Unit testing, integration and deployment : dealing with diversity, interoperability and sustainability of digital corpora

Thibault Clérice, Bridget Almas, Marie-Claire Beaulieu, Stella Dee

Unit testing, integration and deployment : dealing with diversity, interoperability and sustainability of digital corpora

Thibault Clérice[1], Bridget Almas[2], Marie-Claire Beaulieu[2], and Stella Dee[2]

[1]Humboldt Chair of Digital Humanities, Universität Leipzig
[2]Tufts University

October 2015

The Open Philology Project (OPP) at Leipzig and its US affiliate, the Perseus Digital Library at Tufts (PDL), has years of experience developing extensive infrastructures for managing textual data for historical languages. With around 100 million words available on PDL, and millions more words coming through OPP, in a context of opening contributions from wide ranging communities of users, dealing with ingestion of new texts is a matter of security, flexibility and efficiency.

Over the last few years, PDL and OPP have been moving forward in implementing the Canonical Text Service URN norm and the Epidoc subset guidelines to allow for better interoperability and citability of its texts. We are now working towards supporting a scalable workflow centered on continuous curation of these texts, from both within and outside the PDL/OPP ecosystem. Key requirements for such a workflow are ease of maintenance and speedy deployment of texts for use by a wide variety of analytical services and user interfaces.

Drawing on software engineering best practices, we are building an architecture meant for continuous integrations[1]: analogous to the way Travis[2] integrates with Github[3], we are developing a customizable service that test individual files upon each contribution made to our public git repositories. The services can be configured to test and report status on a variety of checkpoints from schema compliance to CTS ready markup.

With a strong continuous integration service, we should be able to deal not only with a wide range of genres and languages, but also with a diversity of contributors. We can delegate the tedious tasks of checking markup to the

[1]`https://en.wikipedia.org/wiki/Continuous_integration`
[2]`https://travisci.org`
[3]`https://github.com`

machine, leaving curators free to focus on the scholarship. We also expect that automating checks on the integrity and the adaptability of textual objects for specific frameworks can reduce the error rate and allow for shorter feedback loops to contributors and users of our corpora.