# Teasing journalistic findings out of heterogeneous sources: a data/AI journey

Ioana Manolescu
Inria and LIX/Ecole Polytechnique
ioana.manolescu@inria.fr

## ABSTRACT

Freedom of the press is under threat worldwide, and the quality of information that people have access to is dangerously degraded, under the joint threat of non-democratic governments and fake information propagation. The press as an industry needs powerful data management tools to help them interpret the complex reality surrounding us. Since 2018, I have been cooperating with journalists from Le Monde, France's leading newspaper, in devising tools for analyzing large and heterogeneuos data sources that they are interested in. This research has been embodied in ConnectionLens, a graph ETL tool capable of ingesting heterogeneous data sources into a graph, enriched (with the help of ML methods) with entities extracted from data of any type. On such integrated graphs, we devised novel algorithms for keyword search, and combine them in more recent research with structured querying. The talk describes the architecture and main algorithmic challenges in building and exploiting ConnectionLens graphs, illustrated in particular on an application where we study conflicts of interest in the biomedical domain. This is joint work with A. Anadiotis, O. Balalau, H. Galhardas and many others. ConnectionLens Web site (papers+code): https://team.inria.fr/cedar/connectionlens/. This research has been funded by Agence Nationale de la Recherche AI Chair SourcesSay (https://sourcessay.inria.fr).

**ACM Reference Format:**
Ioana Manolescu. 2022. Teasing journalistic findings out of heterogeneous sources: a data/AI journey. In *The 16th ACM International Conference on Distributed and Event-based Systems (DEBS '22), June 27–30, 2022, Copenhagen, Denmark*. ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3524860.3544406

## 1 BIOGRAPHY

Ioana Manolescu is a senior researcher at Inria and a part-time professor at Ecole Polytechnique, France. She is the lead of the CEDAR team, focusing on rich data analytics at cloud scale. She is a member of the PVLDB Endowment Board of Trustees, and has been recently an Associate Editor for PVLDB, a president of the ACM SIGMOD PhD Award Committee, and a chair of the IEEE ICDE conference; she has been the first ACM SIGMOD Reproducibility Chair in 2008, and a co-chair in 2009 and 2010. She has co-authored more than 150 articles in international journals and conferences, and co-authored books on "Web Data Management" and "Cloud-based RDF Data Management". Her main research interests algebraic and storage optimizations for semistructured data and in particular data models for the Semantic Web, heterogeneous data integration for data journalism, data models and algorithms for fact-checking, and distributed architectures for complex large data.


*DEBS '22, June 27–30, 2022, Copenhagen, Denmark*

ACM ISBN 978-1-4503-9308-9/22/06.
https://doi.org/10.1145/3524860.3544406