NATIONAL ENDOWMENT FOR THE HUMANITIES

OFFICE OF DIGITAL HUMANITIES

Narrative Section of a Successful Application

The attached document contains the grant narrative and selected portions of a previously funded grant application. It is not intended to serve as a model, but to give you a sense of how a successful application may be crafted. Every successful application is different, and each applicant is urged to prepare a proposal that reflects its unique project and aspirations. Program guidelines also change and the samples may not match exactly what is now required. Please use the current set of application instructions to prepare your application.

Prospective applicants should consult the current Office of Digital Humanities program application guidelines at https://www.neh.gov/grants/odh/digitalhumanities-advancement-grants for instructions.

Applicants are also strongly encouraged to consult with the NEH Office of Digital Humanities staff well before a grant deadline.

Note: The attachment only contains the grant narrative and selected portions, not the entire funded application. In addition, certain portions may have been redacted to protect the privacy interests of an individual and/or to protect confidential commercial and financial information and/or to protect copyrighted materials.

Project Title: Bridge Readability Tools

Institution: Haverford College

Project Directors: Brett Mulligan, Patricia Guardiola, Anna Lacy

Grant Program: Digital Humanities Advancement Grants, Level II

Name (Project Role)

Bret Mulligan (PI/Project Director, Haverford College)

Patricia Guardiola (Co-I, Haverford College)

Anna Lacy (Co-I, Haverford College)

Undergraduate Research Associates: 14 total positions for academic year (part-time) and summer (fulltime) employment, hired annually with hope of some continuity between academic year and summer positions.

6 Full-time Summer URAs: 1 internal-funded per year; 2 grant-funded per year

8 Part-time Academic Year URAs: 2 internal-funded per year; 2 grant-funded per year

Advisory Board

David Bamman, University of California, Berkeley
Rebecca Boyd, George Washington University
Patrick J Burns, New York University
Nava Cohen, Northwestern University
Neil Corrigan, cipolyglot.com
Christopher Francese, Dickinson College
Stella Fritzell, Bryn Mawr College
Clara Hardy, Carleton College
Andy Janco, University of Pennsylvania
Daniel Libatique, Fairfield University
Emily Lewis, South Lakes High School
Ivy Livingston, Harvard University
Dominique Longrée, Université de Liège
Adrienne Lucas, University of Delaware
Hugh McElroy, The Field School
Stephen Sansom, Florida State University
William Turpin, Swarthmore College

1. Overview of the Project

Scholarship on language acquisition in the humanities and beyond confirms the importance of reading at the appropriate level for language learners.Yet teachers and readers of historical languages have little more than anecdotes to help us understand the readability of texts and identify comprehensible assignments. Is this poem, story, author, or genre more accessible to particular students than another? What readings might best prepare students to approach a target text? In the absence of such data, editors, teachers, and professors can only depend on established practices when assaying textual readability. Without accessible tools to support such assessments, pedagogical effectiveness is hindered, innovation curtailed, and proficiency diminished.

This **Level II grant** will advance research on readability and promote the instruction of language acquisition and efficient reading through the swift development of three pedagogical and scholarly tools for the *Bridge* — an established digital ecosystem that supports the digital processing and reading of Ancient Greek and Latin at the secondary, collegiate, and professional levels. These three apps will:

(1) produce practical assessments of textual difficulty (i.e., *readability*) via ***Stats***;
(2) facilitate the discovery of readable texts via ***Oracle***;
(3) democratize and accelerate the accurate encoding of lexical data via ***Lemmatizer2***.

With the right tools, teachers will be able to accurately identify readable, relevant texts for their students. This data-informed approach to text selection will systematically connect course objectives with appropriate materials to create more inclusive, equitable, and effective pedagogies. The *BRT* project will facilitate these connections for historical languages in secondary and collegiate classrooms, as well as for independent learners. Foremost in the minds of the project team will be the task of making these potentially complicated data easily accessible and comprehensible by a wide audience.

2. Enhancing the Humanities

The challenge of identifying compelling and readable texts extends to all learning environments in which reading is a mode of textual engagement, but the issue is especially relevant for the humanities, which center reading and language at their core across most disciplines. We view the ***Bridge Readability Tools*** as a model system that will pioneer the longer-term goal of supporting the teaching and accessibility of other languages, beginning with Ancient Greek and then other historical languages, with extensibility to other commonly taught modern languages, across a global spectrum. During the grant period, we will focus first on developing tools for Latin because of our local expertise, the large audience of Latin teachers and students, the success of the existing *Bridge* ecosystem in supporting Latin pedagogy, and the maturity of supporting digital resources. But since the *BRT* apps will be designed for use with any language for which Natural Language Processing (NLP) resources exist, it has potential applications far beyond its initial target audiences at schools, colleges, and universities around the world.

Why These Tools are Needed The reading and study of many historical languages are on the horns of a dilemma. These languages often comprise vast corpora — in the case of Latin estimated at over a trillion words — yet a typical Latin student might engage texts totaling just a few tens of thousands of words (or a mere 0.000002% of the total corpus). Within this small slice, novice readers routinely move directly from fabricated Latin in textbooks to difficult historical texts, whose reading grade level is akin to college-level texts (Gruber-Miller & Mulligan). To attain full comprehension, readers must typically know 95 to 98% of the words in that text (Hu and Nation). Yet many novice readers routinely know only 25% of the words in commonly-taught texts.

While the statistics vary across language fields, the overarching concerns are the same. Instructors and independent learners have begun to pay attention to this dilemma, but there are no accurate and easily accessible tools to analyze the readability of Latin texts. Additionally, no tools are available to help language instructors identify those texts (or sections of texts) that are most lexically relevant to the texts already taught. Texts, therefore, are typically selected because of tradition or familiarity rather than their ability to promote language acquisition. The *BRT* apps will facilitate the data-informed assessment of

texts for use in language learning, allowing readability — and so efficient language acquisition — to be built fully into course and curriculum development.

What These Tools Will Do The *BRT* apps allow an instructor or independent learner to (A) analyze and compare the readability of texts; (B) discover readable texts for data-informed lesson plans, syllabi, and curricula; and (C) encode texts for analysis in this and other digital ecosystems. While these tools are mutually supporting, they may be used independently, increasing their pedagogical and scholarly utility. As part of the development roadmaps for each *BRT* app, we identify a ***Benchmark*** for the successful implementation of its core functionalities; we then identify ***Stretch*** goals that will guide further development within the period of performance and in the future.

(A) ***Stats*** will be a web-based dashboard that displays information about lexical and syntactic difficulty — i.e., *readability* — for Latin texts, and the effect that user-defined knowledge has on textual readability (**Benchmark**). *Stats* will analyze one or more texts and/or sections for their (1) generic readability; and (2) readability that factors in personalized lexical knowledge. Users can compare generic and personalized statistics for one or more texts. Data summaries and details will be presented numerically and in several visual formats for easier assessment, allowing users to identify more readable text and sections, intentionally select readings, and effectively plan reading activities.

Stats will use high-quality *Bridge Corpus* data to provide basic information about readability: e.g., (1) the number of words in a text; and (2) the size of the text's vocabulary. It will derive standard measures of readability: e.g., (1) word length; (2) word frequency, or the prevalence of very common words; (3) lexical sophistication, or the percentage of rarer words; (4) lexical variation, or the variety of different words; (5) hapax legomena, or words that appear only once; and (6) the corpus frequency of rare and/or unknown words. It will also analyze syntactic measures such as: (1) the number of words per sentence; and (2) the number and length of subordinate clauses. From these data *Stats* will provide standard readability scores (e.g., LIX, RIX, Lambda, ARI, Coleman Liau Index, Modified Dale-Chall, Coh-Matrix, etc.) and scores tailored to Latin (e.g., *LexR*, *SynR*). These analyses can be shared, saved to a *Bridge* user account, or exported for archiving or additional investigation. See Appendix 7.4 for a sample use case, 7.5 for a proof of concept analysis of Latin textual readability, and 7.6 for a sample visualization. **Stretch 1**: support analysis of user-inputted texts using NLP processes to probabilistically lemmatize the text. **Stretch 2**: include support for Ancient Greek and then other languages.

(B) *Oracle* will be a web-based app that allows users to discover lexically readable texts in the *Bridge Corpus* by revealing the authors, texts, and passages that have the highest percentage of familiar vocabulary alongside basic readability data (**Benchmark**). Users will select the author(s), text(s), or genre(s) they would like to explore and then indicate their known vocabulary by selecting textbooks used, lists mastered, and texts previously read. *Oracle* then produces a filterable table of texts and passages ranked by lexical familiarity, as well as visualizations of these data. *Oracle* will allow users to discover readable texts, vocabularies for which can automatically be generated via the *Bridge Lists* app (see Appendix 7.2) and which can be further analyzed and compared via *Stats*. A prototype was developed in 2020, proving the feasibility of the app and allowing the preliminary testing of various use cases. **Stretch 1**: include support for Ancient Greek; **Stretch 2**: support analysis of user-inputted texts.

(C) *Lemmatizer2* will develop an existing tool, *Lemmatizer* (see §2.B, below and Appendix 7.2), into a web-based environment, allowing more rapid, accurate, and detailed lexical and syntactic encoding of texts, and facilitating collaboration by faculty, students, and other contributors (**Benchmark)**.

The protean nature of words in highly inflected languages like Latin thwarts the machine analysis that can be applied to many languages. This hurdle can be overcome by first lemmatizing each word in a text — i.e., matching every word to a unique identifier, often a dictionary headword or *lemma*. Automatic lemmatizers exist but these produce data that are insufficiently accurate for pedagogical resources or analysis. *Lemmatizer2* will improve upon the existing *Lemmatizer* foundation, refining existing automatic lemmatization and integrating this into a web-platform that will enable editors to rapidly complete highly accurate lemmatizations of texts. It will intake a plain-text, demarcate each word, and begin the

lemmatization process by identifying the ~60% of words that match a single lemma. It will also leverage sophisticated machine lemmatization provided by existing resources like *CLTK* and *Latin SpaCy* to provide best guess(es) for the lemmata of ambiguous words, distinguishing these in the output. Users can select the format of the lemmata: either the *Bridge* format (a refined version of the LASLA system that will be linked to the *LiLa* open standard) or the *Morpheus* format (developed by *Perseus Project*), increasing the interoperability of these data.

Lemmatizer2 will also facilitate readability analysis by identifying basic grammatical features (e.g., sentence boundaries, parts of speech, and subordinating elements). Editors can then audit the data, identify remaining ambiguous words, gloss new words, and provide additional markup as required (e.g., grammatical markup, refined subordination tagging, etc.). Texts-in-progress will be stored on-site and linked to *Bridge* user accounts, where they can be worked on by multiple editors, encouraging both collaboration among colleagues and students, and the crowdsourced creation of data. Completed texts can be downloaded for pedagogical or scholarly use or submitted for incorporation into the *Bridge* ecosystem. See Appendix 7.4 for a sample use case. **Stretch 1:** include support for Ancient Greek and then other historical languages. **Stretch 2:** explore how gamification may promote and speed the lemmatization of historical texts.

3. Environmental Scan

Recent developments across diverse disciplines will enable the *BRT* project to expand our knowledge of the readability of Latin texts and foster the use of data in historical language pedagogy and curriculum development. Readability studies are now being applied beyond English to other languages (Xia et. al). Computer analysis of texts has progressed so that analyzing texts by hand is no longer necessary (Feng 2010; Azpiazu et al. 2019; Deutsch et al. 2020). Curated datasets (e.g., the *Bridge Corpus,* LASLA's *Omnia Opera*, Perseus' *Ancient Greek and Latin Dependency Treebank,* the *Concordance Liberation Project*) can expedite the statistical analysis of Latin texts, while the digitization of corpora has created the potential to explore new historical languages and texts (e.g., *CLTK*, the *Open Philology Project*). Meanwhile, trends in Latin pedagogy have refocused attention on the role of vocabulary acquisition in language learning. There exists an interest in assessing the readability of Latin texts at all levels of the curriculum and a newfound capacity to do so.

Yet, there are no resources to analyze historical language readability or discover readable texts. Most existing digital tools are one of two types: **Type 1 Tools** are reading environments focused on providing point-and-click or interlinear vocabulary support; **Type 2 Tools** focus on morpho-syntactic tagging and require technical abilities beyond almost all instructors and students.

Type 1 Tools are frequently used by Latin students to support reading but most use less-than-fully accurate, machine-generated data, and many instructors find that their translation of all words compromises durable language acquisition.

a. **The Perseus Digital Library** has a vocabulary look-up tool but its lexical support is generated from machine algorithms supplemented by crowd-voting, resulting in ambiguous or inaccurate word identifications of homonyms.
b. **Alpheios** is an open-source tool that embeds word lookup and morphological support for any on-line Latin text. It does not differentiate between homonyms because it relies exclusively on machine-generated data.
c. **NoDictionaries.com** is a vocabulary look-up tool that generates interlinear vocabulary based on the **Whitaker's Words** dataset. It does not differentiate between homonyms, except in a small sub-corpus of curated texts.
d. **Hedera** promises to be the only reading environment that reveals the percentage of known words in a text based on a user-defined list of words. The Director of *Hedera* is on our Advisory Board.

Type 2 Tools are sophisticated taggers and analyzers but their technical requirements prevent widespread adoption. Resources from this list may be incorporated into the turnkey *BRT* suite.

e. **Classical Language Toolkit** (*CLTK*) is a Python library that uses NLP to generate linguistic data for many historical languages of pre–modern Eurasia. *Lemmatizer2* will build from *CLTK*'s existing parsers and lemmatizers. A primary developer of *CLTK* is on our Advisory Board.
f. **Latin BERT** and **spaCy** are contextual language models for Latin that achieves state-of-the art part-of-speech tagging and word sense disambiguation for Latin. *Lemmatizer2* will incorporate Latin spaCy's part-of-speech tagging model. The creators of *Latin BERT* and *SpaCy* are on our Advisory Board.
g. **Pie** is a language-independent lemmatizer in Python developed by LASLA that can be trained for many lemmatization tasks. A co-director of *Pie* is on our Advisory Board.
h. **Collatinus** is a lemmatiser and morphological analyser for Latin texts. Its ability to colorize a text according to a list of known words may be incorporated into future versions of *Stats* and/or *Oracle*.
i. **Arethusa** is the tree-banking tool developed by the Perseids Network. It allows the manual tagging of historical texts with a focus on grammatical dependency.

4. History of the Project

The *Bridge Readability Tools* will build from and integrate with the existing infrastructure of the *Bridge* and its corpus of over 200 encoded Latin texts comprising over 2.5 million words. *BRT* will also expand and complement two free lexical resources for Ancient Greek and Latin texts in the *Bridge* ecosystem: (A) *Lists*, and (B) *Lemmatizer* (see Appendix 7.2 for annotated screenshots of each tool and 7.3 for a schematic of the ecosystem). Early in the period of performance, we will improve interoperability with other projects by aligning the *Bridge Dictionary* (and so all words in the *Bridge Corpus*) with the standards recently established by *LiLa* (https://lila-erc.eu/), a Linguistic Linked Open Data project that follows the FAIR principles for scientific data management and stewardship.

(A) *Lists* (bridge.haverford.edu/): *Lists* enables users to generate customized vocabulary lists from its high-quality lexical database of Ancient Greek and Latin texts, core lists, and textbooks. Users can focus on a selection of a work or multiple works and customize their lists to take into account works they have read and words they already know. They can also create lists of words their texts share in common. These lists can then be sorted, searched, and filtered to focus on e.g., part of speech, and then printed or downloaded. The most popular *Bridge* tool, usage averages over 1500 sessions by 600 unique visitors per month. Launched 2014; current version launched 2022. *Lists* was favorably reviewed by the *Society for Classical Studies*, which deemed it "a model of a collaborative digital project that can draw on funding and labor from a number of institutions to create an open resource that helps all teachers and students" (Pistone 2020).

(B) *Lemmatizer* (bridge.haverford.edu/lemmatizer/): The current *Lemmatizer* creates a lemmatization spreadsheet for an Ancient Greek or Latin text, lemmatizes unambiguous words, and downloads the sheet as a CSV file. This rudimentary tool has aided the lemmatization of dozens of Ancient Greek and Latin texts, establishing the feasibility of this method and the support for the product. Launched 2019; current version launched 2020.

Other Integrations The *BRT* project will build on research in Natural Language Processing (NLP) machine learning, and will benefit from existing collaborative relationships with digital humanists and text archives. The project will enhance Haverford's well-established partnership with the *Laboratoire d'Analyse Statistique des Langues Anciennes* (LASLA) at the Université de Liège, which has provided lexical and syntactic encoding for many Latin texts in the *Bridge Corpus.* The PI and Co-Is will work with the student team members to integrate NLP into *Lemmatizer2*, increasing the accuracy of the encoding and analysis of Latin texts and speeding their incorporation into the *Bridge* ecosystem. The project will build on recent technical work — e.g., *Latin spaCy*, a pre-trained NLP model by Burns (on Advisory Board) — to automatically identify parts-of-speech and other linguistic features.

Project History Since the launch of the *Bridge* project in 2014, we have established a track record of rapid implementation, thorough evaluation (via internal testing, and feedback from Board Members,

invited testers, and public users), and quick iteration of popular and effective tools using limited in-house resources.

2014	Planning by Bret Mulligan (PI) and Haverford Librarians Laurie Allen and Michael Zarafonetis. Student developers, Julie Ta and Blair Rush, prototype *Lists* for Latin texts.
2015–2017	Expansion of *Lists* to include Greek texts and iterative improvements by Ta, Jack Raisel, Byron Biney, and Dylan Emery with guidance of Librarian Andrew Janco. Creation of command-line *Lemmatizer* by Raisel and James Faville.
2018–2019	*Lists* revised by Aleena Maryam and GUI prototype of *Lemmatizer* by Noor Fatima.
2020–2023	Faster, more capable *Lists* and *Lemmatizer* released, and prototype of *Oracle* by Carter Langen, Fiona Xu, Samuel Tan, Gulesh Shukla, Yiting Zhou. Collaboration with *MehtA+'s Research-Based Bootcamp* to experiment with NLP processing of the corpus. Current work focused on database implementation in preparation for the *BRT* expansion.

Initial data for the *Bridge* was compiled by Mulligan and undergraduate students from Bryn Mawr and Haverford Colleges. Data contributors now include 11 secondary school students, 23 undergraduates, 14 graduate students, and 13 instructors at high schools, colleges, and universities (credits available at bridge.haverford.edu/about/texts). Additional data development has been provided by the *Laboratoire d'Analyse Statistique des Langues Anciennes, FeminaeRomanae.org,* the *Ancient Greek and Latin Dependency Treebank Project*, and the *Concordance Liberation Project*. Financial support for the development of the *Bridge* was provided by Haverford College (2014–2023), a *Classical Association of the Atlantic States Program Grant* (2015), and a *Mellon Digital Humanities Grant* (2014–2015).

5. Activities and Project Team

The period of performance will feature two mutually-supporting activities. (1) We will develop the three *BRT* apps following our established protocol: (a) ***analysis*** of requirements and existing resources; (b) ***design*** of a prototype; (c) ***implementation*** of a beta; (d) ***testing*** and iterative improvement based on feedback from internal testers, members of the Advisory Board, and invited testers; (e) public ***deployment*** of the tool; and (f) ***maintenance*** and ***improvement*** based on testing, feedback, and fulfillment of stretch goals. (2) We will continue ***maintenance*** and ***improvement*** of existing *Bridge* tools, completing their development roadmaps and improving the data in the *Bridge Corpus*.

Colleagues, members of the Advisory Board, and other invited contributors will evaluate each of the tools during prototyping and beta testing periods. The *BRT* apps will have an open testing period when existing users will be invited to offer feedback. Their impact will be assessed via user surveys targeted to different constituencies and usage metrics. After thorough testing, we will submit the *BRT* apps for peer-review through the *SCS Digital Project Review* and *Reviews in Digital Humanities*. Within 90 days after the period of performance we will submit a white paper to the NEH on lessons learned.

Project Team

Bret Mulligan (PI): academic year salary and benefits provided by Haverford College, of which 1 month will be devoted to the project. The grant will provide two months of summer support to work on the project. The project will advance Mulligan's scholarship on readability and Latin pedagogy.

Patricia Guardiola (Co-I): salary and benefits provided by Haverford College, with 5% work time dedicated to the project (internally funded). The project will help Guardiola build new capacities for web development and NLP in the Haverford College Libraries' Digital Scholarship program.

Anna Lacy (Co-I): Salary and benefits provided by Haverford College, with 5% work time dedicated to the project, including mentorship of the undergraduate students (internally funded).

16 Undergraduate Research Associate(s): Students, mentored by Lacy and Mulligan, will work during the academic year and the summer to develop the *BRT* apps and data, gaining practical skills in programming, text analysis and web development while engaging with language pedagogy and historical sources and gaining experience with academic research. During each year of the project there will be 3

full-time summer URAs (1 internal; 2 grant-funded) and 4 part-time academic year URAs (2 internal; 2 grant-funded). Summer students may continue during the academic year and vice versa.

6. Final Products & Dissemination

Final Products By the conclusion of the period of performance, the three constituent tools of *Bridge Readability Tools* will be freely available on the bridge.haverford.edu website: (1) *Stats* will analyze the readability of texts in a user-friendly dashboard; (2) *Oracle* will help instructors discover comprehensible texts, and foster data-informed syllabi and curricula; (3) *Lemmatizer2* will speed the collaborative lemmatization of the vast corpus of Latin texts (c. 200 BCE – present).

The expansion of the *Bridge* into an online collaborative platform, repository, and analytics suite for historical languages, beginning with Latin as a model system, will benefit scholars, instructors, students, and independent learners. By creating a suite of tools that instructors can use to design and manage their curricula and deliberately connect course linguistic content to course readings, the *BRT* apps will facilitate the teaching of language in context while infusing the curricula of historical languages with the insights of research on language pedagogy.

Dissemination In the spirit of existing collaborations and to expand access to these tools, all work and byproducts produced by the project will be freely available to all users and licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. Project code will be posted to an open GitHub repository with documentation that allows developers to build from the project's codebase. We will continue to store technical documentation in the GitHub repository and follow best practices for test-driven development, docstrings, and code style for Python (PEP8).

Mulligan will promote these tools on social media and apply to present at major academic conferences during AY 2023–26 (e.g., *Classical Association of the Atlantic States, Society for Classical Studies, American Classical League, Classical Association of the Mid, West, and South, Classical Association of New England*). Mulligan will seek opportunities to conduct training sessions at e.g., pre-conference workshops and instructor development days, as he has already done for existing tools. Mulligan will submit a description of the project to the blog of the *Society for Classical Studies* as prelude to a peer-reviewed publication(s) in venues devoted to innovative pedagogy (e.g., *Teaching Classical Languages, Classical Outlook, Classical Journal's* "Forum", etc.). Mulligan will inquire about presenting the project at Liberal Arts Collaborative for Digital Innovation (LACOL) meetings and will seek out other venues in which to disseminate results, partnering with student collaborators as feasible.

Accessibility All tools will be freely available on the bridge.haverford.edu website. Accessibility and user experience design are a key part of our development process and we consistently test and evaluate the accessibility of our tools. The Digital Scholarship program at Haverford regularly works with J.D. Dougherty (Professor of Computer Science, HC) to plan for diverse user needs from the beginning of project work and design. For *BRT* we will continue these protocols and prioritize design that adheres to Web Accessibility Initiative's "Accessibility Fundamentals" (www.w3.org/WAI/). The current version of the *Bridge* uses the *FastAPI* framework, which adheres to the principles of minimal computing design. As a result the *Bridge* is secure, easy to maintain, and minimizes extraneous data transfers, making the tool accessible to users with slower-speed internet connections.

Crediting Collaborators We will continue the *Bridge* project's practice of visibly crediting all contributors. Directors, developers, and advisors appear on the "People" page (bridge.haverford.edu/about/people). Contributors of data are credited on the "About the Texts" page (bridge.haverford.edu/about/texts). Linked icons for institutional partners (LASLA, Dickinson College Commentaries, FeminaeRomanae.org) appear in the footer of every page.

WORKPLAN: ***Bridge Readability Tools for Historical Language Text Analysis and Data-Informed Pedagogy***

An NEH-funded *Bridge Readability Tools* project will drive research on the readability of historical languages and fund the creation of three innovative tools for teachers and scholars during AY 2023/24 & 2024/25. As this project involves multiple constituencies contributing to several independent yet interrelated products, each of the tools (*Stats, Oracle, Lemmatizer2*) will be developed using the same protocols that has allowed us to develop and refine existing *Bridge* tools, while safeguarding against workplan disruptions (see below). Our planning includes ***Benchmarks*** for successful implementation of each tool and ***Stretch*** or aspirational goals that we will strive to complete during the period of performance. This workplan includes timelines for: the analysis, testing, implementation, and maintenance of each tool (1a–c); and (2) the continued development of legacy Bridge tools, data, and resources.

Activities	PERIOD OF PERFORMANCE							
	AY 22/23		AY 23/24			AY 24/25		
	Spr 23	Sum 23	Fall 23	Spr 24	Sum 24	Fall 24	Spr 25	Sum 25
1a. Bridge/Stats								
Analysis: continue readability research; refine use cases & requirements; identify existing resources (readability, tokenizers, syllabizers) and clear, feasible representations and visualizations of data								
Design: create wireframe, UI/UX decisions, & mockup								
Implementation: create prototype of web dashboard that displays text data and lexical readability scores and supporting data for Latin texts; gather feedback from Advisory Board and invited testers								
Phase 2 Analysis, Testing, & Redesign: as above, research more advanced readability metrics (e.g., NLP, cohesion, & topic analysis); create revised wireframe, UI/UX decisions, mockup; gather feedback								
Implementation: ***Benchmark*** : launch beta with lexical and syntactic readability scores and data for Latin texts, modifiable by user-defined knowledge; integration with *Bridge* user accounts. ***Stretch 1*** : support analysis of user-inputed texts; ***Stretch 2*** : support other historical								
Testing: iteratively test all use cases & accessibility; improve based on feedback from Advisory Board and invited testers								
Deployment & Maintenance: release to public; promote; provide production support; develop new features based on public user feedback; correct emergent issues; develop remaining Stretch Goals								
1b. Bridge/Oracle								
Analysis: refine use cases & requirements								
Design: create wireframe, UI/UX decisions, mockup, & prototype; gather feedback from Advisory Board Members and invited testers								
Implementation: ***Benchmark*** : launch app to discover texts with known vocabulary and syntax in context; integration with *Bridge* user accounts; ***Stretch 1*** : include support for Ancient Greek and then other historical languages; ***Stretch 2*** : support analysis of user-inputed texts								
Testing: iteratively test all use cases & accessibility; improve based on feedback from Advisory Board, invited testers, and public users								
Deployment & Maintenance: [same as in 1a]								

Activities	Spr 23	Sum 23	Fall 23	Spr 24	Sum 24	Fall 24	Spr 25	Sum 25
1c. Bridge/Lemmatizer2								
Analysis: establish use cases & requirements; identify existing resources (backoff lemmatizers, machine learning, NLP), model collaboration environments, and gamification models for collaborative markup								
Design: create wireframe, UI/UX decisions, mockup, & prototype; gather feedback from Advisory Board and invited testers								
Implementation: ***Benchmark*** : develop beta with more rapid, accurate, & more detailed lexical & syntactic encoding of Latin texts; integration with *Bridge* user accounts for collaborative lemmatization; ***Stretch 1*** : add support for Ancient Greek; ***Stretch 2*** : gamify lemmatization								
Testing: iteratively test all use cases & accessibility; improve based on feedback from Board Members and invited testers								
Deployment & Maintenance: [same as in 1a]								
2. Legacy Bridge Tools, Data, & Resources								
Lists: Complete FastBridge feature list (e.g. filter by word frequency, enhanced export options); complete database implementation								
Bridge: complete preliminary development of user accounts; *Lists* integration to allow storage of searches & word lists								
Dictionary: Remove near duplicates & augment data & definitions; *LiLa* integration to promote project interoperability								
Dictionary: ***Stretch*** : Cluster lemmata for user-selected dynamic lemmata refinement (e.g., map 'love', 'lovely', 'lover' to each other and 'to love')								
Bridge: Ongoing maintenance, enhancement, & promotion of legacy tools and data; continued addition of new texts to the Bridge Corpus								

Risk Assessment and Mitigation

Software Development Delays: *Risk level: moderate* . The pace of development may lag during some periods because our undergraduate partners begin with variable technical and professional proficiencies. Prior experience allows us to construct realistic timelines. Having multiple personnel engaged in the project will further mitigate this risk. Our design of three independent apps further safeguards against delays in one compromising the others. Indentifying Benchmark and Stretch Goals helps ensures successful implementation. We will continue to document development to help personnel quickly acclimate to the project and reduce delays during personnel transitions.

Personnel Disruption: *Risk level: low.* Mulligan is tenured and (b) (6) during the period of performance; if Guardiola or Lacy should leave Haverford, the other Co-I can cover obligations until a replacement is hired. The cohort structure of the Undergraduate Research Associates fosters continuity, despite some anticipated turnover between years.

Travel Restrictions: *Risk level: minimal.* Promotion at conferences could be done remotely. Consultation with Board Members and other advisors will be done remotely.

DATA MANAGEMENT PLAN

Roles and Responsibilities

The execution of this data management plan is the responsibility of the project PI, Bret Mulligan, and the Co-Is, Patricia Guardiola and Anna Lacy. Project data and digital assets will be maintained in repositories of the Haverford College Libraries' Digital Scholarship Program. Additionally, the Haverford College Archivist will assess project materials for long-term preservation in the College's digital asset management system.

Expected Data

In the course of project work, we will create the following types of data:

- Lemmatized lists of words that appear in the texts. The lists, accessible in the project in a SQLite database, can be serialized and stored as json and CSV. This data will be retained in an open GitHub repository and will be served to the web through the Bridge web application.
- Scripts and web application code for assessing the readability of the texts. All project code will be archived weekly in Digital Ocean droplet, along with snapshots taken at key points in the development process. They will also be maintained in a public GitHub repository as part of the GitClassical organization and maintained by Digital Scholarship following a three-year preservation plan.
- Data from the Bridge user accounts will not be shared or made public. We will use Oauth authentication so that only a temporary token will be granted by Google for user access. All user data, including usernames and passwords, will be managed by the user using Microsoft Azure (Haverford account standard) authentication services.
- Text files which may include original transcriptions from Haverford Special Collections and open source texts from Perseus and similar repositories. The project text corpus will likely be comparable to the corpus used to create LatinBERT, which includes texts from the Corpus Thomisticum (14.1 million tokens), the Internet Archive (561.1M), Latin Library (15.8M), Patrologia Latina (29.3M), Perseus (6.5M) and Latin Wikipedia (15.8M) (see: https://github.com/dbamman/latin-bert). Long-term retention of open-source texts will be decided by the College Archivist. Digitized and transcribed materials belonging to the College will be maintained as part of the College Archives.

Period of data retention

All data will be available during the project period and maintained afterward by Haverford Digital Scholarship librarians as part of the Libraries' portfolio of ongoing digital projects. Software created during the period of performance will be maintained and improved by Haverford's digital scholarship librarians and student developers. Maintenance and development will continue so long as the project is active (it currently has no anticipated date of termination), after which the project will be archived in accordance with Haverford's digital preservation policy (currently under development).