Metadata Report

Katherine Thornton

January 2, 2018

ABSTRACT

This report describes the current state of entity data in the Wikidata knowledge base for the domain of computing as compared to January, 2018.

1 Data Curation

I have been curating data in the domain of computing. I wrote SPARQL queries to return data about sets of resources we reuse in EaaSI.

Items	Jan 18	Jan 19	net change
Software Items	64,925	82,759	17,834
File Format Items	2,834	3,483	649
File Format Items with PUIDs	777	1299	522
File Format Signatures	167	198	31
Emulators	106	115	9
File Systems	146	158	12
Device Drivers	17	28	11
Plugins	155	184	29
Config Emulated Env	0	1	1
File Format with mediatype	936	952	16
Software by dev w ISNI	6,6454	6,6825	371

1.1 Software Items

The Wikidata community added more than 17,000 software items to the knowledge base in 2018. This is a very encouraging number that reflects the strength of interest in this area. Many of these are video game titles (many video game properties created in the last year).

1.2 File Format Items

PRONOM is the a metadata registry about file formats and software created by the National Archives of the UK. Many of these were new items I created for file formats in PRONOM that were not previously in Wikidata.

1.3 File Format Items with PUIDs

The data curation work I prioritized for this quarter was to create new items for file formats described in PRONOM. There are 1,500 resources in PRONOM, a certain number of PUIDs have been deprecated.

1.4 File Format Signatures

File format signatures are strings of text that can be used for format identification. The majority of file format signatures stored in PRONOM remain to be added to Wikidata.

1.5 Emulators

Due to the focus of our grant work on emulation, I am prioritizing getting emulators described as extensively as we need for our work.

1.6 File Systems

Having accurate metadata about file systems in Wikidata will have broad appeal among people who are reusing data from Wikidata. I am prioritizing contributing accurate information with references in this area.

1.7 Device Drivers

Drivers are not yet well documented across repositories. We will add drivers as we create additional configured environments that require drivers.

1.8 Plugins

This is not a current priority, but I am interested in watching strategies people use to describe these resources.

1.9 Configured Software Environments

We created our first software environment item.

1.10 File Formats with mediatype

I am working on changing the datatype of mediatype to item in Wikidata. Tracking the usage of the current string datatype to be sure we can convert all of these later.

1.11 Software titles by dev w ISNI

Tracking this so we can see the impact of adding NSRL data. Many of the developers in the NSRL manufaturers table also have ISNI IDs. On August 10, 2018 I first ran this query and the count was 6,454, in four months 371 titles were added by the Wikidata community.