

Part 2

State of the Art

- 1** Part 0 - Course Organization
- 2** Part 1 - The Reuse Problem
- 3** **Part 2 - State of the Art**
- 4** Part 3 - The Solution iTelos
- 5** Part 4 - The iTelos Methodology

Part 2.1

Data representation SoA

- 1** Data Representation SoA
- 2** Reusable Resources
- 3** Data Integration SoA
- 4** Data Architecture SoA

Data Representation SoA

- To reuse data we need to know **which data, and which types of data, are available**, so that we can identify the most suitable resources for a specific purpose.
- To this end, let's discover
 - the different types of data available;
 - how they are represented, and,
 - which are the existing best practices to enhance data quality and interoperability.

Data Representation SoA - Types of data

Data can be recognized and classified in many different ways. In this course, to describe the available types of data, we focus on 2 key dimensions;

- Cross-sectional data and Time-series data
- Domain data and Person-centric data

Data Representation SoA - Types of data

- **Cross-sectional data:** it carries information about a single moment in time. It doesn't consider the evolution of the data along the time.
"is a type of data collected by observing many subjects (such as individuals, firms, countries, or regions) at the one point or period of time."²
- **Time-series data:** it carries information about multiple moments in time. It describes the properties of one, or more, entities considering their evolution in time.
 - A concrete example of time-series data, is the data collected periodically by using sensors like, gps and accelerometer.

²https://en.wikipedia.org/wiki/Cross-sectional_data

Data Representation SoA - Types of data

- **Domain data:** this kind of data carries information about a specific domain of interest, by describing the entities composing it.
- **Person-centric data:** this kind of data carries information about the human behavior, thus describing a person, and her/his point of view within a specific domain of interest (or context).

The domain data provides the background data space where the person-centric data can be contextualized. In other words the **environment** (domain data) where one, or more, **subjects** (person-centric data) act.

Data Representation SoA - Existing languages and formats

- The data classified as above, is represented in several different languages and formats (as already described discussing the data heterogeneity).
- Sometimes the data is represented by using **tabular formats**, like:
 - JSON (can be used for tables)
 - CSV, TSV
 - Excels spreadsheet
- Other times the data is represented by using **graph-based formats**, like:
 - JSON (can be used for graphs)
 - XML
 - RDF-OWL
- For this course, the graph-based data representation is particularly relevant, because that is the way in which **Knowledge Graphs** (KG) are represented ³

³We will see how to solve the reuse problem by exploiting the Knowledge Graphs ↗ ↘ ↙

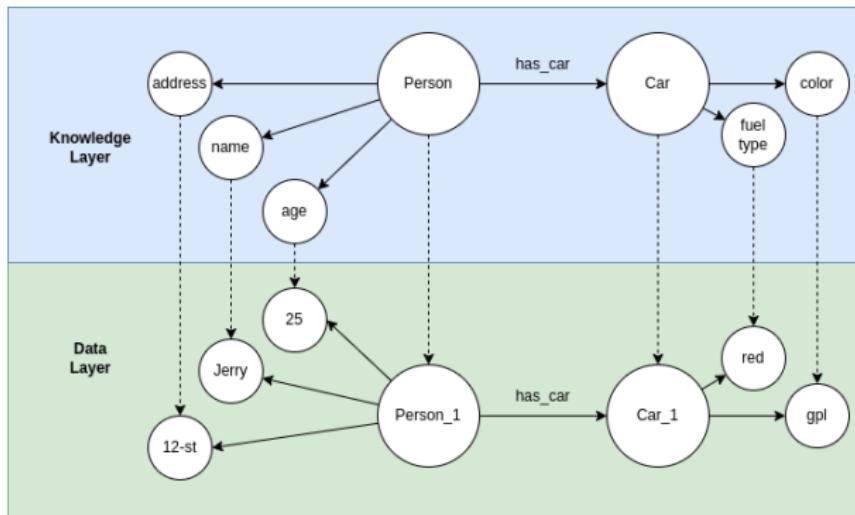
Knowledge Graph definition

A Knowledge Graph K , can be defined as follows:

$$KG = (E, D, R, A)$$

Where:

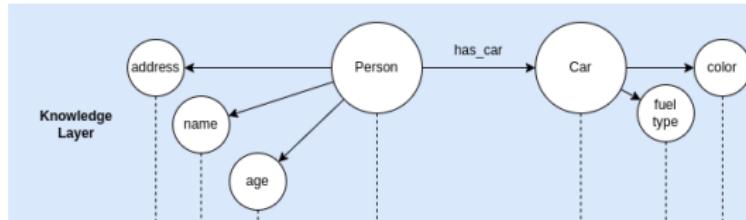
- E : is the set of real-world objects types, called *Entity Types* (or *ETypes*).
- D : is the set of real-world objects representations, called *Entities*. The Entities are *ETypes* instantiation.
- R is the set of properties used to denote the *ETypes*. The elements of R , can be properties related to a single *EType*, called *data properties*, or properties used to define relations among different *ETypes*, called *object properties*.
- A : is the set of property values denoting the attributes of the *Entities*. Each attribute, associated to one and only one property, instantiates the relative *data/object property*.



- $E = \{\text{Person}, \text{Car}\}$
- $D = \{\text{Person_1}, \text{Car_1}\}$
- $R = \{\text{address}, \text{name}, \text{age}, \text{color}, \text{fuel type}, \text{has_car}\}$
- $A = \{12\text{-st}, \text{Jerry}, 25, \text{red}, \text{gpl}\}$

Knowledge Layer

- The KG's Knowledge Layer is composed by the elements of E (ETypes) plus the element of R (properties definition).
- It defines the KG's structure (or schema).
- It is usually defined using an ontology modeled to represent the information to be maintained in the KG.

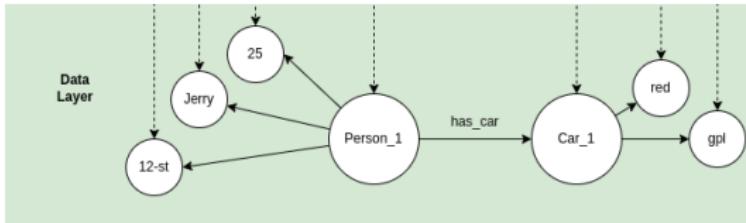


Ontology

- “An ontology is a formal, explicit specification of a shared conceptualization”
-by Gruber (1993) and modified by Studer et. al (1998)
- Ontologies are used to capture knowledge about some domain of interest. An ontology describes the concepts in the domain and also the relationships that hold between those concepts
- Ontologies are crucial for attributing semantics to Knowledge Graphs (KGs) which model ground-truth

Data layer

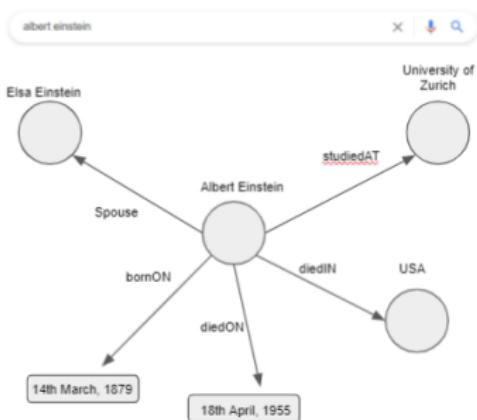
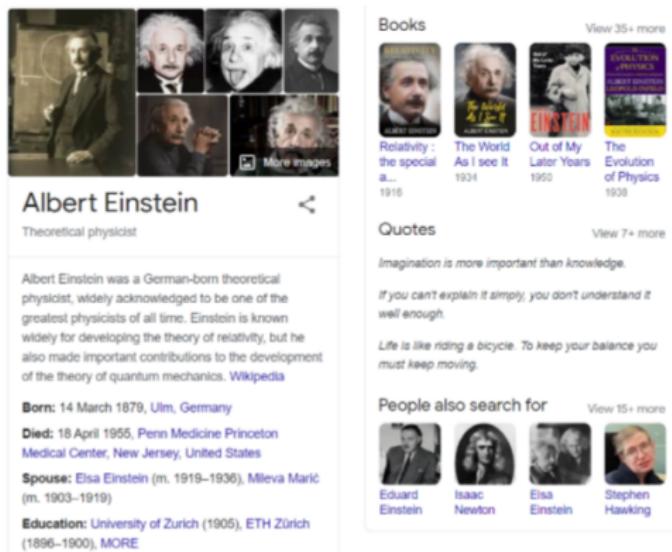
- The KG's Data Layer is composed by the elements of D (Entities) plus the element of A (attributes definition).
- It contains the data values instantiating the KG's structure.



KG-based Apps - examples

■ Google Knowledge Panel

Google Knowledge Panel

Albert Einstein

Theoretical physicist

Albert Einstein was a German-born theoretical physicist, widely acknowledged to be one of the greatest physicists of all time. Einstein is known widely for developing the theory of relativity, but he also made important contributions to the development of the theory of quantum mechanics. [Wikipedia](#)

Books

- Relativity : the special a... 1916
- The World As I see It 1931
- Out of My Later Years 1950
- The Evolution of Physics 1930

Quotes

Imagination is more important than knowledge.
If you can't explain it simply, you don't understand it well enough.
Life is like riding a bicycle. To keep your balance you must keep moving.

People also search for

- Eduard Einstein
- Isaac Newton
- Elsa Einstein
- Stephen Hawking

Figure: Mohit M. A guide to Knowledge Graphs Aug 30, 2021

KG-based Apps - examples

- InteropEHRate EU project

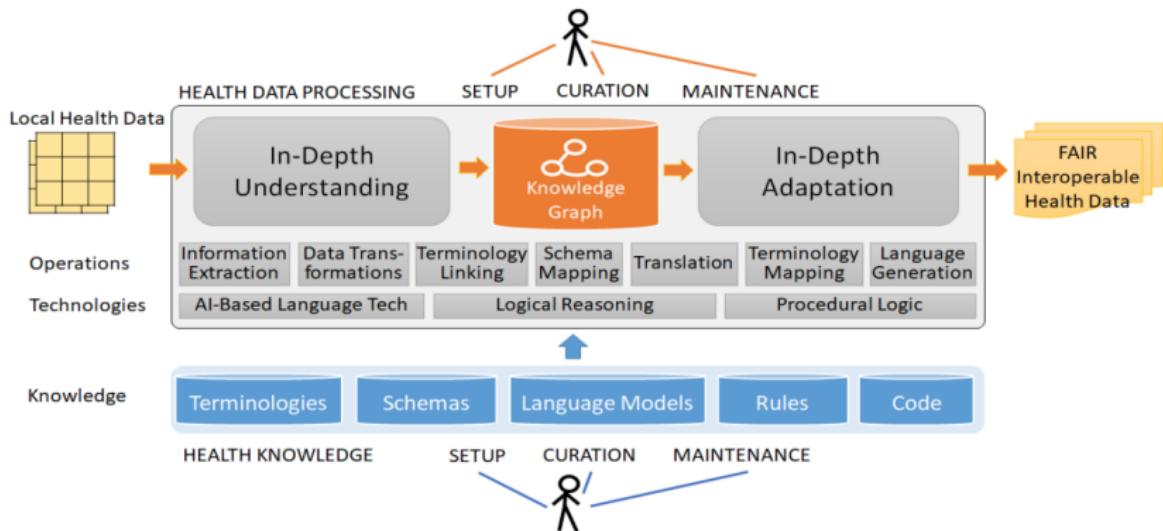


Figure 1: High-level architecture of the InteropEHRate Health Services and the way they are overseen by a human data manager.

KG-based Apps - examples

Many domain specific KGs have been produced supporting tasks like:

- Data Governance
- Automated Fraud Detection
- Knowledge Management
- Insider Trading
- Health Data Interoperability

KG-based Apps - examples

While there are several small-sized and domain-specific KGs, on the other hand, we also have many huge-sized and domain-agnostic KG that contains facts of all types and forms.

- **DBpedia**: is a crowd-sourced community-based effort to extract structured content from the information present in various Wikimedia projects.
- **Freebase**: a massive, collaboratively edited database of cross-linked data. Touted as “an openly shared database of the world’s knowledge”. It was bought by Google and used to power its own KG. In 2015, it was finally discontinued.
- **OpenCyc**: is a gateway to the full power of Cyc, one of the world’s most complete general knowledge base and commonsense reasoning engines.
- **Wikidata**: is a free, collaborative, multilingual database, collecting structured data to provide support for Wikimedia projects.
- **YAGO**: huge semantic knowledge base, derived from Wikipedia, WordNet, and GeoNames.

Data Representation SoA - Quality and Interoperability

- We need reusable resources ..
- To enhance the quality and interoperability of data, some criteria and best practices have been already defined.
 - Open (5*) data
 - FAIR data

Open (5*) data

“The Semantic Web isn’t just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other, related, data.”⁴

- ★ Make your stuff available on the Web (whatever format) under an open license
- ★ ★ Make it available as structured data (e.g., Excel)
- ★ ★ ★ Use non-proprietary open format (e.g., CSV instead of Excel)
- ★ ★ ★ ★ Use URIs to denote things, so that people can point at your stuff
- ★ ★ ★ ★ ★ Link your data to other data to provide context

⁴Tim Berners-Lee, Linked data-design issues, <https://www.w3.org/DesignIssues/LinkedData.html>, 2006.

FAIR data

Findability	F1. (Meta)data are assigned a globally unique and persistent identifier
	F2. Data are described with rich metadata (defined by R1 below)
	F3. Metadata clearly and explicitly include the identifier of the data it describes
	F4. (Meta)data are registered or indexed in a searchable resource
Accessibility	A1. (Meta)data are retrievable by their identifier using a standardised communications protocol
	A1.1. The protocol is open, free, and universally implementable
	A1.2. The protocol allows for an authentication and authorisation procedure, where necessary
	A2. Metadata is accessible, even when the data are no longer available
Interoperability	I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
	I2. (Meta)data use vocabularies that follow FAIR principles
	I3. (Meta)data include qualified references to other (meta)data
Reusability	R1. Metadata is richly described with a plurality of accurate and relevant attributes
	R1.1. (Meta)data are released with a clear and accessible data usage license
	R1.2. (Meta)data are associated with detailed provenance
	R1.3. (Meta)data meet domain-relevant community standards

Figure: FAIR data principles ⁵

⁵<https://www.go-fair.org/fair-principles/>

FAIR is not Open

- “FAIR data and open data are two distinct concepts which are however coming closer and closer. (Mons, et al., 2017) distinguish the concept of FAIR from the concept of open, saying: In the envisioned Internet of FAIR Data and Services, the degree to which any piece of data is available, or even advertised as being available (via its metadata) **is entirely at the discretion of the data owner.**”
- “... moreover, the Council of the European Union concluded that **“as open as possible, as closed as necessary”** is the underlying principle for optimal reuse of research data.”

Cost of non-FAIR data

The European Commission report⁶ (March 2018) indicates that:
"the annual cost of not having FAIR research data costs the European economy at least €10.2bn every year"

⁶European Commission. "Cost of not having FAIR research data-Cost-Benefit analysis for FAIR research data." (2018)

Cost of non-FAIR data - Research activities involved



Figure: Cost of not having FAIR research data-Cost-Benefit analysis for FAIR research data." (2018)

Part 2.2

Reusable Resources

- 1** Data Representation SoA
- 2** Reusable Resources
- 3** Data Integration SoA
- 4** Data Architecture SoA

Data Representation SoA - Reusable resources

- Which data resources are already available to be reused and where we can find them ?
- Depending by the information carried, we can find three different types of reusable data:
 - Linguistic
 - Knowledge
 - Data values

Open data Catalogs

- Where are the reusable resources that we need to build KGs ?
- Several projects and open data portal already exist which allow the users to retrieve useful resources.
- Often such resources are accessible through **Catalogs**. They are open portals collecting information about several resources (i.e. datasets, schemas, ontologies, ...).
- The catalogs doesn't collect the real resources, but instead the **metadata** describing such resources. (Catalogs are supported by backhand repositories)
- The more metadata are associated with a resource, the more detailed it will be on the catalog, thus by consequence, it will be more findable and **reusable**.

Linguistic Resources

A linguistic resource is a dataset which provides data about languages (e.g., meanings, relations between words, ...).

There are two types of mono/multi-lingual resources: (i) online dictionaries and (ii) Wordnet like resources. Wordnets much more useful in data integration as they connect meanings of words in a LKG.

Check the licence (lots of options).

Example

- Global Wordnet Association
- WordNet
- Open Multilingual WordNet
- DataScientia/UKC (forthcoming)

Linguistic Resource Repositories

Global WordNet Association

[Home](#) [About GWA](#) [Home](#) [Resources](#) [Global WordNet Conferences](#) [Contact](#)

Global WordNet Association

**** 10th Conference 2019 ****

A free, public and non-commercial organization
that provides a platform for discussing, sharing
and connecting wordnets for all languages in the
world.

[More info on GWA](#)

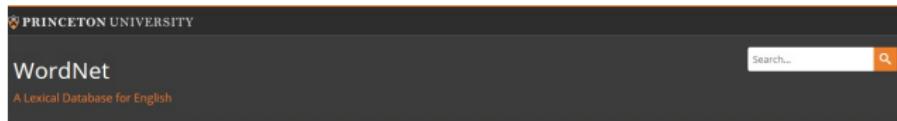


**Global
WordNet
Association**

Figure: Global WordNet Association⁷



Linguistic Resource Repositories



The screenshot shows the WordNet homepage. At the top, there is a navigation bar with links for "Home", "About", "Browse", "Search", "Help", and "Log In". Below the navigation bar, the title "WordNet" is displayed in large, bold letters, followed by the subtitle "A Lexical Database for English". To the right of the title is a search bar with a magnifying glass icon. The main content area contains several sections: "What is WordNet?", "About WordNet", "Structure", and "Note". The "What is WordNet?" section includes a note about funding and staffing issues. The "About WordNet" section provides a detailed description of what WordNet is and how it works. The "Structure" section describes the hierarchical nature of WordNet. The "Note" section contains a note from the creators of WordNet.

- What is WordNet?**
- [People](#)
- [News](#)
- [Use WordNet Online](#)
- [Download](#)
- [Citing WordNet](#)
- [License and Commercial Use](#)
- [Related Projects](#)
- [Documentation](#)
- [Publications](#)
- [Frequently Asked Questions](#)

What is WordNet?

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the creators of WordNet and do not necessarily reflect the views of any funding agency or Princeton University.

When writing a paper or producing a software application, tool, or interface based on WordNet, it is necessary to properly cite the source. Citation figures are critical to WordNet funding.

About WordNet

WordNet® is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the browser®. WordNet is also freely and publicly available for download. WordNet's structure makes it a useful tool for computational linguistics and natural language processing.

WordNet superficially resembles a thesaurus, in that it groups words together based on their meanings. However, there are some important distinctions. First, WordNet interlinks not just word forms—strings of letters—but specific senses of words. As a result, words that are found in close proximity to one another in the network are semantically disambiguated. Second, WordNet labels the semantic relations among words, whereas the groupings of words in a thesaurus does not follow any explicit pattern other than meaning similarity.

Structure

Note

Due to funding and staffing issues, we are no longer able to accept comment and suggestions.

We get numerous questions regarding topics that are addressed on our FAQ page. If you have a problem or question regarding something you downloaded from the "Related projects" page, you must contact the developer directly.

Please note that any changes made to the database are not reflected until a new version of WordNet is publicly released. Due to limited staffing, there are currently no plans for future WordNet releases.

Figure: WordNet Home⁸

Linguistic Resource Repositories

Open Multilingual Wordnet

This page provides access to open wordnets in a variety of languages, all linked to the [Princeton Wordnet of English](#) (PWN). The goal is to make it easy to use wordnets in multiple languages. The individual wordnets have been made by many different projects and vary greatly in size and accuracy. We have (i) extracted and normalized the data, (ii) linked it to Princeton WordNet 3.0 and (iii) put it in one place. The Open Multilingual Wordnet and its components are [open](#): they can be freely used, modified, and shared by anyone for any purpose. There is a fuller list of wordnets at the Global Wordnet Association's [Wordnets in the World page](#).

If you use these wordnets, please cite the original projects who created them (linked in Table 1), if you got value from this aggregation/normalization, please cite [Bond and Paik \(2012\)](#).

You can access the wordnets through the (python) [Natural Language Tool-Kit wordnet interface \(NLTK\)](#).

We have an [extended version](#) with automatically extracted data for over a 150 languages from [Wiktionary](#) and the [Unicode Common Locale Data Repository](#) ([Bond and Foster, 2013](#)).

[Documentation, News and Updates](#)

Search

We have a [simple search interface](#) (search [the extended wordnet](#)). It uses the SQL database originally developed by the Japanese Wordnet.

34 Open Wordnets Merged								
Wordnet	Lang	Synsets	Words	Senses	Core	Licence	Data	Citation
Albanet	als	4,675	5,988	9,599	31%	CC BY 3.0	als.zip (+xml)	cite.als; (.bib)
Arabic WordNet (AWN v2)	arb	9,916	17,785	37,335	47%	CC BY SA 3.0	arb.zip (+xml)	cite.arb; (.bib)
BulTreeBank Wordnet (BTB-WN)	bul	4,959	6,720	8,936	99%	CC BY 3.0	bul.zip (+xml)	cite.bul; (.bib)
Chinese Open Wordnet	cmn	42,312	61,533	79,809	100%	wordnet	cmn.zip (+xml)	cite.cmn; (.bib)

Figure: Open Multilingual WordNet Home

Linguistic Resource Repositories



The lexicons we support



Vision and Mission

The Universal Knowledge Core (UKC) is a psycholinguistic principles based multilingual, high quality, large scale, and diversity aware machine readable lexical resource.

The key design principle underlying the UKC is to maintain a clear distinction between the language(s) used to describe the world as it is perceived and what is being described, i.e., the world itself. The Concept Core (CC) is the UKC representation of the world and it consists of a semantic network where nodes are

Knowledge Resources

A Knowledge resource is a dataset which consists of a KB encoding information about schemas (etypes and properties).

KBs of high quality are usually called ontologies. We call them teleologies (meaning by this, ontologies with metadata which empower their practical use in knowledge and data integration).

Example

- LOV/LOV4IoT
- Schema.org
- DBpedia (schema only)
- DataScientia/liveschema (forthcoming)

Knowledge Resource Repositories

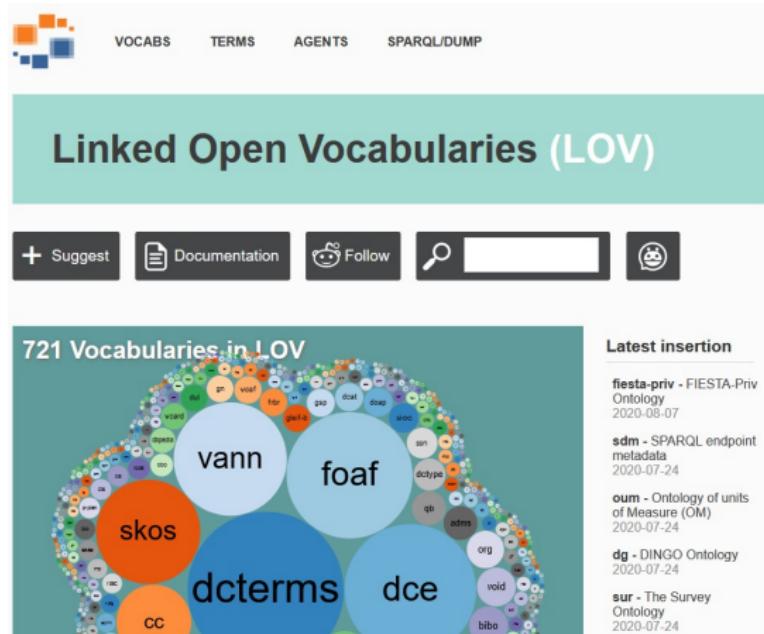


Figure: Linked Open Vocabulary¹¹



Knowledge Resource Repositories

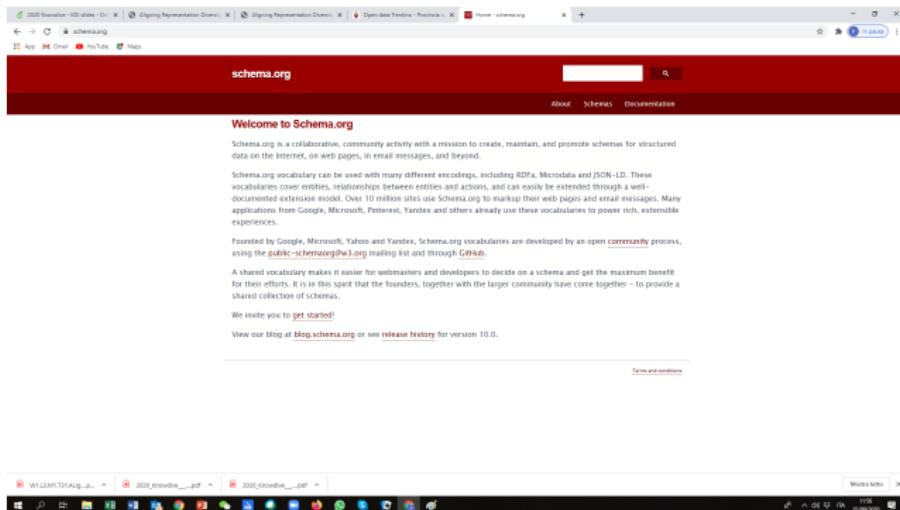


Figure: Schema.org¹²

¹²<http://www.schema.org/>

Knowledge Resource Repositories

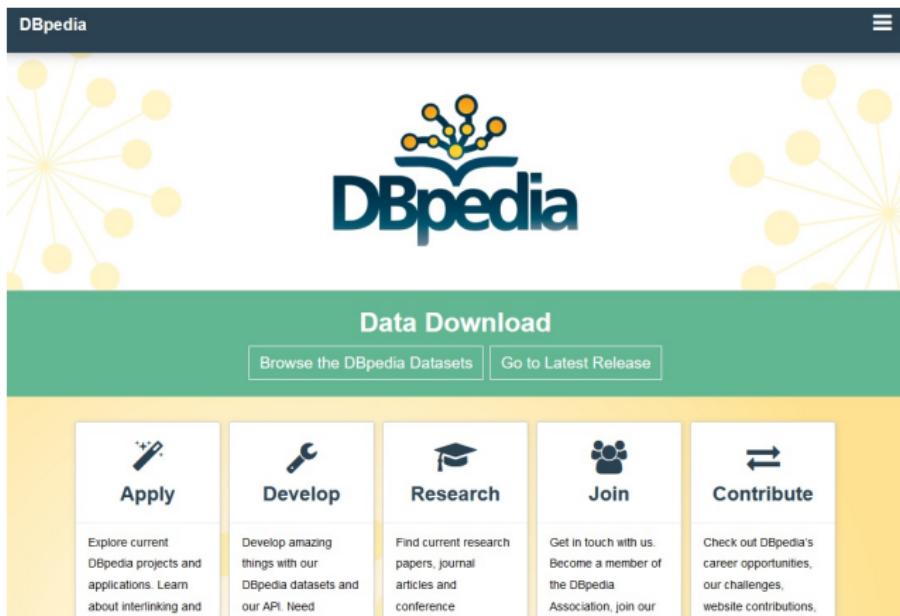


Figure: DBpedia Home¹³

Knowledge Resource Repositories

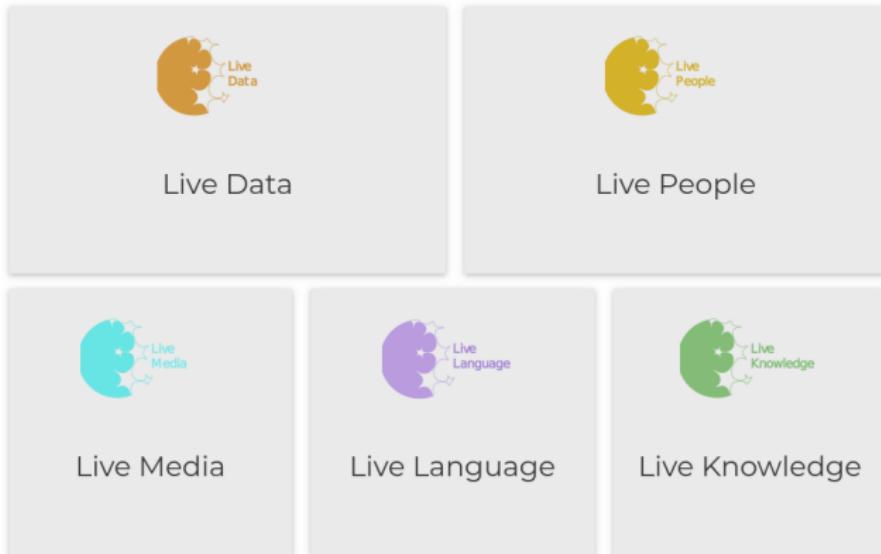


Figure: DataScientia Catalogs¹⁴



Data Resources

A data resource is a dataset which consists of data in some format (tabular, unstructured, entities and property values).

Example

- UK Open Data
- National Bureau of Statistics, China
- data.org
- Opendata Trentino (see, among others, Unitn Open Data)
- Geonames
- Open Street Map
- DBPedia
- Data Hub

Data Resource Repositories

[data.gov.uk | Find open data](#)

[Publish your data](#) [Documentation](#) [Support](#)

BETA This is a new service – your [feedback](#) will help us to improve it

Find open data

Find data published by central government, local authorities and public bodies to help you build products and services

[Business and economy](#)

Small businesses, industry, imports, exports and trade

[Crime and justice](#)

Courts, police, prison, offenders, borders and immigration

[Defence](#)

Armed forces, health and safety, search and rescue

[Education](#)

[Environment](#)

Weather, flooding, rivers, air quality, geology and agriculture

[Government](#)

Staff numbers and pay, local councillors and department business plans

[Government spending](#)

Includes all payments by government departments over £25,000

[Mapping](#)

Addresses, boundaries, land ownership, aerial photographs, seabed and land terrain

[Society](#)

Employment, benefits, household finances, poverty and population

[Towns and cities](#)

Includes housing, urban planning, leisure, waste and energy, consumption

Figure: Open Data UK¹⁵

Data Resource Repositories



The screenshot shows the homepage of the National Bureau of Statistics China. The header features a large graphic of a 3D cube with numbers (1, 2, 3, 4, 5, 6, 7, 8, 9) floating around it, followed by the text "National data 国家数据" and "National Bureau of Statistics". Navigation links include 首页, 月度数据, 季度数据, 年度数据, 普查数据, 地区数据, 部门数据, 国际数据, 可视化产品, 出版物, 我的收藏, and 帮助. A search bar contains "如: 2012年北京GDP" and a pink "搜索" button. Below the search bar are links for GDP, CPI, 总人口, 社会消费品零售总额, 粮食产量, PIM, and PPI. The main content area includes sections for "数据中国 再升级" (with text about the app upgrade), "数据中国 pro App 再升级" (with an image of a hand holding a smartphone displaying the app), and "第三次全国农业普查" (with a thumbnail image of a survey vehicle). Other sections include "统计数据产品", "工作年度报表", "图表中国", "数据中国[中 英]", "如何获取统计数据", and "官方微博(人民 新华 新浪 腾讯)官方微信". A footer navigation bar at the bottom has links for 1 through 15, "更多>>", and icons for back, forward, and search.

Figure: National Bureau of Statistics China

Data Resource Repositories

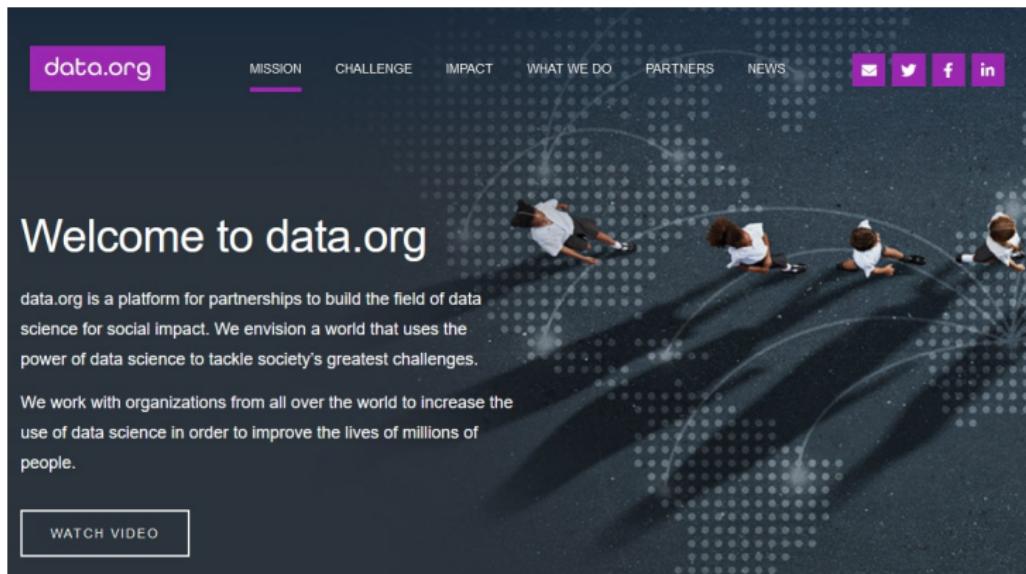


Figure: data.org¹⁷

Data Resource Repositories



The screenshot shows the homepage of the Open Data Trentino portal. At the top, there is a navigation bar with links for "App", "Omni", "YouTube", and "Maps". Below the navigation bar, there is a banner for the "Festival dell'Economia 2020" with a phone number and a link to "Media informazione provincia.it". The main content area features a large "OPENdata TRENTINO" logo. To the left, there is a sidebar with links for "Argomenti", "Provincia Autonoma", "Punto informativo", "Territorio", and "Amministrativa trasparente". A "Partecipo" button is also present. The central part of the page displays a list of datasets, each with a thumbnail, title, and a brief description. Some titles include "più spazio alla scuola", "dati.trentino.it: il portale dei dati aperti del Trentino", and "Open Data Trentino". At the bottom of the page, there is a footer with links for "Stampa" and "Mostra tutta".

Figure: Open Data Trentino¹⁸

Data Resource Repositories



The GeoNames geographical database covers all countries and contains over eleven million placenames that are available for download free of charge.

all countries

[\[advanced search\]](#)

enter a location name, ex: "Paris", "Mount Everest", "New York"

Browse the names	Information	Download
<ul style="list-style-type: none"> • Countries • Postal codes • Country statistics • Recent modifications 	<ul style="list-style-type: none"> • About GeoNames • Data Sources • User manual • Ambassadors and Team • Forum • Blog • Mailing list • Commercial Support and Consulting 	<ul style="list-style-type: none"> • Info • Free Gazetteer Data • Free Postal Code Data • Premium Data
Web Services		
<ul style="list-style-type: none"> • Overview • Documentation • Client Libraries • Premium Web Services 		

Figure: Geonames Home¹⁹

Data Resource Repositories

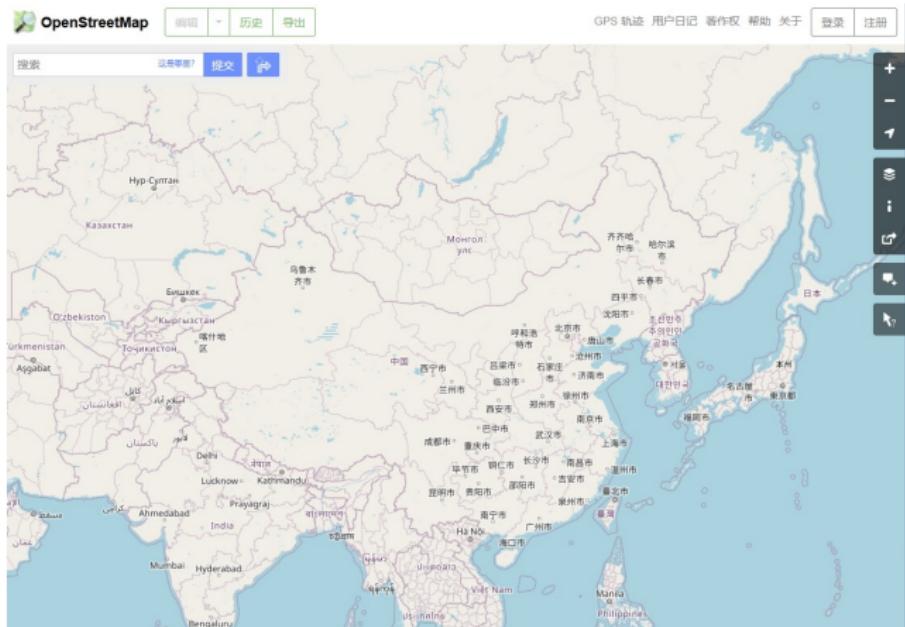


Figure: Open Street Map Home²⁰



Data Resource Repositories

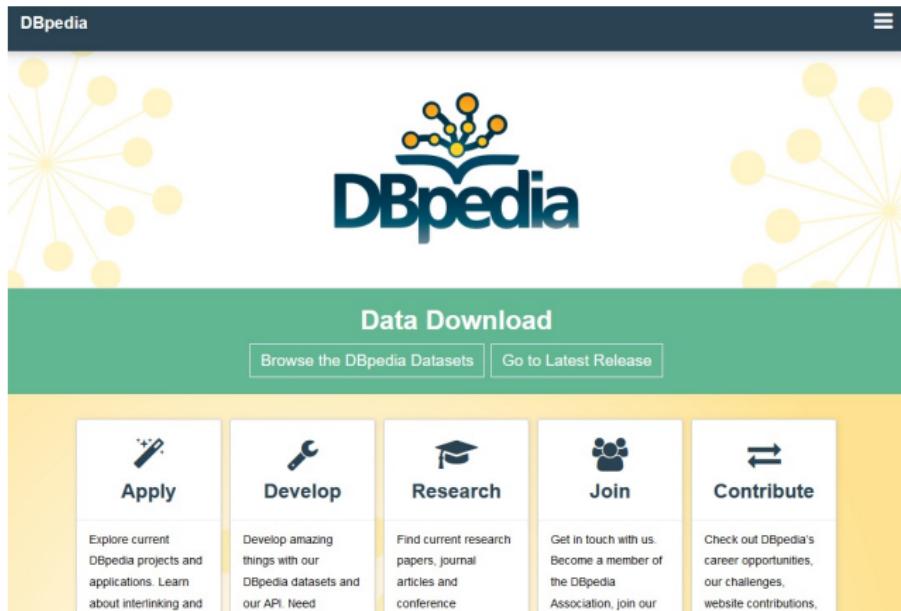


Figure: DBpedia Home²¹

Data Resource Repositories



ABOUT BLOG **FIND DATA** COLLECTIONS DOCS PRICING TOOLS CHAT ● LOGIN JOIN FREE



We help organizations of all sizes to design, develop and scale
solutions to manage their data and unleash its potential.

Let us help you!

[Get in touch now »](#)



Figure: Data Hub Home²²

Data Resource Repositories

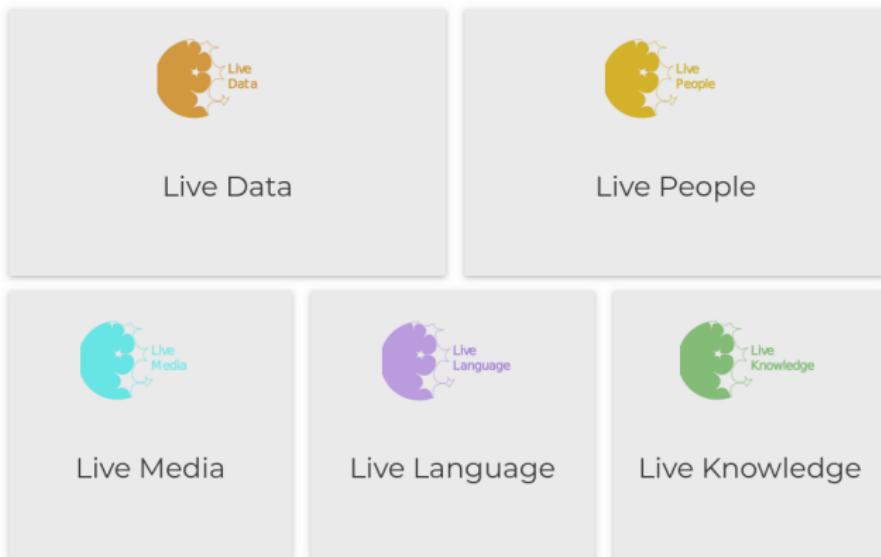


Figure: DataScientia Catalogs²³

Part 2.3

Data Integration SoA

- 1** Data Representation SoA
- 2** Reusable Resources
- 3** Data Integration SoA
- 4** Data Architecture SoA

Data Integration (DI) - Index

- 1** Introduction
- 2** Data adaptation and evolution
- 3** DI Virtualization strategies
 - Ontology Based Data Access (OBDA)
- 4** DI Materialization strategies
 - Knowledge Graph Construction (KGC)

Data Integration (DI) - Introduction

The language, knowledge and data resources, described above, represent the information highlighting the diversity.

Nevertheless, to exploit such a diversity, **language, knowledge and data, need to be integrated**.

The heterogeneity of the resources described above, leads to **different kinds of integration**.

- Integrate resources of the same type.
 - integrate different languages;
 - two or more, data schema, or ontologies;
 - two or more dataset.

Data Integration (DI) - Introduction

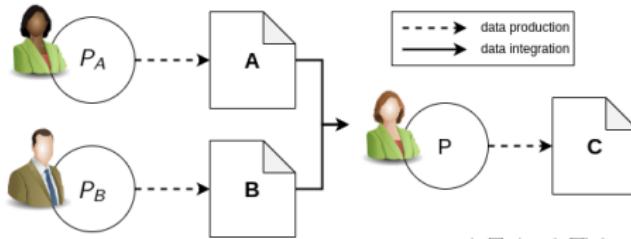
The heterogeneity of the resources described above, leads to **different kinds of integration**.

- Integrate resources of different types.
 - Integrate one dataset with a new language (different from the one used to represent its data).
 - Integrate two datasets by using a third data schema (or ontology) different from the single data schema adopted in the two datasets.
 - Integrate an ontology with a language, to produce multilingual knowledge resources.

Data adaptation and evolution

Regardless of the kind of resource integration, there two phases that always occurs when integrating resources:

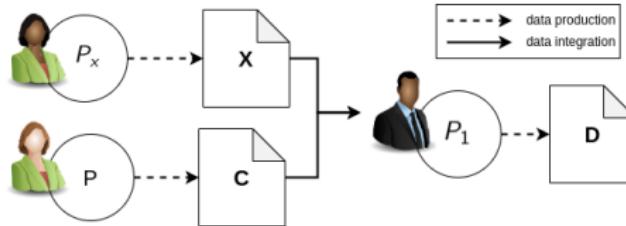
- **Data Adaptation:** this phase defines the first time that two resources, A and B, need to be integrated.
 - The resources A and B, have been created for specific purposes P_A and P_B , and they have not been modified, and/or, integrated with any other resource, to satisfy a different purpose.
 - A and B are then integrated, following a new purpose P, thus producing C as integration output.



Data adaptation and evolution

Regardless the kind of resource integration, there two phases that always occurs when integrating resources:

- **Data Evolution:** in this phase the result of the adaptation integration of A and B, is in turn integrated to satisfy a new purpose P_1 (or an extended version of the P).
 - The adaptation integration output C, is integrated with new resources to satisfy a new purpose P_1 , thus creating the result of the evolution integration D.



DI strategies

The existing DI strategies are mainly divided in two categories:

- **Virtualization strategies:** aim at providing a unique interface for two, or more, data sources, for accessing the data without extraction and transformation data.
- **Materialization strategies:** are based on ETL procedures used to extract the data to be integrated from the respective data sources.

DI Virtualization strategies - LAV

The most known set of virtualization strategies are called:
Ontology Based Data Access (OBDA)

They are all based on the modeling of an **ontology**, or a set of ontologies, interfacing the different data sources:

- **LAV - "Local As a View"** : this family of OBDA techniques assumes to query data as they are provided by the data sources.
 - single-ontology : a single ontology is queried to access all the data sources.
 - multiple-ontology : an ontology for each data source can be queried to access the data.
 - hybrid approach : as the multiple-ontology technique there are more ontologies, but in this case the queries are uniformed by a unique vocabulary used to query the data.

DI Virtualization strategies - GAV

- **GAV - "Global As a View"** : this family of OBDA techniques assumes to query data from the point of view of the application (services, or users) that needs to perform the query, by exploiting an application-specific ontology modeling.

DI Virtualization strategies - Summary

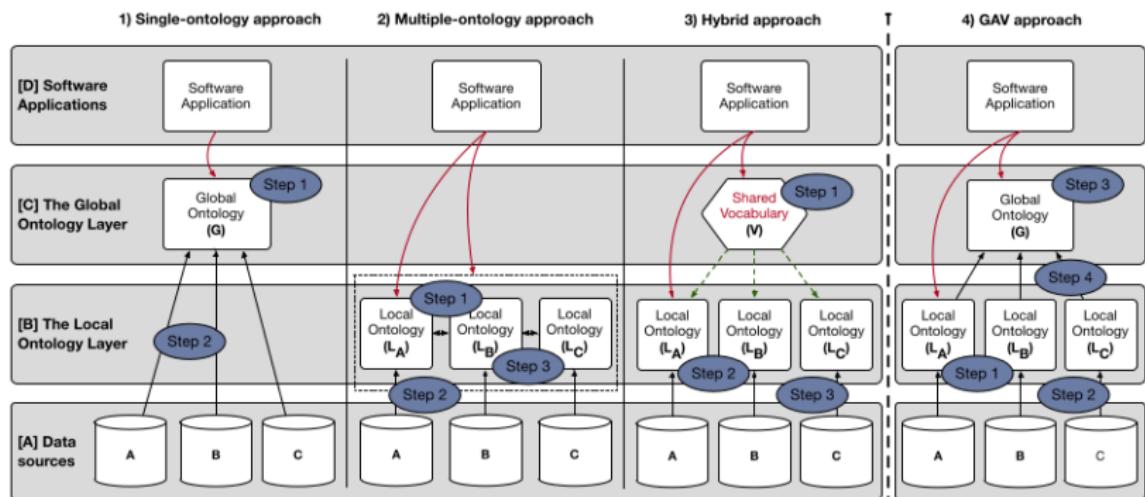


Figure 1: Three variants of OBDI from [75]: (1) single-ontology, (2) multiple-ontology, (3) hybrid, and an additional OBDI variant (4) Global-as-View (GAV).

(Explanation: Red arrows indicate access from an application to data, black arrows represent transformation/virtual access to the data; dotted green arrows represent implicit relations between involved ontologies, and numbered items show the sequence of system development. The dotted rectangle refers to the federation of local ontologies. Section 5.1 explains the additional OBDI variant (4) *Global-as-View* (GAV).)

DI Virtualization strategies - Limitations ²⁴

- In a GAV approach, changes in information sources, or adding a new information source, requires **revisions of a global schema and mappings** between the global schema and source schemas.
- In a LAV approach, automating query reformulation has **exponential time complexity with respect to query and source schema definitions**.

²⁴Xu, Li, and David W. Embley. "Combining the best of global-as-view and local-as-view for data integration." *Information systems technology and its applications*, 3rd international conference ISTA'2004. Gesellschaft für Informatik eV, 2004.



DI Materialization strategies

The usage of Knowledge Graphs increased a lot within the data integration community, thanks to their suitability within different domain of interest. For this reason one of the most famous materialization strategies, is **Knowledge Graph Construction (KGC) DI**.

DI based on KGC is a process that involves different sub activities defined as follows:

- Data collection/extraction
- Schema definition/alignment
- Data cleaning & formatting
- Entity identification & mapping
- Data mapping

DI Materialization strategies - Limitations

- **Missing of standard methodologies** for KG generation;
 - the KGs produced are often, too application-specific, causing an increase of the KG's evolution cost.
- **Missing of frameworks** (tools and application) covering the whole KGC process.
- **Technical skills required** (data management and knowledge modeling) for the KGC process implementation;
 - the KG's final user (who is usually the domain expert) usually doesn't have such expertise.

Part 2.4

Data Architecture SoA

- 1** Data Representation SoA
- 2** Reusable Resources
- 3** Data Integration SoA
- 4** Data Architecture SoA

Data Management Architecture - Index

1 Introduction

2 Architecture history

- 1** data warehouse
- 2** data lake
- 3** data mesh

Data Management Architecture - Introduction

The data, before and after its elaboration (i.e., data integration), needs to be maintained by a dedicated architecture, with **the objective of serving the users, and applications, that exploit such data.**

Different kinds of data management architecture have been studied and implemented, in the past.

- Data warehouse (1960s)
- Data lake (2010)
- Data mesh (2021)

Data Warehouse

Introduced in 1960s, it is a set of centralized framework and data storage systems where:

- data is extracted from many operational databases and sources;
- data is transformed into a universal schema - represented as a multi-dimensional and time-variant tabular format;
- data is loaded into the warehouse tables;
- data is accessed through SQL-like querying operations;
- data is mainly serving data analysts for their reporting and analytical visualizations use cases.

Data Warehouse - Limitations

- Over time, they grow to thousands of ETL jobs, tables and reports that only a specialized group can understand and maintain.
- They don't let themselves to modern engineering practices such as CI/CD and incur technical debt over time and an increased cost of maintenance.
- Single schema, single representation, low interoperability.

Data Lakes

Introduced in 2010, it is a set of centralized framework and data storage systems where:

- data is extracted from many operational databases and sources;
- data is **minimally transformed** to fit the storage format e.g. Parquet, Avro, etc;
- data - as close as the source syntax - is loaded to scalable object storage;
- lake storage is accessed mainly for analytical and machine learning model training use cases and used by data scientists.

Data Lakes - Limitations

- They require complex and unwieldy pipelines of batch or streaming jobs operated by a central team of hyper-specialized data engineers.
- They contain deteriorated and unmanaged datasets, untrusted and some times inaccessible, which provide little value.
- **In other words, too much Noise!**

Data Mesh

"Data Mesh is a sociotechnical approach to share, access and manage analytical data in complex and large-scale environments - within or across organizations." ²⁵

A data mesh is a decentralized data architecture that **organizes data by a specific business domain**, providing more **ownership to the producers** of a given dataset.

²⁵ Dehghani, Zhamak. Data Mesh. Marcombo, 2022.

Data Mesh - Characteristics

- **Domain-oriented ownership:** Decentralize the ownership of sharing analytical data to business domains who are closest to the data.
- **Data as a Product:** Existing or new business domains become accountable to share their data as a product served to data users – data analysts and data scientists.
- **Self-serve Data Platform:** A new generation of self-serve data platform to empower domain-oriented teams to manage the end-to-end life cycle of their data products.
- **Federated Computational Governance:** A data governance operational model that is based on a federated decision making and accountability structure, with a team made up of domains, data platform, and subject matter experts