Drahomira Herrmannova

Research Scientist • Learning Systems
Oak Ridge National Laboratory • 1 Bethel Valley Road, Oak Ridge, TN, USA
Phone: +1 (865) 253 9980 • Email: herrmannovad@ornl.gov
Last updated: September 10, 2021

Summary

I am a Research Scientist in the Learning Systems Group at Oak Ridge National Laboratory (ORNL). The focus of my research is on helping scientists work more effectively by applying Artificial Intelligence methods to improve research workflows and enable intelligent access to the content of research publications. My research interests span Text and Data Mining, Natural Language Processing, Machine Learning, and their applications to biomedical, scientific, and other expert literature and data. My recent work focuses on developing models for literature screening and information extraction from scientific publications in low-resource settings, data extraction from tables in scientific documents, and deploying misinformation detection models to study the effects of misinformation on health outcomes. Prior to my current appointment, I was a postdoctoral researcher at the Knowledge Media Institute, The Open University in Milton Keynes, UK.

Research Experience

2019–now **Applying AI to scientific literature**, Research Scientist (R&D Associate Staff Member), Learning Systems Group, Oak Ridge National Laboratory

- Applying Text and Data Mining, Natural Language Processing, and Machine Learning to scientific literature, misinformation detection, and narrative analysis

2018–2019 **Scholarly data mining**, Research Associate (Postgraduate Researcher), Knowledge Media Institute, The Open University

- Lead study to analyze the proportion of research outputs around the world that are open access and the effect that national policies have on authors making their publications open access
- Developed production-level big data pipeline using Spark for very large collections of research publications (130+ million publications)
- Lead effort to create a labelled dataset and develop a model for identifying sentences from research publications which describe the contribution of the publication

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

2012–2017 **Leveraging text-mining for evaluation of scholarly publications**, PhD Candidate, Knowledge Media Institute, The Open University

- Developed and implemented several content-based 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 content-based methods at scale through funding from Jisc, resulting in Jisc's funding of several new PhD positions to continue research in semantometrics

2013–2015 **Using Machine Learning for predicting student success in online courses**, Full-time consultant (2013), Part-time consultant (2013-2015), 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
- Participated in a 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

2011–2012 **Development of Open Access publication aggregator and search portal**, Software Engineer, 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**, Software Engineer, 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

Education

Doctor of Philosophy, Computer Science, awarded 05/2018
The Open University, Knowledge Media Institute, Milton Keynes, United Kingdom
Thesis topic: *Mining Scholarly Publications for Research Evaluation*
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

2021 **Organizer, Second Workshop on Scholarly Document Processing (SDP 2021)**
- SDP aims to bring together distributed efforts on mining scientific literature to one place, collaboration with organizers of SciNLP, BIRNDL, and BigScholar workshops
- Workshop held at the 2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2021)
- Collaboration between Allen Institute for Artificial Intelligence, IBM Research AI, SRI International, Charles University, Google AI, Open University, Microsoft Research, Elsevier, Leibniz Institute for the Social Sciences, and ORNL

2014–2020 **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, Mendeley and others
- Program committee includes some of the most high-profile researchers in the area

2018 & 2020 **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 Conference

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

Academic Service

2018–2021 **Program committee member**, Joint Conference on Digital Libraries (JCDL)
2019–2021 **Reviewer**, Conference on Computational Natural Language Learning (CoNLL)
2020 **Scientific committee member**, Language Resources and Evaluation Conference (LREC)
2020 **Program committee member**, Workshop on Scientific Knowledge Graphs (SKG) 2020
2020 **Reviewer**, Neural Processing Letters, Springer
2020 **Reviewer**, International Journal of Artificial Intelligence in Education (IJAIED), Springer
2020 **Reviewer**, IEEE Access
2019 **Manuscript peer review**, Scientometrics journal special issue on "Bibliometrics and Information Retrieval"
2019 **Manuscript peer review**, International Journal on Digital Libraries (IJDL)
2019 **Program committee member**, Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL)

2018 **Manuscript peer review**, Journal of the Association for Information Science and Technology (JASIST)

2018 **Program committee member**, Workshop on Semantics, Analytics and Visualisation: Enhancing Scholarly Dissemination (SAVE-SD), The Web Conference 2018

2017 **Manuscript peer review**, Scientometrics Journal

2017 **Manuscript peer review**, Open Journal of Web Technologies (OJWT)

2014 **Manuscript peer review**, Workshop on Machine Learning and Learning Analytics, Learning Analytics and Knowledge Conference (LAK 2014), Indianapolis, IN, USA.

Acomplishments and Awards

2021 **R&D 100 Awards Finalist**, R&D World

2020 **ACM Gordon Bell finalist**, ACM Supercomputing Conference

2019 **Vannevar Bush Best Paper Award**, Joint Conference on Digital Libraries

2016 **Finalist in 2016 WSDM Cup Challenge on assessing query-independent importance of scholarly articles**, Web Search and Data Mining Conference (WSDM), San Francisco, CA; 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

2016 **Best Poster Award**, Joint Conference on Digital Libraries

2015 **Jisc Research Grant**, Semantometrics Demonstrator and Study (co-PI), £15,000

2013 **Finalist in CrossLink-2 Challenge on automatically finding potential links between documents in different languages**, NTCIR-10 Conference, Tokyo, Japan; 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

2012 **Prize of Zdena Rabova** for excellent study and science results (awarded to two students each year), Brno University of Technology

Invited Talks

2018 **Invited speaker** at the Organisation for Economic Co-operation and Development (OECD) workshop in Paris, France, on the topic of Systematic Reviews in the scope of the EDTA Conceptual Framework Level 1

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"

Mentoring

2021 Mohammad Saad Salman, ORISE SULI Internship, currently a sophomore at Loyola Marymount University

2020 Alexander Perry, ORISE SULI Internship, currently a senior at California State University-Sacramento

2019 Tomas Danis, Erasmus Internship, currently a C++ Software Engineer at think-cell Software, Berlin, Germany

2019 Shreyas Shahapur, work experience internship, currently a high school senior

2018 Brett Hagan, SULI Internship, currently graduate student at University of Tennessee

2018 Derek Shafer, SULI Internship, currently graduate student at Tennessee Technological University

Community Outreach

2017–2021	Oak Ridge Computer Science Girls volunteer and instructor, developed and taught a class on text-mining and voluteered for various classes, Oak Ridge, Tennessee
2018	Volunteer and co-instructor at "Introduce Your Daughter to AI" Women in Computing @ ORNL event, Oak Ridge National Laboratory, 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 Computational Linguistics (ACL)
Association for Information Science and Technology (ASIS&T)
ACM Special Interest Group on Information Retrieval (SIGIR)
FORCE11: The Future of Research Communications and e-Scholarship

Publications

Conference proceedings

Herrmannova, D.; Gunaratne, C.; Walker, V.; Rooney, A.; Patton, R.; Wolfe, M.; Schmitt, C. (2021). *Weak Supervision for Scientific Document Relevance Tagging.* Accepted for presentation at the 2021 ACM/IEEE Joint Conference on Digital Libraries.

Conference proceedings

Thakur, G.; Caspersen, J.; **Herrmannova, D.**; Eaton, B.; Burdette, J. (2021). *A Mixed-Method Design Approach for Empirically Based Selection of Unbiased Data Annotators.* Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021.

Workshop proceedings

Herrmannova, D.; Thakur, G.; Grant, J.; Tansakul, V.; Eaton B.; Kotevska, O.; Burdette, J.; Smyth, M.; Smith, M. (2021). *Challenges in Automated Detection of COVID-19 Misinformation.* Workshop on Human Aspects of Misinformation Online at the 2021 ACM CHI Virtual Conference on Human Factors in Computing Systems.

Journal article

Browne, P., de Vries, R. B. M.; **Herrmannova, D.**; LaLone, C. A.; Lam, J.; Marty, M. S.; Stahl, C. G.; Thayer, K. A.; Wheeler, J.; VanDer Wal, L. and Gourmelon, A. (2020). *Use of Existing data and Systematic Review [in the context of the OECD Conceptual Framework] for evaluating Endocrine Disrupting Chemicals.* Under review.

Conference proceedings

Kannan, R.; Piyush S.; Hao L., **Herrmannova, D.**; Patton, R. M.; Potok, T. E.; Thakkar, V. and Vuduc, R. (2020). *Scalable Knowledge-Graph Analytics at 136 Petaflops/s.* Proceedings of the ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis (IEEE Supercomputing). **ACM Gordon Bell finalist.**

Conference proceedings

Herrmannova, D.; Pontika, N. and Knoth, P. (2019). *Do Authors Deposit on Time? Tracking Open Access Policy Compliance.* Proceedings of the 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL 2019), Urbana-Champaign, IL. **Best paper award.**

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.* Proceedings of 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 Analyzing 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. **Best poster award.**

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.

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.

Media Coverage

2021	Our work on quantifying bias in human annotated data featured on ORNL News "New scientific approach reduces bias in training data for improved machine learning" `https://www.ornl.gov/news/new-scientific-approach-reduces-bias-training-data-improved-machine-learning`
2020	Our ACM Gordon Bell Prize submission featured on ORNL News "Computing – Mining for COVID-19 connections" (this article was re-posted by many on-line news sites, including HPC Wire, EurekAlert (AAAS), and SciTechDaily) `https://www.ornl.gov/news/computing-mining-covid-19-connections`
2020	Interviewed for ORNL News Article "COVID-19 forces ORNL researchers to take STEM education online" `https://www.ornl.gov/news/covid-19-forces-ornl-researchers-take-stem-education-online`
2019	Article in Physics Today highlighting the results of a study I lead: `https://dx.doi.org/10.1063/PT.6.2.20190418a`
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://openminted.eu/evaluating-impact-research/`