Visualizing Wikidata: Using Python to Analyze Identity and Representation in Wikidata about Black Art Exhibitions

jay winkler
ICPSR- University of Michigan

Introduction

For many of us, adding "wiki" to the end of a Google search is like a cheat code for getting a quick answer. I know the structure of a Wikipedia article, and if I want an answer to a simple question, like "where was E.B. Lewis born", just finding it on Wikipedia is the fastest way to get it.

But what about more complex questions? What if I wanted to know, say, who are all of the Black artists who were born in Philadelphia? For that, I can turn to Wikidata. Wikidata is an attempt to convert all of the world's information into a structured, queryable dataset.

As part of the 2021 LEADING Fellowship, the team worked a project enhancing and analyzing the available information on Philadelphia's Black artists. While the project was initially focused on the artists themselves, during the research I became interested in the limitations and challenges of Wikidata as a platform.

There are a number of issues surrounding the way that demographic tagging on Wikidata occurs. As a community resource, Wikidata editors preach extreme caution in applying ethnicity properties to Wikidata items. While some of the reasoning behind this is noble, it causes incompleteness in the catalog and hides how diverse the catalog really is. It also obscures whiteness, which is not represented on Wikidata almost at all.

Through use of the Wikidata SPARQL endpoint and Python tools, jay was able to examine some of these challenges by analyzing the data Wikidata provided. This digital poster presentation will examine that process, as well as looking at what Wikidata editors and other libraries are already doing to attempt to mitigate some of these issues.

Objectives

The goal of the project was to examine the information available on Wikidata about Black Artists, with a specific focus on Philadelphia. Our first goal was to apply the Philadelphia Museum of Art Entity ID tag to a large number of Wikidata entries for artists in PMA's collection. We then moved on to the creation of Wikidata entries for a number of Philadelphia Black Artists, including artists in the Charles H. Blockson Afro-American Collection, as well as artists that had been showcased at Mural Arts Philadelphia. While enhancing these Wikidata entries was a goal in of itself, a secondary goal of this process was to familiarize ourselves with Wikidata's possibilities to form research questions. We then started querying the information in Wikidata to gather biographical information about the artists.

When running our initial SPARQL queries that helped us form our research questions, we typically restricted our searches to African American artists. However, we were also interested in comparing African American artists to Wikidata's total population of artists. We removed that parameter and ran a query that would give information about all artists. It was then that we noticed that the vast majority of artists simply did not have any sort of ethnicity tagging. It was this observation that led me to my research question: **How did the use of ethnicity tagging change on Wikidata over time?**

Methodology

While formulating our research questions I became very interested in Wikidata itself, and the gaps and challenges with inclusivity on the platform. The representation of race on Wikidata is an ongoing challenge, and I hoped to focus my querying and visualisation on that challenge. I started thinking about the actual property for ethnicity, and tracking how its use differed over time. Wikidata is a collaborative platform, and its biases reflect the biases of the editors. Additionally, the ethnicity property, by design, has a high bar for usage. Editors are instructed not to use the ethnicity property unless they are absolutely sure, preferably with a source.

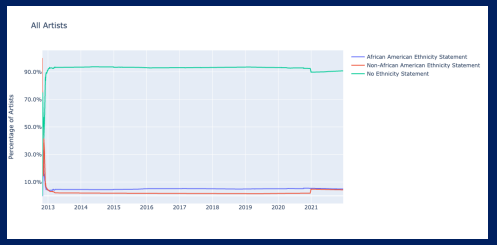
The project team had written a SPARQL query to find all of the African American artists from Philadelphia, and I and the rest of the team added parameters so we could gather a great deal of biographical information for each artist. The code was held in a Google Colab notebook, a shared Jupyter notebook environment. Google Colab allows teams to collaborate on code, and run that code right in the browser. We put that list into a Dataframe, which is a function of Python's Pandas library that can be used to manipulate data and display it as a table.

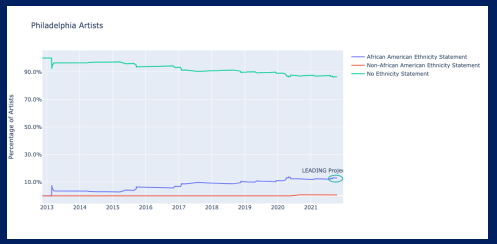
Once I had the Dataframe, my goal was to use Python to iterate through the list and calculate how the percentage of artists that had "African American" as their ethnicity changed overtime. The Wikidata Query Service will not simply return the creation date of an item, which means we needed to turn to Wikidata's API. We started by using SPARQL to gather a list of all artists with American birthplaces. We then iterated through that list using the Wikidata API to gather the creation date of each item. Finally, for each item, we calculated the percentage of Wikidata's collection represented by each ethnic group at the time the given item was added. There were many intermediate steps of data cleaning and wrangling.

The visualizations were created using a Python library called plot.ly. Plot.ly has many tools for gathering data from a Dataframe and charting it, as well as tools to customize those charts.

While our research was focused on the contrast between tagging of Black artists and untagged entries, other ethnic groups were present in the dataset. In the visualizations, other ethnic group tags are combined in a single "Other" category. While most non-African American ethnic groups had too few entries to provide useful data, aggregating the data provides insight on whether or not the ethnic group property is being used at all.

Once I had completed this process on the list of all artists with American birthplaces, I then gathered subsets of the data that covered only artists from a handful of major cities: Philadelphia, New York, Detroit, and Chicago. The goal of this process was to see if there were any differences when broken down by city, as well as to see if I could spot any purposeful efforts, like ours, to improve the diversity of Wikidata's collection.





Results

While we hoped that the data would show change over time, it remained remarkably consistent after some early fluctuation. By 2014, about 6% of all artists on Wikidata were tagged using the African American ethnic group statement. That number stayed within a percentage point of 6% from that point on. For context, around 13% of Americans are African American. Around 89% of all artists had no ethnicity statement assigned. Each individual non-African American ethnic group had extremely small numbers, with most below 10 total uses.

The city level data shows slightly more variation, but most cities we explored individually had numbers fairly close to the national data. The outlier was Philadelphia. Philadelphia always had unusually high representation of African American artists on Wikidata, with the number steadily rising throughout the graph. You can see a tiny spike when our project begins, with it rising from 15% to 18% in a little under a month.

Conclusions

It is difficult to draw conclusions about the data itself. We found that the proportion of artists represented by each ethnicity was quite stable over time, with only minimal differences on a city-by-city basis. However, conclusions can be made about the structure of Wikidata. Wikidata has challenges with correctly capturing demographic information. As mentioned elsewhere, Wikidata has a high bar for applying ethnicities to person records. While the goals of this are noble, it may cause statistical underrepresentation. Editors may have trouble finding explicit sourcing for a person's race, and be reluctant to add it without that sourcing.

Another issue comes from the available options for ethnicity representation. The primary entity that is meant to be used to describe Black Americans is "African-American." This can prevent proper representation for those who claim a more specific identity. Ideally it would be possible to, for example, capture those who identify as Afro-Carribean in a sample of Black Artists, but without specifically asking SPARQL to return all possible African ethnicities, it is difficult to capture Black Americans on aggregate. Overall, some hesitancy in trusting Wikidata's counts for ethnicity is appropriate.

The data shows that most entries on Wikidata simply do not have an ethnicity tag, and that almost no white people carry any ethnicity tagging at all. In some ways, this obscures the overall whiteness of the collection. Furthermore, it shows the biases of Wikidata editors. The convention has become to treat whiteness as the default on Wikidata. For Wikidata to become a more inclusive collection, different practices must be adopted that do not consider whiteness to be the default, non-notable and unnecessary to tag. There are obstacles to this, especially when it comes to sourcing as the ethnicity property requires.

Many of the issues with ethnicity tagging cannot be solved by Wikidata's editors. Defining Blackness is not a task that should be left up to people whose main goal is to solve coding issues. These are not new challenges, and many Wikidata editors are engaged in discussing future steps. Similar efforts are underway to properly define gender within Wikidata, as non-cisgender identities are typically given their own category. "Transgender man" and "transgender woman" are both their own entities distinct from "man" and "women," contributing to an othering effect. Wikidata is designed to be somewhat light on its feet compared to more static catalogs such as the Library of Congress Subject Headings, and the Wikidata community engaging in these efforts (often through the LD4 linked data working groups) will hopefully lead to a more representative catalog.

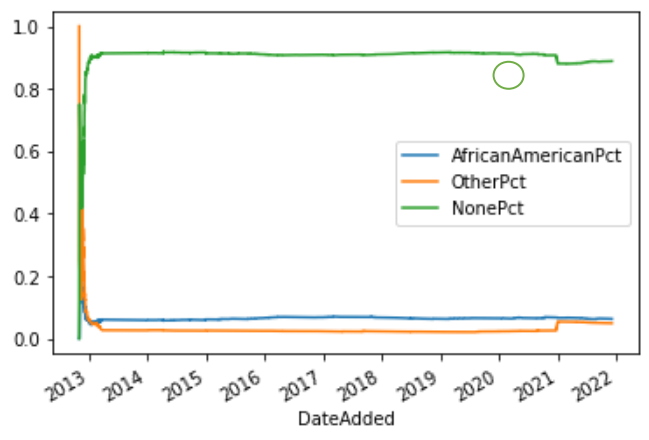
Full Project Team

Mentors (Temple University Library)
Synatry Smith
Holly Tomren
Alex Werner-Colan

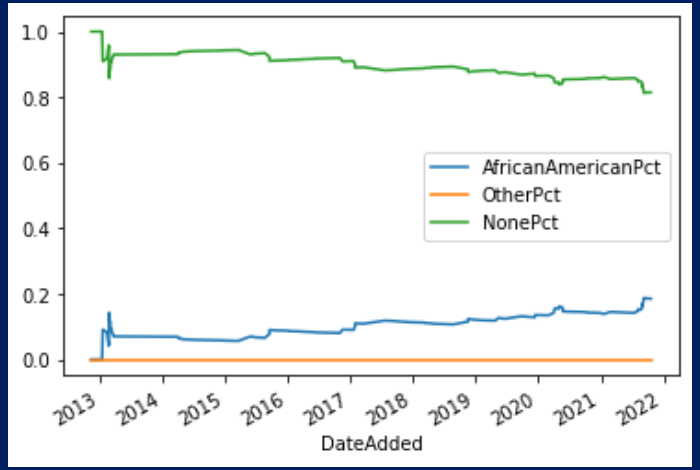
Fellows
Rebecca Bayeck, Schomburg Center at NYPL
jay winkler, ICPSR- University of Michigan

City-by-City Data

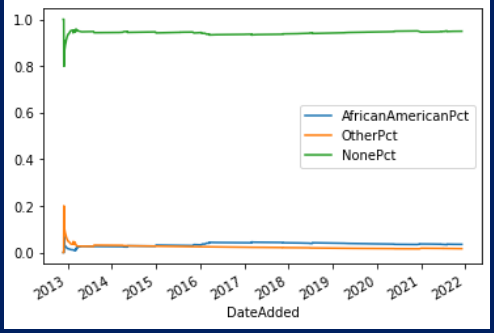
Full Dataset
(n=3683)



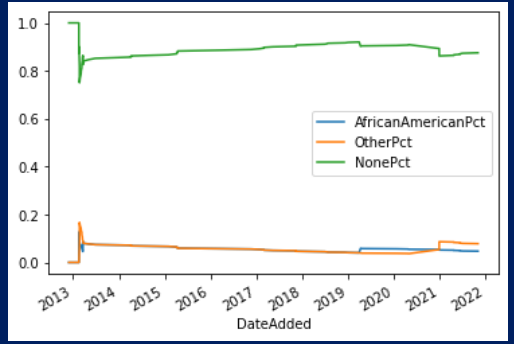
Philadelphia
(n=162)



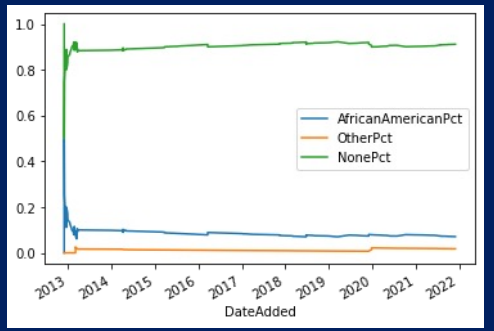
New York City
(n=431)



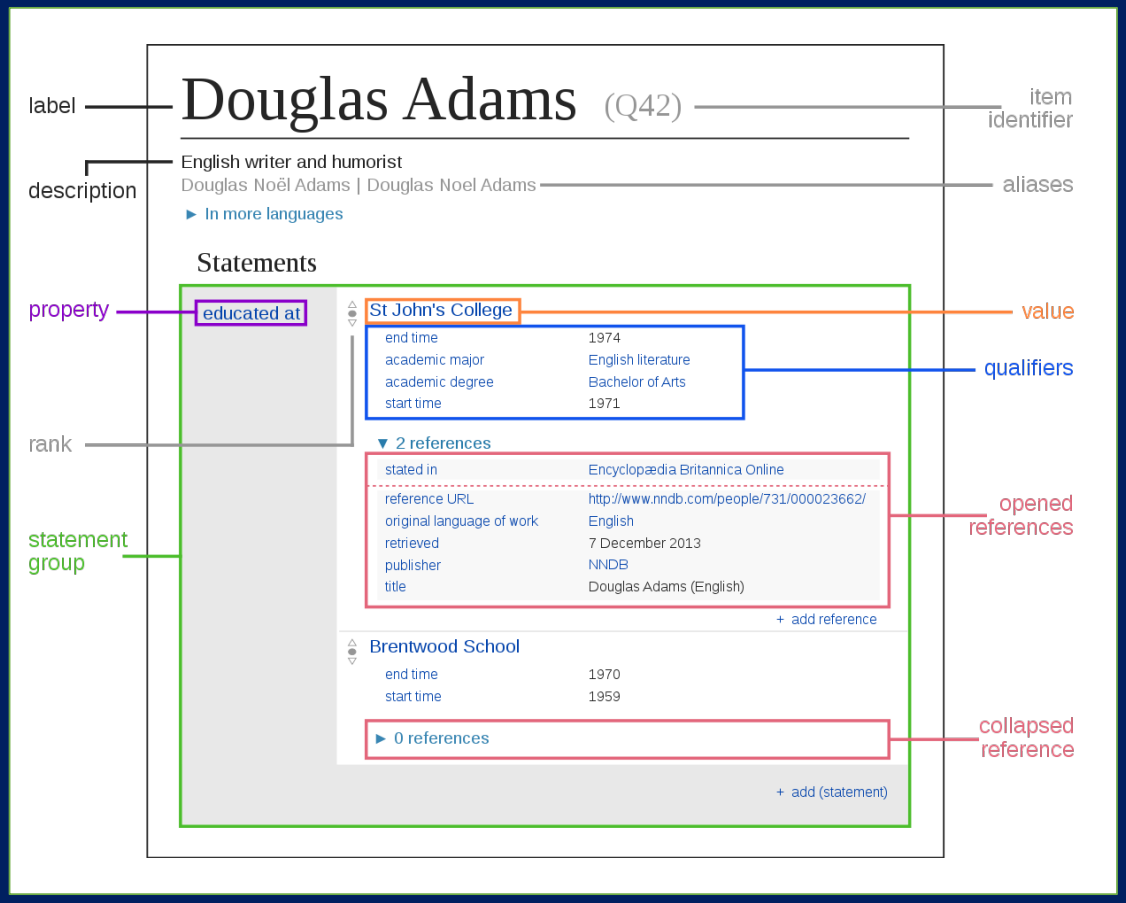
Detroit
(n=64)



Chicago
(n=169)



Wikidata



Eric Battle Q5386100

has property

Occupation P106

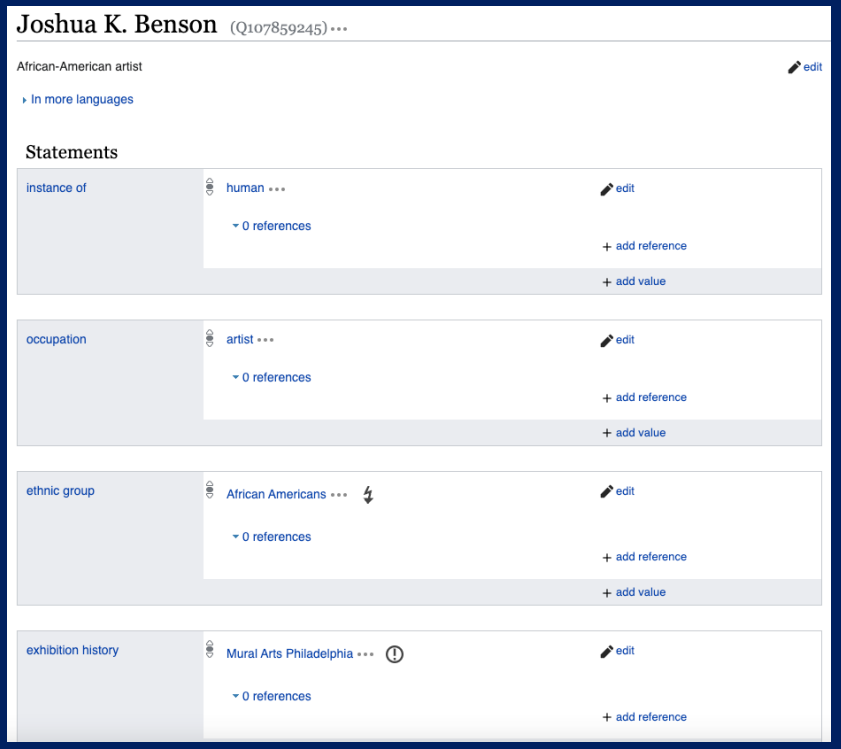
has value

Artist Q483501

Using OpenRefine for Page Creation

I used OpenRefine, a program for data cleaning, to create a table of about 50 artists. I matched the columns to Wikidata properties, and then used OpenRefine's tools to mass ingest the artists into Wikidata.

All	Display Nme	qnumber	date of birth	Name2	date match	year of birth	Begin Date	End Date	Display Bio	ConstituentID	First Name	Last Name	Nationality
1.	Kataoka				N		0	0	NULL	53077	NULL	NULL	NULL
2.	Muneoka tō				N		0	0	NULL	53092	NULL	Muneoka tō	NULL
3.	Alonzo Foringer Choose new match	Q15997631	1878-02-01T00:00:00Z edit		Y	1878	1878	1948	American, 1878 - 1948	53109	Alonzo	Foringer	American
4.	M. Leone Bracker Choose new match	Q37912626	1885-01-01T00:00:00Z		Y	1885	1885	1937	American, 1885 - 1937	53111	M.	Bracker	American
5.	Ishiwa				N		0	0	NULL	53128	NULL	Ishiwa	NULL
6.	Tsujiya Yasubei				N		0	0	NULL	53143	NULL	NULL	NULL
7.	Hayashiya Shōgorō Choose new match	Q68023349			N		0	0	NULL	53145	NULL	NULL	NULL
8.	Marianne Hunter				N		1949	9999	American, born 1949	63745	Marianne	Hunter	American
9.	Janel Jacobson				N		1950	9999	American, born 1950	63750	Janel	Jacobson	American
10.	Jessica Charlesworth				N		1979	9999	English/Canadian, born 1979	63752	Jessica	Charlesworth	English/Canadian
11.	Douglas Frank Choose new match	Q94508460			N		1948	9999	American, born 1948	63777	Douglas	Frank	American
12.	Aaron Martinet Choose new match	Q29909060	1762-01-01T00:00:00Z		Y	1762	1762	1841	French, 1762 - 1841	63786	Aaron	Martinet	French
13.	Louis Roupert Choose new match	Q92344003	1700-01-01T00:00:00Z		N	1700	1666	1699	French, active late 17th century	53115	Louis	Roupert	French
14.	Hariya			Nagasaki	N		0	0	NULL	53122	NULL	NULL	NULL
15.	Alfred Robaut Choose new match	Q2835426	1830-05-20T00:00:00Z		N	1830	0	0	NULL	53124	NULL	NULL	NULL
16.	Hanabusaya Bunzō				N		0	0	NULL	53147	NULL	NULL	NULL
17.	Walter Kidde Sales Co.				N		0	0	NULL	53149	NULL	Kidde	NULL
18.	Mark Perper				N		1900	1954	American, 1900 - 1954	53156	Mark	Perper	American
19.	Ian Hamilton Finlay Choose new match	Q567547	1925-10-28T00:00:00Z		Y	1925	1925	2006	British, 1925 - 2006	53158	Ian	Finlay	British



Google Colab

Querying Wikidata API

This queries for the first timestamp of each artist

```
S = requests.Session()

URL = "https://www.wikidata.org/w/api.php"

finalDate = []

for item in data["QNum"]:

  PARAMS = {
      "action": "query",
      "format": "json",
      "prop": "revisions",
      "titles": item,
      "rvprop": "timestamp",
      "rvlimit": "1",
      "rvdir": "newer"
  }

  R = S.get(url=URL, params=PARAMS)
```

```
sparql.setQuery("""
SELECT
    ?artist ?artistLabel ?sexGenderLabel ?sexualOrientationLabel ?ethnicityLabel
    (group_concat(DISTINCT(?birthPlaceLabel);separator=", ") as ?birthPlaces)

WHERE
{
    ?artist wdt:P106 wd:Q483501 .
    ?artist wdt:P19 ?birthPlace .
    { ?birthPlace wdt:P31/wdt:P279* wd:Q35657. } UNION { ?birthPlace wdt:P31/wdt:P279* wd:Q1093829. } UNION { ?birthPlace wdt:P19 wd:Q30. }
    OPTIONAL { ?artist wdt:P21 ?sexGender. }
    OPTIONAL { ?artist wdt:P91 ?sexualOrientation. }
    OPTIONAL { ?artist wdt:P172 ?ethnicity. }
    SERVICE wikibase:label {
    bd:serviceParam wikibase:language "en".
    ?artist rdfs:label ?artistLabel .
    ?sexGender rdfs:label ?sexGenderLabel .
    ?birthPlace rdfs:label ?birthPlaceLabel .
    ?sexualOrientation rdfs:label ?sexualOrientationLabel .
    ?ethnicity rdfs:label ?ethnicityLabel .
  }
}
GROUP BY ?artist ?artistLabel ?sexGenderLabel ?sexualOrientationLabel ?ethnicityLabel
ORDER BY ?artistLabel
""")
```

The project team used Google Colab to access the Wikidata SPARQL endpoint, and manipulate and plot the data.

Our Colab notebook is available for viewing at tinyurl.com/templecolab.

Read our blogs!

Tinyurl.com/leadingwikidata