

Assignment 5: Query publication data from zbMATH

AI-1 Systems Project (Winter Semester 2022/2023)

Jan Frederik Schaefer

Friedrich-Alexander-Universität Erlangen-Nürnberg, Department Informatik

Topic: Semantic web
Due on: March 28, 2023 (*early track*: January 31, 2023)
Version from: December 30, 2022
Author: Jan Frederik Schaefer

1 Task Summary

Translate a large dataset about mathematical publications into RDF triples and load them into a triplestore. Then use SPARQL queries to answer a set of questions.

Objectives

1. Get hands-on experience with RDF and SPARQL,
2. learn how to find information about standards like RDF and SPARQL,
3. get to know some of the challenges that come from working with large datasets.

Prerequisites and useful methods

1. The XML format and related technologies (in particular, SAX parsing might be useful),
2. RDF, triple stores and SPARQL (the AI lecture introduces these concepts, but there are also plenty of online resources),
3. the basics of HTTP requests (for communicating with the triple store),
4. URLs and percent-encoding (also known as URL encoding).

2 The Dataset

zbMATH [ZBM] collects abstracts and reviews of papers in the area of mathematics and its applications. Recently, a lot of their data was made publicly accessible via an API as well as a dataset [Pet+]. Furthermore, we have created a mini dataset, which contains only a small subset of the entries and might therefore be easier to work with. A download link to the datasets will be posted in the Mattermost channel for this assignment.

The datasets consist of records. Each record corresponds to one publication. The following entries in `metadata/oai_zb_preview:zbmath` are of interest to us:

- `zbmath:document_id` is an identifier for the publication.
- `zbmath:classifications/zbmath:classification` lists the classifications of the publication using the Mathematics Subject Classification [MSC].
- `zbmath:author_ids/zbmath:author_id` lists identifiers for the authors.
- `zbmath:keywords/zbmath:keyword` lists keywords of the publication.
- `zbmath:publication_year` is the year of the publication.

RDF requires that we use URIs to identify resources. These could be anything, but we will use the following URLs for this assignment:

Entry type	Example value	Associated URL
Document id	1448.68463	https://zbmath.org/?q=an%3A1448.68463
Classification	03B10	https://zbmath.org/classification/?q=cc%3A03B10
Author	kohlhase.michael	https://zbmath.org/authors/?q=ai%3Akohlhase.michael
Keyword	semantic web	https://zbmath.org/?q=ut%3Asemantic+web

3 Using Blazegraph

You should translate the datasets into RDF triples and load them into a triple store so that you can run SPARQL queries (you have to solve the problems using SPARQL queries, not using the original dataset). There are many different triple stores and you are free to pick one that works for you. In my (limited) experience, the open source triplestore Blazegraph [BG] is relatively easy to set up and use.

You can start Blazegraph with `java -jar blazegraph.jar`. It then runs on port 9999 (<http://localhost:9999/> gives you access to the blazegraph workbench). It also creates a journal file `blazegraph.jnl` to store the data.

You can update the data with a POST request to `http://localhost:9999/blazegraph/namespace/kb/sparql?update=XYZ` where XYZ is the (URL-encoded) update. For example, `DROP ALL` deletes all triples from the database and `LOAD <file:///path/to/triples.rdf>` loads data from a file.

Similarly, you can send a SPARQL query with a GET request to `http://localhost:9999/blazegraph/namespace/kb/sparql?query=XYZ` where XYZ is the (URL-encoded) SPARQL query.

<pre> <Problems> <Problem id="1" type="..." withquery="true"> ... </Problem> <Problem id="2" type="..." withquery="false"> ... </Problem> ... </Problems> </pre>	<pre> <Solutions> <Solution id="1"> ... <query>...</query> </Solution> <Solution id="2"> ... </Solution> ... </Solutions> </pre>
--	---

Listing 1: Example problems with their solutions. Every problem has an identifier that links it to the solution. The content of the problem and the solution depends on the problem type. For some problems, the SPARQL query that led to the solution should also be included.

4 Problems and Solution Format

Problems are provided in an XML file and the solutions should also be encoded as an XML file. There are separate problem files for the mini dataset and the full dataset. You can find the problem files, along with example problems and solutions at [A5]. Listing 1 illustrates the structure of problems and solutions. Some problems require you to include the used SPARQL query in the solution. The following subsections will describe the different problem types.

4.1 Problem type: coauthors

List all co-authors of someone. This should be fairly straight-forward once you have loaded the data into Blazegraph. Listing 2 shows an example.

4.2 Problem type: msc-intersection

List all publications that have a specific set of classifications. Listing 3 shows an example that looks for publications combining group theory (classification 20) and knowledge representation in artificial intelligence (classification 68T30). One of the publications combining those classifications is 5155089, which has the classifications 68T30, 20F10, 68T37 and 68T05. Note that 20F10 is a subclassification of 20. It might make sense to add the subclassification relation to your dataset.

```

<Problem id="1" type="coauthors" withquery="true">
  <value type="of">https://zbmath.org/authors/?q=ai%3Aeinstein.albert</value>
</Problem>

<Solution id="1">
  <value type="coauthor">https://zbmath.org/authors/?q=ai%3Aschrodinger.erwin</value>
  <value type="coauthor">https://zbmath.org/authors/?q=ai%3Aminkowski.hermann</value>
  ...
  <query>...</query>
</Solution>

```

Listing 2: Example problem and solution solution for problem type coauthors.

```

<Problem id="2" type="msc-intersection" withquery="true">
  <value type="msc">https://zbmath.org/classification/?q=cc%3A20</value>
  <value type="msc">https://zbmath.org/classification/?q=cc%3A68T30</value>
</Problem>

<Solution id="2">
  <value type="paper">https://zbmath.org?q=an%3A5155089</value>
  <value type="paper">https://zbmath.org?q=an%3A647709</value>
  ...
  <query>...</query>
</Solution>

```

Listing 3: Example problem solution solution for problem type msc-intersection.

```

<Problem id="3" type="top-3-keywords" withquery="true">
  <value type="of">https://zbmath.org/authors/?q=ai%3Akohlhase.michael</value>
  <value type="after">1999</value>
  <value type="before">2012</value>
</Problem>

<Solution id="3">
  <value type="keyword">https://zbmath.org?q=ut%3Amathematical+knowledge+management</value>
  <value type="count">4</value>
  <value type="keyword">https://zbmath.org?q=ut%3Asemantics</value>
  <value type="count">2</value>
  <value type="keyword">https://zbmath.org?q=ut%3Acut+elimination</value>
  <value type="count">2</value>
  <query>...</query>
</Solution>

```

Listing 4: Example problem solution solution for problem type top-3-keywords.

4.3 Problem type: top-3-keywords

List the three most common keywords of the publications of an author. There might be additional constraints on the publication year requiring that only consider publications before and/or after a particular year. Listing 4 shows an example.

4.4 Problem type: coauth-dist

Find a minimal connection between two authors X and Y . A connection between X and Y is a sequence of authors a_0, \dots, a_n where $a_0 = X$, $a_n = Y$ and for all $0 \leq i < n - 1$ there is a publication that has both a_i and a_{i+1} as authors.

The length n of such a connection is a generalization of the Erdős number [EN], which measures the co-author distance to mathematician Paul Erdős, who has famously collaborated with very many researchers.

In addition to the authors, you should also list the publications connecting them. Listing 5 shows a minimal connection from Michael Kohlhase to Paul Erdős.

Solving this problem probably requires multiple SPARQL queries.

```

<Problem id="4" type="coauth-dist" withquery="false">
  <value type="from">https://zbmath.org/authors/?q=ai%3Akohlhase.michael</value>
  <value type="to">https://zbmath.org/authors/?q=ai%3Aerdos.pal</value>
</Problem>

<Solution id="4">
  <value type="author">https://zbmath.org/authors/?q=ai%3Akohlhase.michael</value>
  <value type="paper">https://zbmath.org?q=an%3A5576982</value>
  <value type="author">https://zbmath.org/authors/?q=ai%3Adavenport.james-harold</value>
  <value type="paper">https://zbmath.org?q=an%3A3770941</value>
  <value type="author">https://zbmath.org/authors/?q=ai%3Aguy.michael-j-t</value>
  <value type="paper">https://zbmath.org?q=an%3A3643271</value>
  <value type="author">https://zbmath.org/authors/?q=ai%3Aerdos.pal</value>
</Solution>

```

Listing 5: Example problem solution solution for problem type coauth-dist.

5 Submission

At the deadline, we will download a snapshot of your repository. It should contain:

1. All your code.
2. Solution files (both for the mini dataset and the full dataset).
3. The RDF generated for the mini dataset (you can compress it to reduce the file size).
4. A README file explaining how to run your code.
5. A brief summary of how you solved the problem either as a PDF file (\approx 1 page) or as part of your README.md.

6 Random Tips

- Make sure that you understand what RDF is.
- A key challenge is dealing with the large dataset file. One of the problems is that parsing the entire file into a DOM probably won't work for you. Instead, you could consider using a SAX parser. Another option might be to use the streaming capabilities of XSLT 3 (but I haven't tried that myself).
- coauth-dist problems are a bit tricky because the queries might branch out too much.

Consider controlling it a bit using FILTER.

- Have fun :-)

7 Points

The total number of points for this assignment is 100. You can get up to 20 points for the quality of the submission (README, explanation, ...). Furthermore, you get 2 points for every correctly solved problem. There are 20 problems for the mini dataset and 20 problems for the entire dataset, which means that you can get a total of $(20 + 20) \cdot 2 = 80$ points for the solutions. Note that points will be subtracted if the solution files are not in the right format.

If the grading scheme doesn't seem to work well, we might adjust it later on (likely in your favor).

References

- [A5] *Assignment 5: Query math publications*. URL: <https://gitlab.rrze.fau.de/wrv/AlSysProj/ws2223/a5-query-math-publications/assignment>.
- [BG] *Welcome to Blazegraph*. URL: <https://blazegraph.com/> (visited on 02/04/2022).
- [EN] *Erdős number*. URL: https://en.wikipedia.org/wiki/Erd%C5%91s_number (visited on 02/07/2022).
- [MSC] *Mathematics Subject Classification – MSC2020*. URL: <https://zbmath.org/classification/> (visited on 02/04/2022).
- [Pet+] Matteo Petrera et al. “zbMATH Open: API Solutions and Research Challenges”. In: *Submitted to Joint Conference of Digital Libraries 2021 (JCDL '21), September 27–30, 2021, Online, Global*.
- [ZBM] *zbMATH Open*. URL: <https://zbmath.org> (visited on 02/04/2022).