

IASD Data wrangling, Data quality

Homework and Project Instructions

Leonid Libkin

Pierre Senellart

April 22, 2026

1 Description of the homework

For your homework, you have to read in detail one of the articles in the provided list and write a critical report on it: summary of results (reformulated in your own words), discussion on the depth of the work, its impact, interestingness, limitations. This should essentially be what a reviewer of this article would have to do upon submission to a journal or a conference.

In addition, try to also discuss the context of the article: who were the authors and what is their expertise? where and when was it published? has it led to much further work, and, if so, what part of the paper was used? what other work does it rely on?

The entire report should be around 4 pages long.

Practical information

The article can be chosen freely; several students may choose the same article.

The report on the article chosen needs to be submitted by February 26 (23:59) on the Moodle Web site.

2 Description of the final project

For the project, you need to choose a different article from the one you chose for the homework. The project should consist of a practical or theoretical development based on the work of that article. It may be any of the following: implementation work, experimental results (especially for theory papers, showing how algorithms work in practice), extending the results to a different setting or specializing them to a particular case, a comparison with heuristic methods, etc. This list is by no means exclusive; crucially there must be *your own* contribution.

Each student needs to work *individually* on a project. It is fine if several students base their work on the same article, but the work must be completely independent.

The project will be evaluated on the basis of a project defense on March 25 and on a ten-page report that needs to be submitted by April 8. In the presentation and the report, you should focus on your own contribution.

3 List of articles

The references are taken from DBLP (<http://dblp.uni-trier.de>), the main bibliographical source for computer science research. You can search DBLP by authors' names, to find those papers. Once they are found, clicking on the electronic edition icon next to the paper usually gives you access to the source. If for an occasional paper it does not work, try entering the title, in quotes, as well as pdf in a google search, and you should find copies on authors' webpages or websites such as citeseer.

3.1 Information extraction

1. Valter Crescenzi, Giansalvatore Mecca, Paolo Merialdo: RoadRunner: Towards Automatic Data Extraction from Large Web Sites. VLDB 2001: 109-118
2. Arvind Arasu, Hector Garcia-Molina: Extracting Structured Data from Web Pages. SIGMOD Conference 2003: 337-348
3. Bing Liu, Robert L. Grossman, Yanhong Zhai: Mining data records in Web pages. KDD 2003: 601-606
4. Fabian M. Suchanek, Gjergji Kasneci, Gerhard Weikum: YAGO: A Large Ontology from Wikipedia and WordNet. J. Web Semant. 6(3): 203-217 (2008)
5. Michael J. Cafarella, Jayant Madhavan, Alon Y. Halevy: Web-scale extraction of structured data. SIGMOD Record 37(4): 55-61 (2008)

3.2 Data provenance

1. Peter Buneman, Sanjeev Khanna, Wang Chiew Tan: Why and Where: A Characterization of Data Provenance. ICDT 2001: 316-330
2. Todd J. Green, Gregory Karvounarakis, Val Tannen: Provenance semirings. PODS 2007: 31-40
3. J. Nathan Foster, Todd J. Green, Val Tannen: Annotated XML: queries and provenance. PODS 2008: 271-280
4. James Cheney, Laura Chiticariu, Wang Chiew Tan: Provenance in Databases: Why, How, and Where. Foundations and Trends in Databases 1(4): 379-474 (2009)
5. Boris Glavic, Gustavo Alonso: Perm: Processing Provenance and Data on the Same Data Model through Query Rewriting. ICDE 2009: 174-185
6. Yael Amsterdamer, Daniel Deutch, Tova Milo, Val Tannen: On provenance minimization. PODS 2011: 141-152
7. Yael Amsterdamer, Daniel Deutch, Val Tannen: Provenance for aggregate queries. PODS 2011: 153-164
8. Daniel Deutch, Tova Milo, Sudeepa Roy, Val Tannen: Circuits for Datalog Provenance. ICDT 2014: 201-212

9. Antoine Amarilli, Pierre Bourhis, Pierre Senellart: Provenance Circuits for Trees and Treelike Instances. *ICALP (2)* 2015: 56-68
10. Bahareh Sadat Arab, Su Feng, Boris Glavic, Seokki Lee, Xing Niu, Qitian Zeng: GProM - A Swiss Army Knife for Your Provenance Needs. *IEEE Data Eng. Bull.* 41(1): 51-62 (2018)

3.3 Data integration

1. Alon Y. Levy, Alberto O. Mendelzon, Yehoshua Sagiv, Divesh Srivastava: Answering Queries Using Views. *PODS 1995*: 95-104
2. Rachel Pottinger, Alon Y. Halevy: MiniCon: A scalable algorithm for answering queries using views. *VLDB J.* 10(2-3): 182-198 (2001)
3. Oliver M. Duschka, Michael R. Genesereth: Answering Recursive Queries Using Views. *PODS 1997*: 109-116
4. Alon Y. Levy, Anand Rajaraman, Joann J. Ordille: Query-Answering Algorithms for Information Agents. *AAAI/IAAI, Vol. 1* 1996: 40-47
5. Alon Y. Halevy: Theory of Answering Queries Using Views. *SIGMOD Record* 29(4): 40-47 (2000)
6. Xin Luna Dong, Alon Y. Halevy, Cong Yu: Data integration with uncertainty. *VLDB J.* 18(2): 469-500 (2009)
7. Serge Abiteboul, Oliver M. Duschka: Complexity of Answering Queries Using Materialized Views. *PODS 1998*: 254-263
8. Serge Abiteboul, Omar Benjelloun, Ioana Manolescu, Tova Milo, Roger Weber: Active XML: Peer-to-Peer Data and Web Services Integration. *VLDB 2002*: 1087-1090
9. Alin Deutsch, Yannis Katsis, Yannis Papakonstantinou: Determining source contribution in integration systems. *PODS 2005*: 304-315
10. Andrea Cali, Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini: Data integration under integrity constraints. *Inf. Syst.* 29(2): 147-163 (2004)
11. Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini, Riccardo Rosati: Logical Foundations of Peer-To-Peer Data Integration. *PODS 2004*: 241-251
12. Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini: Description Logics for Information Integration. *Computational Logic: Logic Programming and Beyond 2002*: 41-60
13. Andrea Cali, Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini: On the Expressive Power of Data Integration Systems. *ER 2002*: 338-350
14. Maurizio Lenzerini: Data Integration Needs Reasoning. *LPNMR 2001*: 54-61
15. Jayant Madhavan, Shirley Cohen, Xin Luna Dong, Alon Y. Halevy, Shawn R. Jeffery, David Ko, Cong Yu: Web-Scale Data Integration: You can afford to Pay as You Go. *CIDR 2007*: 342-350
16. Shawn R. Jeffery, Michael J. Franklin, Alon Y. Halevy: Pay-as-you-go user feedback for dataspace systems. *SIGMOD Conference 2008*: 847-860
17. Khalid Belhajjame, Norman W. Paton, Alvaro A. A. Fernandes, Cornelia Hedeler, Suzanne M. Embury: User Feedback as a First Class Citizen in Information Inte-

3.4 Data exchange

1. Ronald Fagin, Phokion G. Kolaitis, Lucian Popa: Data exchange: getting to the core. *ACM Trans. Database Syst.* 30(1): 174-210 (2005)
2. Georg Gottlob, Alan Nash: Efficient core computation in data exchange. *Journal of the ACM* 55(2) (2008)
3. Marcelo Arenas, Pablo Barcelo, Ronald Fagin, Leonid Libkin: Solutions and query rewriting in data exchange. *Inf. Comput.* 228: 28-61 (2013)
4. Marcelo Arenas, Pablo Barcelo, Juan L. Reutter: Query Languages for Data Exchange: Beyond Unions of Conjunctive Queries. *Theory Comput. Syst.* 49(2): 489-564 (2011)
5. Ronald Fagin, Phokion G. Kolaitis, Lucian Popa, Wang Chiew Tan: Composing schema mappings: Second-order dependencies to the rescue. *ACM Trans. Database Syst.* 30(4): 994-1055 (2005)
6. Marcelo Arenas, Ronald Fagin, Alan Nash: Composition with Target Constraints. *Logical Methods in Computer Science* 7(3) (2011)
7. Philip A. Bernstein, Todd J. Green, Sergey Melnik, Alan Nash: Implementing mapping composition. *VLDB J.* 17(2): 333-353 (2008)
8. Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, Riccardo Rosati: On reconciling data exchange, data integration, and peer data management. *PODS 2007*: 133-142
9. Ronald Fagin: Inverting schema mappings. *ACM Trans. Database Syst.* 32(4) (2007)
10. Patricia C. Arocena, Boris Glavic, Radu Ciucanu, Renée J. Miller: The iBench Integration Metadata Generator. *PVLDB* 9(3): 108-119 (2015)
11. Bogdan Alexe, Wang Chiew Tan, Yannis Velegrakis: STBenchmark: towards a benchmark for mapping systems. *PVLDB* 1(1): 230-244 (2008)
12. Marcelo Arenas, Jorge Pérez, Juan L. Reutter: Data exchange beyond complete data. *Journal of the ACM* 60(4): 28 (2013)
13. Andre Hernich: Answering Non-Monotonic Queries in Relational Data Exchange. *Logical Methods in Computer Science* 7(3) (2011)
14. Gosta Grahne, Adrian Onet: Representation systems for data exchange. *ICDT 2012*: 208-221
15. Gosta Grahne, Ali Moallemi, Adrian Onet: Intuitionistic Data Exchange. *AMW 2015*
16. Andre Hernich, Leonid Libkin, Nicole Schweikardt: Closed world data exchange. *ACM Trans. Database Syst.* 36(2): 14 (2011)
17. Leonid Libkin, Cristina Sirangelo: Data exchange and schema mappings in open and closed worlds. *J. Comput. Syst. Sci.* 77(3): 542-571 (2011)
18. Marcelo Arenas, Leonid Libkin: XML data exchange: Consistency and query answering. *Journal of the ACM* 55(2) (2008)
19. Shun'ichi Amano, Claire David, Leonid Libkin, Filip Murlak: XML Schema Map-

- pings: Data Exchange and Metadata Management. *Journal of the ACM* 61(2): 12:1-12:48 (2014)
20. Rada Chirkova, Leonid Libkin, Juan L. Reutter: Tractable XML data exchange via relations. *Frontiers of Computer Science* 6(3): 243-263 (2012)
 21. Pablo Barcelo, Jorge Perez, Juan L. Reutter: Schema mappings and data exchange for graph databases. *ICDT 2013*:189-200
 22. Ronald Fagin, Benny Kimelfeld, Phokion G. Kolaitis: Probabilistic data exchange. *Journal of the ACM* 58(4): 15 (2011)

3.5 Incomplete information

1. Tomasz Imielinski, Witold Lipski Jr.: Incomplete Information in Relational Databases. *Journal of the ACM* 31(4): 761-791 (1984)
2. Serge Abiteboul, Paris C. Kanellakis, Gosta Grahne: On the Representation and Querying of Sets of Possible Worlds. *Theor. Comput. Sci.* 78(1): 158-187 (1991)
3. Raymond Reiter: What Should a Database Know? *J. Log. Program.* 14(1&2): 127-153 (1992)
4. Raymond Reiter: A sound and sometimes complete query evaluation algorithm for relational databases with null values. *Journal of the ACM* 33(2): 349-370 (1986)
5. Tomasz Imielinski: Incomplete Deductive Databases. *Ann. Math. Artif. Intell.* 3(2-4): 259-293 (1991)
6. Jan Chomicki, Tomasz Imielinski: Finite Representation of Infinite Query Answers. *ACM Trans. Database Syst.* 18(2): 181-223 (1993)
7. Tomasz Imielinski, Ron van der Meyden, Kumar V. Vadaparty: Complexity Tailored Design: A New Design Methodology for Databases With Incomplete Information. *J. Comput. Syst. Sci.* 51(3): 405-432 (1995)
8. Leonid Libkin: Certain answers as objects and knowledge. *Artificial Intelligence*, Volume 232, March 2016, Pages 1-19
9. Leonid Libkin: SQL's Three-Valued Logic and Certain Answers. *ICDT 2015*: 94-109
10. Amélie Gheerbrant, Leonid Libkin, Cristina Sirangelo: Naive Evaluation of Queries over Incomplete Databases. *ACM Trans. Database Syst.* 39(4): 31:1-31:42 (2014)
11. Amélie Gheerbrant, Leonid Libkin: Certain Answers over Incomplete XML Documents: Extending Tractability Boundary. *Theory Comput. Syst.* 57(4): 892-926 (2015)
12. Pablo Barcelo, Leonid Libkin, Antonella Poggi, Cristina Sirangelo: XML with incomplete information. *Journal of the ACM* 58(1): 4 (2010)
13. Dan Olteanu, Christoph Koch, Lyublena Antova: World-set decompositions: Expressiveness and efficient algorithms. *Theor. Comput. Sci.* 403(2-3):265-284 (2008)
14. Lyublena Antova, Christoph Koch, Dan Olteanu: 10^{10^6} worlds and beyond: efficient representation and processing of incomplete information. *VLDB J.* 18(5):1021-1040 (2009)
15. Robert Fink, Dan Olteanu, Swaroop Rath: Providing support for full relational algebra in probabilistic databases. *ICDE 2011*: 315-326

16. Serge Abiteboul, Benny Kimelfeld, Yehoshua Sagiv, Pierre Senellart: On the expressiveness of probabilistic XML models. *VLDB J.* 18(5): 1041-1064 (2009)
17. Nilesh N. Dalvi, Christopher Re, Dan Suciu: Queries and materialized views on probabilistic databases. *J. Comput. Syst. Sci.* 77(3): 473-490 (2011)

3.6 Inconsistent data; data cleaning

1. Marcelo Arenas, Leopoldo E. Bertossi, Jan Chomicki: Consistent Query Answers in Inconsistent Databases. *PODS 1999*: 68-79
2. Marcelo Arenas, Leopoldo E. Bertossi, Jan Chomicki: Scalar Aggregation in FD-Inconsistent Databases. *ICDT 2001*: 39-53
3. Pablo Barcelo, Leopoldo E. Bertossi: Logic Programs for Querying Inconsistent Databases. *PADL 2003*: 208-222
4. Leopoldo E. Bertossi, Loreto Bravo: Consistent Query Answers in Virtual Data Integration Systems. *Inconsistency Tolerance 2005*: 42-83
5. Jef Wijsen: Database repairing using updates. *ACM Trans. Database Syst.* 30(3): 722-768 (2005)
6. Jef Wijsen: On the consistent rewriting of conjunctive queries under primary key constraints. *Inf. Syst.* 34(7): 578-601 (2009)
7. Jef Wijsen: Certain conjunctive query answering in first-order logic. *ACM Trans. Database Syst.* 37(2): 9 (2012)
8. Gaelle Fontaine: Why Is It Hard to Obtain a Dichotomy for Consistent Query Answering? *ACM Trans. Comput. Log.* 16(1): 7:1-7:24 (2015)
9. Wenfei Fan, Floris Geerts, Jef Wijsen: Determining the currency of data. *PODS 2011*: 71-82
10. Wenfei Fan, Floris Geerts, Xibei Jia, Anastasios Kementsietsidis: Conditional functional dependencies for capturing data inconsistencies. *ACM Trans. Database Syst.* 33(2) (2008)
11. Wenfei Fan, Floris Geerts, Jianzhong Li, Ming Xiong: Discovering Conditional Functional Dependencies. *IEEE Trans. Knowl. Data Eng.* 23(5): 683-698 (2011)
12. Shuai Ma, Wenfei Fan, Loreto Bravo: Extending inclusion dependencies with conditions. *Theor. Comput. Sci.* 515: 64-95 (2014)
13. Leopoldo E. Bertossi, Solmaz Kolahi, Laks V. S. Lakshmanan: Data cleaning and query answering with matching dependencies and matching functions. *ICDT 2011*: 268-279

3.7 Probabilistic databases

1. Omar Benjelloun, Anish Das Sarma, Chris Hayworth, Jennifer Widom: An Introduction to ULDBs and the Trio System. *IEEE Data Eng. Bull.* 29(1): 5-16 (2006)
2. Nilesh N. Dalvi, Dan Suciu: Efficient query evaluation on probabilistic databases. *VLDB J.* 16(4): 523-544 (2007)
3. Nilesh N. Dalvi, Dan Suciu: The dichotomy of conjunctive queries on probabilistic

- structures. PODS 2007: 293-302
4. Serge Abiteboul, Benny Kimelfeld, Yehoshua Sagiv, Pierre Senellart: On the expressiveness of probabilistic XML models. VLDB J. 18(5): 1041-1064 (2009)
 5. Benny Kimelfeld, Yuri Kosharovskiy, Yehoshua Sagiv: Query evaluation over probabilistic XML. VLDB J. 18(5): 1117-1140 (2009)
 6. Dan Olteanu, Jiewen Huang, Christoph Koch: Approximate confidence computation in probabilistic databases. ICDE 2010: 145-156
 7. Abhay Kumar Jha, Dan Olteanu, Dan Suciu: Bridging the gap between intensional and extensional query evaluation in probabilistic databases. EDBT 2010: 323-334
 8. Serge Abiteboul, T.-H. Hubert Chan, Evgeny Kharlamov, Werner Nutt, Pierre Senellart: Capturing continuous data and answering aggregate queries in probabilistic XML. ACM Trans. Database Syst. 36(4): 25:1-25:45 (2011)

3.8 Approximate query answering

1. Yannis E. Ioannidis: Approximations in Database Systems. ICDT 2003: 16-30
2. Surajit Chaudhuri, Bolin Ding, Srikanth Kandula: Approximate Query Processing: No Silver Bullet. SIGMOD Conference 2017: 511-519
3. Surajit Chaudhuri, Gautam Das, Vivek R. Narasayya: A Robust, Optimization-Based Approach for Approximate Answering of Aggregate Queries. SIGMOD Conference 2001: 295-306
4. Pablo Barceló, Leonid Libkin, Miguel Romero: Efficient Approximations of Conjunctive Queries. SIAM J. Comput. 43(3): 1085-1130 (2014)
5. Pablo Barceló, Miguel Romero, Moshe Y. Vardi: Semantic Acyclicity on Graph Databases. SIAM J. Comput. 45(4): 1339-1376 (2016)
6. Leonid Libkin: Certain Answers Meet Zero-One Laws. PODS 2018: 195-207
7. Yang Cao, Wenfei Fan, Tianyu Wo, Wenyuan Yu: Bounded Conjunctive Queries. PVLDB 7(12): 1231-1242 (2014)
8. Yang Cao, Wenfei Fan: Data Driven Approximation with Bounded Resources. PVLDB 10(9): 973-984 (2017)
9. Wenfei Fan, Floris Geerts, Leonid Libkin: On scale independence for querying big data. PODS 2014: 51-62
10. Ronald Fagin, Amnon Lotem, Moni Naor: Optimal aggregation algorithms for middleware. J. Comput. Syst. Sci. 66(4): 614-656 (2003)