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

But what about more complex questions? What if I wanted to know, say, who are all 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 database.

As part of the 2011 LEADING Fellowship, the team worked on a project enhancing and analyzing the available information on Philadelphia's Black artists. While the project was initially focused on the arts themselves, during the research I became interested in the limitations and challenges of Wikidata.

There are a number of issues surrounding the way that demographic tagging on Wikidata occurs. As a community resource, Wikidata edition presents 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.

While our research was focused on the contrast between tagging of Black artists and untagged entries, other ethnic groups were also represented. The data shows that most entries on Wikidata simply do not have an ethnicity tag, and that almost no white people carry any ethnicity tag 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, unnotable and unnecessary to tag. There are obstacles to this, especially when it comes to ensuring that the wiki is not biased by property engine.

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-binary identities are typically given their own category. "Transgender man" and "transgender woman" are both their own entity, as non-binary people carry different identities than their own, as well as looking at what Wikidata editors and other libraries are already doing to attempt to mitigate some of these issues.

**Introduction**

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 Black Artists, with a specific focus on Philadelphia. Our first goal was to apply the American Collection, as well as artists that had been showcased at Philadelphia Black Artists, including artists in the Charles H. Wright African American Museum.

Once I had completed this process on the list of all artists with American birthplaces, I then gathered subsets of the data to answer the question of where they were born. While our research was focused on the contrast between tagging of Black artists and untagged entries, other ethnic groups were represented. The data shows that most entries on Wikidata simply do not have an ethnicity tag, and that almost no white people carry any ethnicity tag 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, unnotable and unnecessary to tag. There are obstacles to this, especially when it comes to ensuring that the wiki is not biased by property engine.

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-binary identities are typically given their own category. "Transgender man" and "transgender woman" are both their own entity, as non-binary people carry different identities than their own, as well as looking at what Wikidata editors and other libraries are already doing to attempt to mitigate some of these issues.

**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 wanted to focus my querying and visualization on that challenge. I started thinking about the actual property for ethnicity, and how it is used differently over time. Wikidata is a collaborative platform, and its bases reflect the biases of the editors. Additionally, the ethnicity property, by design, has a high bar for usage. Editions are instructed not to use the ethnicity property unless they are absolutely sure, preferentially 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, reproducible environment. Google Colab allows teams to collaborate online, 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 and calculate how the percentage of artists that had “African American” as their ethnicity changed over time. 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. 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 these 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 patterns that exist over time that could provide insight into the ethnic group property is being used at all.

While we hoped that the data would show change over time, it remained remarkably consistent after some early fluctuations. 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 or two 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 users.

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, which 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 began, 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 accurately capturing demographic information. As mentioned elsewhere, Wikidata has a high bar for applying attributes 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 identify as Afro-Caribbean to 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 hesitation 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 tag 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, unnotable and unnecessary to tag. There are obstacles to this, especially when it comes to ensuring that the wiki is not biased by property engine.

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-binary identities are typically given their own category. "Transgender man" and "transgender woman" are both their own entity, as non-binary people carry different identities than their own, as well as looking at what Wikidata editors and other libraries are already doing to attempt to mitigate some of these issues.

**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**

**Douglas Adams** (Q42)

**Statements**

- **St John's College**
  - **end time:** 1974
  - **academic major:** English literature
  - **academic degree:** Bachelor of Arts
  - **start time:** 1971

- **Brentwood School**
  - **end time:** 1970
  - **start time:** 1959

**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.
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