— GUIDE —

Assignment 4: Query publication data from zbMATH

AI-1 Systems Project (Winter Semester 2024/2025)

Jan Frederik Schaefer

Friedrich-Alexander-Universität Erlangen-Nürnberg, Department Informatik

This document is intended to help you solve the assignment "Assignment 4: Query publication data from zbMATH" [AS]. You do not have to read it, but we do recommend to at least take a look at the tips and common issues.

1 A Few 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).
- There are different formats for RDF. The most established is probably RDF-XML. You could also use the more human-readable Turtle syntax or N-Triples (a minimal subset of Turtle), which (in my opinion) are more intuitive.
- When processing the large dataset, it can help to log the progress of your program. This way, you can see if your program is still running or if it's stuck somewhere. It also helps you to estimate how long it will take to process the entire dataset.
- You do not have to de-compress the dataset file and can instead read it directly from the compressed file. E.g. in Python you can use the `bz2.open` from the `bz2` module in the standard library. This can save you a lot of disk space.

2 Common problems

- **Malformed XML in solutions:** Please make sure that your solutions are well-formed XML. In particular, this requires that you escape some characters of your SPARQL query.

- **Running out of memory:** This might happen if you use a DOM parser instead of a SAX parser. It can also happen if you read the entire dataset into memory before running the parser. In general, translating the dataset into an RDF file can work with minimal memory usage: you can process the dataset step by step and write the RDF triples to a file as you go.
- **Slow queries:** If your SPARQL queries are slow, you might have to optimize them. There can be many reasons for this. One common reason is the over-use of FILTER. In that case, try to shift more of the work to the triple patterns.

References

[AS] *Assignment 4: Query publication data from zbMATH*. URL: https://kwarc.info/teaching/AISysProj/WS2425/assignment-1.4.pdf.