Award No: NSF ACI 1443013; Project Title: T2C2: Timely and Trustworthy Curator and Coordinator Data Blocks; PI: Nahrstedt

4CeeD: Real-Time Data Acquisition and Analysis Framework for Material-related Cyber-Physical Environments

K. Nahrstedt, S. Konstanty, T. Nicholson, P. Ngyuen, T. Spila, T. O'Brien, M. Chan, A. Schwartz-Duval
N. Aluru, P. Braun, R. Campbell, B. Cunningham, I. Gupta, K. McHenry, J. Rogers
University of Illinois at Urbana-Champaign
Website: http://t2c2.csl.illinois.edu

Studies suggest that it typically takes 20 years to go from the discovery of new materials to fabrication of new and next-generation devices based on the new materials. This cycle must accelerate, and the main challenge is how we collect digital data about materials and how we make the digital data available to computational tools for developing new materials and fabricating new devices and to the research community.

The current state of data capture and storage in materials and semiconductor fabrication domains involves a lot of manual work that contribute to the long cycle from discovery of new materials to fabrication of new devices. First, in terms of *data transfer*, transferring files between materials research lab and researchers' office is often done either via "sneaker-net" techniques using flash-drives, or via emails. During such a slow process, no data conversion is available, which hinders researchers from previewing the results early and making timely decision at the microscope during lab sessions. Second, in terms of *data management and processing*, researchers often store data on their local hard drive or they use cloud-based storage services, such as Box or Google Drive. However, these storage mechanisms do not provide any assistance in organizing and processing data to support efficient search, curation, coordination and management of data. Thus, it often takes an extensive manual effort from researchers to organize their data and log important information about the experiments as they are being interpreted and curated, which leads to poor documentation of results. Third, in terms of *data access and sharing*, despite their strong dependency, materials science and semiconductor fabrication areas have never been digitally connected and the only linkage until now is through published results in publications. Furthermore, it is common that only "best" data results are kept for publishing, while "imperfect" and/or "secondary" data are discarded, even though they could contain vital information that could accelerate the use of a new material for a semiconductor device in the future.

In our T2C2 project, we explore the 4CeeD framework, which provides an expedient mean to capture, transfer, curate, analyze, visualize, coordinate and share the multi-modal data and metadata in real-time and in trusted manner before archiving and publishing experimental results. 4CeeD consists of two main data building blocks and services. At the 4CeeD user tier, the *curator* runs the *curation service* which performs nimble and adaptive data collection from instruments by wrapping data with extensive meta-data in real-time and in a trusted manner. In addition, this service provides advanced data management, annotation, and sharing of the collected data after they have been uploaded to the back-end cloud. At the 4CeeD cloud tier, the *coordinator* runs the *coordination service* to filter data, perform extraction of meta-data from microscope images, analyze and find correlations among the data, and identify new dependency relations between materials and device fabrication processes.

We have implemented and deployed 4CeeD at the University of Illinois at Urbana-Champaign (UIUC) with test users from two major research labs that share research facilities: Micro and Nano Technology Laboratory (MNTL) and Materials Research Laboratory (MRL). The 4CeeD version 1.0 software is available on GitHub. Primary feedback from the test users indicates that 4CeeD has helped them significantly reduce time and cost, hence accelerate their science, at the instruments to upload their heterogeneous data from the instruments, and in their offices to curate, visualize, analyze and search for data and metadata. Our system evaluation shows that 4CeeD deals very well with high-volume and fast-changing data workloads. In addition, by making experimental data available via authorized access, 4CeeD helps to close the communication gap between researchers and prevents unnecessary repetitions of the experimental processes caused by the lack of information in publications.

Remaining challenges and future directions are: (1) addition of new instruments and their metadata extractors connected to 4CeeD; (2) expansion of data upload, curation, workflow functions due to diverse user base (students, faculty, research engineers, lab managers); (3) investigation of correlation algorithms to enable specification and search of interdependencies between materials and semiconductor fabrication data; (4) increase of user base and dissemination of 4CeeeD beyond UIUC organizations; (5) development of 4CeeD training material since 4CeeD is a self-containing tool which can be installed and operated in each organization individually; and (6) installation of production private 4CeeD cloud at UIUC with appropriate storage policies and sustainability charging pricing.