

# Metadata Quality Report

Katherine Thornton

Oct 17, 2019

---

## ABSTRACT

This report describes the current state of entity data in the Wikidata knowledge base for the domain of computing as compared to January of 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 2018	Oct 2019	net change
Software Items	64,925	72,635	7,710
File Format Items	2,834	4,222	1,388
File Format Items with PUIDs	777	1,476	699
File Format Signatures	167	216	49
Emulators	106	149	43
File Systems	146	195	49
Device Drivers	17	28	11
Plugins	155	60	-95
Config Emulated Env	0	1	1
File Format with mediatype	936	1090	154
Software by dev w ISNI	6,677	7,198	521

### 1.1 Software Items

Software items are growing steadily.

### 1.2 File Format Items

There have been more than 1,300 new file formats added to Wikidata since Jan 2018.

### 1.3 File Format Items with PUIDs

The data curation work I prioritized over the last year 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. It looks like there was reorganization of how these are modeled in the past 3 months.

## **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 manufacturers table also have ISNI IDs. This can be considered a proxy for software titles likely to be of interest to EaaSI.