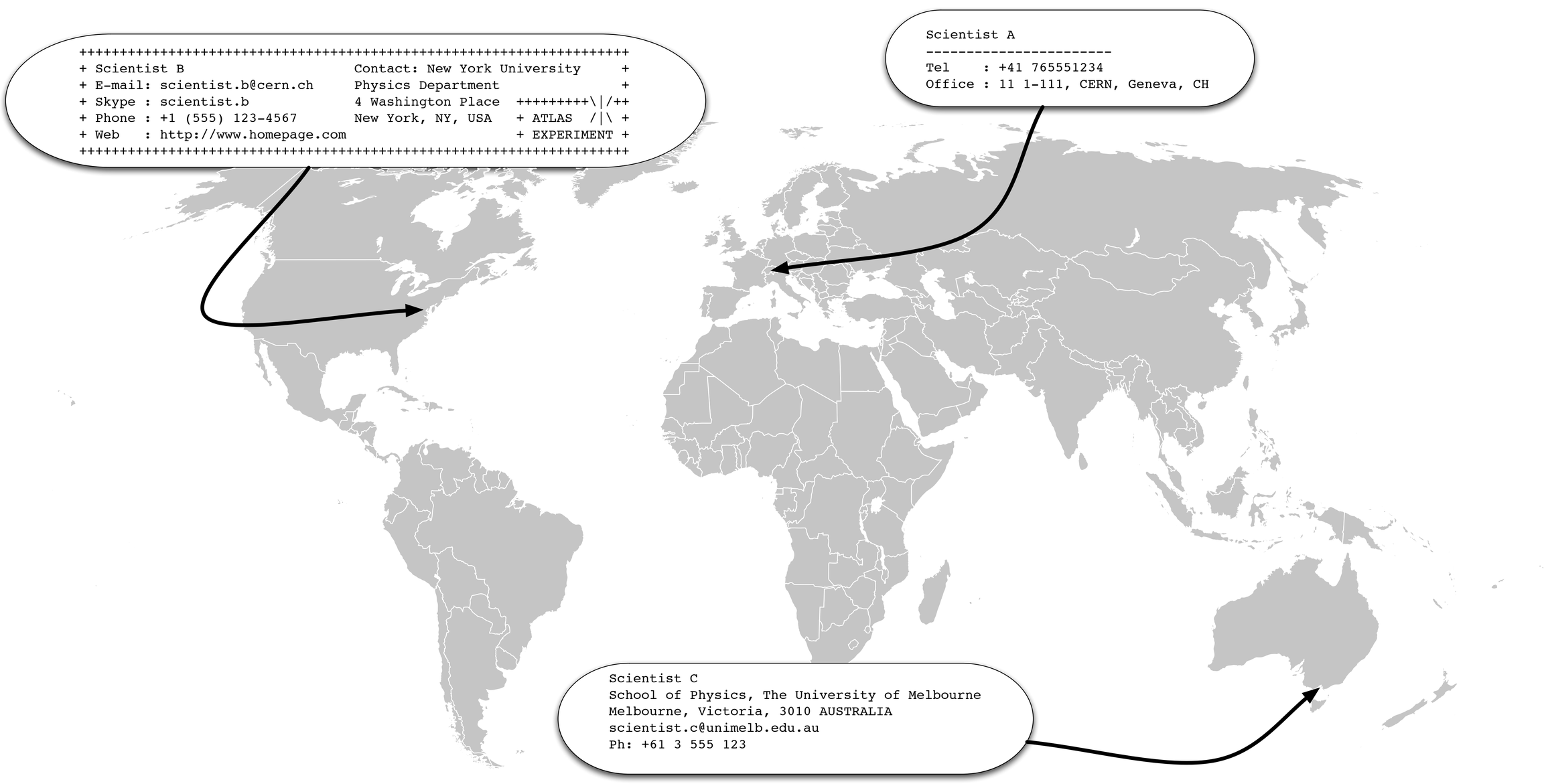
Improving replica placement strategies using information from existing communication infrastructures

Mario Lassnig, Mark Michael Hall
CERN & University of Innsbruck, Austria, mario.lassnig@cern.ch
University of Cardiff, UK, m.m.hall@cs.cardiff.ac.uk

In highly data-driven environments such as the LHC experiments a reliable and high-performance distributed data management system is a primary requirement. Existing work shows that intelligent data replication is the key to achieving such a system, but current distributed middleware replication strategies rely mostly on computing, network and storage properties when deciding how to replicate data-sets across a global set of data centres. We present an approach for improving existing replication strategies based on geographical data available in existing communication infrastructures of scientific experiments. Information on the geographical distribution of scientists is extracted from an existing infrastructure using automated analysis of locational expressions in research documentation, operational logbooks, e-mail correspondence or web presences. Combined with the linking of data-sets to interested users this allows for an anticipatory data replication strategy for data placement at locations close to their interested users.



We have implemented two generic web-service prototypes to parse texts for locational expressions and data-set identifiers. These identifiers map persons to data-sets, therefore expressing interest. In our demo application, we are parsing e-mail signatures for locational expressions; including resolution of facilities, addresses and telephone numbers. The results are returned as a JSON structure, possibly with multiple ranked locations, and the confidence of the resolution. Then, possible sites are matched against the highest-ranking locations and merged as a ranked list for data-sets that should be proactively transferred to the involved sites.