Integrating Digital Papyrology

Roger S. Bagnall (Institute for the Study of the Ancient World,
New York University)

Integrating Digital Papyrology is a joint effort of three existing projects in the field of papyrology: the Duke Databank of Documentary Papyri (DDbDP), the oldest digital resource in the field, now nearly thirty years old; the Heidelberger Gesamtverzeichnis der griechischen Papyrusurkunden Ägyptens (HGV); and the Advanced Papyrological Information System (APIS).[1] DDbDP is a textual databank of Greek and Latin documentary (but not literary) papyri; HGV is a collection of descriptive information (in the category for which we usually use the jargon term "metadata") on published Greek and Latin documents, essentially congruent in its universe with DDbDP but not including the texts themselves; and APIS is a union catalogue of papyrus collections, now numbering about thirty such collections, with metadata and digital images. The home bases of the projects are respectively Duke University, New York University (from 2009, formerly Columbia University), and the University of Heidelberg.[2]

[1] I am indebted to Tom Elliott and Josh Sosin for comments on and significant improvements in early versions of this paper. They are partners in this venture .

[2] DDbDP was initially funded by the Packard Humanities Institute; it has subsequently had funding from NEH, which has also been the main funder for APIS, along with the participating institutions and a

The primary aim of IDP,[3] in its first phase, was to make it possible for researchers to get access to data from all three databases simultaneously, working through a single interface. This goal, which seems less than revolutionary today, was one I proposed in 1992, to the bafflement of most the assembled papyrologists at their international congress in Copenhagen. It has largely been realized with the creation of the Papyrological Navigator (www.papyri.info), although there is still much improving of the capabilities of that engine possible and certain to be achieved in the next couple of years.

Of course, the world of the Web has changed dramatically since 1992, and the possibilities today are much richer than they were then. I would say that these changes have affected the vision and goals of IDP in two principal ways. One is toward openness; the other is toward dynamism. These are linked. We no longer see IDP as representing at any given moment a synthesis of fixed data sources directed by a central management; rather, we see it as a constantly changing set of fully open data sources governed by the scholarly community and maintained by all active scholars who care to participate. One might go so far as to say that we

---

number of private donors. HGV has been funded primarily by German public sources, especially the Academy of Sciences in Heidelberg. IDP has been funded by a planning grant and two implementation grants from the Andrew W. Mellon Foundation.

[3] IDP is co-directed by Joshua D. Sosin, Duke University, who is the principal project leader.

see this nexus of papyrological resources as ceasing to be "projects" and turning instead into a community.

Part of that evolution implies an abandonment of the distinction operative in this and other fields until now between editing texts and maintaining textual databanks. The central feature of the second phase of IDP is the creation of an online editing system that will allow entry of texts into DDbDP and metadata into HGV and APIS by any authorized participant. Even more, it will allow the creation of editions in the first place inside this editorial system, editions that become publicly visible only when the editor chooses to make them so. Instead of someone retyping texts published in print form into the database, the boundary between scholarly creation and the database is a dynamic one controlled by individual contributors, the editing software, and the board of editors—capable of effacement at a keystroke. Where the three source databases at present contain some duplication of information, not to mention some contradictions, the editing system will lead to a gradual blurring of the lines between them.

There is another implication of the direction we have taken: both data and code will be fully exposed. Anyone who wants can write an independent interface, whether it be a cleverer way of doing the same things we are building the Papyrological Navigator to do, or a means to do something we have yet to imagine with our data, perhaps by drawing on a yet richer array of sources from other

places, periods, or genres. (At least two major projects are already reusing our data and code, the Leuven Names project and a British Library Byzantine manuscripts project). In this way we are breaking up the vertical integration of our interface and our data, not by getting out of the interface business, but by allowing users to choose potentially from multiple providers of services based on at least the same set of information.

IDP does not exist in a vacuum, of course. It is founded on the use of an XML encoding standard called EpiDoc, which is a customization of the widely used Text Encoding Initiative tag set.[4] EpiDoc was originally developed by Tom Elliott, who is now associate director for digital programs at ISAW, for the coding of editions of inscriptions, rather than papyri. EpiDoc was built with the larger world in mind, and it is capable of extension to other categories of documents, of which papyri have been the first but not only example. There has now been serious work done on using EpiDoc for recording coins. Because EpiDoc uses standard TEI elements, these documents are all fairly open to searching by a variety of compatible tools, and they will collectively make it possible for us to imagine search interfaces that will interrogate a range of ancient sources of different types. We have in the past year, with funding from the joint NEH-JISC program, developed a prototype for this kind of integration across the boundaries of

[4] Text Encoding Initiative: http://www.tei-c.org; EpiDoc: http://epidoc.sf.net.

papyrology and epigraphy, called Concordia. This reuses various part of our toolkit and anchors the documents to the ancient geographical database Pleiades, which is (as the name is intended to suggest) the daughter of the *Barrington Atlas of the Greek and Roman World*.[5] Pleiades, which is also Tom Elliott's invention, has developed a community editing facility to allow registered users to suggest places to be added, additional information to be incorporated, corrections to locations or other information, and any other improvement needed. These suggestions enter an on-line editorial process in which the contributor can modify them and then submit them for scrutiny. An editorial board oversees a process of community-wide peer review of these contributions before they are accepted and become publicly visible.

The integrated papyrological system will operate on similar principles. Any user may contribute readings, emendations, new texts, translations, or metadata, their proposals being subject to the approval of an editorial board. Because the editing system records and attributes all such interventions, from proposal, through vetting, to ultimate acceptance or rejection, along with (mandatory) prose justifications at each step, the community of both editors and general users may revisit and reverse any action at any time. For this reason, it is possible to adopt a latitudinarian policy on contributors for both Pleiades and IDP; the disruptive or

[5] The first phase of Pleiades was funded by NEH, which currently has an application for its extension under consideration. The *Barrington Atlas*, edited by Richard Talbert, was funded by NEH and a large number of foundations and individuals.

incompetent can be weeded out after this is discovered, rather than subjected to a long up-front interrogation and background check. An advisory group will vet all proposed additions of material and corrections to existing content before they are made “canonical” (before they are promoted to the text or apparatus, proposals can be kept limited-access or made public, at the proposer’s option); after that, any registered user can contribute suggestions, subject again to the same vetting process before incorporation. This also means that ideas that do not pass muster are not irrevocably discarded; rather, they are retained, in the digital 'paper trail', whence in the light of new data or better arguments they may later be revived.

Josh Sosin, who is a man of deep democratic principles, sees this approach as helping to bring an end to the era of hegemonic (or tyrannical, if you prefer) management of projects. As someone with the soul, if not the training, of an economist, I see this as a way of acquiring a lot of free skilled labor. Both are probably true. (And, to be fair, Sosin is an economic historian and also likes free labor.) Hegemony will at least be distributed across a broader and more representative set of papyrological scholars, even if it is not abolished altogether.

Everything I have been talking about is scary at some level and to some people. It certainly frightens some of my colleagues. For one thing, despite all assurances that content in the system will be stamped with its origin and associated rights—branded, in other words—some possible collaborators are reasonably

anxious of the possibility that their visibility will be diminished by the incorporation--which sometimes sounds to them like *absorption*--of their data into a larger universe. This worry was expressed very explicitly and early by another papyrological aggregator, the group based in Leuven. Their Trismegistos provides access to a number of databases, mostly created in Leuven but some created elsewhere (the HGV is included there also), by means of a single identifier added to the record in each database that allows their system to find all hits for that item. But these remain entirely separate; you move from one database to another to find out what each has about an item that interests you. Highly distinct branding is central to their approach. They have built in links to DDbDP and APIS, but again as separate resources to which the user goes. The data are not exposed for web services. But they have begun to incline strongly toward a more open and integrated approach and we have started discussing how to integrate Trismegistos more closely with the PN. These moves are enormously encouraging. It may be worth observing here that the HGV undewent a similar change and was originally motivated by the exact same concerns. We started IDP with the understanding that the online editing environment might operate on DDbDP texts but not on HGV, which has maintained and defended a more rigorous "authorial" mission since its inception. But this position quickly eroded, so that by the close of IDP2 the group will issue archival (developer-ready) XML copies of both HGV and DDbDP, both

stamped with the same creative commons license (CC-BY), both fully editable by the same online tool. The project has now been modified as well to include capabilities for editing and creating APIS records through the same process.

In some, but by no means all, quarters of the epigraphical world, we have found a higher level of resistance to opening up data, although this too is beginning to change. We have been collaborating with the Epigraphische Datenbank Heidelberg on plans for the development of tools that would allow ready interchange of data between their format and EpiDoc, in both directions. Now if we could just persuade those who control the PHI epigraphical corpus to do the same, we would be close to our goal of having the entire ancient Greek and Latin documentary corpus in open form.

These concerns about branding are not idle. Institutions fund projects in part because they give them bragging rights, and cooperation always risks the distinctiveness and distinction sought. Scholars pioneer them, at least in part, out of similar concern for credit and distinction. But keeping data in silos accessible only through one's own interface has risks too, and in my view they are greater—the risk that search engines will ignore you and you will therefore reach a much smaller audience. Our purpose in existing is education; the more we shut out potential users who will come at the world through Google or similar engines, the fewer people we will educate. That to me is an unacceptable cost of preserving the

high relief of the branded silo. Moreover, these resources will never reach their full value to users without extensive interlinkage, interoperation, and openness to remixing by users.

The other major concern that I have heard is quality control. This was expressed to me by a European colleague in respect of the *Berichtigungsliste*, a remarkable research tool in papyrology that collects periodically—there have been twelve volumes since its inception in 1915—all corrections proposed to the texts of papyrus documents (the universe of DDbDP and HGV), new datings or provenances suggested, and a fair amount of bibliography about the documents. It has for two generations been a joint project of Leuven and Marburg, now Leuven and Heidelberg. Before corrections are registered now, the editors of the *BL* do their best to check them to see if they think they are correct; if not, they are reported but with disapproval attached. How, my friend asked, will we prevent people from just putting in fanciful or idiotic proposals, thus lowering the quality of this work?[6]

The answer is partly that one can incorporate the same kind of editorial scrutiny that I have described as inherent in the structures adopted by Pleiades and the papyrological group, partly that this editorial structure can go in and remove

[6] It has to be said that the quality and comprehensiveness of the *BL* have varied markedly over the years. The idiosyncrasies of particular editors have at times colored its coverage.

something malicious or otherwise inappropriate. And the open format of these tools helps ensure that where one or more editors have missed such, the community will let them know about it. These systems are not weaker on quality control, but stronger, inasmuch as they leverage both traditional peer review and newer community-based 'crowd-sourcing' models. The worries, though, are the same ones that we have heard about many other Internet resources (and, if you think about it, print resources too). There's a lot of garbage out there. There is indeed, and I am very much in favor of having quality-control measures built into web resources of the kind I am describing.

But that does not fully address the concern, because this is not about only quality control but control itself. "Ist mein, ist mein!" People who have created or curated projects are possessive. This possessiveness has its good side; it leads to personal investment. But in the end we possess nothing, because we are mortal; and our institutions, even if undying, do not tend to steer straight courses with unvarying purposes and priorities. They abandon our beloved projects when something new comes along. We could all cite examples. Control is the enemy of sustainability; it reduces other people's incentive to invest in something. The same thing could be said of our books; it's just easier to rework and reuse digital content.

I have for some years had my thoughts very much fixed on the argument Don Waters made about the components of sustainability: in a nutshell, a product

people want, a functional governance structure, and a workable financial model. Most of what is written about sustainability is about the last of these, but it is only part of a larger whole. Whether what we have produced and are producing now is something the community will want to support, time will tell. There are encouraging early signs in the form of interest from the projects on Ptolemaic names and Byzantine manuscripts already mentioned, and even from one on Victorian novels. In the area of governance, APIS has developed a successful consortial model. We still need to extend it to the larger partnership, but since there are representatives of Duke and Heidelberg in the APIS board and executive committee, this has not seemed urgent.

On the financial front, we are convinced that too much of the discussion is focused on revenue and not enough on expense. Despite lip service paid to the expense side of so-called business plans, most of the time it is revenue that is the central concern. I do not think there is any viable earned-income option for papyrology. A representative stakeholder group discussed the question already five years ago, and it was clear at that meeting that (1) some proprietors of data were on principle averse to charging; (2) no coherent authority over a large enough part of the data existed to achieve market power to charge users; and (3) the market for papyrology probably is not large enough to yield a serious revenue stream. Most revenue ideas other than direct subscription charges that I have read about over the

intervening years would in my view generate greater costs in cash and leadership time than they would repay.

We passed subsequently through stages that will be familiar to many present. Like everyone, I suppose, we started with the idea of creating an endowment for APIS. As it became clear that the world would soon be full of dozens of campaigns for similar endowments, like the one the APA is currently conducting for the American Office of *L'Année philologique*, this seemed less and less brilliant. The next stage was for APIS and DDbDP to think about a joint endowment. This was a better idea, but not good enough. What stopped us from proceeding, however, was not doubts—however justified—about whether we could raise the money. It was the lack of a coherent idea of how much money would be needed. What exactly did we want to fund in perpetuity? The more rapidly our work developed and we saw the essential identity of what was needed for papyri and inscriptions, as well as other types of evidence, the more we realized that not only is APIS or DDbDP not a defensible silo, neither is papyrology. If we share the essential data structures and the tools needed to edit and exploit these types of evidence, why should the projects themselves, the offspring of a moment in time, be central to our concern? Even the disciplines themselves are arbitrary divisions of a seamless spectrum of written expression that includes graffiti, lead tablets, inscribed potsherds, wooden tablets, tablets with wax surfaces, bones, and

so on. If we could get past the vectors of resistance that I have described, could we accomplish much more together?

That thinking has brought us to the view that sustainability for papyrological editing and research will come in the first instance from sharing in an organizational and technological infrastructure maintained to serve a much wider range of resources for the ancient world (and perhaps not necessarily limited to antiquity, either). The costs of the infrastructure should be much the same whether there are two or fifty databases using the same data structures, search engine, and data management tools. By the same token, papyrology's continued viability will not depend solely on the size of that one small field.

But the technological infrastructure is only part of the cost of these projects, or ex-projects perhaps. Content creation is the other. This is where the community maintenance features come into play. If the original work of scholars, which they do in the course of their ordinary studies and employment, is captured in our databases rather than having to be retyped or reformatted by paid staff, the other side of the cost base essentially goes away.

All of this is no doubt too simple. We are still trying to analyze what will be needed to create a shared infrastructure, which we have for now code-named Backstop. We believe this will be highly attractive to other projects wondering about their long-term viability. Its costs will be mainly management and the

technological maintenance of the data structures. We imagine that major leaps forward and the inclusion of new resources will continue to require finite project-type funding. But we do not think that finding endowment support for something serving a wide range of resources should be impossible.

At this point, I have wandered some distance from papyrology itself. The reason is primarily my conviction that the future of papyrological projects lies in transcending the limited scale of the discipline and its separateness. I believe that this would actually be justifiable on a purely scholarly basis. As I have remarked, the demarcation of papyrology as a field separate from epigraphy and other neighboring subjects is, although not completely unnatural, by no means necessary. It has, over time, tended to go beyond convenience to something approaching principle, and that is bad for papyrological scholarship. When the default on papyrological searches is covering inscriptions on stone and other materials, papyrological scholarship will improve. For example, editors of papyri will see at once that the name or word they are struggling with turns up in documents from other parts of the ancient world. Intellectual relationships now obscured by disciplinary boundaries will come to light. In this way, an evolution driven in considerable part by the economics of sustainability will turn out to lead to better scholarship, not merely cheaper scholarship.