

LINQ – Querying DBLP Data

DBLP is an online bibliographical database for computer science containing around 1 million references. Its content is publically available in XML format.

Since this content is more than 400 MB a small excerpt of this data will be used for this exercise.

The DBLP collection follows the BibTeX format and contains the following types of references: article, inproceedings, proceedings, book, incollection, phdthesis, masters-thesis, and www. The fields describing the above types of references are the following: author, editor, title, booktitle, pages, year, address, journal, volume, number, month, url, ee, cdrom, cite, publisher, note, crossref, isbn, series, school, and chapter. Notice that the not all fields are allowed in all reference types; please refer to the DTD file for this information.

Write the LINQ code for the following queries:

1. The list of author names
2. The number of authors
3. The list of author names that are also editors
4. The number of publications by author
5. The authors ordered by the number of publications, in descending order
6. The author(s) having the highest number of publications
7. Give for each author the number of publication types
8. Give for each author the total number of publications and the number of publications by type
9. The list proceedings that have at least one editor that is also author of at least one article in the proceedings
10. Give for each author the number of co-authors and the number of joint publications with each of them
11. For each proceedings give its title and the titles of articles appearing in it
12. Give the transitive co-authors of Frank Manola
13. Give the distance of Frank Manola with respect to other authors. Two authors that write together a publication have distance 0. If an author a write a publication with author b and if author b write a publication with author c, then a is at distance 1 from c if a and c have not published together. If an author a is at distances d_1 and d_2 from an author b, where these distances are obtained following different paths, the minimum value of d_1 and d_2 will be given as answer.