System and Methods for Big and Unstructured Data – Group 20

Alessio Buda
Leonardo Cesani
Fausto Lasca
Gabriele Munafò
Matteo Paraboschi

POLITECNICO
MILANO 1863

Project presentation & assumptions

Project presentation

Relational tables

Graphs

Documents

Keys

Columns

Project presentation

Graph model

Documental model

Resilient Distributed Dataset (RDD)

Project presentation

dblp
computer science bibliography

- 4.84 millions of publications
- Different types of publication
- Relevant information on each publication

Assumptions

Conference edition

vs

Journal edition

Organizations' affiliations

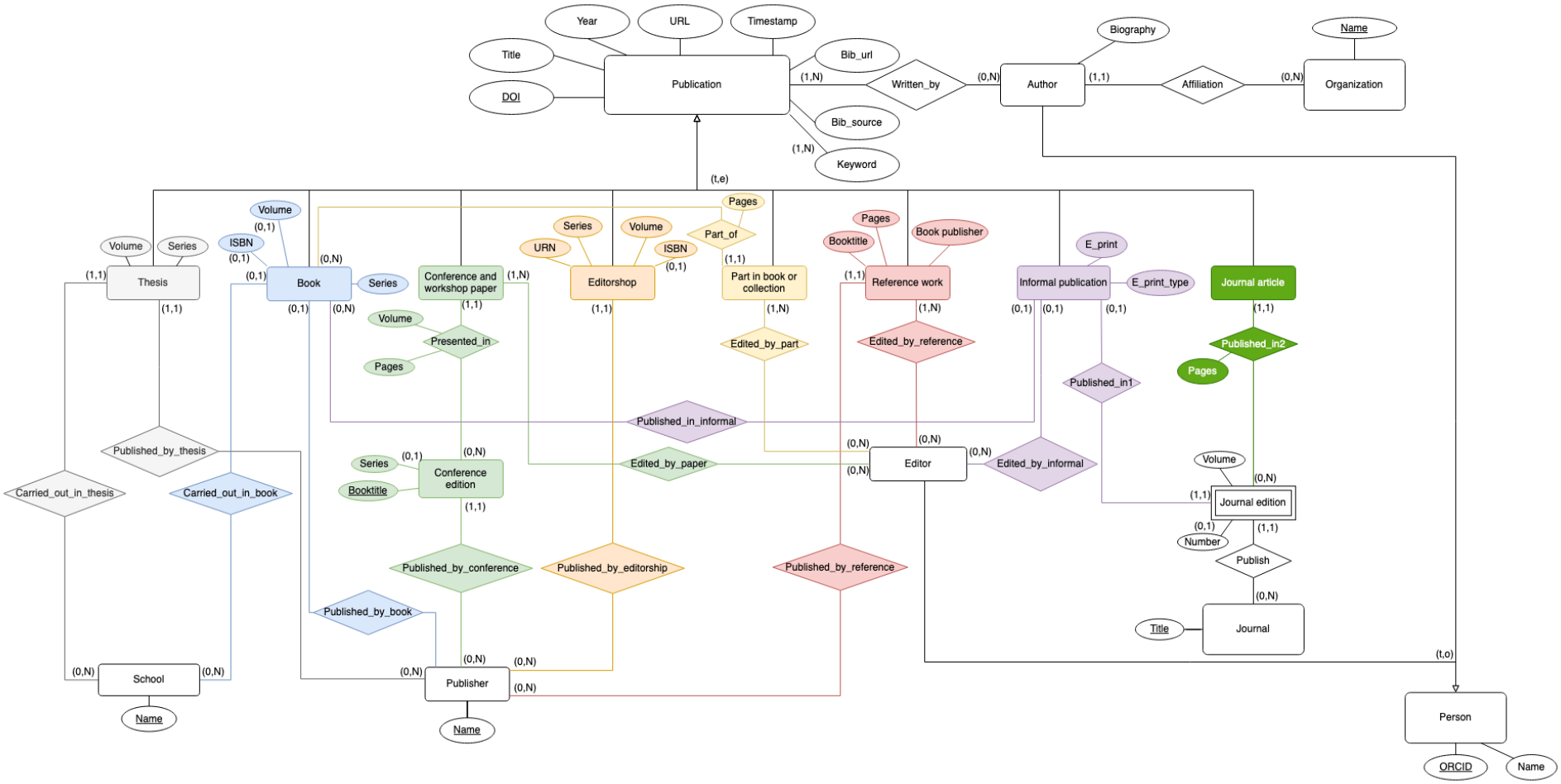
Unique attributes

POLITECNICO
MILANO 1863

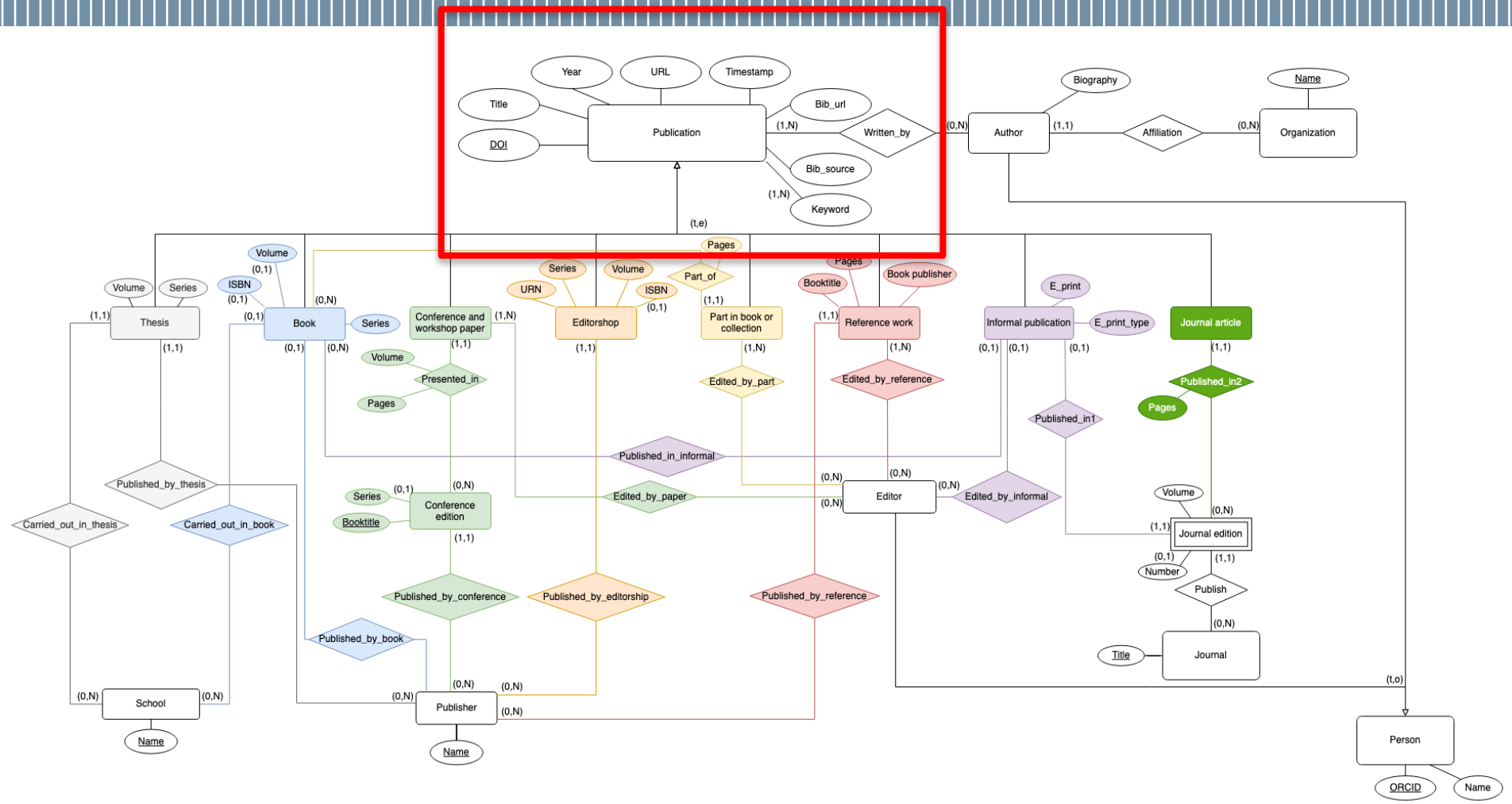
ER model

Conceptual data model

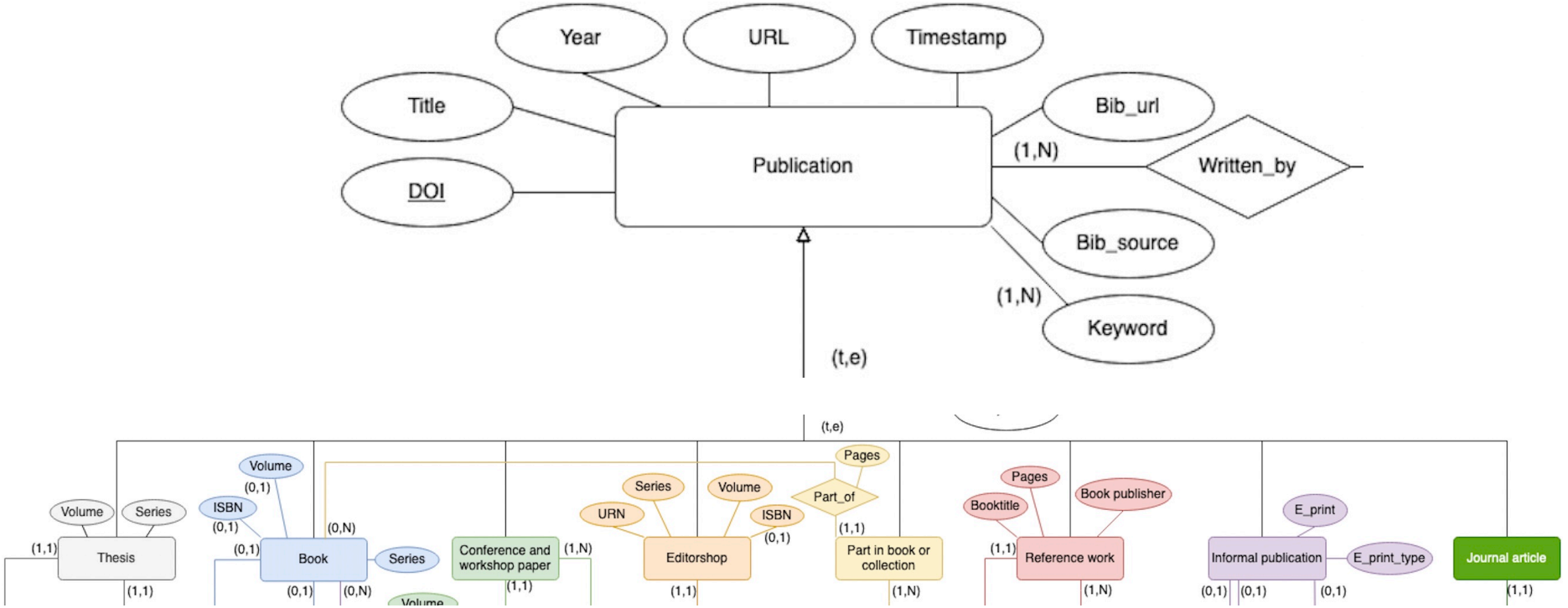
ER Model


Year
URL
Timestamp
Title
DOI
Publication
Bib_url
Bib_source
Keyword
(1,N)
Written_by
(0,N)
Author
Biography
(1,1)
Affiliation
Name
(0,N)
Organization
(t,e)
Volume
Series
ISBN
Thesis
Book
Conference and workshop paper
Pages
Part_of
Part in book or collection
Series
URN
Volume
ISBN
Editorshop
Pages
Booktitle
Book publisher
Reference work
E_print
Informal publication
E_print_type
Journal article
Published_in2
Pages
Presented_in
Edited_by_part
Edited_by_reference
Published_in1
Published_in_informal
Published_by_thesis
Carried_out_in_thesis
Carried_out_in_book
Series
Booktitle
Conference edition
Edited_by_paper
Editor
Edited_by_informal
Volume
Journal edition
Number
Publish
Journal
Title
Published_by_conference
Published_by_editorship
Published_by_reference
Published_by_book
School
Name
Publisher
Name
(t,o)
Person
ORCID
Name

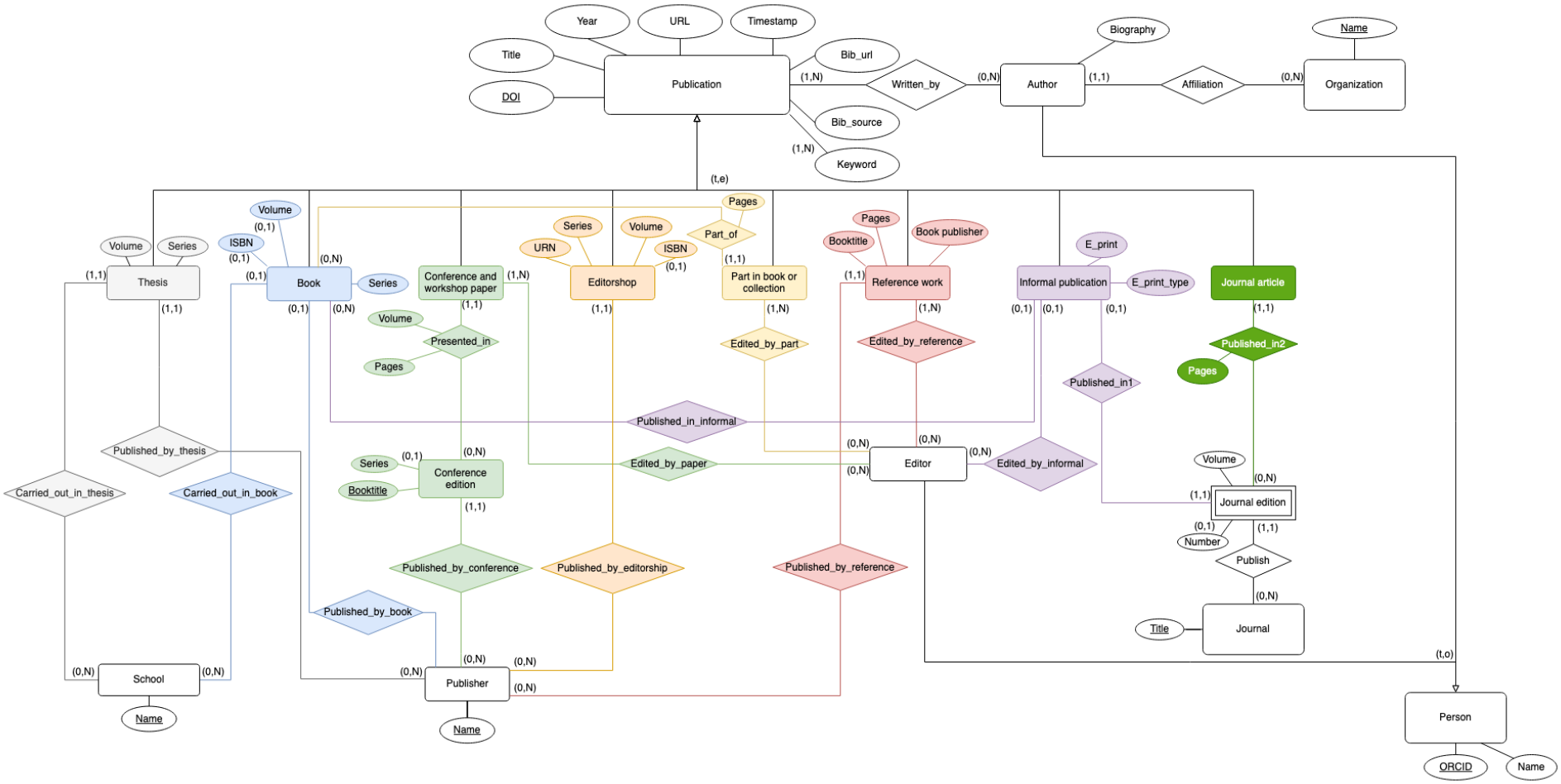
ER Model


Year
URL
Timestamp
Title
DOI
Publication
Bib_url
Bib_source
Keyword
(1,N)
Written_by
(0,N)
Author
Biography
(1,1)
Affiliation
Name
(0,N)
Organization
(t,e)
Pages
Volume
(0,1)
Volume
Series
ISBN
(0,1)
Thesis
(1,1)
(0,1)
Book
(0,N)
Series
Conference and workshop paper
(1,N)
(1,1)
Series
URN
Volume
ISBN
(0,1)
Editorshop
(1,1)
Part_of
(1,1)
Part in book or collection
(1,N)
Pages
Booktitle
(1,1)
Book publisher
Reference work
(1,N)
E_print
Informal publication
(0,1)
(0,1)
E_print_type
(0,1)
Journal article
(1,1)
Volume
Presented_in
Pages
Edited_by_part
Edited_by_reference
Published_in2
Pages
Published_in1
Published_in_informal
Published_by_thesis
Carried_out_in_thesis
Carried_out_in_book
Series
(0,1)
Booktitle
Conference edition
(0,N)
(1,1)
Edited_by_paper
(0,N)
(0,N)
Editor
(0,N)
Edited_by_informal
Volume
(0,N)
(1,1)
Journal edition
(0,1)
(1,1)
Number
Publish
(0,N)
Title
Journal
Published_by_conference
Published_by_editorship
Published_by_reference
Published_by_book
(0,N)
(0,N)
(0,N)
Publisher
(0,N)
(0,N)
School
(0,N)
(0,N)
Name
Name
(t,o)
Person
ORCID
Name

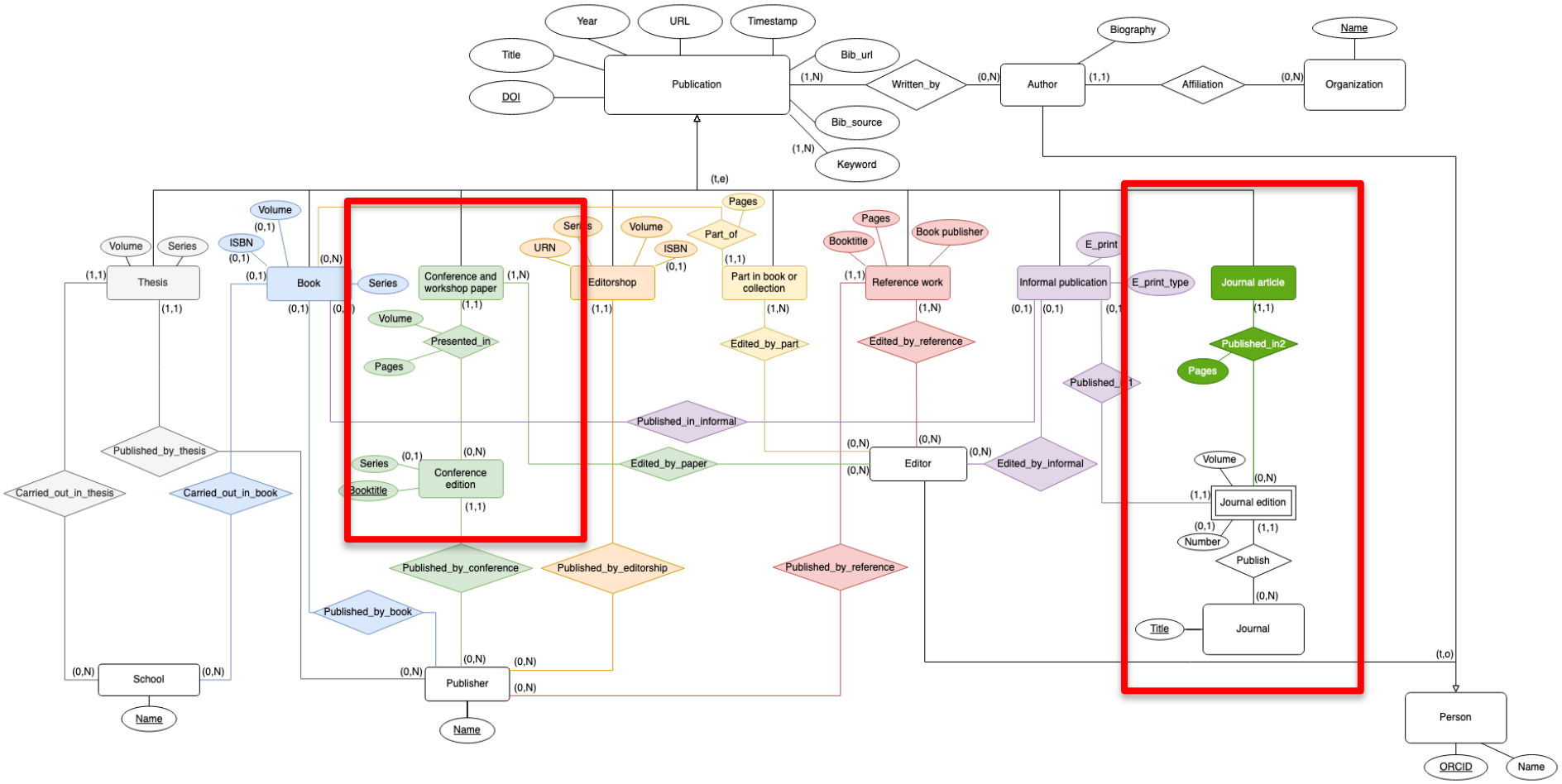
ER Model


Year
URL
Timestamp
Title
DOI
Publication
Bib_url
(1,N)
Written_by
Bib_source
(1,N)
Keyword
(t,e)
(t,e)
Volume
Series
(1,1)
Thesis
(1,1)
Volume
(0,1)
ISBN
(0,1)
Book
(0,1)
(0,N)
(0,1)
(0,N)
Series
Conference and workshop paper
(1,1)
(1,N)
Pages
Series
Volume
Part_of
URN
ISBN
(0,1)
(1,1)
Editorshop
(1,1)
Part in book or collection
(1,N)
Pages
Booktitle
Book publisher
(1,1)
Reference work
(1,N)
E_print
Informal publication
E_print_type
(0,1)
(0,1)
(0,1)
Journal article
(1,1)

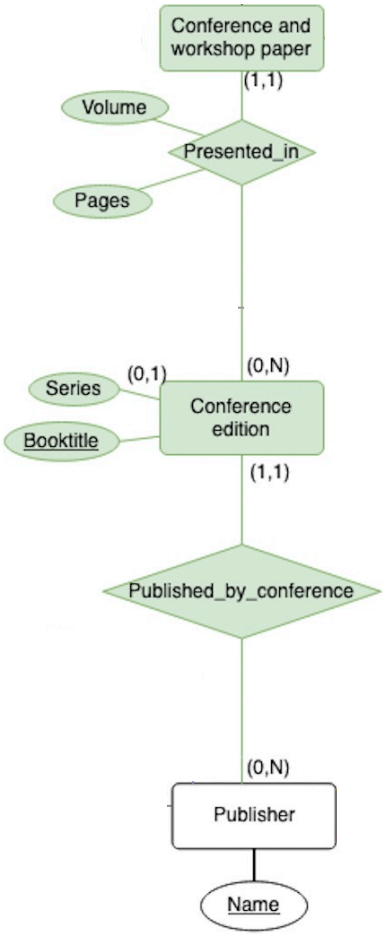
ER Model

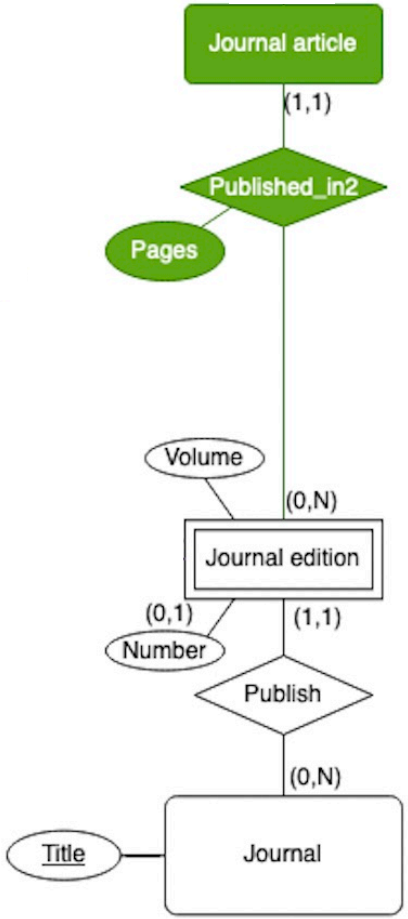

Year
URL
Timestamp
Title
DOI
Publication
Bib_url
Bib_source
Keyword
(1,N)
Written_by
(0,N)
Author
Biography
(1,1)
Affiliation
Name
(0,N)
Organization
(t,e)
Volume
Series
ISBN
Thesis
Book
Conference and workshop paper
Pages
Part_of
Part in book or collection
Series
URN
Volume
ISBN
Editorshop
Pages
Booktitle
Book publisher
Reference work
E_print
Informal publication
E_print_type
Journal article
Published_in2
Pages
Presented_in
Edited_by_part
Edited_by_reference
Published_in1
Published_in_informal
Published_by_thesis
Carried_out_in_thesis
Carried_out_in_book
Series
Booktitle
Conference edition
Edited_by_paper
Editor
Edited_by_informal
Volume
Journal edition
Number
Publish
Journal
Title
Published_by_conference
Published_by_editorship
Published_by_reference
Published_by_book
School
Name
Publisher
Name
(t,o)
Person
ORCID
Name

ER Model

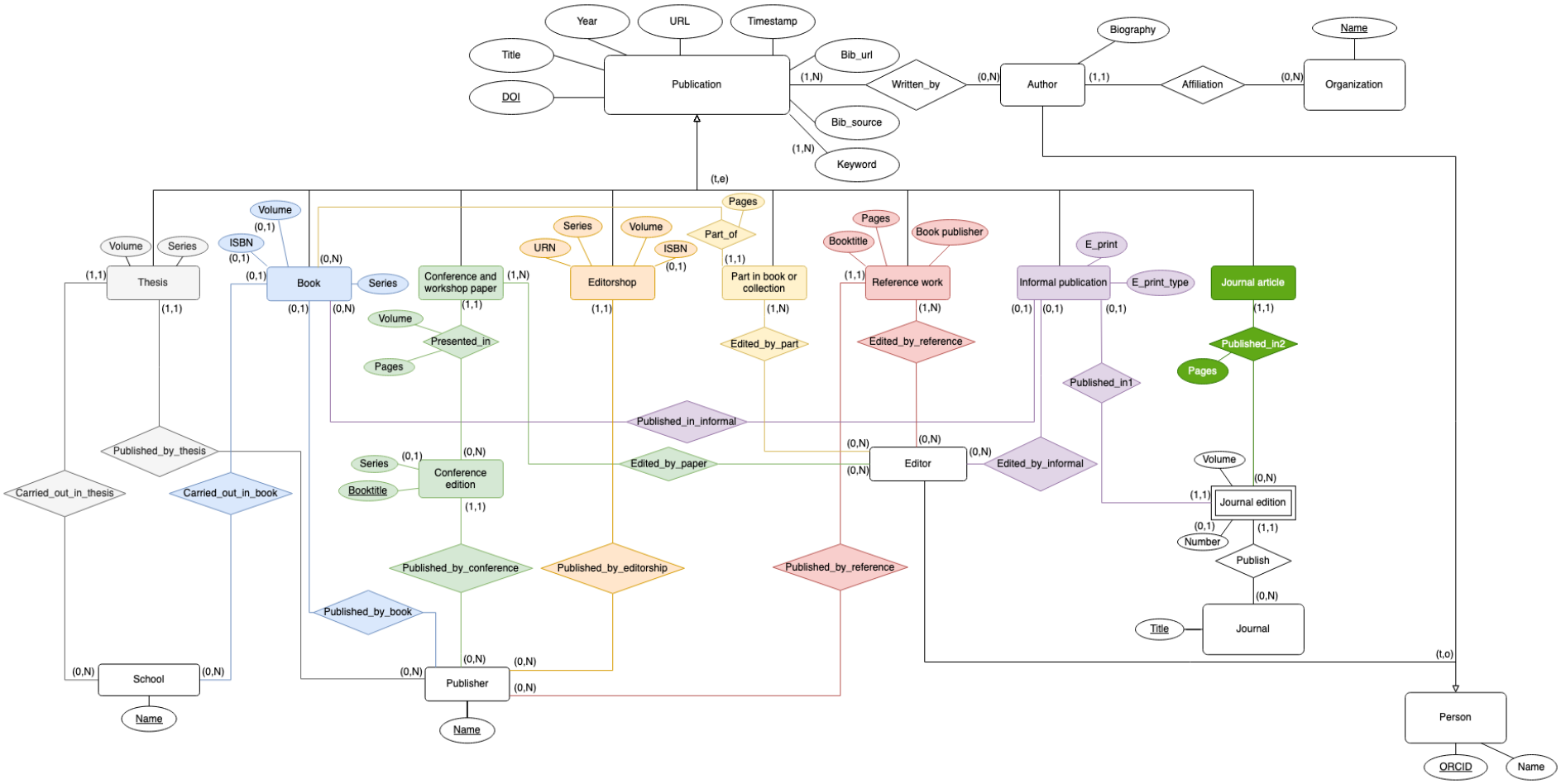

Year
URL
Timestamp
Title
DOI
Publication
Bib_url
Bib_source
Keyword
Written_by
Author
Biography
Affiliation
Name
Organization
(t,e)
Thesis
Book
Conference and workshop paper
Editorship
Part in book or collection
Reference work
Informal publication
Journal article
Conference edition
Journal edition
Journal
Editor
Publisher
School
Person
Published_by_thesis
Carried_out_in_thesis
Carried_out_in_book
Published_by_conference
Published_by_editorship
Published_by_reference
Published_by_book
Published_in_informal
Edited_by_paper
Edited_by_part
Edited_by_reference
Edited_by_informal
Presented_in
Published_in2
Publish
Volume
Series
ISBN
URN
Booktitle
Pages
Part_of
Book publisher
E_print
E_print_type
Number
ORCID
(t,o)

ER Model

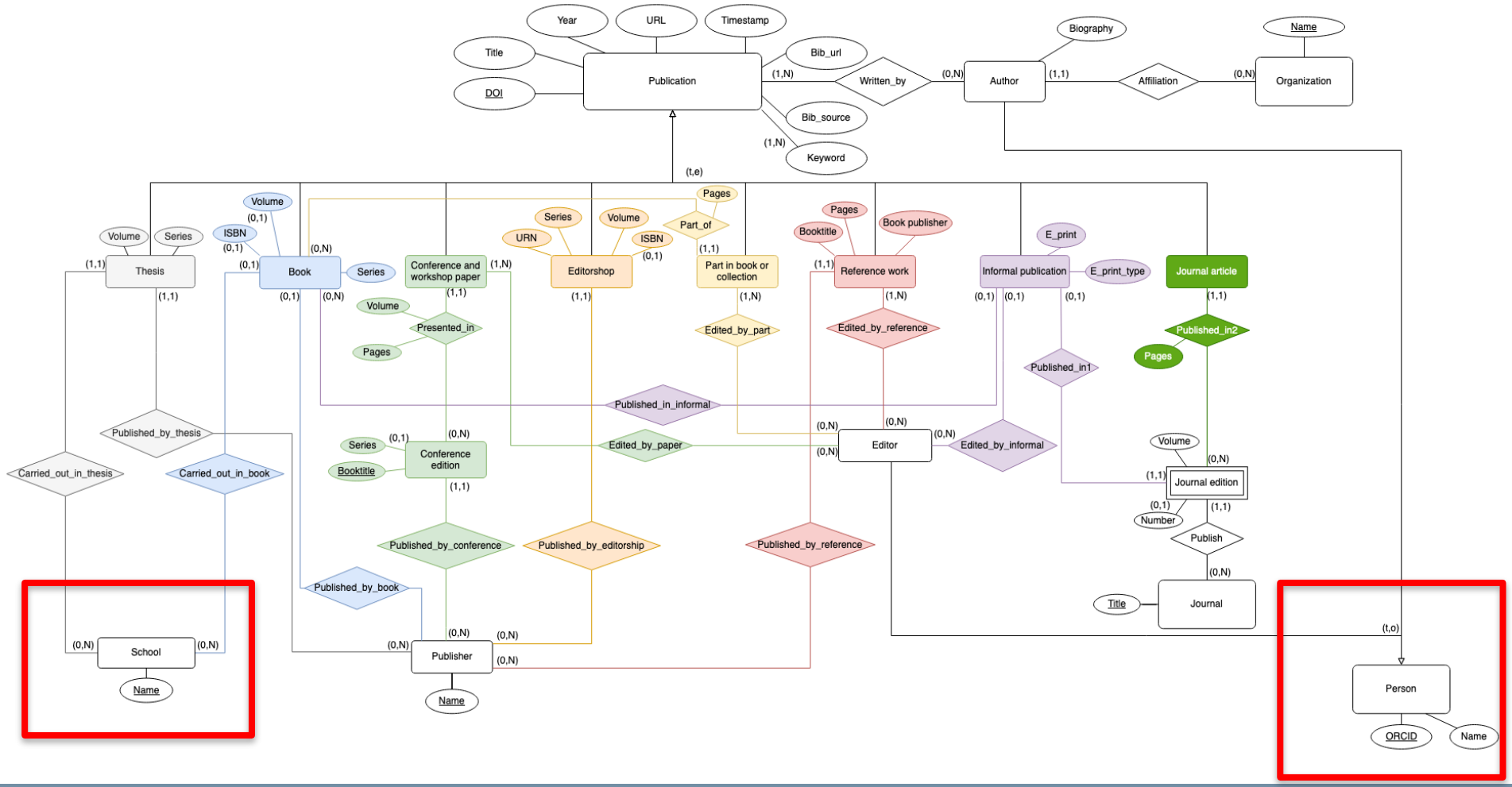

Conference and workshop paper
(1,1)
Volume
Presented_in
Pages
(0,N)
(0,1)
Series
Conference edition
Booktitle
(1,1)
Published_by_conference
(0,N)
Publisher
Name


Journal article
(1,1)
Published_in2
Pages
Volume
(0,N)
Journal edition
(0,1)
(1,1)
Number
Publish
(0,N)
Title
Journal

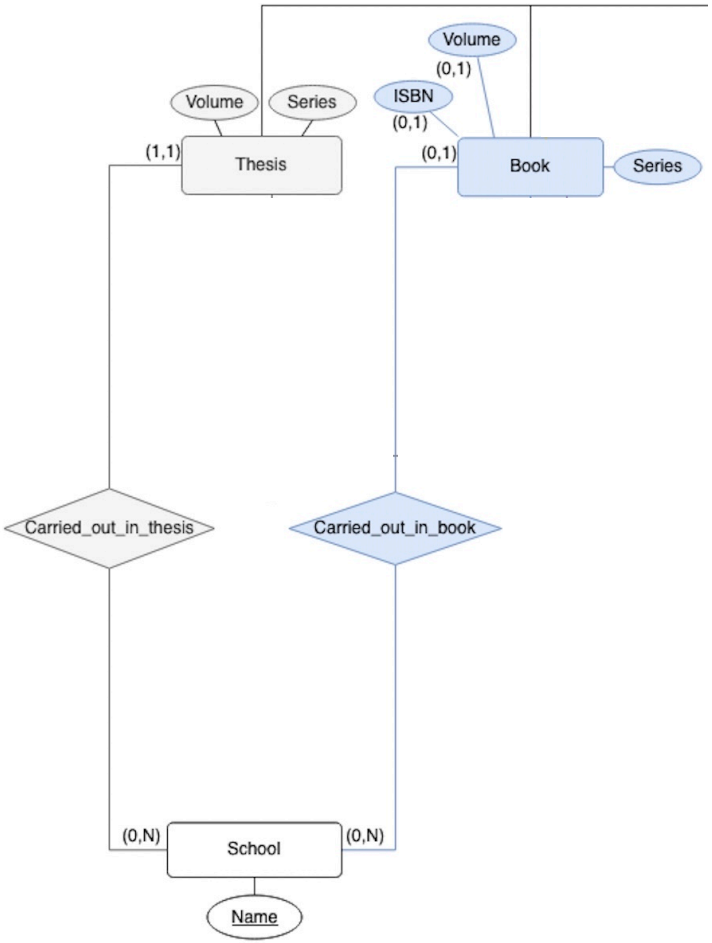
ER Model

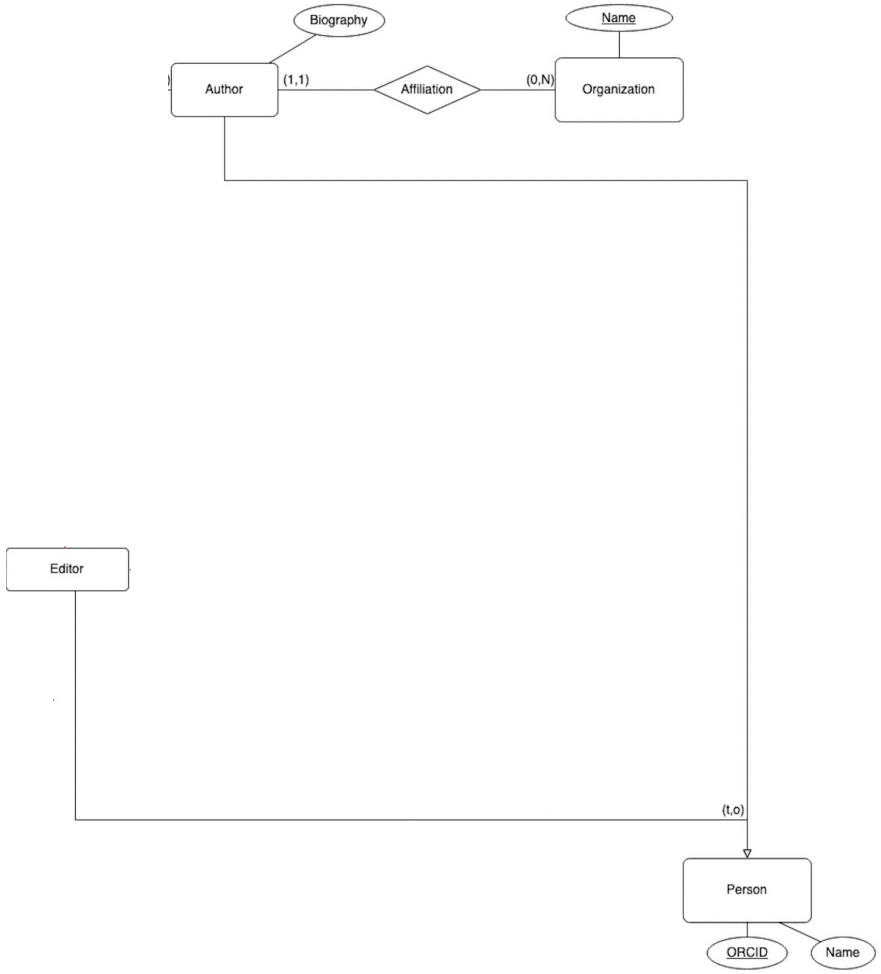

Year
URL
Timestamp
Title
DOI
Publication
Bib_url
Bib_source
Keyword
(1,N)
Written_by
(0,N)
Author
Biography
(1,1)
Affiliation
Name
(0,N)
Organization
(t,e)
Volume
Series
ISBN
Thesis
Book
Conference and workshop paper
Pages
Part_of
Part in book or collection
Series
URN
Volume
ISBN
Editorshop
Pages
Booktitle
Book publisher
Reference work
E_print
Informal publication
E_print_type
Journal article
Published_in2
Pages
Presented_in
Edited_by_part
Edited_by_reference
Published_in1
Published_in_informal
Published_by_thesis
Carried_out_in_thesis
Carried_out_in_book
Series
Booktitle
Conference edition
Edited_by_paper
Editor
Edited_by_informal
Volume
Journal edition
Number
Publish
Journal
Title
Published_by_conference
Published_by_editorship
Published_by_reference
Published_by_book
School
Name
Publisher
Name
(t,o)
Person
ORCID
Name

ER Model


Year
URL
Timestamp
Title
DOI
Publication
Bib_url
Bib_source
Keyword
(1,N)
Written_by
(0,N)
Author
Biography
(1,1)
Affiliation
Name
(0,N)
Organization
(t,e)
Volume
Series
ISBN
Thesis
Book
Conference and workshop paper
Pages
Part_of
Part in book or collection
Series
URN
Volume
ISBN
Editorshop
Pages
Booktitle
Book publisher
Reference work
E_print
Informal publication
E_print_type
Journal article
Published_in2
Pages
Published_in1
Presented_in
Edited_by_part
Edited_by_reference
Published_in_informal
Published_by_thesis
Carried_out_in_thesis
Carried_out_in_book
Series
Booktitle
Conference edition
Edited_by_paper
Editor
Edited_by_informal
Volume
Journal edition
Number
Publish
Journal
Title
Published_by_conference
Published_by_editorship
Published_by_reference
Published_by_book
School
Name
Publisher
Name
(t,o)
Person
ORCID
Name

ER Model

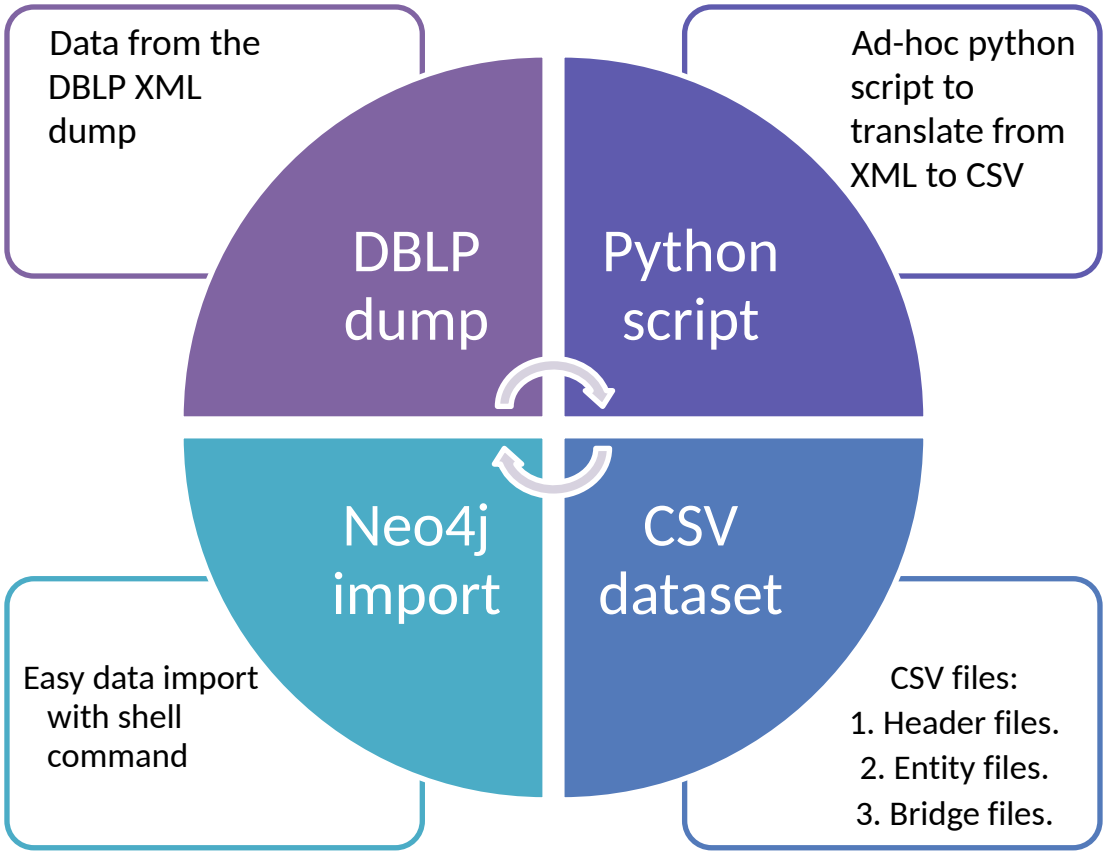

Volume
(0,1)
Volume
Series
ISBN
(0,1)
(1,1)
Thesis
(0,1)
Book
Series
Carried_out_in_thesis
Carried_out_in_book
(0,N)
School
(0,N)
Name


Biography
Name
Author
(1,1)
Affiliation
(0,N)
Organization
Editor
(t,o)
Person
ORCID
Name

POLITECNICO
MILANO 1863

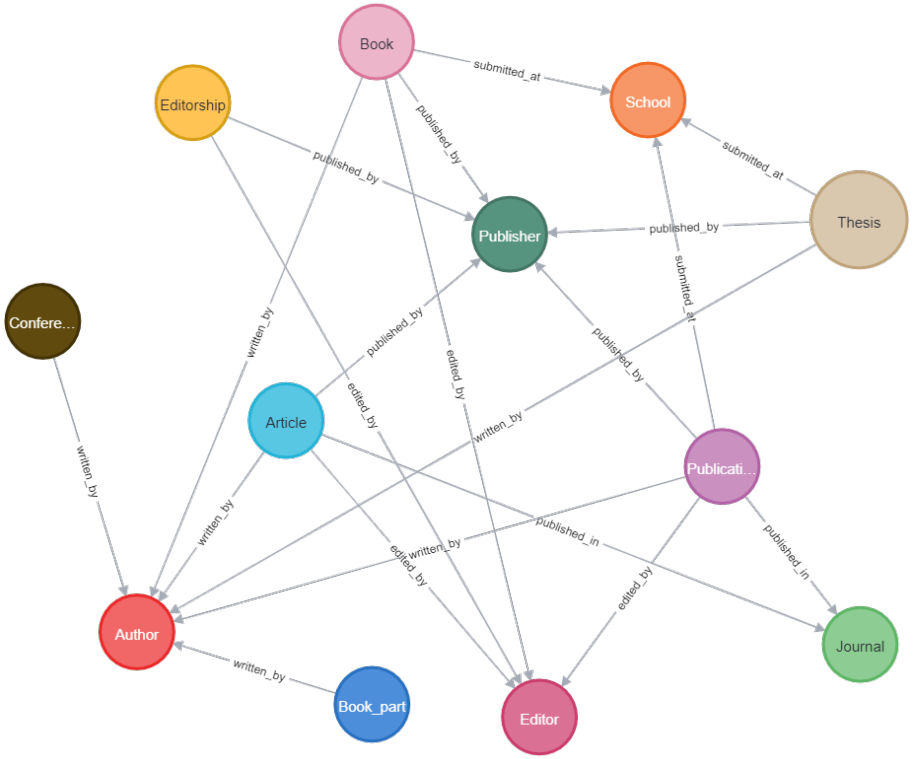
Neo4j

Neo4J : dataset generation and import

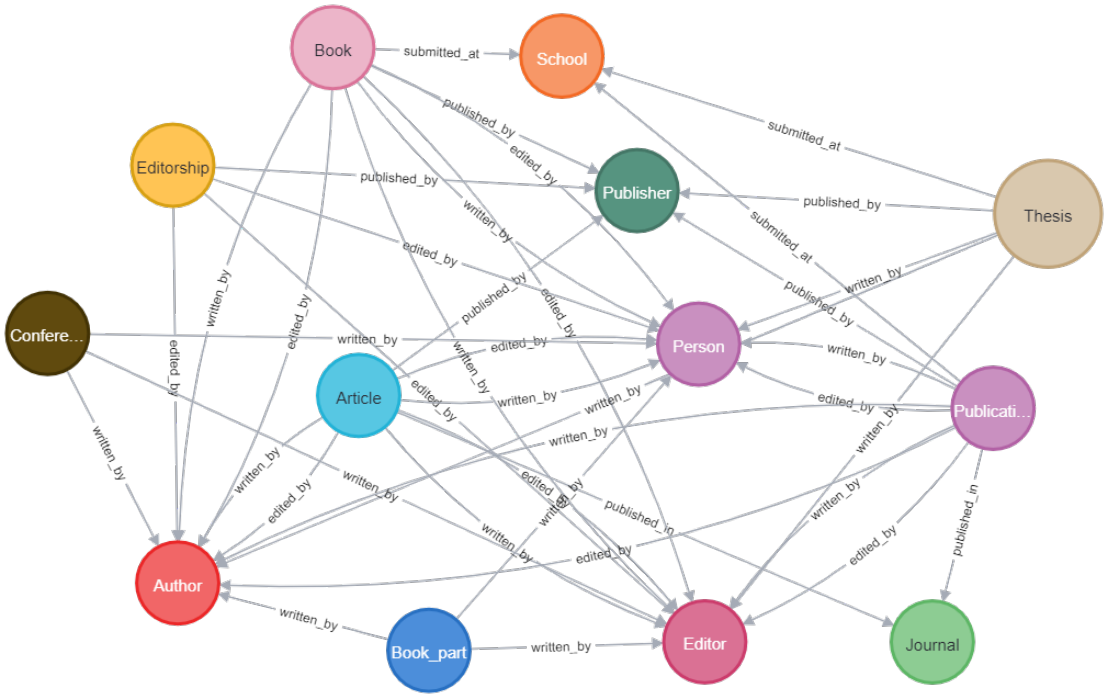


Neo4j: data structure

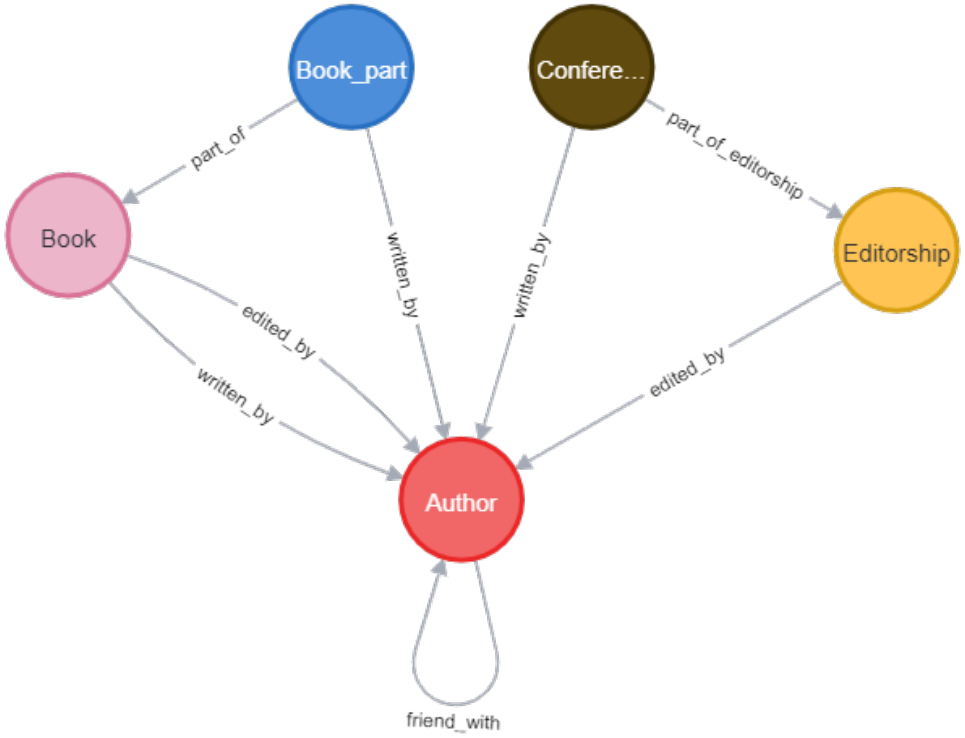
Neo4j: data structure



Neo4j: data structure

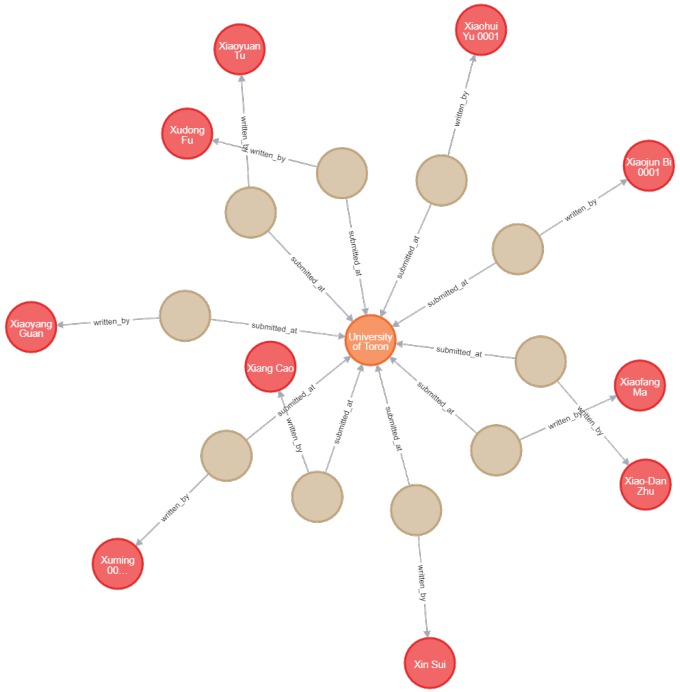


Neo4j: data structure



Neo4j: query

```
1  match (per:Person)<-[]-(pub:Publication)-[s:submitted_at]->(school:School)
2  where per.author starts with 'X'
3  with school, count(s) as submissions, collect(pub) as publications, collect(per) as people
4  return school, people, publications, submissions order by submissions desc limit 1
```



POLITECNICO
MILANO 1863

MongoDB

MongoDB: Dataset generation

MongoDB: Dataset generation

We started by defining the structure of the document

MongoDB: Data structure

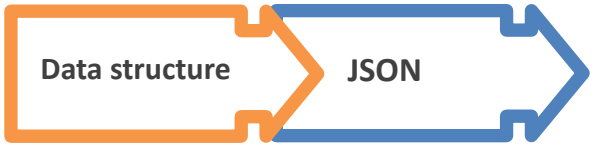

```
0:
  _id:                  "63776ddd297f4328218ebc59"
  title:                "On the Penrose process for rotating black holes"
  authors:
    0:
      name:             "Salinas Bauer"
      affiliation:      "Geekol"
      email:            "salinasbauer@geekol.com"
      bio:              "Veniam ea velit in nulla…trud exercitation sunt."
    1:                  {…}
    2:                  {…}
    3:                  {…}
    4:                  {…}
    5:                  {…}
  keywords:
    0:                  "Information Technology/Informatics"
    1:                  "Research Networking/GRID"
    2:                  "Process automation"
  journal:
    name:               "Optim. Lett."
    volume:             11
    number:             58
    date:               "1987-05-\t\t\t\t\t23"
    pages:              "197 - 388"
  abstract:             "  Penrose described a pr… of 1-1/sqrt{2} =29%.\n"
  sections:
    0:                  {…}
    1:                  {…}
    2:
      title:            "Chain enumeration of $k$…s of classical\n  types"
      paragraphs:
        0:              "Went here for brunch tod… or beer in the future."
      figure:           […]
      subsections:
        0:
          title:        "Green-Tao theorem in function fields"
          paragraphs:   […]
          figure:       {…}
  bibliography:
    0:                  "63776dddcbf3d894a5f8aff3"
    1:                  "63777073d939a5250d1c80ef"
    2:                  "63776ddef403a87eea28accf"
```

MongoDB: Dataset generation

We started by defining the structure of the document

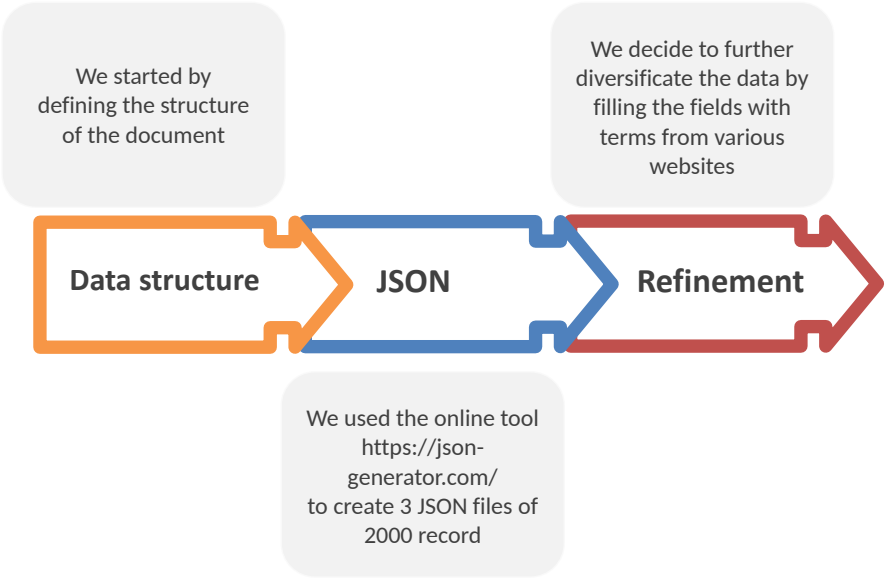
MongoDB: Dataset generation

We started by defining the structure of the document

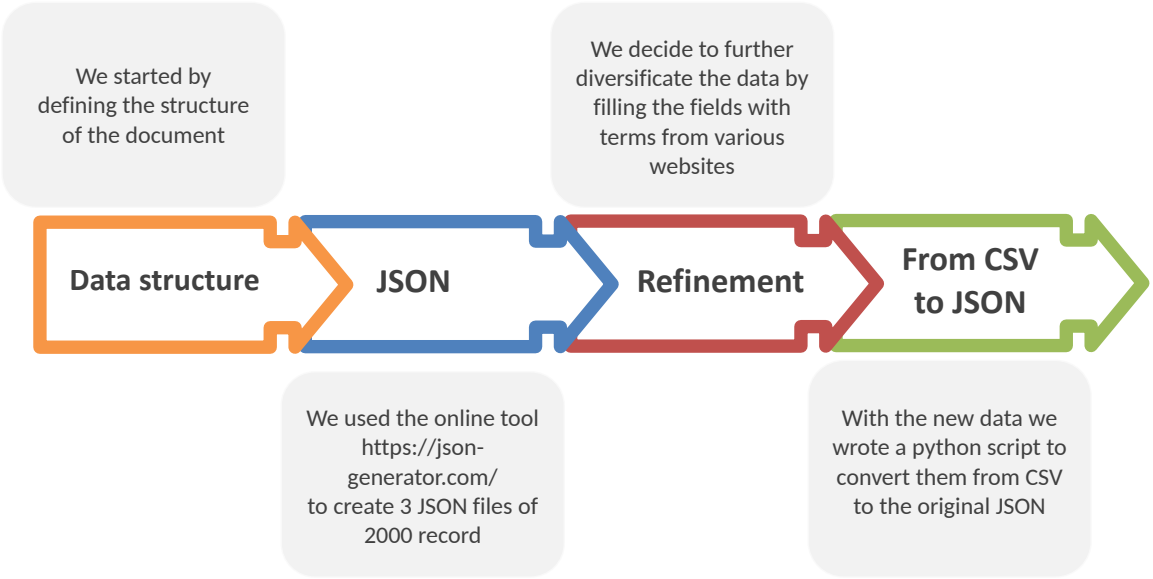


We used the online tool https://json-generator.com/ to create 3 JSON files of 2000 record

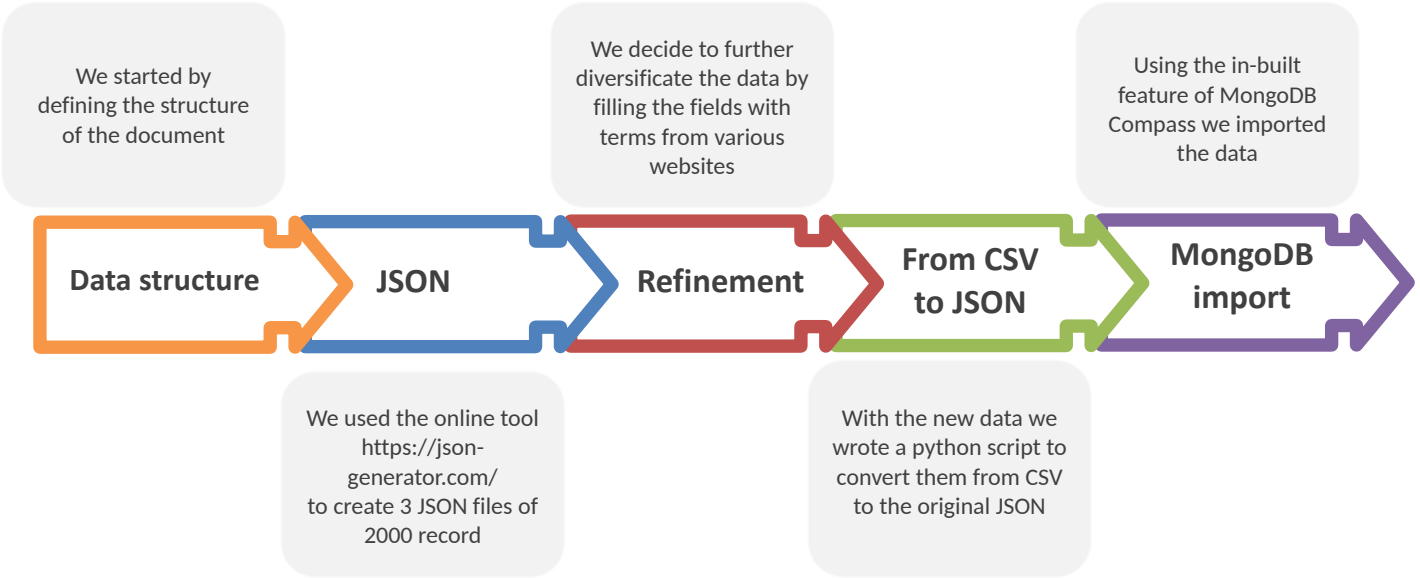
MongoDB: Dataset generation



MongoDB: Dataset generation



MongoDB: Dataset generation



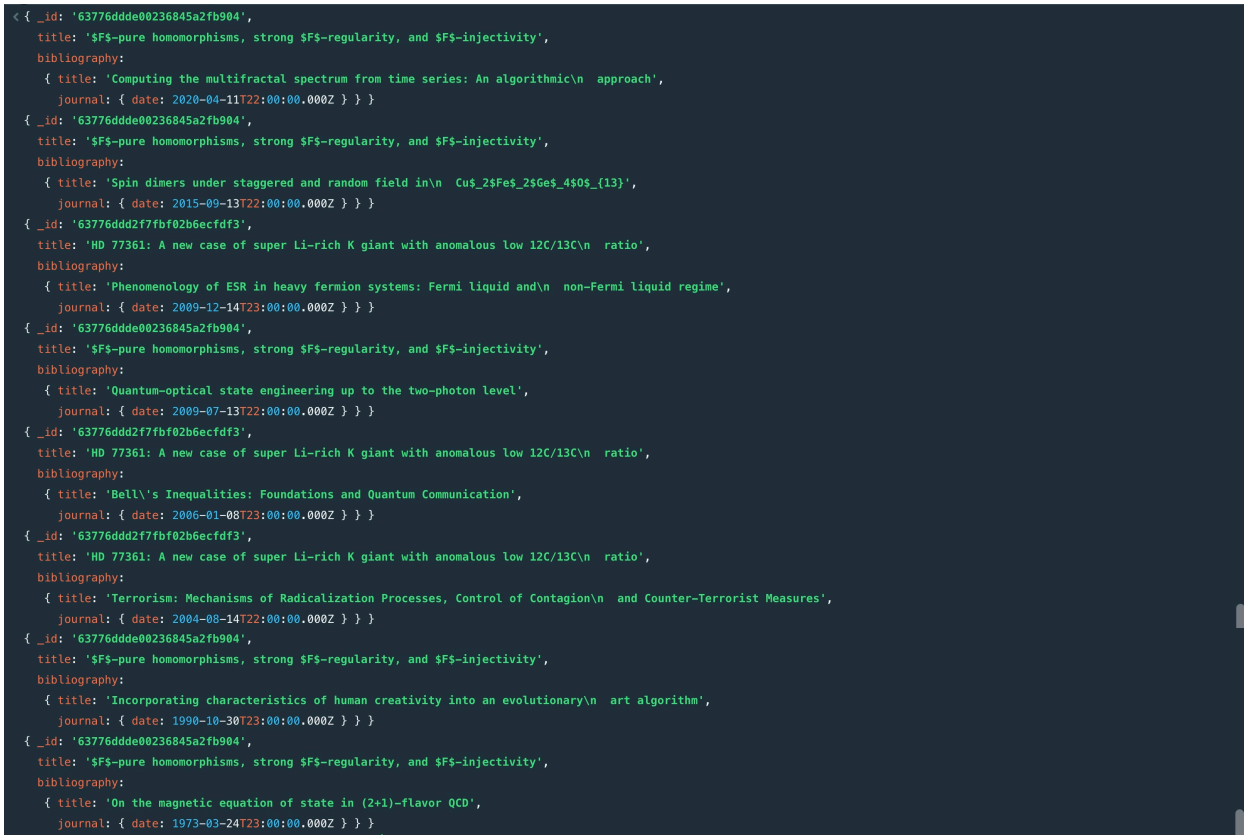
MongoDB: Query

```
Atlas atlas-l7mawr-shard-0 [primary] journal_articles> db.articles.aggregate([
                                                       {
                                                         $match: {$or: [{"_id": ObjectId("63776ddde00236845a2fb904")},{"_id": ObjectId("63776ddd2f7fbf02b6ecfdf3")}]}
                                                       },{
                                                         $lookup: {
                                                         from: "articles",
                                                         localField: "bibliography",
                                                         foreignField: "_id",
                                                         as: "bibliography" }
                                                       },{
                                                         $unwind: "$bibliography"
                                                       },{
                                                         $sort: {"bibliography.journal.date": -1}
                                                       },{
                                                         $project: {"title": 1, "bibliography.title":1, "bibliography.journal.date": 1}
                                                       }])
```

```
{ _id: '63776ddd2f7fbf02b6ecfdf3',
  title: 'HD 77361: A new case of super Li-rich K giant with anomalous low 12C/13C\n  ratio',
  bibliography:
   [ '63777073bc1aeb06de1b7492',
     '63776dde07a64b8cb303d4e9',
     '63776dde6a9de277c1ce0ae1' ] }
{ _id: '63776ddde00236845a2fb904',
  title: '$F$-pure homomorphisms, strong $F$-regularity, and $F$-injectivity',
  bibliography:
   [ '63776dde84264e473ba7c413',
     '63776dde697c3f791c68642d',
     '63776dde6efbe362b4909aba',
     '637770ac8fdfb72787bb6363',
     '63776ddef5d187cd65b423b2' ] }
```

MongoDB: Query

MongoDB: Query

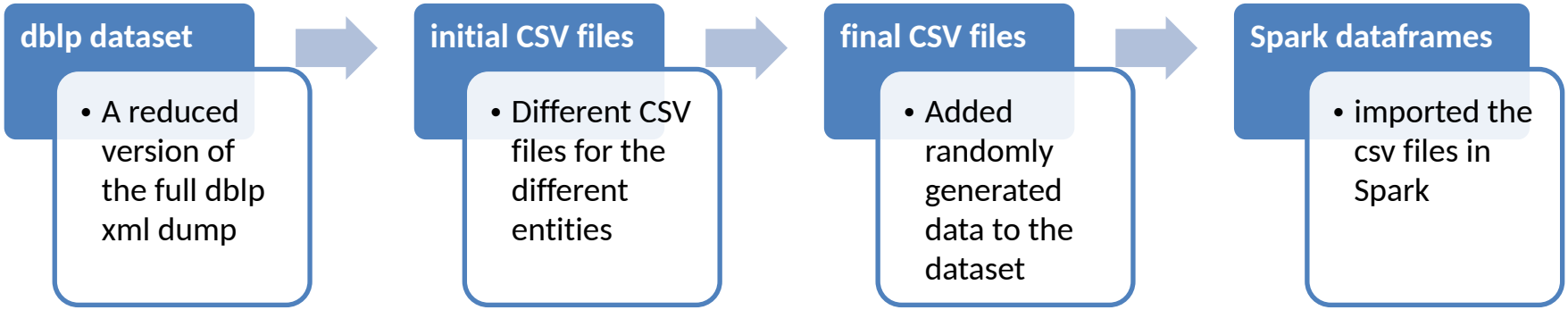


POLITECNICO
MILANO 1863

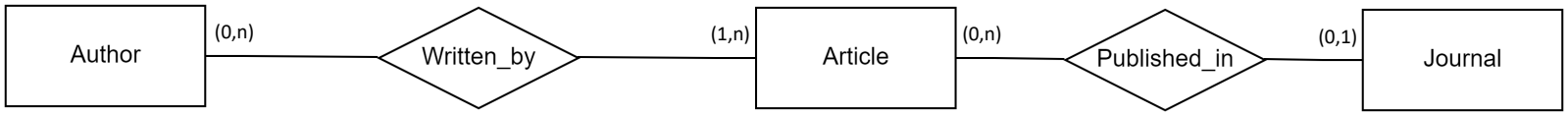
Apache Spark

Spark: dataset generation and import

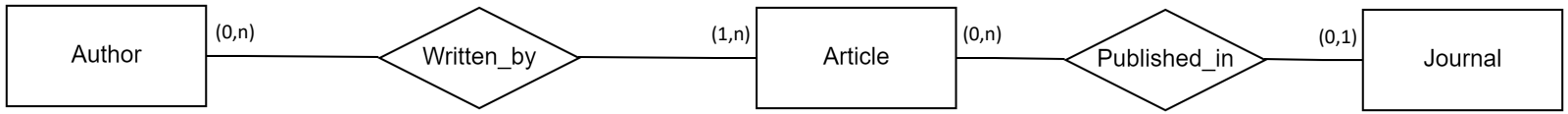
Similar process as for the Neo4j dataset



Spark: data structure



Spark: data structure

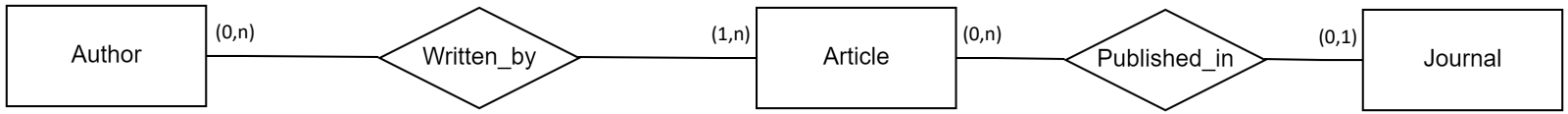



```
root
 |-- ID: integer (nullable = true)
 |-- Name: string (nullable = true)
 |-- BirthDate: date (nullable = true)
 |-- Nationality: string (nullable = true)
```

```
root
 |-- ID: integer (nullable = true)
 |-- Authors: string (nullable = true)
 |-- Journal: integer (nullable = true)
 |-- Key: string (nullable = true)
 |-- Date: date (nullable = true)
 |-- Publisher: string (nullable = true)
 |-- Type: string (nullable = true)
 |-- Title: string (nullable = true)
```

```
root
 |-- ID: integer (nullable = true)
 |-- Name: string (nullable = true)
 |-- FoundationDate: date (nullable = true)
 |-- Type: string (nullable = true)
```

Spark: data structure




```
root
 |-- ID: integer (nullable = true)
 |-- Name: string (nullable = true)
 |-- BirthDate: date (nullable = true)
 |-- Nationality: string (nullable = true)
```

```
root
 |-- Article: integer (nullable = true)
 |-- Author: integer (nullable = true)
```

```
root
 |-- ID: integer (nullable = true)
 |-- Authors: string (nullable = true)
 |-- Journal: integer (nullable = true)
 |-- Key: string (nullable = true)
 |-- Date: date (nullable = true)
 |-- Publisher: string (nullable = true)
 |-- Type: string (nullable = true)
 |-- Title: string (nullable = true)
```

```
root
 |-- ID: integer (nullable = true)
 |-- Name: string (nullable = true)
 |-- FoundationDate: date (nullable = true)
 |-- Type: string (nullable = true)
```

Spark: query

```
authors.filter(col("Name").like("L%")) \
    .join(written_by, written_by.Author == authors.ID, "inner") \
    .groupby(col("Name")).count() \
    .select(col("Name"), col("count").alias("#Articles")) \
    .sort(col("#Articles").desc()) \
    .show(10)
```

```
+--------------+---------+
|          Name|#Articles|
+--------------+---------+
|   Lajos Hanzo|      124|
|    Leo Storme|       41|
|Lie-Liang Yang|       21|
| Lothar Breuer|       17|
|  Lin Cai 0001|       16|
|      Lijun Ji|       16|
|     Lian Zhao|       13|
|   Luca Benini|       13|
|     Luxi Yang|       13|
|Ludger Humbert|       12|
+--------------+---------+
only showing top 10 rows
```

The query returns the Authors with the name that start with an «L», ranked in descending order w.r.t. the number of articles they wrote