Documenting and preserving programming languages and software in Wikidata

John Samuel, Katherine Thornton, Kenneth Seals-Nutt

CPE Lyon, EaaSI

SWIB 2018, Bonn, 27th November, 2018

Programming Languages

Python

python™

Paradigm	Object-oriented, imperative, functional, procedural, reflective
Designed by	Guido van Rossum
Developer	Python Software Foundation
First appeared	1990; 28 years ago[1]
Stable release	3.7.0 / 27 June 2018; 2 months ago[2] 2.7.15 / 1 May 2018; 4 months ago[3]
Typing discipline	Duck, dynamic, strong **since version 3.5:** Gradual[4]
License	Python Software Foundation License
Filename extensions	.py, .pyc, .pyd, .pyo (prior to 3.5),[5] .pyw, .pyz (since 3.5)[6]
Website	www.python.org

Java

Java™

Paradigm	Multi-paradigm: object-oriented (class-based), structured, imperative, generic, reflective, concurrent
Designed by	James Gosling
Developer	Sun Microsystems (now owned by Oracle Corporation)
First appeared	May 23, 1995; 23 years ago[1]
Typing discipline	Static, strong, safe, nominative, manifest
License	GNU General Public License, Java Community Process
Filename extensions	.java, .class, .jar
Website	oracle.com/java/

C

THE C PROGRAMMING LANGUAGE

The C Programming Language[1] (often referred to as "K&R"), the seminal book on C

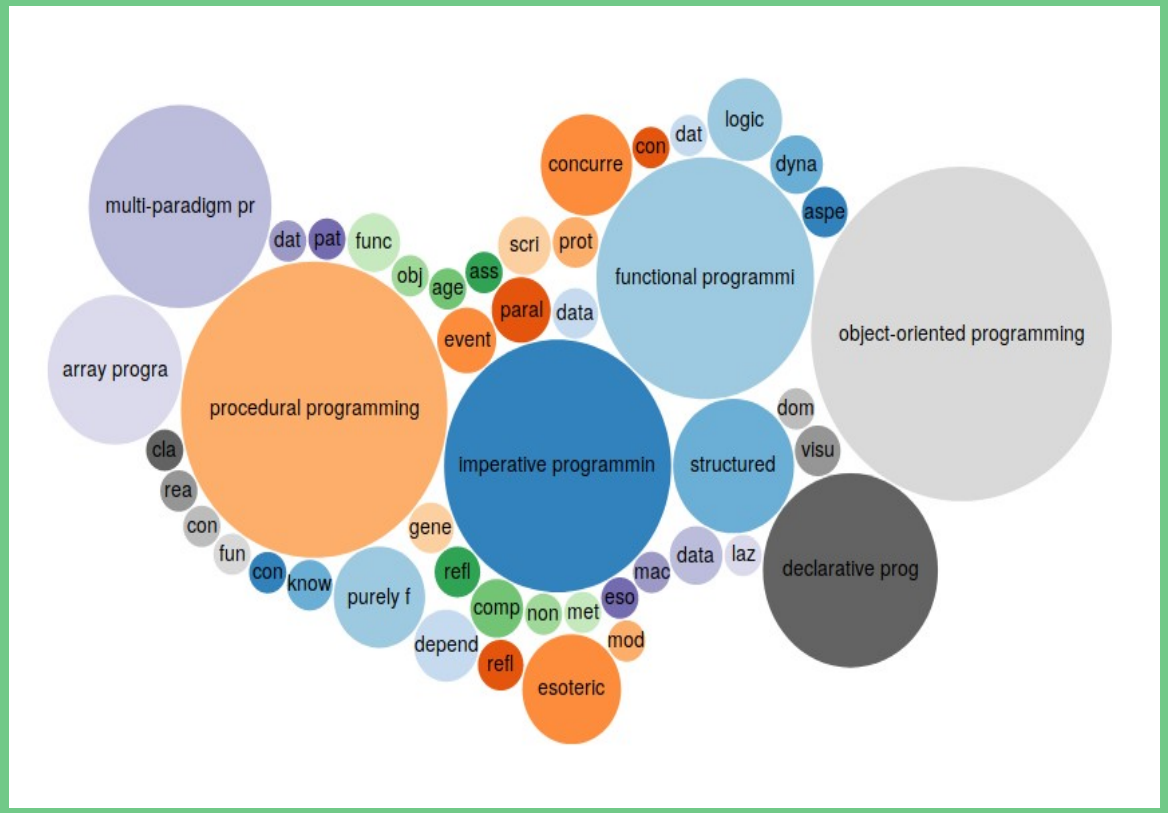
Paradigm	Imperative (procedural), structured
Designed by	Dennis Ritchie
Developer	Dennis Ritchie & Bell Labs (creators); ANSI X3J11 (ANSI C); ISO/IEC JTC1/SC22/WG14 (ISO C)
First appeared	1972; 46 years ago[2]
Stable release	C11 / December 2011; 6 years ago
Typing discipline	Static, weak, manifest, nominal
OS	Cross-platform
Filename extensions	.c, .h

English Wikipedia Infoboxes of Programming Languages

Programming Languages

Programming Languages with the most multilingual labels

Programming Languages

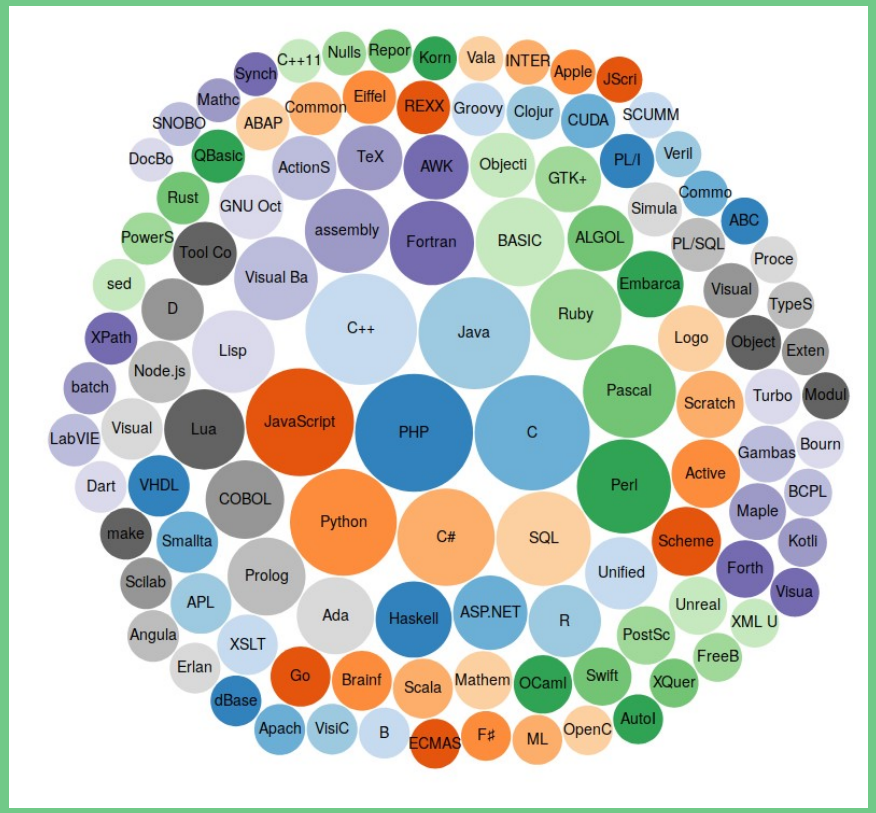


Programming Language Paradigms

Programming Languages

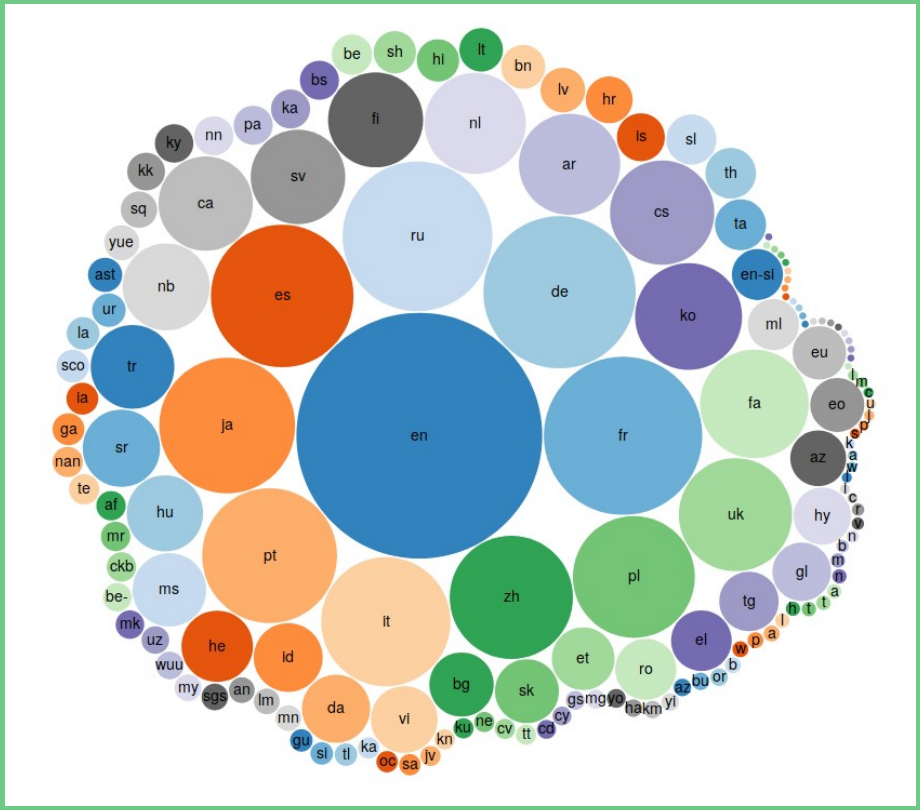
Programming Languages with the most number of different paradigms

Programming Languages



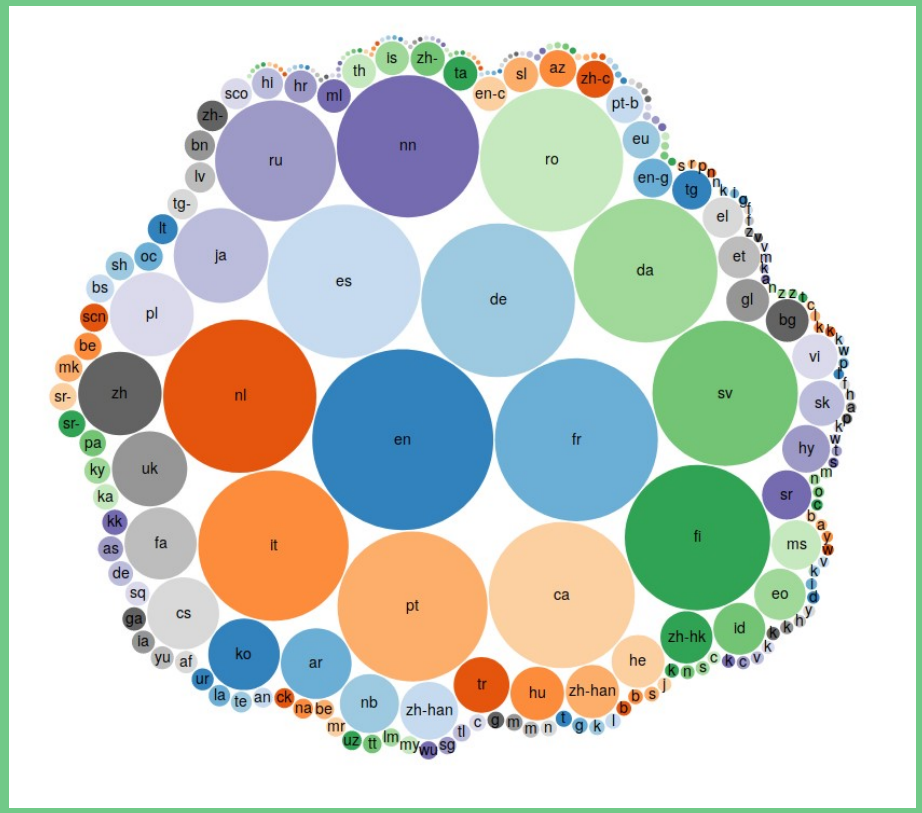
Programming languages with most number of multilingual Wikpedia articles

Programming Languages



Wikipedia languages with the most number of articles on programming languages

Programming Languages



Languages with the most number of labels of programming languages

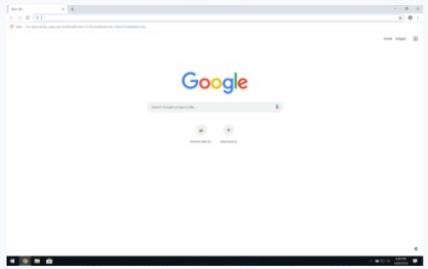
Software

Mozilla Firefox

Firefox 57.0 running on Windows 10

Developer(s)	Mozilla Foundation and its contributers Mozilla Corporation
Initial release	September 23, 2002; 15 years ago
Stable release(s) [±]	
Standard	62.0 / September 5, 2018; 14 days ago[1]
Extended Support Release	60.2.0 / September 5, 2018; 14 days ago[2]
Preview release(s) [±]	
Beta & Developer Edition	63.0beta / September 5, 2018; 14 days ago[3][4]
Nightly	64.0a1 / September 5, 2018; 14 days ago[5][6]
Repository	https://hg.mozilla.org/mozilla-central/

Google Chrome



Google Chrome on Windows 10

Developer(s)	Google LLC
Initial release	September 2, 2008; 10 years ago
Stable release(s) [±]	
Windows, macOS, Linux	69.0.3497.100 / September 17, 2018; 2 days ago[1]
Android	69.0.3497.100 / September 17, 2018; 2 days ago[2]
iOS	69.0.3497.91 / September 11, 2018; 8 days ago[3]

English Wikipedia Infoboxes of Software

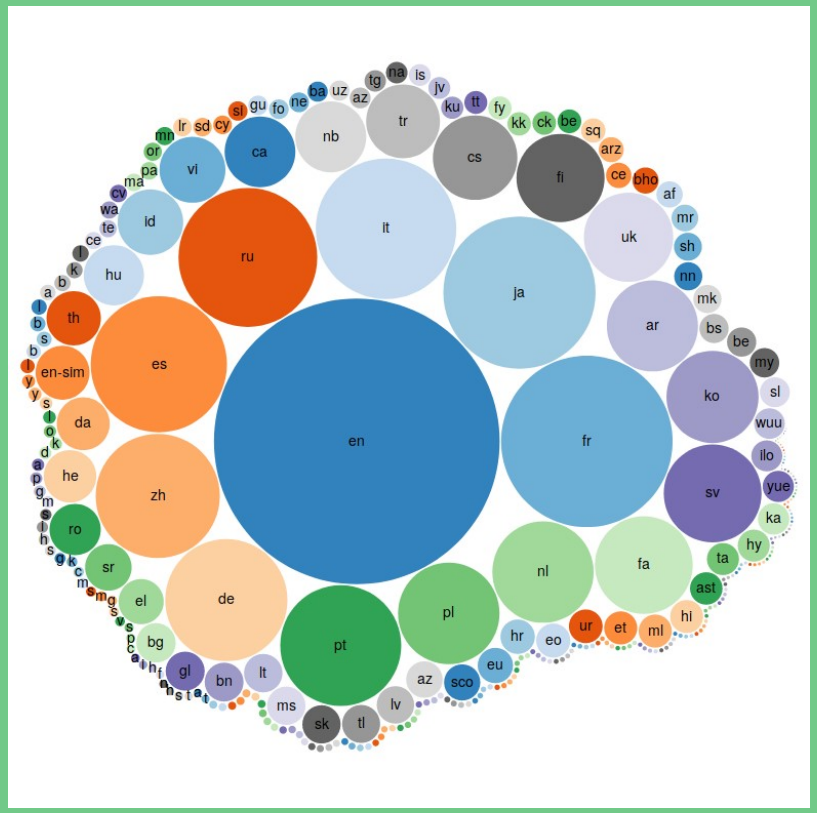
Software

Software with the most number of labels on Wikidata

Software

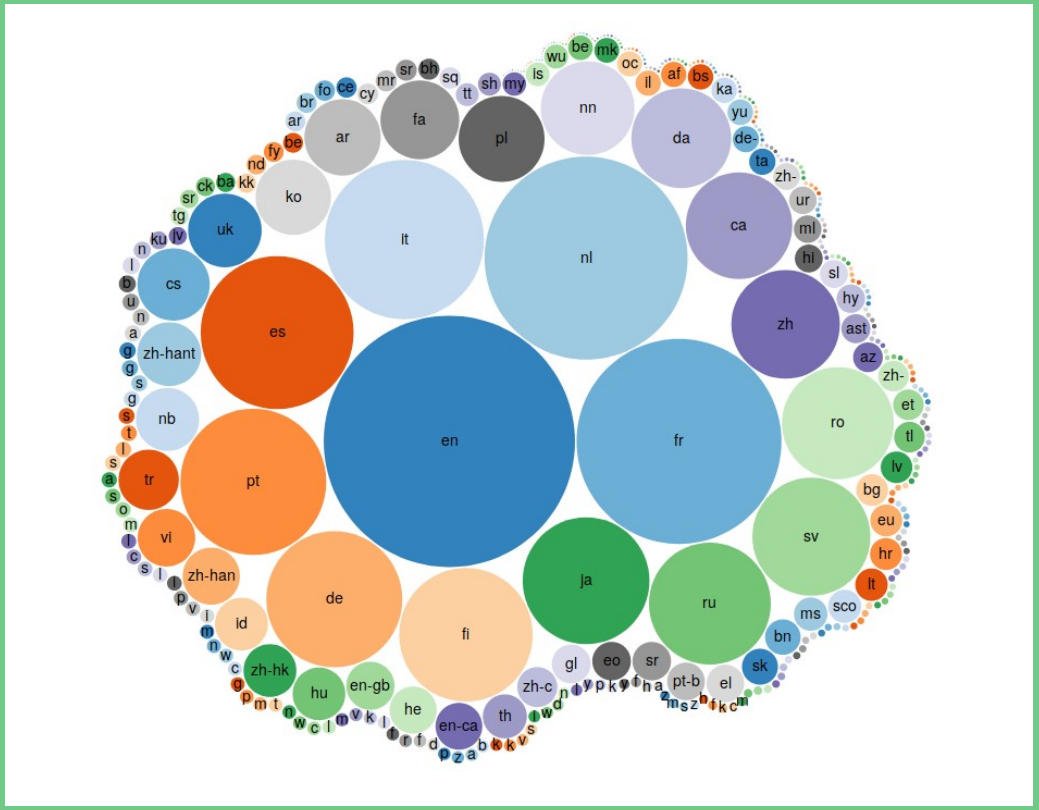
Software with the most number of articles on Wikipedia

Software



Languages with the most number of articles on Wikipedia

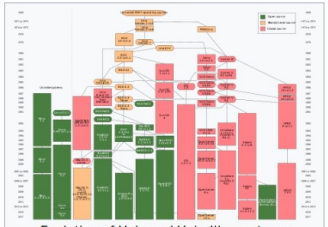
Software



Languages with the most number of Software labels on Wikidata

Operating Systems

Unix



Evolution of Unix and Unix-like systems

Developer	Ken Thompson, Dennis Ritchie, Brian Kernighan, Douglas McIlroy, and Joe Ossanna at Bell Labs
Written in	C and assembly language
OS family	Unix
Working state	Current
Source model	Historically closed-source, while some Unix projects (including BSD family and illumos) are open-source
Initial release	Development started in 1969; 49 years ago First manual published internally in November 1971[1] Announced outside Bell Labs in October 1973[2]
Available in	English
Kernel type	Varies; monolithic, microkernel, hybrid

Linux

Tux the penguin, mascot of Linux[1]

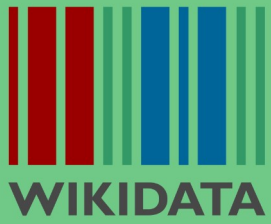
Developer	Community Linus Torvalds
Written in	Primarily C and assembly
OS family	Unix-like
Working state	Current
Source model	Mainly open-source, proprietary software is also available.
Initial release	September 17, 1991; 27 years ago
Marketing target	Personal computers, mobile devices, embedded devices, servers, mainframes, supercomputers
Available in	Multilingual

English Wikipedia Infoboxes of Operating Systems


WIKIDATA

Digital Preservation

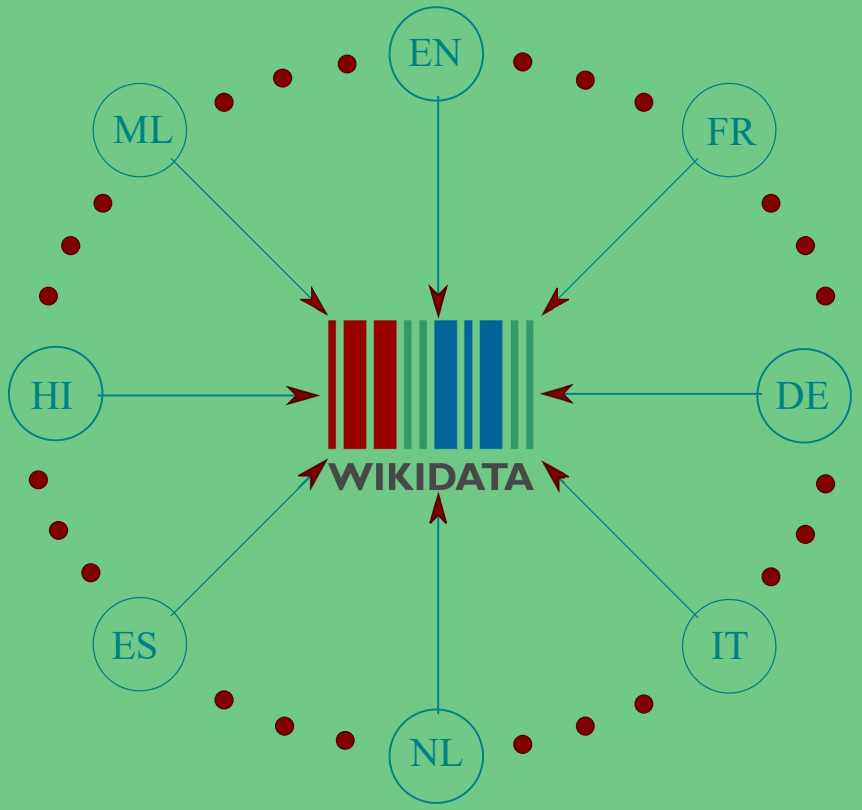
- Digital Preservation
  - OPF
  - Software Heritage
  - EaaSI



Wikidata

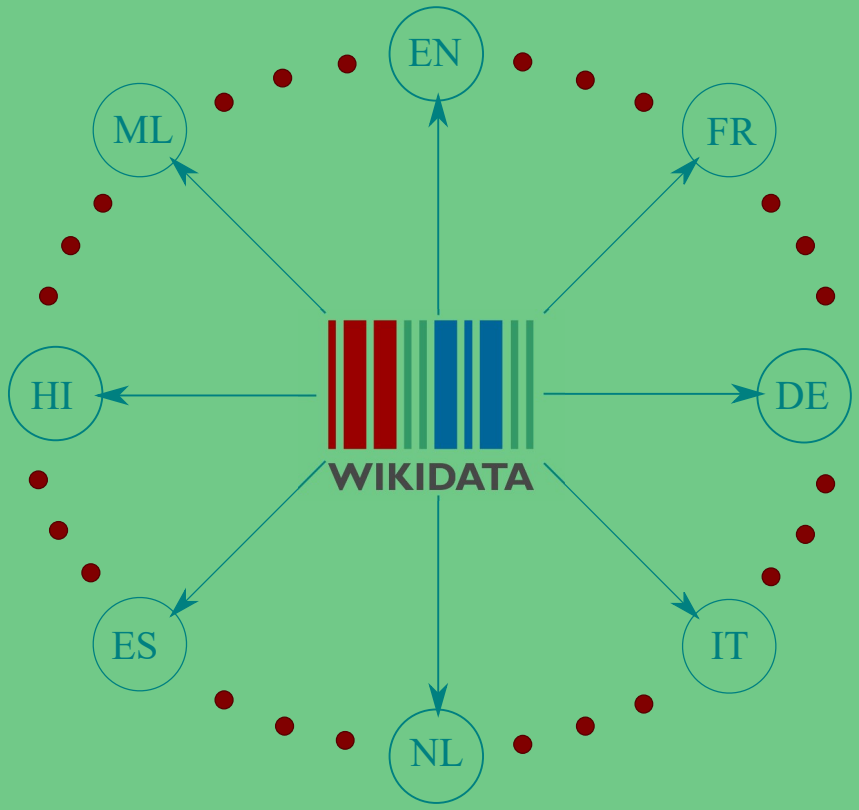
- Wikidata
  - Started in 2012
  - is free, open, linked, structured, collaborative and multilingual knowledge base
  - From multi-(sub)domain multilingual Wikipedia sites to a single-domain multilingual website
  - Collaborative Multilingual Multi-domain Ontology development

Wikipedia to Wikdiata



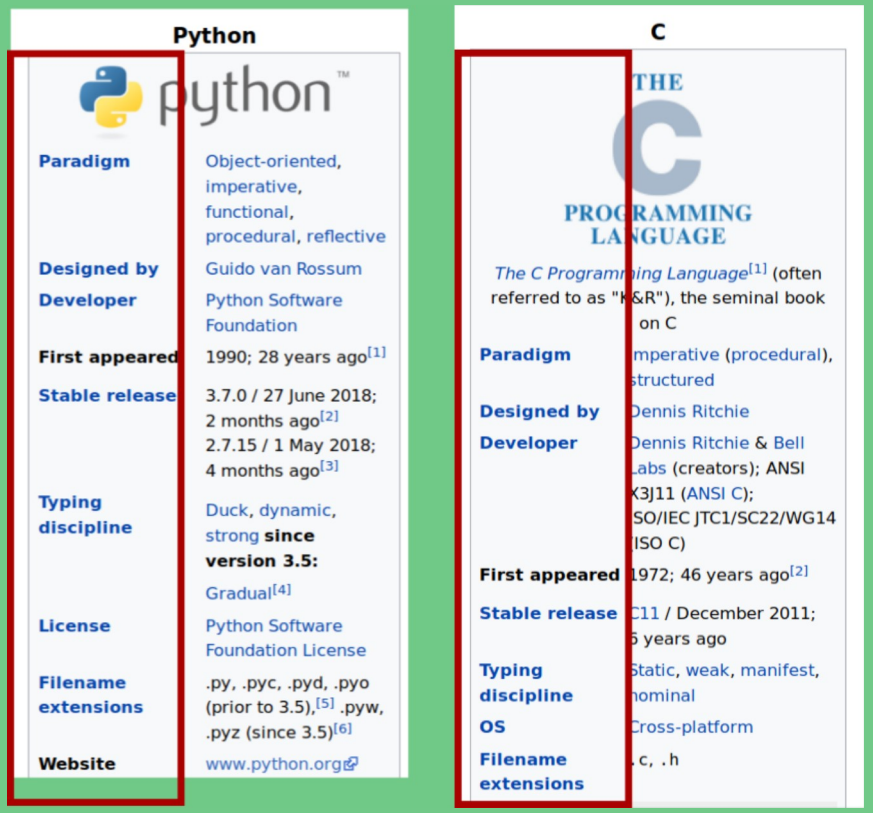
Importing structured data from Wikipedia Infoboxes to Wikidata

Wikdiata to Wikipedia



Exporting data from Wikidata to multiple multilingual Wikipedia articles

Wikipedia Infobox Properties



Python

Paradigm	Object-oriented, imperative, functional, procedural, reflective
Designed by	Guido van Rossum
Developer	Python Software Foundation
First appeared	1990; 28 years ago[1]
Stable release	3.7.0 / 27 June 2018; 2 months ago[2] 2.7.15 / 1 May 2018; 4 months ago[3]
Typing discipline	Duck, dynamic, strong **since version 3.5:** Gradual[4]
License	Python Software Foundation License
Filename extensions	.py, .pyc, .pyd, .pyo (prior to 3.5),[5] .pyw, .pyz (since 3.5)[6]
Website	www.python.org

C

	The C Programming Language[1] (often referred to as "K&R"), the seminal book on C
Paradigm	Imperative (procedural), structured
Designed by	Dennis Ritchie
Developer	Dennis Ritchie & Bell Labs (creators); ANSI X3J11 (ANSI C); ISO/IEC JTC1/SC22/WG14 (ISO C)
First appeared	1972; 46 years ago[2]
Stable release	C11 / December 2011; 6 years ago
Typing discipline	Static, weak, manifest, nominal
OS	Cross-platform
Filename extensions	.c, .h

Existing English Wikipedia Infobox Properties of Programming Languages

Wikidata

Python (Q28865)

general-purpose, high-level programming language

edit

Python language | Python programming language

In more languages Configure

Language	Label	Description	Also known as
English	Python	general-purpose, high-level programming language	Python language Python programming langu...
French	Python	langage de programmation objet, multi-paradigme et multi-plateformes	langage Python langage de programmation ...
Spanish	Python	lenguaje de programación de alto nivel	Python 3.0 Código python Lenguaje python Piton Pitón .py Codigo python Lenguaje de programacion ... URBI Lenguaje de programación ... Python 3 0
German	Python	Programmiersprache	Programmiersprache Python

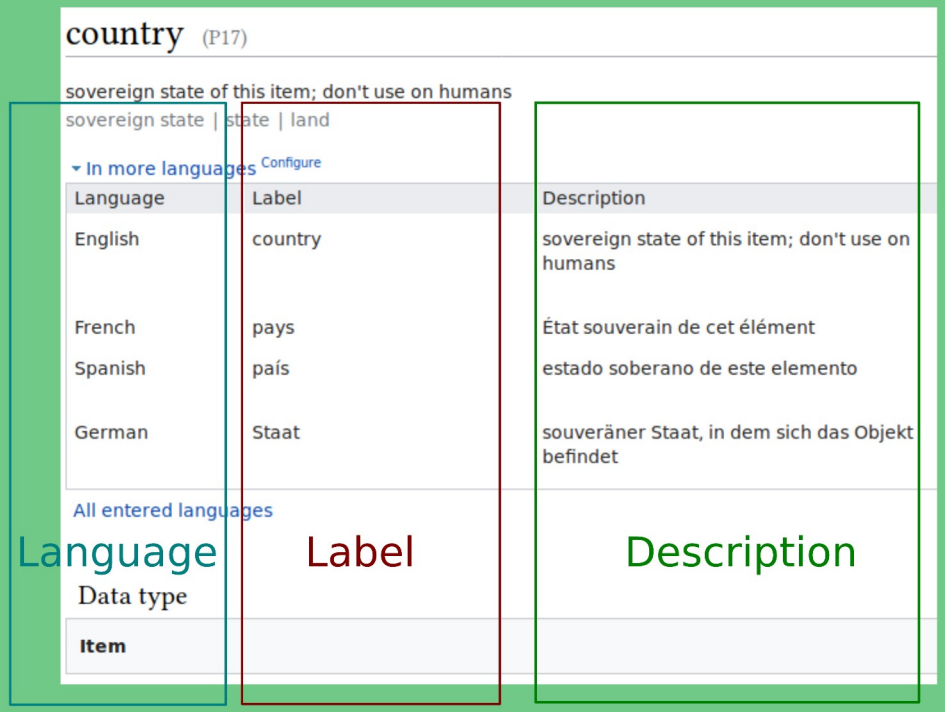
All entered languages

Wikidata entry of Python Programming Language (labels)

Wikidata Properties

Wikidata entry of Python Programming Language (property values)

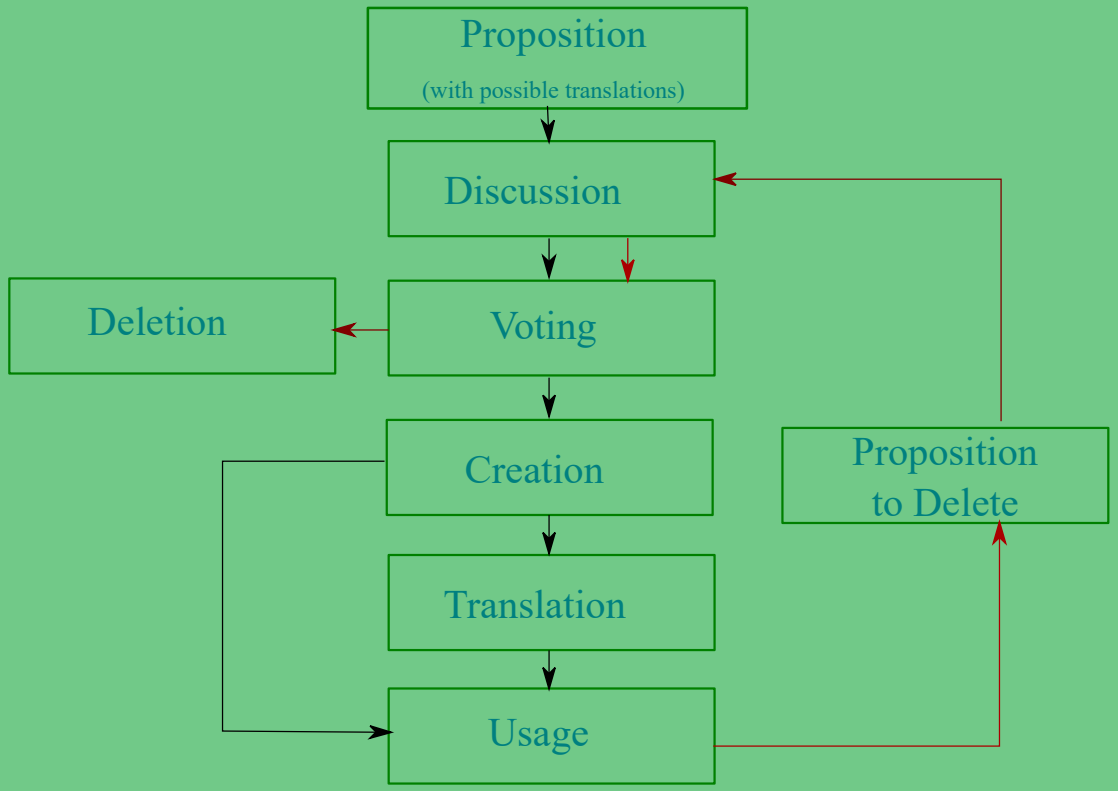
Wikidata Properties



https://www.wikidata.org/wiki/Property:P17

Example of Wikidata Property

Wikidata Properties



Property Creation on Wikidata

Wikidata Projects

Wikidata:WikiProject Informatics/Programming Language

< Wikidata:WikiProject Informatics

The aim of of the **WikiProject Informatics/Programming Language** is to map properties of Infobox programming language to existing properties of Wikidata and t
programming languages can be found here.

Contents [hide]

1 Key Properties
2 Additional Properties
3 External Identifiers
4 Mapping
5 Participants
6 See also

Key Properties [edit]

Title	ID	Data type	Description
image	P18	Commons media file	illustration: image of relevant illustration of the subject; if available, use more specific properties (sample: coat of arms image, locator map, flag image, signature image, logo image, collage image); only images which exist on Wikimedia Commons are acceptable
instance of	P31	Item	instance of and set membership: that class of which this subject is a particular example and member (subject typically an individual member with a proper name label); different from P279; using this property as a qualifier is deprecated—use P2868 or P3831 instead
programming paradigm	P3966	Item	programming paradigm: programming paradigm in which a programming language is classified

Example Wikidata WikiProject

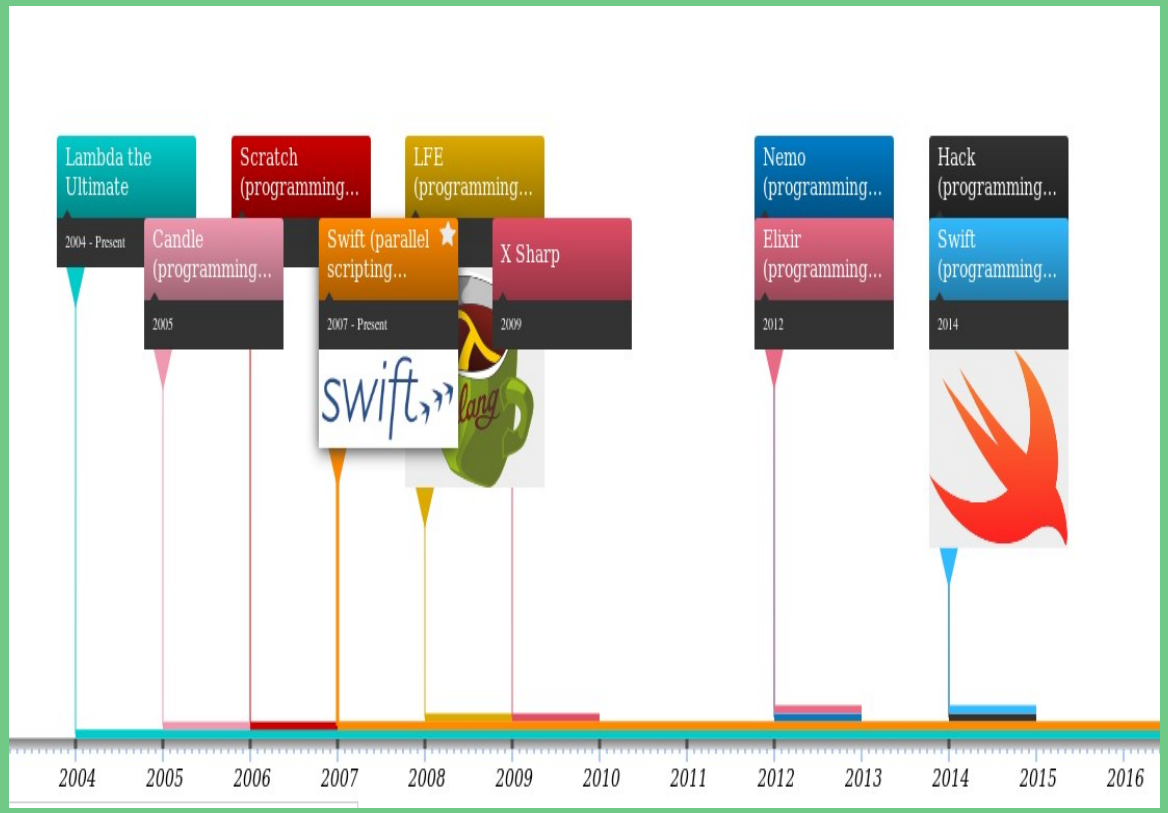
Wikidata Projects

Properties [edit]

Title	ID	Data type	Description
Arch Linux package	P3454	External identifier	name of the official Arch Linux package
based on	P144	Item	based on: the work(s) used as the basis for subject item
based on with qualifier removed feature	P756	Item	software feature: which feature was removed by this version of a product item
binding of software library	P1372	Item	language binding: software library in another programming language provided by the subject software binding
bug tracking system	P1401	URL	bug tracking system: bug report page where bugs, issues, and feature requests are filed for a particular software
creator	P170	Item	creator: maker of this creative work or other object (where no more specific property exists)
Debian stable package	P3442	External identifier	deb: name of the official Debian stable package
depends on software	P1547	Item	subject software depends on object software
developer	P178	Item	organisation or person that developed the item
developer with qualifier instance of	P31	Item	instance of and set membership: when the business model of the developer (P178) of a software is different than the business model of the company considered as a whole, it should be set to an instance of (P31) subclass of (P279) business model (Q815823)
DistroWatch ID	P3112	External identifier	identifier for an operating system at DistroWatch.com

Example Wikidata WikiProject and Property Suggestions

Tools: Histropedia



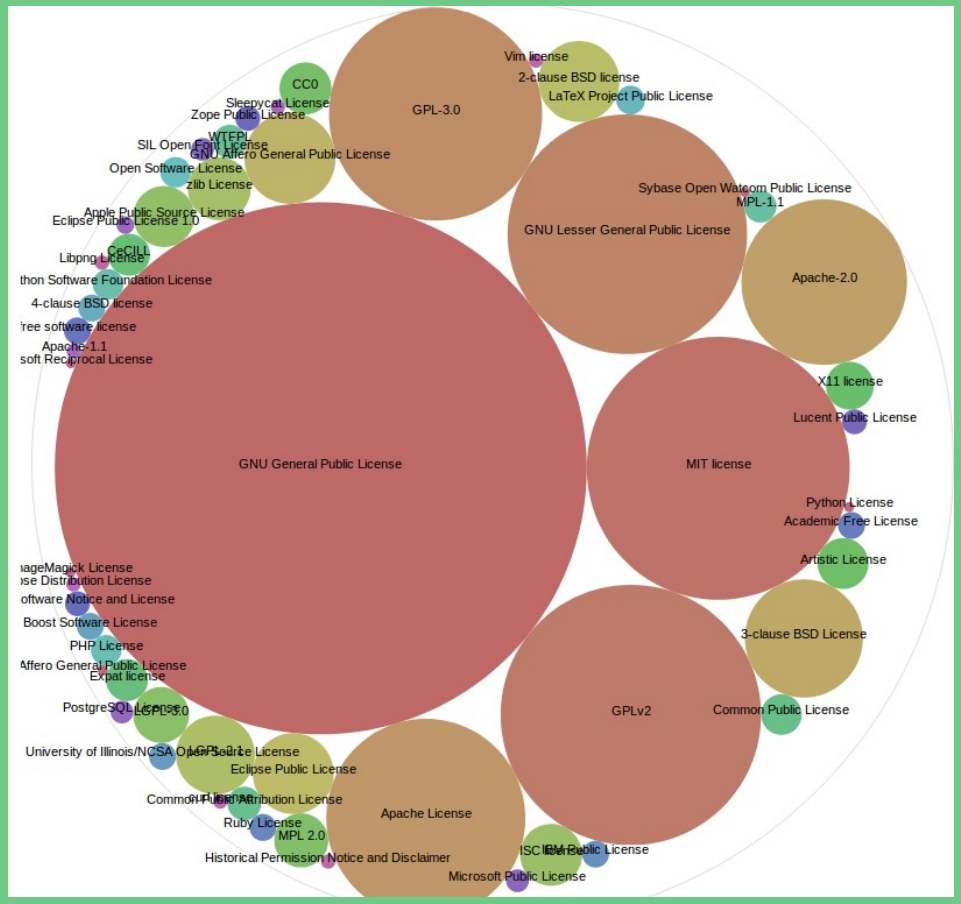
Timeline of Programming Languages

http://histropedia.com/timeline/d98rtpg9bg0t/Programming-languages

Status of software data

- Wikidata
  - 85,000
  - destop applications
  - research software
  - FLOSS

Licenses approved by the Free Software Foundation



Licenses approved by the Free Software Foundation by count of software titles available under each

UNIX utilities

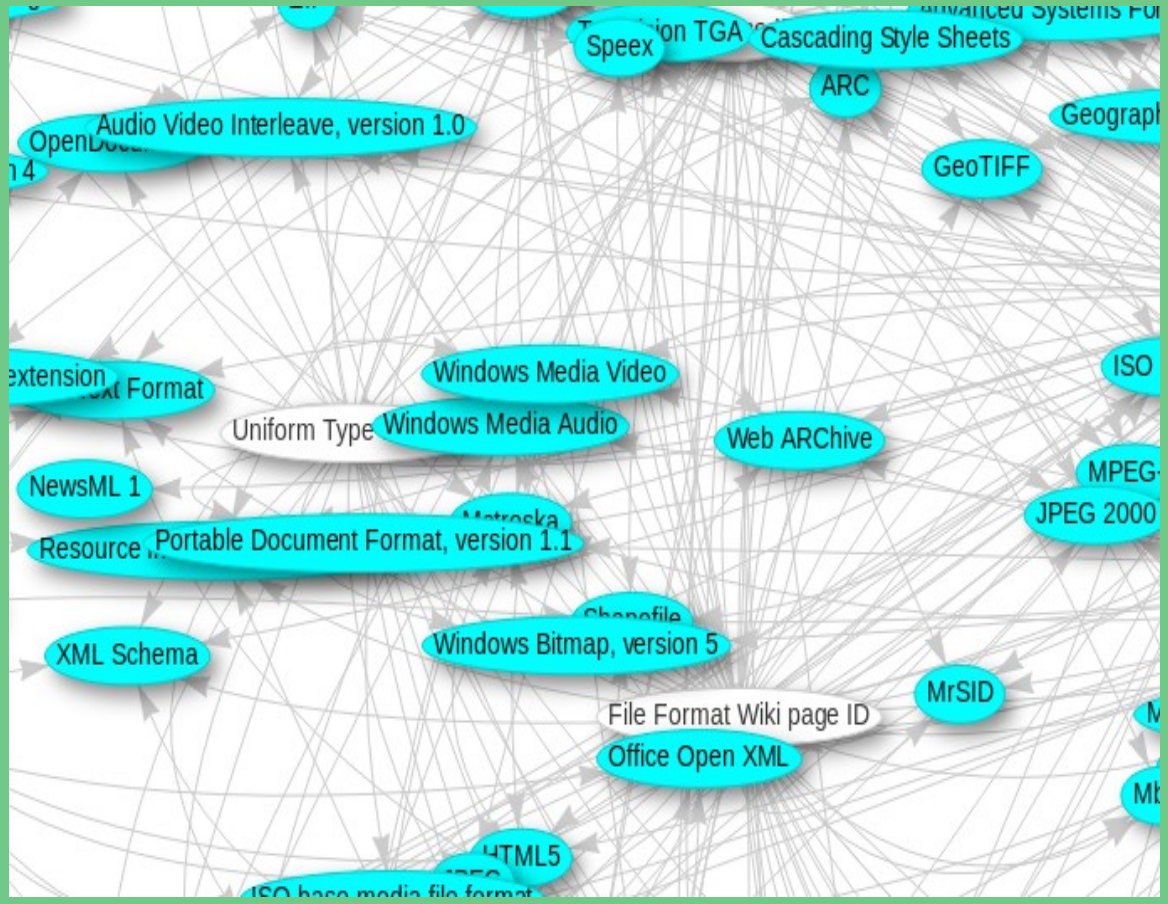
item	itemLabel	LCNAF	GND
wd:Q283302	grep		7692411-7
wd:Q300867	make	n92090404	4334388-0
wd:Q1047966	UUCP	n91047280	4303684-3
wd:Q305932	yacc		4293012-1
wd:Q305876	sed		4262086-7
wd:Q2141501	Source Code Control System		4247001-8
wd:Q213970	AWK	sh87003812	4242961-4
wd:Q214743	vi		4195682-5

Some unix utilities have their own identifiers in the LoC Name Authority File or in the GND

Deutsches Forschungsnetz

Software developed by members of Deutsches Forschungsnetz

File format items



File format items that have a LoC FDD identifier, along with all other identifiers

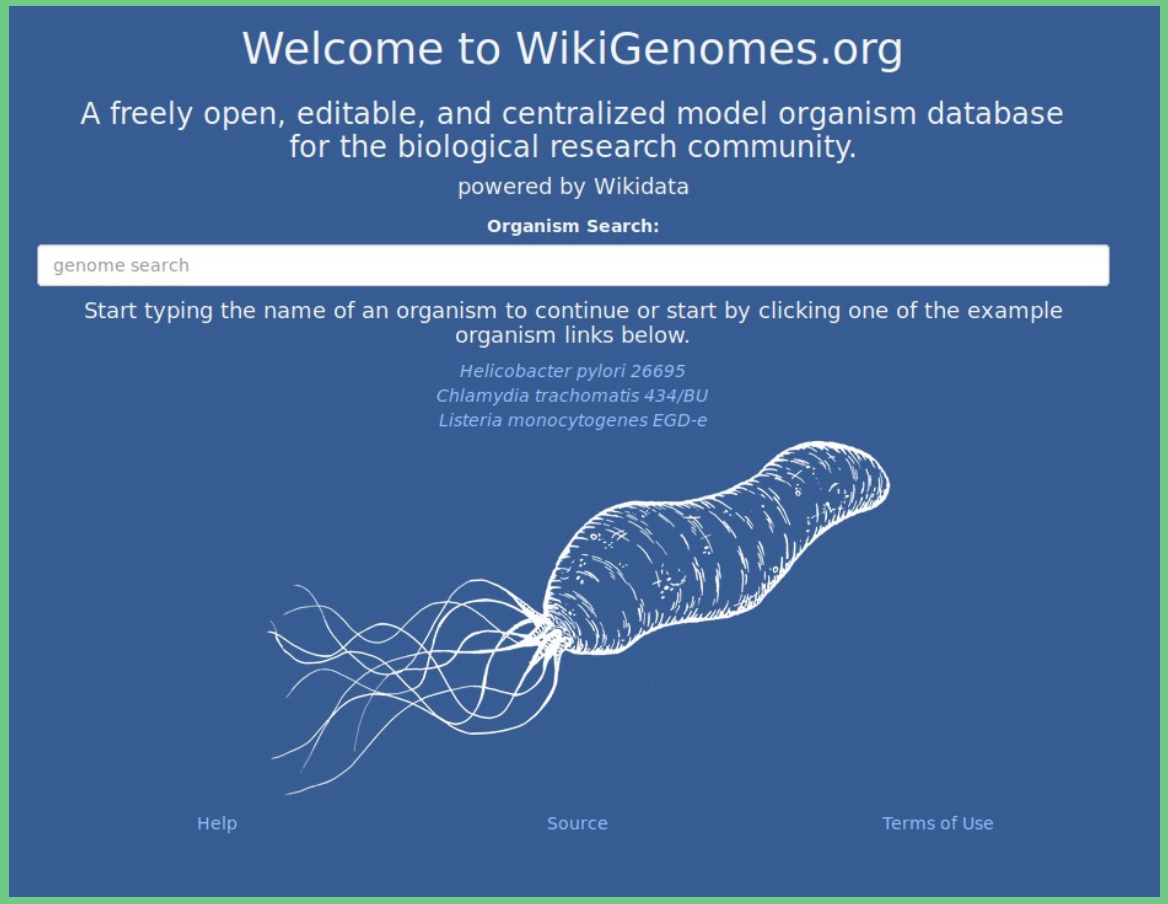
Wikidata for Digital Preservation

- Wikidata
  - Inspired by WikiGenomes
  - Streamlined interface
  - Property checklists tailored to digital preservation
  - Specialty searches (PUID, mimetype)

Wikidata for Digital Preservation

- Development Team
  - Kenneth Seals-Nutt: software engineer
  - Katherine Thornton: Data curation, data models, SPARQL queries
  - Carl Wilson: technical mentor
  - Euan Cochrane: digital preservation program of work

WikiGenomes



wikigenomes.org

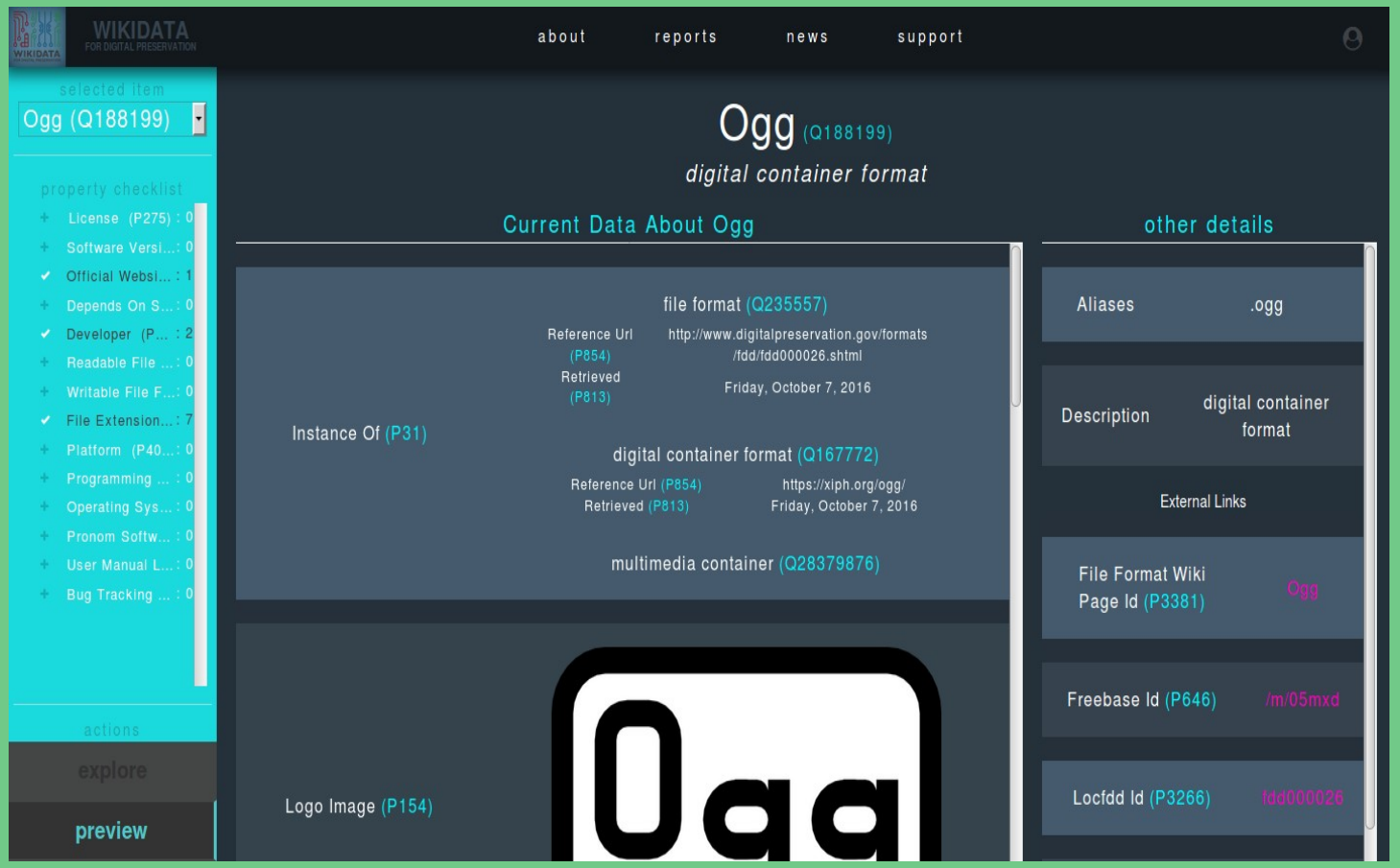
Role of Portals

- About **5,000** properties in Wikidata
- Data models are not pre-defined
- Portal has a domain-specific property checklist

Technologies

- Python
- Flask
- SPARQL
- Wikidata Integrator
- MediaWiki API

WikiDP.org



Screenshot of search results in the WikiDP portal

WDProp

1. WDProp:
   - Collaborative Multilingual Multi-domain Ontology development: is it possible to achieve a truly multilingual experience?
2. Goals:
   - Understanding Wikidata property proposal, creation and translation
   - Available templates and their usage
   - Providing real-time statistics to (multilingual) contributors

WDProp

Information on Wikidata Properties

WDProp

- WDProp
  - Get real-time translation statistics
  - Navigate supported languages, properties, datatypes, classes
  - Compare translation statistics
  - Find available properties for an entity
  - Uses Wikidata SPARQL endpoints and Mediawiki API
- URL
  - https://tools.wmflabs.org/wdprop

Conclusion

- Digital Heritage
  - Wikidata: Multilingual, Structured Knowledge Base
  - Need for Digital Preservation
  - Digital Preservation on Wikidata
  - Community participation: Property proposition, translation and item description
  - Tools using SPARQL endpoints and/or MediaWiki API

Tools and Projects

- Tools
  - Wikidata SPARQL query endpoint
  - MediaWiki API
  - Wikidata Integrator
  - Histropedia
  - wdtaxonomy
  - WDProp
  - WikiDP
  - WikiDP Portal (Github)

Tools and Projects

- WikiProjects
  - WikiProjects
  - WikiProject Informatics
  - WikiProject Informatics/Programming Language
  - WikiProject Informatics/Software/Properties
  - WikiProject Informatics/Operating System

References

1. Kaffee, L. A., Piscopo, A., Vougiouklis, P., Simperl, E., Carr, L., & Pintscher, L. (2017, August). A glimpse into Babel: an analysis of multilinguality in Wikidata. In Proceedings of the 13th International Symposium on Open Collaboration (p. 14). ACM.
2. Müller-Birn, C., Karran, B., Lehmann, J., & Luczak-Rösch, M. (2015, August). Peer-production system or collaborative ontology engineering effort: What is Wikidata?. In Proceedings of the 11th International Symposium on Open Collaboration (p. 20). ACM.
3. Samuel, J. (2017) Collaborative Approach to Developing a Multilingual Ontology: A Case Study of Wikidata. In : Research Conference on Metadata and Semantics Research. Springer, Cham, 2017. p. 167-172.
4. Samuel, J. (2018). Towards Understanding and Improving Multilingual Collaborative Ontology Development in Wikidata. In: WikiWorkshop 2018
5. Thornton, K., Cochrane E., Ledoux T. (2017). Modeling the Domain of Digital Preservation in Wikidata . In: iPRES 2017

Thank you

Questions?

SPARQL Query

Programming paradigms with the count of programming languages

```
SELECT ?paradigmLabel (count(?prog) as ?count)
{
  ?prog wdt:P31 wd:Q9143;
        wdt:P3966 ?paradigm.
  SERVICE wikibase:label { bd:serviceParam
          wikibase:language "[AUTO_LANGUAGE],en". }
}
GROUP by ?paradigmLabel
HAVING (?count>1)
```

SPARQL Query

Programming languages with the count of programming paradigm

```
SELECT ?progLabel (count(?paradigm) as ?count)
{
  ?prog wdt:P31 wd:Q9143;
        wdt:P3966 ?paradigm.
  SERVICE wikibase:label { bd:serviceParam
          wikibase:language "[AUTO_LANGUAGE],en". }
}
GROUP by ?progLabel
HAVING (?count>2)
```

SPARQL Query

Programming languages with the count of multilingual labels

```
SELECT ?languageLabel (count(?label) as ?count) {
  {
    SELECT DISTINCT ?languageLabel ?label (lang(?label) as ?langLabel) {
      ?language wdt:P31/wdt:P279* wd:Q9143;
                rdfs:label ?label.
      SERVICE wikibase:label { bd:serviceParam
          wikibase:language "[AUTO_LANGUAGE],en". }
      }
  }

}
GROUP by ?languageLabel
HAVING (?count > 50)
ORDER by DESC(?count)
```

SPARQL Query

Software with the count of multilingual labels

```
SELECT ?softwareLabel (count(?label) as ?count) {
  {
    SELECT DISTINCT ?softwareLabel ?label (lang(?label) as ?langLabel) {
      ?software wdt:P31/wdt:P279 wd:Q7397;
                rdfs:label ?label.
      SERVICE wikibase:label { bd:serviceParam
          wikibase:language "[AUTO_LANGUAGE],en". }
      }
  }

}
GROUP by ?softwareLabel
HAVING (?count > 40)
ORDER by DESC(?count)
```

SPARQL Query

Programming language with the count of multilingual labels

```
SELECT ?langLabel (count(?language) as ?count) {
  {
    SELECT DISTINCT (lang(?label) as ?langLabel) ?language {
      ?language wdt:P31/wdt:P279* wd:Q9143;
                rdfs:label ?label.
      }
  }

}
GROUP by ?langLabel
ORDER by DESC(?count)
```

SPARQL Query

Language with the count of software labels

```
SELECT ?langLabel (count(?software) as ?count) {
  {
    SELECT DISTINCT (lang(?label) as ?langLabel) ?software {
      ?software wdt:P31/wdt:P279* wd:Q7397;
                rdfs:label ?label.
      }
  }

}
GROUP by ?langLabel
ORDER by DESC(?count)
```

SPARQL Query

Languages with the count of Wikipedia articles on programming languages

```
SELECT DISTINCT ?languageLabel ?sitelinks {
      ?language wdt:P31/wdt:P279* wd:Q9143;
                wikibase:sitelinks ?sitelinks.
       FILTER(?sitelinks > 20)
       SERVICE wikibase:label { bd:serviceParam
          wikibase:language "[AUTO_LANGUAGE],en". }
}
ORDER by DESC(?sitelinks)
```

SPARQL Query

Languages with the count of Wikipedia articles on software

```
SELECT DISTINCT ?softwareLabel ?sitelinks {
      ?software wdt:P31/wdt:P279* wd:Q7397;
                wikibase:sitelinks ?sitelinks.
       FILTER(?sitelinks > 100)
       SERVICE wikibase:label { bd:serviceParam
          wikibase:language "[AUTO_LANGUAGE],en". }
}
ORDER by DESC(?sitelinks)
```

SPARQL Query

Languages with the count of Wikipedia articles on programming languages

```
SELECT ?lang (count(?progLanguage) as ?count) {
  {
    SELECT DISTINCT ?progLanguage ?lang {
      ?progLanguage wdt:P31/wdt:P279* wd:Q9143.
      [] schema:about ?progLanguage;
         schema:inLanguage ?lang.
     }
  }
}
GROUP BY ?lang
ORDER BY DESC(?count)
```

SPARQL Query

Languages with the count of Wikipedia articles on Software

```
SELECT ?lang (count(?software) as ?count) {
  {
    SELECT DISTINCT ?software ?lang {
      ?software wdt:P31/wdt:P279* wd:Q7397.
      [] schema:about ?software;
         schema:inLanguage ?lang.
     }
  }
}
GROUP BY ?lang
ORDER BY DESC(?count)
```

SPARQL Query

Licenses approved by the Free Software Foundation by count of software titles available under each

```
SELECT ?item ?itemLabel (COUNT(DISTINCT ?software) AS ?count) WHERE {
  ?software (wdt:P31/wdt:P279*) wd:Q7397.
  ?software wdt:P275 ?item.
  ?item wdt:P790 wd:Q48413.
  SERVICE wikibase:label { bd:serviceParam
          wikibase:language "[AUTO_LANGUAGE],en". }
}
GROUP BY ?item ?itemLabel
ORDER BY DESC(?count)
```

SPARQL Query

UNIX utilities with identifiers in the LoC Name Authority File or in the GND

```
SELECT ?item ?itemLabel ?LCNAF ?GND
WHERE
{
  ?item wdt:P31 wd:Q18343316.
  OPTIONAL {?item wdt:P244 ?LCNAF}.
  OPTIONAL {?item wdt:P227 ?GND}.
  SERVICE wikibase:label { bd:serviceParam
          wikibase:language "[AUTO_LANGUAGE],en". }
}
```

SPARQL Query

Software developed by members of Deutsches Forschungsnetz

```
SELECT ?member ?memberLabel ?software ?softwareLabel WHERE {
  ?member wdt:P463 wd:Q2514863.
  ?software wdt:P178 ?member.
  SERVICE wikibase:label { bd:serviceParam
          wikibase:language "[AUTO_LANGUAGE],en". }
}
```