Francesco Tosoni, PhD

Lossless Compression | Open Data Infrastructure | Green Algorithm Engineering

francesco.tosoni@santannapisa.it francescotosoni.it www.santannapisa.it/en/francesco-tosoni Wikidata
LinkedIn GitHub GitLab ORCID: 0000-0001-8457-3866 dblp Google Scholar Scopus
Semantic Scholar ResearchGate IRIS Sant'Anna ARPI Pisa Web of Science

L'EMbeDS, Scuola Superiore Sant'Anna p.zza Martiri della Libertà 33, 56127 Pisa, Italy

Computer scientist (PhD) working at the intersection of *lossless data compression*, *compressed data structures*, and *green, energy-aware algorithm engineering*. My current research designs scalable, energy-aware compression, indexing, and semantic-search techniques for large source-code archives (Software Heritage), grounded in a commitment to open science and open-data infrastructures.

EDUCATION

PhD in Computer Science (EQF 8)

University of Pisa

Nov 2020 – Jan 2024 Pisa, Italy

I was a member of the Acube Laboratory (acube.di.unipi.it), directed by Prof. Paolo Ferragina.
My doctoral research focused on *computation-friendly* lossless compression techniques, adapting general-purpose lossless compression algorithms (grammar-based, zip-like, entropy-based) to application contexts such as machine-learning inference and string indexing. The proposed compressed representations enable the direct execution of linear algebra operations within time proportional to their compressed size. My research interests included compressed data structures, sparse matrix-vector multiplication (SpMV), key-value stores, compacted tries, stringology, and graph algorithms. I was advised by Prof. P. Ferragina and Prof. G. Manzini, authors of the FM-index and recipients of the 2022 ACM Paris Kanellakis Theory and Practice Award.

Thesis title: *Computation-Friendly Compression of Matrices and Tries*.
Advisors: P. Ferragina, G. Manzini. Defence date: 6 May 2024.
Grade: Excellent.

MSc in Computer Science and Networking (EQF 7)

University of Pisa *and* Sant'Anna School of Advanced Studies

Nov 2017 – Mar 2020 Pisa, Italy

The MSc programme in Computer Science and Networking (conducted in English) was a joint initiative of the TeCIP Institute (Sant'Anna School of Advanced Studies) and the University of Pisa (Departments of Computer Science and Information Engineering), attracting international students.

Achieved 30/30 *cum laude* in five courses.
Served as a *student representative* for the Master's programme (2018–2020).

Thesis title: *Algorithms and Data Structures for Efficient Ride-Sharing Platforms*.
Advisors: P. Ferragina, A. Marino. Defence date: 6 Mar 2020.
Grade: 110/110 *cum laude*.

BSc in Computer and Electronic Engineering (EQF 6)

University of Perugia

Sep 2014 – Oct 2017 Perugia, Italy

The BSc in Computer and Electronic Engineering provides a robust foundation in mathematics and physics, equipping students with the engineering skills to solve complex problems and integrate diverse knowledge for roles in electronics or ICT.

Achieved 30/30 *cum laude* in seven courses.

Thesis title (translated): *Implementation of a Distributed Graph Clustering Algorithm on the Giraph Platform Following the TLAV Paradigm*.
Advisor: E. Di Giacomo. Defence date: 27 Oct 2017.
Grade: 110/110 *cum laude*.

AWARDS

Sant'Anna School — Best Graduate Award (Master's in Computer Science & Networking)
Ranked 2nd in the cohort (2017-2018) | High GPA | Fastest completion.

Award details

---

con.Scienze — National Best Master's Thesis Award 2020
Selected among all Italian technical universities | August 2019 - July 2020 cohort.

Award announcement

PUBLICATIONS

Link legend: DOI • arXiv • author PDF • Wikidata • / code • data • SWHID • fix • licence. *Venue rankings:* ICORE 2026 (conferences; workshops are not ranked) and SCImago SJR 2025 (journals; best-category quartile), each linked to its source. *Author order:* alphabetical, as customary in theoretical computer science, in j2, j3, c1, c2, c4, w2.

Journals

[ ]

[j4]. **F. Tosoni**, P. Bille, V. Brunacci, A. De Angelis, P. Ferragina, G. Manzini.
"Toward Greener Matrix Operations by Lossless Compressed Formats".
IEEE Access [SJR 2025: Q1, 0.884], vol. 13, pp. 56756–56779, 2025.
DOI • arXiv • WD (paper) • WD (code) • code
As first and corresponding author, I initiated and led this interdisciplinary research project, bridging theoretical computer science and electronic engineering. My work involved enhancing Google's Zuckerli compressed matrix formats (Github fork) to enable computation-friendly matrix-to-vector and PageRank kernel operations. Published as Gold open access under CC BY 4.0.

Author PDF • available under CC BY 4.0

[j3]. A. Boffa, P. Ferragina, **F. Tosoni**, G. Vinciguerra.
"CoCo-trie: Data-aware Compression and Indexing of Strings".
Information Systems (IS) [SJR 2025: Q1, 0.981], vol. 120, art. 102316, 2024.
DOI • WD (paper) • WD (code) • code • SWHID
Extended version of the SPIRE 2022 paper [c1]. I contributed to both the theoretical framework and the codebase of the CoCo-trie.

Author PDF • available under CC BY-NC-ND 4.0

[j2]. P. Ferragina, G. Manzini, T. Gagie, D. Köppl, G. Navarro, M. Striani, **F. Tosoni** (alphabetical by convention; the published order was altered during typesetting).
"Improving Matrix-Vector Multiplication via Lossless Grammar-Compressed Matrices".
Proc. VLDB Endowment (PVLDB) [SJR 2025: Q1, 1.767], vol. 15, no. 10, pp. 2175–2187, 2022.
DOI • arXiv • WD (paper) • WD (code) • code • SWHID
As the corresponding author, I developed the theoretical framework and codebase for column-reordering optimisation algorithms based on TSP heuristics. I also conducted extensive comparisons with baselines, notably with Apache SystemDS's Compressed Linear Algebra (CLA), for which I corresponded with Prof. Matthias Boehm (IBM Research), who later refereed my PhD dissertation.

Author PDF • available under CC BY-NC-ND 4.0

[j1]. **F. Tosoni**, P. Ferragina, A. Marino, G. Resta, P. Santi.
"Locality Filtering for Efficient Ride-Sharing Platforms".
IEEE Transactions on Intelligent Transportation Systems (TITS) [SJR 2025: Q1, 2.608], vol. 23, no. 7, pp. 7785–7804, 2022.
DOI • WD (paper)
Based on my MSc thesis, this work, developed in collaboration with the MIT Senseable City Lab, earned the con.Scienze National Best Master's Thesis Award 2020.

Conferences

[ ]

[c5]. **F. Tosoni**, G. Mencagli.
"Streaming Right Multiplication over Grammar-Compressed Matrices: A Memory-Bounded GPU Engine for Genotype and Graph Data".
SIAM Symposium on Algorithm Engineering and Experiments (ALENEX) [ICORE 2026: A], 2027 *(to appear)*.
arXiv • code • data • SWHID

As first and corresponding author, I designed and implemented a memory-bounded GPU engine for right-multiplication over grammar-compressed matrices, targeting genotype and graph data. The code is openly released and archived in Software Heritage; the accompanying data are deposited on Zenodo.

[c4]. C. C. Pellizzari di San Girolamo, **F. Tosoni**.
"From Code Archival to Knowledge Graph: Bridging Software Heritage, COAR Notify and Wikidata".
Wikidata Workshop 2026, co-located with *ISWC 2026*, *(to appear)*.
arXiv • code
Work on bridging Software Heritage code archival with the Semantic Web via COAR Notify and Wikidata reconciliation.

[c3]. L. Venuta, **F. Tosoni**, P. Ferragina.
"Candidate Recall in Large-Scale Code-to-Code Retrieval: A Comprehensive Benchmark".
Web Information Systems Engineering (WISE) [ICORE 2026: B], Springer, 2026 *(to appear)*.
arXiv • code
A comprehensive benchmark of candidate recall for large-scale code-to-code retrieval.

[c2]. P. Ferragina, **F. Tosoni**.
"The Energy–Throughput Trade-off in Lossless-Compressed Source Code Storage".
GREENVOLVE Workshop, co-located with *IEEE International Conference on Software Analysis, Evolution and Reengineering* (SANER), in *SANER 2026 Companion Proceedings*, pp. 157–164, 2026.
DOI • arXiv • WD (paper) • WD (code) • code • fix
I wrote and curated the codebase, contributed to the scientific framework and writing and managed contacts with the Software Heritage team at Inria Paris.

[c1]. A. Boffa, P. Ferragina, **F. Tosoni**, G. Vinciguerra.
"Compressed String Dictionaries via Data-Aware Subtrie Compaction".
String Processing and Information Retrieval (SPIRE) [ICORE 2026: C], pp. 233–249, 2022.
DOI • WD (paper) • WD (code) • code
I contributed to both the theoretical framework and the codebase of the CoCo-trie.

Submitted manuscripts

[ ]
Note: Manuscripts submitted to peer-reviewed journals, at the stage indicated for each; not yet accepted.

[w4]. **F. Tosoni**, P. Ferragina.
"The Energy–Throughput Trade-off in Lossless-Compressed Source Code Storage".
Extended manuscript of the conference paper [c2].
Submitted (28 Aug 2026): *Information and Software Technology* (IST) [SJR 2025: Q1, 1.054], Elsevier, special issue on Green Software Evolution.
WD (code) • code

[w3]. **F. Tosoni**.
"MediaWiki Code2Code Search: Neural Retrieval for the Semantic Discovery of Open-Source Software Entities".
Minor revision: SoftwareX [SJR 2025: Q2, 0.499], Elsevier; revised version resubmitted on 14 Sep 2026.
arXiv • WD (code) • code • SWHID • Zenodo • Hugging Face

[w2]. G. Carmona, P. Ferragina, G. Manzini, **F. Tosoni**.
"Extended Depth-First Representations of k^2-trees".
Extends "Depth-First Representations of k^2-trees" (SPIRE 2025).
Under review: Theory of Computing Systems (ToCS) [SJR 2025: Q3, 0.295], Springer.
arXiv

[w1]. **F. Tosoni**.
"An Analysis and Implementation of Seam Carving for Content-Aware Image Resizing".
Under revision: Image Processing On Line (IPOL) [SJR 2025: Q3, 0.297], peer-reviewed article with reproducible code.
arXiv • WD (code) • code

RESEARCH POSITIONS & SCHOLARSHIPS

Note: For each position, I report the associated publications (c1,c2,j1,j2,j3,j4) produced as outcomes.

Research Contractor
Health Science Interdisciplinary Research Centre
Sep 2025 – Aug 2027 Sant'Anna School, Italy

Research conducted at the Sant'Anna School of Advanced Studies, on solutions for code archival, language-to-code, code-to-code, and semantic search (website). (c2)

Postdoctoral researcher

Project in collaboration with Software Heritage

Jul 2024 – Jun 2025 | Univ. of Pisa, Italy

Research conducted at the University of Pisa, on energy-aware compression schemes and parallel and I/O-efficient compression and indexing techniques for large source-code archives (website). I collaborated with the team at Software Heritage, including Martin Kirchgessner, David Douard, Benoît Chauvet, and Roberto di Cosmo. (j4)

PhD Scholarship

PhD in Computer Science, 36° cycle

Nov 2020 – Oct 2023 | Univ. of Pisa

Recipient of a three-year PhD research grant from the University of Pisa (Department of Computer Science). (c1, j1, j2, j3)

Research Scholarship

Citypost S.p.A.

Jun – Oct 2020 | Univ. of Pisa

Title: *Algorithms and Data Structures for Urban Mobility Platforms*. Duration: five months. Grant: Citypost S.p.A. Researched graph-based algorithmic solutions for vehicle routing and mobility problems, as part of Acube Lab's 2018–2020 research collaboration.

Offer declined: fixed-term junior researcher position (RTDA, art. 24(3)(a) L. 240/2010) in Computer Science (INFO-01/A), Free University of Bozen-Bolzano, 2025, with Prof. Johann Gamper; project on scalable database solutions for temporal data and time-series processing. (website)

FUNDED PROJECTS SUPPORTING MY RESEARCH

Note: [Inter]national projects under which my positions or research were funded, as a team member (not as PI). For each project, I report the associated publications (c1,j1,j2,j3,j4) produced as outcomes.

Alfred P. Sloan Foundation

Technology: Open Source in Science — Grant G-2025-25193

2025–ongoing | Grant `G-2025-25193`

Funding for the project to design and implement in Software Heritage new algorithmic solutions for compression and code-based searches.
Grant #G-2025-25193 (sloan.org) (c2)

NextGenerationEU—National Recovery and Resilience Plan (PNRR)

SoBigData.it-Strengthening the Italian RI for Social Mining and Big Data Analytics — Avviso (3264 del 28/12/2021)

2022–ongoing | Grant `IR0000013`

Funding for the project "SoBigData.it-Strengthening the Italian RI for Social Mining and Big Data Analytics."
Visit pnrr.sobigdata.it/ (j3, j4)

European Union-NextGenerationEU-PNRR

ICSC-Centro Nazionale di Ricerca in High-Performance Computing, Big Data and Quantum Computing

2022–ongoing | Spoke "Future HPC and BigData"

Funding for the Spoke "Future HPC and BigData."
Learn more at www.supercomputing-icsc.it. (j3, j4)

European Union-Horizon 2020 Program

SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics

2020–ongoing — Grant 871042

Funded through the Scheme "INFRAIA-01-2018-2019—Integrating Activities for Advanced Communities." More information can be found at www.sobigdata.eu. (j2, j3, j4)

European Union-Horizon 2020 Program

HumanE AI Network

2020–ongoing — Grant 952026

Funding through the project "HumanE AI Network."
See www.humane-ai.eu for details. (j2)

NextGenerationEU – PNRR / MUR PRIN Project

Multicriteria Data Structures and Algorithms: From Compressed to Learned Indexes, and Beyond

2019–2023 — Grant n. `2017WR7SHH`

Funding from the Italian Ministry of University and Research (MUR) under the "Progetti di Rilevante Interesse Nazionale" (PRIN) programme for the project "Multicriteria Data Structures and Algorithms." Extended research includes compressed/learne indexes and beyond. Project website: learned.di.unipi.it. (j2, j3)

MIT-UniPI Grant

Using Graph Compression for Shortest Path Computation in Urban On-Demand Mobility

2019–2021

Grant on "Using Graph Compression for Shortest Path Computation in Urban On-Demand Mobility" (website). (j1)

Travel & open-access grants

- **2026** Wikimedia CH, open-access publication grant (1,000 CHF): article-processing charges for "MediaWiki Code2Code Search" (w3). (grant page)
- **2026** Wikimedia Italia, travel microgrant (€1,704, joint with C. C. Pellizzari di San Girolamo): presenting (c4) at the Wikidata Workshop @ ISWC 2026, Bari. (grant page)

INTERNATIONAL RESEARCH VISITS

Short research visit (4 days)

GRICAD

25 Nov – 28 Nov 2025 — Grenoble, France

Collaboration on the integration of RocksDB-PPC into the Software Heritage object store, with the goal of scaling to handle 30 billion files. Focus on dataset construction, partitioning strategies, and alternatives to filename-based heuristics, using the Kraken cluster for benchmarking. (signed letter)

Visiting Researcher

Software Heritage, Inria

1 Jul – 31 Jul 2025 — Paris, France

Collaboration with the Software Heritage team on efficient code compression for storage and retrieval (shard format, Terabyte-sized caching system), specifically linked to the CodeCommons project supervised by Prof. Roberto Di Cosmo and Prof. Stefano Zacchiroli. This visit aims to strengthen the existing collaboration and apply research directly to Software Heritage's vast source code archive. (signed letter)

Visiting PhD student

University of Chile

Sem. I, A.Y. 22/23 — Santiago de Chile

Three-month stay at the laboratory directed by Prof. Gonzalo Navarro. I worked on applications of the k^2-tree data structure for the storage and operation of large and sparse graphs. (signed letter)

TALKS & PRESENTATIONS

2026

- **31 Aug** *"Unpacking Software Heritage Open Science"* — DCW Conversation Hour #44, Deoband Community Wikimedia, online. [virtual] (slides • Commons • Zenodo)
- **23 Jul** *"India's Software Stories: preserving code as heritage with Software Heritage, Wikimedia Commons and Wikidata"* — Wikimania 2026, Paris, France. Joint with P. Santhosh Surampudi. (slides • Commons • video • Zenodo)
- **2 May** *"MediaWiki Code2Code Search: Semantic Search to Find Code by Under-the-Surface Similarity"* — Wikimedia Hackathon 2026 (unconference), Milan, Italy. (website • slides • Commons • Zenodo • Phabricator)
- **26 Mar** *"The Software Heritage Open Science Ecosystem: Preserving software as executable knowledge"* — CTS Seminar Series, Univ. of Pisa & IMT Lucca. [invited] (website • slides)
- **17 Mar** *"The Energy–Throughput Trade-off in Lossless-Compressed Source Code Storage"* — GREENVOLVE Workshop @ SANER 2026, Limassol, Cyprus. Joint with P. Ferragina. (website • slides)
- **21 Jan** *"Lossless-Compressed Key-Value Store for the Software Heritage Archive"* — A³Lab Inaugural Workshop (Session I), Scuola Superiore Sant'Anna, Pisa. (website • poster • slides)

2025

- **21 Nov** *"The Energy–Throughput Trade-off in Lossless-Compressed Source Code Storage"* — DBS Group seminar (Prof. J. Gamper), NOI Techpark, Free Univ. of Bozen-Bolzano. [invited] (slides)
- **22 Sep** *"Object Storage Compression"* — CodeCommons Plenary Event (Infrastructure & Unified Data Model track). [invited] (website • slides)
- **2 Jul** *"Lossless-compressed data storage for SWH: Compressed, tunable & energy-aware"* — Software Heritage team seminar, Inria Paris Centre, Paris. [invited] (slides)
- **4 Jun** *"Toward Greener Matrix Operations by Lossless Compressed Formats"* — Workshop "Learning from large, complex and structured data", Sant'Anna School, Pisa. [invited] (website • slides)
- **27 Feb** *"Verso operazioni più green su matrici tramite formati compressi senza perdita"* — Google Developer Group (GDG) Pisa, Polo Fibonacci, Univ. of Pisa. [invited] (website • slides)
- **30 Jan** *"Measuring impact by extracting knowledge of software assets"* (poster) — 2025 Software Heritage Community Workshop, Inria, Paris. (Zenodo • website)
- **28 Jan** *"Enhancing SWH Object Storage with Compressed and Dynamic Solutions"* — CodeCommons Kickoff, Inria, Paris. [invited] (website • slides)

PEER REVIEWING

- **2026** Program Committee member, **WISE 2026** [ICORE 2026: B] (27th Int. Conf. on Web Information Systems Engineering), research track, Venice, Italy — 2 papers reviewed. (program committee)
- **2025** Reviewer, **The Journal of Supercomputing** [SJR 2025: Q2, 0.729], Springer Nature — 1 manuscript. (certificate)

TECHNICAL EXPERTISE

Research focus

Compression & compressed data structures

lossless compression • compressed data structures • computation-friendly compression • string indexing • stringology

Compressed linear algebra on GPUs & accelerators

compressed linear algebra • memory-bounded GPU algorithms • grammar-compressed matrices • SpMV & PageRank kernels • genotype & graph data

Code archives & semantic search

source-code archives • semantic code search • code-to-code retrieval • digital preservation • open-data infrastructures

Cross-cutting

green, energy-aware algorithm engineering

Core Tech Stack

C/C++	(Expert; C++11/14/17, CMake, pthreads, parallel patterns)
CUDA	(GPU kernels, memory-bounded)
Python	(NumPy, pybind11)
Bash	(Scripting/Automation)
Java	(Prior experience)
JavaScript	

Tools

- CLion (CMake, Debugging)
- VS + Copilot + git
- RocksDB, Boost
- Docker/Compose (Good)

TEACHING

Tutor

“Laboratorio II”

Sem. I, A.Y. 23/24 — University of Pisa

Tutoring (conducted in Italian) for the undergraduate course of “Laboratorio II” (24 hours).

Tutor

“Laboratorio I-A”

Sem. I, A.Y. 21/22 — University of Pisa

Tutoring (conducted in Italian) for the undergraduate course of “Laboratorio I-A” (20 hours).

NATURAL LANGUAGES

I get by with several languages, with varying levels of active engagement and passive comprehension. These include the following.

English ●●●●●
CEFR C1. Professional proficiency in all communication skills.

Deutsch [German] ●●●○○
CEFR B1. Certified language proficiency from the Centre for Applied Linguistics (CLA), UNIPG, Perugia, Italy (July 2017).

Castellano [Spanish] ●●●○○
CEFR B1-B2. Strong listening and reading comprehension; experience attending scientific seminars in Spanish; AI-assisted writing.

हिन्दी [Hindi] ●○○○○
CEFR A2. Basic grammar and vocabulary knowledge. Gboard-assisted writing; no handwriting.

Italiano [Italian] ●●●●●
Native speaker fluency.

OPEN DATA CONTRIBUTIONS & COMMUNITY ENGAGEMENT

Maintainer of open-source tools on Wikimedia Toolforge, including Code2Code Search (w3) and the Nablabot cross-wiki consistency bot. Member of the Indic MediaWiki Developers User Group (Wikidata reconciliation; student mentoring at the Indic Wikimedia Hackathon 2026). Co-manager of the Wikidata community on Zenodo, curating inclusion requests for Wikidata-related research outputs.

All of the above statements are true in accordance with the regulations on substitute declarations as per Article 46 et seq. of the Italian Presidential Decree 445/2000.

Pisa (Italy), the 2nd of October 2026