Funded by the National Science Foundation, Grant #2127548

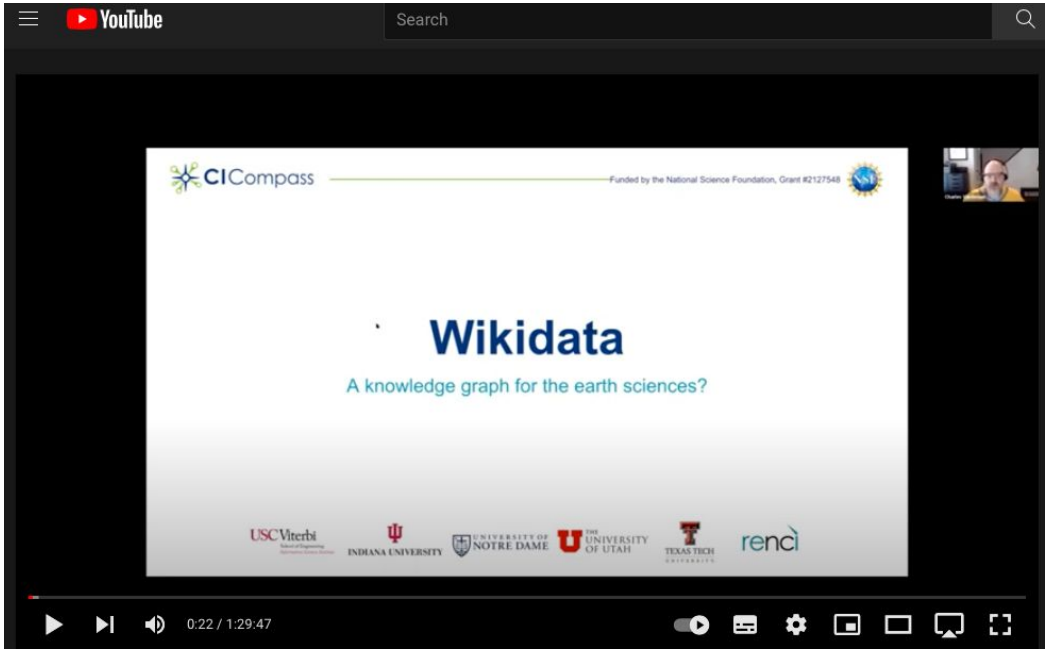
Wikidata

A knowledge graph for the earth sciences?

USC Viterbi School of Engineering Information Sciences Institute
INDIANA UNIVERSITY
UNIVERSITY OF NOTRE DAME
THE UNIVERSITY OF UTAH
TEXAS TECH UNIVERSITY
renci

ESIP Winter Meeting Session

https://youtu.be/3oN67CfirDI



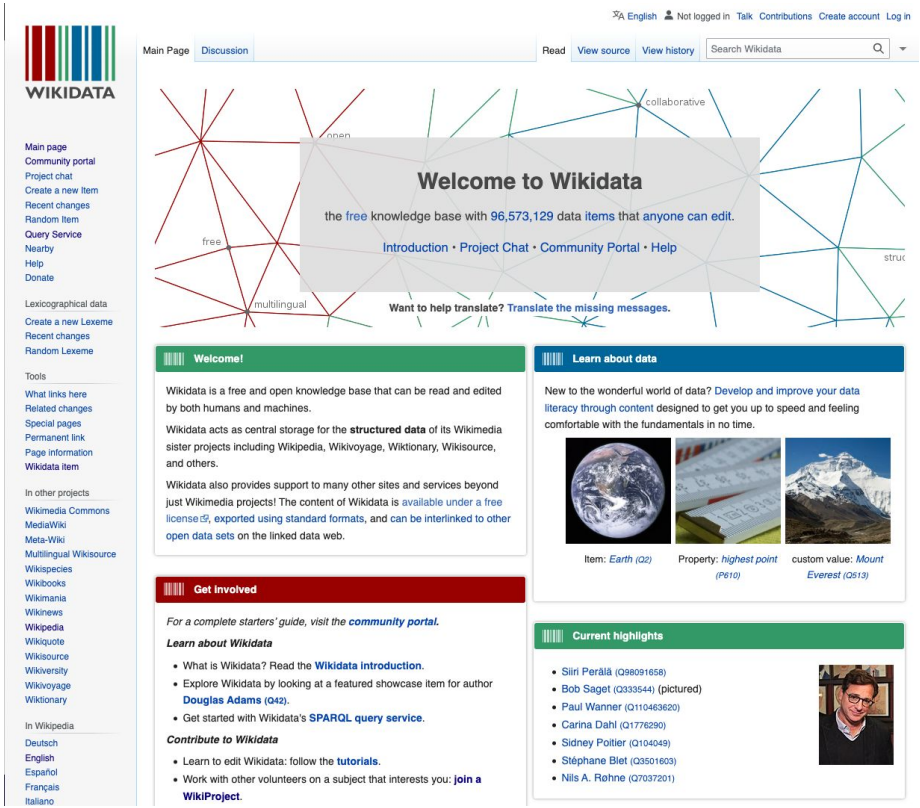
What is Wikidata?

“Wikidata is a free and open knowledge base that can be read and edited by both humans and machines.

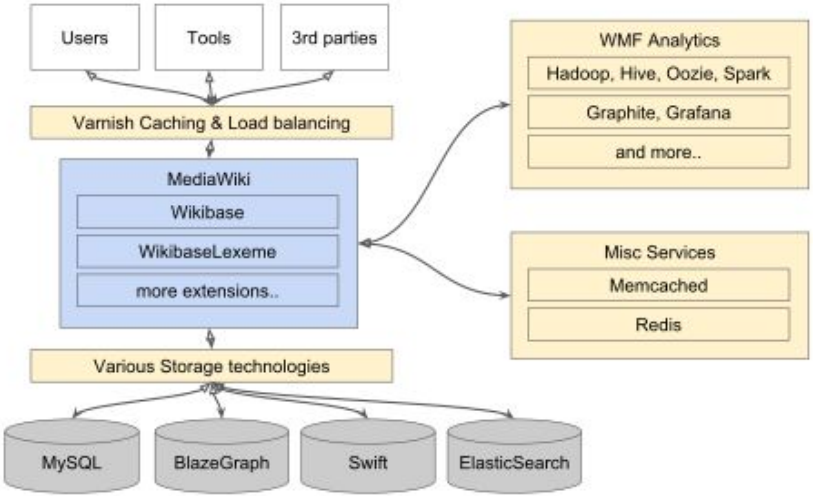
Wikidata acts as central storage for the **structured data** of its Wikimedia sister projects including Wikipedia, Wikivoyage, Wiktionary, Wikisource, and others.

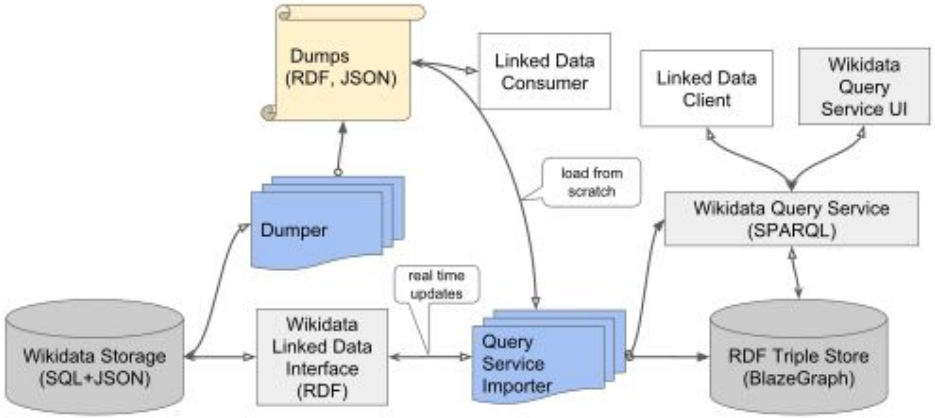
Wikidata also provides support to many other sites and services beyond just Wikimedia projects! The content of Wikidata is available under a free license, exported using standard formats, and can be interlinked to other open data sets on the linked data web.”

https://www.wikidata.org/wiki/Wikidata:Main_Page



Wikidata High Level Architecture: Storage and Query





https://w.wiki/4fei
https://addshore.com/2018/12/wikidata-architecture-overview-diagrams/

Motivating Article from the Life Sciences…

Computational and Systems Biology

Science Forum: Wikidata as a knowledge graph for the life sciences

Andra Waagmeester, Gregory Stupp, Sebastian Burgstaller-Muehlbacher, Benjamin M Good, Malachi Griffith, Obi L Griffith, Kristina Hanspers, Henning Hermjakob, Toby S Hudson see all »

Micelio, Belgium; Department of Integrative Structural and Computational Biology, The Scripps Research Institute, United States; Center for Integrative Bioinformatics Vienna, Max Perutz Laboratories, University of Vienna and Medical University of Vienna, Austria; McDonnell Genome Institute, Washington University School of Medicine, United States; Institute of Data Science and Biotechnology, Gladstone Institutes, United States; European Bioinformatics Institute (EMBL-EBI), United Kingdom; School of Chemistry, The University of Sydney, Australia; Division of Allergy and Infectious Diseases, Department of Medicine, University of Washington, United States; Wellcome Trust Sanger Institute, United Kingdom see all »

Feature Article · Mar 17, 2020

Cited 17 Views 6,492 Annotations 0

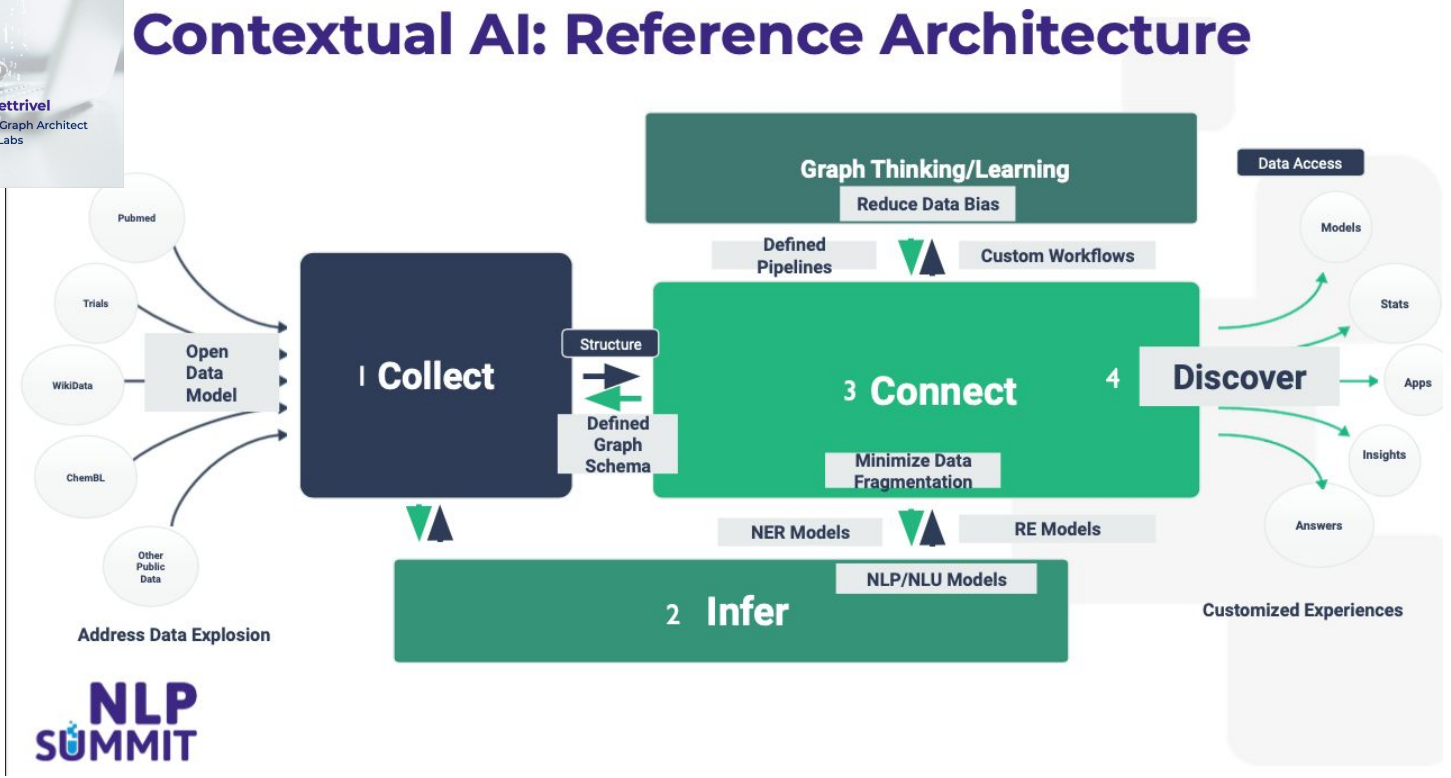
Cite as: eLife 2020;9:e52614 DOI: 10.7554/eLife.52614

Abstract

Wikidata is a community-maintained knowledge base that has been assembled from repositories in the fields of genomics, proteomics, genetic variants, pathways, chemical compounds, and diseases, and that adheres to the FAIR principles of findability, accessibility, interoperability and reusability. Here we describe the breadth and depth of the biomedical knowledge contained within Wikidata, and discuss the open-source tools we have built to add information to Wikidata and to synchronize it with source databases. We also demonstrate several use cases for Wikidata, including the crowdsourced curation of biomedical ontologies, phenotype-based diagnosis of disease, and drug repurposing.

https://elifesciences.org/articles/52614

Wikidata use by ML Communities



https://www.nlpsummit.org/nlp-2021/

Collaborative, Reusable Tooling

Wikidata:Recoin

Translate this page

Other languages: Deutsch • **English** • dansk • español • polski • اردو • سنڌي • 中文 • 日本語

Recoin

Recoin ("Relative Completeness Indicator") is a script that extends Wikidata entity pages with information about the *relative completeness* of the information. Relative completeness refers to the extent of information found on an item in comparison with other similar items.

The indicator aggregates the extent of information into a colored progress bar, showing 5 possible color-coded levels of completeness that range from *very detailed information* to *very basic information*.

Recoin is intended to both help authors to know where to potentially focus their attention, and to make data consumers aware of the degree of information found in a specific article.

Recoin adds a status indicator (top right) and two expandable lists of important absent properties and IDs to Wikidata (center). Here shown for Abbey Road, for which data is very detailed.

Max Planck Institute for Informatics: Detailed information

Arno Kompatscher: Basic information

Contents [hide]

1 Motivation
2 What it shows
3 How it works
  3.1 Architecture
  3.2 Computation
    3.2.1 Determination of absent properties and IDs
    3.2.2 Status indicator computation
    3.2.3 Treatment of multi-class-membership
    3.2.4 Special cases
  3.3 Multilinguality
4 Installation
  4.1 Main Gadget
  4.2 Special version: IDs only
5 APIs
  5.1 Per-entity access
  5.2 Per-class access
6 Data Dumps
7 Further information

Motivation [edit]

Recoin is intended to assist both authors and consumers of Wikidata.

For *users* (consumers), it provides a handy summary of the degree of completeness of information in Wikidata, which may help them in deciding whether to rely on Wikidata or not in order to satisfy their information need. This is because judging purely by article length may not always be a good idea, as for instance the chess player Jeff Sarwer (Q3494327) has a long article due to lots of statements about his Elo rating, but until recently was missing even very basic information such as citizenship or family name.

For *authors*, similarly it provides information about which persons' information is more complete than others', thus allowing them to focus attention on more incomplete persons. For an individual person, it allows them to see the most important properties that are missing, which authors they might focus on completing, or, if no values for these properties exist, might mark this with a *novalue* assertion.

cyberinfrastructure center of excellence, major facilities data lifecycle

(Q108615661)

National Science Foundation Cyberinfrastructure Center of Excellence for NSF Major Facilities Data Lifecycle edit

CI-Compass | cicompass

Recoin: Most relevant properties which are absent

Property ID	Label	Relative	Add Claim
P2769	budget	52.73%	
P276	location	45.77%	+
P571	inception	9.7%	
P1476	title	8.47%	
P361	part of	6.42%	+
P1416	affiliation	6.01%	+
P373	Commons category	5.05%	
P101	field of work	4.92%	+
P1344	participant in	4.51%	+
P18	image	3.55%	

In more languages

Configure

Language	Label	Description	Also known as
English	cyberinfrastructure center of excellence, major facilities data lifecycle	National Science Foundation Cyberinfrastructure Center of Excellence for NSF Major Facilities Data Lifecycle	CI-Compass cicompass
Spanish	No label defined	No description defined	
Traditional Chinese	No label defined	No description defined	
Chinese	No label defined	No description defined	

Statements

instance of	research grant	edit
	0 references	+ add reference
	research project	edit
	0 references	+ add reference

Integrative Queries

```
SELECT DISTINCT ?compound ?compoundLabel where {

  # gene has genetic association with a respiratory disease
(i)  ?gene       wdt:P31    wd:Q7187 .
  ?gene       wdt:P2293  ?diseaseGA .
  ?diseaseGA  wdt:P279*  wd:Q3286546 .

  # gene product is localized to the membrane
(ii)  ?gene     wdt:P688             ?protein .
  ?protein  wdt:P681             ?cc .
  ?cc       wdt:P279*|wdt:P361*  wd:Q14349455 .

  # gene is involved in a pathway with another gene ("gene2")
(iii)  ?pathway  wdt:P31   wd:Q4915012 ;
           wdt:P527  ?gene ;
           wdt:P527  ?gene2 .
  ?gene2    wdt:P31   wd:Q7187 .

  # gene2 product has a Ser/Thr protein kinase domain AND
  #       known enzyme inhibitor
(iv)  ?gene2     wdt:P688  ?protein2 .
  ?protein2  wdt:P129  ?compound ;
(v)          wdt:P527  wd:Q24787419 ;
           p:P129    ?s2 .
  ?s2        ps:P129   ?cp2 .
  ?compound  wdt:P31   wd:Q11173 .
  FILTER EXISTS {?s2 pq:P366 wd:Q427492 .}

  SERVICE wikibase:label { bd:serviceParam wikibase:language "en". }
}
```

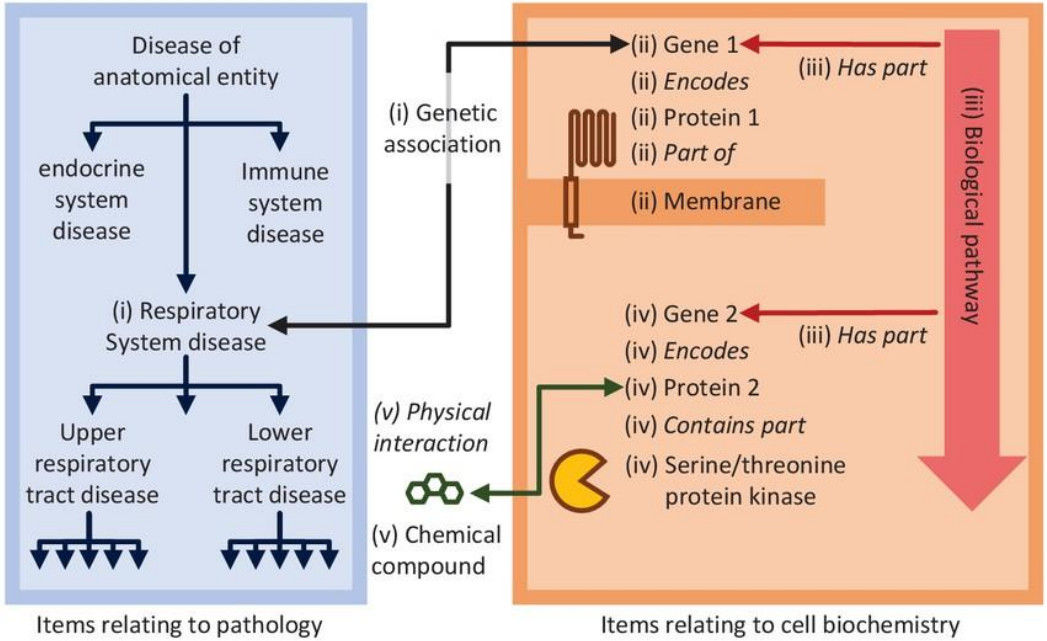
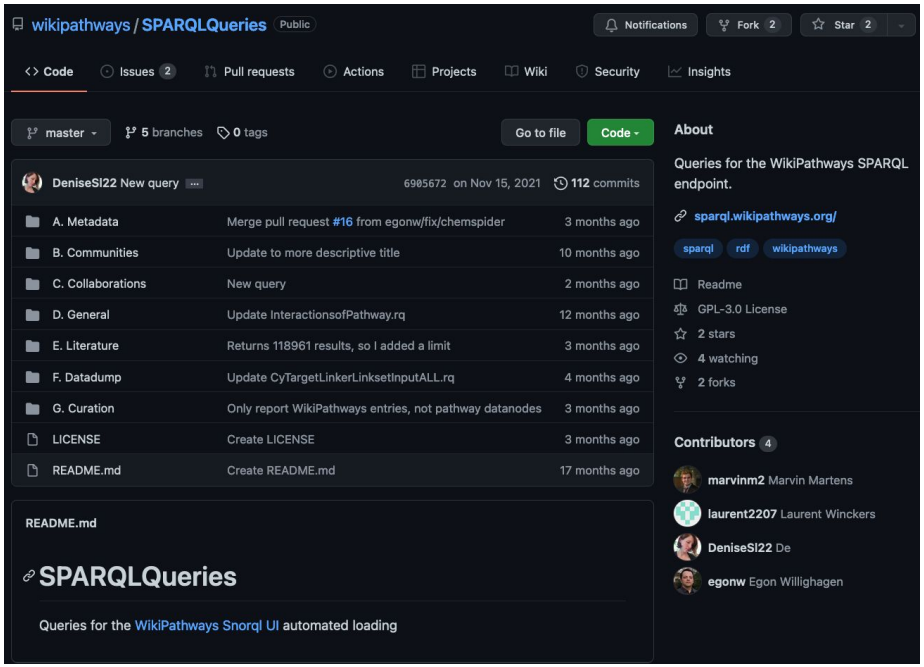
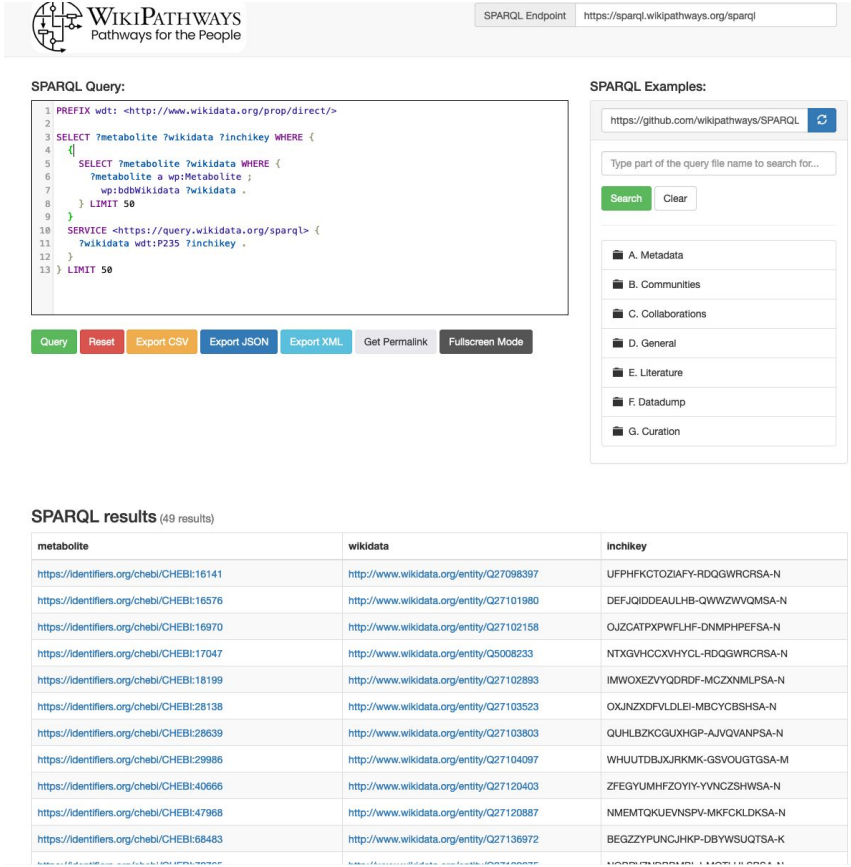


Figure 3, “Science Forum: Wikidata as a knowledge graph for the life sciences”,

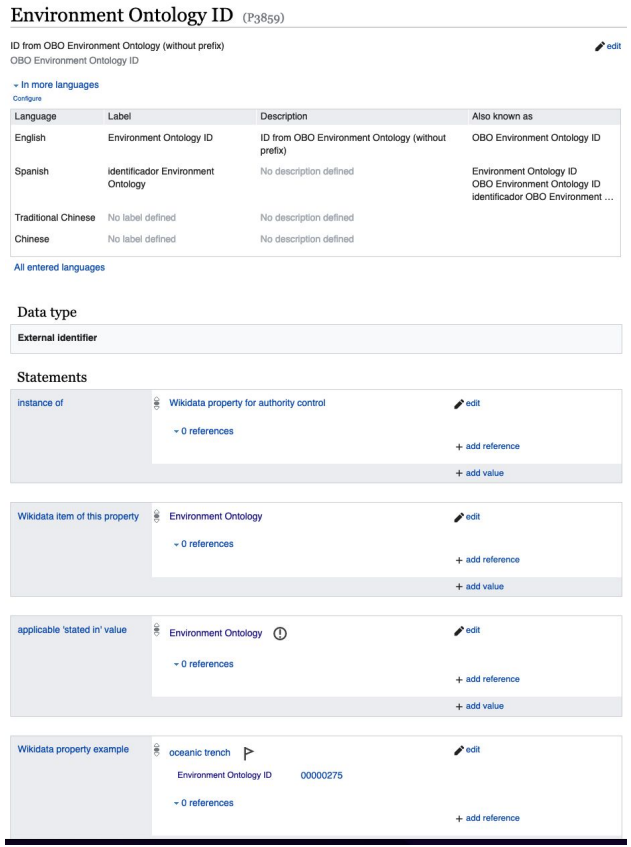
Wikipathways tooling reuse





https://www.wikipathways.org/index.php/WikiPathways

Wikidata external identifier property for crosswalks



Environment Ontology ID (P3859)

ID from OBO Environment Ontology (without prefix) edit

OBO Environment Ontology ID

In more languages

Configure

Language	Label	Description	Also known as
English	Environment Ontology ID	ID from OBO Environment Ontology (without prefix)	OBO Environment Ontology ID
Spanish	identificador Environment Ontology	No description defined	Environment Ontology ID OBO Environment Ontology ID identificador OBO Environment …
Traditional Chinese	No label defined	No description defined	
Chinese	No label defined	No description defined	

All entered languages

Data type

External identifier

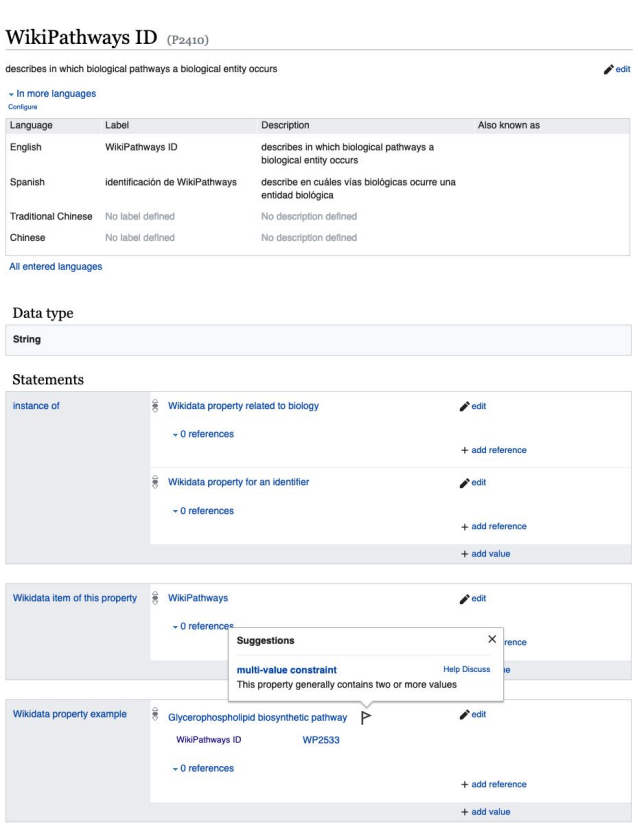
Statements

instance of — Wikidata property for authority control — edit
0 references
+ add reference
+ add value

Wikidata item of this property — Environment Ontology — edit
0 references
+ add reference
+ add value

applicable 'stated in' value — Environment Ontology — edit
0 references
+ add reference
+ add value

Wikidata property example — oceanic trench — edit
Environment Ontology ID 00000275
0 references
+ add reference



WikiPathways ID (P2410)

describes in which biological pathways a biological entity occurs edit

In more languages

Configure

Language	Label	Description	Also known as
English	WikiPathways ID	describes in which biological pathways a biological entity occurs	
Spanish	identificación de WikiPathways	describe en cuáles vías biológicas ocurre una entidad biológica	
Traditional Chinese	No label defined	No description defined	
Chinese	No label defined	No description defined	

All entered languages

Data type

String

Statements

instance of — Wikidata property related to biology — edit
0 references
+ add reference

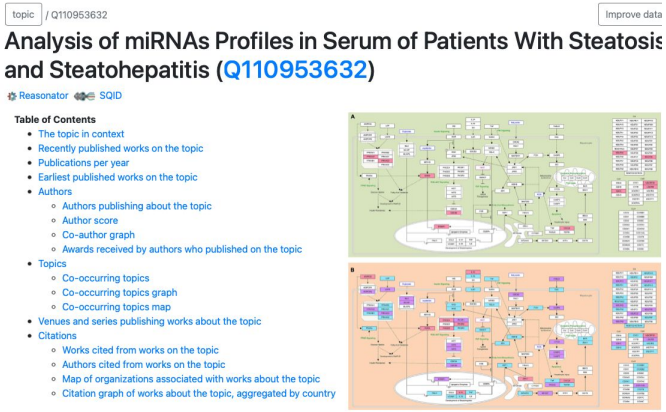
Wikidata property for an identifier — edit
0 references
+ add reference
+ add value

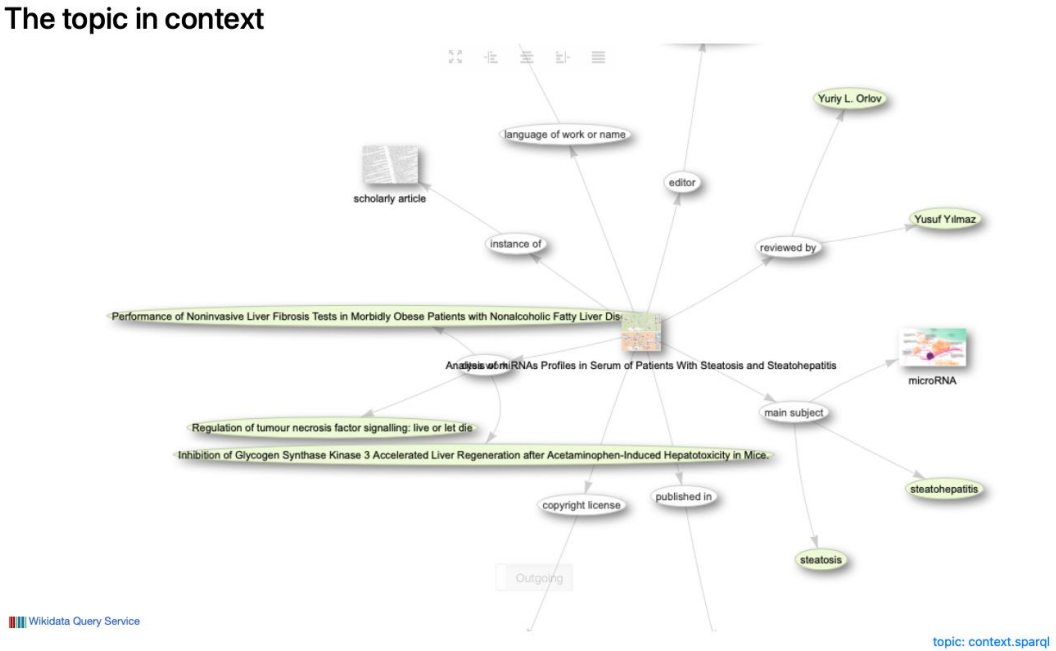
Wikidata item of this property — WikiPathways — edit
0 references

Suggestions
multi-value constraint Help Discuss
This property generally contains two or more values

Wikidata property example — Glycerophospholipid biosynthetic pathway — edit
WikiPathways ID WP2533
0 references
+ add reference
+ add value

Linking Publications to Data Sets Though Scholia and Wikidata





https://scholia.toolforge.org/topic/Q110953632

Linking Publications to Data Sets Though Scholia and Wikidata

work / Q110953632 Improve data

Analysis of miRNAs Profiles in Serum of Patients With Steatosis and Steatohepatitis (Q110953632)

Search:

Description	Value
Title	Analysis of miRNAs Profiles in Serum of Patients With Steatosis and Steatohepatitis
Language	English
Editors	Anastasia Efimenko
Reviewers	Yuriy L. Orlov, Yusuf Yılmaz
Published in	Frontiers in cell and developmental biology
Publication date	2021-09-09
Topics	microRNA, steatosis, steatohepatitis
DOI	10.3389/FCELL.2021.736677 ↗
Full text	https://www.frontiersin.org/articles/10.3389/fcell.2021.736677/pdf ↗

Wikidata Query Service work: data.sparql

https://scholia.toolforge.org/topic/Q110953632

Nonalcoholic fatty liver disease (Q74069909)

Nonalcoholic fatty liver disease (NAFLD) is a condition in which too much fat is stored in the liver. Nonalcoholic steatohepatitis (NASH) is a type of NAFLD marked by liver inflammation, which may progress to scarring and irreversible damage simi... edit

In more languages

Configure

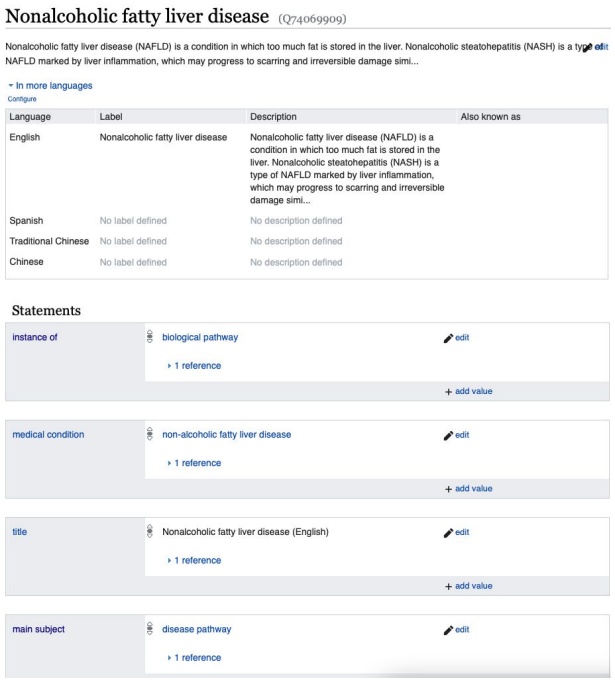
Language	Label	Description	Also known as
English	Nonalcoholic fatty liver disease	Nonalcoholic fatty liver disease (NAFLD) is a condition in which too much fat is stored in the liver. Nonalcoholic steatohepatitis (NASH) is a type of NAFLD marked by liver inflammation, which may progress to scarring and irreversible damage simi...	
Spanish	No label defined	No description defined	
Traditional Chinese	No label defined	No description defined	
Chinese	No label defined	No description defined	

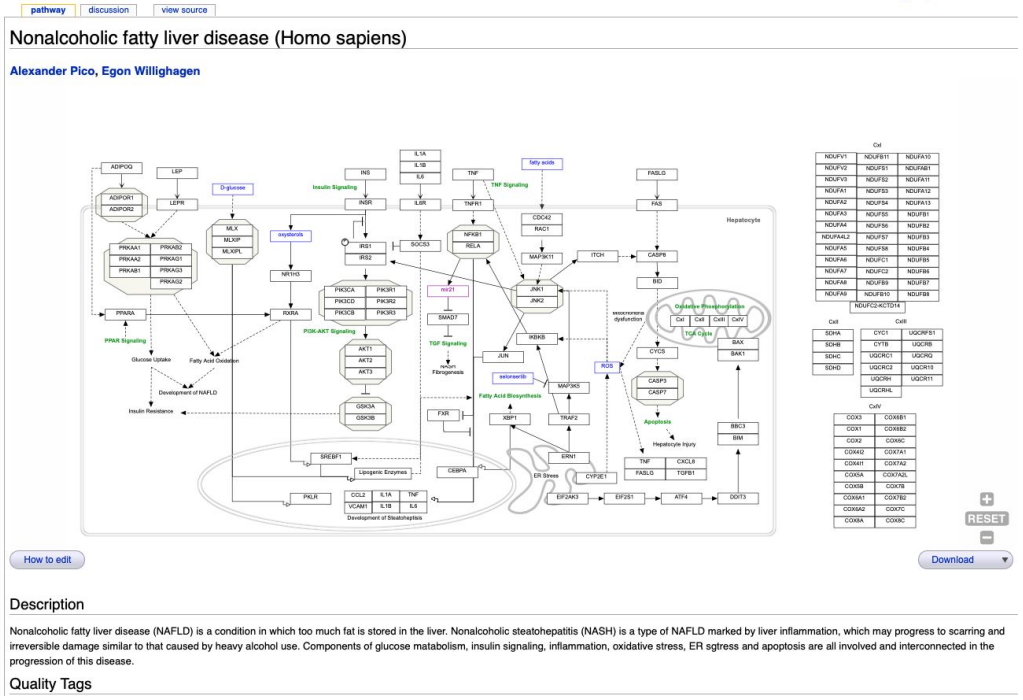
Statements

instance of	biological pathway 1 reference	edit add value
medical condition	non-alcoholic fatty liver disease 1 reference	edit add value
title	Nonalcoholic fatty liver disease (English) 1 reference	edit add value
main subject	disease pathway 1 reference	edit

https://www.wikidata.org/wiki/Q74069909

Linking Publications to Data Sets Though Scholia and Wikidata





https://www.wikidata.org/wiki/Q74069909

https://www.wikipathways.org/instance/WP4396

Example of leveraging Wikidata for building applications from the science drone community

Linked-data API for Networked Drones (Q109626199)

Cyberinfrastructure for unmanned systems scientific platforms edit

LANDRS

Recoin: Most relevant properties which are absent

In more languages

Configure

Language	Label	Description	Also known as
English	Linked-data API for Networked Drones	Cyberinfrastructure for unmanned systems scientific platforms	LANDRS
Spanish	No label defined	No description defined	
Traditional Chinese	No label defined	No description defined	
Chinese	No label defined	No description defined	

Statements

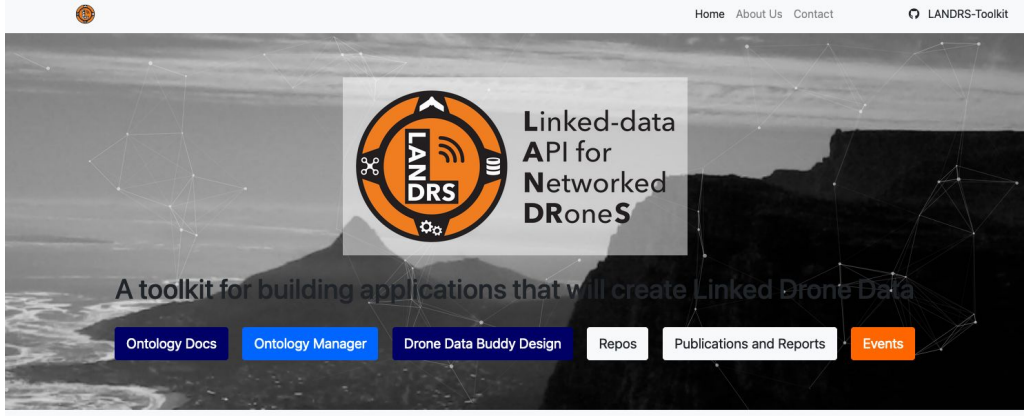
instance of
- research grant edit
  - 1 reference
- research project edit
  - 1 reference
- add value

official name
- Linked-data API for Networked Drones (English) edit
  - 1 reference
- add value

short name
- LANDRS (English) edit
  - 0 references
  - add reference
- add value

country
- South Africa edit
  - 0 references
  - add reference
- United States of America edit
  - 0 references

https://www.wikidata.org/wiki/Q109626199



Cyberinfrastructure for making Drone data F.A.I.R

Researchers are increasingly making small Unmanned Aircraft Systems [sUAS] -- aka drones -- a standard platform for scientific data capture. As a nascent and emerging tool, the need for drone data management cyberinfrastructure has increased as adoption has grown. For the many reasons there is much to be gained by academia and society by meeting this need. This project is working to build open source, community driven, foundational data management cyberinfrastructure so as to enable drone data to be easily made FAIR [Findable Accessible Interoperable Reusable].

For more on why we think this is particularly important for drone data see Emergent Challenges for Science sUAS Data Management Fairness through Community Engagement and Best Practices Development

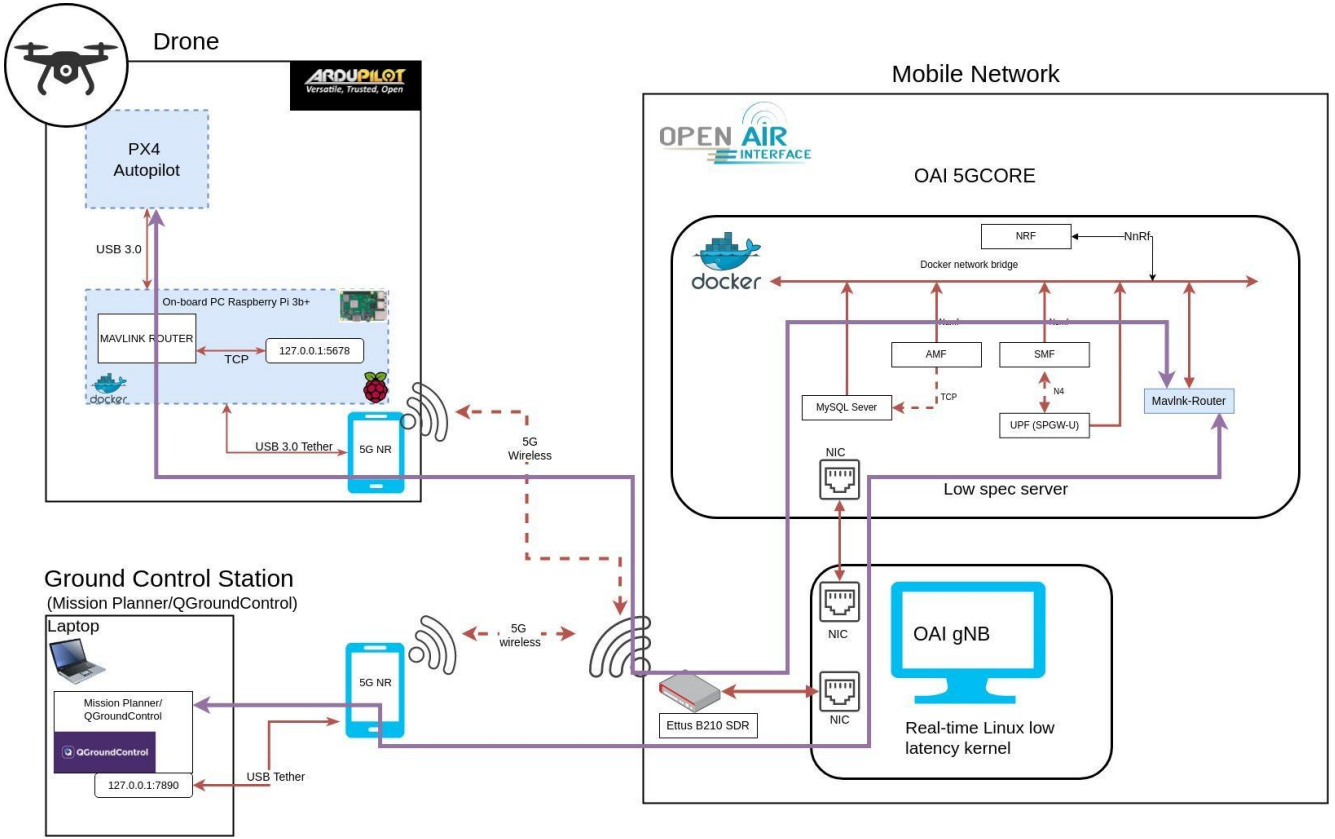
Project Goals

This Sloan foundation funded project is working to create a suit of tools and example applications to facilitate the building of CI for drone data management via:

- An ontology that aligns with existing best practice ontologies such as OGC's SOSA, NASA's SWEET, and OBO Foundary ontologies
- RESTFUL APIs for building web-native data handling tools
- A set of proposed drone data best practices developed through global community discussion

https://www.landrs.org/

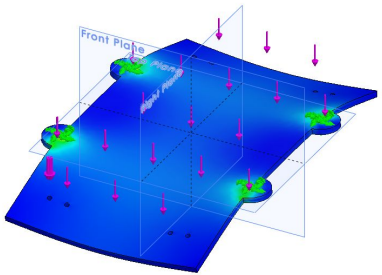
Open Software Defined Radio based 5G Drone Data Streaming

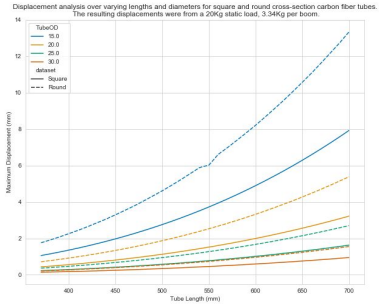


Msc student Ngoni Mombeshora
mmbngo003@myuct.ac.za
University of Cape Town
Co-supervisor: Joyce Mwangama
Built in LANDRS Tooling
Delivery: July 2022

Open Science Multirotor Design

- Formally designed & tested
- Flexible: Quad&Hex
- Open Source & Documented
- Target: Science Data Capture Applications
- Multifunction attachment
- On wikidata: https://www.wikidata.org/wiki/Q110041198





Msc student Mauro Borrageiro
brrmau002@myuct.ac.za
University of Cape Town
Built in LANDRS Tooling
Delivery: July 2022

LANDRS Science Drone (Q110040336)

Specification for open source multirotor drone for scientific observations edit

LSD

Recoin: Most relevant properties which are absent

In more languages

Configure

Language	Label	Description	Also known as
English	LANDRS Science Drone	Specification for open source multirotor drone for scientific observations	LSD
Spanish	No label defined	No description defined	
Traditional Chinese	No label defined	No description defined	
Chinese	No label defined	No description defined	

Statements

instance of	electronic device model series	edit
	0 references	add reference
		add value
subclass of	Remotely Piloted Aircraft System	edit
	0 references	add reference
	open source hardware	edit
	0 references	add reference
	technical standard	edit
	0 references	add reference
	robot	edit
	0 references	add reference
		add value
developer	Linked-data API for Networked Drones	edit
	0 references	add reference
		add value

https://www.wikidata.org/wiki/Q110040336

1 drone/
2 drone/
many drone?

LANDRS Science Drone Hex1 (Q110041008)

First iteration of the LANDRS Hexacopter Science Drone edit

LSD-H1

Recoin: Most relevant properties which are absent

In more languages

Configure

Language	Label	Description	Also known as
English	LANDRS Science Drone Hex1	First iteration of the LANDRS Hexacopter Science Drone	LSD-H1
Spanish	No label defined	No description defined	
Traditional Chinese	No label defined	No description defined	
Chinese	No label defined	No description defined	

Statements

instance of	electronic device model	edit
	0 references	add reference
	robot	edit
	0 references	add reference
	Remotely Piloted Aircraft System	edit
	0 references	add reference
	open source hardware	edit
	0 references	add reference
		add value
subclass of	multirotor	edit
	0 references	add reference
		add value
part of the series	LANDRS Science Drone	edit
	0 references	add reference
		add value

https://www.wikidata.org/wiki/Q110041008

Cube Orange Standard Set (Q109997567)

Open source autopilot hardware derived from pixhawk that is part of the Cube autopilot model series edit

Cube Orange

Recoin: Most relevant properties which are absent

In more languages

Configure

Language	Label	Description	Also known as
English	Cube Orange Standard Set	Open source autopilot hardware derived from pixhawk that is part of the Cube autopilot model series	Cube Orange
Spanish	No label defined	No description defined	
Traditional Chinese	No label defined	No description defined	
Chinese	No label defined	No description defined	

Statements

instance of

- electronic device model — edit
  - 0 references — add reference
- autopilot — edit
  - 0 references — add reference
- computing platform — edit
  - 0 references — add reference
- embedded system — edit
  - 0 references — add reference
- technical standard — edit
  - 0 references — add reference

add value

image — edit

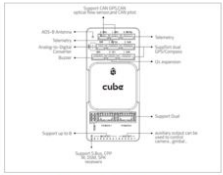
Cubepilot.png
640 × 472; 223 KB

uses

- Ardupilot — edit
  - 0 references — add reference
- PX4 autopilot — edit
  - 0 references — add reference

add value

schematic — edit



Cubepilot specifications.png
1,010 × 789; 217 KB

1 reference

add value

has part

- ICM-20649 — edit
  - 1 reference
    - reference URL: https://docs.cubepilot.org/user-guides/autopilot/the-cube-module-overview
  - add reference
- MEAS MS5611 — edit
  - 1 reference
    - reference URL: https://docs.cubepilot.org/user-guides/autopilot/the-cube-module-overview
  - add reference

add value

MEAS MS5611 (Q110016968)

Barometric Pressure Sensors edit
MS561101BA I MS5611-01BA
▸ Recoin: Most relevant properties which are absent

▾ In more languages
Configure

Language	Label	Description	Also known as
English	MEAS MS5611	Barometric Pressure Sensors	MS561101BA MS5611-01BA
Spanish	No label defined	No description defined	
Traditional Chinese	No label defined	No description defined	
Chinese	No label defined	No description defined	

Statements

instance of
- sensor — edit
  - ▾ 0 references
  - + add reference
- integrated circuit — edit
  - ▾ 0 references
  - + add reference
- barometer — edit
  - ▾ 0 references
  - + add reference

+ add value

manufacturer
- TE Connectivity — edit
  - ▾ 0 references
  - + add reference

+ add value

user manual URL
- https://www.te.com/commerce/DocumentDelivery/DDEController?Action=srchrtrv&DocNm=MS5611-01BA03&DocType=Data+Sheet&DocLang=English — edit
  - language of work or name: English
  - ▾ 0 references
  - + add reference

+ add value

measures
- atmospheric pressure — edit
  - ▾ 1 reference
    - QUDT unit ID: MilliBAR
  - + add reference

+ add value

protocol
- I²C bus — edit
  - ▾ 0 references
  - + add reference
- I²C bus — edit
  - ▾ 0 references
  - + add reference
- Serial Peripheral Interface — edit
  - ▾ 0 references
  - + add reference

+ add value

uses
- Quad Flat No-leads package — edit
  - ▾ 0 references
  - + add reference

+ add value

has part
- analog-to-digital converter — edit
  - ▾ 0 references
  - + add reference

+ add value