Sharing Civil Engineering Data: Technical, Cultural, and Practical Lessons from Two Open Databases

Lucio Soibelman, Ph.D.

Bio

Dr. Lucio Soibelman holds the Fred Champion Chair in Engineering and serves as Professor of Civil and Environmental Engineering and the Spatial Sciences Institute at the University of Southern California Viterbi School of Engineering. He earned his Ph.D. from MIT (1998) and previously held faculty positions at the University of Illinois and Carnegie Mellon University. His research centers on information technology for construction management, process integration in large-scale engineering systems, artificial intelligence, data mining, machine learning, and advanced infrastructure systems. With over 150 publications, his work has been funded by NSF, NASA, DOE, and industry partners. He was elected to the National Academy of Construction (2019), is an ASCE Fellow (2013), and received the 2022 ASCE Peurifoy Construction Research Award for pivotal contributions to information technologies, smart buildings, and AI in civil engineering.

Abstract

Open datasets are increasingly essential for advancing data-driven research in civil engineering, yet creating and sustaining them requires significant technical and organizational effort. In this talk, lessons learned from developing and publishing two public engineering datasets in different domains are shared.

The first dataset addressed a critical gap in the detailed electrical characterization of domestic appliances. Electrical load disaggregation algorithms rely on high-resolution appliance fingerprints, but most prior monitoring campaigns reported only low-resolution metrics such as average power. To overcome this limitation, synchronized voltage and current waveforms at 30 kHz were collected for over one hundred appliances. While early publication attempts (circa 2010) were challenging due to limited data-sharing infrastructure, later releases (2016–2018) benefited from mature repositories and data journals. The peer-review process for a dedicated data publication significantly improved both the dataset and its documentation.

The second case study is a large-scale 3D computer vision dataset released with BMVC 2022. Public availability greatly amplified its visibility, leading to thousands of downloads, widespread benchmarking use, and two international challenges at ECCV 2022 and ICCV 2023. Alongside these benefits come ongoing responsibilities: maintenance, user support, and clear documentation. A key lesson is that investing in high-quality documentation from the outset substantially improves reuse and reduces long-term support burdens.