Cartography M.Sc.

Integrating Geospatial Linked Open Data and Knowledge Networks into Business Intelligence

OLHA OSTROVNA

Supervisors: Dipl.-Ing. Dr. Markus Jobst
Univ.Prof. Mag.rer.nat.
Dr.rer.nat. Georg Gartner
Dr.-Ing. Ekaterina Chuprikova

29.09.2020

Outline

1. Introduction

Motivation and problem statement
Research goals and objectives
Research questions

2. Theoretical background

Geospatial Linked Open Data and Knowledge Networks
Location Business Intelligence

3. Methodology

Thematic scenarios development
Overview of LOD providers and leaders of Location BI market

4. Implementation, Results and Discussion

Data Access Interfaces in Linked Data providers
Geospatial Linked Data integration within visualization frameworks in Business Intelligence tools

5. Conclusion

Motivation and problem statement

Geospatial Linked Open Data

- Interlinks location data with other data on the Web.
- Prevents data of being hidden and isolated.
- Increases accessibility, transparency, reuse, interoperability and value of open data.
- Combines a large variety of data and queries data at the source.

Location Business Intelligence

- Helps to understand the significance of location information.
- Offers powerful functionalities and technologies for data processing.
- Serves as an optimal Graphical User Interface (GUI) for visualization and interaction.
- Provides end-users with meaningful insights and industry-specific solutions.

Geospatial LOD + **Location BI**

= **use-cases**, **valuable decisions** and **solutions across domains**

BUT, integration gap between **Geospatial LOD providers** and **Location BI software !**

Research goals and objectives

1. Analysis of current **leading Business Intelligence tools** in terms of **Linked Open Data Integration**

2. Analysis of relevant **Linked Open Data providers** in terms of **integration with BI tools**

3. Visualization framework for the integrated **Linked Open Data** within the selected **BI tools**

4. Identified **reasons** limiting the **integration** of **Geospatial LOD** to **BI software**

Research questions

What **data transfer technologies** currently define the leading **BI tools integrations' support**?

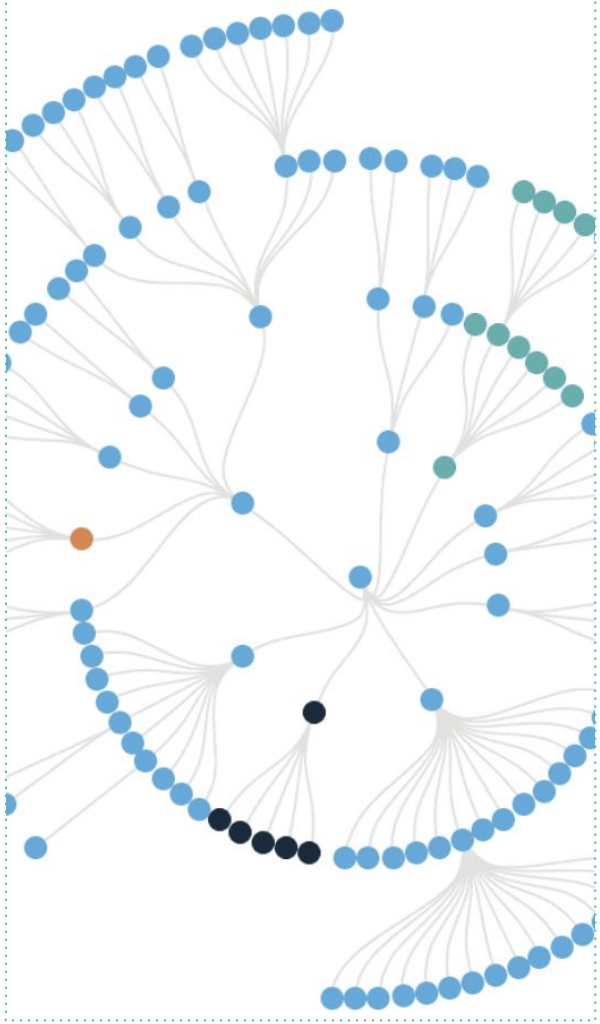
What specific **data transfer technologies** have to be in focus for the **Linked Open Data platforms** to provide **successful integration use-cases**?

What are the key components of the **Linked Open Data concept** required to be integrated in **BI software** to provide access to the **global Geospatial Linked Open Data** and **Knowledge Network platforms**?

Geospatial Linked Open Data and Knowledge Networks

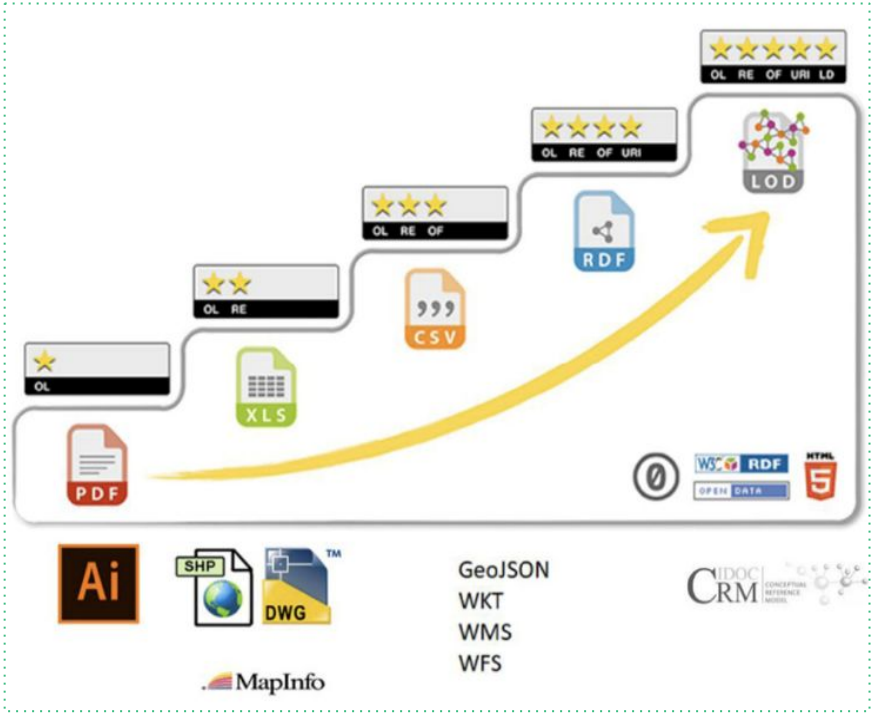
Linked Data refers to the provision of **integrated access** to data which is **interlinked** with other data from a wide range of distributed and heterogeneous data sources *[Bizer et al. 2009]*, collectively forming a **network of knowledge** *[Mangaladevi et al. 2017]*.

Knowledge Networks are expressed through graphs of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent **relations** between these entities *[Hogan et al. 2020]*.



Linked Data *[PoolParty Demo]*

Geospatial Linked Open Data and Knowledge Networks



The Five Star Open Data model *[Berners-Lee 2006; Berners-Lee 5]* applied to spatial data types *[McKeague et al. 2020]*

★ Available on the web under an open license

★★ Available as machine-readable structured data in proprietary formats

★★★ Available as machine-readable data, but in a non-proprietary format (e.g. CSV instead of XSLT)

★★★★ Using open web standards from W3C (IRIs for identifiers, RDF for data model, SPARQL for querying)

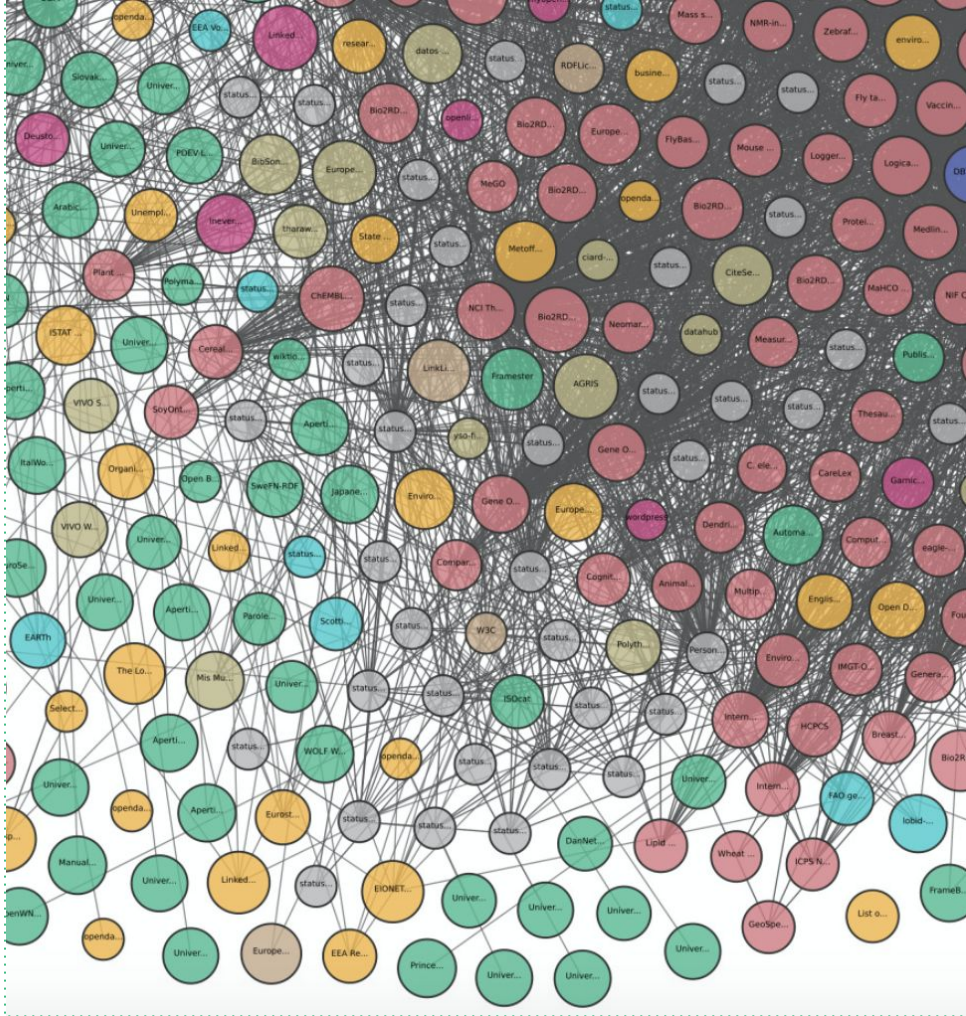
★★★★★ Linked to other Linked Open Data on the web providing content

Geospatial Linked Open Data and Knowledge Networks

Geospatial Linked Open Data is **location data interlinked** with other Web data and available under **open licenses** from a wide range of distributed and heterogeneous data sources using **RDF*** and **SPARQL**** technologies.

* **RDF** [Resource Description Framework]

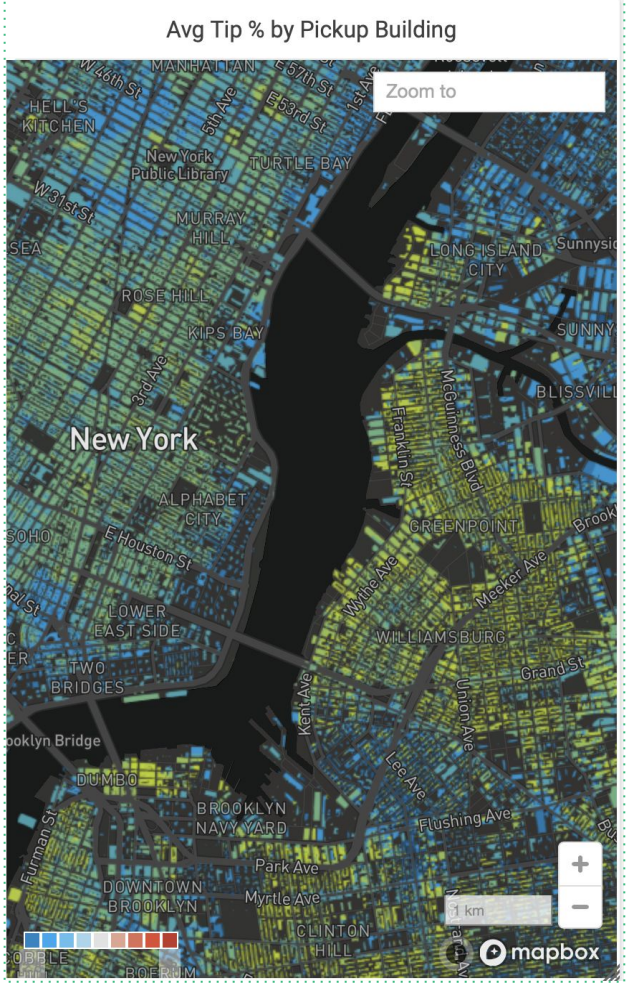
** **SPARQL** [Query language for linked data]



The LOD Cloud *[LOD Cloud Demo]*

Location Business Intelligence

Location Business Intelligence (Location BI) is a broad category of **applications**, **frameworks** and **technologies** of combining **geospatial data** with **business data** & **processes** in order to create a **user-** and **industry-specific solutions** and help users make **better decisions**.



BI Tools *[OmniSci Demo]*

Thematic Scenarios Development

1

Potential impact of COVID-2019 crisis on EU labor markets

Used data:

Business demography and high growth enterprise by NACE rev.2 and NUTS 3 regions

Population in EU countries by regions

Area data for NUTS-3 regions

NUTS-3 regions geometry for EU countries

2

Closeness of bike-sharing services to the areas of Alpine ibex colonies concentration within Switzerland

Used data:

Bike-sharing and bicycle hire locations in Switzerland

Distribution of ibex colonies in Switzerland

Overview of LOD providers and leaders of Location BI market

Linked Open Data providers

eurostat

EUROPEAN DATA PORTAL

Business Intelligence tools

tableau

esri

CARTO

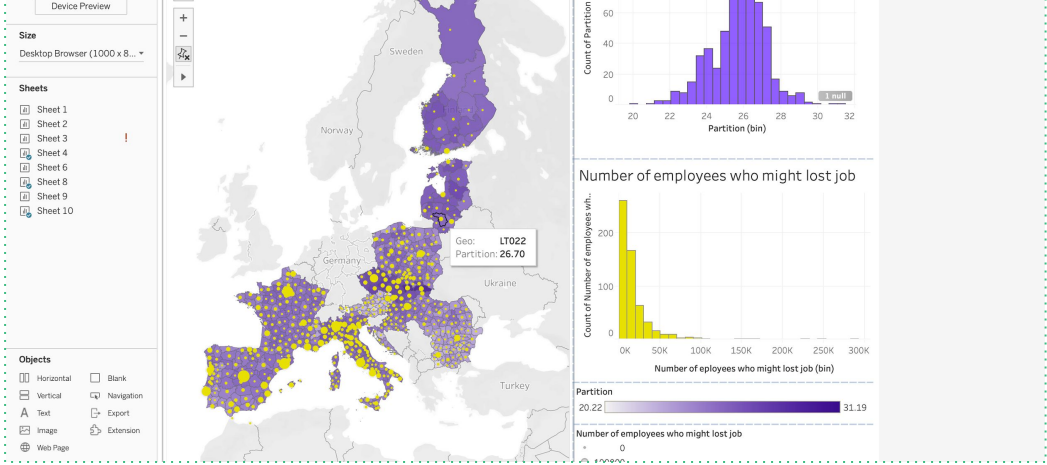
omni·sci

+ a b | e a u

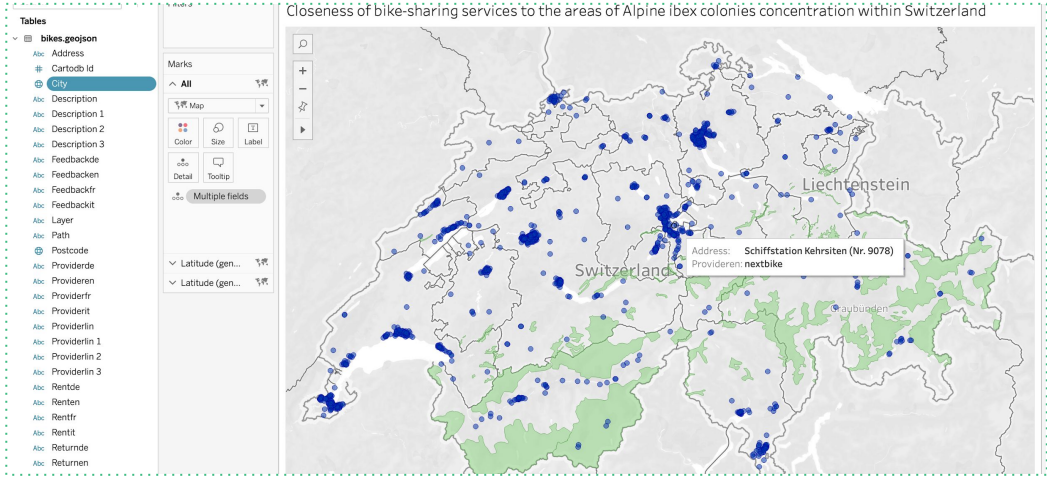
Data standards according to The Five Star Open Data model applied to spatial data formats available within Tableau

- ★ PDF*
- ★★ XLSX*, Esri Shapefile, MapInfo, Geodatabase
- ★★★ CSV*, TXT*, GeoJSON, TopoJSON, KML, KMZ
- ★★★★ SPARQL
- ★★★★★

* Although, PDF, XLSX, CSV, TXT are not geospatial data formats by their nature, they might be used for storage of geospatial data



Potential impact of COVID crisis on EU labour markets



Bike-sharing with Alpine ibex colonies concentration

CARTO

Data standards according to The Five Star Open Data model applied to spatial data formats available within CARTO.

★

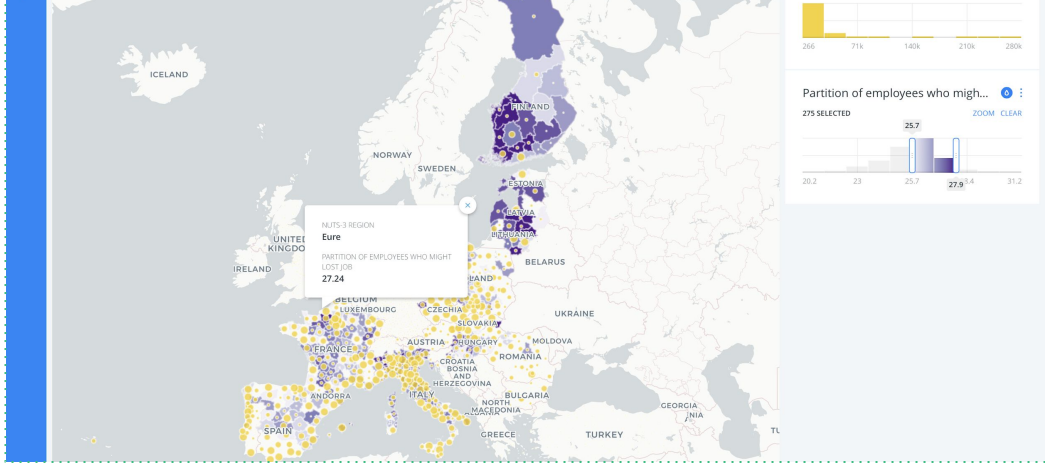
★★ XLSX*, Esri Shapefile, MapInfo, Geodatabase

★★★ CSV*, GeoJSON, KML, KMZ, OSM, Geopackage, GPX, WMS, WMTS

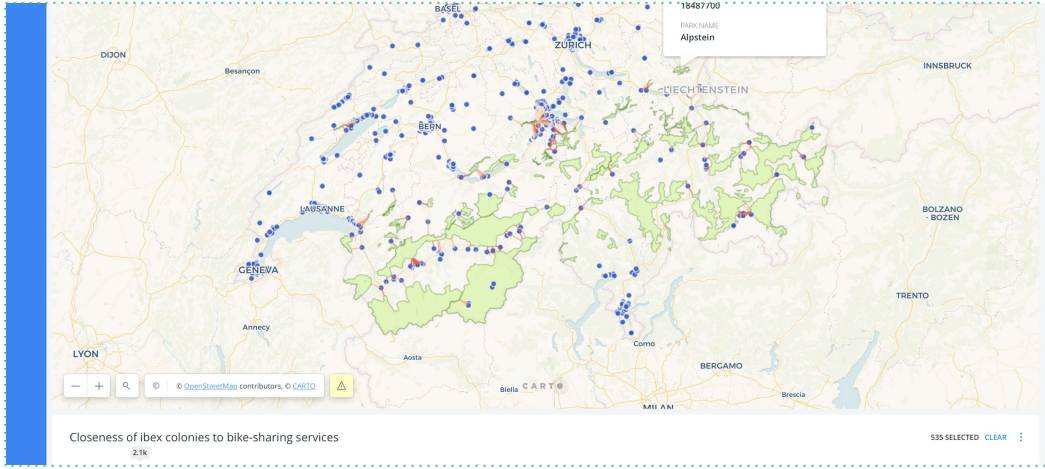
★★★★

★★★★★

* Although, XLSX, CSV are not geospatial data formats by their nature, they might be used for storage of geospatial data



Potential impact of COVID crisis on EU labour markets



Bike-sharing with Alpine ibex colonies concentration

omni·sci

Data standards according to The Five Star Open Data model applied to spatial data formats available within OmniSci.

★

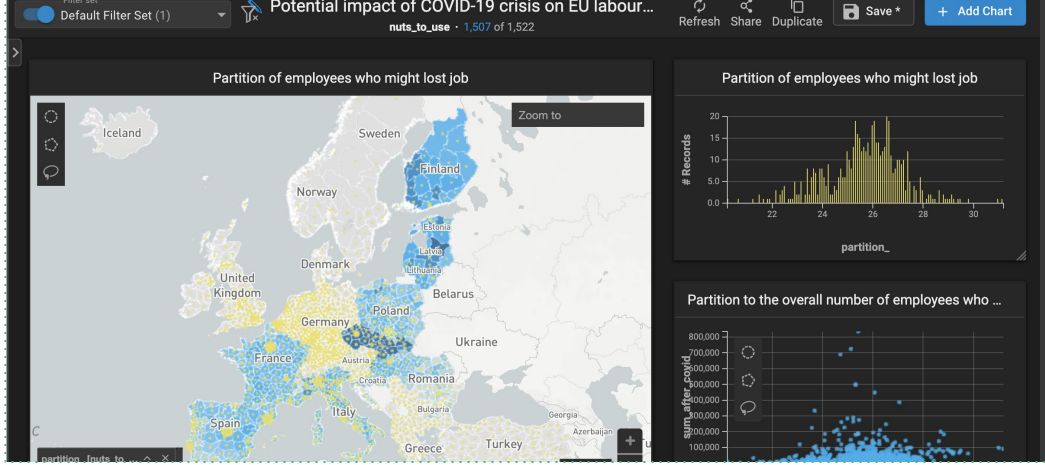
★★ Esri Shapefile

★★★ CSV*, TSV*, TXT*, GeoJSON, JSON, KML, KMZ

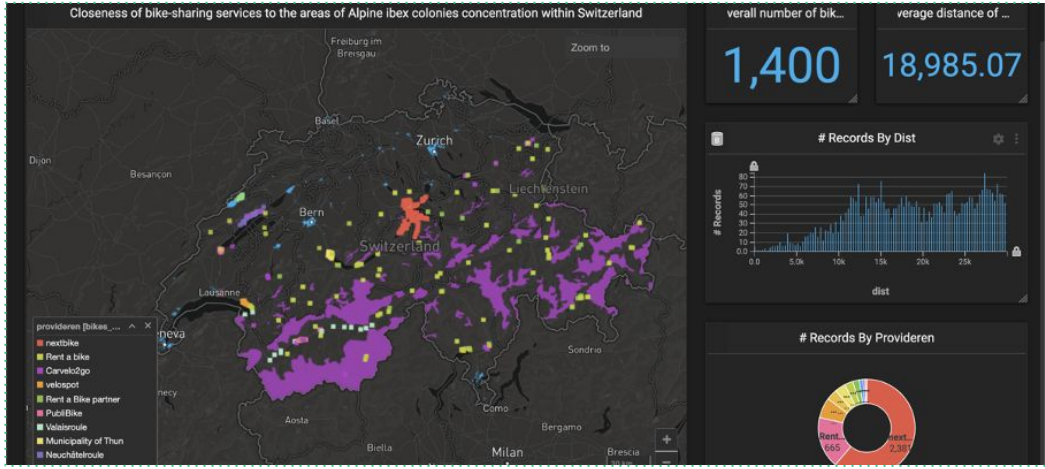
★★★★

★★★★★

* Although, CSV, TSV, TXT are not geospatial data formats by their nature, they might be used for storage of geospatial data



Potential impact of COVID crisis on EU labour markets



Bike-sharing with Alpine ibex colonies concentration

Data standards according to The Five Star Open Data model applied to spatial data formats available within ArcGIS Insights.

★

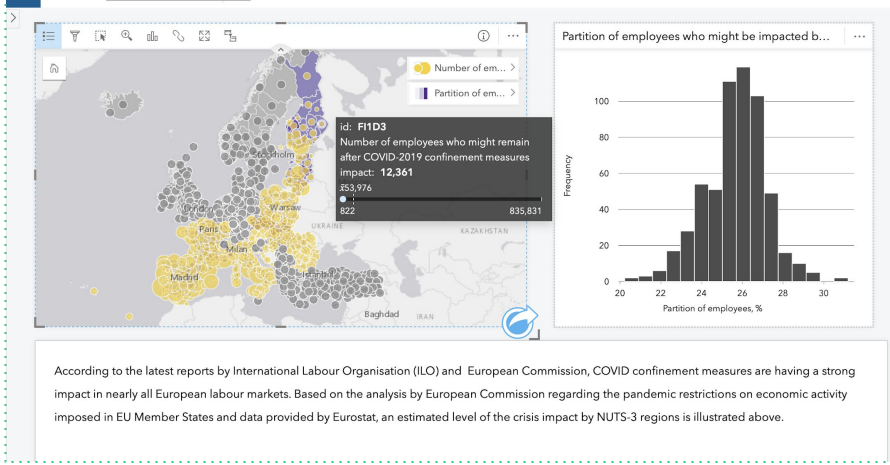
★★ XLS*, Esri Shapefile, Map image layers

★★★ CSV*, GeoJSON

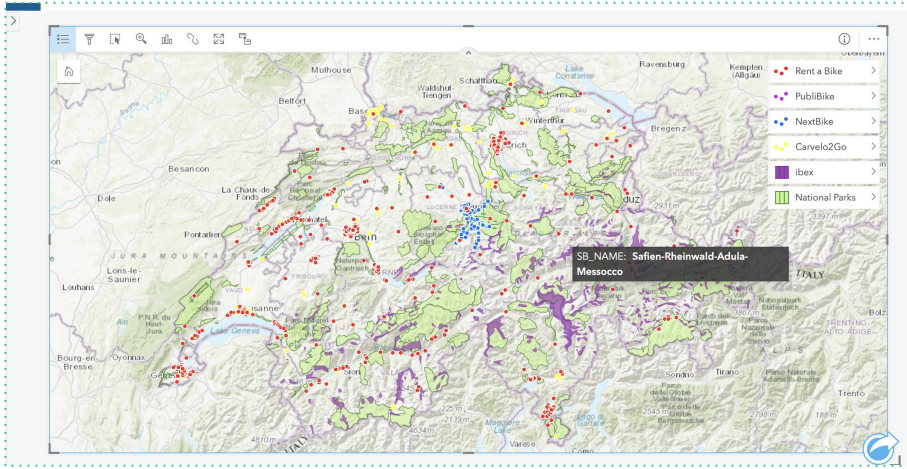
★★★★

★★★★★

* Although, XLS, CSV are not geospatial data formats by their nature, they might be used for storage of geospatial data



Potential impact of COVID crisis on EU labour markets



Bike-sharing with Alpine ibex colonies concentration

eurostat

Database. The majority of Eurostat's statistics may be accessed from the data navigation tree, which is structured according to statistical themes.

Web Services. Eurostat's datasets can be accessed through:

- SDMX Web Services, as well as
- JSON and Unicode Web Services.

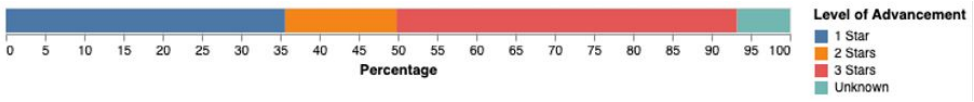
SPARQL endpoint. Eurostat provides its data as Linked Open Data via the SPARQL endpoint at the EU Open Data Portal.

Stars	Spatial Open Data formats available
★	PDF*
★★	XLSX*, Esri Shapefile, GDB, SPSS*
★★★	CSV*, GeoJSON, TopoJSON, TSV*
★★★★	RDF
★★★★★	Yes

* Although, PDF, XLSX, SPSS, CSV, TSV are not geospatial data formats by their nature, they might be used for storage of geospatial data

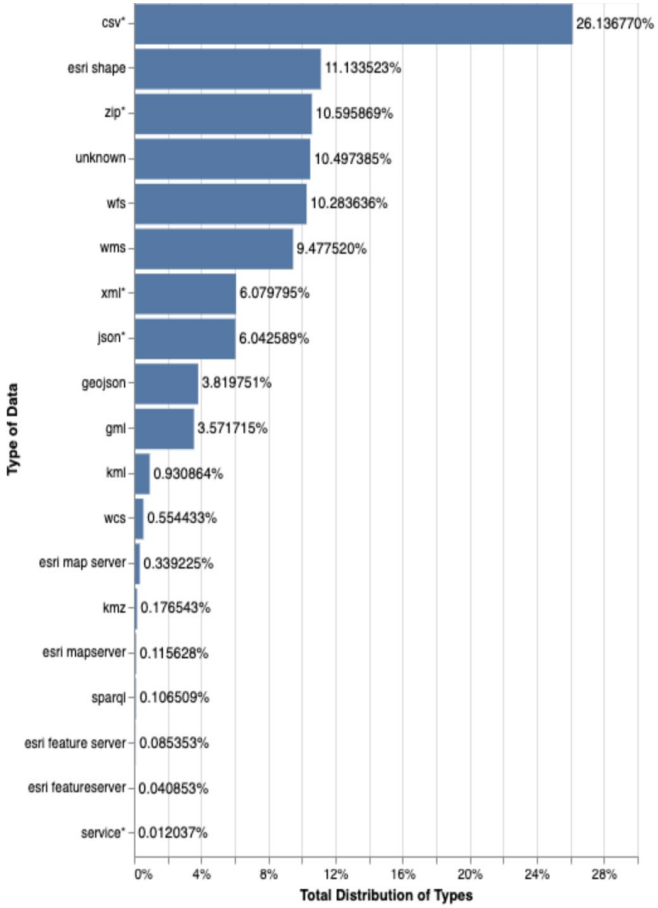
EUROPEAN DATA PORTAL

The majority of data at the European Data Portal distributed on the **third** level of advancement



Data Types usage aggregated by the Five Star Open Data model.
Based on the data sample collected via EDP API:

- 153951 Unique Datasets
- 390757 Unique Distributions
- 139259 Distributions (35,6%) with no Geospatial formats



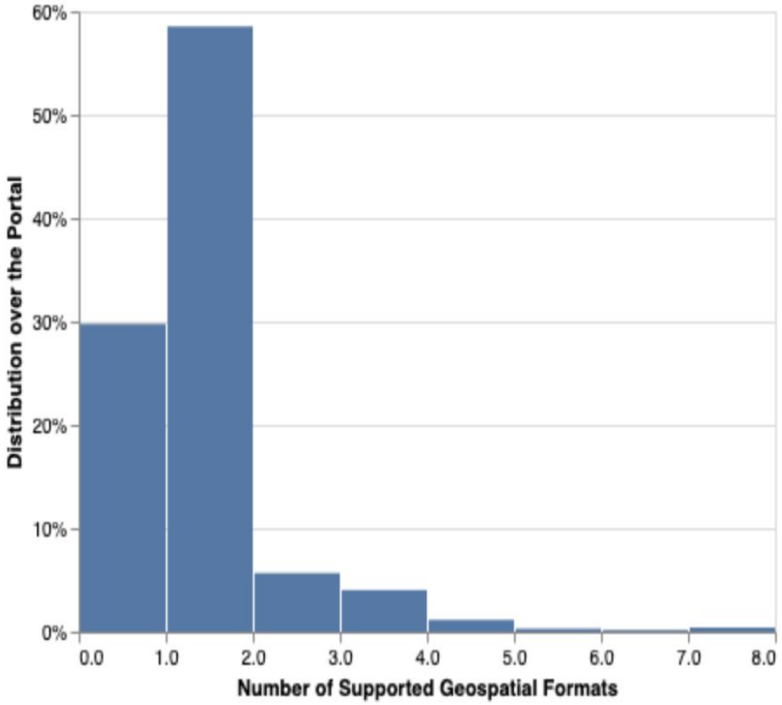
Data Types usage in the Datasets' Distributions

EUROPEAN DATA PORTAL

The significant majority of data sources provide only **two** or **one** types of data distribution.

Based on the data sample collected via EDP API:

- 153951 Unique Datasets
- 390757 Unique Distributions
- 139259 Distributions (35,6%) with no Geospatial formats



Distribution of the Number of Supported Geospatial Formats over datasets

Conclusion

- Geospatial data integration in BI is a fundamental step for the creation of maps and other visualization products.
- LOD remains as an isolated technology for the outside use, in particular in Location BI.
- BI tools mostly focus on geospatial data transfer and integration technologies that are common among end-users (ETL, API, Web Services and SQL).
- The existing (Geospatial) Open Data Models must be reviewed, updated and advanced to provide not only theoretical insights but rather real practical applications.
- Successful integration of LOD and BI would allow direct transformation of the potential of Geospatial Linked Open Data into valuable decisions and solutions across various domains.
- It is highly reasonable to support Location Business Intelligence with a cartographic methodology of the LOD integration.

Research Questions

What **data transfer technologies** currently define the leading **BI tools integrations' support**?

Majority of users are following a typical **Extract, Transform, Load** (ETL).

All the BI software allow integrating data via **API** and **SQL** queries on databases.

Research Questions

What specific **data transfer technologies** have to be in focus for the **Linked Open Data platforms** to provide **successful integration use-cases**?

Database integration as an opportunity to run **SQL queries**. Clear **API access**

GeoJSON, TopoJSON, KML, KMZ, **GeoPackage**, **WMS**, **WMTS**

Research Questions

What are the key components of the **Linked Open Data concept** required to be integrated in **BI software** to provide access to the **global Linked Open Data** and **Knowledge Network platforms**?

Support of **Web Services** as a way of **streaming** geospatial data **directly** from a source.

Adoption of **SPARQL technology**

Thank You!