**Some Information about dblp data**

The original data from the website [DBLP](https://web.archive.org/web/20170721110641/https://kdl.cs.umass.edu/display/public/DBLP) is an xml file that is uploaded in the Google drive. It contains the entire data. It is hard and requires some time to make sense out of the data by looking at the xml file. Since it is extremely huge, It needs to be opened with an IDE like visual studio code. I have also uploaded a sample of that data to give an idea of how it is actually structured.

The data can also be obtained from BigQuery. There are three main components of the data, objects, links and attributes. Every attribute has its own object\_id. The object component consists of all the objects\_ids in the dataset. The data is originally a graph data that connects two objects with links. While the link component is what that connects two object ids using a link\_id. The attribute component contains the raw data (each object id and its value).

Big Query can be used for data processing before it can be uploaded in the database of your choice. An another way can be using a programming language like python, scala or similar to do the required preprocessing to upload the data in the database.

For the purpose of Project Task 1, understanding the data is required. The dataset is explained clearly in the above link.