

2 Relational Database Schema for Wikipedia Articles

Exercise Sheet 3

Big Data Analytics

Tobias Machnitzki und Finn Burgemeister, WiSe 2017/2018

- Create a relational schema for Wikipedia data in (at least) third normal form.

Lecture „Databases and Data Warehouses“:

1 NF (normal form)	It follows our notion of a table. No collections in the table. A primary key exists.
2 NF	No redundancy of data, i.e., entities of many-to-many relations are stored in separate tables. Every column must depend on each candidate key and not a subset.
3 NF	Columns are not functional dependent to sth. else than a candidate key.
4 NF	Do not store multiple relationships in one table.

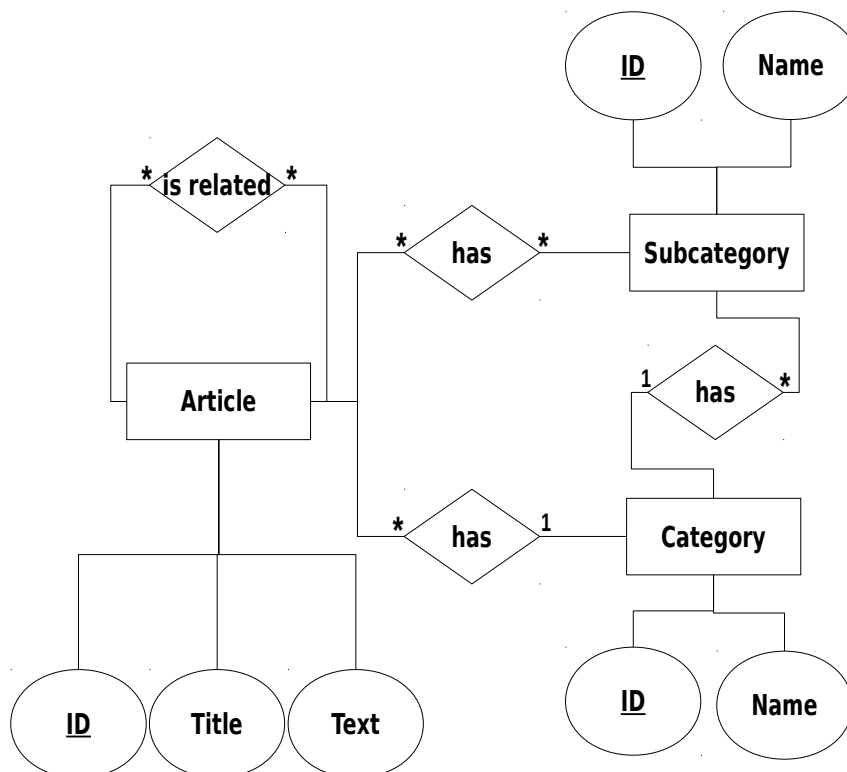


Figure 1: Entity-relationship diagram for Wikipedia data in fourth normal form (description on next page)

Description of entity-relationship diagram for Wikipedia data in fourth normal form

We have to create the following **tables**:

Article, Category, Subcategory, Articles_related*, Article_Subcategory*,
Article_Category* (*representing a relation)

The following **key** definition is valid for every table:

ID : Integer - Primary key (also surrogate key)

The following **cardinalities** have to be implemented:

- (i) Many Articles are related to many articles
- (ii) Many Articles have many subcategories
- (iii) One Category has many articles
- (iv) One Category has many subcategories

- *Discuss the usefulness of indexes for each of the relations*

It is required to use IDs (indexes) for each entry, because a title does not need to be unique. In a category 'biography' are names not a possible indicator/key.