

Drahomira Herrmannova

Knowledge Media Institute • The Open University
Walton Hall, Milton Keynes, MK3 6EL, United Kingdom
Phone: +44 (0) 7907 903227 • Email: emailme@drahomira.net
Last updated: November 17, 2018

SUMMARY

Research scientist with 6+ years of academic and professional experience. I work in a unique area between text-mining and research evaluation. My research directly resulted in the development of a new area called *semantometrics*, which uses full-text for research evaluation. I have developed several methods which demonstrate the benefits and suitability of using full-text in automated impact assessment, which led to a research grant and two fully funded PhD scholarships for my university. I am well connected in the academic community and have been organizing research workshops since 2014 to bring together different communities working on related problems in text mining of research publications. I have been a finalist in two international research competitions, participated in numerous research projects, and have undertaken a two year internship at Oak Ridge National Laboratory.

RESEARCH EXPERIENCE

- 2018–now **Scholarly Data Mining**, Research Associate (Postgraduate Researcher), Knowledge Media Institute, The Open University
- Leading a study to analyze the proportion of research outputs around the world that are open access and the effect that open access has on number of citations and other research evaluation metrics
 - Developing production-level big data pipeline using Spark for very large collections of research publications (130+ million publications)
 - Developing methods for classifying research publications into subject categories based off abstracts and document metadata such as author information; techniques include deep learning and graph-based approaches
- 2016–2018 **Text-mining research publications**, ORISE Postgraduate Intern/Researcher, Oak Ridge National Laboratory
- Worked with NIEHS scientists to automatically extract study descriptors from research documents, developed an unsupervised method to help identify relevant text segments, and tested a number of supervised classification methods (traditional machine learning as well as deep learning) in terms of their potential to support annotation and extraction
 - Prototyped a novel system for monitoring research performed around the world and for analyzing collaboration patterns in specific disciplines; this prototype led to a collaboration with the U.S. Department of Energy through which it is being expanding into an operational system
 - Developed and applied new research publication evaluation methods combining bibliometrics and altmetrics to enable the Oak Ridge Leadership Computing Facility to better evaluate research publications that make use of the facility

- 2012–2017 **Leveraging text-mining for evaluation of scholarly publications**, PhD Candidate, Knowledge Media Institute, The Open University
- Co-founded semantometrics (<http://semantometrics.org>), a new research direction in scientometrics that, unlike traditional methods, leverages full-text for research evaluation and incorporates semantics into evaluation
 - Developed and implemented several semantometric methods for evaluation of research publications which demonstrate the benefits of using full-text based metrics compared to traditional metrics
 - Evaluated and demonstrated effectiveness of semantometrics methods at scale through funding from Jisc, resulting in Jisc’s funding of several new PhD positions to continue research in semantometrics
- 2016 **Text-mining and enrichment of CORE Open Access collection**, Research Assistant, Knowledge Media Institute, The Open University
- Merged two large (70+ and 120+ million publications) collections to improve search and recommendation of research publications in CORE
 - Developed solutions for data engineering challenges ranging from correct matching of citations between the two collections to efficient data updates
- 2016 **Text-mining research publications to support research publication evaluation**, ORISE Postgraduate Intern/Researcher, Oak Ridge National Laboratory
- Developed new methods for assessing the importance of research publications by exploring knowledge transfer between disciplines, and then tested the methods on a set of publications that used ORNL’s Titan supercomputer
 - Developed the first openly available dataset for validating research evaluation metrics that contains qualitative evaluations
- 2014–2015 **Text-mining and enrichment of CORE Open Access collection**, Technical Support, Knowledge Media Institute, The Open University
- Developed a new version of an API for text-mining a large research publication collection
 - Developed and tested a new method for characterizing types of research collaboration that focuses on identifying collaborations with potential for high future impact
- 2013–2015 **Creating Data-Driven predictive models for improving student retention at the Open University**, Part-time consultant, Knowledge Media Institute, The Open University
- Developed and tested new machine-learning based methods for early identification of students at risk of dropping out of courses
 - Designed and implemented a dashboard with visualizations for presenting predictions to course tutors
- 2013 **Using Machine Learning for predicting student success in online courses**, Full-time consultant, Knowledge Media Institute, The Open University
- Pilot project in collaboration with Microsoft Research Cambridge
 - Implemented predictive models using Infer.NET, a tool for Bayesian Inference in graphical models, for predicting student outcomes in Open University courses

EDUCATION

Doctor of Philosophy, Computer Science, graduated 12/2017

The Open University, Knowledge Media Institute, Milton Keynes, United Kingdom

Thesis topic: *Mining Research Publication Networks for Impact*

Research interests: research evaluation, scientometrics, text-mining, information retrieval

Advisers: Professor Zdenek Zdrahal, The Open University & Dr. Petr Knoth, The Open University

Master of Science, Information Systems, graduated 08/2012

Brno University of Technology, Faculty of Information Technology, Brno, Czech Republic

Thesis topic: *A Relation/Topic-Based Visualisation to Aid Exploratory Search in Large Collections*

Received Dean's prize for excellent diploma thesis and Prize of Zdena Rabova for outstanding study and science results

Bachelor of Science, Information Technology with Cum Laude Honors, graduated 08/2010

Brno University of Technology, Faculty of Information Technology, Brno, Czech Republic

Thesis topic: *Social Network Integration into an Information Portal*

COMMUNITY BUILDING

2014–2018 **Organizer of Workshop on Mining Scientific Publications (WOSP)**

- Focus on text-mining scholarly publications
- Every year includes a data challenge
- Receive submissions each year from top institutions in the area
- Participation by Allen Institute for Artificial Intelligence, Microsoft, Elsevier, Harvard University, University of Cambridge, UC Berkley, Mendeleev and others
- Program committee includes some of the most high-profile researchers in the area
- WOSP 2018 held as part of Language Resources and Evaluation Conference (LREC) 2018 in Miyazaki, Japan

2018 **Organizer of Smoky Mountains Computational Sciences and Engineering Conference (SMC) Data Challenge**

- Worked with scientists at Oak Ridge National Laboratory to develop a challenge focused on big data and scientific datasets
- Received submissions from top U.S. institutions
- Data challenge co-located with SMC 2018 Conference in Gatlinburg, TN

2017 **Organizer of Workshop on Scholarly Web Mining (SWM)**

- Due to high popularity of WOSP, SWM was founded as a new workshop that focuses on web mining in addition to text-mining scholarly publications
- Record high attendance with participants from CMU, Cornell University, Peking University, KAIST, Microsoft, Google, Apple, Baidu, Yahoo and others

PROFESSIONAL EXPERIENCE

2011–2012 **Development of Open Access publication aggregator and search portal**, Internship, Knowledge Media Institute, The Open University

- Worked on both front- and back-end development of a research publication aggregator and search portal
- Implemented several parts of the system including a reference extractor, an API for text-mining and a visual search interface

- 2010–2011 **Design, development, testing and maintenance of internal software applications**, Application Development Support, Honeywell, Czech Republic
- Participated in all stages of production software development including design, administration, maintenance and user support
 - Developed software for monitoring internal processes from manufacturing to license management
 - Received a special award for developing a software utility for monitoring license usage which enabled the company to save money by purchasing fewer licenses

NOTABLE ACHIEVEMENTS

- 2016 **Finalist in 2016 WSDM Cup Challenge**, Web Search and Data Mining Conference, San Francisco, CA
- Developed a method for providing a static rank for scholarly articles using data from Microsoft Academic Graph
 - Ranked in the top 8 out of 32 teams, invited to compete in the second round of the challenge and presented at the WSDM Cup workshop in San Francisco
- 2013 **Finalist in CrossLink-2 Challenge**, NTCIR-10 Conference, Tokyo, Japan
- Together with my PhD advisor, developed a method for providing a ranked list of potential links between documents in different languages
 - Our methods achieved the best overall results in the English to Chinese, Japanese, Korean subtask and were one of the top performers in the Chinese, Japanese, Korean to English subtask

FELLOWSHIPS & GRANTS

- 2015 **Jisc Research Grant**, Semantometrics Demonstrator and Study (PI), £15,000
- 2012 **PhD Grant**, The Open University, United Kingdom, £45,000
- 2011 **Travel Grant**, Board of European Students of Technology
- 2011 **Travel Grant**, Technological Educational Institute of Crete, Greece
- 2007–2010 **Scholarship for excellent study results**, Brno University of Technology

AWARDS & HONORS

- 2016 **Best poster award**, Joint Conference on Digital Libraries, Newark
- 2012 **Prize of Zdena Rabova**, Brno University of Technology
- 2012 **Dean’s prize** for excellent diploma thesis, Brno University of Technology
- 2011 **Special Award** for creating a utility for license usage monitoring, Honeywell

INVITED TALKS

- 2016 **Invited speaker** at Jisc Open Citations Workshop in London, UK, on the topic “Semantometrics: Towards full-text based research evaluation”
- 2016 **Invited speaker** at Jisc Digital Festival Digifest 2016 in Birmingham, UK, on the topic “Towards full-text based research metrics: exploring Semantometrics”

ACADEMIC SERVICE

- 2018 **Manuscript peer review**, Journal of the Association for Information Science and Technology (JASIST)
- 2018 **Program committee member**, Joint Conference on Digital Libraries (JCDL) 2018
- 2018 **Program committee member**, Workshop on Semantics, Analytics and Visualisation: Enhancing Scholarly Dissemination, The Web Conference 2018
- 2017–2018 **Manuscript peer review**, Scientometrics Journal
- 2017 **Manuscript peer review**, Open Journal of Web Technologies
- 2014 **Manuscript peer review**, Workshop on Machine Learning and Learning Analytics, Learning Analytics and Knowledge Conference (LAK), Indianapolis, IN, USA.

COURSES AND CERTIFICATES

- 2017 **Neural Networks and Deep Learning**, deeplearning.ai @ Coursera, grade: 100%
- 2014 **Computing for Data Analysis**, Johns Hopkins University @ Coursera

COMMUNITY OUTREACH

- 2018 Oak Ridge Computer Science Girls instructor, developed and taught a class on text-mining
- 2018 Volunteer and co-instructor at “Introduce Your Daughter to AI” Women in Computing @ ORNL event, Oak Ridge National Laboratory, Tennessee
- 2017–2018 Oak Ridge Computer Science Girls volunteer, Oak Ridge, Tennessee
- 2017 DaVinci Art & Science Fair judge, Jefferson Middle School, Oak Ridge, Tennessee
- 2017 Volunteer at “Introduce Your Daughter to Code” Women in Computing @ ORNL event, Oak Ridge National Laboratory, Tennessee

PROFESSIONAL AFFILIATIONS

Association for Information Science and Technology (ASIS&T)
International Society for Scientometrics and Informetrics (ISSI)
ACM Special Interest Group on Information Retrieval (SIGIR)
IEEE Technical Committee on Digital Libraries (TCDL)
IEEE & ACM Member
ACM Women Member

PUBLICATIONS

Workshop proceedings

Herrmannova, D.; Young, S. R.; Patton, R. M.; Stahl, C. G.; Kleinstreuer, N. C. and Wolfe, M. S. (2018) *Unsupervised Identification of Study Descriptors in Toxicology Research: An Experimental Study*. Accepted for presentation at the 9th International Workshop on Health Text Mining and Information Analysis (LOUHI 2018) at the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018), Brussels, Belgium.

Journal article

Herrmannova, D.; Patton, R. M.; Knoth, P. and Stahl, C. G. (2018). *Do Citations and Readership Identify Seminal Publications?* Scientometrics 115: 239.

Conference proceedings

Herrmannova, D.; Knoth, P. and Patton, R. M. (2018). *Analyzing Citation-Distance Networks for Evaluating Publication Impact*. Proceedings of the Language Resources and Evaluation Conference (LREC) 2018, Miyazaki, Japan.

Workshop proceedings

Stahl, C. G.; Young, S. R.; **Herrmannova, D.**; Patton, R. M. and Wells, J. C. (2018) *DeepPDF: A Deep Learning Approach to Extracting Text from PDFs*. Proceedings of the 7th International Workshop on Mining Scientific Publications (WOSP) at Language Resources and Evaluation Conference (LREC) 2018, Miyazaki, Japan.

Workshop proceedings

Herrmannova, D.; Knoth, P.; Stahl, C. G.; Patton, R. M. and Wells, J. C. (2018) *Text and Graph Based Approach for Analyzing Patterns of Research Collaboration: An analysis of the TrueImpactDataset*. Proceedings of the 1st Workshop on Computational Impact Detection from Text Data (CIDTD) at Language Resources and Evaluation Conference (LREC) 2018, Miyazaki, Japan.

Thesis

Herrmannova, D. (2018) *Mining Scholarly Publications for Research Evaluation*. PhD Thesis. The Open University.

Workshop proceedings

Herrmannova, D.; Patton, R. M.; Knoth, P. and Stahl, C. G. (2017) *Citations and readership are poor indicators of research excellence: Introducing TrueImpactDataset, a new dataset for validating research evaluation metrics*. Proceedings of the 1st Workshop on Scholarly Web Mining (SWM) at Tenth ACM International Conference on Web Search and Data Mining (WSDM 2017), Cambridge, UK. ACM ICPS.

Workshop proceedings

Patton, R. M.; **Herrmannova, D.**; Stahl, C. G.; Wells, J. C. and Potok, T. E. (2017) *Audience Based View of Publication Impact*. 6th Workshop on Mining Scholarly Publications (WOSP) at 2017 ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL), Toronto, Canada.

Technical report

Herrmannova, D. and Knoth, P. (2016) *Towards full-text based research metrics: Exploring semantometrics: Report of Experiments*. Jisc repository. Jisc Report 6376.

Conference proceedings

Herrmannova, D. and Knoth, P. (2016) *Semantometrics: Towards fulltext-based research evaluation*. Proceedings of IEEE/ACM Joint Conference on Digital Libraries (JCDL 2016), Newark, NJ, USA.

Workshop proceedings

Herrmannova, D. and Knoth, P. (2016) *Simple Yet Effective Methods for Large-Scale Scholarly Publication Ranking: KMi and Mendeley (team BletchleyPark) at WSDM Cup 2016*. Proceedings of WSDM Cup 2016 - Entity Ranking Challenge Workshop at International Conference on Web Search and Data Mining (WSDM 2016), San Francisco, CA, USA.

Journal article

Herrmannova, D. and Knoth, P. (2016) *An Analysis of the Microsoft Academic Graph*. D-Lib Magazine, 22, (9/10). Corporation for National Research Initiatives.

Journal article

Herrmannova, D. and Knoth, P. (2015) *Semantometrics in Coauthorship Networks: Fulltext-based Approach for Analysing Patterns of Research Collaboration*. D-Lib Magazine, 21, (11/12). Corporation for National Research Initiatives.

Conference proceedings

Herrmannova, D.; Hlosta, M.; Kuzilek, J. and Zdrahal, Z. (2015) *Evaluating Weekly Predictions of At-Risk Students at The Open University: Results and Issues*. Proceedings of European Distance and E-Learning Network Conference (EDEN 2015). Barcelona, Spain.

Conference proceedings

Herrmannova, D. and Knoth, P. (2015) *Semantometrics: Fulltext-based Measures for Analysing Research Collaboration*. Proceedings of International Society for Scientometrics and Informetrics Conference 2015, Istanbul, Turkey.

Conference proceedings

Kuzilek, J.; Hlosta, M.; **Herrmannova, D.;** Zdrahal, Z. and Wolff, A. (2015) *OU Analyse: analysing at-risk students at The Open University*. Learning Analytics Review, 1-16.

Journal article

Knoth, P. and **Herrmannova, D.** (2014) *Towards Semantometrics: A New Semantic Similarity Based Measure for Assessing a Research Publication's Contribution*. D-Lib Magazine, 20, (11/12). Corporation for National Research Initiatives.

Workshop proceedings

Wolff, A.; Zdrahal, Z.; **Herrmannova, D.;** Kuzilek, J. and Hlosta, M. (2014) *Developing predictive models for early detection of at-risk students on distance learning modules*, Proceedings of the 2nd Workshop on Machine Learning and Learning Analytics at Learning Analytics and Knowledge Conference (LAK 2014), Indianapolis, IN, USA.

Workshop proceedings

Hlosta, M.; **Herrmannova, D.;** Vachova, L.; Kuzilek, J.; Zdrahal, Z. and Wolff, A. (2014) *Modelling student online behaviour in a virtual learning environment*. Workshop on Machine Learning and Learning Analytics at Learning Analytics and Knowledge Conference (LAK 2014), Indianapolis, IN, USA.

Conference proceedings

Knoth, P. and **Herrmannova, D.** (2013) *Simple Yet Effective Methods for Cross-Lingual Link Discovery (CLLD) - KMI @ NTCIR-10 CrossLink-2*. NTCIR-10 Evaluation of Information Access Technologies, Tokyo, Japan.

Book chapter

Wolff, A.; Zdrahal, Z.; **Herrmannova, D.** and Knoth, P. (2013) *Predicting Student Performance from Combined Data Sources*. Educational Data Mining: Applications and Trends, 524, Springer.

Thesis

Herrmannova, D. (2012) *A Relation/Topic-Based Visualisation to Aid Exploratory Search in Large Collections*. Brno University of Technology.

Journal article

Herrmannova, D. and Knoth, P. (2012) *Visual Search for Supporting Content Exploration in Large Document Collections*. D-Lib Magazine, 18, (7/8), Corporation for National Research Initiatives.

Book

Herrmannova, D. (2011) *Social Network Integration into an Information Portal: Social Media in Practice*. LAP Lambert Academic Publishing.

MEDIA COVERAGE

- 2018 **Article about an event for daughters of ORNL staff I helped to run:**
<https://web.archive.org/web/20180717193126/https://www.olcf.ornl.gov/2018/07/17/introduce-your-daughter-to-ai-event-sees-olcf-participation/>
- 2018 **Article about the team of volunteers who helped to prepare and run an event for daughters of ORNL staff titled “Introduce Your Daughter to AI”:**
<https://web.archive.org/web/20180717194636/http://www.techgirlz.org/ornl-the-team-behind-the-artificial-intelligence-how-computers-learn-techshop/>
- 2017 **Article about me on the news page of my alma mater (in Czech):**
<https://web.archive.org/web/20180717193927/http://zvut.cz/lide/lide-f38102/poznatky-pribyvaji-tak-rychle-ze-neni-v-lidskych-silach-informace-tridit-mysli-si-absolventka-fit-vut-drahomira-herrmannova-d157545>
- 2017 **Interview with me about my research (video):**
<http://openminded.eu/evaluating-impact-research/>