NATIONAL ENDOWMENT FOR THE HUMANITIES

SAMPLE APPLICATION NARRATIVE

JISC/NEH Transatlantic Digitization Collaboration Grants
Institution: Tufts University and Imperial College Internet Centre

The advent of immense digital libraries has changed not only the way humanities researchers and students access content, but also user expectations as well. A variety of sophisticated tools and processes offer enhanced collection support, linking digital works with lexica, authority lists, word analysis, gazetteers, maps, and more. Obscure references and allusions that may have once required multiple trips through library stacks are now found with the click or roll over of a mouse. Presenting gigabytes of data in a single, queriable system has made possible previously impractical forms of hypothesis and visualisation.

Enabling easy and fast access from anywhere on the Internet to powerful content enrichment tools is the next great challenge in the evolution of digital libraries. Advances in Internet technology now make it possible to consider extending these previously website-bound services and data aggregations to the entire Internet community. We can now envisage a user who gains instant access to contextualization services simply by dropping a scan of a book into a folder on the desktop. And that user's annotations, corrections, disambiguations could be made available not just in the context of a particular digital library or website, but globally, benefitting researchers worldwide.

To be sure, a mature global e-infrastructure for the humanities capable of delivering instant access to rich web environments will not come cheaply or easily: it can only be the result of years of investment and effort by many professionals working together in a coordinated fashion. But it *is* possible to take the first steps at fairly low cost. Even a basic prototype for an Open Services Environment for the Humanities will have the potential to encourage other humanities digital projects to publish services as well as data.

Philogrid, a collaboration of the Perseus Digital Library at Tufts University in the United States and the Internet Centre at Imperial College London in the UK, proposes to create an expandable, Grid-enabled, web service-driven virtual research environment for Greco-Roman antiquity based initially upon open-source texts and services from the Perseus Digital Library. **First**, we will add to the Perseus DL Greek historians who exist only in fragmentary form. This task goes beyond simple data entry: we will create the first major digital collection of fragmentary authors designed from the start to interact with multiple source editions. **Second**, we will create a repository of philological data about the Greco-Roman world seeded with twenty years' worth of Perseus materials. The objects that we create will not only include books but every labeled object within each logical document. **Third**, we will convert the workflow that has evolved over the past ten years to process textual materials in Perseus into a grid-enabled workflow based on web services that can be applied to and customized for many collections. Although this project will concentrate upon the classics collections in the Perseus DL, the new workflows will also process non-classical Perseus content, and will thus from the start demonstrate their generality.

The foundation of this work is mature digital library technology, available as open source in a standard format. The Perseus Project is one of the earliest web-enabled digital libraries, with tools and workflows that have been tested on material spanning a wide range of domains with the Humanities and Cultural Heritage over many years. The development process will follow a strategy already successfully employed in e-Science projects at the Imperial College Internet Centre. It will consist of conversion of the Perseus workflow and tools into a web service environment, in which the Perseus workflow is analysed into steps, each of which is published as a web service with a configurable API. A workflow mirroring the Perseus workflow will then be composed using a standard workflow, then fine-tuned using model data from Perseus and sample data from the digitisation within the current scope of this project. At the same time, we will match the steps of the Perseus workflows to other open-source tools with similar functionality, which will be mounted as "competitor" services; this will enable us to compose and compare workflows with nearly identical outputs but different implementations.

Thus, at fairly low cost, the proposed Philogrid would provide both a ground-breaking prototype for the open services approach in the Humanities and a test-case for bringing sophisticated but project-bound digital library technology into an Internet Web Service space.

Table of Contents

Significance

B1. Philogrid proposes to create an expandable, Grid-enabled, web service-driven virtual research environment for Greco-Roman antiquity based initially upon texts and services from the Perseus Digital Library that are available under Creative Commons and open source license. To this end, we will perform three tasks.

B2. **First**, we will add to the Perseus Digital Library Greek historians who exist only in fragmentary form. These fragments will be based upon earlier work by Müller (Müller 1848) and Jacoby (Jacoby 1923) and upon contemporary editions of Greek fragmentary authors published in the past several years in the series "I frammenti degli storici greci." These fragments will, however, not be static quotations but will be instead dynamic excerpts which can automatically be found, align themselves against and compare themselves with any editions of the authors from which these fragments are drawn. This task goes beyond simple data entry: we will create the first major digital collection of fragmentary authors designed from the start to interact with multiple source editions, whether these have careful XML transcription or only uncorrected OCR of classical Greek. The fragments in this collection will not only automatically link themselves to editions of Greek texts already available but will be able to find new editions of their source texts and secondary sources that discuss them.

B3. **Second**, we will create a repository of philological data about the Greco-Roman world, seeding this collection with materials developed over twenty years by the Perseus Digital Library. All digital objects created by the Perseus Digital Library over the past ten years – and many objects from the preceding decade of work — are available for download under a Creative Commons license – we thus design the output of this funded project as both useful in itself and as a starting point for further work that can be conducted without the need for additional permissions from anyone. While the Fedora-based Tufts Digital Library will be the long term primary home for the objects that we create, we will follow the emerging guidelines developed by the Fedora group for Open Archives Initiative Object Re-use and Exchange Initiative (OAI-ORE).[1] We hope in this way to make it as easy as possible for third parties to extract parts or all of this Creative Commons collection.
The objects that we create will not only include books but every labeled object within each logical document: thus, we will not only deposit major lexica of Greek and Latin but each separate entry and each separately labeled sense. We will add not only Greek and Latin source texts but each citable chunk within each text: thus, we will include not only multiple editions of Livy but the text of, for example, Livy's *History of Rome*, Book 1, chapter 3, section 2 as it appears in multiple editions. We will also deposit born digital information, such as the database of Greek and Latin stems and endings that drives the Perseus morphological analyzer, corrected indices of Greek and Latin proper names from classical source texts which can be used as training data for automated named entity identification, and the results of automated text mining conducted according to documented procedures (e.g., automatic calculation of the most common English equivalents for a given Greek or Latin word across multiple authors).

B4. **Third**, we will convert the workflow that has evolved over the past ten years to process textual materials in Perseus into a grid-enabled workflow based on web services that can be applied to, and that can be customized for, many collections. In this project, we will concentrate upon the classics collections in the Perseus Digital Library. The workflow, however, that we have already released includes collections on non-classical subjects such as the History and Topography of London, the American Civil War, and early modern English literature. The new workflows will therefore also process the same non-classical materials and will thus from the start demonstrate their generality.

[1] The American Principal Investigator serves on the advisory committee of this group. For more on the work of the OAI-ORE, please see, (Van de Sompel and Lagoze 2007).

Fragments of Greek Historians

B5. Classicists have been developing corpora of Greek and Latin texts since the *Thesaurus Linguae Graecae* (TLG) was founded in 1972. Most work so far has focused upon (1) providing broad coverage, (2) creating semantic markup for particular texts, both literary (e.g., the TEI XML Greek and Latin sources in Perseus) and documentary (e.g., the papyri in the Duke Databank of Documentary Papyri, the inscriptions encoded in EpiDoc).

B6. While the bulk of our surviving Greek and Latin comes from works that have been preserved through manuscript tradition, the vast majority of ancient works were lost and are known only insofar as surviving authors quote, summarize or allude to their contents. Editions of such fragmentary authors pose particular problems. First, while an increasing number of fragments derive from literary papyri (more or less damaged texts or scraps of texts recovered from the sands of Egypt), editions of fragmentary authors generally consist of excerpts from other authors. For example, a "fragment" of the historian Ion of Chios can be found in a passage in Plutarch's *Life of Cimon*:

> And Ion actually mentions the phrase by which, more than by anything else, Cimon prevailed upon the Athenians, exhorting them "not to suffer Hellas to be crippled, nor their city to be robbed of its yoke-fellow." (Plut. Cimon 16.8, trans. Perrin)

B7. Fragmentary editions are thus largely second order, meta-editions. Classicists have already developed EpiDoc as an extension of the Text Encoding Initiative (TEI) Guidelines so that they could represent documentary texts derived from stone and, more recently, papyri. In creating a corpus of Greek fragmentary historians we will address enough concrete problems and provide enough useful content to provide an initial model for a *FragDoc*, a description of how to build on the TEI to systematically represent fragmentary authors – essentially, annotated excerpts from other editions.

B8. Editions of fragmentary authors have long been problematic because they involve extracting passages from an open set of authors, where the original source editions may be unavailable except in a relatively few research libraries. Already more than a decade ago the American ancient historian Glenn Bowersock suggested that a net-based environment would provide a more satisfactory space within which to explore the problem of fragmentary Greek authors. While his comments particularly address Felix Jacoby's monumental edition of fragmentary Greek historians, the overall conclusion applies to all fragmentary editions:

> "The methodological problems of Jacoby's Fragmente illustrated here may arguably not warrant a wholesale condemnation of his enterprise. His collection can be viewed as a kind of ladder borrowed from Wittgenstein's philosophy: one uses it to climb up and then throws it away. Or again it may be seen to resemble navigational software for the Internet. We can perhaps locate useful material more quickly with it than without it. But one point is absolutely secure, and that is the necessity to leave Jacoby behind and to examine the original sources for historical fragments (however defined) before bringing any scholarly research on them to a conclusion."[2]

B9. Fragmentary editions should consist not of isolated quotations but of pointers to the original contexts from which the editor has chosen the fragments. Of course, editors should be able to define the precise chunks of text that they feel to be relevant and to be able to annotate these texts in ways that they think relevant (e.g., distinguishing what they consider to be paraphrase from direct quotation). But such

[2] Glenn Bowersock, quoted in (Most 1997), pg. 185, quoted also by John Gibert in *Bryn Mawr Classical Review* 1.23 (1998) available at http://ccat.sas.upenn.edu/bmcr/1998/98.1.23.html.

fragments should be dynamically linked to their original contexts and to up-to-date contextualizing information.

B10. A modern version of the above should allow for at least the following functions:

- Citation as machine actionable link: All links should be able to retrieve the full text of Plutarch, *Life of Cimon*, chapter 16, section 10 from the original edition on which the fragment is based.

- Citations as portals into multiple editions: First, we need to be able to label citations with the editions on which they are based: e.g., a machine actionable equivalent to "Plut. Cim. 16.10 Ziegler" (which scholars recognize is a citation to Ziegler's Teubner edition). Second, we need to able to call up other editions besides the source for the citation. Third, where the cited edition is unknown, the system should be able to use metadata about the date and location of publication in which the citation occurs (1) to rule out all editions published after the citation was printed and (2) to infer the most likely edition(s) from which the source was probably drawn.

- Alignment of citation schemes: Where citation schemes differ, the system should, in collating multiple editions, have aligned multiple citation schemes. In some cases, citation schemes are completely different: e.g., page breaks in an older canonical edition vs. logical chapter/section divisions. In the Ion Fragment above, the Ziegler Teubner edition and the Perrin Loeb edition share chapter but not section breaks: thus, section 10 in Ziegler corresponds to section 8 in Perrin. Thus, Plut. Cim. 16.10 Ziegler should automatically retrieve Plut. Cim. 16.8 Perrin.

- Fragment as search query: We should automatically be able to use the excerpted text to find the corresponding passage as it appears in all on-line editions, even when careful transcriptions of these editions are not available and we have to rely upon machine generated OCR text. In the latter case, the goal will be to generate links between the excerpted text and the page image of multiple editions for, and secondary sources about, the same passage.

- Dynamic collation: The system should be able to identify and prioritize points where the excerpt as published in the fragment edition differs from the corresponding text in various editions. Prioritization should include ability to distinguish variants in punctuation from variations on the same dictionary word (e.g., indicative vs. subjunctive), the choice of wholly different dictionary words and more substantial changes.

- Secondary source identification/summarization: The system should identify passages in journal articles, monographs, commentaries and other sources that discuss either the same fragment (e.g., Müller Ion of Chios, fr. 7; Jacoby FGrH 392 fr. 14") or the primary source from which that fragment is drawn (e.g., Plutarch, *Life of Cimon*, chapter 16, section 10 Ziegler). Where these passages are numerous, clustering should assemble these passages in thematically coherent groups, summarization techniques should identify common ideas associated with this passage, and similar text mining routines should search for other significant patterns (e.g., trends visible in scholars before and after Jacoby).

From Project to E-infrastructure: an Open Services Environment for the Humanities.

B11. When a researcher or student in the humanities today wishes to consult a digital version of a text he will most often end up in a digital library such as the Perseus Project. It is now standard for such libraries to offer the user "collection support" in the form of enriched content which links the digital text to a much broader context through a variety of sophisticated tools and processes such as authority lists, morphological analysers, and place name gazetteers. As a result, where a researcher would once have had to move back and forth across a library digging through piles of books to chase up references and allusions, he can now access them with a mouse-click. And speed is not the only difference: concentrating all of this information in a single, queriable system means that new forms of hypothesis and visualisation become possible, which were not only difficult but *impossible* with pre-digital formats.

B12. What if a user could have instant access to these services and enriched contexts simply by dropping a scan of a book into a folder on the desktop? And what if that user's own contributions—annotations, corrections, disambiguations—could be made available not only in the context of a particular digital library or website, but globally, so that entire communities of users could benefit mutually from each individual's contributions?

B13. Making this possible is the next great challenge in the evolution of digital libraries. Advances in next-generation Internet technology now make it possible to consider extending these previously website-bound services and data aggregations to the entire Internet community.[3]

B14. To be sure, a mature global e-infrastructure for the humanities capable of delivering instant access to rich web environments will not come cheaply or easily: it can only be the result of years of investment and effort by many professionals working together in a coordinated fashion. But it *is* possible to take the first steps at fairly low cost. And a prototype for an Open Services Environment for the Humanities, even a very basic one, will have enormous potential as a trail-blazer to encourage other humanities digital projects to publish services as well as data.

B15. The key is to start with a mature digital library technology that is available as open source in a standard format. This is provided by the Perseus Project's recent open source all-java release on SourceForge.[4] The Perseus Project is one of the earliest web-enabled digital libraries, with tools and workflows that have been tested on material spanning a wide range of domains with the Humanities and Cultural Heritage over many years. The maturity of the technology and the workflows guarantees an absolutely firm baseline from which to begin iterative development of web-based workflows.

B16. The development process will follow a strategy already successfully employed in e-Science projects at the Imperial College Internet Centre. It will consist of conversion of the Perseus workflow and tools into a web service environment, in which the Perseus workflow is analysed into steps, each of which is published as a web service with a configurable API. A workflow mirroring the Perseus workflow will then be composed using a standard workflow editor such as Taverna[5], then fine-tuned using model data from Perseus and sample data from the digitisation within the current scope of this project. At the same time, we will match the steps of the Perseus workflows to other open-source tools with similar functionality, which will be mounted as "competitor" services; this will enable us to compose and compare workflows with nearly identical outputs but different implementations.

B17. Thus, at fairly low cost, the proposed Philogrid would provide both a ground-breaking prototype for the open services approach in the humanities and a test-case for bringing sophisticated but project-bound digital library technology into an Internet Web Service space.

[3] For the requirements of a Humanities Grid, see (Crane, Fuchs, and Iorizzo 2007), (Fuchs 2007a), (Fuchs 2007b).

[4] http:// SourceForge.net/projects/perseus-hopper

[5] http://taverna. SourceForge.net/

History, Scope and Duration

History

B18. This project involves three groups. The Perseus Digital Library at Tufts University in the United States and the Internet Centre at Imperial College London in the UK are the institutions that will conduct the primary work, but they will be working with an established network of scholars from Europe and North America already engaged in creating new scholarship about fragmentary Greek historians.

B19. **Perseus Digital Library:** The Perseus Digital Library has been under constant development since 1987. While its core collections focus upon the Greco-Roman world, Perseus has worked extensively with collections in the history of science, early modern studies, American history, and other areas of the humanities. The earliest versions of Perseus were published by Yale University Press on CD ROM in 1992 and 1996 but developed a web presence in 1995. In 2005, Perseus began making its TEI-compliant texts available for download under a Creative Commons license. In November 2007 the full code base for the most recent version of the Perseus became available in SourceForge.

B20. The work proposed here complements a recent grant from the NEH Advancing Knowledge Digital Partnership program to develop *Scalable Named Entity Identification in Classical Studies*. This on-going effort will augment our ability to identify and index personal and place names in the fragments of Greek historians on which the work proposed here will focus.

B21. **Imperial College Internet Centre:** The Imperial College Internet Centre (http://www.internetcentre.imperial.ac.uk) was established in 2005, with seed-corn funding from Imperial College, to continue and expand the activities of the London e-Science Centre (http://www.lesc.ic.ac.uk/). The Internet Centre aims to develop the applications and industries of the next-generation Internet. Taking as its model the successful UK e-Science Programme, the Centre promotes a linked programme of application development, generic computing research and software development. In addition, as the next-generation Internet will clearly be an important economic and social arena, the Centre also promotes a research agenda investigating the economic, social and legal of the next-generation Internet. Through its Industrial Forum, the Internet Centre also works closely with leading commercial and academic stakeholders in the next-generation Internet, such as Vodafone, the BBC, and the Science Museum.

B22. The Internet Centre is at the forefront of research in developing tools and infrastructures for the next-generation Internet and mobile services. It has developed techniques in semantic discovery of Internet resources, service authoring (via service composition or mashups) and in service publishing, that constitute a comprehensive Internet service development and deployment toolkit. (http://www.internetcentre.imperial.ac.uk/projects)

B23. The Internet Centre also continues to pursue its successful research and development programme in component-based Grid Middleware begun under the London e-Science Centre: Iceni (http://www.lesc.ic.ac.uk/iceni/), an infrastructure for building component-based network-level applications, Gridsam (http://gridsam. SourceForge.net), a job-submission web service for grids, E-MAAS (http://www.lesc.ic.ac.uk/projects/microarray.html), a distributed grid-enabled system for microarray data analysis and management.

B24. **Pinakes:** On May 16-17, 2007, Holy Cross College hosted a conference *Doing Fragmentary History in a Global Context: International Efforts to Preserve Ancient Greek Historians*, which brought scholars from Europe and North America together to talk about the problem of managing fragmentary historians. All the ancient historians were involved in, or looking for a solution for, electronic publication of fragmentary historians. Ian Worthington represented the New Jacoby, a digital project supported by the publisher Brill. North American classicists Thomas Martin, Neel Smith, Bruce Robertson, Ross Scaife, Gregory Crane, and Mary Ebbott explored the issues of representing complex fragmentary texts in a digital environment. European classicists Monica Berti, Virgilio Costa, Donatella Erdas, Eugenio Lanzillota, Gabriella Ottone, and Guido Schepens talked about the collaborative efforts involving colleagues across Europe to create a new scholarly infrastructure for fragmentary historians that would reach the widest possible audience. This workshop established new collaborations – Professors Berti and Costa visited the US again in October 2007, visiting Tufts as well as Holy Cross to develop the nascent collaboration between the US and European scholars representing more than thirty colleges and universities on both sides of the Atlantic. This scholarly community provides both current scholarly source material for, and a dedicated audience for the results from, the work proposed here.

B25. The European collaboration is well established. For fifteen years a group of scholars from four European universities has been intensely devoted to the study of fragmentary Greek Historiography. These academic institutions are: Università di Roma “Tor Vergata” (prof. Eugenio Lanzillotta); Katholieke Universiteit Leuven (prof. Guido Schepens); Universidad de Sevilla (prof. José Maria Candau Morón); Université “Marc Bloch”, Strasbourg (prof. Dominique Lenfant). These scholars started their researches independently; subsequently, through the organization of meetings, seminars, workshops, etc., they began to share the results of their experiences and now also cooperate in several common projects. Their studies embrace complementary domains: origins of Greek historiography and regional history (Università di Roma Tor Vergata); historiography of Asia Minor (Université “Marc Bloch”, Strasbourg); antiquarian literature and biography (Katholieke Universiteit Leuven); Hellenistic historiography and historical geography (Universidad de Sevilla). A relevant part of the results of these studies has flowed into the scientific and editorial project of the series “I frammenti degli storici greci” (“The fragments of the Greek historians”), established and directed by Eugenio Lanzillotta (Università di Roma Tor Vergata, Greek history, chair). Since its inception, scholars of 26 European universities have chosen to participate in *Pinakes*.

Scope and Duration

B26. We are seeking support for one year of work but this year of support will set in motion a process that will, we hope, extend long into the future. The fragments of Greek historians will be sustainable indefinitely as a long-term collection within the Tufts Digital Library. The ICL will convert the Perseus Hopper (already open source) into an open, Grid-enabled workflow, built on top of web services. Third parties can use the workflow as implemented at ICL and Tufts or download the code for use within another high performance environment. An international community of scholars, already engaged in publishing editions and studies of these fragmentary Greek historians, is looking for the digital infrastructure within which to conduct their own researches more effectively and to make their work more accessible, intellectually and physically.

Methodology and Standards

B27. **Software Development Methodology:** Agile Programminng Methodology will be used as the main development methodology on this project, including the following techniques: Use case development, Iterative prototyping, and Test-led Programming. We will use an iterative development cycle, with a close monitoring of progress against various metrics and tests. We will manage technical risk by early prototyping of areas of uncertainty using model data. ICL has a long and successful track-record in Agile development and web service based projects.

B 28. **Hardware and Software:** The web services mounted for this project will run on the Internet Jupiter installation—24 1.2GHz UltraSparcIV processor Sun E6900 with 96GB memory, running Solaris 10, which is currently used to host development projects. The system is extremely reliable and recovers well from system failures. The Perseus Project’s all-java build on SourceForge (http://hopper.SourceForge.net) will furnish the baseline for the code. The Perseus "hopper" will be mounted on our Jupiter installation; QA test will be established to ensure the same QoS as the reference Perseus Project Digital Library installation at Tufts University. Web services will be mounted and maintained with Glassfish 2 (https://glassfish.dev.java.net/), an open source version of the Sun Application Server 9.1. The ICL team has proven experience implementing successful web service projects with Sun Application Server. Workflows will be composed using Taverna (http://taverna.SourceForge.net/), a workflow editor originally developed for the bioinformatics community, but with wide acceptance among UK Grid users. ICL are active developers of Taverna-related components. The ICL development team uses Eclipse as its IDE, with nightly maven builds, a subversion repository and a linked tracker (JIRA). The Perseus Project uses a subversion repository for versioning of text-enrichment. Software development is done in java, with some scripting languages (perl, jruby) used for prototyping and rapid deployment when appropriate. Perseus has a well-established and extensively tested workflow for feature tagging in TEI XML, employing scripting (Perl), XML editors (emacs + xml-mode), and validation with Xerces and libxml2 XML parsers. For OCR, Perseus will use PrimeRecognition (http://www.primerecognition.com/).

B 29. **Preparation and processing of material:** The Perseus Project will use TEI format XML(http://www.tei-c.org) for content tagging. They have a long-established and successful record in employing several TEI flavors, including tei_ms, tei_dictionaries, and tei_drama, and have themselves made substantial contributions to the improvement of the standards. TEI is used because it provides the best guarantee of longevity and onward sustainability for historical textual material. The entire Perseus Project Digital Library is tagged in TEI.

B 30. **Organisation of and access to material:** Digitised material will be published in the Perseus Project Digital Library and accessible via the same rich toolset and customisable online environment as other Perseus Project textual resources. Publication and access will conform to the Perseus Project's well-established standards. All digitised material will be available as open source, with appropriate creative commons licenses.

B 31. **Storage, maintenance and protection of data**: Data at the Perseus Project is backup up daily on the Tufts University system. The entire ICL Internet Centre's cluster storage is backed up daily both at the Departmental and College level. For coding, the Internet Centre uses a professional development system, with subversion check-in, nightly maven builds, and a linked project tracker (JIRA Atlassian). ICL has over 20 years experience in maintaining hpc installations for research; the services developed by this project will continue to run on the ICL machines at the customary high post-project-completion service level. All software developed over the course of the project will be made available as open source software on SourceForge. The Perseus Project web services will become part of the standard SourceForge offering for the Perseus Project hopper.

B32.

Risks	Likelihood (L/M/H)	Impact (L/M/H)	Mitigation
ORE specification changes during project scope	L	M	Revise ORE use during development and at project end.
OCR QA falls below prediction	L	L	Reduce size of sample dataset to increase quality of data.
Implicit state cannot be easily refactored out of Perseus workflow	M	L	Continue development with state reqs made explicit; create storage staging areas for incremental update of dbs.
Comparison web services	L	M	Revise workflows; write temporary glue code to

for feature extraction etc. are not interoperable with Perseus services			force interop; seek alternative underlying implementations.

Work Plan

B33. **Immediate Work:** Because the Perseus Digital Library has already published its content and code-base under open source licenses, the Internet Centre at Imperial College can begin its work immediately. ICL will begin immediately after the project plan is published with design specifications that will guide development; these will be periodically updated as needed. At the same time, ICL will install a version of the Perseus hopper on its Solaris machines for QA.

B34. Perseus is already planning to have Müller's *Fragmenta Historicorum Graecorum* scanned into the OCA and PDF versions of the new editions from the series "I frammenti degli storici greci" are already available. We can thus immediately send this content out for data entry for initial XML transcription. Within one month, we will have enough data from the data entry firm to begin refining the results and creating a version of the final XML transcription. We will send a copy of Jacoby for the citations to be keyed in within two months.

B35. Consultation with our colleagues in the project will allow us to draw up the initial list of editions and important source texts that we will submit to the OCA. As we analyze the citations from Müller and Jacoby we will add additional editions and authors to this list. Work on aligning the names of Greek fragmentary historians with the LC NAF and the FRBR catalogue will begin immediately.

B36. **Workshops:** We propose not only to enter data but to extend our ability to represent fragmentary authors in a digital environment. We will disseminate our plans from the beginning but we have also allocated support for workshops in both the US and in the UK to support face to face discussion. The US workshop will take place in the autumn of 2008; the UK workshop will take place in the winter of 2009.

Deliverable 1: Professional data entry and careful XML Markup (c. 20 mbytes):

B37. We will create an initial collection of fragments that point readers to the general context from which each is drawn; we will collate the fragments as published in both of these monumental works with the sources from which they are drawn (e.g., Müller Ion of Chios fr. 7 = Jacoby *FGrH* 392 fr. 14 = Plutarch, *Life of Cimon* 16.10 Ziegler). This stage will entail double keyboarding of and careful TEI-compliant XML tagging of the following:

- An index of sources in Felix Jacoby (*Die Fragmente der griechischen Historiker*, 18 vols.). We will index the sources on which Jacoby draws for each fragment. (c. 1 mbyte).

- Karl Müller, *Fragmenta Historicorum Graecorum*, 5. Vols. First, editors of new editions will be able to download the Müller fragments as a starting point for their own work. Second, editors of fragmentary authors in drama, Roman history etc. can examine the encoding that we have used for Müller as they design their own projects. Third and most immediately important, the excerpts in Müller provide a set of queries with which we can search for other full text versions of the sources on which Müller draws. Thus, we can use the text from the fragment of Ion of Chios to find editions of the original passage from Plutarch's *Life of Cimon* as well as secondary sources that quote this passage. (10 mbytes of mixed Greek and Latin).

- The three volumes already published in the series "I frammenti degli storici greci": Gabriella Ottone, *Libyka. Testiomonianza e frammenti* (Rome 2002, pp. xxix + 707); Donatella Erdas, *Crater oil Macedone. Testimonianza e frammenti* (Rome 2002, pp. xvi + 338); Virgilio Costa, *Filocoro di Atene. I: I frammenti dell' Atthis* (Rome 2007, pp xiv + 526). These three volumes are the first of more than thirty additional volumes in preparation (http://antichita.uniroma2.it/frammstg.htm). Careful XML encoding of the information within these three volumes will lay the foundations for electronic publications within the rest of this series and, we hope, for many publications within classics and beyond. We consider the fact that the primary language of these volumes is Italian to be a major advantage. Most of our work in Perseus has focused upon English language scholarship and this project will allow us to address the problem of multilinguality directly. (5 mbytes of Greek, Latin, Italian).

Deliverable 2: Image-books of classical editions from which fragments of Greek historians are drawn and of other works relevant to Greek fragmentary historians (c. 500 volumes)

B38. The fragments created in Deliverable 1 are designed from the start both to be of use by themselves but also to serve as a gateway into the sources from which they are drawn. While many of these sources are available in Perseus and the TLG, both Perseus and the TLG provide only a single edition and, with a few exceptions in Perseus, they provide only the reconstructed text and not the textual variants and other scholarly apparatus.

B39. We propose to digitize c. 500 books of core importance to the study of Greek fragmentary authors. These will be available as image books – i.e., page images with text generated by OCR (see below). Our first priority will be to include at least one, and where possible more than one, edition for the most important authors from which Greek fragmentary historians are known. In this project, we will focus on works that are in the public domain (e.g., editions by European editors who have been dead for 70 or more years). This opens up to us most of the textual infrastructure developed through the early part of the twentieth century and gives us more than enough material for this project.

B40. We will use the Open Content Alliance (OCA) digitization workflow. We have already tested this workflow: the University of Toronto has digitized several hundred works on Greek and Latin, as well as Old Norse, Sanskrit, and Syriac.[6] The cost is $0.10US per page and $5 for pulling each book from and returning it to the shelf. Since the average book contains c. 300 pages, it costs c. $35US to produce a high-quality image book that can be run through OCR software, such as the Prime Recognition package[7], or printed via emerging print-on-demand services such as Lulu and Booksurge. While we will store all image books created for this project within the Tufts Digital Library, they will also remain available as part of the OCA library hosted by the Internet Archive.

B41. While Google Books includes some relevant editions, Google places restrictions on the ways in which these books can be used (e.g., you have to ask for permission to apply them to OCR). Furthermore, we have found the quality of scanning in Google Books to be uneven, with poorly scanned and even missing pages. An investment of $35US per book thus guarantees both unrestricted access and consistent quality.

Deliverable 3: FRBR compliant MODS cataloguing for open content editions (c. 1,000 volumes)

B42. Libraries have been cataloguing classical works for two millennia. We have therefore authoritative names for major classical authors: Cicero is “Cicero, Marcus Tullius” and searching an OPAC (online publicly accessible catalogue) for “Cicero, Marcius Tullius” will retrieve books by or about Cicéron as

[6] For an evolving list, see http://www.archive.org/bookmarks/AliB; for an example of an image book, see the *Stoicorum Veterum Fragmenta* at http://www.archive.org/details/stoicorumveterum01arniuoft.
[7] For more information on Prime Recognition, see http://www.primerecognition.com/.

well as works with titles such as “Ciceros Rede für T. Annius Milo” or “M. Tullii Ciceronis De Officiis.” The Library of Congress Name Authority File (LC NAF) focuses upon the authors or subjects of whole volumes. Since most Greek and Latin authors of whom we are aware have left behind works that are very brief (e.g., a few poems in the *Greek Anthology*) or fragments (as in the case of the historians on whom we are focusing), they do not have authoritative names in the LC NAF. Libraries cover a vast array of subjects beyond classics and thus provide breadth but do not provide the depth of coverage that professional scholars need.

B43. Classicists have therefore created their own lists of authors and identifiers independent of the existing library infrastructure. Cicero, for example, is “M. Tullius Cicero” in the Packard Humanities Institute Latin (PHI) CD ROM bibliography. Furthermore, the bibliographies of ancient works on which classicists depend tend to be checklists for particular works such as the editions cited by comprehensive lexica or the particular Greek and Latin editions included in a specific digital corpus. Thus, as of mid-November 2007, the on-line TLG Canon of Greek Authors lists only the 1972 D. L. Page editions of Aeschylus’ plays and not the 1955 Gilbert Murray edition (which the TLG included in its first digital releases), the 1998 M. L. West edition or any other edition published before or after.

B44. To address these issues, we have already begun creating a catalogue of Greek and Latin works. As of mid-November 2007, we have catalogued 920 authors and 2,200 individual works in 600 editions of Greek and Latin authors. In some cases, a single work extends across multiple volumes (e.g., the two-volume Oxford Classical Text of the *Iliad*). In other cases, a single volume will include the works for many different authors (e.g., the *Greek Anthology*). We will include the database of 800 Greek fragmentary authors already created by the European team led by Eugenio Lanzillota. The catalogue already:

- Integrates authorized names from the LC NAF (where these exist, e.g. “Cicero, Marcus Tullius”) with the identifiers from more detailed disciplinary bibliographies (e.g., Latin author number 474 in the PHI bibliography of Latin, stoa0087 in the *Canon of Latin Literature* from the Stoa.org). We will add the 800 entries in the European database of Greek fragmentary historians to the authority lists that we are already aligning to the LC NAF.

- Manages multiple editions, translations, commentaries, indices, lexica etc. focused on the same author or work: we draw upon the Functional Requirements of Bibliographic Records (FRBR) developed by the International Federation of Library Associations as a framework within which to capture the relationships between, for example, Thucydides’ *History of the Peloponnesian War* as a general work (TLG author #3, work #1), various Greek editions, translations into English and other languages, commentaries, Betant’s specialized lexicon of the language of Thucydides (*Lexicon Thucydideum*) etc. Within the scope of this JISC/NEH project, we will only be able to catalogue the most important scholarly works for Greek fragmentary historians but the foundation can be absorb a much more comprehensive bibliography.

- Includes analytical cataloguing that covers individual authors and their works as they appear in a given volume: thus, the record for Gow’s Oxford Classical Text for the Greek Bucolic Poets (*Bucolici Graeci*) includes records for Theocritus (LC NAF: “Theocritus”; TLG #0005), Bion (“Bion of Phlossa near Smyrna” in the LC NAF; TLG #0036); and Moschus (LC NAF: “Moschus”; TLG #0035). Even with works that include only a few authors, standard cataloguing is uneven: the MARC record in the Tufts OPAC for Gow’s *Bucolici Graeci* lists Theocritus but not Bion or Moschus. Within the context of this project, analytical cataloguing will include the testimonia and fragments that Müller, Jacoby, and other scholars have identified within surviving works. Thus, Ion of Chios would appear in the record for Plutarch’s *Life of Cimon* with Müller and Jacoby as authorities) as they appear in each of these collections.

- Provides a framework that can gradually include comprehensive coverage of all versions, whether manuscript or printed scholarly edition, for a given work: a digital environment can provide access to images of any written text. As a starting point for this process, we will include identifiers for the manuscripts listed in G. W. Hall's *Companion to Classical Texts*, augmented as time allows with reference to later editions.

- Includes cataloguing data and the URL and/or URN with which to locate on-line versions of a work: this catalogue thus includes those works available not only from the Perseus Digital Library and the OCA but also Google Book Search and other sources.

- Builds upon existing library metadata: we begin work with MODS records downloaded from the Library of Congress. We correct these (e.g., Latin works are occasionally labeled as Italian) and augment them with information described here but preserve the MODS XML format.[8] The records that we create can therefore be integrated back into standard library infrastructure that exists now and are designed to become part of emerging Cyberinfrastructure now envisioned for the humanities.[9]

Deliverable 4: Searchable Greek text from open content image-books with substantial classical Greek text (c. 5,000 volumes)

B45. Image books have become fundamental instruments of scholarship because OCR software can generate searchable text at scale. Thus, Making of America and JSTOR set the stage for even more massive projects such as the Million Book Library, the OCA and Google Book Search, all of which rely upon OCR generated from scanned page images with error correction routines but no manual editing. In such "image front" systems, users search the text but then retrieve the original image, with text acting as an index.

B46. Classicists have, however, always assumed that the image front model was of limited value to them, because no one had produced OCR software that could generate usable results from classical Greek. Experiments conducted during late 2006 and published in mid 2007, however, demonstrated that the current generation of OCR, when coupled with intelligent error-correction routines, could generate useful searchable classical Greek text from page images. Because one major OCR engine covers modern Greek but does not manage classical Greek (and its more complex system of accentuation), in our initial work, we chose to ignore Greek accentuation and to focus on accurate transcription of the characters: this would support most string based searching (e.g., look for any word that begins with the string σωφρ) and would have a negligible impact upon lemma based searches (e.g., find all possible forms of πέμπω).[10] We found that on modern Greek fonts (e.g., Loeb, OCT) we could achieve accuracy of up to 99.94%. Even for older Greek fonts that modern readers find difficult the results are encouraging: we found that we could generate transcriptions of the Bekker Aristotle with character level accuracy of 99.20%

B47. As one element of our work, we will run OCR software tuned for classical Greek not only over those books that we have digitized but over as many image books as we can find within the OCA (and from any other source that becomes available). The basic method will be simple. We will assemble a list of all texts that seem likely to contain Greek (e.g., books on classics, New Testament studies, etc.), run

[8] MODS stands for "Metadata Object Description Schema". http://www.loc.gov/standards/mods.
[9] For more on the need for a humanities cyberinfrastructure, please see (ACLS 2006).

OCR software over them, and then scan the output for words that our Greek morphological analyzer recognizes as valid Greek.

Deliverable 5: Object Reuse and Exchange – Repository Ready Data

B48. In November 2007, the Perseus Project released the first installment of its collections under a Creative Commons license. This release included not only TEI-compliant XML, but datasets such as the morphological analyses generated for Greek and Latin in the Perseus collections, TEI-compliant catalogues as well as structured metadata for the art and archaeological collections and all the metadata upon which the Perseus Digital Library depends.

B49. We will not only add the data produced under this project but will expand the materials that we publish. We will create objects for reuse and exchange among various repositories. To this end we will:

- Add the philological data within the Perseus Digital Library to the Fedora-based Tufts Digital Library, where they will be available for download under appropriate Creative Commons licenses.

- Add the same data to a D-Space repository at the Perseus Digital Library as an exercise in portability.

- Make available the scripts for and documentation about ingesting this data into both repositories

- Make the data that we ingested into Fedora and D-Space available as a separate download. This data will include:

    - The FRBR compliant authority records for personal names and literary works (created as MADS records).[11] All names for Greek fragmentary historians and historical works will be attributed to their scholarly sources such as Müller, Jacoby and contemporary scholars. Variants for the same name will be aligned. Where scholarly authorities suggest fundamentally different names for authors or works, or where they attribute fragments to different authors and works, the catalogue will include these attributions along with their scholarly source in a machine actionable form.

    - Image-books suitable for OCR as well as the OCR that we have generated. This will include all image books created as a part of this project and any other relevant materials from the OCA and any other suitable source that may become available.

    - TEI-compliant XML text for all books. Where available, these XML files will contain carefully transcribed text and manually tagging (level 4 and 5 markup). For many image books, the TEI XML will consist of level 1 text (i.e., TEI header followed by uncorrected OCR output with page breaks and markup for content for which automated methods are suitable –e.g., identification of Latin and Greek, proper names, citations, basic page elements of page layout such as the separation of text from textual notes). We will include for convenience the following categories of data automatically extracted from these files.

[11] MADS stands for "Metadata Authority Description Schema", a LOC XML schema for authority data that serves as a companion to MODS. http://www.loc.gov/standards/mads/

- Indices of all people, places, citations and other automatically extracted entities. Each index entry will label the source of an index entry as machine generated and/or human verified. Indices will include byte offsets to the file as distributed as well as the best logical citation data available (i.e., chapter/verse, book/chapter/section etc. for TEI XML transcriptions, page numbers for image books). FIGURES

- Separate entries for every labeled chunk in every reference work. This includes each headword in every lexicon, encyclopedia and other reference work. We will also publish the underlying structure of each article (e.g., not only the 116,000 and 52,000 entries in the Liddell Scott Jones Greek-English and Lewis and Short Latin-English lexica but the 161,000 and 102,000 hierarchically arranged senses).

- Machine actionable tables of contents reflecting multiple overlapping hierarchies: e.g., by chapter/section as well as by Stephanus/Bekker page. FIGURES

Deliverable 6: Open Source, Grid-enabled, Service Oriented Middleware for Humanities Materials

B50. Brian Fuchs at Imperial College Internet Centre (ICL) will direct the creation of open-source, grid-enabled middleware for humanities texts that is based upon web services and workflows. A first working version of this task is feasible within this project because of the code base that the Perseus Digital Library released under an open source license on SourceForge in November 2007. This code base provides a complete, end-to-end workflow that applies a number of standard processes for humanities texts. The published system comes with collections about the History and Topography of London, American History and Literature, Early Modern English, and Nineteenth Century American Newspapers, as well as with the classics collections on which the work proposed here will concentrate. The code base incorporates years of development. Between 1995 and 2004, David A. Smith, Anne Mahoney, Jeff Rydberg Cox and others developed the original version of the standard pre-processing workflow in Perl. Starting in 2003, David Mimno, Gabriel Weaver, and Adrian Packel led the development of the Java-based revision. The first version of this system became available in May 2005. Rashmi Singhal tested, documented, and edited the new code base from May 2007 until its release on SourceForge in November of the same year.

B51. The Perseus ingestion workflow includes the following processes. **(1) Chunkification**: Given an XML document, identify and index all citable chunks of text, including not only the dominant hierarchy but also schemes based upon milestones. **(2) Citation Extraction and Link Generation**: Extract tagged citations from an XML document, converting them into bi-directional links (i.e., if a document cites Thucydides, Book 1, chapter 86, create a link from Thuc. 1.86 back to the citing document). **(3) Named entity identification:** Discover personal and place names within untagged text and match these against authority lists. **(4) Named entity indexing**: Given tagged proper names in a document, generate a flexible index (e.g., George Washington, G. Washington, Washington). **(5) Morphological analysis**: Apply morphological analysis to words tagged in various languages. At present, these include Greek, Latin, and Italian.

B52. Each of these services will be abstracted into a web service and the entire “pipeline” will be reconstituted as a grid workflow using Taverna. Additionally, in order to provide end-to-end coverage, we will wrap the Perseus OCR process, currently a pre-processing step, as a web service. Upon completion of processing via the workflow, texts will be ingested into the Perseus Project Digital Library for access within existing UIs. No attempt will be made within the scope of this project to mount these UIs as web services.

B53. The service environment and workflows will be validated against three sets of sample data: first, Greek and Latin texts chosen from existing Perseus Project material, which will be used as a benchmark; secondly, the sample data produced by the digitisation and content development scope, which will furnish a test of the system's ability to successfully process new complex material; thirdly, a large and diverse sample set chosen from existing Perseus material, which will furnish a test of the system's ability to handle complex content semantics.

B54. We propose to develop the middleware for Philogrid in short cumulative iterations, following the standard Agile Programming methodology used in Internet Centre development projects. There will be 9 iterations, each of a month's duration (see below). A working environment will be ready by the end of iteration 3: it will then be iteratively tested and expanded.

B55. **Iterations: 1.** modeling I: Abstract the Perseus hopper workflow. **2.** web services I: Wrap hopper classes as web service. **3.** workflows I: recreate Perseus workflow in Taverna using wrapped services **4.** test I: test workflows on existing Perseus model data. **5.** Web services II: Add comparison web services from external components.. **6.** Workflows II: introduce variant workflows based on semantic criteria for comparison testing **7.** Test II: test sample dataset generated by digitization. **8.** Workflows III: Develop workflow store with semantic access layer based on MODS, FRBR + CIDOC-CRM **9.** test III: validate semantic content-based access against model data from Perseus and sample fragments dataset.

B56. Two important features of the development plan are comparison testing and semantic access.

B57. **Comparison Testing:** The comparison testing iterations (5 and 6) aim to test the "openness" of the developed infrastructure by introducing web services based on external software that "match" the functionality of the steps in the Perseus workflow. By using such "competitor" services, we ensure that there are no hidden state issues that may affect a user's ability to compose services freely, and can validate the concept that web services for humanities texts can be usefully grouped semantically.

B58. **Semantic Access**: In the final development iteration, we will establish a basic workflow store with a semantic access layer. The store will provide a mechanism for publishing workflows and for retrieving them based on metadata about the content of the texts that they have been used to process. The metadata will be provided by library records MODS format and will be fed into a semantic model provided by the CIDOC CRM standard (http://cidoc.ics.forth.gr/), which allows querying across nested temporal data. The aim is to make possible in a future development cycle automatic retrieval of workflows based on pre-existing metadata.

Nature of Collaboration and Staff

B59. The Perseus Digital Library has already created the code base that will provide the starting point for the work that the Internet Centre at Imperial College London will perform. The Internet Centre has a long record of proven success in converting project-based software to web service distributed environments. Members of the Perseus Project (Gregory Crane) and the Internet Centre (Brian Fuchs) have a history of collaboration on ground-breaking digital projects that spans nearly ten years.

B60. Key project developers will maintain weekly teleconferences. In addition, there will be a semi-annual meeting of principals. A project website will be set up immediately, which will include a project wiki that will be used for coordination. In addition, ICL will open an installation of its tracker (Atlassian JIRA) for the project as a whole in addition to its own software development tracker.

B61. **Gregory Crane**, Editor in Chief of the Perseus Project, Professor of Classics and Winnick Family Chair of Technology and Entrepreneurship, will direct the work done at Tufts University. **Alison Jones Babeu** holds an MLS and has worked as a research associate at Perseus for three years. She is responsible for developing the authority lists and for the cataloguing. **David Bamman** has an MA in computational linguistics and has worked as a research associate at Perseus for two years. He will be responsible for overseeing the day-to-day programming and advanced markup at Perseus. **John Darlington**, Director of the Imperial College Internet Centre and Professor of Programming Methodology in the Department of Computing at Imperial College, London, will direct the work done at Imperial College. **Brian Fuchs**, Coordinator of the Imperial College Internet Centre in the Department of Computing at Imperial College, London, will oversee the day-to-day infrastructure development at Imperial College. Beyond the core staff, the project will have an advisory committee with Monica Berti (University of Turin), Christopher Blackwell (Associate Professor of Classics, Furman University), Virgilio Costa (University of Rome Tor Vergata), Donatella Erdas (Scuola Normale Superiore di Pisa), Eugenio Lanzillota (University of Rome Tor Vergata), Thomas Martin (Jeremiah W. O'Connor Professor of Classics, College of the Holy Cross, USA), Bruce Robertson (Professor of Classics, Mount Allison University, Canada), Ross Scaife (Professor of Classics, University of Kentucky), Neel Smith, Associate Professor of Classics (College of the Holy Cross).

Dissemination

B62. We will, of course, disseminate the results of this work through publications and presentations. Source code will made available under an open source license on SourceForge. Image Books will be available without restrictions both from the OCA and from the Tufts Digital Library (which at present can, for example, provide easier access to individual pages than can the OCA). XML versions of recent scholarship will be made available under a Creative Commons license that allows free distribution. XML versions of older scholarship (such as Müller's *Fragmenta)* will allow for both free distribution and for the creation of derivative works.

References

B63. ACLS. "Our cultural commonwealth: the final report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities & Social Sciences." 2006. http://www.acls.org/cyberinfrastructure/OurCulturalCommonwealth.pdf

B64. Crane, Gregory, B Fuchs, D Iorizzo. "The Humanities in a Global e-Infrastructure: a Web –Services Shopping-List", *UK e-Science AHM*, Nottingham, 10-13 September 2007. (http://www.allhands.org.uk/2007/proceedings/papers/887.pdf)

B65. Jacoby, Karl. *Die Fragmente der griechischen Historiker (F Gr Hist)*. Berlin, Weidmann, 1923-58.

B66. Most, Glenn (ed.). *Collecting Fragments-Fragmente Sammeln.* Göttingen, 1997.

B67. Müller, Karl Otfried. *Fragmenta Historicorum Graecorum*. Parisiis : Ambrosio Firmin Didot, 1848-1878. 5 volumes.

B68. Sompel, Herbert Van de and Carl Lagoze. "Interoperability for the Discovery, Use, and Re-Use of Units of Scholarly Communication." *CTWatch Quarterly* 3 (August 2007). http://www.ctwatch.org/quarterly/articles/2007/08/interoperability-for-the-discovery-use-and-re-use-of-units-of-scholarly-communication/

B69. Stewart, Gordon, Gregory Crane, and Alison Babeu. "A new generation of textual corpora: mining corpora from very large collections." *JCDL '07: Proceedings of the 2007 conference on Digital libraries*. New York, NY, USA: ACM, 2007, 356-365.