[topic: Classical Languages in a Million Book Library]
David Bamman and David Smith

With the rise of large open digitization projects such as the Internet Archive and Google Books, we are witnessing an explosive growth in the number of source texts becoming available to researchers in historical languages. The Internet Archive alone contains over 12,585 texts catalogued as Latin, including classical prose and poetry written under the Roman Empire, ecclesiastical treatises from the Middle Ages, and dissertations from 19th-century Germany written – in Latin – on the philosophy of Hegel. At 1.7 billion words, this collection eclipses the extant corpus of Classical Latin by several orders of magnitude. In addition, the much larger collection of books in English, German, French, and other languages already scanned contains unknown numbers of translations for many Latin books, or parts of books.

The sheer scale of this collection offers a broad vista of new research questions, and we will focus in this paper on both the opportunities and challenges of computing over such a large space. The availability of massive corpora documenting over two millennia of usage begins to offer insight into grand questions such as the evolution of a language over both time and space, but we must contend as well with the noise inherent in a corpus that has been assembled with minimal human intervention, along with the massive computing power that such a scale often requires.