

This presentation considers the foundations for a new generation of Greek and Latin textual corpora and scholarly services, based upon very large collections of image-books and the emerging elements of a cyberinfrastructure for the humanities in general and for classics in particular. Large new collections offer several opportunities and challenges.

- They will include the entire text of critical editions, thus making it possible for scholars for the first time to search not only for the reconstructed text but for the variants as well. Since the variants can account for 1-10% of the Greek and Latin words in a standard edition, this is a substantial step forward.
- They will include multiple editions of the same work. These can be compared against each other and the differences between them automatically calculated. For the first time scholars will be able to compare multiple editions, measuring their relationship to one another and visualizing the change of texts over time.
- They will include very large bodies of primary materials. While the corpus of classical Latin is fairly small (5 million words up through 200AD), the body of Latin expands over time: Johann Ramminger has collected 100 million words of early modern Latin, already in electronic form, from Petrarch to 1700 (<http://www.neulatein.de/>). We will be able to extract from these libraries a larger and more diverse corpus of Latin than has ever existed.
- They will include monographs and articles which contain many citations of, and quoted excerpts from, the primary sources. These citations and excerpts can be automatically mined and linked to the primary sources on which they comment.
- They will include commentaries, grammars, lexica, translations, and other reference materials that shed light upon the primary sources and that are broken down into smaller logical chunks that can be abstracted and linked to the passages on which they comment.
- They will contain substantial older scholarship of enduring value, reflecting the antiquity and early professionalization of classics as a field. An infrastructure built with older materials is useful in itself. There are multiple public-domain critical editions for virtually all major classical authors. For many less widely studied authors, the nineteenth century editions remain standard. In addition, the form of classical editions has changed relatively little since the nineteenth century. Thus, an initial collection built around older materials will provide concrete models and working code that could be applied to collections with materials under copyright, whether by commercial entities such as Google and Microsoft or libraries with access to substantial materials digitized by Google and Microsoft (e.g., Michigan, California and Cornell).