AHRC ICT Methods Network Workshop

OPEN SOURCE CRITICAL EDITIONS

centre for computing in the humanities, king's college london, 22 september 2006

Report by Gabriel Bodard and Juan Garcés

Introduction

The Open Source Critical Editions Workshop was held on 22 September 2006 at the Centre for Computing in the Humanities, King's College London under the auspices of the AHRC ICT Methods Network; the meeting was also supported in part by the Perseus Project and the Digital Classicist.[1] This workshop was set up with the aim of exploring the possibilities, requirements for, and repercussions of a new generation of digital critical editions of Greek and Latin texts with underlying code made available under an open license such as Creative Commons or GPL.[2] This topic broached many technological, legal, and administrative issues, and the participants were selected for their interest and/or expertise in these areas, and asked to consider how such editions advance classical philology as a whole, both in terms of the internal value to the subject itself, and in terms of outreach, interdisciplinarity, and the value of philology to the wider world outside the academy.

Technological questions discussed at this event included: the status of open critical editions within a repository or distributed collection of texts; the need for and requirements of a registry to bind together and provide referencing mechanisms for such texts (the Canonical Texts Services protocols being an obvious candidate for such a function[3]); the authoritative status of this class of edition, whether edited by a single scholar or collaboratively; the role of e-Science and Grid applications in the creation and delivery of editions.

Legal issues largely revolved around the question of copyright and licensing: what status should the data behind digital critical editions have? It was an assumption of this group that source texts should be both Open Source and Public Domain, but the specifics remain to be discussed. Attribution of scholarship is clearly desirable, but the automatic granting of permission to modify and build upon scholarly work is also essential. There were also questions regarding the classical texts upon which such editions are based: what is the copyright status of a recently-published critical edition of a text or manuscript, that the editor of a new edition needs to incorporate?

Administrative questions posed by open critical editions included: issues of workflow and collaboration (in which Ross Scaife of the Stoa Consortium has considerable experience, for example through the Suda Online and other projects[4]); protocols for publication and reuse of source data: a genealogy of reuse and citation could be generated using version control tools, or a system of passive link-back generating an automatic citation index through a web search engine. Issues of peer review and both pre- and post-publication validation of scholarship were also discussed.

[1] The Methods Network description of this workshop is at <http://www.methodsnetwork.ac.uk/activities/act9.html>; the project description, including the text of positioning papers and responses, is available from the Digital Classicist Wiki at <http://wiki.digitalclassicist.org/Open_Source_Critical_Editions>.

[2] Creative Commons licences defined at <http://www.creativecommons.org/licenses/>; the GNU General Public License at <http://www.gnu.org/copyleft/gpl.html>.

[3] The Canonical Text Services specification is at <http://chs75.harvard.edu/projects/diginc/techpub/cts>.

[4] Suda On-line is at <http://www.stoa.org/sol>; other Stoa projects have built on the editorial work of this groundbreaking initiative.

Report

Delegates from Germany, the United States, as well as around Britain delivered eight positioning papers on pertinent topics organized in three sessions. All of the position papers were made available in advance through the Digital Classicist Wiki and read by all delegates; accordingly the presentations on the day were relatively brief summaries, leaving time for considered responses in each case and then thorough and in-depth discussion from the group as a whole.

Critical editions

Charlotte Roueché addressed the value added by electronically-published texts, particularly, but not exclusively, of inscriptions and other documentary material. The first part of her paper focused on the issue of increased accessibility for a non-academic and non-English-speaking audience and its impact on wider popularity and international collaboration. The second part dealt with the way such editions might open up the scholarly decision-making process, a transparency that might make some scholars nervous, but should help to elucidate the intellectual weight and distribution of scholarly tasks.

Bodard's paper defined markup as the addition of information in relation to the construction of the text, references to entities like persons and places, as well as cross-references to other texts. What makes an edition truly ‘critical’ is then ‘the explicit recording of editorial decisions, critical apparatus, and notes’. In view of the aim to build a distributed collection of compatible texts that can be handled, displayed, searched, and processed by a single corpus or database, he recommended the adherence to a consistent markup scheme, such as TEI, or, alternatively, machine-readable documentation of the schema used, for example in ODD files.[5] Notis Toufexis responded by raising the need to further define the ‘critical’ in critical editions, a task that is further complicated by the shift to the digital medium. He supported an approach that allows the full reconstruction of the individual witnesses and the machine-readable documentation of local uses derived from a standard like TEI.

Technologies

Stuart Dunn offered an introduction to e-Science as ‘the development and deployment of a networked infrastructure and culture through which resources can be shared in a secure environment’. Collaborative critical editing could benefit particularly from VREs and virtual organization facilitated by grid networks.

Gregory Crane presented a strategy for creating primary source corpora that are permanent, openly accessible, multi-versioned, and funded by academic libraries. They would profit from large-scale digitization projects like Google Library, the Open Content Alliance, and Europe's i2010[6] to create a digital library by adding semantic markup and collation tools, advancing in cycles of critical editorial work from the raw image files.

Ross Scaife discussed the need for technology to support Humanists involved in collaborative editing projects in three areas: (1) the ability to build editions encompassing texts, images, and annotations; (2) the management of access and version control; and (3) accessibility for a geographically-distributed group of scholars, illustrating each of these areas with concrete examples.

[5] The Text Encoding Initiative's ODD language is documented at <http://www.tei-c.org.uk/ODD/Manual/>.

[6] Google Print: <http://books.google.com/> Open Content Alliance: <http://www.opencontentalliance.org/>; the EU's i2010 scheme <http://ec.europa.eu/information_society/eeurope/i2010/index_en.htm>.

Protocols

Sayeed Choudhury talked about the new challenges and opportunities libraries face in relation to the development of infrastructure in support of digital collections. He advocated in principle the development of repositories that are technologically open systems, but might be constrained by local policies and legal frameworks. He also argued that for the scholarly community only completely open standards and open access to digital texts would create a digital environment that fosters collaborative critical editions. This would have consequences for organization roles, business models, and reward structures.

Neel Smith proposed a method of constructing an architecture for a distributed digital library of critical editions, which consists of moving from a conceptual model of digital publications to the analysis of its distinctive features, and subsequently translates the resulting functional requirements into technical ones. In dealing with these tasks he strongly recommended the reliance on well-established technical protocols—HTTP as the transport mechanism, XML for service requests and replies—in order to provide indexing services and basic citational functionality on different levels of detail. His proposed architecture has been implemented in the Canonical Text Services project. Garcés pointed out the need to further develop protocols for referring to work-in-progress editions across the different evolving stages and the careful distinction between the archival role of digital repositories and the different implementations of the texts within them.

Daniel Deckers focused on the authority of collaboratively-edited digital critical editions and the role of peer review in quality control. On the subject of transparently-encoded interpretative contributions, he suggested a dual strategy that combines the mapping of widely-recognized expertise onto a management system with vetting capabilities. He further suggested freezing work-in-progress editions at milestones decided by editorial committees, in order to produce authoritative and quotable versions.

There was an additional session for general dialogue and strategy discussion at the end of the day, in which various options for funded projects were raised. It was agreed that any such project should be fronted by the learned societies for Greek and Roman studies, and further discussion was set to continue at venues to be determined.

Dichotomies

During the course of the OSCE workshop and the discussions that preceded and followed it, it became clear that there are several areas in the field of online text publication that are sources of debate. These areas contain interesting dichotomies that will often be found to cohabit naturally or to lead to useful compromizes. There is of course room for a multiplicity of positions and opinions within the area, and any technologies and protocols we develop or adopt should be able to cope with several approaches. It is important that we discuss these issues not only to better understand the needs and approaches of those around us, but to further examine our own attitudes and assumptions. We shall summarize three of the main contrasting pairs here.

Scale and depth

The first major issue to emerge was the conflict between the vast benefits to be found in large-scale collections of text such as the Google Print library, the *Thesaurus Linguae Graecae* (although neither of these are open in the sense discussed in this paper), or some of the proposals the Perseus Project, on the one hand, and the significant possibilities of close markup of edited texts with apparatus, editorial commentary and so forth made possible with TEI XML, for example, on the other.

Crane argued that the benefits of scale were incalculably great, and indeed that many of the tasks that would take a human editor a lifetime could be achieved by machine translation, data mining, automated learning based on huge amounts of plain text in a Grid environment, and other scaleable operations. Bodard observed, against this, that the detailed work enabled by EpiDoc and other text-editing standards was potentially leading to the creation of a significant amount (although tiny by the standards of the Million Book Project) of highly enriched textual material. It would be unfortunate, even wasteful, to neglect this body of available, open source text in favour of pure machine brute force. Indeed, he went on to argue, without some capacity for dealing with apparatus criticus and commentary, an electronic text could not be said to be a 'critical edition' at all.

It was agreed by all parties, including those with the most entrenched views expressed here, that any protocol or set of standards adopted or developed for the archiving and dissemination of open source critical editions should be flexible enough to incorporate both deeply marked-up critical texts and large numbers of machine-accessible text pages. The detailed mark-up of the human-edited texts may be

ignored (but not, of course, flattened) in the course of machine searches, data mining, statistical and pattern recognition processes that depend on huge scale and uniformity for the power and applicability of their results. Even more importantly, however, the levels of detail and unambiguous scholarly information added by human editors in a body of such canonical critical editions would serve as a template to vastly enrich the process of machine-analysing large bodies of consistently-encoded text. The two faces of electronic text publishing, the vast scale made possible by technology and the detailed work that is traditionally the philologist's art, are both invaluable to one another and to the endeavour of open source critical editions as a whole.

Texts and Manuscripts

The second conflict that became evident in the course of discussing electronic texts was the question of publishing critical texts in the usual form of edited, combined or eclectic editions (for example the *Iliad* of Homer), usually as the reconstruction of the *Urtext*, as against making available edited versions of individual manuscripts or source texts: papyri, inscriptions, mediaeval codices, and the like. Images and transcriptions of the individual manuscript texts, argued Garcés and Fuchs, for example, are important not only because they are under threat—they exist in limited, fragile, physical form, and many of them have never been digitized or otherwise reproduced—but because they can then be used to contribute to machine-collated critical editions.[7]

Two concepts of the definition of 'text' are at work in these two approaches and neither should have to exclude the other. The eclectic critical edition operates at a more abstract level and understands text as a sequence of linguistic signs, while the critical edition of a manuscript operates at a more specific level and has to consider the physical condition of a text, as well as the contemporary oral and scriptural linguistic conventions and the idiosyncrasy of the scribe(s). Arguably the abstract text has to be derived critically from the specific instance(s) of that text. Building up critical editions of texts from the collation of specific critical editions of manuscript instances has the advantage of linking decisions made on the latter to the reconstruction on the former, as well as allowing for eclectic editions that, rather than asking for the original version of the text, reconstruct textual versions in relation to specific periods or regions. The apparatus of such an edition can be seen as a condensation of the manuscript evidence pertinent to the specific aims of an eclectic edition, which – and this would be the advantage of this approach – can be linked dynamically to the far richer evidence of each instance, as Toufexis argued (above).

Development and Reapplication

[7] The Online Critical Pseudepigrapha <http://ocp.acadiau.ca/> is an example of the power of such collated editions.

The third dichotomy is the question of how much technology and protocol classical scholars ought to develop and define for their own, unique needs, and how much the models created by digital humanities and even the sciences as a whole ought to be borrowed and applied to their ends. Crane has gone on record with the claim that classicists (and perhaps even humanities scholars) have no need to invent tools, technologies, or working methods, because everything we are trying to do is achievable with the tools that better-funded and better-organized sciences have created.[8] This is perhaps superficially at odds with thinking behind the Canonical (originally 'Classical') Text Services that classical scholars such as Neel Smith have been developing, although CTS protocols may of course be used in many other humanities disciplines.

There are in any case two possible counterpoints to this position. The first of these being that perhaps in an ideal world, all possible tools that are useful or necessary for scholarship would have been (or can easily be) created by somebody with the resources to do so; they are after all the same requirements of all scholars and disciplines in the academy. But there are still many gaps in the toolbox available to academics today, in the humanities and sciences alike. This leads to the second argument: that just as Classics has subject matter and methodologies that have been developed over its long history as a discipline, there will be needs and vacancies that differ from those of other disciplines, or at least are not currently being met by them.

Clearly the answer is not for classicists to spend a lot of time and even more scarce money on creating all new tools and protocols for every need that arises. But likewise, we should not be shoe-horning our problems and academic questions into the models that social scientists and biologists have created technological solutions for. We should reapply existing tools as far as they exist, and adapt rather than recreate them when they only cover 90% of our needs (assuming that the tools are Open Source, for example). And when we do need to develop new technologies or protocols, they should be built within the framework of existing standards, toolsets, and technologies: as, for example, the EpiDoc Guidelines for publication and interchange of Greek and Latin epigraphic documents in XML built upon the TEI rather than creating their own schema from scratch.[9] Classicists should take advantage of models, tools, and systems developed in other fields, without underestimating the extent to which the approaches developed or improved within their own framework might be fed back successfully to the lively exchange of cross-disciplinary ideas and experiences.

Conclusions

The Open Source Critical Editions workgroup and its first workshop bear witness to the perceived need to provide twenty-first century Classicists—and Humanists in general—with primary resources that take advantage of contemporary technologies to offer maximum accessibility and quality. The workshop foregrounded only some of the most pertinent technological, legal, and administrative issues still to be addressed. Of the saccount by any digitization protocol or strategy are: (1) the integration of the human-tagged editions with deep markup, and a large scale collection of machine-digitized texts for data-mining and the like; (2) the complex and potentially powerful relationship between traditional eclectic editions on the one hand, and critical texts and images of individual manuscripts and witnesses on the other.

It is therefore clear that, while the format of and contributions to the workshop were a success, there is still need to further develop and test some of the discussions initiated at the workshop, before undertaking a large, collaborative, and international project.

[8] e.g. Gregory Crane, 'Classics and the Computer: An End of the History' in edd. Schreibman, Siemens, Unsworth, *A Companion to Digital Humanities*, 2004.

[9] EpiDoc: <http://epidoc.sourceforge.net/>; Text Encoding Initiative: <http://www.tei-c.org/>