KNOWDIVE

KGE - Knowledge Graph Engineering

The Reuse Problem

Fausto Giunchiglia

Contents

1 **Build a KG by reuse**

2 Reusable Resources

3 Existing Reusable Resources

4 Data reuse - Producer & Consumer

KG Purpose

- Both data producer and consumer consider their own objective when building KGs.
- Such an objective implicitly includes the user "point of view", the representation that the user uses to model (a portion of) the world, where the desired information lives.
- We define the user objective, **The Purpose** which will lead the entire KGE process.

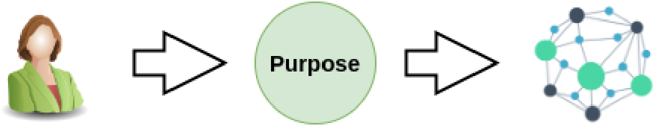
KG Purpose — producers / consumers

- KG producers: publish their data in some catalogues/ repositories maximizing reuse. They strive for maximum sharability (use of ontologies, most of the time). Purpose with stronger schema requirements, data defined by the applications generating them, cleaned with as minimal as possible cleaning (EML data level requirements)
- KG consumers: use existing data (and schemas) to produce their own (e.g., market application) data. Purpose with stronger data requirements, schema mainly defined by the application, with as minimal as possible tweaks (EML schema level requirements)

NOTE: Always, the schema is the *means*, data is the *goal*.

Purpose-specific KGs

- Each user who desires to build a KG, will have her own Purpose.
- That is why, most of the KGs produced are **purpose-specific**.
- Even trying to build general purpose KGs, it is impossible to extract the KG produced from a specific (more/less general/specific) context.
- **Context**: a vision of the world we live in, which can change geographically, socially as well as considering its evolution in time.



Contents

1 Build a KG by reuse

2 **Reusable Resources**

3 Existing Reusable Resources

4 Data reuse - Producer & Consumer

Reusable Resources

- The user purpose is reflected on the resources used to build the KG.
- Three categories of reusability are defined for such resources, depending on their relevance for the purpose itself: **Core**, **Common** and **Contextual** resources are used to build KGs.
- In order to define the 3 categories, we will follow the example having the following purpose:

"The user want build a KG able to support the access to the health facilities in Trento (Italy) and all the medical care that they can offer to the citizens."

Common resources

- **Common resources**: this category involves resources carrying information common to several contexts (or domain of interest), thus they can be resources not strictly related to the user's Purpose, but essential to support it in the KG.
- Some example of common resources for the Purpose declared above, can be datasets for:
  - Trento location.
  - Trento public transportation.
  - Trento parking areas.
  - ...

Core resources

- **Core resources**: this category involves resources carry information about the most important aspects considered by the purpose, information without which it would be impossible to build the KG.

- Some example of core resources for the Purpose declared above, can be datasets for:
  - Trento Hospitals.
  - Trento Pharmacies.
  - Drugs.
  - Patients.
  - Doctors.
  - ...

Contextual resources

- **Contextual resources**: this last category involves resources which carry specific, possibly unique, information related to the user's Purpose. These are the resources whose main goal is to create added value. If core resources are necessary for a meaningful application, contextual resources are the ones which can make the difference with respect to the competitors.

- Some example of contextual resources for the Purpose declared above, can be datasets for:
  - Drug's components.
  - Medical Prescriptions.
  - Hospital Departments.
  - Doctors Specializations.
  - Patient Allergies.
  - ...

Reusable Resources

- Depending by its category, a resources can be more or less reusable.
- Common resources are **the most reusable**, due to their usage shared among several contexts (so even among different purposes)
- Core resources are **less reusable** (even less findable) respect the common ones, due to their specificity on the contexts/domain of interest.
- Contextual resources are **the least reusable** due to their specificity on the Purpose. Moreover, **they are often created** from scratch because not available from other sources.

The Purpose of KG producers / consumers

- KG producers: Focus on common and core data (and schemas as a consequence). contextual data possibly to be dropped and not published.

- KG consumers: Focus on core and contextual data (and schemas as a consequence): maximum reuse of common and core data. Contextual data possibly generated on purpose.

Contents

1. Build a KG by reuse

2. Reusable Resources

3. **Existing Reusable Resources**

4. Data reuse - Producer & Consumer

Open data Catalogs

- Where are the reusable resources we need to build KGs ?
- Several projects and open data portal already exist which allow 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)
- More metadata are associated to a resource, more detailed it is on the catalog, thus by consequence, it will be more findable and even **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

Global WordNet Association

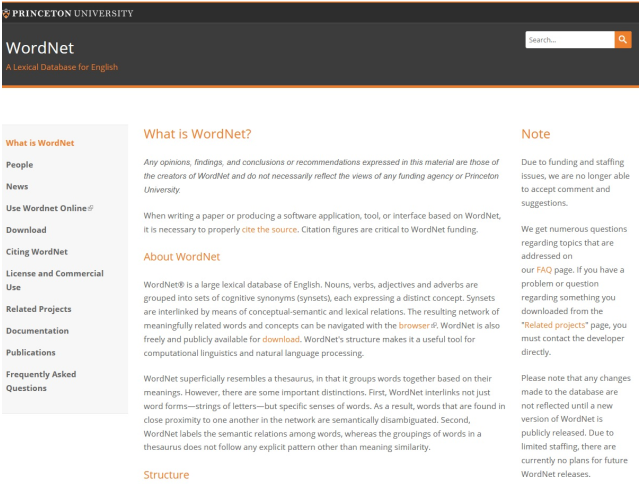
** 10th Conference 2019 **

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

Figure: Global WordNet Association[1]

[1] http://globalwordnet.org/

Linguistic Resource Repositories



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.

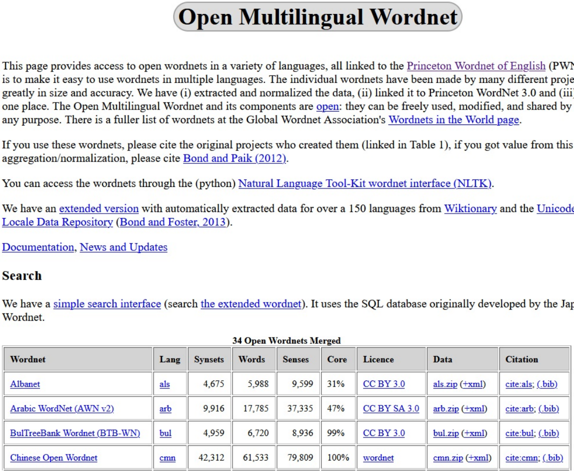
We get numerous questions regarding topics that are addressed on our FAQ page. If you have a problem or question regarding something you downloaded from the "Related projects" page, you must contact the developer directly.

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 Home[2]

[2] https://wordnet.princeton.edu/

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.

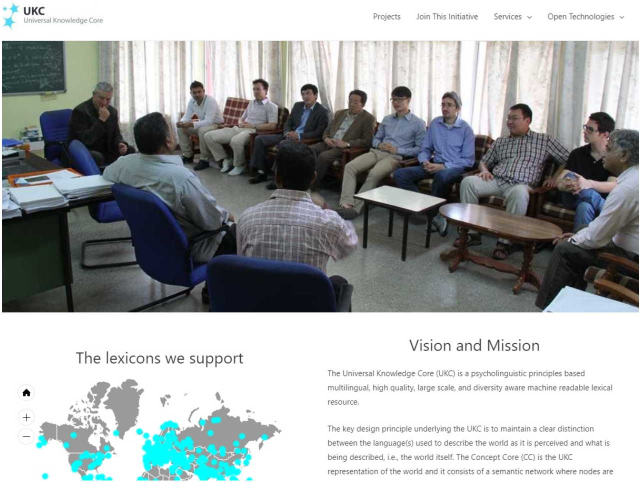
34 Open Wordnets Merged

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[3]

[3] http://compling.hss.ntu.edu.sg/omw/

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

Figure: UKC Home[4]

[4] http://ukc.disi.unitn.it

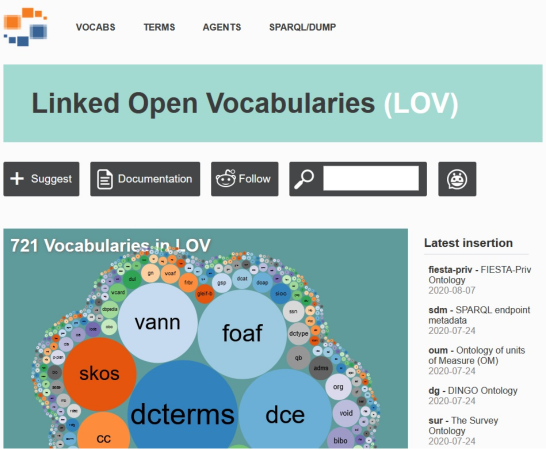
Knowledge Resources

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



Linked Open Vocabularies (LOV)

721 Vocabularies in LOV

Latest insertion

fiesta-priv - FIESTA-Priv Ontology
2020-08-07

sdm - SPARQL endpoint metadata
2020-07-24

oum - Ontology of units of Measure (OM)
2020-07-24

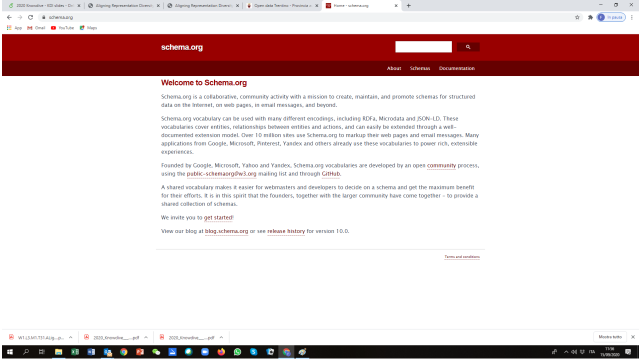
dg - DINGO Ontology
2020-07-24

sur - The Survey Ontology
2020-07-24

Figure: Linked Open Vocabulary[5]

[5] https://lov.linkeddata.es/dataset/lov/

Knowledge Resource Repositories



2020 Knowdive - KDI slides - Onl
Aligning Representation Diversity
Aligning Representation Diversity
Open data Trentino - Provincia a
Home - schema.org
schema.org
In pausa
App
Gmail
YouTube
Maps
schema.org
About
Schemas
Documentation

Welcome to Schema.org

Schema.org is a collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet, on web pages, in email messages, and beyond.

Schema.org vocabulary can be used with many different encodings, including RDFa, Microdata and JSON-LD. These vocabularies cover entities, relationships between entities and actions, and can easily be extended through a well-documented extension model. Over 10 million sites use Schema.org to markup their web pages and email messages. Many applications from Google, Microsoft, Pinterest, Yandex and others already use these vocabularies to power rich, extensible experiences.

Founded by Google, Microsoft, Yahoo and Yandex, Schema.org vocabularies are developed by an open community process, using the public-schemaorg@w3.org mailing list and through GitHub.

A shared vocabulary makes it easier for webmasters and developers to decide on a schema and get the maximum benefit for their efforts. It is in this spirit that the founders, together with the larger community have come together - to provide a shared collection of schemas.

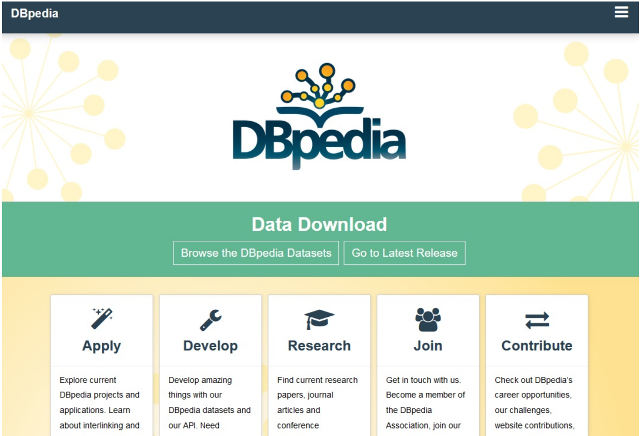
We invite you to get started!

View our blog at blog.schema.org or see release history for version 10.0.

Figure: Schema.org[6]

[6] http://www.schema.org/

Knowledge Resource Repositories



DBpedia

DBpedia

Data Download

Browse the DBpedia Datasets

Go to Latest Release

Apply

Explore current DBpedia projects and applications. Learn about interlinking and

Develop

Develop amazing things with our DBpedia datasets and our API. Need

Research

Find current research papers, journal articles and conference

Join

Get in touch with us. Become a member of the DBpedia Association, join our

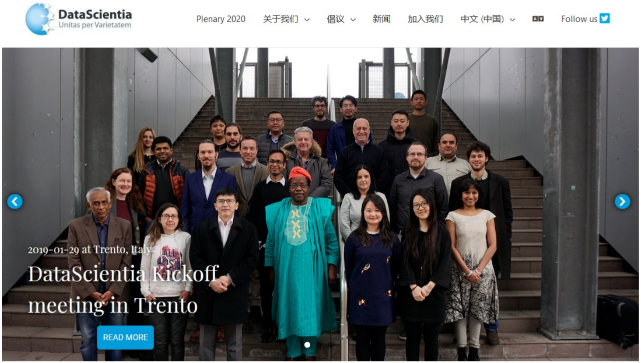
Contribute

Check out DBpedia's career opportunities, our challenges, website contributions,

Figure: DBpedia Home[7]

[7] https://wiki.dbpedia.org/

Knowledge Resource Repositories



2019-01-29 at Trento, Italy

DataScientia Kickoff meeting in Trento

Figure: DataScientia Home[8]

[8] http://datascientia.disi.unitn.it/zh/home-zh/

Data Resources

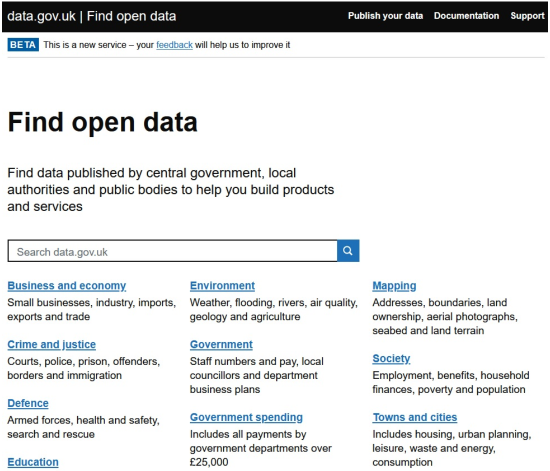
A data resource is a dataset which consists of data in some format (tabular, unstructured, entities and property values).

Open Data: data freely available. Check the licence (lots of options).

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

Search data.gov.uk

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

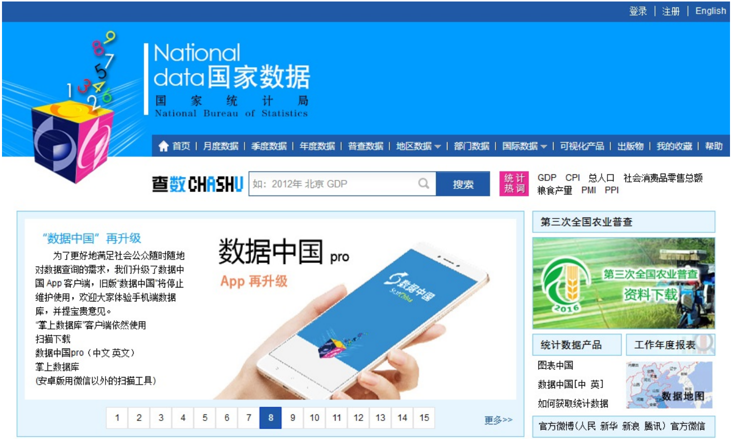
Society
Employment, benefits, household finances, poverty and population

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

Figure: Open Data UK[9]

[9] https://data.gov.uk/

Data Resource Repositories



“数据中国”再升级

为了更好地满足社会公众随时随地对数据查询的需求，我们升级了数据中国 App 客户端，旧版“数据中国”将停止维护使用，欢迎大家体验手机端数据库，并提宝贵意见。

“掌上数据库”客户端依然使用

扫描下载

数据中国pro（中文 英文）

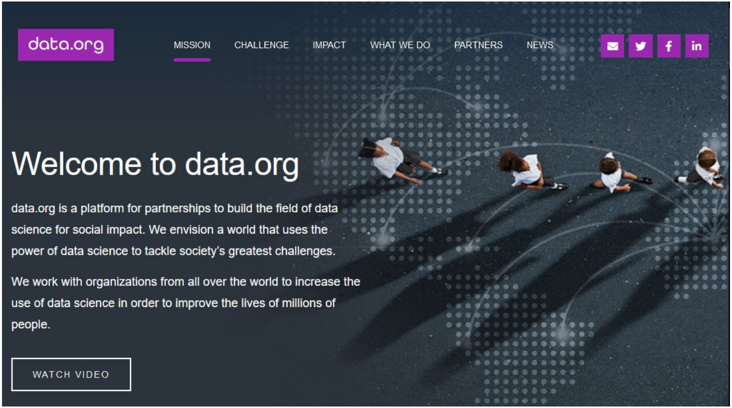
掌上数据库

(安卓版用微信以外的扫描工具)

Figure: National Bureau of Statistics, China[10]

[10] http://data.stats.gov.cn/

Data Resource Repositories



Welcome to data.org

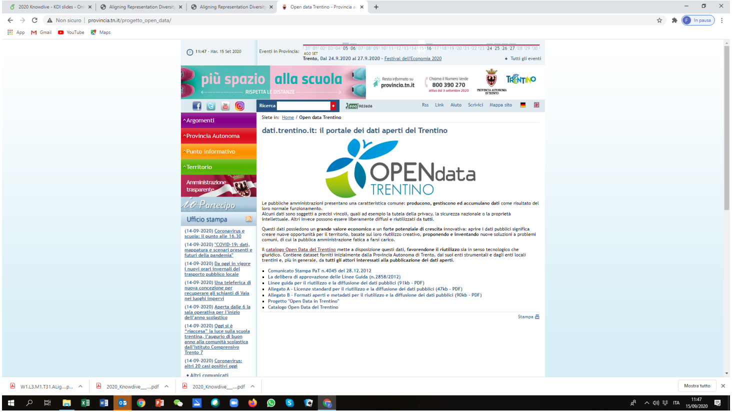
data.org is a platform for partnerships to build the field of data science for social impact. We envision a world that uses the power of data science to tackle society's greatest challenges.

We work with organizations from all over the world to increase the use of data science in order to improve the lives of millions of people.

Figure: data.org[11]

[11] https://www.data.org/

Data Resource Repositories



Trento, Dal 24.9.2020 al 27.9.2020 - Festival dell'Economia 2020

Tutti gli eventi

Siete in: Home / Open data Trentino

dati.trentino.it: il portale dei dati aperti del Trentino

Le pubbliche amministrazioni presentano una caratteristica comune: **producono, gestiscono ed accumulano dati** come risultato del loro normale funzionamento.
Alcuni dati sono soggetti a precisi vincoli, quali ad esempio la tutela della privacy, la sicurezza nazionale o la proprietà intellettuale. Altri invece possono essere liberamente diffusi e riutilizzati da **tutti**.

Questi dati possiedono un **grande valore economico** e un **forte potenziale di crescita** innovativa: aprire i dati pubblici significa creare nuove opportunità per il territorio, basate sul loro riutilizzo creativo, **proponendo e inventando** nuove soluzioni a problemi comuni, di cui la pubblica amministrazione fatica a farsi carico.

Il catalogo Open Data del Trentino mette a disposizione questi dati, **favorendone il riutilizzo** sia in senso tecnologico che giuridico. Contiene dataset forniti inizialmente dalla Provincia Autonoma di Trento, dai suoi enti strumentali e dagli enti locali trentini e, più in generale, da **tutti gli attori interessati alla pubblicazione dei dati aperti.**

- Comunicato Stampa PaT n.4045 del 28.12.2012
- La delibera di approvazione delle Linee Guida (n.2858/2012)
- Linee guida per il riutilizzo e la diffusione dei dati pubblici (91kb - PDF)
- Allegato A - Licenze standard per il riutilizzo e la diffusione dei dati pubblici (47kb - PDF)
- Allegato B - Formati aperti e metadati per il riutilizzo e la diffusione dei dati pubblici (90kb - PDF)
- Progetto "Open Data in Trentino"
- Catalogo Open Data del Trentino

(14-09-2020) Coronavirus e scuola: il punto alle 16.30

(14-09-2020) "COVID-19: dati, mappatura e scenari presenti e futuri della pandemia"

(14-09-2020) Da oggi in vigore i nuovi orari invernali del trasporto pubblico locale

(14-09-2020) Una teleferica di nuova concezione per recuperare gli schianti di Vaia nei luoghi impervi

(14-09-2020) Aperta dalle 6 la sala operativa per l'inizio dell'anno scolastico

(14-09-2020) Oggi si è "riaccesa" la luce sulla scuola trentina, l'augurio di buon anno alla comunità scolastica dall'Istituto Comprensivo Trento 7

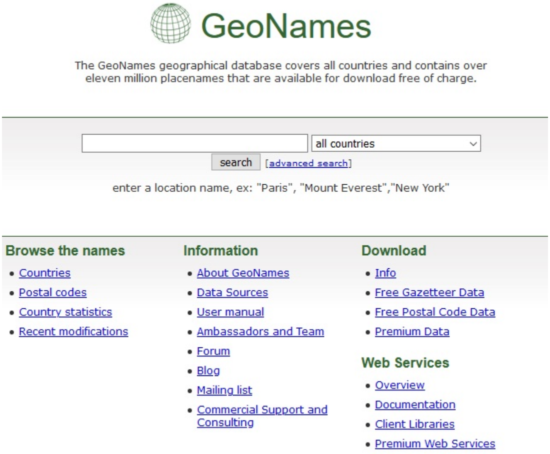
(14-09-2020) Coronavirus: altri 20 casi positivi oggi

Altri comunicati

Figure: Open Data Trentino[12]

[12] http://dati.trentino.it/

Data Resource Repositories



GeoNames

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

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

Browse the names

- Countries
- Postal codes
- Country statistics
- Recent modifications

Information

- About GeoNames
- Data Sources
- User manual
- Ambassadors and Team
- Forum
- Blog
- Mailing list
- Commercial Support and Consulting

Download

- Info
- Free Gazetteer Data
- Free Postal Code Data
- Premium Data

Web Services

- Overview
- Documentation
- Client Libraries
- Premium Web Services

Figure: Geonames Home[13]

[13] https://www.geonames.org/

Data Resource Repositories

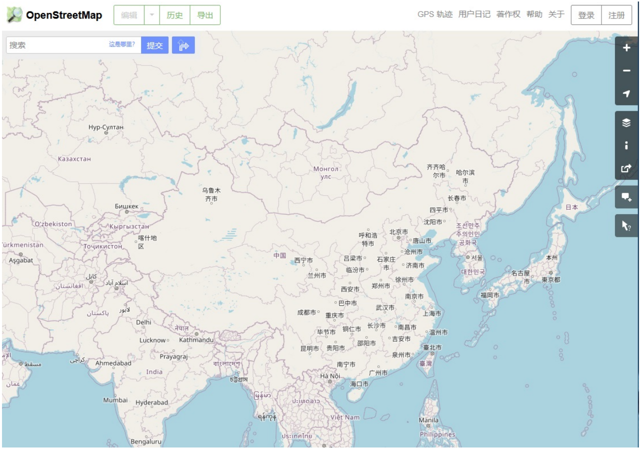
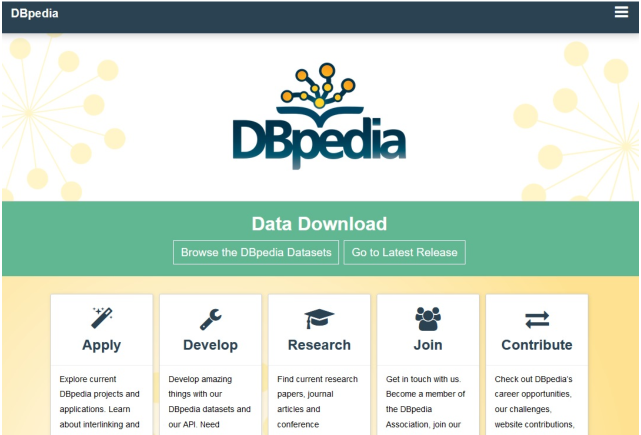


Figure: Open Street Map Home[14]

[14] https://www.openstreetmap.org/

Data Resource Repositories



DBpedia

DBpedia

Data Download

Browse the DBpedia Datasets | Go to Latest Release

Apply

Explore current DBpedia projects and applications. Learn about interlinking and

Develop

Develop amazing things with our DBpedia datasets and our API. Need

Research

Find current research papers, journal articles and conference

Join

Get in touch with us. Become a member of the DBpedia Association, join our

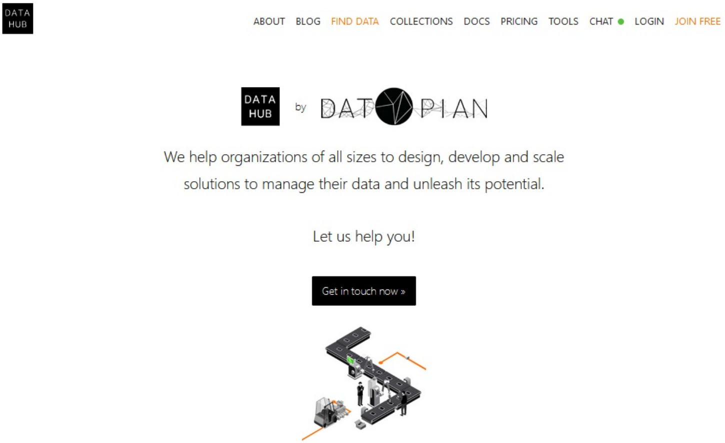
Contribute

Check out DBpedia's career opportunities, our challenges, website contributions,

Figure: DBpedia Home[15]

[15] https://wiki.dbpedia.org/

Data Resource Repositories



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

Let us help you!

Figure: Data Hub Home[16]

[16] https://datahub.io/

Contents

1 Build a KG by reuse

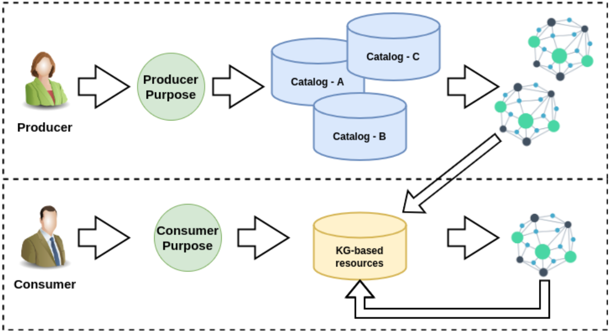
2 Reusable Resources

3 Existing Reusable Resources

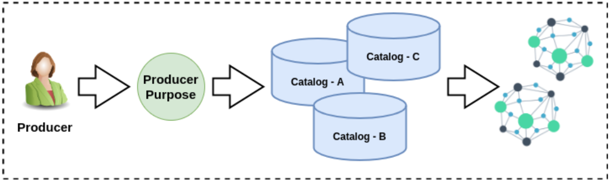
4 **Data reuse - Producer & Consumer**

Data reuse - Producer & Consumer

- The Purpose changes between producer and consumer users.
- Different kinds of resources are considered if the Purpose aims to produce **new structured resources** (suitable to be reused by KGE processes), or to exploit as much as possible, **already existing KG-based resources**.



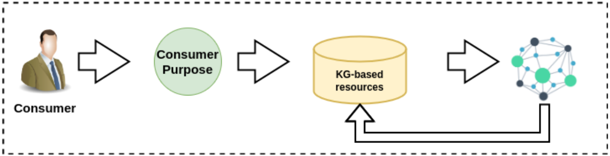
Reuse for the Data Producer



- The producer collects resources from catalogs, and/or produces from scratch the resources she needs, with the objective of **produce KG-based version of such resources**, which can be exploited for specific purposes.

- The producer has to deal with the data **heterogeneity** present in the existing catalogs resources:
  - Low quality resources.
  - Noisy resources.
  - Resources not described by metadata.
  - Missing values.
  - Syntactic heterogeneity (see next lectures)

Reuse for the Data Consumer



- The consumer collects KG-based resources with the objective of **produce KG-based new resources by composition**, supporting her specific purpose.
- The consumer has to deal mostly with the data semantic heterogeneity (see next lectures).
- The KG-based resources already have some useful features like:
  - More quality.
  - Structured resources (Knowledge + Data layer).
  - Mandatory metadata description.
  - Syntactically aligned (see Syntactic heterogeneity lecture)

KGE - Knowledge Graph Engineering

Fausto Giunchiglia

The Reuse Problem