For all theories of legal interpretation, it matters what words and sentences conventionally mean. And for Originalists and Non-Originalists, Textualists and Intentionalists alike, it is relevant what the words and sentences of the U.S. Constitution and its amendments conventionally meant – and implied – in the times and places of their passage.

As an empirical foundation for interpreting foundational American texts, and especially the U.S. Constitution and the Bill of Rights, we propose to create the Founders’ Corpus. This digital collection will include as much as possible of the English-language material that educated readers of the late 18th and early 19th centuries might have encountered. It will also include accurate metadata descriptions that allow modern researchers to focus on patterns of usage in relevant subsets, chosen by date, genre, topic, and so on. We believe that it will be easy to obtain on-going external support for this enterprise, once its credibility has been established by a pilot project.

Including books, periodicals, legislative and legal documents, letters, and so on, the relevant collection is potentially very large and not clearly bounded; but a useful starting point is now available in scattered form. However, at least four problems stand in the way of practical use.

First, available works are scattered in many places. Thus the digital text of a 1984 edition of Thomas Jefferson’s 1781 Notes on the State of Virginia is available in html form from the Electronic Text Center of the University of Virginia Library. The same work is available from Google Books in several versions, e.g. as a facsimile of an 1832 edition, along with a good-quality plain text version created by optical character recognition. And this work is also available in several versions from the Internet Archive, along with OCR-created text of generally poor to fair quality.

Second, there are many diverse formats. In some cases, material may only be available as facsimile images, which must be mapped to digital text by optical character recognition (OCR) or human transcription or both. Digital texts may be in html, xml, plain text, or pdf form. Mark-up (such as html or xml) is variable and difficult to remove in a consistent way, and even plain text forms may include copyright notices, page headers and footers, and other interruptions of the content. Coherent indexing and search requires mapping into a consistent format. Some users will want a well-designed search interface, such as the one that Mark Davies has constructed at corpus.byu.edu; other users will want their own copy of the entire dataset.

Third, the quality of available metadata and of the texts themselves is often poor. Thus the Google Books metadata for the cited edition of Notes on the State of Virginia lists the publication date as 1832 (which is true for the cited edition but not for the original work), and the genre of the work is given as “Tobacco Industry”. And the quality of OCR is often fair at best – see the samples listed below from works in the Internet Archive’s digitization of the John Adams Library.

Fourth, the intellectual property issues are often complex, even though the underlying works are out of copyright. Thus the Conditions of Use for the Electronic Text Center of the University of Virginia state that “These ebooks, texts and images may not be re-published in print or electronic form without permission from the Electronic Text Center”.

The goal of the proposed pilot project is to establish viable methods for dealing with these four problems, and to create a core dataset of high quality and significant size. The first steps will therefore be to collect a significant portion of the large amount of material that is more-or-less
openly available, to negotiate solutions to any legal access issues that may arise, to put the material into a consistent form as high-quality digital text, to provide it with appropriate metadata, and to index it for immediate use.

Our current plan is to start with two lists: (1) The 2,640 books in Thomas Jefferson’s 1783 Catalog of Books (the Catalog is available in facsimile from the Massachusetts Historical Society, and in transcribed form from the Thomas Jefferson Foundation; and (2) the John Adams Library, comprising more than 2,700 books, which was deposited with the Boston Public Library in 1894, and has recently been digitized in full by the Internet Archive thanks to a grant from the Sloan Foundation. In both cases, works that are not in English will be excluded.

These two lists pose different problems. In the case of Jefferson’s catalog, we will need to transform entries like this

![Image of a book entry]

into a fuller form (in this case, William Roscoe, *The Life of Lorenzo de’ Medici, called The Magnificent*). Then we will need to determine which editions might be available in which form (in this case, at least four editions are available from Google Books, the earliest a “third edition, corrected”, dated 1797, while the sixth edition, dated 1825, is also available from the Internet Archive), what format and quality of digital text are available (in this case, good-quality OCR), under what legal restrictions (in this case, Google Books general Terms of Service, which says that “You may not sell, rent, lease, distribute, broadcast, transfer, or assign your rights to the Digital Content or any portion of it to any third party except as expressly permitted by Google. Provided, however, that nothing in the Terms of Service shall prohibit any uses of Digital Content that would otherwise be permitted under the United States Copyright Act.”).

We will then choose an edition, download whatever forms are available, negotiate any permissions that may be needed, and add a version in a standard digital-text format to our collection, with appropriate metadata.

In the case of the John Adams Library, the problems are somewhat different. The digitization of the catalog and the library itself has already been done. However, the quality of the OCR text leaves something to be desired – these are typical samples:

Mostinfecls, from the time of their being excluded from the egg, till they cease from feeding, wear a different form from that which they put on in their perfect state.

An aft providing a naval armament” patted p. 12. on the first day of July 1797, expired at the end of the last session of Congress.
There are several alternative solutions. One is to have human editors correct the OCR output, with reference to the original page images when needed. Another is to have people transcribe the material over again from the page images. A third is to post-process the OCR output using “language models” trained on typical text of the period, and on the observed statistical relationship between OCR output and the true text.

A fourth alternative is to re-do the OCR using better (or better-trained) systems.

In past projects, LDC staff have used all of these methods – the choice must be made on the grounds of relative cost and quality, and the outcome depends on the specific characteristics of the material.

Working with these two collections should enable us to explore plausible solutions for the problems arising in all aspects of the project, including the metadata, the format(s) of the text, the IPR and other legal issues, the acquisition/transformation of digital text from sources of various types, and so on. This exploration should give us a solid basis for obtaining additional funds to continue the project.

In addition, we propose to include a small research project, to validate the benefits of such a collection to evidence-based interpretation. It is well known that norms of comma placement were different in the 18th century – this issues arises, for example, in the case of the second amendment, which reads

A well regulated Militia, being necessary to the security of a free State, the right of the people to keep and bear Arms, shall not be infringed.

Although there is a significant scholarly literature on the history of punctuation in English, there has not to our knowledge been a careful corpus-based study of comma usage in the 18th century. Even in the early stages of our project, when the proposed collection contains only a few hundred books, there should be enough material of enough diversity to support a careful quantitative study of the syntactic and discourse-structural patterns of comma placement.

The proposed project will clearly require a great deal of skilled labor. But luckily, many of the skills involved are among those routinely required of Penn undergraduates; and many of the remaining skills are fairly easy for students to learn. The Linguistic Data Consortium (LDC) has more than 20 years of experience in creating large text and speech datasets, by recruiting, training, and supervising research assistants who are often Penn undergraduates. As each LDC project proceeds, some of those involved move up to train and supervise others; some become engaged in more technical aspects of the project; and some become co-authors of associated conference papers and of the final datasets themselves.

We expect to use a similar approach for this pilot project, using a quarter of the time of an experienced staff manager at the LDC, along with ten to fifteen undergraduate research assistants. The undergraduates will be encouraged to take on more complex and independent technical and scholarly aspects of the project, as their interests and skills permit. All needed equipment, facilities, and technical support, along with help negotiating legal agreements, will be provided gratis by the LDC.
For encoding digital texts, we plan to use a form of the Text Encoding Initiative (TEI) standard, and for the metadata, we will use a form of the Machine Readable Cataloging (MARC) standard from the Library of Congress, or (if it seems mature enough) a form of the new Bibliographic Framework (BIBFRAME) now under development at LOC.

Overall technical and editorial guidance will be provided by the members of an expert Advisory Board, whose members now include Rick Beeman, (John Welsh Centennial Professor of History at Penn) Mark Davies (Professor of Corpus Linguistics at BYU), Bryan Garner (author of *Reading Law: The Art of Interpreting Legal Texts*), Neal Goldfarb (Senior Counsel at Butzel Long Tighe Patton, and author of the Law n Linguistics blog), Jerry Goldman (Professor emeritus of Political Science at Northwestern University, and founder and director of the Oyez Project, www.oyez.org), Barbara Grosz (Higgins Professor of Natural Sciences at Harvard University, and former director of the Radcliffe Institute for Advanced Study), James Herbert (former director of the Division of Education and Research, National Endowment for the Humanities), Brewster Kahle (founder and director of the Internet Archive), Tony Kroch (Edmund J. and Louise W. Kahn Endowed Term Professor at the University of Pennsylvania, and developer of the Penn-Helsinki Parsed Corpora of Historical English), Erez Lieberman Aiden (co-founder of Harvard’s Cultural Observatory and co-developer of the Google Books ngram viewer), Joan Maling (NSF program director for Linguistics), Peter Norvig (Director of Research, Google Inc.), Geoffrey Nunberg, Martha Palmer (Professor of Linguistics at the University of Colorado at Boulder), Geoffrey Pullum (Professor of Linguistics at the University of Edinburgh and co-author of *The Cambridge Grammar of the English Language*), Stu Shieber (James O. Welch, Jr. and Virginia B. Welch Professor of Computer Science and Director of the Office of Scholarly Communication at Harvard University), and Lawrence Solum (John Carroll Research Professor of Law at Georgetown University, and editor of the Legal Theory Blog).