Integrating Digital Papyrology 2

Duke Databank of Documentary Papyri
Advanced Papyrological Information System
Heidelberger Gesamtverzeichnis der griechischen
Papyrusurkunden Ägyptens

Project Leaders: Roger Bagnall (New York University), Deborah Jakubs (Duke University), Joshua Sosin (Duke University). Co-PIs: Jakubs / Sosin

[*Partners*: Department of Classical Studies and Duke University Libraries, Duke University; Center for Visualization and Virtual Environments, University of Kentucky; Institute for the Study of the Ancient World, New York University; Zentrum für Altertumswissenschaften Institut für Papyrologie, Universität Heidelberg; consultants at the Centre for Computing in the Humanities, King's College London]

Note: The following proposal is offered as the sequel to a project currently funded by The Andrew W. Mellon Foundation, "Integrating Digital Papyrology" (08/01/2007–07/31/2008, at $500,000 USD); some of the following introductory material is adapted from that project proposal.

Among humanistic fields, papyrology is notably well provided with digital resources for accessing and controlling the primary materials of its discipline, including texts, metadata, and images of the papyri, ostraca, and tablets preserved in Greek, Latin, Arabic, various forms of ancient Egyptian, and several other languages. Until recently, growth in the digital domain had proceeded on parallel tracks, discrete projects pushing ahead under independent leadership, in different—usually proprietary and often incompatible—technology platforms, on different—mostly closed and one-off—standards. But in the last few years several projects have begun to strive for greater interoperation, more efficient use of broadly distributed efforts and expertise, and sustainable development and growth. The holistic view manifested and sustained by recent collaborations, including the currently funded project, is in fact as old as the discipline itself. Since its birth in the late nineteenth century papyrology has insisted, much more so than its older sister, epigraphy, on a rigorous and comprehensive standard of scholarly presentation and documentation, including text and translation, images, and a full range of metadata, at the level of the physical object (provenance, acquisition, custodial history, inventory numbers, dimensions, preservation, etc.), its content (genre, author, palaeography, date, historical significance, etc.), and its scholarly legacy (publication, emendation, etc.). The discipline's fruits are most abundant when scholars can navigate and compare these different types of data easily and effectively. Forging a digital environment that supports the full range of papyrological use behaviors is, in an important sense, the next step on a path that the discipline has been blazing since its inception in the 1890s.

Under generous support from The Andrew W. Mellon Foundation, the pace of this change has accelerated in the last year. Even before the currently funded project the two most important digital papyrological resources in North America, the Advanced Papyrological Information System (**APIS**) and the Duke Databank of Documentary Papyri (**DDbDP**), had already taken a number of steps toward greater integration and stabilization. APIS is a multi-institutional, international database of papyrus collections, covering all languages, with extensive metadata, digital images, and translations; the DDbDP is a textual database of Greek and Latin documents preserved on papyrus. Both had already developed close links to a third resource, the Heidelberger Gesamtverzeichnis der griechischen Papyrusurkunden Ägyptens (**HGV**), which includes extensive metadata for essentially the same set of documents included in the DDbDP, and overlaps with many thousands of papyri accessible via APIS; moreover, the HGV had begun—and now has completed—an exhaustive mapping of its data to Leuven's Trismegistos, a comprehensive portal to multiple databases of metadata concerning

papyrological and epigraphical materials found in the Nile valley from roughly 800 BC to AD 800, ranging in content from bibliography to geographical, prosopographical and onomastic information (http://www.trismegistos.org). Finally, taking advantage of fresh technological developments, and the recent collaboration between HGV and Trismegistos, the APIS central interface technical team had begun development of a prototype search/browse interface, the Papyrological Navigator (**PN**), with the intermediate goal of supporting sophisticated search and display of data drawn from APIS, DDbDP, and HGV.

At the start of the currently funded project the major tools of digital papyrology had begun to converge on a course toward common access under a common interface. What was needed was the unifying force of a common technological standard. Thus, in the past year APIS, the DDbDP, and HGV have been developing and implementing a strategy for interoperation and sustainability, by erecting a standards-based, open technical architecture that will soon establish these three assets as a tightly integrated cluster, on the basis of whose rigorous, transparent, and collaboratively designed standards the rapidly growing and maturing host of other papyrological resources will be able to interoperate and co-evolve. The thrust of this work has been foundational: (a) converting from SGML (Standard Generalized Markup Language; DDbDP texts), FMPro (FileMaker Pro; HGV metadata), and idiosyncratic XML (eXtensible Markup Language; HGV translations) to a single EpiDoc XML standard, (b) building a first-generation front-end to enable searching of Greek and Latin texts, mapping common objects across multiple databases, and concurrent display of distributed materials (text, translation, metadata, and image side-by-side), and (c) improving and bundling tooling to enable other, even next-generation, projects to redeploy not only the data but also the processes constructed under the project.

The bulk of our efforts under a second phase, which we propose to call **IDP2** (following the current Integrating Digital Papyrology grant, which we call here **IDP1**), will move from framework to implementation, from foundational to operational concerns. Our goal is to build tools and workflows that will vest content-control of these papyrological resources in the papyrological community worldwide (from creation to peer-review to diachronic tracking), as a step toward migrating the full range of papyrological work behavior to the digital environment. In support of this goal we request **$813,619.80 USD** for two years (09/2008-08/2010; 12 months software development, followed by 12 months of documentation and training), in realization of the longer term goals enumerated and work accomplished under the first phase of this project, upon which we elaborate further below.

Our primary goal is to open up creation and control of the principal papyrological data-set to the community of scholars, to create new opportunities for adding and improving content. Our motivations and objectives are **economic**, **ideological**, and **scholarly**. The DDbDP's mode of data population, hand-keying of Greek from published print editions, has long since ceased to be economically viable; it is slow, costly, inefficient. But it is also ideologically archaic in its exclusive concentration of effort and expertise in a single center, rather than in the discipline as a whole. It might be argued that current hegemonic editorial strategy works well for discrete corpora but we believe that it is unsuited to the public representation of *all* corpora; put another way, the current orientation of the DDbDP purports to deliver answers (here is what the texts say), rather than provide a forum in which questions may be formulated, tested, and answered, collaboratively. Moreover, the current scholarly mission of the DDbDP is retrospective; it is essentially a box into which already-edited texts are deposited, even as the field's state of knowledge and scope are ever shifting with the landscape of scholarly debate. Similar observations can be made of the HGV, which also relies on centralized content creation, a somewhat retrospective approach, and even tends—for example, with regard to variant dates—toward a kind of descriptive agnosticism, reflecting a variety of unweighted opinion rather than providing a forum in which the community can strive for consensus dynamically. Even APIS takes a similar view, for its partly curatorial mission has always meant (but need not forever do so) the centralized creation of static

records. Catalogue records are only meaningful to the extent that they keep pace with the progress of papyrological knowledge, so that the need for a technological bridge between the sorts of expertise that papyrologists and curators, now separately, bring to bear on these ancient materials is acute. Now that these resources are tightly bound to each other—the DDbDP and HGV fully integrated with each other and both fully mapped to APIS—the scholarly community requires a robust mechanism to control the cascading effects of improvements made to any one piece of data in the network. Collaborating partners need not only a common standard in which to encode data—completed for DDbDP and HGV and begun for APIS under IDP1—but also a common mechanism for allowing the community of papyrologists, ancient historians, etc. to control and improve content as the state of scholarship grows and interpretations evolve.

Under IDP2 we aim to address these three objectives—each of them the inheritance of the pre-internet era and longstanding institutional funding mechanisms—by developing a technological infrastructure that will make maintenance of core papyrological assets **economically viable**, **vest content-control** in the hands of the papyrological community, and carefully **track scholarly evolution** of the papyrological state of the art. To accomplish these goals we shall build a collaborative, web-based, fully audited, version-controlled editing environment, which will be opened to wide participation. The tooling will be called the Son of Suda On Line (**SoSOL**), named after its ideological and methodological forebear (though it will employ a newly-developed code base), the Suda On Line (**SOL**) project, a groundbreaking browser-based, collaboratively authored, peer-reviewed, translation of the massive tenth-century encyclopedia of Greco-Roman antiquity, called the Suda. Although the Suda On Line is now 10 years old, and still operating essentially on its original software platform, it remains the best example of a collaborative editing environment within the domain of Classical Studies, or even beyond. It is, even today, the only substantial body of ancient material that is authored collaboratively, on line, and with all stages of work—from proposal to vetting, to acceptance and incorporation—fully and transparently audited. Its mode, we believe, is the way of the future for Classics, even if its actual software platform is now badly out of date. In other words, SOL furnishes the model and inspiration for the next-generation editing environment that we shall build and call SoSOL, even though it will not be possible to repurpose its code base, due to its age and incompatibility with the modern platform upon which we plan to build.

SoSOL will extend the accomplishments of IDP1 to allow teams of DDbDP-HGV editors to add new EpiDoc XML-encoded texts and metadata directly to the DDbDP-HGV, whether converted from files prepared for print publication, or created in XML. This will have the immediate effect of driving data-entry costs down to levels approaching zero, by recycling the labor invested by the editors themselves, who have already keyed the data (in their preferred desktop editor). SoSOL will also allow the wider community of scholars to propose and justify emendations and enhancements to texts, translations, and metadata, and even new editions of previously unedited papyri, collaboratively and in real time. SoSOL, in other words, will begin to transform the DDbDP-HGV from static reflection of closed teams' selective representation of print scholarship to a dynamic, collaborative, and transparent arena for presenting, contesting, and honing the primary data-set of an entire discipline, as the venue in which texts are edited, revised, debated, joined, translated, classified, dated and re-dated, etc., online and in real time. In so doing SoSOL will enable the papyrological community to begin the process of migrating traditional print workflows to the digital world.

The SoSOL editing environment will be a portlet within the Papyrological Navigator (PN). But beyond the PN itself, which has been extended under IDP1 to support Greek text searching, SoSOL will also make use of several tools that have been developed or greatly enhanced under IDP1. A major operating principle of IDP1 has been that tools should be created with a view to extension and modification toward integration with other related projects or processes. Under IDP2 we seek to

implement that principle in practice, extending several tools to become integral features of regular SoSOL-PN processes.

EpiDoc: First, we elected to encode the DDbDP not in a highly customized one-off schema but in a robust, widely accepted standard used by papyrology's sister-discipline, epigraphy. How well EpiDoc would suffice as a format for digital Papyrology was an unknown at the beginning of the project, but it has proved more than adequate to the task. A few extensions have been applied to the guidelines and schema to accommodate DDbDP distinctions not previously identified in epigraphical markup, but on the whole the two communities of practice are now largely working with the same set of tags and recommendations. Thus, all editing and searching processes developed against the DDbDP, for use with SoSOL and the PN, will function with little need for modification on epigraphic data-sets.

CHET-C: The first incarnation of the Chapel Hill Epigraphic Text Converter, created by T. Elliott in 2003, was based on Microsoft Access and provided a form-based interface that allowed Epigraphic texts to be entered in Leiden form and then transformed via a series of regular expressions into EpiDoc XML. The Access version was followed by a completely rewritten Java version by H. Cayless, and subsequently by a Javascript port that allowed the Leiden to EpiDoc conversion to run entirely in a web browser. IDP1 required both a quantum leap in sophistication, as the converter would be required to support not just Leiden, but also Beta Code escapes, and a script-based port that could be executed as part of the Runner pipeline. Under IDP1 R. Viglianti at King's College has been responsible for the greater part of CHET-C's conversion to the Python scripting language and for its added Beta Code support. CHET-C will only grow in importance in IDP2 because a vital component of the SoSOL editor will be the ability to display a Leiden view of transcriptions for editing, but transform them to EpiDoc for storage. CHET-C will require further extension under IDP2 to accommodate editorial shorthand for indicating classes of semantic mark-up not explicitly handled by Leiden, to enable papyrologists to work in a "tagless" environment.

Transcoder: The Transcoder is a tool originally developed in 2003 by H. Cayless to address the problem of converting between Greek encoded in Beta Code, GreekKeys, and other legacy formats, to Unicode. One of the aspects of the tool that makes it different from others in this space is its ability to perform bidirectional encoding shifts. This feature, and the fact that it is implemented as a Java library, have led to its being incorporated in a number of projects over the years, including Demos and Perseus. The IDP1 grant has allowed a thorough reworking of the Beta Code translation part of the software, including a number of bug fixes, and a vastly improved ability to support the transcoding of text embedded in XML documents. The updated transcoder has already been released at its home on SourceForge under an LGPL license. The Transcoder will continue to be an important tool in the SoSOL system now that the base texts are in Unicode. A significant number of scholars prefer to work in Beta Code, because they are used to it and it doesn't require a change of keyboard. The SoSOL editor may need a Beta Code view, mediated by the Transcoder. In addition it will be a component available for the import of new texts, which may use Beta, or other legacy encodings. Under IDP2 the Transcoder proper will not require significant extension, but rather runtime integration with the SoSOL editing software, to enable papyrologists to work in an environment that accommodates multiple encoding preferences.

Crosswalker Tool: The Crosswalker is a tool originally prototyped by G. Bodard, H. Cayless, and T. Elliott at an EpiDoc worksprint in 2006. Its original purpose was to facilitate the export of epigraphical databases (RDBMS systems) such as the Epigraphic Database Roma (EDR) to EpiDoc XML for purposes of archiving and interchange and also the import from EpiDoc files into EDR and similar databases. The initial specifications for further crosswalker development were developed as part of IDP1, but the work was deemed out of scope for this phase of development. There is a clear need, however, for a tool that can manage the interchange between the SoSOL system and external formats, and the crosswalker specification has been further refined to support this kind of activity. One of the early tasks of IDP2, therefore, will be to implement the crosswalker specification developed during IDP1. Thus, under IDP2 we aim to develop this tool with a view to supporting the conversion of MS

Word to EpiDoc XML for import into SoSOL and to act as a framework for the import of other data sources into SoSOL (these will include the HGV and APIS imports described below).

HGV Metadata Crosswalk: The HGV metadata crosswalk, as created for the IDP1 project, is a conversion process written principally in XSLT 2.0, by Z. Au, that takes as input the XML export from the HGV Filemaker database, and creates from it 56,000 TEI XML files (one per record) containing the metadata marked up according to the EpiDoc recommendations. There are several records or groups of records in HGV that require complicated and idiosyncratic processing (in particular dates, bibliography, and placenames), and therefore specialised scripting, to move from one complex schema to another. This crosswalk process will form a core part of the ongoing DDbDP/HGV integration workflow, and will also serve as the primary example of functionality required by the Crosswalker tool. Under IDP2, the HGV Metadata Crosswalk will not require significant extension per se, but rather careful integration with the data population workflow that feeds the PN; for example processes by which new HGV records generate corresponding DDbDP text stubs and vice versa.

APIS Translation Crosswalk: A conversion process was built for IDP1 to convert a sample of translations from the APIS database to EpiDoc XML (via another set of replacements and XSLT to be plugged into CHET-C). The APIS translation Crosswalk will be an essential asset, as the community migrates, piecemeal, APIS translations from their current text-only format to EpiDoc XML, for search and display via PN. Under IDP2, the APIS Translation Crosswalk will not need significant modification, but rather close integration with SoSOL-PN processes; for example hard-wiring automatic attribution of APIS translations on ingest, and only then presenting them in SoSOL for editing by interested translators.

XML Aggregator: The XML Aggregator is an XSLT-based tool that pulls together disparate EpiDoc XML files containing DDbDP texts, HGV and APIS translations, and HGV metadata that relate to the same papyrological text (as mapped by bibliographic identifiers), and collates these several streams into a single EpiDoc XML file. This functionality will be an integral part of the underlying SoSOL architecture, especially since some of these streams (the HGV metadata) will continue to be authored outside of the XML editing stream and need regularly to be recombined with the rest of the data. Under IDP2, work on the XML Aggregator will concentrate on integration with the back-end processes that dynamically consolidate the multiple XML streams for indexing by PN and editing by SoSOL.

Runner and test environment: The Runner is a set of Python and Shell scripts that muster and serialize the various elements of the DDbDP to EpiDoc conversion process (Transcoder, file splitter, Python and XSLT modules of CHET-C, Greek Number Converter) and the translation and metadata conversions. This tool also includes a testing mechanism for comparing output of selected files and patterns with expected output. Although the DDbDP texts and HGV translations will be authored in EpiDoc XML from the end of IDP1 onwards, the HGV metadata will be an ongoing conversion process. The testing mechanisms especially can be abstracted and incorporated into the working process of parts of SoSOL. The Runner itself may not be recycled by SoSOL but its testing environment will surely be reused and extended by SoSOL to enable users to verify that data have been entered and/or converted correctly (whatever input path they follow; i.e. whether users are feedng SoSOL Word documents or entering XML directly).

EpiDoc Stylesheets: The EpiDoc standard XSLT stylesheets have been a core part of the EpiDoc toolset for several years, maintained by T. Elliott, G. Bodard, and H. Cayless. For the IDP1 project these stylesheets were completely rewritten in a more modular and parameterized form by Z. Au, with a view especially to rendering papyrological-style Leiden output from the DDbDP texts. These stylesheets are both free-standing, and will be needed for Leiden-rendering of EpiDoc texts in SoSOL, Papyrological Navigator, and any other tools that need to display papyrological texts from the XML. Extensions to the stylesheets will be made to conform with enhancements executed on the EpiDoc encoding under IDP2 (more below).

Thus, the second phase of work intended to lightly extend and tightly integrate these tools has suggested itself in precisely the way we had hoped during the formulation of the IDP1 workplan. By accomplishing almost all of the conversion work via repurposeable tools rather than one-off processes we have in effect produced a series of modules that can be modified to work together under different but related processes. Inasmuch as the back end of the SoSOL editor will need to perform regular bidirectional encoding conversions and the back end of the PN will have to perform regular crosswalking and aggregating routines, this suite of tools will prove essential and labor-saving to both the work of IDP2 and the ongoing maintenance of the data themselves. Moreover, some of these tools, particularly a more generalizable Crosswalker and the Transcoder, will prove invaluable to the community of epigraphists should it choose to follow the papyrologists' lead and migrate existing database formats and encodings to Unicode EpiDoc. Just as enhancement of some of these tools was part of the core mission under IDP1, their greater consolidation and interoperation with the PN and SoSOL frameworks is central to the plan of IDP2.

Son of Suda On Line (University of Kentucky)

The work accomplished under IDP1 has created a basic infrastructure, in the form of standards and tools, upon which a new form of digital scholarship for papyrology can be built. Our motives are several. Collaborative scholarly editing is a force multiplier, a positive alternative to the current operational and intellectual bottleneck wherein a single office manages the community's data set. The digital environment allows the community to preserve and attribute the contributions of its members. At the same time, the publication life cycle of primary data can be enhanced and accelerated by creating a digital publication channel dedicated to new editions, while maintaining essential rigorous and transparent peer-review and credit mechanisms. Additionally, the content of the repository can be distributed in ways that make its preservation and sustainability far more likely.

One of the primary aims of IDP2 is to build on the work of IDP1 to further facilitate reengineering of the full life cycle of professional content creation and management for the digital environment, to embrace the range of social interaction and behaviors that are often disjoint, implicit, and obscured from the community within a technical architecture that renders them coherent, explicit, and discoverable. To that end, the University of Kentucky's Center for Visualization and Virtual Environments (**Vis Center**) proposes under IDP2 to build a browser-based collaborative editing environment designed to manage the editorial workflow for the publication of standards-based XML documents: the Son of Suda On Line (SoSOL). The SoSOL editing workflow encompasses the production and enhancement of new and existing content, the process of peer review by domain experts and refinement by the community of scholars, as well as the reporting of scholarly activity by, e.g., sub-discipline, time, or author. The SoSOL system will audit and publish contributions to the data set from initial proposal and justification to approval or rejection, and grounds thereof, by the Editorial Board, thus adapting traditional scholarly social structures to the digital environment. But in so tracking this information SoSOL will also be able to furnish reporting data at a scale of detail well beyond current practice, from, e.g., quantifying levels of editorial activity within the area of Byzantine papyrology as against Ptolemaic, to the scope of disciplinary engagement with documents of a certain genre or content, to the profile or success rates of emendations proposed by a given scholar.

At the core of SoSOL is a **collaborative XML editor**, an editorial **workflow manager**, and a **version control system**. The XML editor will feature hierarchical access control and a cascading capability set, allocating and controlling privileges to Super-Editors, an Editorial Board, and the user community at large. Authorized Super-Editors will be empowered to upload new texts, execute corrections to data and encoding, and exert control over any publication to the system, in effect multiplying the role of the current directors and distributing it to leaders of the discipline. An appointed and rotating Editorial Board will vet corrections and emendations proposed by the user community. Other users can suggest readings, translations, emendations to the DDbDP texts, or even propose

editions of unpublished papyri to a 'sandbox' (i.e. an unpublished but visible environment where changes will not affect the published version until and unless they are approved by the editors), all of which are in turn vetted and approved or rejected by the Editorial Board. All changes within the system are fully audited and discoverable by all users (e.g. who uploaded what when; who proposed what emendations; emendations' state of vetting and grounds for approval or rejection). SoSOL in effect transforms the DDbDP from a black box into an information commons. SoSOL will operate with identical capabilities and analogous workflows on HGV metadata as well as translations, whether from HGV, APIS, DDbDP Super-Editors, or the user community at large.

The impact of SoSOL on the DDbDP in particular, and the field of ancient studies in general, will be immediate and revolutionary. The DDbDP is at present unable to keep pace with the rapid acceleration of papyrological print publication because of its tightly centralized workflow and data population model (it is now several thousand documents behind current print publications). Part of the Foundation's investment in the initial IDP effort has freed the DDbDP from the out-of-date and idiosyncratic formats that kept it in its own box. SoSOL will complete the transformation by thoroughly reorganizing the untenable labor calculus. It will place in the hands of papyrologists worldwide the ability to keep the database current by adding new texts and improving existing ones. Scholars in multiple time zones will be able to work together to construct definitive, born-digital editions—transforming the DDbDP and its affiliated tooling and databases from a retrospective digitization effort to a full-fledged digital publication environment.

But the potential impact of SoSOL reaches well beyond this immediate need. By creating SoSOL and offering the example of the working, worldwide collaboration that it will enable, we hope to catalyze similar transformations of many other textual databases, papyrological, epigraphic, Latin, Greek, or otherwise. Most of these are presently discrete resources—data silos of the sort that digital humanities infrastructure thinkers seek to modernize and integrate—but their combined texts exceed in number that of the DDbDP-HGV by an order of magnitude: there are perhaps 800,000 print-published Latin and Greek inscriptions, and the number of digitally published Latin and Greek inscriptions is already several times greater than documents in the DDbDP-HGV; to this add the growing digital corpora of Arabic and Egyptian documentary and literary papyri, as well as Greek and Latin literary and para-literary texts. Moreover, SoSOL could be deployed on other sorts of data, e.g. bibliographic controls such as the Checklist of Greek, Latin, Demotic and Coptic Papyri, Ostraca and Tablets (http://scriptorium.lib.duke.edu/papyrus/texts/clist.html), next-generation specialized papyrological lexica, or the like. It is clear, then, that a single, online, standards-based, collaborative, language- and discipline-agnostic text-editing tool (especially one designed to feed a similarly generalizable search engine, PN) will revolutionize the study of ancient documents. We are confident that its potential for scholarly and functional impact in our field is even greater than that wrought by access to affordable desktop publishing and database software in the 1980s and 1990s. SoSOL will be a critical asset for digital humanities infrastructure, and strong support of this effort by New York University's Institute for the Study of the Ancient World (**ISAW**) is motivated in no small part by this potential: its reusability in the context of its ambitious sustainability initiative (see more below under Sustainability), which aims to rescue, revitalize and sustain key digital resources for ancient studies. Its plan goes beyond the simple preservation of files to embrace standards-based stewardship of projects and their digital publications. As a result, ISAW must employ generalizable tools like SoSOL, not just as aids to scholarship but as cost containers and force multipliers: direct routes to efficiency of scale in project management, technical support, periodic software migrations, etc. In other words, the success of this critical initiative depends in significant measure on the creation of SoSOL and tools like it.

Hierarchical Access and Cascading Capability Set

SoSOL users consist of Super-Editors, an Editorial Board, and all other registered users (in effect, anyone who registers), whose access and responsibilities are based on a cascading capability set. Access

and responsibilities are controlled by an abstraction layer built on top of the version control system. The abstraction layer is responsible for translating commands between the editorial tools and the version control system, providing flexibility for future modifications to the software (replacing the Subversion version control system with different system, for example) as well as making communication between the different layers of the SoSOL transparent to the end users.

- All **registered users** may:
  - suggest readings, translations, and emendations, including links to external Internet resources
  - 'publish' unpublished papyri to the sandbox
  - comment on their own proposed modifications as well as those of others
  - view a snapshot of the entire history of a text, including all modifications, editorial decisions etc.
  - view the original source text in various display flavors of Greek (including transliterated, SGreek, SMK GreekKeys Athenian, Unicode, and TLG "beta code")
  - send comments or queries directly to Super-Editors, Editorial Board, or other contributors
  - update their profiles (select Greek font display etc.)
  - generate reports of the contributions of participants
- Members of the **Editorial board** may also:
  - vet readings, translations, emendations, and 'publications' to the sandbox
  - assign status to readings, translations, emendations, and 'publications' to the sandbox (draft, low, high)
  - promote emendations, new texts and translations for publication to the main corpus
- **Super-Editors** may also:
  - add new texts and translations, whether previously published or not, without first 'publishing' them through the sandbox
  - oversee communication to categories of participants
  - review the progress made by particular contributors
  - see the status of particular texts
  - assume another participant's identity if necessary, e.g. to reset a lost password

The abstraction layer will allow for the configuration of access levels and for flexible assignment of responsibilities that differ from the three-level hierarchy of the IDP project. In other words, it will not be hard coded to allow only three levels of access with only these specific responsibilities. This design makes the SoSOL system generalizable, and so more useful to other projects with different editorial requirements.

Technical Description

The beating heart of the SoSOL design is an innovative, browser-based **collaborative XML editor** that enables papyrological users to create and modify the complex, semantic structure of EpiDoc document objects via the familiar and efficient Leiden conventions. Although both browser-based and standalone editors exist to support creation and editing functions for text, structured documents, and XML, none meets—out of the box—the specific requirements presented by the intersection of a computationally-actionable semantic document format for papyrology, namely a familiar and efficient set of visual formatting conventions in the editing environment, and the necessity of integrated version control and collaborative review. The editor will need to support an editable, Leiden-based view of the text and at the same time enable the embedding of EpiDoc features not supported by Leiden sigla. For example, Leiden furnishes a convenient and universally accepted shorthand for distinguishing ancient lacunas ([...]), modern editorial insertions (<...>), ancient erasures ([[...]]), modern expansion of ancient

abbreviations ((...)), modern omissions of extraneous ancient letters ({...}), but does not have a means of representing inline apparatus entries, material causes of lacunas, degrees of certainty in restoration, and a wide array of other semantically significant information, traditionally rendered in human-readable verbose prose commentary. This requirement implies a need for, at the least, the extensive customization of an existing in-browser editing tool.

In constructing the editor, the UKY team will (a) extend an existing server-installed XHTML editor, FCKEditor (http://www.fckeditor.net), with functionality to be added for robust XML editing, including validation against both general and papyrology-specific schemas, and real-time verification of tagless input, i.e. iterative background validation of XML that will be generated by Leiden entry, building on existing user interface elements and adapting the back end to output EpiDoc XML rather than XHTML, and (b) integrate existing EpiDoc tooling, developed and upgraded during IDP1, for managing conversions on the Leiden-EpiDoc axis and for handling font encoding differences across browser versions and locales, adding translations or metadata to existing texts or texts to existing translations or metadata (CHET-C, EpiDoc XSLTs, Transcoder, Aggregator, Runner, etc., on all of which see IDP1 Resources and Tooling above). These components will also be deployed as services to support the direct ingestion and conversion of legacy texts (instantly ready for subsequent refinement in the editor), as well as export capabilities to facilitate simultaneous publication in print-oriented contexts, like books and journals, that are expected to persist for some time (nothing would prevent interested parties to deploy SoSOL and lightly modified EpiDoc stylesheets to generate two outputs, camera-ready copy for print book production as well as direct entry to the DDbDP). This XML editor will be surfaced to the user as a portal within an instance of the Papyrological Navigator, which was prototyped during IDP1 and which is receiving continuing improvements with NEH funds under APIS 6. The PN will provide users of the SoSOL editor with robust tools for access to relevant papyrological imagery and bibliography and for the discovery and analysis of comparanda. The PN is written in Java, using the Apache JetSpeed portal application framework. SoSOL will exploit the out-of-box JetSpeed architecture and components, as well as the interface components customized and created for the PN. This will ensure seamless integration with the PN, as well as the use of GUI metaphors already understood and approved by the papyrological user community. We shall also make comprehensive use of the Fluid Infusion User Experience Toolkit to guide and test the user interface development process (http://fluidproject.org/index.php/what-is-infusion).

Although the PN will provide the SoSOL XML editor and its users with information and tools drawn from the wider papyrological cyberinfrastructure, we must establish new infrastructure to facilitate the digital transmission of scholarly contributions constructed in the XML editor to an editorial board, and ultimately to publication via the DDbDP. To accomplish this goal, the SoSOL editor will be embedded in a service-oriented infrastructure that provides the digital **workflow management**, shared data repository and archiving functions necessary to support a globally distributed team of scholarly contributors, editors, and administrators. Atop the proven, open-source **version control system** called Subversion, we shall erect a workflow management system, using the OpenWFE framework, that enables contributors to submit content generated or modified in the XML editor, together with necessary notes and justificatory prose, to the editorial board for review. The workflow layer will notify the editors, facilitate their review of the proposal, and permit further annotation while recording editorial decisions concerning the publication-readiness of the proposal. The editorial workflow management interface will also be surfaced via the PN interface, although its unique interface components must be developed.

In order to support both the day-to-day interactions of contributors and editors, as well as the professional needs of the broader community, SoSOL will also provide a range of reporting functions, drawing on data and metadata artifacts of the content and workflow it manages. Both editors and contributors will be able to draw upon activity histories, showing the quantity and frequency of individual's contributions, together with the percentages of these that have been accepted by the editors

for publication. Such reports will link directly to the content in question, and will also be designed for ready incorporation into CVs or other similar documentation. Simple, stable URLs for these reports will facilitate ready verification by third parties. To generate these reports we shall integrate JasperReports, an open-source Java reporting library that outputs a wide variety of formats including PDF, HTML, Microsoft Excel, RTF, ODT, Comma-separated values, and XML. These reporting functions, combined with public access to the contents of the Subversion repository will surface the conduct and artifacts of peer review processes that are normally hidden from public view, thereby encouraging a broader culture of content-oriented discussion and peer-monitored peer review processes.

Functionality and Usability

The SoSOL interface will have the look and feel of any other Papyrological Navigator portlet, but a completely different back end for editing files, built on a **version control system** and supported by a **workflow manager**. The workflow manager controls actions related to sending texts through the vetting process, making vetted and unvetted texts available to the PN search, and saving those texts back to the main repository. A visually common interface for both searching and editing will help editors feel more comfortable with the environment and so facilitate adoption. Because the majority of SoSOL users will not be computing specialists, usability of the system is a driving concern of the development team. The entire process, from selection via PN search, import, or creation, through the entire editing process, through the vetting process and final 'publication', must be completely transparent for the end user. By design, all interaction between users/editors and the technology will be web-based. In an effort to maximize user-friendliness the Kentucky team will enlist the expertise of the Vis Center's Information Design and Usability Lab in the form of a Graduate Student Assistant dedicated to usability concerns (hereafter "User Experience Expert"). It will also take advantage of the efforts of the Fluid Project (http://www.fluidproject.org) for the design and usability testing of SoSOL; both in tandem with regular and intensive consultation with papyrologists (**J. Cowey, J. Sosin**) to ensure that user options conform to disciplinary expectation and practice.

From a given text view in the PN, registered users can elect to call up the text in the SoSOL portlet to edit the text and/or translation. The SoSOL editing view will include four basic PN components: the image viewer (to consult while checking transcriptions), the metadata viewer (for viewing, but not at present editing, APIS and HGV metadata), the SoSOL XML editor (for editing transcriptions and translations), and the PN text viewer (for viewing transcriptions while editing or creating translations).

SoSOL will provide both an XML view, for users who prefer to work directly in the XML, and a non-XML (tagless) view in which editors can use Leiden conventions, which are transferred to EpiDoc upon commission (using the incorporated Crosswalker tool). The XML editing functionality will be built on existing open-source software. On investigation of several XML and XHTML editors (both server-side and stand-alone) we have decided to deliver SoSOL XML editing functionality via extension of an existing server-installed XHTML editor (FCKEditor), with functionality to be added for robust XML editing (validation against schemas, etc.). This will build upon the user interface elements already present in the XHTML editor, while adapting the back end so that the output will be the appropriate EpiDoc XML rather than XHTML. The non-XML editing view will incorporate technologies developed under IDP1, including CHET-C and the Crosswalker spec (for Leiden-to-EpiDoc conversions), and EpiDoc-to-Leiden XSLT stylesheets. As the existing XHTML editor does not focus on ease of use for editing the raw XML output, significant additional tooling (XML tree view, etc.) will also be developed for users who prefer to edit in this mode. During the development process, the Kentucky team will consult regularly and intensively with EpiDoc specialists (**G. Bodard, H. Cayless, T. Elliott**) and papyrologists (**J. Cowey, J. Sosin**) to test for functionality and usability, as well as consulting with the Crosswalker development team and the developers of the XSLT stylesheets built under IDP1 (more under IDP1 Resources and Tooling above).

SoSOL will support creation/editing of (1) existing DDbDP/HGV files encoded in EpiDoc and saved in the main repository, (2) published texts yet to be added to the DDbDP, and (3) texts that are born digital in the DDbDP. First, all registered users can propose edits to existing transcriptions and translations. The user navigates via PN to a given text and elects to edit it in the SoSOL PN portlet. The user edits the text or translation (at one sitting or over time), saving the file in his/her own branch in the version control system. Once the edit is complete the user marks the text as "ready for vetting," and the editorial board is automatically alerted. These unvetted suggestions are to be indexable by the Papyrological Navigator. The Editorial Board vets suggestions and merges both accepted and rejected suggestions into the main repository for publication. This feature enhances the DDbDP from a static database to a structured, peer-reviewed public forum for honing the collective knowledge of the discipline in real time. It also applies two distinct levels of peer-review: the Editorial Board applies a traditional gatekeeping function, deciding which emendations are promoted to text and which are deprecated for representation in the apparatus alone. The papyrological community, however, will always retain the ability to weigh in collectively, whether to lobby for subsequent rejection of readings accepted by the Editorial Board, or the opposite.

Second, Super-Editors may add texts, whether previously published (print or otherwise) or not, straight to the main repository, without the approval of the Editorial Board or prior deposit in the sandbox. These texts can be batch-imported in various formats (MS Word, Open Office, EpiDoc XML, other TEI XML schemas). The majority of these texts are expected, at least in the near future, to come via modification and import of files prepared as camera-ready copy for print publication, in other words, texts that have already undergone traditional peer review processes. Following import, these texts may be edited via SoSOL by any user, as described above. Though Super-Editors will be able to publish texts that have not been published, or are not about to be published, in another setting directly to the main corpus, we expect that this stream will start small. In other words, the small group of Super-Editors may elect to rely on either the Editorial Board and the wider community for peer review, or on the community alone.

Finally, any user can add new texts to a 'sandbox', to be vetted by the Editorial Board for promotion/publication to the main corpus. As envisaged, the sandbox is not a physical location in the system but rather a label applied to new texts when they are created/imported. These texts are located physically in the individual users' branches but are treated differently from texts being modified from the main repository. This practice of labeling particular texts for different treatment can be extended for other projects that may have other types of information that require special editorial treatment. As with proposed emendations, once these born-digital texts have been marked "ready for vetting," they are indexable by PN, and matched with an automatically generated comment board to which any user can post comments during the vetting period. Once the new texts are approved (or denied) and merged into the main repository, the complete history of their approval is available for viewing (including all comments and any changes made to them during the vetting process). This set of features allows the DDbDP to become an environment in which texts are submitted, peer-reviewed publicly both by editors and the wider papyrological community, revised, and "published" directly to the disciplinary repository of primary source material.

This system of dual layers of editorial control, Editorial Board and Super-Editors, opens new avenues of collaborative, distributed content-creation, without sacrificing exacting standards of professional review and 'quality control,' and provides a technological basis for re-framing editorial decision-making and peer-review as ongoing discussions rather than summary judgments to be fossilized in print.

Version Control

Sophisticated version control is a core requirement for SoSOL. The main repository will be under the version control system called Subversion. The editorial requirements of the project (for example, that

the Editorial Board and Super-Editors can commit changes directly to the main repository, but other registered users may not) necessitate special controls between the editing view and the version control system, which will be provided by an abstraction layer built on top of the version control system. The benefits of building this abstraction layer are three-fold: the details of the version control system are hidden from the users, the higher-level editing layers will have no direct dependency on the underlying version control system should the repository change in the future, and the abstraction layer can enforce finer-grained controls on user actions for the editorial process.

The abstraction layer will include rules for concurrent editing. Multiple users may work on one text (transcription or translation) concurrently. A user may want to work on a single transcription or translation for an extended period of time before marking it as "ready for vetting", and we do not wish to impose locks in the system preventing a text from being modified by more than one user at a time. At the same time, users working on a text must be aware when other users are working on that text and be able to see others' changes as they are made. For purposes of maintaining editorial integrity, in-process changes by individual editors will be kept separate (i.e., not constantly merged, as in a Google Docs-type collaborative editor), but the different working copies will be viewable by other users who are editing the same text even before they are marked as "ready for vetting". To enable these requirements:

- Every registered user will have his or her own branch.
- Users editing the same text at the same time will have the ability to view each other's work as it proceeds, even before it is ready for vetting.
- When the user decides that a particular text in the branch is ready for vetting, it will be made available for viewing by all other users, indexable by PN, and released to the Editorial Board for vetting.
- All users, Editorial Board, and Super-Editors will be able to comment on a text. Comments will be presented as threaded discussions. Comment boards will be automatically generated when a text is released for vetting, and there will be a central index of all texts currently being vetted with links to all those texts and to the comment boards.
- Once approved or denied, the Editorial Board will flag the text appropriately and commit it from the user's branch into the main repository.

Merges, both between modified texts and between branches and the main repository, will be handled at the level of the Editorial Board. Conflicts between texts will be neither seen nor dealt with by general users. Although users will know when other users are working on the same text, and they will have the option of viewing this work, their different versions will not be merged until after they are vetted. In other words, final editing and quality control remain the responsibility of the Editorial Board.

Every change made to a text will be noted, time-stamped, and attributed. Change log information will be appended to the EpiDoc XML when changes are made. Thus, there will be two parallel sets of change metadata, one provided by the version control system and a second located within the XML itself. All changes must be approved by the Editorial Board, flagged as "unvetted" prior to approval and "approved" afterwards. Rejected proposals will persist in the apparatus criticus for the text, but will be flagged as such.

We cannot assume that any user of the SoSOL (whether a registered user, member of the Editorial Board, or Super-Editor) is technically proficient. Therefore, interaction with the version control system must be as transparent, simple, and intuitive as the editor. All version control will be handled through a user friendly web application.

Deployment

During the 12-month development period, regular builds of the SoSOL software will be hosted on a server at the University of Kentucky for development, reference, and testing. The 1/2 FTE systems

administration budget line covers related essential support. During the fourth quarter of the project, UKY and ISAW teams will collaborate on integration testing of SoSOL with the production PN and VCS at NYU. At the end of the fourth quarter, the production version of SoSOL code will be running publicly along with the production PN at NYU (ISAW committing to its continued system administration from internal funds). Concurrent with the deployment of production software at NYU, the software development stream will be migrated to SourceForge (including code and comment histories). At this time the SoSOL code will be publicly released under the GNU Public License via the SourceForge file release system. SourceForge is already home to the EpiDoc Community's code and documentation, and therefore provides the most appropriate context for public release and future collaborative code maintenance. Public queries and questions about the code will thereafter be handled via the EpiDoc community's existing "markup" listserv.

Documentation and Training

Developer-focused documentation will be created as one of the project deliverables for all components to be built or enhanced as part of IDP2 (this includes not only SoSOL, but other integrated components described below). It will consist of standard programming-language-appropriate elements including docstrings, readme files and annotated unit tests and test datasets. This documentation will be included with the code release and development history transferred to SourceForge at the end of the 12-month development cycle.

The papyrological user community will require comprehensive user documentation for SoSOL and its components. It will be encoded in XML using the industry-standard Darwin Information Typing Architecture (DITA), and will comprise appropriate concept, task and reference descriptions that are structured and delivered to the user in both HTML and PDF as a comprehensive User's Guide and a series of tutorials, "How Tos" and FAQs. This documentation will be prepared by the User Experience Expert and other project staff, in regular and close consultation with J. Sosin concerning domain-specific use cases. It will be delivered to users over the web alongside the production PN/SoSOL installation.

The project will employ three 5-day training sessions in order to introduce practicing papyrologists to the use of SoSOL. This approach follows the successful model—pioneered by G. Bodard in London—employed by the EpiDoc community to introduce epigraphers to its encoding specifications and related tooling. An initial session will be held in Durham, NC, in October 2009, with a second session in London in April 2010 and a final session, again in Durham, NC, in June 2010 (as part of the Summer Institute of Papyrology, to be held at Duke University that summer, sponsored by the American Society of Papyrologists). By the mid-point of the project's first year, the PIs will have identified 12 persons (6 North American and 6 European) to train as Super Editors. These individuals will be invited to the nearest training session(s), and the project will pay for their travel and subsistence. Remaining seats in these sessions will be offered to other practicing papyrologists on a first-come, first-served basis. Following the EpiDoc model, self-identified participants will be required to secure their own travel and subsistence. IDP2 will also fund travel and subsistence for instructors, as follows:

- Durham 1: User Experience Expert, Bodard, Cowey, Porter (Sosin will be local)
- London: User Experience Expert, Cowey, Sosin (Bodard will be local)
- Durham 2: User Experience Expert, Cowey (Cayless and Sosin will be local)

University of Kentucky (UKY) Workplan

The work to be carried out at the University of Kentucky Center for Visualization and Virtual Environments (**Vis Center**) involves the design and development of the SoSOL collaborative editing system.

Development will proceed over three interacting tracks: XML editor development, Version

Control Management, and Workflow Management. One full-time programmer will be assigned principally to each track, to ensure that each receives complete attention and top priority for the full twelve months of the development phase of the project. Each track represents a core and essential component of SoSOL functionality, contributing equally to its unique impact on collaborative digital scholarship. The XML editor is the most visible part of the SoSOL and is all that most editors will ever see of the tool. Existing "tagless" XML editors function by representing tags with formatting, while the SoSOL XML editor will specifically take advantage of tooling from IDP1 to provide "tagless" editing that is based on the Leiden conventions, with translations back and forth between the conventions and EpiDoc XML. The Version Control Management aspect of SoSOL ensures that scholars can work on the same text at the same time, viewing one another's changes and making suggestions as they proceed. Version control is a very well-understood problem in application development, with many available solutions, but the accommodation of the versioning system to the Workflow Management and the development involved in effectively displaying version information to users and editors of the tool will be considerable. The Workflow Management aspect of SoSOL is especially important, as it controls what happens to a text from the moment it is brought into the system, through the editing and vetting process, and to final publication. Workflow Management that aims to accommodate all traditional editing principles of an entire discipline is a large and complex task and we are not aware of any comparable systems for the humanities. While work has been done in each of these areas before, no other project has attempted to tightly integrate such components into a single seamless interface usable by non-computing experts. The scale of the project necessitates a substantial amount of programming support. Individually, each task is substantial in and of itself (for example, the creation of tagless editor to operate on the vest array of semantic mark-up accommodated by EpiDoc), but combining the three together requires significantly more resources due to the need for seamless integration of the resulting components.

The programmers working on version control and workflow management will work together closely and will share tasks across tracks when appropriate, while the XML editing programmer will interact principally with the papyrologists (**J. Cowey, J. Sosin**) and EpiDoc consultants (**G. Bodard, H. Cayless, T. Elliott**). A Graduate Student Assistant programmer will work alongside the full-time programmers, providing support for specific tasks and also assisting and coordinating with XSLT development (**Z. Au**). The programming team will work under the direction of a Programmer Coordinator (10 hours/week) who will be responsible for working with consultants and the Project Coordinator to guide and supervise the work of the programming team. This person will not be responsible for any hands-on programming, although s/he will need to be an experienced programmer, knowledgeable enough to be able to work with and understand classicists and the needs of the project. More advanced guidance on software development and programming will be provided by Professors of Computer Science Raphael Finkel, developer of the original Suda On Line, and James Griffioen, whose research focuses on computer networks and distributed systems. Finkel is on sabbatical 2008-2009 but has volunteered to consult on an as-needed basis, as will Dr. Griffioen. Throughout the development process, functionality and user testing of the individual components and of the entire system will be undertaken by a Graduate Student assistant from the Vis Center's Information Design and Usability Lab. The Project Coordinator (**D. Porter**) will be responsible for working with the project director, the entire UKY team, consultants, other project partners, and anyone else involved to keep the development of SoSOL on track and running smoothly. The coordinator will need to be knowledgeable of every aspect of the project.

XML Editor: Development of the XML editor will begin with interface design and implementation extensions to FCKEditor, an existing open-source server-side XHTML editor. Substantial work will need to be done on the back end of the XHTML editor to enable it to handle generic XML and normal schema validation (so that any XML schema can be used in the editor; it will not be hard coded to work with EpiDoc). Once this is complete, after the first four months, basic

functionality will exist for users to add new (non-published) texts directly into the tool. At the same time, substantial work will also need to be done to design and implement interaction between the XML editor and the version control system. This interaction will enable users to save files in their version control branches and to access those files for future editing. Next, to provide more usability for editing Greek, we shall integrate the Transcoder, which will enable a user to choose which "flavor" of Greek he or she wishes to view and edit. During the second four months we shall work with the PN development team to extend the shared interface and facilitate interaction with the PN search. Once this is complete, users will be able to search the PN and pull found texts into the editing system. Months nine and ten will focus on developing functions for importing (both single files and batches of files) pre-published texts into the system, and testing and refinement of the editor (and the system as a whole) will take place during months ten through 12.

Version Control Management: Development of the version control management system will begin with setting up a local version control system and copying the EpiDoc files from the production server at NYU. Files will be periodically updated throughout the development period. Concurrently we shall begin the design of the abstraction layer, which serves as an intermediary between the XML editor and the version control system and also between the version control system and the workflow manager. Providing an abstraction layer on top of the version control system, rather than building interactions directly between the software components, will make it much easier to switch out components if needed (for example, providing a different version control system or other type of database in place of Subversion). Major tasks in the development of the abstraction layer include: designing the abstraction layer, setting up a method for per-user branching (in consultation with the workflow management programmer, who will be concurrently designing and implementing a system for user registration), and designing systems to allow commits from the XML editor into the user's branch, for merging user branches into the main repository, and handling conflicts between user branches and the main repository. As managing conflicts between concurrently edited files is one of the most difficult and typically operator-intensive parts of version control, we shall dedicate many resources during the design period to create a system that is as simple and unobtrusive for the end-user as possible. Integration with the production server at NYU will take place during months 11 and 12.

Workflow Manager: The workflow manager is a critical facet of the SoSOL system, as it controls the entire editing process from the moment a text is first imported, created, or pulled in from the PN, through the vetting process, and to final approval and publication. We plan to integrate an existing open source workflow engine, OpenWFE (http://www.openwfe.org), to implement the SoSOL workflow management system. For the first four months, we shall focus on designing the workflow and designing and implementing the user registration system. The second four-month period will concentrate on integrating the workflow manager with the version control system and beginning to implement the workflow, and building the web application interface for the editors to interact with the version control system. The last four months will complete workflow implementation, test the workflow, and design and implement reporting capabilities.

Documentation and Training: Developer-focused documentation will be created and maintained by all developers working on the code as part of their day-to-day duties. The creation of user documentation and training materials require discrete attention. As interaction and user-interface design winds down in the fourth quarter of the project, the User Experience Expert will take the lead in developing user documentation and training materials, drawing on the expertise of the PIs and the entire IDP2 team as appropriate. S/he will also liaise with personnel at Columbia and NYU who will be working concurrently on the APIS6 NEH grant. Because the content of these materials is critically dependent on the specific layouts and behaviors of the software user interface, best practice dictates that its creation begin late in the development cycle and extend beyond it. Such a schedule permits the documentation team to address the effects of late-breaking changes that, while minor in terms of code, are significant in terms of user task description (e.g., altered button labels or relocated menu entries).

For IDP2, it will also be essential to capture and address the questions raised and difficulties encountered by participants in planned training events (see above). For both of these reasons, the User Experience Expert will continue to refine and expand documentation and training materials for a full 12 months following the end of the initial year-long development cycle. The User Experience Expert will also attend, and participate in the leadership of, all three training events and will liaise with G. Bodard, J. Sosin, and other instructors to prepare all presentation and handout materials. Additional programmer resources (25% for 12 months) will also be assigned to support documentation development, examples for training, and the correction of minor software bugs surfaced during this period. J. Sosin to coordinate feedback capture from training sessions and relay of such, along with use case examples, to the User Experience Expert.

Papyrological Navigator

The Papyrological Navigator (PN) combines web-based text, image and metadata viewers with a customized search engine, capable of searching and displaying information from multiple related sites. It is intended to replace the current production applications for APIS and DDbDP, and to provide an alternative mode of access to the full content of HGV records. It is a custom web application, prototyped (http://papyri.info) by the Columbia University Libraries Digital Program with funds allocated in part by Roger Bagnall from his 2003 Mellon Distinguished Achievement Award, in part by the National Endowment for the Humanities (APIS 5 grant), in part by Columbia University Libraries and in part by Mellon funds from the initial IDP grant.

As of 1 July 2008 New York University's Institute for the Study of the Ancient World the Digital Libraries office will assume responsibility for the PN and for hosting the new DDbDP-HGV, including version control. Moreover, the PN is undergoing concurrent extension under two separately funded projects (APIS 6 and Concordia, on both of which see the Appendix below), so that while the UKY IDP2 team will take the lead in specifying and implementing changes to the PN software and its server-side architecture necessary to fully integrate the SoSOL editor, management of the additional, concurrent requirement stream from IDP2, levies additional design, implementation and test burdens for which we seek funding under IDP2. With all PN work under the ISAW aegis, we shall be able to set a single development team on the task of coordinating PN improvement to accommodate the demands of IDP2 with those of the other initiatives.

Thus, **under IDP2 the ISAW team requests funding** to support one half-time programming position (1) to interface with the PN programming and development team that is already overseeing ongoing PN-development to ensure integrity of the whole, (2) to optimize version control and all other operational aspects of the production DDbDP-HGV Subversion installation, and (3) to enhance the PN search and display interface to accommodate, greater XML-aware searching, more sophisticated concurrent text-metadata searching.

Papyrological Navigator Workplan

Under the present proposal, we seek additional funding for PN-related programming at ISAW (50% FTE on average for one year, though workloads at certain points in the cycle will approach 80% near the end of the project, with 20% maxima during earlier lull periods). In large part this labor expands central PN development for critical issues not currently addressed by other PN-dependent projects. This funding is essential to support the UKY SoSOL development team in the area of SoSOL/PN integration (months 2, 5-6 and 10-12), as well as version control system peering with the production DDbDP/HGV Subversion installation at NYU (months 1-2, 7 and 11-12). Modifications to the PN search and display interfaces will also be required to interact with the SoSOL-related version control system interface (months 8-10) and to streamline user interface components for the fully integrated software package (months 10-12). Over the course of the entire grant period, approximately 10% of this individual's time weekly will go to general cross-component integration and system administration duties supporting the

entire team (especially in the context of PN interoperation). Where appropriate, this developer will assist the separately-funded APIS developers at Columbia, full-time PN developer at NYU and IDP2-funded developers at UKY in design, troubleshooting, bug fixing and user-driven feature enhancement work across the software family supporting IDP2 and its partners. **T. Elliott**, in consultation with leadership at other institutions, will lead and coordinate priorities to ensure core IDP2 deliverables are met.

EpiDoc Optimization / Crosswalker Tooling

During the course of the year in which the first phase of the current project was carried out (2007/08), the Text Encoding Initiative (whose XML schema EpiDoc follows) published a major new release, version P5, which includes many significant enhancements and refinements. Some of these refinements are already in EpiDoc recommendations (indeed some were EpiDoc innovations later adopted by TEI), but when the EpiDoc community starts seriously to consider the actual upgrade to the P5 schema after the end of the first phase of this project, there will be a certain amount of automated upgrading both of core EpiDoc tooling and of DDbDP texts and processes to be carried out. For the second phase of this project, therefore, we shall be retaining a fraction of time from the core group of EpiDoc/EpiDoc XSLT experts who worked on the original conversion under IDP1 (**Z. Au, G. Bodard, H. Cayless, T. Elliott**; domain-specific assistance by **J. Cowey**, **J. Sosin**).

Their efforts will be also essential in development and implementation of the Crosswalker, a customizable application, composed mostly of XSLT modules within a Java framework that automates conversion between the EpiDoc interchange format and various XML, database, and other schemata. Crosswalker mappings can be defined, using a simple syntax, for a variety of schemata: these mappings express how fields or elements in the target schema are to be translated into the XML structure of EpiDoc, and vice versa. Compliant EpiDoc files are those either that conform to the interchange schema or for which a Crosswalker mapping has been provided to perform the lossless translation into EpiDoc. Complete implementation of the Crosswalker was envisaged as part of the core tooling of IDP1, with a view to converting HGV metadata to EpiDoc, but full development was deemed out of scope for the very specific needs of IDP1. The SoSOL system, however, has a clear dependency on this component, so that full development and implementation of the Crosswalker becomes a necessity under IDP 2. The Crosswalker will be made up of both static and configurable components. The configurable components consist of a mapping file that specifies the data sources and formats and maps them to a key-value format that will in turn be used as a source for generating the EpiDoc files. The mapping file, once configured for a particular type of data source, will be transformed into an executable script to effect the transformation of the data source information into a key-value XML file. The second configurable component is an EpiDoc template with pointers to keys in the XML file. When the Crosswalker process is run, values from the key-value XML file will be inserted into one or more EpiDoc files generated using the template. Source types will include databases (such as HGV), Word documents, HTML documents, and other formats. The Crosswalker component will rely on both CHET-C and the transcoder, as well as new services to be created, notably an MS Word to XML converter and an HTML parser. (**Z. Au, G. Bodard, H. Cayless**)

The dominant thrust of Heidelberg's efforts under IDP2, in collaboration with Duke, will address enhancements to the EpiDoc texts that will require a combination of iterative, labor-intensive hand- and batch-fixing. EpiDoc admits much more and deeper semantic mark-up than was originally encoded in Beta and so representable by SGML. For example, crucial structural features of documents (recto/verso, column, fragment, sub-document identifier), and even concatenations of multiple structural features may be encoded with a single tag; one result of this suboptimal granularity of structural markup is that HGV references to fragments, columns, sub-documents etc. (often the unit of analysis in the HGV) cannot be mapped precisely to the corresponding DDbDP XML fragment. Such examples are numerous, and some of them rate-limiters on complete HGV-DDbDP integration. The current EpiDoc instance of the DDbDP has offered major enhancements consistent with costs and benefits under the first phase of the project.

Under the second, we aim to disambiguate all major cases of semantic 'lumping' and fully document the rest, so that the community of users may subsequently begin to tackle what remains; some documentation and reporting was generated as part of the IDP1 effort, which will greatly facilitate the process. Heidelberg will also continue the process of populating the merged DDbDP-HGV EpiDoc library with translations, including both production of new ones and the migration of existing translations from the APIS records. It will also make enhancements to the existing glossary of technical terms, which will be opened to the user community via SoSOL. (**Z. Au, J. Cowey**, **J. Sosin**)

EpiDoc/Crosswalker Workplan

P5: The conversion of DDbDP EpiDoc to P5 has an obvious dependency on the upgrade of the EpiDoc Guidelines themselves to the new TEI release. This effort began during the development of the TEI P5 Guidelines themselves during 2007, as a voluntary initiative of the EpiDoc community; indeed, several major EpiDoc needs have already been addressed directly in the TEI P5 Guidelines. Incremental work by the EpiDoc community aimed at achieving full EpiDoc compliance is already underway, and will accelerate during summer and fall of 2008. Funding is not requested for the upgrade of EpiDoc itself, but as the EpiDoc community settles major remaining questions, it will be possible to inventory the DDbDP for features that need to be converted. Following that exercise, it will be necessary to develop a TEI ODD (definition for the customized schema), and to convert the XML and XSLT to the new schema. These specific DDbDP-related tasks will require dedicated person-time from the IDP2 team. (**G. Bodard**, **Z. Au**; domain-specific assistance by **J. Cowey**, **J. Sosin**)

Crosswalker: Development on the crosswalker will begin with a gathering of additional requirements for the schemas for the mapping files, since the initial prototype addressed only a limited set of mappings from Epigraphical databases to EpiDoc and did not directly address the need for textual transfomations (e.g. transcoding and Leiden to EpiDoc conversion). The SoSOL workflows will entail more of these types of complex mapping and multi-stage conversion and therefore the schema will need to be modified to enable the expression of these mappings. This will be followed by development of the code needed to generate an executable pipeline from a mapping file, this might entail, in a simple example, the generation of an XSLT that would handle the conversion from an XML source, plus an executable script that would call that XSLT with the correct parameters. This will be followed by the integration of processes that will need to be available to the generated code, including the transcoder and CHET-C, and the development and integration of an MS Word conversion service, probably based on OpenOffice, which will read Microsoft Office documents and convert them to an XML format. (**Z. Au**, **G. Bodard**, **H. Cayless**)

DDbDP EpiDoc Optimization: Improvements to structural markup are highest priority. We shall transform the current system of "flat" milestones to properly nested document divs that reflect sub-document structures (e.g. columns, recto/verso, fragments, etc.) and correspond to HGV analysis (a crucial feature of the current trajectory on which different selection and analysis criteria long used by HGV and DDbDP are being collapsed and merged). Reporting and scoping done under IDP1 will facilitate identification and classification of all such structural features. Transformation will require a combination of hand-fixing and batch transformation, on regular iteration between papyrologists and XSLT specialist (**J. Cowey**, **J. Sosin**, and **Z. Au**), regular consultation with the EpiDoc team (**G. Bodard, H. Cayless**) to ensure EpiDoc compliance, and finally reporting to the UKY team (**D. Porter** et al.) to ensure that the SoSOL system accommodates all improvements to EpiDoc implementation. As a result of IDP1 the DDbDP and HGV are now fully integrated and representable in single merged XML documents; addressing these issues of structure will propel the two projects toward our long-term goal of a precise 1:1 relationship between all units of analysis.

In three other areas enhancements will be made to the XML with a view to supporting the goal of transforming the DDbDP-HGV from a static guide to printed editions to an organic and evolving record of scholarly, editorial processes. First, limitations in the manner in which SGML handled apparatus

criticus entries that were broken by line-breaks or other structural markup prevented, given the scope of IDP1, comprehensive optimization of all such instances in the EpiDoc XML. Second, we shall enhance the Greek Number Converter (produced under IDP1) in order to cope better with partially preserved numerals, whose DDbDP representation as Arabic numerals was not perfectly transformable under IDP1 without significant hand work. Third, the same sorts of improvements will be made to markup of the semantically complicated abbreviations of small monetary values. All three sets of tasks will require of the same personnel the same combination of iterative hand- and batch-fixing, along with the same processes of consultation and reporting as described above. Inasmuch as one of the chief goals of IDP2 is to transform the DDbDP-HGV from a static representation of single conspectus of editorial decision-making to an organic and evolving record of scholarly, editorial processes, it is essential that the corpus of texts be able to accommodate even the most semantically complicated markup. (**J. Cowey**, **J. Sosin**, and **Z. Au**)

Work on translations will also proceed apace, with priority given to the entry of some one hundred and fifty prepared translations according to the newly created and now standard EpiDoc mark up system. New translations will also be prepared and entered as possible (**J. Cowey**). Also, the HGV translation glossary will to be reviewed with a view to better automating current display processes and alienating the glossary as a tool editable by the community via SoSOL (**G. Bodard, H. Cayless, J. Cowey, J. Sosin, Z. Au, D. Porter**).

Sustainability

The sustainability of digital humanities projects remains an area of great concern for which no universal, long-term solution can yet be offered. Nonetheless, IDP2 achieves significant advances in this domain by building upon a range of earlier investments in refactoring and capitalizing on fresh opportunities provided by technological change and the new collaborative models now emerging. We are taking a multi-pronged approach to the critical issue of sustaining not only the DDbDP-HGV-APIS cluster, but also a host of related resources that will grow over time. Some of these resources exist already, whereas others will emerge once the scholarly opportunities made feasible by a more diverse and agile approach have been realized. We expand the number of institutions committed to the maintenance of each of these resources, engage a much larger body of scholars in not just the use but the creation and maintenance of these resources, and establish a non-exclusive, flexible technical architecture for the dissemination of readily reusable data and the deployment of special-purpose software.

Our approach tackles two separate but related problems of scaling that can compromise the longterm viability of digital projects in the academy. First, individual scholars create more or less self-contained projects, usually built to customized and closed standards. But these projects, unlike books, take on a life of their own and the best ones quickly grow beyond the capacity of their founders to maintain, expand, and enhance the content. Second, within the community of any given university or even discipline, this story plays out repeatedly: individual, self-contained, faculty-led projects quickly multiplying beyond the capacity of institutional sponsors to maintain, expand, and enhance the technical infrastructure and architecture. This puzzle of scaling is in an important sense the natural and predictable result of traditional institutional behaviors in humanistic settings: scholars build things (print publications), sometimes together, but mostly alone or in small groups, and institutions see to the preservation of single instances of them (books housed in libraries). This pattern is not news to anyone, but that does not make it easy to break.

We aim to depart from this familiar pattern by re-focusing attention from *projects* to *infrastructure* and *methods*. This recalibration occurs on two axes: (i) a consolidative approach to economizing effort through development and sponsorship of common tools and centralized tool-management, -development, -migration, and (ii) direct application of a collaborative, crowdsourcing model that enables entire scholarly disciplines to take charge of their core intellectual assets instead of acting as passive consumers of a dataset whose curation is the responsibility of a special academic staff.

Thus, the same attention to architecture allows us to ameliorate, at once, the scaling challenge faced by universities, who are increasingly hard put to manage the rapid growth of self-contained projects, as well as that familiar problem of scaling faced by individual scholars, whose projects grow in both size and duration beyond the limits of any one person or team to see to their survival.

There is no silver bullet in the sustainability landscape, but if we can get a discipline to agree on a shared approach to data encoding (first steps accomplished under IDP1, and over the years by APIS), a shared set of tools for accessing that data (first steps accomplished by PN and in part under IDP1), a shared set of tools for controlling that data (the goal of IDP2), release and licensing of both tools and data for the widest possible range of re-use; and a new mode of doing work (digital not paper, collaborative not solo; also the goal of IDP2), we shall have accomplished a lot. Then, the most pressing issue remaining is organizational: how do we ensure that encoding standards, tools, and social infrastructure evolve in mutually aware fashion and who will manage this process such that economies and efficiencies can be brought to bear? No solution can be permanent, but we aim to solve at least this piece of the puzzle for papyrology, for at least this generation. Alongside its funded deliverables under IDP2, ISAW plans, both using its own existing funding and forging paths to new funding streams, to generalize this approach and tooling to address what is essentially the same set of sustainability and interoperability problems in the epigraphic subdiscipline, and even beyond. That this step can even be contemplated is an illustration of the validity of the approach and a validation of its earliest steps: papyrological use of the EpiDoc standard (originally developed for epigraphy) means that tools built for papyrological work can be repurposed for epigraphic use at a fraction of the cost that would attend creation of epigraphic tooling from scratch.

Key Aspects and Rationale

The tightly interwoven dimensions of this sustaining architecture are fourfold: social, technological, archival, and institutional.

The current data-population model of the DDbDP, and to some extent that of the HGV and APIS too, is neither sustainable nor, in the light of new technological opportunities, intellectually and professionally optimal. Centralized, top-down, post-print reflection of scholarship that has already happened is slow, error-prone and intellectually authoritarian. It is vulnerable to acute, systemic failures, and is extraordinarily expensive. The creation of SoSOL is an essential step toward rectifying these liabilities. SoSOL is in effect the air and water of a new disciplinary ecosystem in which members assume much broader control of the integrity of core data-sets and the pace at which these materials are uploaded and improved. It will accelerate the speed at which errors (whether in content or technical framework) are identified and the creativity with which they are rectified. It will drive down the untenable concentration of costs necessary to centrally maintain data-sets, while ensuring that these remain maximally useful to the community of scholars

SoSOL and the new collaborative environment that it will foster is a force-multiplier, in effect redistributing the current editorial workload to a multiplicity of stakeholders. From the point of view of individual data-sets, it should eventually drive the cost of content-creation to near zero (if editors of print publications can upload their data with a minimum of fuss, they will; and if they do not, someone else interested in having access to the material will). These effects are quantitative. But the SoSOL system brings the entire community of scholars so much more closely into the editorial circle, even allowing peer-reviewed publication directly to the shared data-set, that the global and emergent effects are qualitative: the DDbDP-HGV is transformed from an environment in which scholarship is mirrored to one in which scholarship happens; the traditional black box of peer review becomes a transparent forum in which review is controlled sometimes by editors, sometimes by the community, and the consensus of both constituencies are able to inform each other iteratively and in real time.

In other words, at the heart of our sustainability plan is a piece of disciplinary social reengineering, all of which requires a robust and generalizable technological underpinning. The new

collaborative behaviors and workflows of a more fully digital papyrology will require flexible, extensible, standard-based tooling for creating and changing data (SoSOL, EpiDoc Crosswalker), accessing and manipulating known relatable data (PN, informed by mapping work conducted under IDP1, and to be conducted under Concordia and APIS 6), and establishing relationships between and exchanging data across known but parallel data-sets (Concordia). Consolidation and economization of such tooling under the aegis of a dedicated programming and support team, while open-sourcing all of it to the wider community of papyrologists and digital humanists, will produce efficiencies of tool-development, -maintenance, and -migration, similar to those achieved for papyrologists by the tools themselves.

It is not enough, however, to provide the discipline with distributed control of ever-evolving core data-sets and robust tooling with which to manage them. In digital settings, as in analog, the fundamental scholarly act remains reference, and this demands a robust mechanism for stable, archived, release of both data and tooling. All IDP participants will ensure that tools and data (including marked up texts, scholarly apparatus, and translations) created under the auspices of this program will be properly licensed and released. For software, this means the assignment of the GNU Public license and the alienation of a complete release package through SourceForge or an equivalent third-party software release management system. The GPL is an obvious choice because SoSOL will incorporate code already so licensed (including the EpiDoc tools) and because the project principals believe that requiring any future, derivative developments of the system proposed here to release the source code contributes to the sustainability of the code base. We do not plan at this point to foster a new community of software developers (not that such would be at all unwelcome), but we can easily envisage other potential uses and users of the software. We wish, therefore, to ensure that if a developer community should arise it will continue the same policy of transparency and openness that IDP participants hold as a core operating principle. For the EpiDoc-encoded texts, apparatus, and translations we will create a contributor agreement that assigns to the DDbDP-HGV-SoSOL host a nonexclusive, irrevocable right to republish and distribute contributions as long as the contributor is cited. We will obtain similar agreements with participating institutions for preexisting material that will be loaded into the initial incarnation of SoSOL. This agreement will give the project sufficient rights to assign a Creative Commons Attribution license (http://creativecommons.org/licenses/by/2.0/) to the material distributed online. We hope and expect the texts and translations to have wide distribution and we wish by attaching an appropriate license to the data to help ensure that it is uncomplicated for scholars and other processing or delivery environments freely to use and re-use the material we produce, as expected in academic practice. The project will also prepare appropriate metadata for each release in conformance with the newly-promulgated JISC Version Identification Framework (defined dates, unique identifiers, version numbering, version labels and textual description). Procedures for the creation and application of this metadata will be incorporated into appropriate project workflows so that they persist beyond the end of the IDP grant period. Furthermore, ISAW will provide stable, long-term access to all versioned releases of IDP-associated datasets from an NYU server (on the model already employed by the Stoa Consortium and the EpiDoc community for its guidelines and DTD: http://www.stoa.org/epidoc) and will ensure ingest of copies into the NYU Faculty Digital Archive, the Internet Archive and, as appropriate, other archives such as the UK Archaeological Data Service.

Institutional commitments are essential to the success of any scholarly enterprise; however the most common institutional approach to digital humanities *projects* must be complemented by equally aggressive promotion of standards-based architecture (for both economization of technical overhead and force-multiplication of content-creation efforts), if the sustainability challenge is to be met. Our complementary model will in effect leverage a shared technical architecture to subdivide the costs of projects into smaller bundles more readily accepted by university and college administrations as the day-to-day cost of doing scholarly business, while paving the way for projects to thrive—nourished by the wider user-community—beyond the attention and lifespan of their scholarly creators.

ISAW's Commitment to IDP2 and Future Directions in Sustainability: Backstop

ISAW views its commitment to the hosting of the technical infrastructure resulting from IDP2 as a long-term, core feature of its institutional responsibility to the discipline of ancient studies. Thus, collaborative digital papyrology and epigraphy are the tip of the spear in ISAW's newly-launched "Backstop" initiative.

ISAW has made digital innovation a top priority and within that context, is allocating both institutional and externally-sourced resources to help retool the discipline for more sustainable digital scholarship and publication. As part of this work, ISAW is developing this major collaborative initiative focused on the preservation and long-term vitality of born-digital scholarly works for ancient studies. Although this initiative starts from the premise that even the basic survival of digital files cannot yet be taken as assured, it grows more directly from the insight that long-term sustainability of digital resources depends above all on a user community that both relies on a tool or dataset and is willing to invest its efforts in seeing these remain vital and current. From this perspective, even the survival of digital files depends on robust organizational and institutional structures as much as on technology or finance.

ISAW looks to develop the basis for long-term management of digital assets for the study of antiquity on the intersection of organization-building and technology in the creation of scholarly communities to sustain common tools while diversifying authorial and editorial control. In the first stage, these structures are being created around two of the most complex bodies of material, in the belief that they can be generalized far more broadly from that base. One of these initial digital clusters is the body of digital papyrological resources currently managed under the APIS, DDbDP and HGV rubrics. The other, digital historical geography as organized around the Pleiades project (http://pleiades.stoa.org), supports many subdisciplines, including papyrology and epigraphy.

Key tenets of the approach taken by ISAW and its partners are the development and distribution of free, open-source software for common scholarly tasks and the creation of distributed, collaborative frameworks whereby communities of practice are empowered to take direct and active control of their digital output and work materials. In collaboration with the NYU Digital Libraries Team and other institutions, ISAW pursues this mission in five ways:

- assisting communities of scholars, students, and enthusiasts in creating and selecting broadly collaborative regimes for the sustainable creation, editing and dissemination of resources;
- identifying exemplary resources for immediate bit-preservation through deposit into multiple repositories, with a focus on public-domain and open-licensable materials;
- facilitating the rescue of essential digital works whose originating institutions or authors can no longer host or manage them by embedding them in a robust, multi-institutional matrix of shared tooling and dissemination mechanisms;
- working with individual scholars, institutional partners, and third-party entities to facilitate the conversion of legacy resources from obsolescent, idiosyncratic and proprietary formats to open, standards-based versions that can be more confidently preserved and redistributed;
- developing guidance and services for creating, disseminating, preserving, and updating the varied modes of digital scholarly output, especially via elements of the emerging cyberinfrastructure for digital humanities.

The sustainability rationale outlined above was pioneered by DDbDP and also by Pleiades, but has now expanded to embrace HGV, APIS and other entities through the combination of NEH and Mellon support. It is an exemplary model, and must be brought to fruition through the steps outlined in this proposal, not only for the sake of papyrology, but as a prototype for the large number of similar renovations of other endeavors that must be undertaken over the next 20-40 years if we are to avert their

loss or obsolescence.

Because ISAW leadership sees the value of this example, as well as the high percentage of reusable, cost-saving tooling that will emerge from IDP2, it is committing its own resources (in the form of 20% of T. Elliott's time) to design, coordination, organization, standards-verification, release management and software troubleshooting needs across the span of the proposed work. ISAW also commits to the long-term hosting of the production DDbDP and HGV (EpiDoc version) dataset repository, as well as the Papyrological Navigator and associated software, as well as archive deposit of all released tools and datasets. Baseline archiving of content will be guaranteed by the deposit of copies into the NYU Faculty Digital Archive, whose maintenance is already an institutional priority at NYU. Full APIS production support, including the core APIS database, image repository and collections management application will to migrate from Columbia to NYU in 2010. ISAW plans to deploy a combination of internal budget, 10-20-year spend-down funds and permanent endowment (to be raised), alongside short-term grants for specific projects, to facilitate on-going maintenance and periodic upgrade of common tools and formats. This approach will be supplemented as necessary by separate grant- and/or internal-funded initiatives (in collaboration with content communities as appropriate) to exploit opportunities presented by technological change.

From the perspective of the digital papyrological projects, this is a better, more efficient, approach than that originally envisaged under IDP1, namely the creation of a specific endowment to support APIS and DDbDP. ISAW establishes and maintains core technical systems at a stable institutional base (and a few partner institutions) to support a large number of resources and datasets, thereby achieving significant economies of scale and ensuring ready availability of humanities-centric programmers and tools. Impact is increased further by combining this "core technology services" model with globally distributed content creation and maintenance procedures. This shifts most of the content-related costs from small groups of scholars and institutions (where the aggregation of said costs makes many projects unsustainably expensive from an institutional perspective, especially as faculty in specialties move or retire) to the home institutions of most of the active scholars in an entire discipline (where the much smaller fractions of the cost can be absorbed simply as part of the day-to-day research, publication and professional activity of the faculty). No doubt, from time to time, specific research and publication projects will require the assignment of time-limited content- or technology-related personnel. ISAW will act as a coordinator, institutional host or subcontractor for such employees and consultants as needed in consultation with individual user communities, just as it now commits to do for IDP2.

Coordination

Regular coordination will be necessary and the whole team plans to continue the regular practice developed under IDP1, managing regular workflow through a combination of (a) weekly whole-team conference calls, as well as smaller assemblies as needed, (b) regular reporting via email list, (c) group-authored documentation in shared work spaces (a project Wiki and GoogleDocs), and (d) storage of tools and data in Subversion repository. Inasmuch as D. Porter (UKY Project Coordinator) has long since been read into existing IDP1 communication processes, the transition from IDP1 to IDP2, as far as coordination goes, will be seamless.

Several components of the work can be conducted on a schedule more or less independent of certain others; for example, the results of the EpiDoc optimization must be fed eventually to the PN development programmer and to the SoSOL development team, but time-sensitive dependencies here are relatively flexible. Nevertheless, at certain points in the process members of the team will need to coordinate more closely. Thus, under IDP2 we propose to continue another mechanism that has proven invaluable to IDP1, periodic weeklong worksprints, at which larger groups of the whole team assemble to engage in design or redesign of components, data formats, and processes, optimize workflow, identify and manage unforeseen dependencies or requirements, adjust internal scheduling, troubleshoot bugs, etc.

Conducting these exercises in person, in an environment temporarily removed from day-to-day distractions, effectively jumpstarts the next stage of development and is an indispensable part of the project work.

We shall hold worksprints at three critical junctures in the project. In October 2008, the UKY team will have commenced to modify an editor for SoSOL and will have begun the process of interface design, both of which suggest dependencies on other processes, not least of which preparations for production installation at ISAW and constraints on implementation of papyrological behaviors and conventions. UKY will also have established transfer protocols with the VCS and NYU, which, along with concurrent P5 development and the start of EpiDoc optimization process, will require careful collaboration with regard to workflow management and associated burdens on ISAW. This will also be the time for SoSOL designers and papyrologists to review SoSOL workflow contingencies, whose design will at this point be a quarter to a third done.

In February 2009, the first round of the XML editor's interface design will be done and PN integration will be commencing; this will be the moment for papyrologists and SoSOL designers to work very closely to ensure that the design specs have met papyrological conventions and expectations. This will also be the moment for ISAW staff to coordinate very closely with SoSOL designers to ensure that the design of the XML editor's interaction with NYU version control is optimized for production. This sprint will also provide SoSOL designers, ISAW staff, and papyrologists the opportunity to assess progress on the interface for the workflow manager, whose challenging design will by this point be halfway complete, as well as providing papyrologists and SoSOL designers the chance for careful review the Crosswalker design and framing of any late breaking design needs for the MSWord-to-XML converter, before its production commences. Finally, EpiDoc optimization will at this point be nearing completion such that this sprint will provide the final opportunity to ensure that the effect of improvements on SoSOL and PN performance are fully mapped and accommodated.

Our last sprint will be held in June 2009, at which point the XML editor will be about to enter its final phase of interface development, implementation of Crosswalker will be just completed and SoSOL testing will commence. The VCS will be about to start its final phase of production implementation at NYU. This will be the team's last opportunity to troubleshoot unforeseen problems, to spec and implement any late-breaking enhancements to Crosswalker, to identify and begin to address any unforeseen challenges to SoSOL or PN performance.

Sosin will shoulder project management duties, coordinating all regular communication channels, facilitating timely information exchange, tracking internal deadlines, coordinating iterative review and feedback for all components, tracking and controlling cross-project dependencies, arranging logistics and agenda-formulation for worksprints, and all related tasks.

Chronological Distribution of Tasks

		Year 1 (2008/09)												Year 2 (2009/10)											
		S	O	N	D	Ja	F	M	A	M	J	J	A	S	O	N	D	Ja	F	M	A	M	J	J	A
SoSOL																									
XML Editor	XML editor component interface design	x	x	x	x	x						x	x												
	Modify to handle generic XML		x	x	x																				
	Design interaction with version control			x	x	x																			
	Schema Validation						x																		
	Integrate Transcoder					x	x	x																	
	Integrate XML editor with PN interface (image, metadata)						x	x																	
	Build interaction with PN search							x	x	x															
	Implement Crosswalker as SoSOL plugin									x	x														
	Testing and refinement										x	x	x												
Version Control Management																									
Version Control System	Set up VCS and transfer files from SVN at NYU	x	x																						
	Integrate with workflow manager					x	x																		
Abstraction Layer	Design abstraction layer	x	x																						
	Set up per user branching in VCS			x	x																				
	Allow commits from XML editor to user branch					x	x																		
	Merge user branches to trunk							x	x	x															
	Handle conflicts between user branches and trunk										x	x	x												
	Integrate with production SVN at NYU											x	x												
Workflow Manager																									
	Design the workflow	x	x	x	x			x	x																
	Build on top of Version Control System					x	x																		
	Implement the workflow								x	x															
	Test the workflow									x	x														
	Design reporting capabilities											x													
	Implement reporting capabilities												x												
Account Management	Design system for user registration	x	x																						
	Implement system for user registration			x	x																				
Web User Interface	Build web app/interface for editors				x	x	x	x	x																
System	Complete testing				x				x		x	x	x												

	Deployment										x	x	x														
	Developers' documentation	x	x	x	x	x	x	x	x	x	x	x	x														
	User documentation										x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	Training													x	x							x		x			
PN	Cross component integration	x	x	x	x	x	x	x	x	x	x	x	x														
	System administration	x	x	x	x	x	x	x	x	x	x	x	x														
	Bug fixing and feature enhancement			x	x	x			x	x	x																
	Streamline user interface components										x	x	x														
	SoSOL/PN/Portlet integration and test		x			x	x				x	x	x														
	Specific PN/SoSOL and PN/VCS integration and deployment tasks in partnership with UKY team (dates as indicated above, sub "SoSOL - XML Editor" -> "System")																										
EpiDoc: P5	Identify features that must be converted to P5	x	x																								
	Write an EpiDoc ODD		x	x																							
	Create and iterate on conversion XSLT			x	x	x																					
Crosswalker	Design mapping file schema (dependent upon P5 features)		x	x																							
	Develop code generation from mapping files				x	x	x																				
	MS Word to XML converter							x	x																		
	develop interfaces to SoSOL API									x	x																
EpiDoc: Optimization	structural clean up and improvements	x	x	x	x	x	x	x																			
	display modifications								x	x	x																
	glossary modification and integration					x	x	x																			
	translation entry and integration		x	x	x						x	x	x														

NOTE: Months featuring Work Sprints in GREY

APPENDIX: Unfunded Partnerships

Concordia

Concordia is a collaboration of the Institute for the Study of the Ancient World at NYU and the Centre for Computing in the Humanities at King's College, London. Under the co-direction of Roger Bagnall (ISAW) and Charlotte Roueché (King's), this project will establish geographic and textual interoperability services for a number of document collections, including the IDP1 version of the integrated DDbDP-HGV data-sets. Funding for Concordia has been provided by joint grants from the National Endowment for the Humanities and the UK Joint Information Systems Council (April 2008 - March 2009). Geographically, the project embraces more than 830,000 square kilometers: from Tripolitania (southern Tunisia and western Libya) along the coast eastward to the Nile delta and then southward up the river to include the Fayum and several of the best documented ancient administrative regions (the Oxyrhynchite, Kynopolite, Herakleopolite and northern Arsinoite nomes). The temporal coverage of the geographic data extends from the Greek archaic period to the 5th century CE, but will be expanded to address later toponymy and topography attested in the papyri.

Concordia will employ the new Object Reuse and Exchange (ORE) specification, promulgated by the Open Archives Initiative, to underpin open-access web services across separately hosted digital collections. ORE linking of texts and metadata records to Pleiades place records will facilitate automated geographic correlation, and dynamic mapping alongside more traditional search modes (metadata fields, substring). When Concordia is complete, it will be possible to map the findspots of texts in the DDbDP-HGV dataset and to browse from one papyrus text to inscriptions, coins, or other papyri found nearby.

Facsimile mapping

SoSOL will include simple image viewing functionality (allowing editors to view images of papyri alongside the transcription editing window), and development of SoSOL will be conducted with a view to the later addition of text-image mapping functionality. Collaborators from the UKY group and the University of North Carolina at Chapel Hill are currently seeking partners and use cases for a project to develop a system for incorporating images into the collaborative, networked editing process. This project will address the issues of image manipulation and comparison, image annotation, linkage of images and text (automated, semi-automated, and by hand), and building the metadata framework(s) that support these other operations. We anticipate incorporating these image-focused tools in SoSOL at a later date, so that all SoSOL development conducted with that eventual goal in mind in effect paves the way for the creation of this tool, which has long been a desideratum not only of papyrologists.

Confirmed partners for this image mapping project in addition to IDP include John Walsh at Indiana University Bloomington (Swinburne Project and the Chymistry of Isaac Newton Project), Stephanie Wood at the University of Oregon (Central American Mapas Project), Christopher Blackwell at Furman University and Neel Smith at College of the Holy Cross (Homer Multitext Project) and Aaron Kleist at Biola University (Electronic Aelfric Project). We are submitting a proposal to support the design and implementation of these image mapping tools to the National Endowment for the Humanities, under the "Preservation and Access: Humanities Collections and Resources program (Research and Development focus)," with a deadline of July 31, 2008.

APIS 6

During the next round of NEH support, APIS 6 (July 2008 - June 2010), APIS will refine and enhance existing components of the central system's internal architecture in order to facilitate addition of new content and to improve the general quality of the metadata. As a major enhancement to the internal

architecture, APIS will reengineer the current online editing interface to allow remote input and update of records by its partners, which will especially benefit those institutions that do not have technology staff available to maintain a local database, and will have the significant added benefit of driving down costs currently shouldered by the APIS technology host, by displacing the burden of upload and any necessary data-massaging from the host to contributing institutions (a model that is consistent with that envisaged for DDbDP-HGV). Similarly, APIS plans to investigate and, if feasible, implement XML interoperability with non-APIS digital projects such as the German Papyrus Portal and so extend the APIS data model for use in Europe. Moreover, to improve the quality of the metadata, APIS will also work with the HGV to formalize controlled vocabulary for genre types and subjects, and then to map corresponding German/English terms in order to permit bilingual keyword searching.

One of APIS's largest contributions to scholars and other interested parties has been the ability to offer high quality images of ancient objects. APIS 6 will extend that ability by continuing to expand the repository of TIFF images, allowing APIS both to offer ongoing preservation support to its partners and to present high resolution images in the PN's FSI viewer, whose level of functionality is superior to other image viewers. APIS intends also to work together with the multi-spectral imaging project at Brigham Young University to explore ways of displaying BYU's multi-spectral images.

A critical part of APIS's ongoing integration with other projects hosted within the Papyrological Navigator is the development of a strategy for feeding new content, whether via batch contributions or harvesting, to the PN, which will become the sole public interface for the APIS project.

Under APIS 6 the addition of new material will continue to take place at Columbia, while the responsibility for technical development will be assumed by staff at NYU. At the close of APIS 6, in 2010, NYU will become the new APIS technology host and the entire database and supporting software will migrate there.