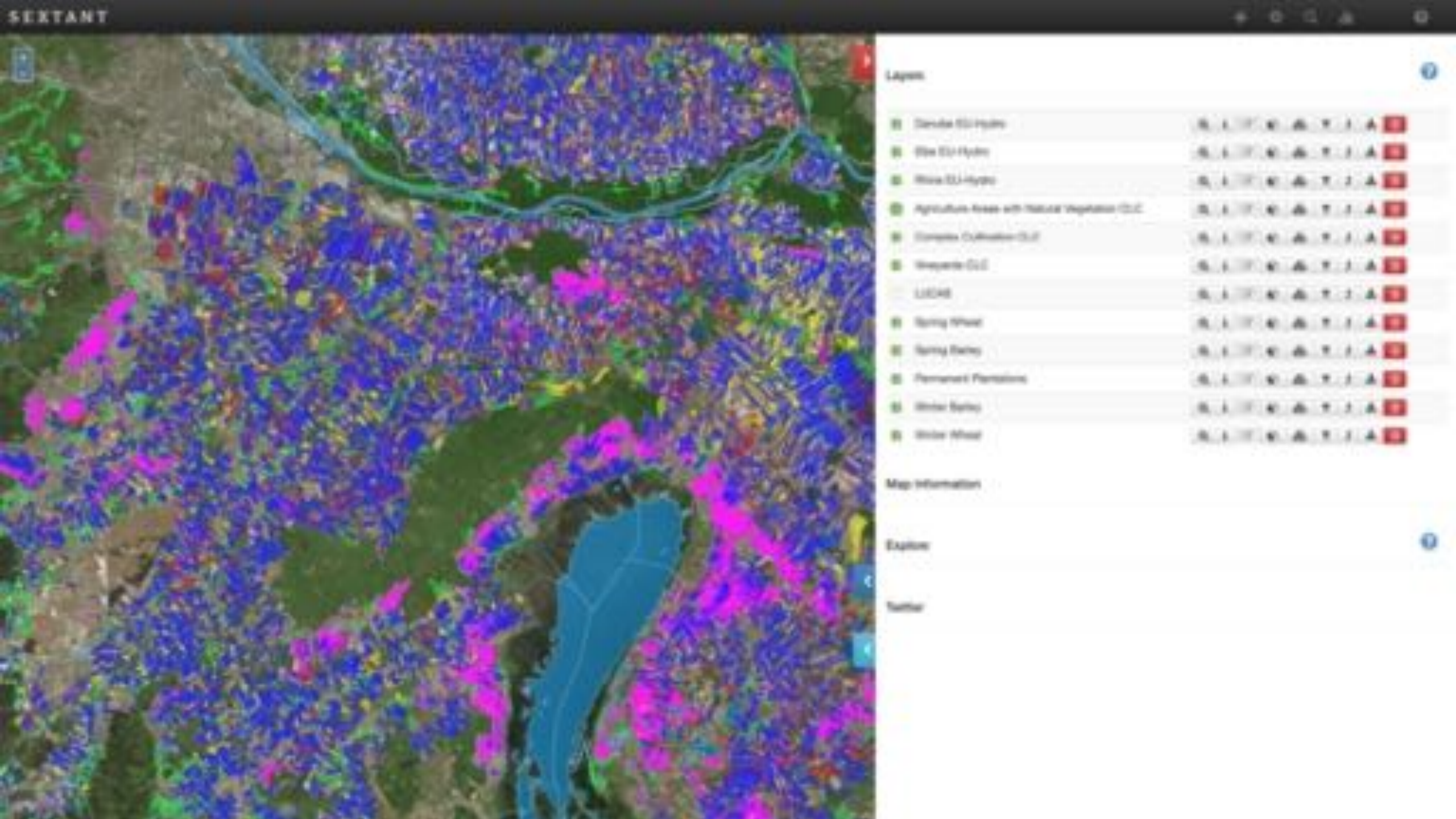
Sextant Architecture

Find more at:
http://sextant.di.uoa.gr



Thematic Maps

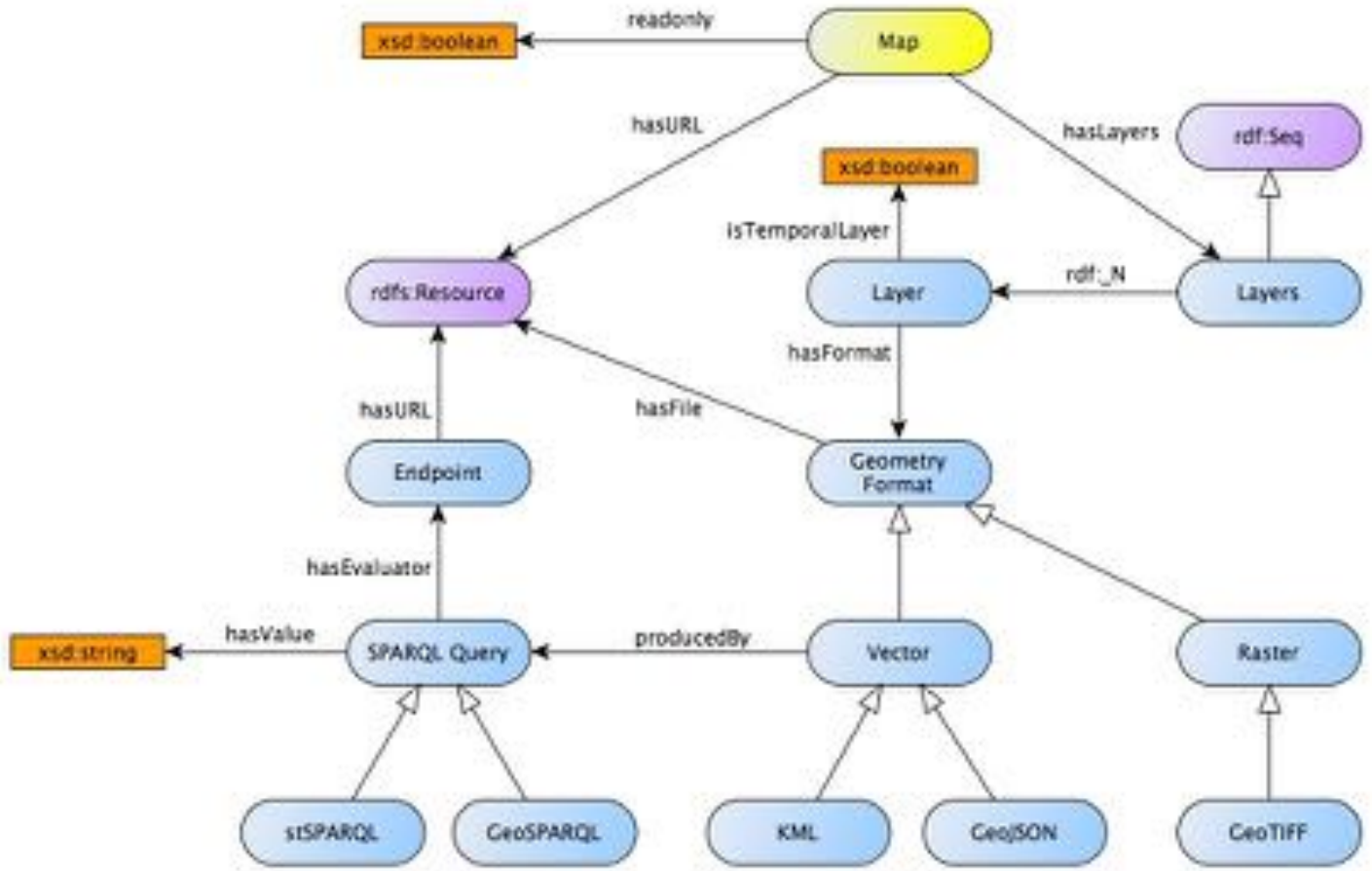
Definition: A thematic map is a type of map designed to show a particular theme connected with a specific geographic area. These maps can portray physical, social, political, cultural, economic, sociological, agricultural, or any other aspects of a city, state, region, nation, or continent.


SEXTANT
Layers
Map Information
Explore
Toolbar

Map Ontology

The ontology allows us to describe each map in the RDF format and allows us to store the information in an RDF store and query it using SPARQL.

This enables the construction of Map Registries as dedicated SPARQL endpoints that store our maps.



Creating Layers

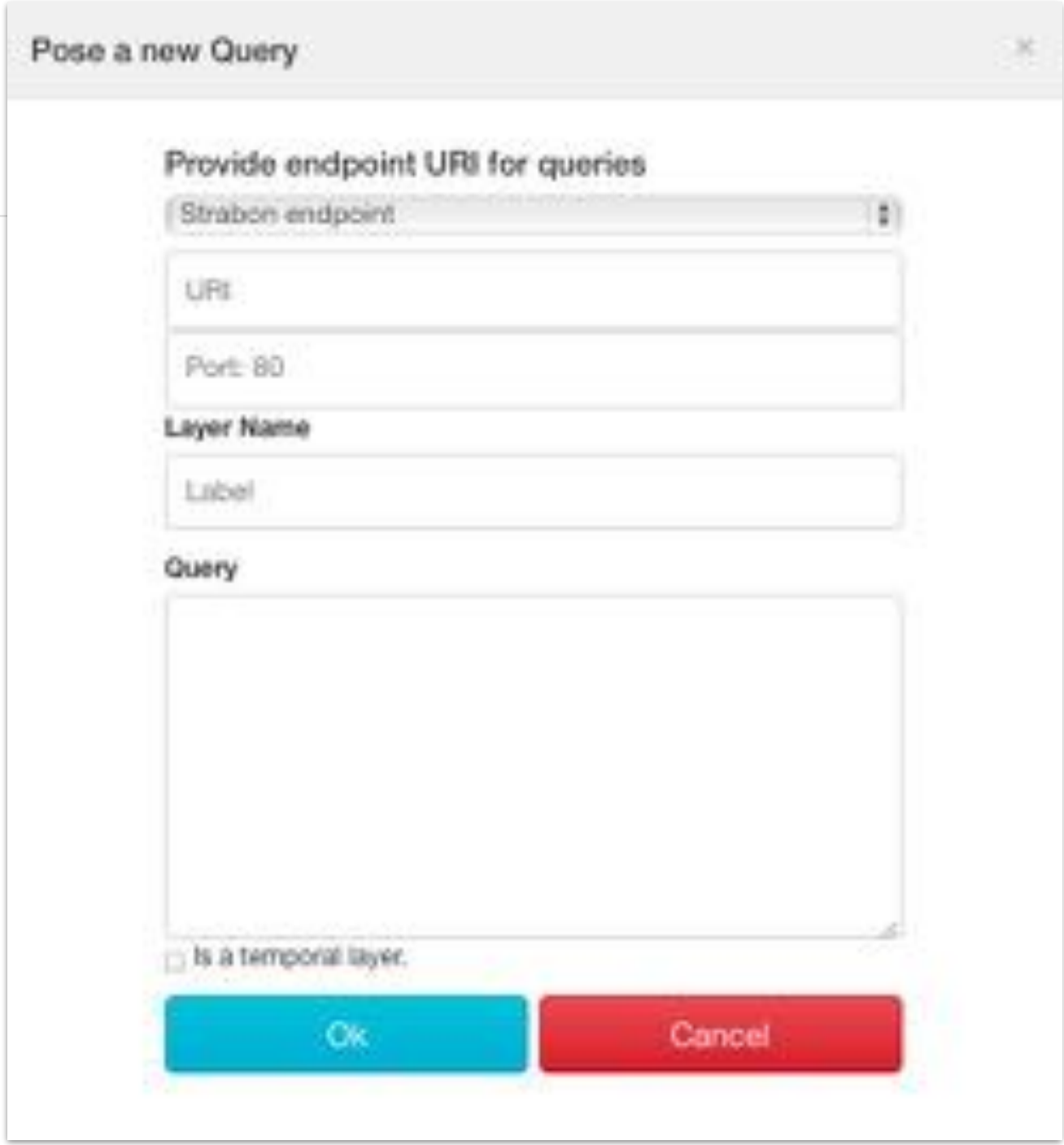
The core feature of Sextant is the ability to create thematic maps by combining geospatial and temporal information that exists in a number of heterogeneous data sources:

- GeoSPARQL - stSPARQL endpoints
- KML
- GeoJSON - TopoJSON
- GeoTIFF
- WMS

Layers

GeoSPARQL and stSPARQL

We can pose GeoSPARQL or stSPARQL queries to endpoints and visualize the results as a layer on the map, using this modal in the UI of the application.

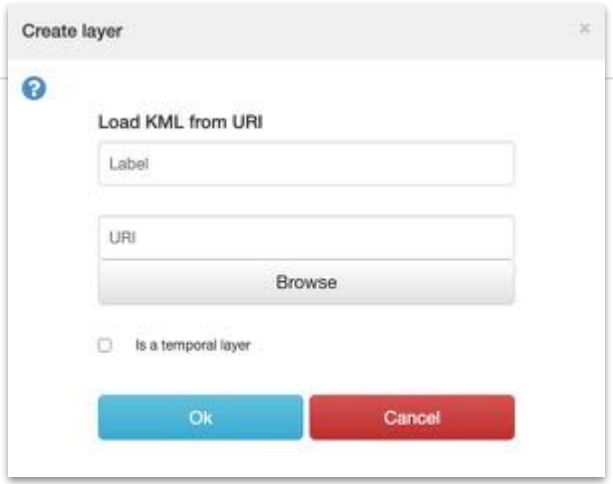


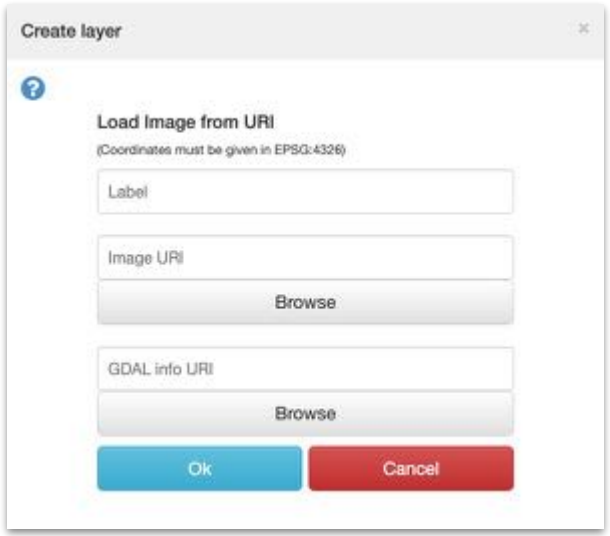
Layers

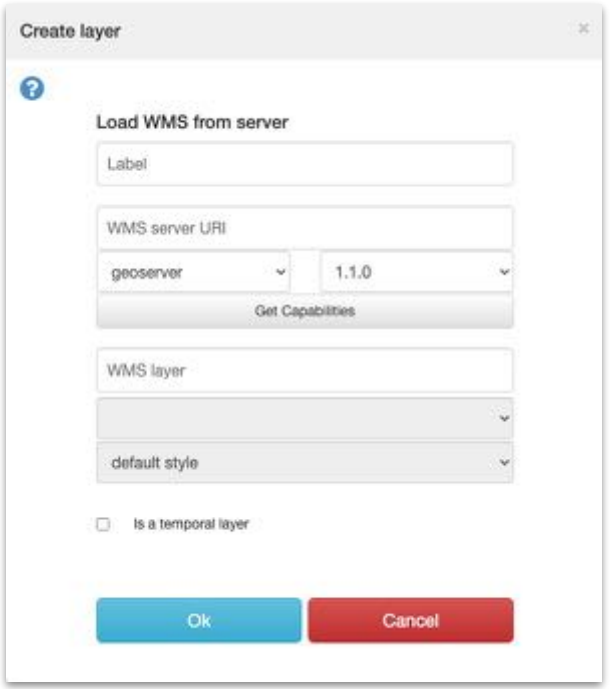
Popular GIS Formats

Users can also create layers utilizing existing popular GIS file formats, like KML, JSON, GeoTIFF and WMS.

Each format comes with a different modal in the UI of the application.







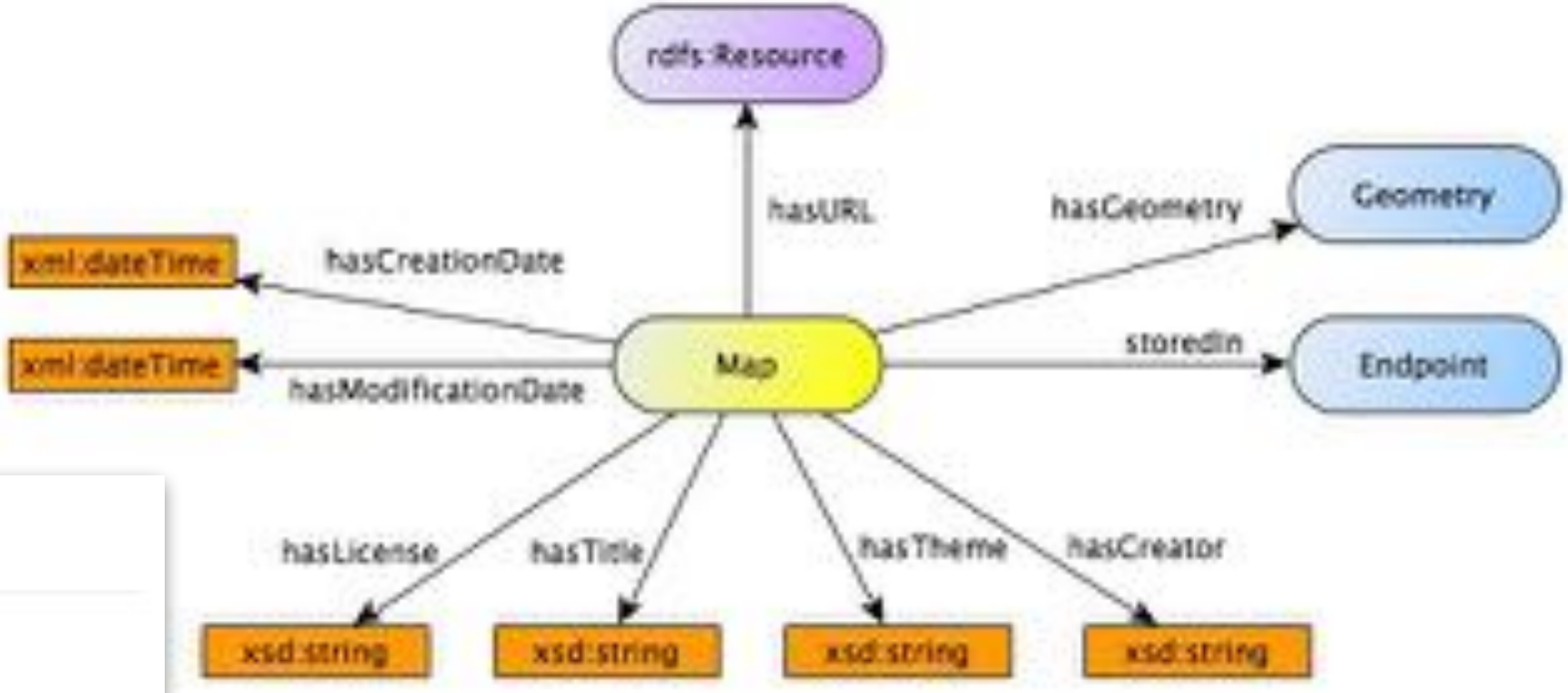
Layers

Layer functions

Each layer according to its type, can be further manipulated with some function buttons in the UI of the application:

- Zoom
- Info
- Update Query
- Global Styles
- Feature Styles
- Spatial Filter
- Move-on-top
- Download as KML

Map Metadata



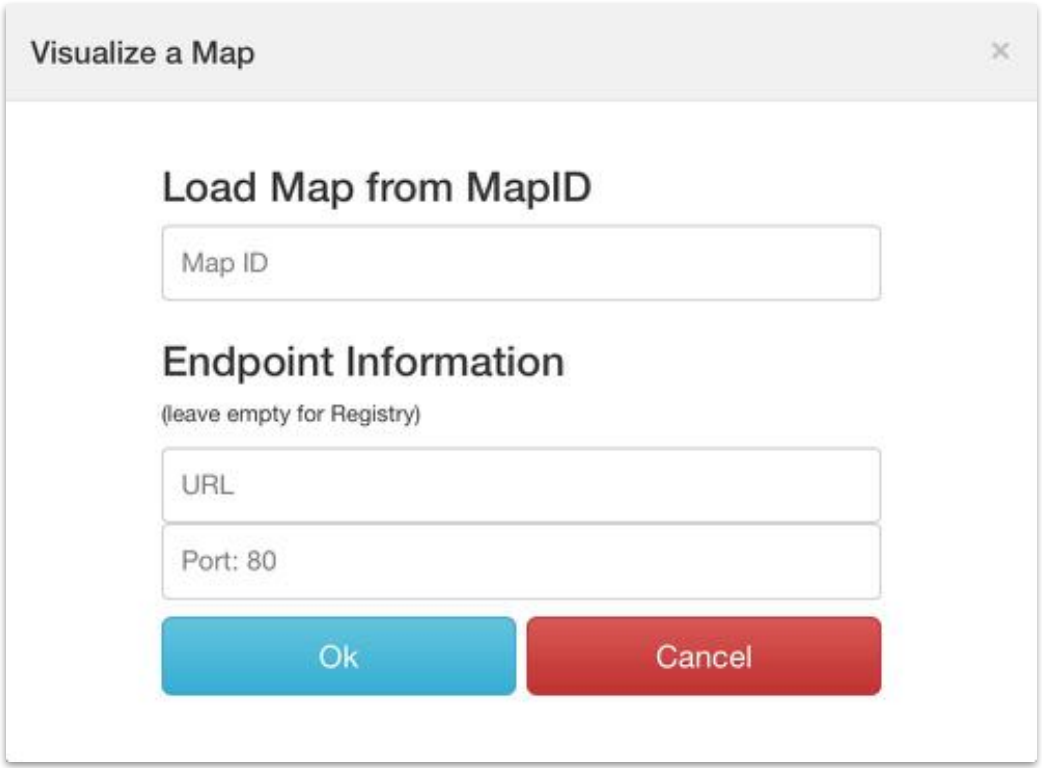
Map Information

Title:	testMap
MapID:	m8ib3721fpfugf9v_
Endpoint:	registry
Number of layers:	8
Number of charts:	3
Date of creation:	2015-03-27T17:24:06
Date of modification:	2015-03-27T21:53:03
Creator:	gstam
License:	none
Theme:	general
Extent:	<http://www.opengis.net/def/crs/EPSG/0/4326> POLYGON((33.66507289999983 26.62779254012927, 33.66507289999983

Map Sharing

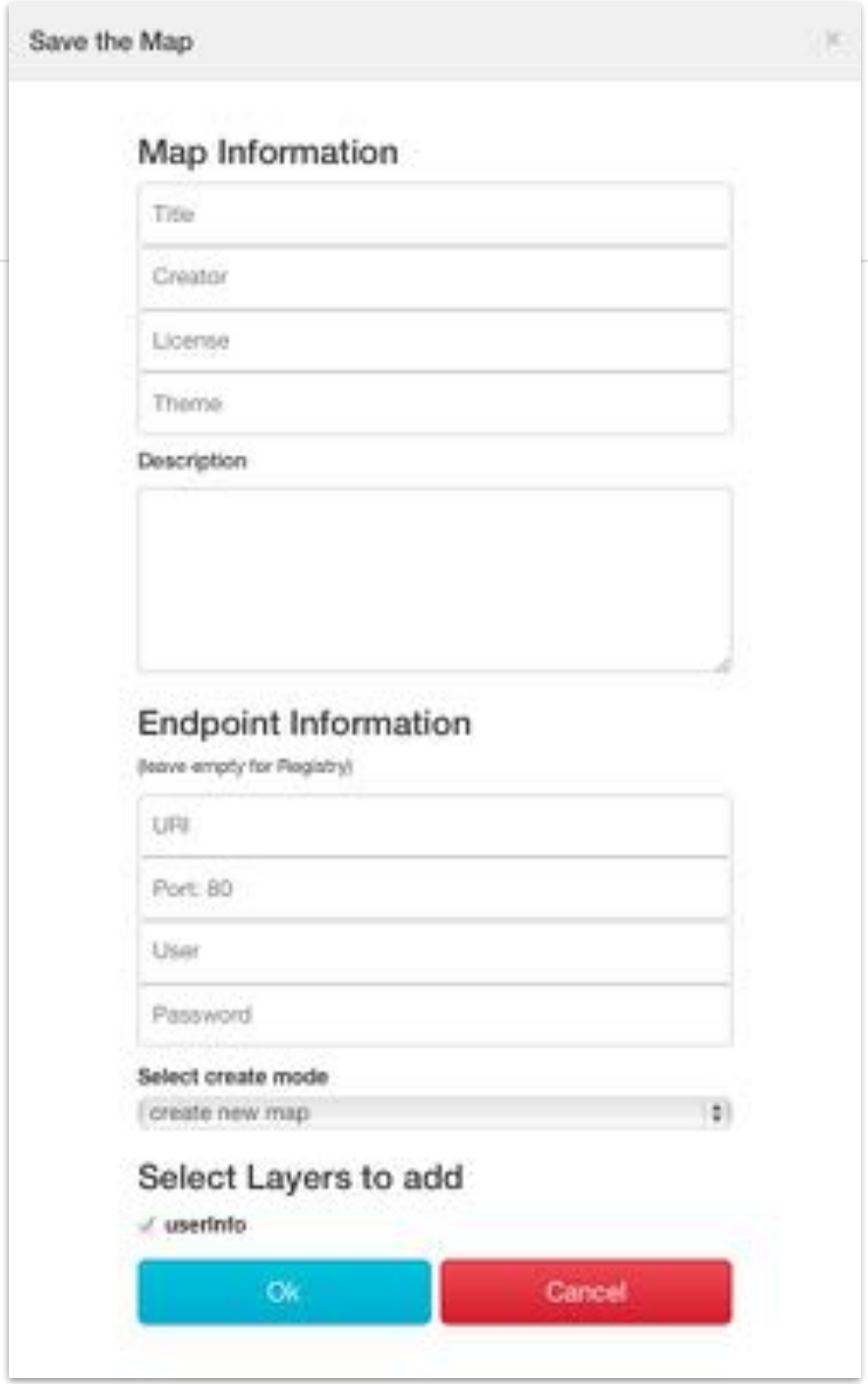
Share maps using map URI,
or use the load map modal in the UI.

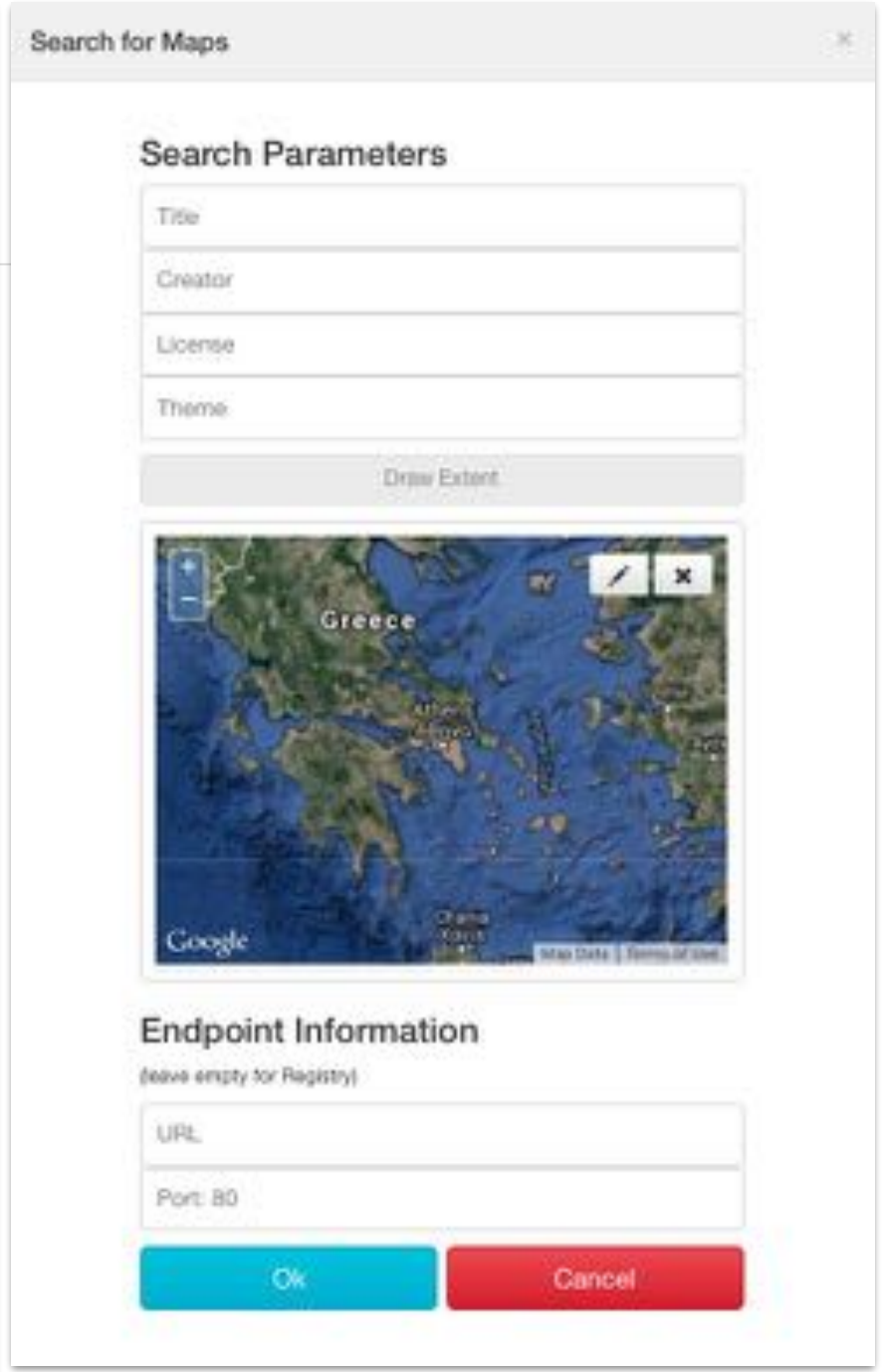
Map URI:
http://<domain>/Sextant/?mapid=<mapID>



Map Registry

Definition: Map registries, are SPARQL endpoints that hold all the map information and metadata to assist us in saving and retrieving the maps.

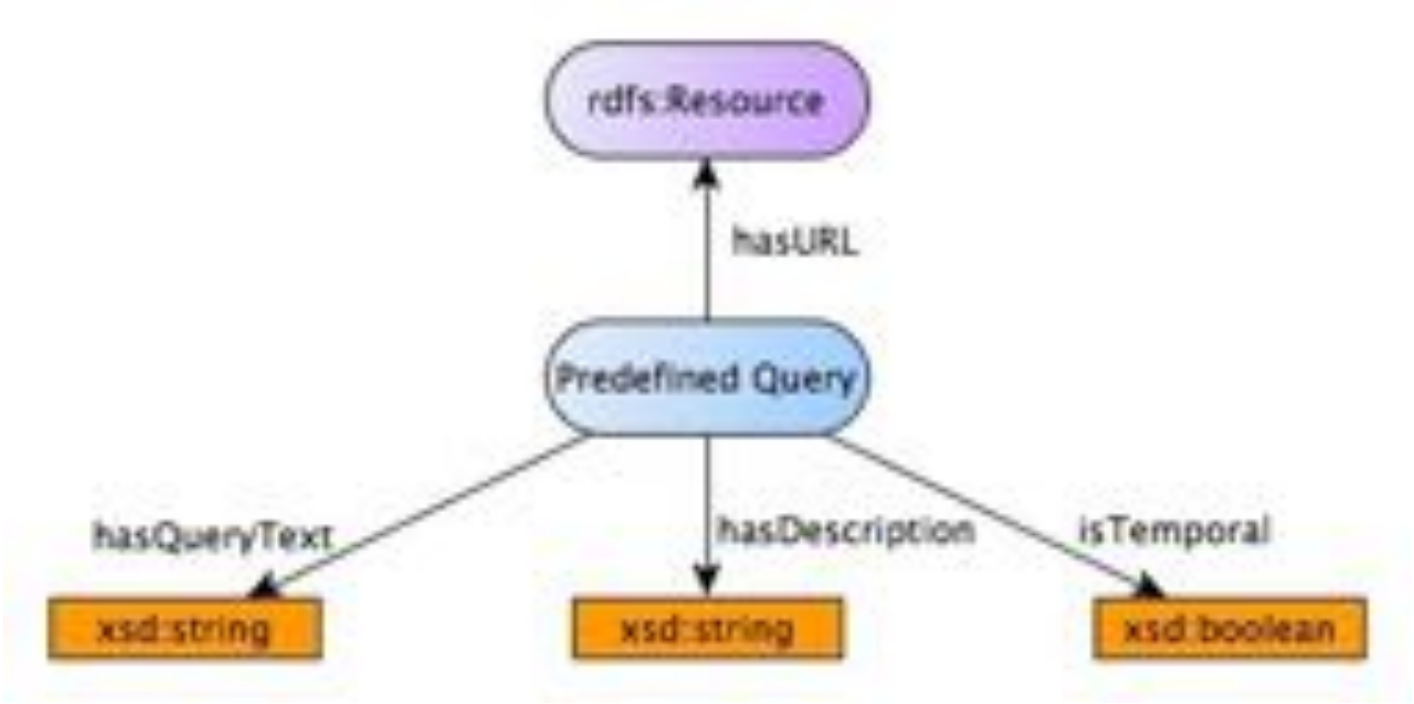




Predefined Queries

Queries that are created by an expert and are stores as triples in a SPARQL endpoint.

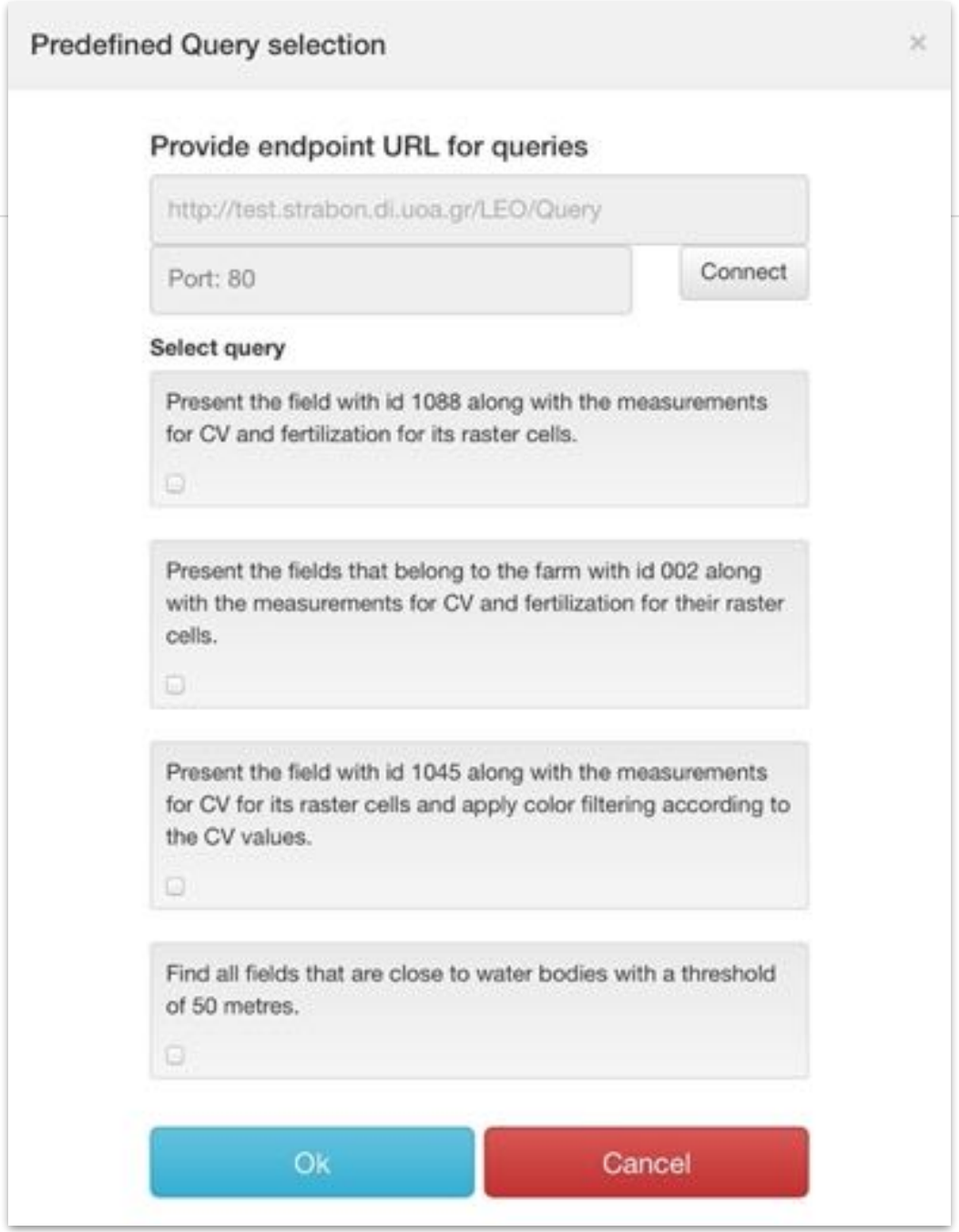
Non-expert users can provide the URL of the endpoint and get a list of the descriptions of all the predefined queries available, then select one and visualize it on the map.



Predefined Queries

Once the URL of a SPARQL endpoint is provided, the system searches for existing predefined queries and presents them to the user.

We can select each of the available queries to visualize on the map.

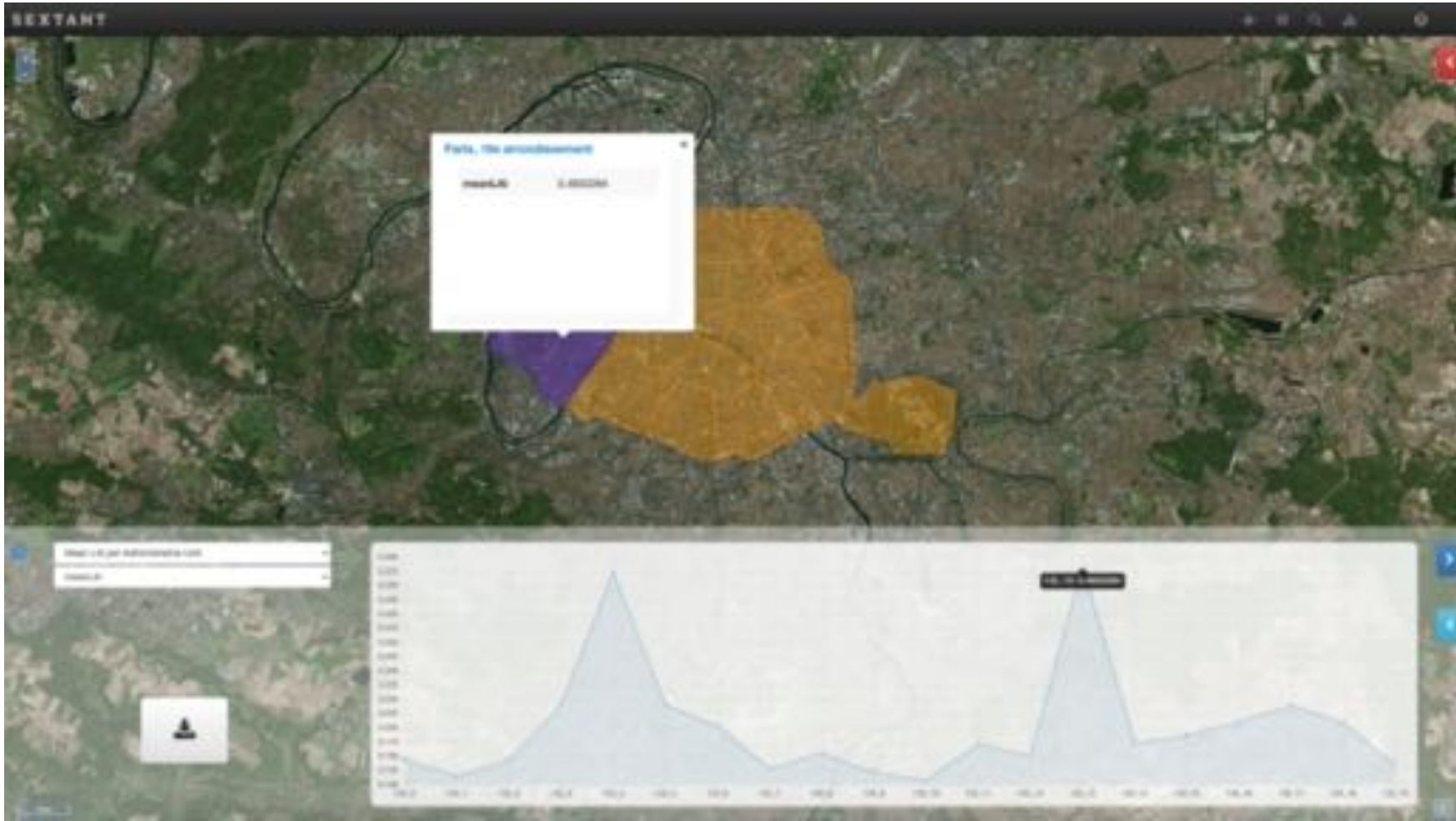


Statistical Charts

Automatic creation of charts over a layer’s attribute

Users select a layer and a numeric attribute from the UI to create an interactive chart.

Each value in the chart takes us to the geometric feature on the map.



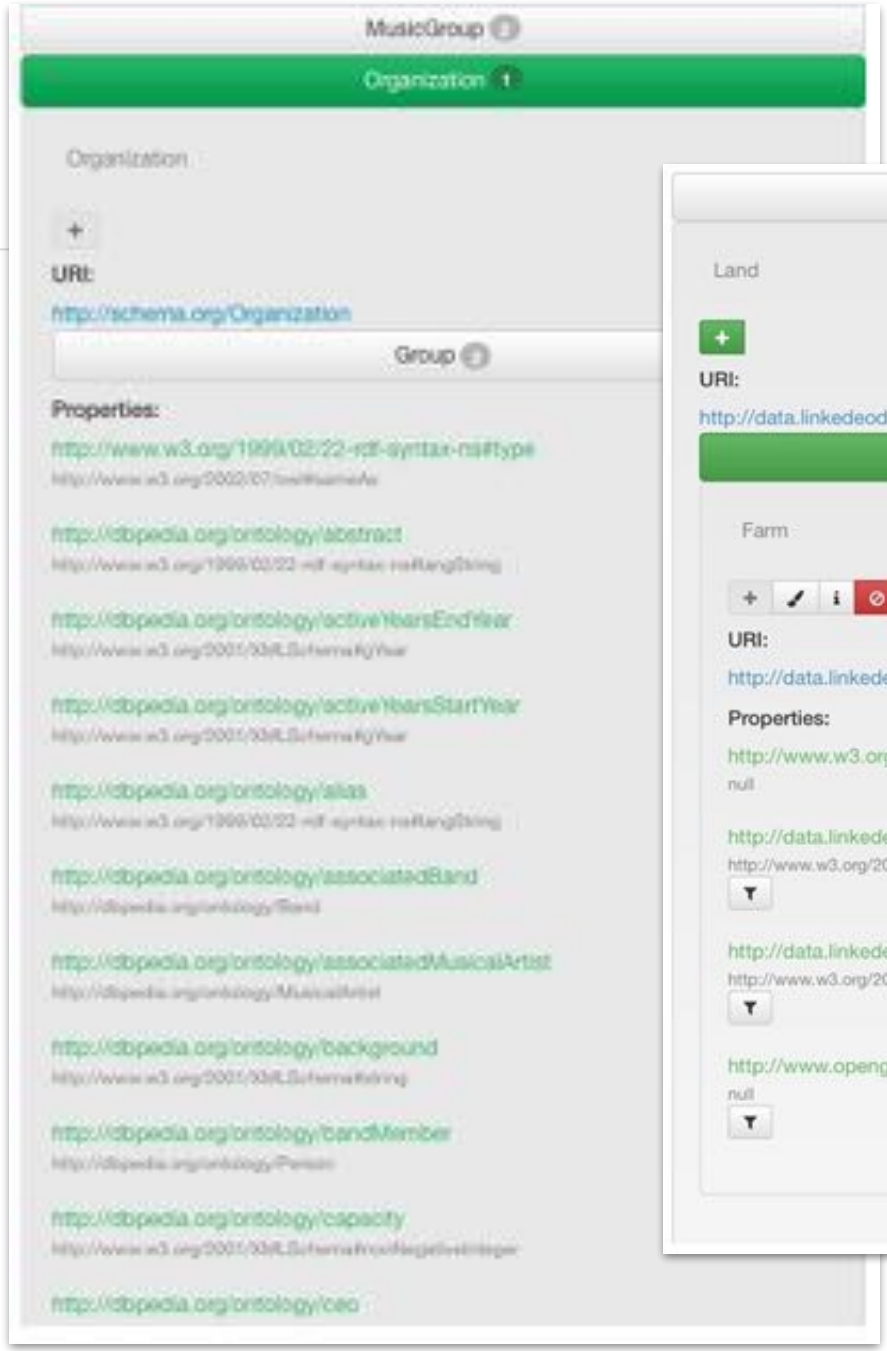
Explore Panel

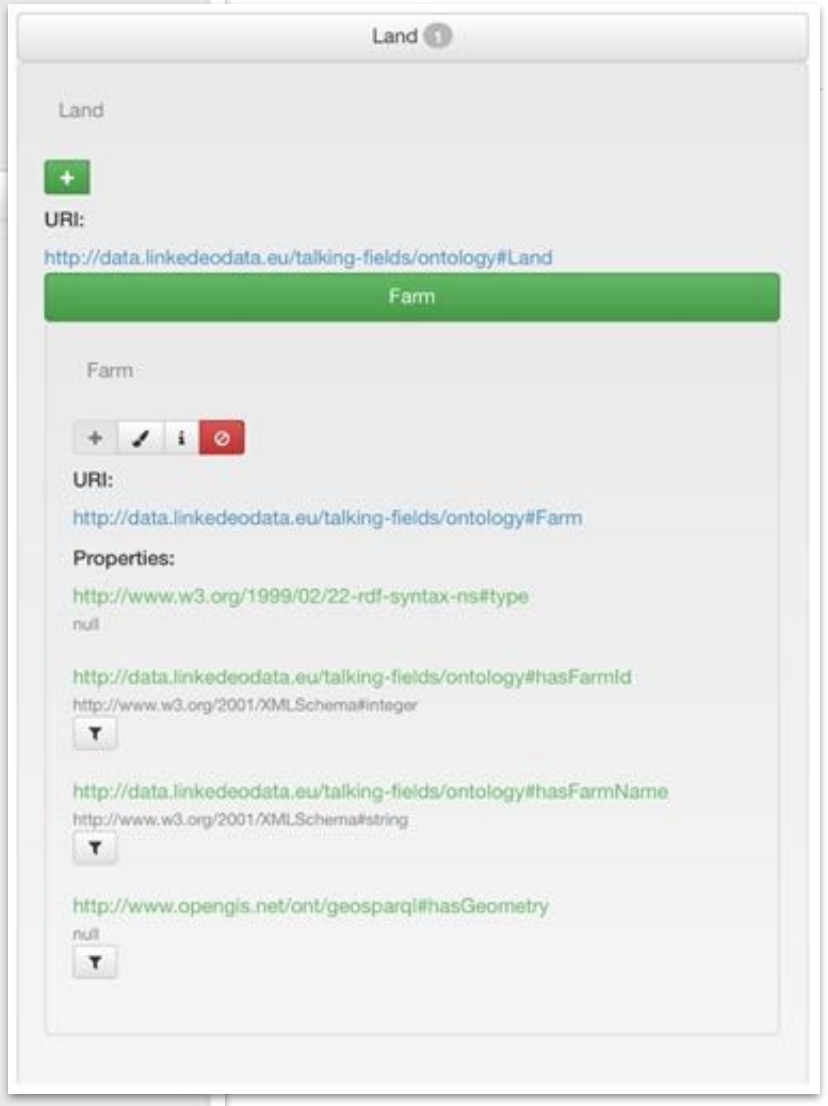
By providing the URL of a SPARQL endpoint in the Explore panel, we can view the underlying ontology in a tree form.

The number next to each class denotes the number of subclasses.

Explore

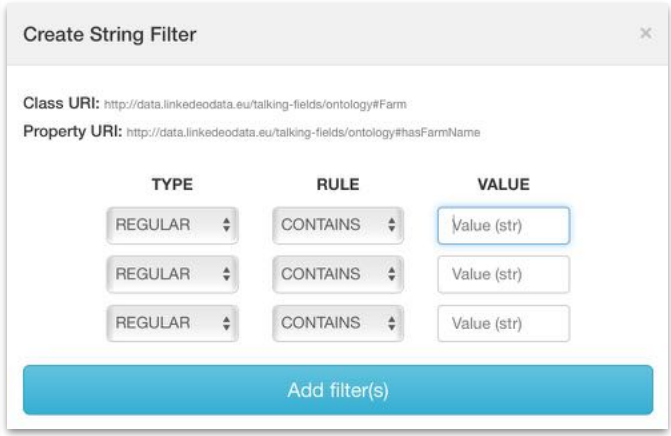
Classes and Properties

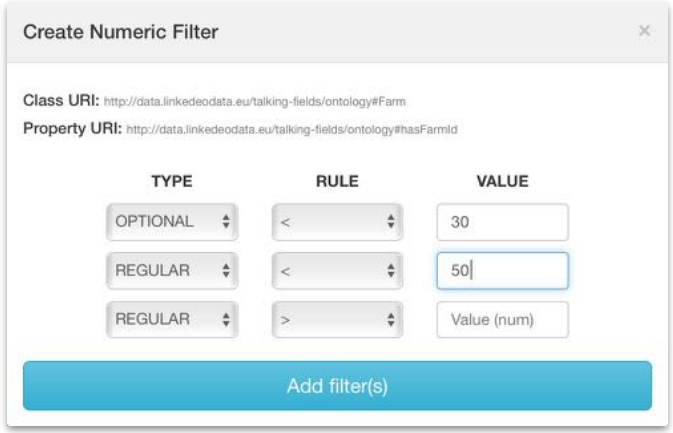


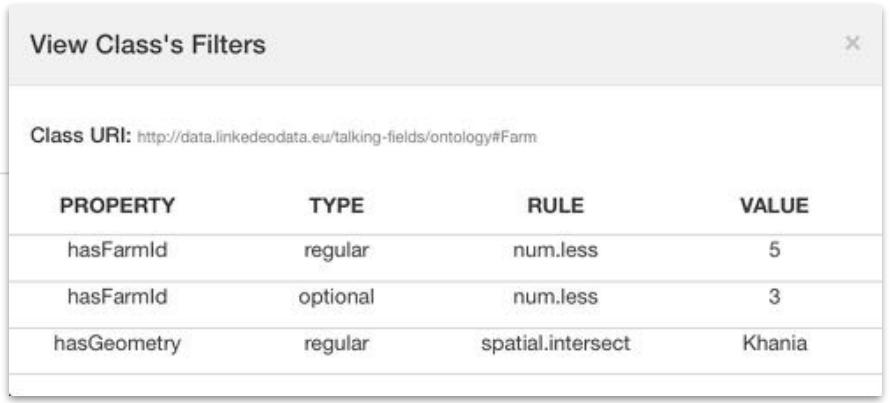


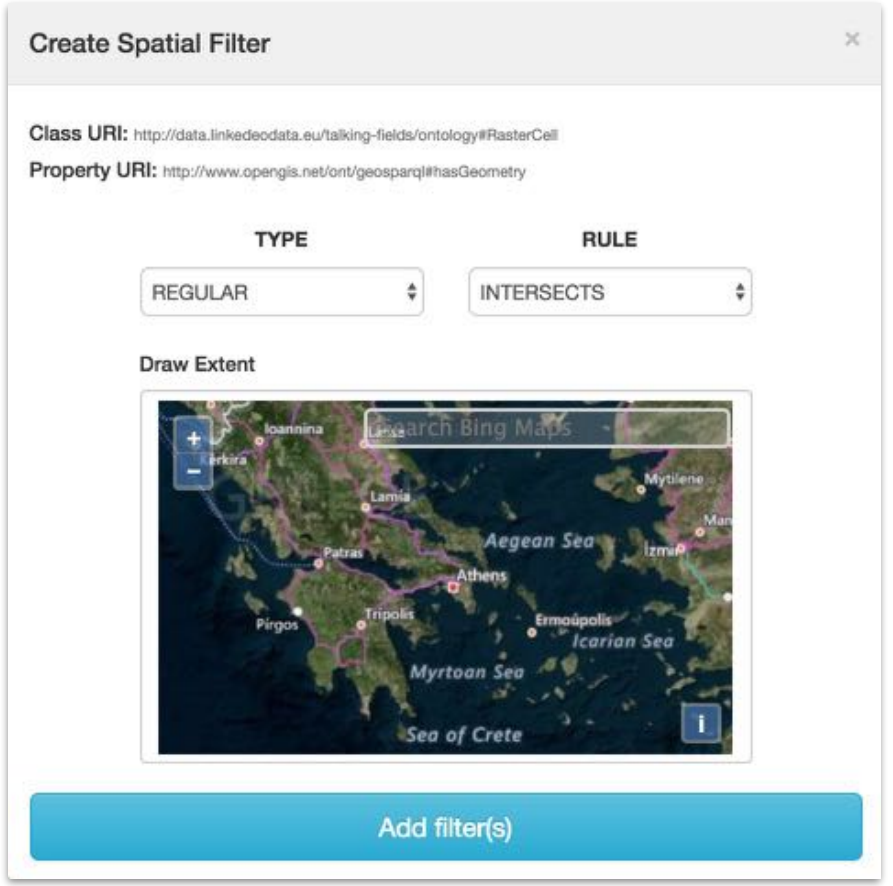
Explore

Visual Query Builder









Describe

Describe results

About: http://schema.org/Organization

Subject	Predicate	Object
http://dbpedia.org/resource/3Com	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://schema.org/Organization
http://dbpedia.org/resource/7-Eleven	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://schema.org/Organization
http://dbpedia.org/resource/Aardman_Animations	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://schema.org/Organization
http://dbpedia.org/resource/About.com	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://schema.org/Organization
http://dbpedia.org/resource/Academy_of_Motion_Picture_Arts_and_Sciences	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://schema.org/Organization
http://dbpedia.org/resource/Acorn_Computers	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://schema.org/Organization
http://dbpedia.org/resource/Activision	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://schema.org/Organization
http://dbpedia.org/resource/Ad_Lib,_Inc.	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://schema.org/Organization
http://dbpedia.org/resource/Adnams_Brewery	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://schema.org/Organization

Describe

Describe results

About: http://dbpedia.org/resource/Activision

Subject	Predicate	Object
http://dbpedia.org/resource/Activision	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#Thing
http://dbpedia.org/resource/Activision	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://dbpedia.org/ontology/Company
http://dbpedia.org/resource/Activision	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#Agent
http://dbpedia.org/resource/Activision	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#SocialPerson
http://dbpedia.org/resource/Activision	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.wikidata.org/entity/Q24229398
http://dbpedia.org/resource/Activision	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.wikidata.org/entity/Q43229
http://dbpedia.org/resource/Activision	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://dbpedia.org/ontology/Agent
http://dbpedia.org/resource/Activision	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://dbpedia.org/ontology/Organisation

Future Work

- New version of Sextant that offers both a frontend UI and a backend API to access data layers as a service
- New visualization tool that allows faceted browsing of data sources and allow non-expert users to construct custom datasets based on their selections over different resources

Open Questions

Open Questions

- How can we build GeoSPARQL and GeoSPARQL+ query processors that **scale to the extreme volumes of geospatial data, information and knowledge**:
  - Owned by a national cartographic agency
  - Produced by Copernicus

- Strabo 2 is a good start! 😊

EXTREME EARTH

Open Questions (cont'd)

- How do we build **spatiotemporal query answering systems** that scale to extreme volumes of spatiotemporal data?
- The work of Nikitopoulos et al. (2021) is a good start!

P. Nikitopoulos et al. Parallel and scalable processing of spatio-temporal RDF queries using Spark. GeoInformatica 25(4): 623-653 (2021)

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
UNIVERSITY OF PIRAEUS

Open Questions (cont'd)

- How do you develop **question answering systems for the dataset discovery step of our pipeline** (discovering Earth Observation datasets)?
- See our forthcoming **EarthQA** system!

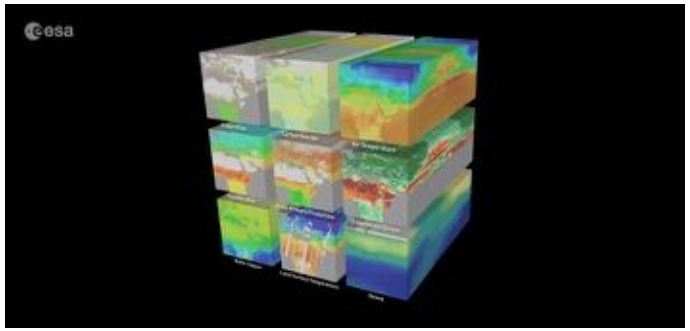
AI4 copernicus

Open Questions (cont'd)

- How do we build **spatiotemporal question answering systems** over **spatiotemporal knowledge graphs**?
- See forthcoming work of our team with colleagues from L3S, Hannover and University of Bonn.

Open Questions (cont'd)

- How do we develop **semantic data cube systems**?
- See our forthcoming system Plato.



D. Bilidas et al. Plato: A semantic data cube system. Forthcoming.

DEEP CUBE

Thanks! Questions?

- Thanks to all our colleagues for their contributions.
- For more, see our web pages:
  - http://ai.di.uoa.gr for Manolis, Despina, Dimitris, George, Theofilos, Dharmen and George.
  - https://rsim.berlin/ for Begüm.

Thank you!