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
<table>
<thead>
<tr>
<th>Programming Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Python</strong></td>
</tr>
<tr>
<td><strong>Java</strong></td>
</tr>
<tr>
<td><strong>C</strong></td>
</tr>
</tbody>
</table>

**Paradigm**
- Python: Object-oriented, imperative, functional, procedural, reflective
- Java: Multi-paradigm: object-oriented (class-based), structured, imperative, generic, reflective, concurrent
- C: Imperative (procedural), structured

**Designed by**
- Python: Guido van Rossum
- Java: Sun Microsystems (now owned by Oracle Corporation)
- C: Dennis Ritchie

**Developer**
- Python: Python Software Foundation
- Java: Sun Microsystems (now owned by Oracle Corporation)
- C: Sun Microsystems

**First appeared**
- Python: 1990; 28 years ago
- Java: May 23, 1995; 23 years ago
- C: 1972; 46 years ago

**Stable release**
- Python: 3.7.0 / 27 June 2018; 2 months ago
- Java: 2017; 2 years ago
- C: 2011; 5 years ago

**Typing discipline**
- Python: Duck, dynamic, strong since version 3.5
- Java: Static, strong, safe, nominative, manifest
- C: Static, weak, manifest, nominal

**License**
- Python: Python Software Foundation License
- Java: GNU General Public License, Java Community Process
- C: X11 (ANSI C); ISO/IEC 9899:1990 (ISO C)

**Filename extensions**
- Python: .py, .pyc, .pyd (prior to 3.5), .py3, .py (since 3.5)
- Java: .class, .jar
- C: .c, .h
Programming Languages with the most multilingual labels
Programming Language Paradigms
Programming Languages with the most number of different paradigms
Programming Languages with most number of multilingual Wikipedia articles
Wikipedia languages with the most number of articles on programming languages
Languages with the most number of labels of programming languages
### English Wikipedia Infoboxes of Software

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

**Google Chrome**
- **Developer(s)**: Google LLC
- **Initial release**: September 2, 2008; 10 years ago
- **Stable release(s)**: 69.0.3497.100 / September 17, 2018; 2 days ago[^1]
- **Windows**: 69.0.3497.100 / September 17, 2018; 2 days ago[^2]
- **macOS, Linux**: 69.0.3497.100 / September 17, 2018; 2 days ago[^2]
- **Android**: 69.0.3497.91 / September 11, 2016; 8 days ago[^3]

[^1]: [Source](https://en.wikipedia.org/wiki/Mozilla_Firefox)
[^2]: [Source](https://en.wikipedia.org/wiki/Google_Chrome)
Software with the most number of labels on Wikidata
Software with the most number of articles on Wikipedia
Languages with the most number of articles on Wikipedia
Languages with the most number of Software labels on Wikidata
### Operating Systems

<table>
<thead>
<tr>
<th>Developer</th>
<th>Unix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Writer in</td>
<td>C and assembly language</td>
</tr>
<tr>
<td>OS family</td>
<td>Unix</td>
</tr>
<tr>
<td>Working state</td>
<td>Current</td>
</tr>
<tr>
<td>Source model</td>
<td>Historically closed-source, while some Unix projects (including BSD family and others) are open-source</td>
</tr>
<tr>
<td>Available in</td>
<td>English</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Developer</th>
<th>Linux</th>
</tr>
</thead>
<tbody>
<tr>
<td>Written in</td>
<td>Community Linus Torvalds</td>
</tr>
<tr>
<td>OS family</td>
<td>Unix-like</td>
</tr>
<tr>
<td>Working state</td>
<td>Current</td>
</tr>
<tr>
<td>Source model</td>
<td>Mainly open-source, proprietary software is also available.</td>
</tr>
<tr>
<td>Initial release</td>
<td>September 17, 1991, 27 years ago</td>
</tr>
<tr>
<td>Marketing target</td>
<td>Personal computers, mobile devices, embedded devices, servers, mainframes, supercomputers</td>
</tr>
</tbody>
</table>

[1] Available at: [Linux info](https://en.wikipedia.org/wiki/Linux)
Digital Preservation

- Digital Preservation
  - OPF
  - Software Heritage
  - EaaSI
• 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
Importing structured data from Wikipedia Infoboxes to Wikidata
Exporting data from Wikidata to multiple multilingual Wikipedia articles
Wikipedia Infobox Properties

Existing English Wikipedia Infobox Properties of Programming Languages

Digital Preservation | John Samuel, Katherine Thornton
## Wikidata entry of Python Programming Language (labels)

<table>
<thead>
<tr>
<th>Language</th>
<th>Label</th>
<th>Description</th>
<th>Also known as</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Python</td>
<td>general-purpose, high-level programming language</td>
<td>Python language</td>
</tr>
<tr>
<td>French</td>
<td>Python</td>
<td>langage de programmation objet, multi-paradigme et multi-plateformes</td>
<td>langage Python, langage de programmation</td>
</tr>
<tr>
<td>Spanish</td>
<td>Python</td>
<td>lenguaje de programación de alto nivel</td>
<td>Python 3.0, Codigo python, lenguaje python, Piton, .py</td>
</tr>
<tr>
<td>German</td>
<td>Python</td>
<td>Programmiersprache</td>
<td>Programmiersprache Python</td>
</tr>
</tbody>
</table>
## Wikidata Properties

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>inception</td>
<td>20 February 1991</td>
</tr>
<tr>
<td>named after</td>
<td>Monty Python</td>
</tr>
<tr>
<td>movement</td>
<td>free software movement</td>
</tr>
<tr>
<td>publisher</td>
<td>Core Python Programming</td>
</tr>
<tr>
<td>influenced by</td>
<td>ALGOL 68</td>
</tr>
</tbody>
</table>

Wikidata entry of Python Programming Language (property values)
### Wikidata Properties

<table>
<thead>
<tr>
<th>Language</th>
<th>Label</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>country</td>
<td>sovereign state of this item; don't use on humans</td>
</tr>
<tr>
<td>French</td>
<td>pays</td>
<td>État souverain de cet élément</td>
</tr>
<tr>
<td>Spanish</td>
<td>pais</td>
<td>estado soberano de este elemento</td>
</tr>
<tr>
<td>German</td>
<td>Staat</td>
<td>souveräner Staat, in dem sich das Objekt befindet</td>
</tr>
</tbody>
</table>

https://www.wikidata.org/wiki/Property:P17  
Example of Wikidata Property
## Wikidata:WikiProject Informatics/Programming Language

The aim of the [WikiProject Informatics/Programming Language](https://wikidata.org/wiki/Wikidata:WikiProject_Informatics/Programming_Language) is to map properties of [infobox programming language](https://wikidata.org/wiki/Q150552) to existing properties of Wikidata and Wikisource. Information on programming languages can be found [here](https://commons.wikimedia.org/wiki/MediaWiki:WikiProject_Informatics/Programming_Language).

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

### Key Properties

<table>
<thead>
<tr>
<th>Title</th>
<th>ID</th>
<th>Data type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>image</td>
<td>P18</td>
<td>Commons media file</td>
<td>Illustration: image of relevant illustration of the subject; if available, use more specific properties (example: coat of arms image, locator map, flag image, signature image, logo image, collage image); only images which exist on Wikimedia Commons are acceptable.</td>
</tr>
<tr>
<td>instance of</td>
<td>P31</td>
<td>Item</td>
<td>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 P3833 instead.</td>
</tr>
<tr>
<td>programming paradigm</td>
<td>P3964</td>
<td>Item</td>
<td>Programming paradigm; programming paradigm in which a programming language is classified.</td>
</tr>
</tbody>
</table>
### Wikidata Projects

**Properties**

<table>
<thead>
<tr>
<th>Title</th>
<th>ID</th>
<th>Data type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arch Linux package</td>
<td>P3454</td>
<td>External identifier</td>
<td>name of the official Arch Linux package</td>
</tr>
<tr>
<td>based on</td>
<td>P144</td>
<td>Item</td>
<td>based on: the work(s) used as the basis for subject item</td>
</tr>
<tr>
<td>based on with qualifier</td>
<td>P756</td>
<td>Item</td>
<td>removed feature: which feature was removed by this version of a product</td>
</tr>
<tr>
<td>binding of software library</td>
<td>P1372</td>
<td>Item</td>
<td>language binding: software library in another programming language provided</td>
</tr>
<tr>
<td>bug tracking system</td>
<td>P1401</td>
<td>URL</td>
<td>bug tracking system: bug report page where bugs, issues, and feature requests filed for a particular software</td>
</tr>
<tr>
<td>creator</td>
<td>P170</td>
<td>Item</td>
<td>creator: maker of this creative work or other object (where no more specific property exists)</td>
</tr>
<tr>
<td>Debian stable package</td>
<td>P3442</td>
<td>External identifier</td>
<td>deb: name of the official Debian stable package</td>
</tr>
<tr>
<td>depends on software</td>
<td>P1547</td>
<td>Item</td>
<td>subject software depends on object software</td>
</tr>
<tr>
<td>developer</td>
<td>P178</td>
<td>Item</td>
<td>organisation or person that developed the item</td>
</tr>
<tr>
<td>developer with qualifier</td>
<td>P31</td>
<td>Item</td>
<td>instance of and set membership: when the business model of the developer (P318) 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 (Q815923)</td>
</tr>
<tr>
<td>DistroWatch ID</td>
<td>P3112</td>
<td>External identifier</td>
<td>identifier for an operating system at DistroWatch.com</td>
</tr>
</tbody>
</table>
Tools: Histropedia

Timeline of Programming Languages

http://histropedia.com/timeline/d98rtpg9bg0t/Programming-languages
Status of software data

- Wikidata
  - 85,000
  - desktop applications
  - research software
  - FLOSS
Licenses approved by the Free Software Foundation by count of software titles available under each license.
### UNIX utilities

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

<table>
<thead>
<tr>
<th>item</th>
<th>itemLabel</th>
<th>LCNAF</th>
<th>GND</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q wd:Q283302</td>
<td>grep</td>
<td></td>
<td>7692411-7</td>
</tr>
<tr>
<td>Q wd:Q300867</td>
<td>make</td>
<td>n92090404</td>
<td>4334388-0</td>
</tr>
<tr>
<td>Q wd:Q1047966</td>
<td>UUCP</td>
<td>n91047280</td>
<td>4303684-3</td>
</tr>
<tr>
<td>Q wd:Q305832</td>
<td>yacc</td>
<td></td>
<td>4293012-1</td>
</tr>
<tr>
<td>Q wd:Q305876</td>
<td>sed</td>
<td></td>
<td>4262086-7</td>
</tr>
<tr>
<td>Q wd:Q2141501</td>
<td>Source Code Control System</td>
<td></td>
<td>4247001-8</td>
</tr>
<tr>
<td>Q wd:Q213970</td>
<td>AWK</td>
<td>sh87003812</td>
<td>4242961-4</td>
</tr>
<tr>
<td>Q wd:Q214743</td>
<td>vi</td>
<td></td>
<td>4195682-5</td>
</tr>
</tbody>
</table>

Digital Preservation | John Samuel, Katherine Thornton
Deutsches Forschungsnetz

Software developed by members of Deutsches Forschungsnetz
File format items that have a LoC FDD identifier, along with all other identifiers
• Wikidata
  ○ Inspired by WikiGenomes
  ○ Streamlined interface
  ○ Property checklists tailored to digital preservation
  ○ Specialty searches (PUID, mimetype)
• 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
Welcome to WikiGenomes.org

A freely open, editable, and centralized model organism database for the biological research community.

powered by Wikidata

Organism Search:

Start typing the name of an organism to continue or start by clicking one of the example organism links below.

* Helicobacter pylori 26695
* Chlamydia trachomatis 434/BU
* Listeria monocytogenes EGD-e

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
Screenshot of search results in the WikiDP portal
1. **WDP**
   - 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
Translation statistics by languages

- Supported languages
- Translated Labels
- Translated Descriptions
- Translated Alias
- Overall Translation Statistics
- Missing Translation Statistics

Navigate properties

- Browse all properties
- Browse by datatype

Information on Wikidata Properties
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


Thank you

Questions?
**Programming paradigms with the count of programming languages**

```sparql
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)
```
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)
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)
```
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)
```
Languages with the count of Wikipedia articles on software

```sparql
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)
```
Languages with the count of Wikipedia articles on programming languages

```sparql
SELECT ?lang (count(?progLanguage) as ?count) {
  {
    SELECT DISTINCT ?progLanguage ?lang {
      [] schema:about ?progLanguage;
    }
  }
}
GROUP BY ?lang
ORDER BY DESC(?count)
```
Languages with the count of Wikipedia articles on Software

SELECT ?lang (count(?software) as ?count) {
    {
        SELECT DISTINCT ?software ?lang {
            [] schema:about ?software;
        }
    }
} 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 {
  ?item wdt:P790 wd:Q48413.
  SERVICE wikibase:label {
    bd:serviceParam
    wikibase:language "[AUTO_LANGUAGE],en".
  }
}
GROUP BY ?item ?itemLabel
ORDER BY DESC(?count)
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". } }
Software developed by members of Deutsches Forschungsnetz

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