



# Denilson Barbosa

Associate Professor  
Department of Computing Science  
University of Alberta

denilson@ualberta.ca

 @ddmbarbosa  <https://www.linkedin.com/in/denilsonbarbosa/>

---

## PERSONAL

### Education

- PhD in Computer Science, University of Toronto, Canada; thesis: “Incremental Validation of XML Documents and Mappings”, supervised by Alberto O. Mendelzon. March 2005.
- MSc in Computer Science, Univ. Federal de Minas Gerais, Brazil; March 1999.
- BSc in Computer Science. Univ. Federal de Pernambuco, Brazil; December 1996.

### Employment History

- Associate Professor, Department of Computing Science, University of Alberta; July 2011–present.
- Assistant Professor and Ingenuity New Faculty, Department of Computing Science, University of Alberta; July 2008–June 2011.
- Assistant Professor, Department of Computer Science, University of Calgary; January 2005–June 2008.

### Honors and Awards

- Supervisor of the winner of the 2017 TextExt - DBpedia Open Extraction Challenge
- Visiting Scientist, Max Planck Institute of Informatics, Saarbrücken, Germany; July 2014–April 2015
- Best Reviews of 2013 Award; ACM Computing Reviews
- Supervisor of the winners the 2012 ACM SIGMOD Undergraduate Student Competition
- Distinguished Service Award in Teaching; Department of Computing Science, University of Alberta (2010)
- Best Paper Award—2010 IEEE International Conference on Data Engineering
- Visiting Professor, Faculty of Computer Science, Free University of Bozen-Bolzano; April 2008–July 2008
- Alberta Ingenuity Fund New Faculty Award (2006)
- IBM Faculty Award (2005)

## ACADEMIC INTERESTS

My research is on knowledge management, information retrieval, and natural language processing, and I work on machine learning methods applied to these fields. I have supervised graduate level research on the problems of named entity recognition, typing and disambiguation; open relation extraction; understanding social processes in Wikipedia article authoring; mining citation networks; and semistructured data management.

## FUNDING SUPPORT AND PARTICIPATION IN NATIONAL PROJECTS

### Major Research Support (shared)

- Co-Principal Investigator of the **Linked Infrastructure for Networked Cultural Scholarship (LINCS)** (2020–2023): CAD\$5,000,000.  
Funded by the Canadian Foundation for Innovation (CFI), provincial matches by Alberta, Ontario and British Columbia, the provinces of Alberta, Ontario and British Columbia, and industry through various contracts, LINCS is a consortium of Canadian Universities and Libraries, with partnerships in Europe and the United States to build an Open Knowledge Graph of Canadian Cultural Heritage materials.
- Lead Researcher in the **Scotiabank Artificial Intelligence Research Initiative** at the University of Alberta (2020-2022): CAD\$360,000  
Multiple projects on explainable AI, customer management, detection of fraudulent transactions.
- Co-Principal Investigator and lead of the Data Quality Theme of the **NSERC Business Intelligence Network** (2009–2014): CAD\$6,000,000  
The network involved 6 Canadian universities (Alberta, Carleton, Dalhousie, Toronto, UBC, and Waterloo) and several industrial partners, including IBM (Cognos team) and SAP (Business Objects team).

### Research Support (individual)

- NSERC Discovery Grant (2005-2022): CAD\$ 387,000
- MITACS Accelerate Grants:
  - Analytica Advisors (2020-2021), CAD\$45,000; “COVID19 and Disclosure of Risk in Energy Exposed Sectors”.
  - The Canadian Energy and Climate Nexus (CECN) (2020-2021), CAD\$100,000 (33%); “A Toolkit for Analyzing Online Conversations for Solutions Based Policy Development”.
  - MatchWorks (2028-2019), CAD\$30,000; “Data Driven Assessment for Self-Determined Employment”.
  - Granify (2018–2019), CAD\$75,000 (33%); “Semantic Information Processing to Improve e-Commerce Shopper Experience”.
- NSERC Engage Grant: CAD\$25,000 each
  - Bioware/EA Games (2019): Text Analysis for Understanding Gamer Social Behavior.
  - HireGround Software Solutions (2017): Deep Learning and Word Embeddings for Information Retrieval.
  - Kira Systems Inc (2016): Document Structure Driven Named Entity Recognition.
  - Apropos Information Systems (2013): Information Extraction on Indigenous Knowledge.
- Other
  - Grant from Amazon AWS (2019), USD\$40,000.
  - Grant from Servus Credit Union (2018): CAD\$50,000; partnership development and initial modeling of AI solution for identifying customer profiles.
  - Gift from **DiffBot.com** (2017), USD\$55,000.
  - Gift from **NVIDIA** (2017), CAD\$2,000.
  - IBM Alberta Centre for Advanced Studies (2013–2014), CAD\$40,000.
  - Gift from **GE Intelligent Systems** (2011), CAD\$10,000.
  - **Alberta Ingenuity New Faculty Award** (2007–2010), CAD\$294,000.

- Start-up grant at the University of Alberta (2008–2012), CAD\$40,000.
- **IBM Faculty Award** (2006), CAD\$11,500.
- Start-up grant at the University of Calgary (2005–2006), CAD\$50,000.

**Leadership in other Major Research Initiatives** – **CWRC–The Canadian Writing Research Collaboratory**. Project Leader since 2010 and Chair of the Technical Board since 2011.  
 CWRC is funded through the Canadian Foundation for Innovation (CAD\$1.5M).  
<http://www.cwrc.ca/>

## FULLY REFEREED PUBLICATIONS

Trainees under my supervision highlighted in bold.

- Books**
- R. T. Ng, P. C. Arocena, D. Barbosa, G. Carenini, L. Gomes-Jr, S. Jou, R. A. Leung, E. Milios, R. J. Miller, J. Mylopoulos, R. A. Pottinger, F. Tompa, E. Yu. *Perspectives on Business Intelligence*. Synthesis Lectures; Morgan and Claypool, 2013. DOI:10.2200/S00491ED1V01Y201303DTM034.
- Book Chapters**
- **D. Serrano**, E. Stroulia, D. Barbosa, and V. Guana. “SociQL: A Query Language for the Social Web”. In *Advances in Network Analysis and its Applications* (E. Kranakis, ed.). 30 pp. Mathematics in Industry; Springer Verlag, 2013.
- Journal Articles**
- **Z. Guo**, D. Barbosa. “Robust Named Entity Disambiguation with Random Walks”. *Semantic Web*. Preprint, pp. 1-21, 2017. DOI: 10.3233/SW-170273.
  - **H. Sepehri-Rad**, D. Barbosa. “Identifying Controversial Wikipedia Articles using Editor Collaboration Networks”. *ACM Trans. on Inf. Sys. and Tech.* 6(1): 5:1-5:24, 2015.
  - **M. Moraes**, C. Heuser, V. Moreira, D. Barbosa. “Pre-Query Discovery of Domain-specific Query Forms: A Survey”. *IEEE Trans. on Data and Knowledge Engineering*, 25(8); pp. 1830-1848, 2013.
  - **Y. Merhav**, **F. Mesquita**, D. Barbosa, W. G. Yee, O. Frieder. “Extracting Information Networks from the Blogosphere”. *ACM Transactions on the Web*, 6(3): 33 pp., 2012.
  - N. Augsten, D. Barbosa, M. Böhlen, and T. Palpanas. “Efficient Top-k Approximate Subtree Matching in Small Memory”. *IEEE Trans. on Data and Knowledge Engineering*, 23 (August 2011): 1123-1137, 2011.
  - **G. Leighton** and D. Barbosa. “Access Control Policy Translation, Verification, and Minimization within Heterogeneous Data Federations”. *ACM Trans. Inf. Syst. Secur.* 14(3): 27 pp., 2011.
  - D. Barbosa, and A. O. Mendelzon. “Declarative Generation of Synthetic XML Data.” *Software: Practice and Experience* 36(10), pp. 1051-1079, 2006. John Wiley & Sons.
  - D. Barbosa, L. Mignet, and P. Veltri. “Studying the XML Web: Gathering Statistics from an XML Sample”. *World Wide Web* 8(4), pp. 413-438, December 2005, Springer Science + Business Media.
- In Proceedings (Chronological)**
- **V. Ranganathan**, S. Suresh, Y. Mathur, N. Subramanyam, and D. Barbosa. “Grcluster: a score function to model hierarchy in knowledge graph embeddings”. In *SAC '20: The 35th ACM/SIGAPP Symposium on Applied Computing*, pp. 964–971. ACM, 2020.
  - **C. Pang**, A. Hindle, and D. Barbosa. “Understanding DevOps education with grounded theory.”. In *Proceedings of the 41st International Conference on Software Engineering: Software Engineering Education and Training.*, pp. 8. IEEE, 2020.

- F. Mesquita, M. Cannaviccio, J. Schmidek, P. Mirza, and D. Barbosa. “KnowledgeNet: a benchmark dataset for knowledge base population”. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp 749–758. ACL, 2019.
- **P. Xu** and D. Barbosa. “Connecting language and knowledge with heterogeneous representations for neural relation extraction”. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp 3201–3206. ACL, 2019.
- **D. Caminhas, D. Cones, N. Hervieux**, and D. Barbosa. “Detecting and correcting typing errors in DBpedia”. In *Proceedings of the 1st International Workshop on Challenges and Experiences from Data Integration to Knowledge Graphs co-located with the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Volume 2512 of CEUR Workshop Proceedings. CEUR-WS.org, 2019. URL: <http://ceur-ws.org/Vol-2512/paper6.pdf>.
- **M. Cannaviccio**, D. Barbosa, and P. Merialdo. “Towards Annotating Relational Data on the Web with Language Models”. In *Proc. of the 2018 World Wide Web Conference*. pp 1307-1316. ACM, 2018.
- **P. Xu**, and D. Barbosa. “Neural Fine-Grained Entity Type Classification with Hierarchy-Aware Loss”. In *Proc. of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 10 pp. ACL, 2018.
- **M. Cannaviccio**, L. Ariemma, D. Barbosa, and P. Merialdo. “Leveraging Wikipedia Table Schemas for Knowledge Graph Augmentation”. In *Proc. of the 21st Int. Workshop on Web and Databases*. 9 pp. ACM, 2018.
- **P. Xu**, and D. Barbosa. “Matching Résumés to Job Descriptions with Stacked Models”. In “Proc. of the 31st Canadian Conference on Artificial Intelligence”. 6 pp. Springer, 2018.
- **M. Cannaviccio**, D. Barbosa, and P. Merialdo. “Lector+ Entry for the TextExt - DBpedia Open Extraction Challenge”. In *CEUR-WS.org - CEUR Workshop Proceedings*, 2017.
- M. Yahya, D. Barbosa, K. Berberich, Q. Wang, and G. Weikum. “Relationship Queries on Extended Knowledge Graphs”. In *Proceedings of the 9th ACM International Conference on Web Search and Data Mining*. ACM, 2016.
- **M. Cannaviccio**, D. Barbosa, and P. Merialdo. “Accurate fact harvesting from natural language text in wikipedia with Lector”. In *Proceedings of the 19th International Workshop on Web and Databases*, San Francisco, CA, USA, June 26, 2016, 9. ACM, 2016. doi:10.1145/2932194.2932203.
- **I. Polato**, D. Barbosa, A. Hindle, and F. Kon. “Hadoop energy consumption reduction with hybrid HDFS”. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, Pisa, Italy, April 4-8, 2016, 406–411. ACM, 2016. doi:10.1145/2851613.2851623.
- D. Barbosa, H. Wang, and C. Yu. “Inferencing in information extraction: techniques and applications”. In *Proceedings of the IEEE 31st International Conference on Data Engineering*. IEEE, 2015.
- **I. Polato**, D. Barbosa, A. Hindle, and F. Kon. “Hadoop Branching: Architectural Impacts on Energy and Performance”. In *Sixth International Green and Sustainable Computing Conference, IGSC 2015*, Las Vegas, NV, USA, December 14-16, 2015, 1–4. IEEE Computer Society, 2015.
- **Z. Guo** and D. Barbosa. “Entity recognition and linking on tweets with random walks”. In *Proc. of the the 5th Wkshp. on making sense of microposts co-located w/ 24th WWW Conf.*, pp. 57–58. ACM, 2015.
- J. Fagnan, O. R. Zaiane, and D. Barbosa. “Using triads to identify local community structure in social networks”. In *2014 IEEE/ACM Conf. on Adv. in Soc. Net. Analysis and Mining*, pp. 108–112, 2014.
- **Z. Guo** and D. Barbosa. “Robust Entity Linking via Random Walks”. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pp. 499-508. ACM, 2014.

- **Y. A. Sekhavat, F. di Paolo**, D. Barbosa and P. Merialdo. “Knowledge Base Augmentation using Tabular Data”. In 7th Workshop on Linked Data on the Web. CEUR-WS.org, 2014.
- **Z. Guo** and D. Barbosa. “Entity linking with a unified semantic representation”. In *23rd International World Wide Web Conference, Companion Volume*, pp. 1305-1310. ACM, 2014
- **J. Schmidek**, D. Barbosa. “Improving Open Relation Extraction via Sentence Re-Structuring.” In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA), Reykjavik, Iceland, 2014.
- **F. Mesquita, J. Schmidek**, D. Barbosa. “Effectiveness and Efficiency of Open Relation Extraction”. In *EMNLP 2013: Conference on Empirical Methods in Natural Language Processing*. pp 447–457. ACL, 2013.
- **M. Moraes**, C. A. Heuser, V. Moraes, D. Barbosa. “Automatically Training Form Classifiers”. In *Proceedings of the 14th International Conference on Web Information System Engineering*. 14 pp. Springer, 2013.
- **H. He**, D. Barbosa, and G. Kondrak. “Identification of Speakers in Novels”. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1312-1320. Association for Computational Linguistics, 2013.
- **Y. Xu**, M.-Y. Kim, K. Quinn, R. Goebel, and D. Barbosa. “Open Information Extraction with Tree Kernels”. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 12 pp. ACL, 2013.
- **H. Sepehri-Rad**, and D. Barbosa. “Identifying controversial articles in Wikipedia: A comparative study”. In *7th International Symposium on Wikis and Open Collaboration*, 10 pp.; ACM, 2012.
- **H. Sepehri-Rad**, A. Makazhanov, D. Rafiei and D. Barbosa. “Leveraging editor collaboration patterns in wikipedia”. In *Proceedings of the 23rd ACM conference on Hypertext and social media*, pp. 13-22. ACM, 2012.
- **S. Husby** and D. Barbosa. “Topic Classification of Blog Posts Using Distant Supervision”. In *Proceedings of the Workshop on Semantic Analysis in Social Media*, pp. 28-36. Association for Computational Linguistics, 2012.
- **M. Bronzi, Z. Guo, F. Mesquita**, D. Barbosa and P. Merialdo. “Automatic Evaluation of Relation Extraction Systems on Large-scale”. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, pp. 19-24. Association for Computational Linguistics, 2012.
- **T. Sliwkanich, D. Schneider, A. Yong, M. Home**, and D. Barbosa. “Towards scalable summarization and visualization of large text corpora (abstract only)”. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 863, 2012. **(Best Undergraduate Poster Award)**
- **F. Mesquita** and D. Barbosa. “Extracting Meta Statements from the Blogosphere”. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 2011. (11% acceptance rate.)
- C. E. Pires, P. Vieira, M. Saraiva and D. Barbosa. “Generating Synthetic Database Schemas for Simulation Purposes”. In *Proceedings of the 22nd International Conference on Database and Expert Systems Applications*, Springer-Verlag, 2011.
- **H. Sepehri-Rad** and D. Barbosa. “Towards identifying arguments in Wikipedia pages”. In *Proceedings of the 20th International Conference on World Wide Web*, pp. 117-118, ACM, 2011.
- N. Augsten, D. Barbosa, M. Böhlen, and T. Palpanas. “TASM: Top-k Approximate Subtree Matching”. In *26th International Conference on Data Engineering (ICDE)*; pages 353-364, IEEE Computer Society, 2010 (13% acceptance rate, **Best Paper Award**).
- **G. Leighton** and D. Barbosa. “Access Control Policy Translation and Verification within Heterogeneous Data Federations”. In *Proceedings of the 15th ACM Symposium on Access Control Models and Technologies*, pp. 173–182, ACM, 2010 (23% acceptance rate).

- A. Celikyilmaz, D. Hakkani-Tur, **H. He**, G. Kondrak and D. Barbosa. “The Actor-Topic Model for Extracting Social Networks in Literary Narrative”. In *Proceedings of the 2010 Neural Information Processing Systems in Machine Learning for Social Computing Workshop* (NIPS 2010 - MLSC), 7 pages. December 2010. (45.8% = 11/24 acceptance rate)
- **F. Mesquita**, **Y. Merhav** and D. Barbosa. “Extracting Information Networks from the Blogosphere: State-of-the-Art and Challenges”. In *4th Int’l AAAI Conference on Weblogs and Social Media–Data Challenge Track*, 8 pages, AAAI, 2010.
- **P. Cappellari**, D. Barbosa and P. Atzeni. “A Framework for Automatic Schema Mapping Verification Through Reasoning”. In *Proceedings of the 2010 International Workshop on Data Engineering meets the Semantic Web* (in conjunction with ICDE), pages 245-250, IEEE Computer Society, 2010 (37% acceptance rate).
- **M. Hasan**, E. Stroulia, D. Barbosa, M. Alalfi. “Analyzing Natural-Language Artifacts of the Software Process”. In *26th IEEE International Conference on Software Maintenance* (Early Results Track), 5 pages, IEEE Computer Society, 2010 (41% acceptance rate).
- **Y. Merhav**, **F. Mesquita**, D. Barbosa, W. G. Yee, O. Frieder. “Incorporating Global Information into Named Entity Recognition Systems using Relational Context”. In *33rd annual ACM SIGIR conference*, 2 pages, ACM, 2010 (30% acceptance rate).
- **V. Ganev**, **Z. Guo**, **D. Serrano**, B. Tansey, D. Barbosa, E. Stroulia. “Exploring and Visualizing Academic Social Networks”. In *19th ACM International Conference on Information and Knowledge Management* (CIKM), 2 pages, ACM, 2010.
- **V. Ganev**, **Z. Guo**, **D. Serrano**, B. Tansey, D. Barbosa, E. Stroulia. “An environment for building, exploring and querying academic social networks”. In *International ACM Conference on Management of Emergent Digital EcoSystems*, pages 282-289, ACM, 2009 (40% acceptance rate).
- **G. Leighton**, D. Barbosa. “Optimizing XML Compression”. In *Database and XML Technologies, 6th International XML Database Symposium*, pages 91-105, Springer-Verlag, 2009.
- **P. Cappellari** and D. Barbosa. “Towards Automatic Schema Mapping Verification Through Reasoning”. In *International Workshop on Logic in Databases*, pages 43-56, Rotschild University, 2009.
- A. S. da Silva, D. Barbosa, J. M. B. Cavalcanti, and **M. Sevalho**; “Labeling Data Extracted from the Web”. In *Proceedings of the 6th International Conference on Ontologies, DataBases, and Applications of Semantics*; pages 1099-1116, Springer, 2007. (23% acceptance rate).
- **F. Mesquita**, D. Barbosa, **E. Cortez**, and Altigran S. da Silva; “FleDEx: Flexible Data Exchange”. In *Proceedings of the 9th ACM International Workshop on Web Information and Data Management*; pages 25-32, ACM, 2007. (25% acceptance rate).
- **J. Park**, and D. Barbosa. “Adaptive Record Extraction From Web Pages”. In *Proceedings of the 16th International Conference on World Wide Web* (WWW), pages 1335-1336, ACM, 2007.
- D. Barbosa, **G. Leighton**, and **A. Smith**. “Efficient Incremental Validation of XML Documents After Composite Updates”. In *Proceedings of the Fourth International XML Database Symposium*, pages 107-121. Springer-Verlag, 2006.
- D. Barbosa, J. Freire, and A. O. Mendelzon. “Designing Information-Preserving Mapping Schemes for XML”. In *Proceedings of the 31st International Conference on Very Large Databases* (VLDB); pages 109-102. ACM, 2005. (16% acceptance rate).
- M. P. Consens, D. Barbosa, **A. M. Teisanu**, and L. Mignet. “Goals and Benchmarks for Autonomic Configuration Recommenders”. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data* (SIGMOD); pages 239-250. ACM, 2005. (15% acceptance rate.)
- D. Barbosa, A. O. Mendelzon, L. Libkin, L. Mignet, and M. Arenas. “Efficient Incremental Validation of XML Documents”. In *Proceedings of the 20th International Conference on Data Engineering* (ICDE); pages 671-682. IEEE Computer Society, 2004, (14% acceptance rate).

- D. Barbosa, J. Freire, and A. O. Mendelzon. “Information Preservation in XML-to-Relational Mappings”. *Lecture Notes in Computer Science*, vol 3186 (Database and XML Technologies: Proceedings of the Second International XML Database Symposium, Toronto, Canada), pages 66-81. Springer-Verlag, 2004.
- D. Barbosa and A. O. Mendelzon. “Finding ID Attributes in XML Documents”. *Lecture Notes in Computer Science*, vol 2824 (Database and XML Technologies: Proceedings of the First International XML Database Symposium, Berlin, Germany), pages 180-194. Springer-Verlag, 2003.
- L. Mignet, D. Barbosa and P. Veltri. “On the Characterization of the XML Web (extended abstract).” In *Proceedings of the Eleventh Italian Symposium on Advanced Database Systems*. Cetraro, Italy, 2003.
- L. Mignet, D. Barbosa and P. Veltri. “The XML web: a first study”. In *Proceedings of the 12th international conference on World Wide Web (WWW)*; pages 500-510. ACM, 2003. (13% acceptance rate).
- D. Barbosa, A. O. Mendelzon, J. Keenleyside and K. Lyons. “ToXgene: a template-based data generator for XML”. In *Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data*, pp. 616, 2002, ACM.
- D. Barbosa, A. O. Mendelzon, J. Keenleyside and K. Lyons. “ToXgene: a template-based data generator for XML”. In *Proceedings of the Fifth International Workshop on the Web and Databases*; pages 49-54. Madison, WI, USA, 2002.
- D. Barbosa, A. Barta, A. O. Mendelzon, G. Mihaila, F. Rizzolo and P. Rodriguez-Gianolli. “ToX - the Toronto XML Engine”. In *Proceedings of the Workshop on Information Integration on the Web*. Rio de Janeiro, Brazil, 2001.
- D. Barbosa, J. P. Kitajima and W. Meira Jr. “Parallelizing MPEG Video Encoding using Multiprocessors”. In *Proceedings of the XII Brazilian Symposium on Computer Graphics and Image Processing*. Campinas, SP, Brazil. IEEE Computer Society, 1999.
- D. Barbosa, J. P. Kitajima and W. Meira Jr. “Real-time MPEG Encoding in Shared-Memory Multiprocessors”. In *Proceedings of the 2nd International Conference on Parallel Computing Systems*. Ensenada, Baja California, Mexico, 1999.
- A. Macêdo, E. S. Silva, D. Barbosa, M. A. Cristo, J. P. Kitajima, B. Ribeiro-Neto, G. Navarro and N. Ziviani. “Experimental Analysis of a Parallel Quicksort-based Algorithm for Suffix Array Generation”. In *Proceedings of the 3rd International Meeting on Vector and Parallel Processing*. Porto, Portugal, 1998.

## OTHER PUBLICATIONS

### Lightly Refereed and/or Invited

- D. Barbosa, P. Bohannon, J. Freire, C.-C. Kanne, I. Manolescu, V. Vassalos, M. Yoshikawa. “XML Storage”. In *Encyclopedia of Database Systems*, pages 3627-3634, Springer-Verlag, 2009.
- D. Barbosa, I. Manolescu, J. X. Yu. “XML Benchmarks”. In *Encyclopedia of Database Systems*, pages 3576-3579, Springer-Verlag, 2009.
- D. Barbosa, I. Manolescu, J. X. Yu. “Application Benchmark”. In *Encyclopedia of Database Systems*, pages 99-100, Springer-Verlag, 2009.
- D. Barbosa, I. Manolescu, J. X. Yu. “Microbenchmark”. In *Encyclopedia of Database Systems*, page 1737, Springer-Verlag, 2009.
- D. Barbosa, A. Bonifati, Z. Bellahsene, E. Hunt, R. Unland (Eds.): “Database and XML Technologies, 5th International XML Database Symposium, XSym 2007, Vienna, Austria, September 23-24, 2007, Proceedings”. *Lecture Notes in Computer Science 4704* Springer 2007, ISBN 978-3-540-75287-5.
- D. Barbosa, **G. Elliott**; “Generating Synthetic XML Documents by Example”. Demo at the *2007 Conference of the IBM Center for Advanced Studies on Collaborative research*, 2007.

- D.Barbosa. “On the World Wide Web, XML and Open Hypermedia Systems”. CITO Knowledge Network Conference. October 2-3, 2000.

### Technical Reports

- **G. Leighton**, D. Barbosa. “Optimizing XML Compression”. CoRR abs/0905.4761, 2009.
- **F. Mesquita**, D.Barbosa, **E. Cortez**, and A. S. da Silva. “A Lightweight Framework for Exchanging Web Data”. Technical Report 2007-876-28, University of Calgary; July, 2007.
- D.Barbosa, M. P. Consens, L. Mignet. “Experimental Evaluation of Autonomic Indexing”. Technical Report CSRG-495, University of Toronto, July 2004.
- D.Barbosa and A. O. Mendelzon. “CASCON 2003 Workshop Report–XML Query Processing”. IBM Technical Report TR-74.193-21, January 2004.
- D.Barbosa. “ToXgene Template Specification Language”. Technical Report CRG-474. University of Toronto, July 2003.
- D.Barbosa and A. Mendelzon. “CASCON 2002 Workshop Report–XML Data Management”. IBM Technical Report TR-74.188-14, January 2003.

### Theses

- “Incremental Validation of XML Documents and Mappings”. Ph.D. Thesis, Department of Computer Science, University of Toronto, 2005, 149 pages, ISBN:0-494-02722-3. Advisor: Alberto O. Mendelzon (Ph.D., 1979, Princeton University).
- “Compressão de Vídeo Digital MPEG em Paralelo”. M.Sc. Thesis (in Portuguese), Departamento de Ciência da Computação, Universidade Federal de Minas Gerais, 1999. Advisor: João Paulo Kitajima (Ph.D., 1994 Institute National Polytechnique de Grenoble).

### SOFTWARE AND OTHER SCHOLARLY WORKS

- **PRIMA** (<https://github.com/U-Alberta/prima>) Personal Research Management for the Internet Archive. PRIMA is an open source suite of information retrieval algorithms implemented in Python and integrated with the Internet Archive’s own toolkit for downloading and contributing collections.
- **EXEMPLAR** (<https://github.com/U-Alberta/exemplar>). EXEMPLAR is an open relation extraction system originating from our work presented at the 2013 EMNLP conference paper. EXEMPLAR takes text files as input and extracts relations with two or more arguments.
- **Walking NED** (<https://github.com/U-Alberta/wned>). WNED is an random walk-based named entity disambiguation systems that accepts text with marked named entities and links them to their referent entities in a knowledge base (e.g. Wikipedia).
- **Lector+** (<https://github.com/miccia4/LectorPlus>) Lector is an extraction tool originated from a joint research project between Roma Tre University and University of Alberta. The tool is able to extract facts from English Wikipedia article text, learning the expressions (i.e. phrases) that are commonly used to describe instances of relations between named entities in the text. It reaches an estimated precision of 95%.
- **Entity Miner** (<http://durville.nis.ualberta.ca:8081/>): online browsable and searchable database of named entities and relations among them, automatically extracted from online text.
- **ReaSoN** (<http://hypatia.cs.ualberta.ca/reason/>): a portal for visualizing, exploring, querying, and integrating academic social networks, extracted from bibliographic and citation databases. ReaSoN allows the automatic extraction of social networks from bibliographic and citation databases, and allows the formulation of exploratory queries that may integrate data from external sources. It’s current version contains over 350K publications, 1.2M citations, and 379K authors.



- **ToXgene**: a declarative, template-based tool for the generation of arbitrarily large collections of complex, semantically-correlated collections of synthetic XML documents. ToXGene has had over 3800 downloads and has been used in over 50 universities and over 200 companies and research labs around the world (as of 2005, when I stopped tracking). It is also the data generator of XBench (Yao et al., ICDE 2004), the most comprehensive XML benchmark to date developed at the University of Waterloo; TPoX V1.0 (Nicola et al., SIGMOD 2008), an industry-standard benchmark developed by IBM<sup>1</sup>; and the MemBeR<sup>2</sup> repository for XML benchmarks, developed by INRIA, The University of Amsterdam, and The University of Antwerp.
- **ToXedit**: a tool for semi-automatically producing ToXgene data generation templates from existing XML documents. ToXedit allows the creation and editing of ToXgene templates, as well as annotating them with probabilistic information necessary for producing realistic test data; such information are automatically extracted from real XML documents. This tool was developed with funds from an IBM Faculty Award, and was part of the training of two undergraduate students at the University of Calgary (Y. Zholudz and G. Elliott).
- **A parallel MPEG video encoder**. This tool was developed as part of my MSc work, using C++ and PThreads, on a Sun multiprocessor with 32 CPUs. It was subsequently transferred to the Electrical and Computer Engineering Department at Carnegie Mellon University, where it is being used in the development of a high-level multiprocessor simulator.
- **A Parallel Suffix-Array Indexer**: This work concerned the design and tuning of a Quicksort-based parallel text indexing algorithm using a shared-nothing, message passing architecture. This software was transferred to a commercial Internet search engine in Brazil<sup>3</sup>, subsequently acquired by Google.

## INVITED TALKS, PANELS, TUTORIALS, AND VISITS

- “Knowledge Extraction From the Web”
  - University of Regina, October 2015
  - Federal University of Pernambuco, Brazil, June 2015
  - Dalhousie University, June 2015
  - University of the Peloponnese, Greece, May 2015
  - Universta Roma Tre, Italy, May 2015
  - Ludwig Maxmilliam Universitat, Germany, November 2014
  - Technische Universitat Darmstadt, Germany, September 2014
- “Inferencing in Information Extraction: Techniques and Applications”–Refereed Tutorial at the IEEE 31st International Conference on Data Engineering, 2015.
- “Shallow Information Extraction for the Knowledge Web”–Refereed Tutorial at the 29th IEEE International Conference on Data Engineering, 2013
- “Text Analytics for Social Media Analysis”–Tutorial at the 2013 UFRJ/EMC Summer School on Big Data, February 2013
- “Information Extraction for Social Media Analysis”–Refereed Tutorial at the 6th AAAI Conference on Weblogs and Social Media, June 2012.
- “Towards Summarizing and Making Sense of the Blogosphere”
  - Pontifical Catholic University, Rio de Janeiro, Brazil, August 2012.
  - Roma Tre University, June 2012.
  - MoMiNIS Seminar Series at Dalhousie University, March 2012.
- Google Faculty Summit, July 2010.
- “Top-k Approximate Subtree Matching”–Roma Tre University, July 2010.

<sup>1</sup><http://tpox.sourceforge.net/>

<sup>2</sup><http://tilcara.science.uva.nl:8080/exist/benchmarks/index.xq>

<sup>3</sup><http://www.todobr.com.br>

- “A Framework for Automatic Schema Mapping Verification Through Reasoning”–Roma Tre University, July 2010.
- “Building, Exploring and Querying Social Networks”.
  - Roma Tre University, June 2010,
  - National Institute for Web Science and Technology, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil, November 2009.
  - Universidade Federal de Pernambuco, Recife, Brazil, November 2009.
- “Evaluating the Effectiveness of Autonomic Data Management Tools”–Free University of Bozen-Bolzano, Bolzano, Italy, July 2008.
- “Towards Schema-Free Exchange and Update of XML Data”– University of Waterloo, Waterloo, March 2008.
- “Semantics and Information Loss in XML Data Exchange”–INRIA Lille Nord Europe, Lille, France, February 2008.
- “Panel on XML Updates”–Fourth International XML Database Symposium, Seoul, Korea, September 2006.

## **SERVICE**

### **National Societies**

- Canadian Artificial Intelligence Association – Executive Committee Member (2019 to 2021).

### **Select Editorial Service**

- Associate Editor, Distributed and Parallel Databases, 2020–2026.
- Associate Editor, IEEE Transactions on Knowledge and Data Engineering, 2015–2018.
- Associate Editor, Computational Intelligence, 2015–2018.
- D. Barbosa, E. Milios (Eds.). “Proc. of the 28th Canadian Conference on Artificial Intelligence”. ISBN 978-3-319-18355-8. Springer 2015.
- Associate Editor, ACM SIGMOD Record, 2010 – 2014.
- Reviewer for ACM Computing Reviews, 2012–
- SIGMOD Associate Information Director, 2006–2012.

### **Conference Proceedings Edited**

- D. Barbosa, K. LeFevre, E. Terzi (Eds.). “Proc. of the Second ACM SIGMOD Workshop on Databases and Social Networks”. ISBN 978-1-4503-1495-4. ACM, 2012.
- D. Barbosa, R. Pottinger, (Eds.). “Proc. of the 3rd International Workshop on Data Engineering Meets the Semantic Web”. IEEE Press, 2012.
- D. Barbosa, G. Miklau and C. Yu (Eds.). “Proc. of the First ACM SIGMOD Workshop on Databases and Social Networks”. ISBN 978-1-4503-0650-8. ACM, 2011.
- D. Barbosa, A. Bonifati, Z. Bellahsene, E. Hunt, R. Unland (Eds.): “Database and XML Technologies, 5th International XML Database Symposium”, Proc. . Lecture Notes in Computer Science 4704 Springer 2007, ISBN 978-3-540-75287-5.

**Program Committee  
Co-Chairing**

- 2016 – Applications Track of the 3rd IEEE International Conference on Data Science and Advanced Analytics
- 2015 – 28th Canadian Conference on Artificial Intelligence
- 2012 – Second ACM SIGMOD Workshop on Databases and Social Networks  
– 3rd International Workshop on Data Engineering Meets the Semantic Web
- 2011 – First ACM SIGMOD Workshop on Databases and Social Networks
- 2007 – The Fifth International XML Database Symposium

**Select Conference  
Program Committee  
Participation**

- 2021 – WSDM: 14th ACM International Conference on Web Search and Data Mining  
– EDBT: 24th International Conference on Extending Database Technology
- 2020 – EMNLP: The 2020 Conference on Empirical Methods in Natural Language Processing  
– ACL: The 58th Annual Meeting of the Association for Computational Linguistics  
– WWW: The Web Conference 2020  
– CIKM: 29th ACM International Conference on Information and Knowledge Management  
– EDBT: 23rd International Conference on Extending Database Technology
- 2019 – WSDM: 12th ACM International Conference on Web Search and Data Mining  
– LDK: Language, Data and Knowledge 2019.
- 2018 – VLDB: 44th Very Large Data Bases Conference  
– SIGMOD: 2018 Conference on Management of Data  
– NAACL: 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies  
– EDBT: 21st International Conference on Extending Database Technology
- 2017 – WWW: 26th World Wide Web Conference – Web Mining and Content Analysis  
– WSDM: 10th ACM International Conference on Web Search and Data Mining
- 2016 – SIGMOD: 2016 ACM International Conference on Management of data  
– ICDE: 32nd IEEE International Conference on Data Engineering  
– NAACL: 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies
- 2015 – VLDB: 41st Very Large Data Bases Conference  
– ICDE: 31st IEEE International Conference on Data Engineering

- 2014
  - WEBDB: 17th International Workshop on the Web and Databases
  - EDBT: 17th International Conference on Extending Database Technology
- 2013
  - WWW: 22nd World Wide Web Conference – Practice and Experience Track
  - ISWC: 12th International Semantic Web Conference – Evaluation Track
  - CAI: 26th Canadian Conference on Artificial Intelligence
  - WebDB: 16th International Workshop on the Web and Databases
- 2012
  - ICWSM: 6Th International AAAI Conference on Weblogs and Social Media
  - EDBT: 15th International Conference on Extending Database Technology
  - ISWC: 11th International Semantic Web Conference – Evaluation Track
  - SBBD: 27th Brazilian Symposium on Databases
- 2011
  - SIGMOD: 2011 ACM International Conference on Management of data
  - CIKM: 20th ACM International Conference on Information and Knowledge Management
  - APWEB: 13th Asia-Pacific Web Conference
  - WEBIST: 7th International Conference on Web Information Systems and Technologies
  - SBBD: 26th Brazilian Symposium on Databases
- 2010
  - VLDB: 36th International Conference on Very Large Data Bases
  - WWW: 19th International World Wide Web Conference
  - EDBT: 13th International Conference on Extending Database Technology
  - SBBD: 25th Brazilian Symposium on Databases
- 2009
  - ICDT: 2009 International Conference on Database Theory
  - WebDB: 12th International Workshop on The Web and Databases
  - SBBD: 24th Brazilian Symposium on Databases
- 2008
  - ICDE: 24th IEEE International Conference on Data Engineering
  - VLDB: 34th International Conference on Very Large Databases (Demos)
- 2007
  - ICDE: 23rd IEEE International Conference on Data Engineering
  - WWW: 16th International World Wide Web Conference (Posters)
  - CIKM: 16th Conference on Information and Knowledge Management
  - WebDB: 10th International Workshop on The Web and Databases
  - SBBD: 22nd Brazilian Symposium on Databases

- 2006
  - ICDE: 22nd IEEE International Conference on Data Engineering
  - WWW: 15th International World Wide Web Conference (Posters)
  - WebDB: 9th International Workshop on The Web and Databases
  - SBBB: 21st Brazilian Symposium on Databases
  - WIDM: 8th ACM Int'l Workshop on Web Information and Data Management
- 2005
  - WWW: 14th International World Wide Web Conference (Posters)
  - SBBB: 20th Brazilian Symposium on Databases

**Select External  
Reviews**

- Natural Science and Engineering Research Council of Canada
  - Discovery Grants (2/year on average): 2005–2017
  - Site Visit Committee Member, NSERC Industrial Research Chair: 2017
  - Panel member of an NSERC Canada Research Chair committee: 2016
  - Collaborative Research and Development grant (1/year on average): 2011–2017
- Swiss National Science Foundation, 2016
- CINECA, Italian Consortium for Computing, 2014, 2016
- Portuguese Foundation for Science and Technology (FCT), 2012, 2016
- British Columbia Innovation Council: 2010
- Israel Science Foundation: 2008
- The VLDB Journal (VLDB Endowment/Springer): 2008, 2009, 2010, 2014
- ACM Transactions on Database Systems: 2003, 2004, 2005, 2006, 2010
- ACM Transactions on the Web: 2007, 2009
- ACM Computing Surveys: 2008
- IEEE Transactions on Data and Knowledge Engineering: 2005—2008, 2010–2011
- IEEE Internet Computing: 2011
- IEEE Transactions on Dependable and Secure Computing: 2011
- Software: Practice and Experience (John Wiley & Sons): 2005, 2008
- Information Systems (Elsevier); 2005, 2007, 2008, 2011
- Journal of Computer and Systems Sciences (Elsevier); 2009
- Journal of Systems and Software (Elsevier); 2008
- Data & Knowledge Engineering (Elsevier); 2007, 2008

**External Examination  
Committees**

- MSc Thesis, University of Melbourne, 2017
- Dr-Ing Dissertation, University of Saarland/Max Planck Inst. for Informatics, 2015
- PhD Thesis, University of Sidney, 2014
- PhD Thesis, Roma Tre University, 2012
- Phd Thesis, Dalhousie University, 2012
- PhD Thesis, Roma Tre University, 2008
- MSc Thesis, University Federal de Minas Gerais, Brazil, 2012

- MSc Thesis, University of Amsterdam, 2010
- MSc Thesis, City University Dublin, 2007
- MSc Thesis, Universidade Federal do Amazonas (Brazil), 2007
- MSc Thesis, University of Calgary, 2006

#### **Other Service**

- Web Chair of ACM SIGMOD 2008.
- Publicity Chair of WWW 2007.
- Local organization/proceedings Chair of the 19th Scientific and Statistical Database Management Conference (SSDBM), 2007.
- Associate Information Director of ACM SIGMOD, 2006–2014.
- Co-chair of the *Fourth Canadian Database Research Workshop*, co-located with the 30th International Conference on Very Large Databases; August 2006.

#### **Event Organization**

- 2019 Reverse Expo, February 19; over 200 attendees, sponsored by AltaML, DeepMind, BioWare, Intuit, Levven, Scotiabank, Servus Credit Union, Startup Edmonton.
- 2018 Reverse Expo, February 16; over 130 attendees, sponsored by Alberta Innovates, AltaML, Deloitte, Google, Intuit, Servus Credit Union, and Startup Edmonton. Generated over 2K in revenue for other graduate student activities in Computing Science.
- 2017 Reverse Expo, February 24; over 70 attendees, funded by NSERC.

#### **Department, Faculty and University Committees**

- Department: Industrial Relations Director, since 2019;
- Department: Teaching Assistantship Coordinator, since 2019;
- Departmental Industrial Relationship Committee, 2016–2019;
- Departmental Director of the Science Internship Program, 2013–2014;
- Faculty IT Oversight Committee, 2012–2013;
- Departmental Operations Executive Committee, 2010–2012;
- Undergraduate Curriculum Committee, 2009–2011;
- Undergraduate Affairs Committee (U. of Calgary) 2005–2007.

#### **Programming Team Coach (U. of Calgary)**

- 2007 – ACM ICPC World Finals: honorable mention
- ACM ICPC Rocky Mountain Regional (3 teams): 1st, 14th, and 22nd
- Alberta Collegiate Programming Contest (3 teams): 1st, 5th, and 6th
- 2006 – ACM ICPC Rocky Mountain Regional (3 teams): 1st, 9th, and 13th
- Alberta Collegiate Programming Contest (3 teams): 2nd, 5th, and 6th

## **SUPERVISION**

## PDFs, RAs, etc

- Dr. **A. E. dos Santos**, PhD from University of Regina; *post-doctoral fellow* since 2020, working on explainable AI.
- Dr. **Y. Sekhavat**, PhD from Memorial University of Newfoundland; *post-doctoral fellow* 2013–2014, working on information extraction from the Web; Assistant Professor, Tabriz Art University
- **M. Cannaviccio**, *visiting MSc student* from Roma Tre University, 2014;
- **M. Bronzi**, *visiting PhD student* from Roma Tre University, 2011  
Senior NLP Research Engineer at Nuance Communications (Montreal);
- **M. Moraes**, *visiting PhD student* from Fed. Uni. of Rio Grande do Sul (Brazil), 2011  
Software engineer at HP;
- **C. Bigonha**, *visiting MSc student* from Fed. Uni. of Minas Gerais (Brazil), 2011;
- Dr. **P. Cappellari**, PhD in Computer Science from Roma Tre University, 2007; *post-doctoral fellow* 2008–2009, working on semantic verification of schema mappings.
- **Y. Merhav**, *visiting PhD student* from Illinois Institute of Technology, 2009/2010;
- **E. Cortez**, *visiting PhD student* from Fed. Uni. of Amazonas (Brazil), 2009;
- **F. Mesquita**, *visiting MSc student* from Fed. Uni. of Amazonas (Brazil), 2007.

## Graduate Level (completed)

- **C. Pang**, PhD 2019 co-supervised with A. Hindle; thesis: “Grounded Theory for DevOps Education”.  
Assistant Professor at MacEwan University.
- **E. Macdonald**, MSc 2019, thesis: “Neural Relation Extraction on Wikipedia Tables for Augmenting Knowledge Graphs”.  
Software Engineer at Intuit.
- **D. Caminhas**, MSc 2019, thesis: “Detecting and correcting typing errors in open-domain knowledge graphs using semantic representation of entities”.  
Software Engineer at Amazon.
- **W. Ni**, MSc 2019, thesis: “Multiple-Choice Question Answering Over Semi-Structured Tables”.  
Machine Learning Engineer - YITUTech.
- **P. Xu**, MSc 2018, *honorable mention for Departmental MSc Achievement Award*, thesis: “Towards Neural Information Extraction without Manually Annotated Data”.
- **M. Cannaviccio**, PhD 2018 (Roma Tre University), thesis: “Augmenting Knowledge Graphs with Natural Language Evidence”, co-supervised with P. Merialdo.
- **Z. Guo**, PhD 2018, thesis: “Toward an Accurate, Robust, and Scalable Named Entity Disambiguation System”.
- **H. Sepehri-Rad**, PhD, Alberta 2015, thesis: “Analyzing Controversy in Wikipedia”.  
Machine Learning Researcher at Granify
- **F. Mesquita**, PhD, Alberta 2015, thesis: “Extracting Information Networks from Text”.  
Machine Learning Engineer at Diffbot
- **A. Anandan**, MSc, Alberta 2012; dissertation: “Efficient Algorithms for Hierarchical Agglomerative Clustering”.  
Product Manager at Square
- **G. Leighton**, PhD, Alberta 2011; thesis: “Security Issues in Heterogeneous Data Federations”.  
Director of R&D at Pharm3r

- **S. Husby**, MSc, Alberta 2011; research paper: “Topic Classification of Blog Posts Using Distant Supervision”.  
Post-Secondary Instructor, MacEwan University
- **H. He**, MSc, Alberta 2011, co-supervised with Prof. G. Kondrak; dissertation: “Automatic speaker identification in novels”.
- **D. Serrano**, MSc, Alberta 2010, co-supervised with Prof. E. Stroulia; dissertation: “SociQL: a query language for the social web”.
- **M. Hasan**, MSc, Alberta 2010, co-supervised with Prof. E. Stroulia; dissertation: “Extracting Structured Knowledge from Textual Data in Software Repositories”.
- **Z. Guo**, MSc, Alberta 2009; dissertation: “Entity Resolution for Large Relational Datasets”. Currently PhD student under my supervision.
- **J. Walny**, MSc, Alberta 2009; dissertation titled “A framework for semantically verifying schema mappings for data exchange”.

**Graduate Level  
(current)**

- **P. Yao**: PhD student since 2019.
- **V. Raghantan**: MSc student since 2019.
- **N. Pei**: MSc student since 2019.
- **G. Alexander**: MSc student since 2019.
- **S. Ravishankar**: MSc student since 2019.
- **A. Whittaker**: MSc student from 2018.
- **T. Renwick**: MSc student from 2018.
- **A. Hejazizo**, MSc student since 2017.