Perseids: Experimenting with Infrastructure for Creating and Sharing Research Data in the Digital Humanities

Bridget Almas, Tufts University

Overview

The Perseids project, funded by the Andrew W. Mellon Foundation, provides a platform for creating, publishing, and sharing research data, in the form of textual transcriptions, annotations and analyses. An offshoot and collaborator of the Perseus Digital Library (PDL), Perseids is also an experiment in reusing and extending existing infrastructure, tools, and services. This essay discusses infrastructure in the domain of digital humanities (DH). It outlines some general approaches to facilitating data sharing in this domain, and the specific choices we made in developing Perseids to serve that goal. It concludes by identifying lessons we have learned in the process, noting some critical gaps in infrastructure for the digital humanities, and suggesting some implications for the wider community.

Perseids evolved to fill a critical need of the vibrant digital classics community of scholars and students (Bodard and Romanello 2016): infrastructure that supports textual transcription, annotation, and analysis at a large scale in both scholarly and pedagogical contexts. Such infrastructure would give us the ability to work with text-centric publications containing a variety of different data types, and would include:

- stable, persistent identifiers for all publications
- a versioned, collaborative editing environment
- the ability to extend the environment with data type-specific behaviors and tools
- customized review workflows

We wanted not only to support our scholarly workflows, but also to be sure that the outputs would be fully sharable and preservable. Perseids currently serves an active user base, averaging between one and two thousand user sessions per month during the academic year, the majority of which come from five active DH communities including Tufts, the University of Nebraska at Lincoln, the College of Letters and Science of the Sao Paulo State University, the University of Leipzig, the University of Lyon, and the University of Zagreb. Several external projects also connect to Perseids's tools and review workflow via its API.

Infrastructure and Data Sharing

General approaches

What constitutes infrastructure, and how does it facilitate data sharing in the domain of digital humanities (DH), and in the Perseids project in particular? According to Mark Parsons, Secretary General of the Research Data Alliance (RDA), infrastructure can be defined as 'the relationships, interactions and connections between people, technologies, and institutions that help data flow and be useful.'

In the realm of DH, any of the following might be considered infrastructure: original digital collections, linked data providers, general purpose and domain-specific platforms, content management systems (CMSs), virtual research environments (VREs), online tools and services, repositories and service providers, aggregators and portals, APIs and standards. Table 1 provides some specific examples of these in the DH and digital classics (DC) communities, illustrating the diversity and breadth of infrastructure in this community.

Infrastructure type	**Examples in DH and DC**
Original digital collections	PDL, Papyri.info, NINES, Digital Latin Library, Coptic Scriptorium, Roman de La Rose
Linked data providers and gazetteers	Pleiades, PeriodO, Syriaca.org, VIAF, Getty, Trismegistos, DBPedia
General purpose platforms, CMS, VREs, tools and services	Omeka, MediaWiki, Heurist, TextGrid, Voyant, Mirador, CollateX, JUXTA, Neatline
Domain-specific platforms, CMS, VREs, tools and services	Perseids, Recogito, Symogih, PECE
Repositories and service providers	CLARIN, DARIAH, EUDAT, MLA Commons/CORE, HumaNum, Hathi Trust Research Center, California Digital Library
Aggregators and portals	Europeana, Digital Public Library of America, HuNi, EHRI
APIs and standards	IIIF, OA, TEI, OAUTH, Shibboleth/SAML, CTS

Table 1: Examples of infrastructure in digital humanities and digital classics

Enabling data sharing includes ensuring that data objects have persistent, resolvable identifiers, providing descriptive and structural metadata, providing licensing and access information, and using standard data formats and ontologies. The recent W3C recommendation 'Data on the Web

Best Practices' (Loscio, et. al. 2016) cites many strategies such as providing version history, provenance information, and data quality information.

The Perseids strategy

One goal of infrastructure is to connect what already works, adding value and capacity without reinventing solutions. Perseids evolved, in part, out of a prior ambitious, but ultimately unsuccessful, infrastructure effort in the humanities, Project Bamboo (Dombrowski 2014). One of the aims of Project Bamboo was to develop a Service Oriented Architecture (SOA) that could serve a wide variety of use cases and requirements for textual analysis and humanities research. This accorded with the goal of the PDL: to begin to decouple the many services making up the Perseus 4 application, so that they could be recombined and reused to build new applications (Almas 2015). The PDL's contribution to Bamboo included development (and implementation) of APIs for morphological analysis and syntactic annotation. These services, intended to be shared on the Bamboo Services Platform, reused code from two main sources: the PDL's web application and the Alpheios Project's reading environment, and were designed to be easily extended to serve additional languages and use cases.

These services provided essential functionality for textual analysis and annotation, but a critical missing component was a platform for management of the data and scholarly workflow which would allow for full peer and professorial review. This was to have been provided by Bamboo. We looked instead to another domain-specific infrastructure to fill this role, the Son of SUDA Online (SoSOL), which served as the core for the Papyri.info site. SoSOL, a Ruby on Rails application built on top of a Git repository, provides an open-access, version-controlled, multi-author web-based editing environment that supports working with collections of related data objects as publications. It was developed by the Integrating Digital Papyrology project, a multi-institution project aimed at supporting interoperability between five different digital papyrological resources (Baumann 2013). SoSOL is now maintained jointly by the Duke Collaboratory for Classics Computing and the Perseids project.

One thing that prevented Bamboo from succeeding was the assumption that scholars would be willing to give up their domain-specific tools and services for a more general infrastructure to which everyone would contribute (Dombrowski 2015). Humanities use cases are far too diverse for that, and technologies move too fast. Learning from this experience, we decided that Perseids would support a looser coupling of existing tools and services.

Our development approach was based on three principles:

1. tool interoperability
2. flexibility and agility

3. data integrity

Tool interoperability

Decoupling data creation tools from the sources and destinations of the data was a key part of our approach. APIs and standards are critical components of infrastructure, and integration and sharing requires that data be retrievable from and persistable to any source (Hilton 2014).

Perseids offers an API for Create, Read, Update, and Delete update operations for all data types supported by the platform. API clients can authenticate using the OAuth 2.0 protocol. This enables integration with specific tools and services, such as the Arethusa Annotation Framework and the Alpheios Alignment Editor, as well as external projects such as the Syriaca.org Gazetteer.

We also offer a lightweight URL-based API which lets individual scholars and smaller projects, particularly those that don't have time or skills to develop client software, pull their own data in or integrate Perseids with their application. Professors such as Robert Gorman at University of Nebraska Lincoln (Gorman and Gorman, forthcoming) are using this feature to produce templates for new annotations that they publish on their university Learning Management Systems (LMS). They then include links to Perseids in their syllabi that instruct Perseids to pull the templates from the LMS to create a new annotation publication. Other applications such as Digital Athenaeus and Sematia use Perseids's URL API to offer links to Perseids with specific content already identified for transcribing, translating, or annotating.

Perseids also uses external APIs to pull data from other infrastructures. We use the Canonical Text Services URN protocol and API (Smith and Blackwell 2012) to identify and retrieve textual transcription, translation and annotation targets. We have also implemented a workflow for Marie-Claire Beaulieu's Journey of the Hero course which allows students to use the Hypothes.is annotation tool to annotate named entities and social networks of mythological characters from Smith's Dictionary of Greek Names. This workflow uses the Hypothes.is API to pull the annotations into Perseids for review and publication.

The Perseids/EAGLE integration (Liuzzo 2014) uses a combination of both of these pull strategies: links from EAGLE to Perseids identify a resource on the EAGLE site, and trigger a callback to the EAGLE MediaWiki API to pull metadata and data from that resource into new translation publications on Perseids.

We also use external APIs to push data to external repositories. For the EAGLE project integration, Perseids uses the Mediawiki API to publish data to the EAGLE repository once it has passed through a review workflow. Through a new NEH-funded collaboration with the

Syriaca.org project, we have developed a service which allows us to push data to external GitHub repositories at the end of the review workflow. Eventually we'd like to be able to support pushing data to any external API endpoint.

Flexibility and Agility

From the outset, we have taken an agile approach to development of Perseids. While we do not use sprints and formal iterations, we approach planning in short increments, guided by a long-term vision and goals. We also deploy code as soon as it is in a usable state, so that we can get feedback from its users and stakeholders. We do this not only for internal-facing features, but also to prototype new integrations with external services and projects. This flexibility allows us to try many things, keeping those that work and prove to be useful and deprecating those that do not.

To support this approach, we could not commit to a specific set of hardware requirements in advance, as we needed the flexibility to extend and reduce resources used as development proceeded. We therefore chose to budget for cloud-based resources on the Amazon Web Services (AWS) platform rather than using university IT resources. Full ownership and control over our infrastructure allowed us to experiment with features and integrations that otherwise would not have been possible; however, it did have some drawbacks and unexpected costs. These are described in the 'Lessons learned' section below.

Data Interoperability

The final component of our strategy for enabling data reuse was to take steps to ensure data interoperability through the use of stable identifiers and standard formats. Publications produced on Perseids can be thought of as research objects (Bechhofer, et. al. 2013), where the object of the research is a passage or passages of canonically-identifiable text. We use Canonical Text Service URNs to identify these targets and the CapiTainS infrastructure to resolve them to text. These URNs can be considered stable identifiers, but do not quite qualify as persistent identifiers as they are not universally resolvable or guaranteed to be available. We hope to find solutions to this problem, for example by mapping CTS URNs to the handle system (Almas and Schroeder, forthcoming), but in the absence of this piece of infrastructure, the CTS URNs do provide stable, machine actionable identifiers which are technology independent and their use is part of our strategy to make the data produced on our platform sustainable and reproducible.

We also use other types of stable identifiers within our annotations and texts, including the URIs published by the Pleiades Gazetteer. We are working towards ensuring that any data published by the platform has a persistent identifier as well. We are therefore participating in the Research Data Alliance's Research Data Collections working group to develop a multidisciplinary,

collections-based approach to data management that supports persistent identifiers for the collections themselves, and for the items within a collection.

The second part of our strategy for ensuring data interoperability is to use standard data formats and ontologies for our data and to validate all objects against these. The primary data format standards supported on the platform include the TEI Epidoc Schema for textual transcriptions and translations, the Open Annotation protocol for annotations, the ALDT/ALGT schemas for treebank data, the Alpheios Alignment Scheme for translation alignments, and the SNAP ontology for social network annotations.

Lessons Learned

We have learned much about infrastructure building throughout the course of this project. The technical hurdles to interoperability and sharing are usually much less difficult to overcome than those of social issues, funding, and governance. Even where there was a clear interest in interoperability and it was technically possible, we failed sometimes to implement or sustain an integration because doing so wasn't in the funded mandate of the partner project. This was the case for us with the Recogito application from the Pelagios Project. But even where explicit funding support doesn't exist, interoperability can still succeed if one project can fill a key gap in another, and if there are people willing to champion the effort to ensure its success. One example was our integration with the EAGLE project, where Perseids provides a review workflow for EAGLE (Liuzzo 2014). As official funding on both projects winds down, it remains to be seen whether this collaboration can continue without explicit support. This is an area where more formal governance structures, such as those offered by larger research infrastructures such as CLARIN and DARIAH (Lossau 2012) could be useful. The key challenge for the community is to encourage and support ad-hoc collaborations to get initial solutions working, and then move from there to more formal agreements to ensure sustainability.

Laura Mandell talks about the various models being considered for where and how to position DH, and points out that the question of how to support diverse infrastructure needs is still unsolved (Dinsman 2016). A second lesson we have learned from our experience on Perseids is that for development of interoperable infrastructure to succeed and be sustainable, we need better collaborative models for working with our university Information Technology departments and libraries. We knew we needed the flexibility to change our hardware requirements as we developed, and to deploy new code and services quickly to support rapid prototyping. This allows us to develop and try out new solutions more rapidly than we would have been able to if we had to go through university policies and procedures, but it also involved a lot of extra system administration work we had not anticipated, leaving us with a somewhat over-complicated infrastructure at the end of the first phase of the project. Accordingly, in the second phase we

built in funding for a devops consultant, who helped us move to a fully configuration-managed system, so that the Perseids platform can be deployed easily by others and sustained for the long term. This is a critical characteristic for software-related infrastructure - building it must be automatable and reproducible by others. In hindsight, having such consultancy from the outset would have been beneficial; collaboration between developers and the IT staff responsible for deploying and sustaining software is a more viable model than throwing code 'over the wall' at the end of a project (Arundel 2016). As cloud computing becomes increasingly cost-efficient, there is a need for models in which university IT departments can partner with projects to provide this sort of expertise, regardless of whether code is deployed on the university infrastructure or on a cloud platform.

It is very important to us that the research data produced with Perseids be preserved. However, our data models and approach to publications are constantly evolving, making coordination with the university library to preserve this data challenging, as they don't necessarily fit the data models the library is already able to support. As a publicly available and open infrastructure, we also have many users from many institutions across the world, and it is not clear what responsibility Tufts, the university hosting the infrastructure, should have for data created by external users. We mitigate this with Perseids by ensuring that users can always access and download their data, and encouraging them to take responsibility for publishing and preserving it on their own. We continue to explore general models such as the Research Object (Belhajjame, et. al. 2015), or BagIt, which will enable users to export data in a format that is ready to store in a repository. Another question is that of software preservation (Rios 2016). As the Perseids software is under active development, it is difficult to keep the code for digital publications up to date with all the underlying services providing the data. We need to plan better for this preservation, including taking into account the need to represent interdependencies between visualizations and the underlying services and software (Lagos and Vion Dury 2016).

One important requirement is incorporating provenance information in our publications. We have made some progress on this, and one of our motivations for supporting the Shibboleth/SAML protocol for authentication on Perseids was to be able to ensure a chain of authority for university repository systems. However, capturing and recording provenance information reliably across a diverse ecosystem of tools and services is a big job, and we need general-purpose solutions that we can reuse. As articulated by Padilla (2016): "A researcher should be able to understand why certain data were included and excluded, why certain transformations were made, who made those transformations, and at the same time a researcher should have access to the code and tools that were used to effect those transformations. Where gaps in the data are native to the vagaries of data production and capture, as is the case with web archives, these nuances must be effectively communicated." While we have taken significant

steps in this direction, we recognize that there is a great deal more to do, but it is a goal that we, and all data infrastructures, should aim towards.

Conclusion

Infrastructure for interoperability and data sharing in the humanities takes many forms. With Perseids, we have explored an agile approach to infrastructure development, emphasizing reuse of both software and data. This has been successful on many levels. Reuse of existing infrastructure components leads to collaborations which increase chances for sustainability, such as the joint maintenance of the SoSOL application. Low-overhead approaches to cross-project integration also benefit all parties involved. However, transitioning to more formal governance models and increased engagement with host institutions will be essential to longer term success.

References

Bibliography

Almas, Bridget 2015 The Road to Perseus 5 - why we need infrastructure for the digital humanities. Blog post on the Perseus Updates Blog. 18, May 2015. Available at URL http://sites.tufts.edu/perseusupdates/2015/05/18/the-road-to-perseus-5-why-we-need-infrastructure-for-the-digital-humanities/

Almas, Bridget and Schroeder, Caroline T Applying the Canonical Text Services Model to the Coptic SCRIPTORIUM. *Data Science Journal* Special Issue on Data Models (forthcoming).

Arundel, John 2016 Build bridges not walls: devops is about empathy and collaboration. Available at URL http://bitfieldconsulting.com/bridges-not-walls

Baumann, Ryan 2013 The Son of Suda Online. In: Dunn, Stuart and Simon Mahoney (eds) *The Digital Classicist 2013*. Offprint from BICS Supplement-122. London. The Institute of Classical Studies University of London. pp. 91-106.

Bechhofer, Sean; Ainsworth, John; Bhagat, Jitenkumar; Buchan,Iain; Couch, Phillip; Cruickshank, Don; Delderfield,Mark; Dunlop,Ian; Gamble, Matthew; Goble, Carole; Michaelides,Danius; Missier,Paolo; Owen,Stuart; Newman, David; De Roure, David; Sufi, Shoaib 2013 Why Linked Data is Not Enough for Scientists. *Future Generation Computer Systems*, 29(2):599-611. DOI:http://dx.doi.org/10.1016/j.future.2011.08.004

Belhajjame, Khalid; Zhao, Jun; Garijo,Daniel; Gamble, Matthew; Hettne,Kristina; Palma, Raul; Mina,Eleni; Corcho,Oscar; Gómez-Pérez,José Manuel; Bechhofer,Sean; Klyne,Graham; Goble, Carole 2015 Using a suite of ontologies for preserving workflow-centric research objects. *Journal of Web Semantics*, 32(05.2015):16-42. DOI:http://dx.doi.org/10.1016/j.websem.2015.01.003

Bodard, Gabriel, and Matteo Romanello (eds) 2016 *Digital Classics Outside the Echo-Chamber*. London. Ubiquity Press.

Dinsman, M 2016 The Digital in the Humanities: An Interview with Laura Mandell - Los Angeles Review of Books. April 24, 2016. Available at URL https://lareviewofbooks.org/article/digital-humanities-interview-laura-mandell/

Dombrowski, Q. 2014 What Ever Happened to Project Bamboo? *Literary and Linguistic Computing*, 29(3): 326–339.

Gorman, Robert and Gorman, Vanessa. Approaching questions of text reuse in Ancient Greek using computational syntactic stylometry. *Open Linguistics* Topical Issue on Treebanking and Ancient Languages (forthcoming).

Hilton, James L. 2014 Enter Unizin. *EDUCAUSE Review*,49(5).

Lagos, Nikolaos and Vion-Dury, Jean-Yves 2016 Digital Preservation Based on Contextualized Dependencies. Doc Eng. September 13-16. Available at URL http://www.xrce.xerox.com/content/download/93294/1307736/file/2016-031.pdf

Liuzzo, Pietro Maria 2014 Translating Greek and Latin Inscriptions. Presentation for Greek and Latin in an Age of Open Data, University of Leipzig. December 2014. Available at URL http://www.dh.uni-leipzig.de/wo/wp-content/uploads/2014/11/Pietro-Liuzzo-Translations-of-Greek-and-Latin.pdf

Loscio, Bernadette Farias; Burle, Caroline; Calegari, Newton. W3C. 2016 Data on the Web Best Practices. W3C Candidate Recommendation. 30 August 2016. Available at URL https://www.w3.org/TR/2016/CR-dwbp-20160830/

Lossau, N. 2012 An Overview of Research Infrastructures in Europe - and Recommendations to LIBER. *LIBER Quarterly,*21(3-4):313–329. DOI: http://dx.doi.org/10.18352/lq.8028

Padilla, Thomas 2016 Humanities Data in the Library: Integrity, Form, Access. *D-Lib Magazine*, 22(3/4)

Parsons, Mark 2015 e-Infrastructures & RDA for data intensive science. 22 September 2015. Available at URL https://rd-alliance.org/sites/default/files/attachment/Infrastructures,%20relationship,%20trust%20and%20RDA_MarkParsons.pdf

Rios, Fernando 2016. The Pathways of Research Software Preservation: An Educational and Planning Resource for Service Development. *D-Lib Magazine*, 22 (7/8). DOI: http://dx.doi.org/10.1045/july2016-rios

Smith, Neel and Blackwell, Christopher W 2012 Four URLs, limitless apps: Separation of concerns in the Homer Multitext architecture. In *Donum natalicium digitaliter confectum Gregorio Nagy septuagenario a discipulis collegis familiaribus oblatum: A Virtual Birthday Gift Presented to Gregory Nagy on Turning Seventy by His Students, Colleagues, and Friends*. Boston. The Center of Hellenic Studies of Harvard University

Vierros, Marja and Henriksson, Erik 2016 Preprocessing Greek Papyri for Linguistic Annotation. Hal-01279493. Preprint. Available at URL https://hal.archives-ouvertes.fr/hal-01279493

Projects, Websites, Software

The Ancient Greek and Latin Dependency Treebank by PerseusDL [WWW Document], n.d. URL https://perseusdl.github.io/treebank_data/ (accessed 9.29.16).

Alpheios [WWW Document], n.d. URL http://alpheios.net/ (accessed 9.29.16).

alpheios-project/arethusa: Arethusa: Annotation Environment [WWW Document], n.d. URL https://github.com/alpheios-project/arethusa (accessed 9.29.16).

alpheios-project/alignment-editor: Alpheios Alignment Editor [WWW Document], n.d. URL https://github.com/alpheios-project/alignment-editor (accessed 9.29.16).

draft-kunze-bagit-08 - The BagIt File Packaging Format (V0.97) [WWW Document], n.d. URL https://tools.ietf.org/html/draft-kunze-bagit-08 (accessed 9.29.16).

CapiTainS [WWW Document], n.d. URL http://capitains.github.io/ (accessed 9.29.16).

Digital Athenaeus - A digital edition of the Deipnosophists of Athenaeus of Naucratis [WWW Document], n.d. URL http://digitalathenaeus.org/ (accessed 9.29.16).

PonteIneptique/flask-github-proxy: Github proxy to push resource to github [WWW Document], n.d. URL https://github.com/PonteIneptique/flask-github-proxy (accessed 9.29.16).

Journey of the Hero [WWW Document], n.d. URL http://perseids.org/sites/joth/#index (accessed 9.29.16).

Hypothesis [WWW Document], n.d. URL https://hypothes.is/ (accessed 9.29.16).

Morphological Analysis Service Contract Description - v1.1.1 - Project Bamboo - UCB Confluence [WWW Document], n.d. URL https://wikihub.berkeley.edu/display/pbamboo/Morphological+Analysis+Service+Contract+Description+-+v1.1.1 (accessed 9.29.16).

pleiades.stoa.org [WWW Document], n.d. URL https://pleiades.stoa.org/ (accessed 9.29.16).

Research Data Collections WG [WWW Document], n.d. URL https://rd-alliance.org/groups/pid-collections-wg.html (accessed 9.29.16).

RECOGITO [WWW Document], n.d. URL http://pelagios.org/recogito (accessed 9.29.16).

Sematia [WWW Document], n.d. URL http://sematia.hum.helsinki.fi (accessed 9.29.16).

Syntactic Annotation Service Contract Description - v1.1.1 - Project Bamboo - UCB Confluence [WWW Document], n.d. URL https://wikihub.berkeley.edu/display/pbamboo/Syntactic+Annotation+Service+Contract+Description+-+v1.1.1 (accessed 9.29.16).

Ontology | Standards for Networking Ancient Prosopographies [WWW Document], n.d. URL https://snapdrgn.net/ontology (accessed 9.29.16).

Syriaca.org: The Syriac Reference Portal [WWW Document], n.d. URL http://syriaca.org/ (accessed 9.29.16).

EpiDoc Guidelines 8.22 [WWW Document], n.d. URL http://www.stoa.org/epidoc/gl/latest/ (accessed 9.29.16).