Dimitris Kontokostas

Veria, 59100, Greece | +30 694 579 6255 | jimkont@gmail.com
kontokostas.com | LinkedIn | Google Scholar | GitHub

Profile

Principal software engineer and knowledge graph architect with a PhD in large-scale knowledge extraction and 15+ years at the intersection of knowledge representation, data quality, and large-scale data integration. Currently building AI agent infrastructure and data standardization services for Medidata's clinical trial platform. Co-editor of the W3C SHACL specification, co-author of *Validating RDF Data* (Morgan & Claypool), with 50+ peer-reviewed publications and 7,500+ citations (h-index 22). Career spans building and validating knowledge graphs from web-scale (Diffbot, 10B+ entities) to enterprise (Medidata, clinical trials) to open community (DBpedia, Wikipedia's structured data). Combines deep academic grounding in semantic technologies with production engineering across JVM (Java, Kotlin, Scala) and Python stacks.

Core Expertise

- **Knowledge Graphs & Ontologies:** RDF, OWL, SPARQL, SHACL, ShEx, property graphs, graph databases (Virtuoso, Stardog, Neo4j), ontology design, entity resolution, knowledge fusion
- **AI & LLM Integration:** Multi-agent AI architectures, RAG systems, prompt engineering, LLM-grounded knowledge retrieval, schema-guided extraction, agentic workflows (MCP, AG-UI, ADK)
- **Data Engineering:** Snowflake, distributed data pipelines, Kafka, Spark, Kubernetes, AWS, GraphQL, SQL optimization, data quality at scale
- **Languages:** Java (primary, 15+ years), Kotlin, Scala, Python (current working language), shell/bash
- **Research & Standards:** W3C SHACL co-editor, 50+ publications, 7,500+ citations, h-index 22, peer review, EU project leadership

Employment History

Principal Software Engineer — Data Fabric & AI, Medidata (Dassault Systèmes) — *Remote, Greece*
April 2022 – Present

- Drive data standardization initiatives across the Medidata platform; co-built the enterprise data catalog powering Medidata Data Connect and Clinical Data Studio
- Design and implement a unit conversion service grounded in the QUDT ontology for consistent measurement handling across clinical trial datasets
- Build AI agents enabling customers to query and analyze their clinical trial data through natural language, including the analytics agent and supporting agent infrastructure; contribute to broader AI agent strategy and architecture initiatives across the platform

Key technologies: Python, Kotlin, Snowflake, Kubernetes, GraphQL, RDF/SHACL, LLMs, MCP, AG-UI, ADK, AWS

Lead Software Engineer — Knowledge Graph, Diffbot — *Remote, Greece*
March 2019 – March 2022

- Led data quality, coverage, and freshness improvements for the Diffbot Knowledge Graph (10B+ entities, 1T+ facts extracted from the open web)
- Built a scalable data quality evaluation platform to monitor knowledge graph performance across entity types and data dimensions

- Developed machine learning algorithms for knowledge fusion and entity reconciliation across heterogeneous web sources
- Designed tools and algorithms to identify and prioritize the refresh of data sources and web pages at scale

Key technologies: Java, Scala, Knowledge Graphs, Graph Databases, ontologies, data quality, knowledge fusion, ML

Tech Lead / Senior Software Engineer — Data Platform, GeoPhy — *Remote, Netherlands*

April 2017 – March 2019

- Hands-on lead engineer shaping the technical direction of the data ingestion, transformation, and enrichment platform for commercial real estate analytics
- Implemented streaming RDF data processing agents on Apache Kafka and co-developed data pipelines for the GeoPhy Knowledge Graph
- Incorporated data quality techniques from academic research into production semantic platform workflows

Key technologies: Kafka, AWS, Scala, Java, Docker, Kubernetes, RDF, SPARQL, Ontologies, CI/CD

Head of Technical Developments, DBpedia Association — *Remote, Leipzig*

March 2014 – April 2017

- Led the technical evolution of DBpedia, the canonical structured data extraction of Wikipedia, serving millions of queries daily across the linked data ecosystem
- Improved data extraction precision and recall through iterative processes; incorporated new sources (Wikidata, Commons) into the DBpedia data stack
- Administered Google Summer of Code projects (2014–2018) and coordinated the global DBpedia developer community

Key technologies: Java, Scala, Knowledge Graphs, RDF, SPARQL, Ontologies, community coordination

Research Fellow / PhD Researcher, University of Leipzig (AKSW/KILT) — *Leipzig, Germany*

September 2012 – March 2017

- PhD research on large-scale knowledge extraction, publishing, and quality assessment (magna cum laude)
- Created RDFUnit, an open-source RDF unit testing framework with significant industry adoption
- Led the ALIGNED EU Horizon 2020 project on behalf of Leipzig University, coordinating data quality engineering across the consortium
- Co-editor of W3C SHACL specification; active contributor to ShEx community group (Chair since 2017)

Key technologies: Java, Scala, RDF, SPARQL, OWL, SHACL, ShEx, data quality, EU project management

Earlier Career

R&D Software Engineer, Aristotle University of Thessaloniki (2012) — Semantic Web / Linked Data platform development

ICT Teacher, Ministry of Education, Greece (2006–2011) — Secondary education

R&D Software Engineer, Technical University of Crete (2001–2003) — XML parsers and databases

Software Engineer C++, TESAE, Athens (2004–2005) — Compiler implementation (FLEX/BISON)

Education

PhD (Dr.-Ing.) in Semantic Web — University of Leipzig, Germany (2012–2018)
Dissertation: Multilingual large-scale knowledge extraction, publishing and quality assessment. Grade: magna cum laude

MSc in Web Science — Aristotle University of Thessaloniki, Greece (2009–2012)
Grade: 9.2/10

Diploma in Electronic & Computer Engineering — Technical University of Crete, Greece (1999–2004)
Grade: 8.26/10

Selected Publications & Standards

- **W3C SHACL Specification** — Co-editor. Shapes Constraint Language, W3C Recommendation (2017). 733 citations
- *DBpedia — A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia*. Semantic Web Journal (2015). 4,846 citations
- ***Validating RDF Data*** — Co-authored book, Morgan & Claypool (2017). 183 citations
- *Test-driven Evaluation of Linked Data Quality*. WWW Conference (2014). 393 citations
- 50+ peer-reviewed papers and articles. Full list: scholar.google.gr/citations?user=lbWDQ7EAAAAJ

Notable Open Source Projects

- **RDFUnit** — Java RDF unit testing framework with significant industry uptake (github.com/AKSW/RDFUnit)
- **DBpedia Extraction Framework** — Technical lead of the structured data extraction pipeline for Wikipedia (github.com/DBpedia)

Languages

English: Excellent | Greek: Native