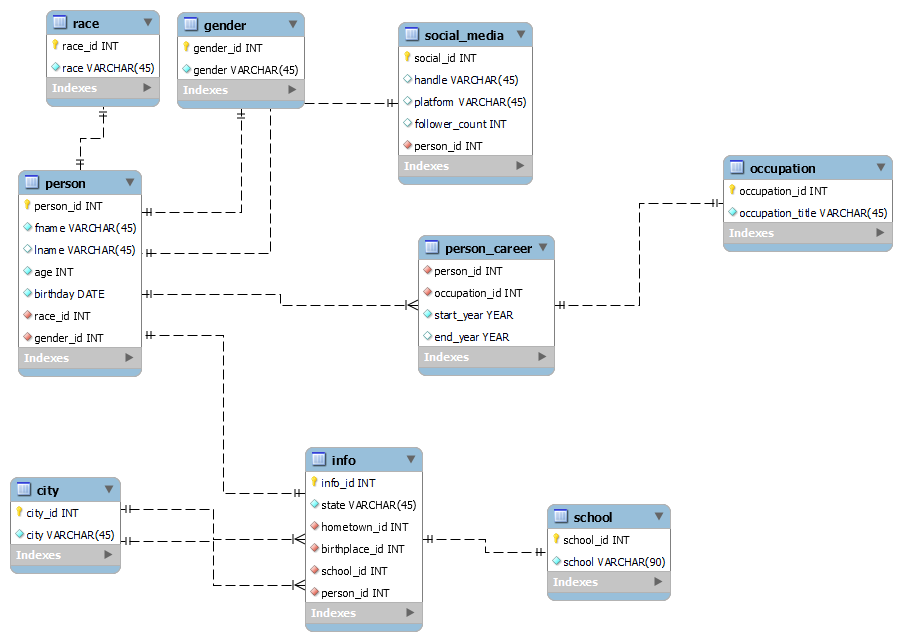
**Introduction and Database Description:**

For this project, we wanted to create a database containing well-known public figures in the DMV area. Thousands of well-known actors, actresses, presidents, and artists reside in the DC, Maryland, and Virginia area, and compiling them into one database would make finding information for specific individuals easier. Having everything in one database makes it much easier to gather information about an individual instead of checking multiple sites.

The type of data that we included in our database was information like name, age, race, gender, birthday, occupation, hometown, city, state they were born in and any social media accounts. We normilized the data and were able to separate the tables into appropriate categories, to help with the organization of the data. We decided because of time restrictions that we were only going to include the 20 most well known public figures in the DMV area, rather than thousands of rows of public figures, as we deemed that would not be feasible for this project. After loading our data into the database, we began creating our views that had example queries that a user might want to use. Example queries include looking up the occupations of figures 50 years and younger. Another example might be finding the average followers of all the social media accounts in the database. The table below lists the queries that we made, along with the requirement that it satisfies. The views that we made attempted to showcase what type of information was in our database, and the type of people, occupations, and states that they reside in.

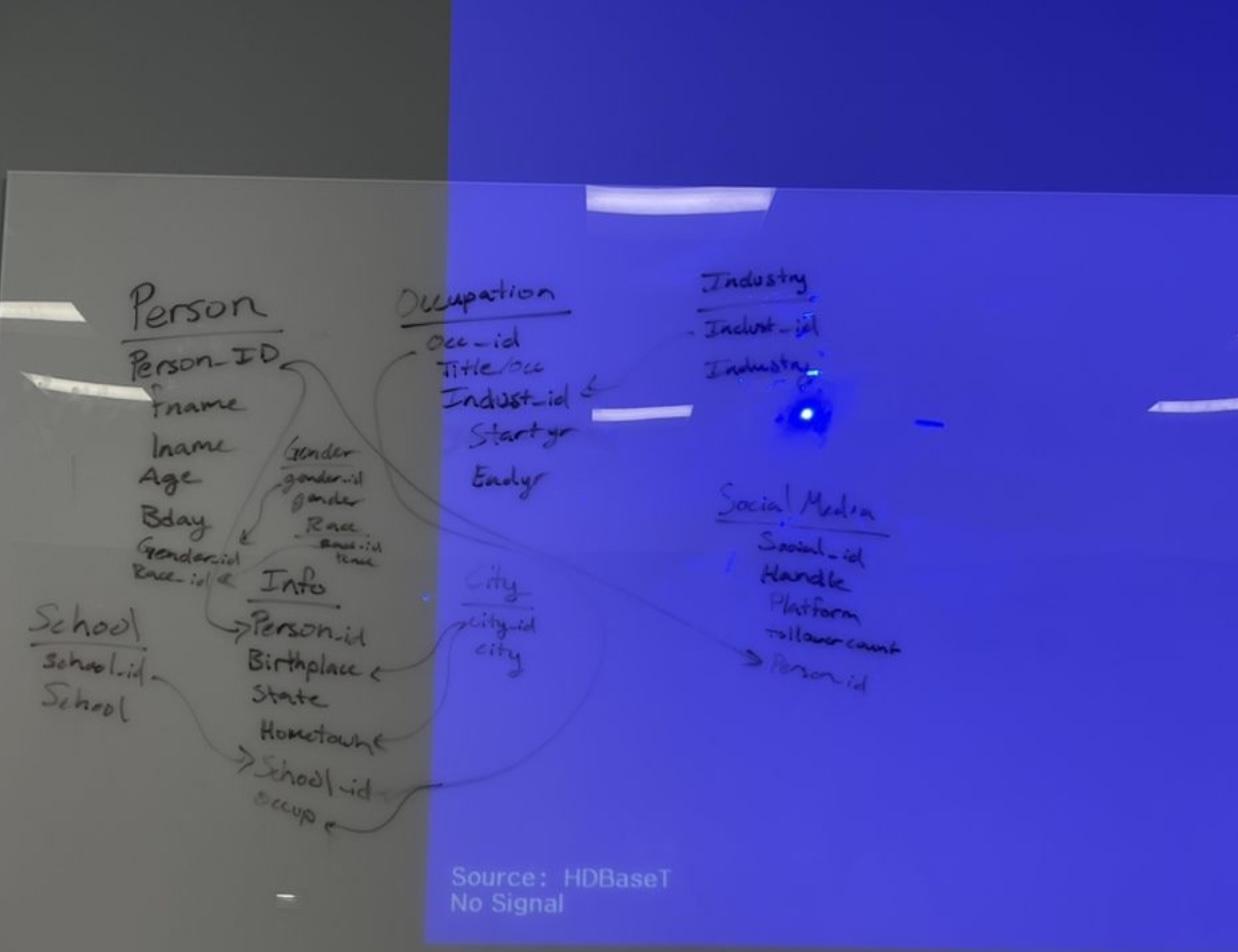
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **View/Query Name** | **Req A** | **Req B** | **Req C** | **Req D** | **Req E** |
| person\_occupation | X | X | X | X |  |
| social\_media\_count | X | X | X |  |  |
| occupation\_count | X | X | X |  |  |
| state\_born | X | X | X |  |  |
| occupation\_type |  | X |  | X | X |

There are many sites out there that have the names and information of public figures. But there isn’t a site that has all of the metadata of each of those figures together in one place. We chose to create this database because it would allow users to access data altogether rather than having to jump around and search for information. Being able to have all of the data for one person in one place makes researching significantly easier. In the ERD image below, we created several tables highlighting important information about a specific public figure, including race, occupation, social media accounts, the school they attended, etc. By creating this database, it hopefully makes information easier to access and more concise as well.



**Changes from Original Design**

This is what our start up sketch looked like. We used time in lab to work with the TA on perfecting our database and making sure we had the correct information for our ERD. The TA drew up this sketch as we listed out what our ERD would contain and he helped figure out what the lines would connect with. We decided to get rid of the industry table, since we found that information to be redundant with occupation. We also added a linking table, person\_career, that we could link the occupation with the person, and this would make it easier to keep track of multiple occupations for one person. The info table was also given a new primary key, so that information was easier to track across the entire database.

****

**Database Ethics Consideration**

We planned for out database to be as inclusive as possible, so we limited our domain to figures only well-known in the DMV area. This database is as transparent as possible regarding containing the correct information about each individual. During this project, we attempted to pull data from as many public figures as possible, so there is no bias in favoring specific actors/actresses, for example. This means any public figure who is disabled, part of the LGBTQ+ community, racially diverse, etc. is added to our database. This database and the tables inside the database are the most accurate and up-to-date during the time of its creation. As stated in the sample data plan, we planned to choose who we want to put into our database as a group, so that the database can be as diverse as possible with limited bias. We wanted the database to accurately reflect the diversity of the DMV area, so the only way to do this was to include people of all different types of industries, occupations, race, etc.

Since our database uses personal information relating to people within our database, there are some ethical considerations to consider. One consideration we understand is that there are concerns regarding safety, relating to home robberies, stalking, or cyberbullying. There has been an ongoing issue relating to home robberies of individuals with certain levels of fame and fortune due to their homes being publicly known. Another consideration we had related to confidentiality. Unfortunately, this information is made public by the nature of paparazzi and public records leaving this issue at hand. We believe that there are no data elements that might lead to compromised privacy in a case of a flawed user access policy or data breach. This data is general information so there is no case relating to a data breach. We did consider whether our data plan might pose copyright or fair use risks. We found no substantial information that stated what our database presents is a breach of copyright or fair use violations. Our sample plan does not include data from proprietary or closed data sources. We plan to use data from sources that are widely available to use by the public.

In the US, some laws prevent unintentional privacy from public organizations called the “Privacy Act of 1974”. The act guarantees a code of fair information practices that manage the collection, maintenance, use, and dissemination of information about individuals in systems of records by federal agencies. All of the public data collected must be counted on to be used in the right manner. Since we are not using data like home addresses, phone numbers, or email addresses, we see no issue with breaking this act.

**Lessons Learned**

Some lessons we learned were that we need to review what is implemented into the database before everything is added in because we either forgot to add individuals or there were some typos in certain demographic data. We simply either updated the data by update statements or within the actual script as some of the typos were easy to fix. Another lesson we learned is making sure we review subqueries as we had issues and the way to resolve these issues were resolved by TA’s help. Another lesson that was learned as a result of TA help was when forward engineering there were some errors that were presented that we weren't able to troubleshoot by ourselves. We learned that we should make sure that the datatypes are the same in our CSV files so that importing the data into the database is as easy as possible. We also had some issues with out ERD early on, which was also resolved by working with the TA to get the tables fixed and making the database the best that it could be.

**Potential Future Work**

If we had more time to work on this in the future, we could make our ERD a little more complex. We could try color coding the different elements in the ERD. For example, different colors for the primary keys. We also could spend more time figuring out how to create a subquery multiple times, without facing any issues. Being able to do this assignment with the help of a TA along the way really helped but I feel like if we needed to do this project again, we should be able to try it and get far, without the help of a TA. Another thing we could do in the future is possibly adding notes to our ERD. This is a very optional thing we could do but having notes in the ERD helps everyone understand the different relationships and whats going on in the ERD. The notes could also contain reasoning behind our decisions, like why we chose a specific datatype. We could also potentially add more people to our database, expanding the information that is housed inside of the database. By adding more people, it would diversify our database even more and allow for users to see all of the different types of people that grew up or reside in the DMV area. Since we were limited with time, we could only add a select number of people, but if we had more time to expand on our ideas, we would be able to make this a much bigger database. If we could change something that we wish we did differently, it would probably be designing our ERD more efficiently and establishing the type of information each table would house. Overall, this was a really good learning experience and taught us a lot about how to design, import and normalize a relational database.