# A Modeling, Loading, Evolving

## A.1 Modeling

1. The metadata/schema visual representation of the graph is shown below.

A screenshot of a computer

Description automatically generated with medium confidence

2. The design decisions made are explained below.

* We decided to represent the corresponding author as an edge because such person doesn’t have any differences in properties between them and other types of authors.
* Because there is a query that distinguishes conference editions, we decided to represent it as a node, while the volume of a journal is being represented in a simpler way as a property of the edge published\_in, because we don’t have queries specifying different volumes.
* Since we will eventually query a community of keywords we decided it’s best to represent them as nodes.

We have made the following assumption.

* Since we are modeling a graph for publications, we assume that rejected articles (and therefore not published) won't be represented in this database.

## A.2 Instantiating/Loading

1. Explanations on how we generated the data and any other assumptions about the solution.

The Cypher expressions to create and instantiate the solution are in a file named load.cypher. It is being executed by a Python application attached to the project.

We obtained the base for our data from the BYU Engineering Publications in Scopus 2017-21 Kaggle dataset[[1]](#footnote-1) as a csv file, which presents data about publications and already provided real data regarding authors, publication title, publication year, volume, DOI, access link, author’s affiliations, keywords, document type (if article, conference paper or other), and publication stage. The Python script generate\_data.py processes this file with the following steps:

1. Only use articles and conference papers;
2. Discard data that does not have one of the fields needed (such as affiliation, keywords, duplicated rows, etc.);
3. Remove the edition information such as year and edition number from the conference names
   1. For example,

“Proceedings - 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science, IRI 2020”

became

“Proceedings – IEEE International Conference on Information Reuse and Integration for Data Science, IRI”;

1. Parse the author’s affiliations and generate a separate csv file with them;
2. Because we lacked sufficient papers in the same conference but in four different years (for one of the queries requested), we also synthetically created extra papers for the previous and/or next years;
3. We synthetically created the citation links between articles and stored them in a separate csv file. For creating such citations, we made sure that the papers being cited were from a previous year and had at least one keyword in common with the paper citing them. A paper could have from 0 to 20 citations.
4. Since we didn’t have reviewer data, we chose up to 3 authors that have at least one keyword in common with the paper and assigned them as reviewers, saving this information in a separate csv file. We have also generated some text for the reviews themselves and stored them as a separate csv. Since we assumed all publications would be accepted in the end, there can be up to 1 reviewer rejecting each of them.

1. <https://www.kaggle.com/dpixton/byu-engineering-publications-in-scopus-201721/version/1> [↑](#footnote-ref-1)