Digital Libraries of Scholarly Editions

Buchanan, George

g.r.buchanan@gmail.com
School of Informatics, City University, London

Bohata, Kirsti

K. Bohata@swansea.ac.uk
Centre for Research into the English Literature and Language of Wales (CREW), Swansea University

Digital libraries are a key technology for hosting large-scale collections of electronic literature. Since the first digital library (DL) systems in the early 1990s, the sophistication of DL software has continually developed. Today, systems such as DSpace and Greenstone are in use by institutions both large and small, providing thousands of collections of online material. However, there are limitations even to "state-of-the-art" DLs when considering digital humanities.

Contemporaneously with the growth of DL technology, digital scholarly editions of significant texts and archival material have emerged. In contrast to digital libraries, where there are a number of readily available generic software systems, critical editions are largely reliant on bespoke systems, which emphasise internal connections between multiple versions. Whilst useful editions are online, and are increasingly used in scholarly endeavour, digital scholarly editions suffer from "siloing": each work becoming an island in the ocean of the web.

Like 'physical' libraries, digital libraries provide consistent support for discovering, reading and conserving documents in large collections. For scholarly editions, these features present a potential solution to "siloing". Without trusted digital repositories, preservation and maintenance are endemic problems, and providing consistent experiences and unified workspaces across many sites (i.e. individual texts) is proving highly challenging. However, current DL systems lack critical features: they have too simple a model of documents, and lack scholarly apparatus.

Digital library systems can readily contain electronic forms of "traditional" critical editions in static forms where each work is a separate and indivisible document. However, search facilities cannot then reliably distinguish between commentary and the primary content. Using XML formats such as TEI can permit this distinction to be made, but only with extensive configuration. Furthermore, the reading experience in such a DL is likely to fall far below scholars' requirements of digital editions.

European initiatives, such as DARIAH, focus on facilitating access to existing scholarly editions, with longer-term aims of fostering standards and interoperability of data. This approach presumes that each existing site (and hence, typically, edition) remains autonomous, and remains a discrete entity, which is then aggregated through a centralised service. It also admits the absence of standardised, highly functional storage and publication systems. Furthermore, this approach has been attempted in "federated" DLs, with only limited success. In federated DLs, unless every member uses the same software configured in the same manner, the appearance of each library differs and – worse – preservation remains in the hands of individual sites, and cross-site services (e.g. search) can only operate at a very rudimentary level.

There have been projects to develop generic scholarly edition software, but success has been limited. Shillingsburg [Buchanan 2006] highlights a number of such systems up to the mid-2000s. Few of these initiatives engaged with computer science, and the software systems have proved hard to maintain.

Digital library systems provide a potential route for providing collections of digital scholarly editions. However, they are not yet an answer. When a digital edition supports discussion between scholars, grounded on and linked to the text, the standard DL infrastructure requires extensive modification. Data structures are required to capture and store scholarly discourse, relate each item of discourse in detail to part of a complex document structure, and provide this through a seamless and consistent user interface. Multiple structures and complex document relationships fit uneasily within current DL software [Buchanan *et al.*

2007,Rimmer *et al.* 2008]. For instance, most DL software requires or assumes that any collection of documents is homogenous in terms of the interior structure of each document. This simply cannot be true of a collection including – say – diaries, journals, letters and novels. We need software that provides DL collection support with the ability to provide for complex document structure.

1. Current Work

The goal of our research is to develop software that transcends the current limitations of DL systems in supporting digital scholarly editions for the humanities. Our intention is that in turn organisations and publishers who seek to provide series of critical material can build upon software that is scalable, systematically engineered and sustainable. This software will also support the necessary complexity of critical editions and possess a rich apparatus to support contemporary digital practices, not simply digitised forms of practice from the print era. Whilst no single system is likely to provide all the requirements of all possible circumstances, our aim is to create software that can provide the technical core of any collection of critical editions, with a minimum of effort. Adapting the system for a specific need may require extensive work, but only for more unusual circumstances. This would bring us to a point comparable to the support that current DLs give for simpler texts and scholarly practices. For users of scholarly editions – i.e. the research community – the presence of a common infrastructure and the increased ease of working across sites will very likely increase research activity across multiple 'editions'.

2. Context and Motivation

This project has identified Wales as presenting an interesting case study. It is a distinct cultural entity with an abundance of valuable written cultural material and a sizable scholarly community researching the cultural life and output of the nation. Reflecting the bilingual linguistic identity of the nation, there are extensive archives and printed matter in national, university, local government and private hands in both Welsh and English (as well as other languages). Wales suffers from a poor physical infrastructure, and this has

motivated the provision of digital access to cultural material, from the early days of the National Library of Wales's digitisation projects (e.g. 'Campaign' 1999) to the present.

While the National Library of Wales has done outstanding pioneering work in digitisation of its collections and remains an asset in our selection of Wales as a 'case study', their remit does not extend to the interpretation of their collections. Despite considerable demand from scholars in Wales and beyond for digital critical editions of Welsh material (in both languages) no one project has access to the technical expertise to create software that embodies the requirements of the scholarly community. The motivation of our project is to build a common infrastructure that both enables each project to produce high-quality scholarly work, and provides for consistent access and preservation of that work.

3. Understanding User Needs

To undertake this work requires not only technical expertise, but also a systematic study of the requirements of scholarly practice in the digital age. To date, we have reviewed the existing literature, and gained an initial set of requirements from a retrospective analysis of data from the recent User Centred Interactive Search (UCIS) project at University College London [Rimmer *et al.* 2008].

The UCIS project revealed that many technical difficulties emerged when configuring DL systems, even with relatively simple digital humanities material. Humanists do not necessarily search for material that directly corresponds to the "book" or "document" level of a particular library. Items may be sought that constitute part of a single document (e.g. a poem in a collection of poetry), and conversely larger works may be realised in several separate "documents". Search and browse facilities typically work only at one level, typically consonant with either a book or article. However, collections are frequently heterogeneous and multi-layered. In the case of critical editions, the complexity of document structure and users' tasks is even greater.

A second problem is that humanists often require different variants of one work. Though library infrastructures can relate these together,

using standard features alone is insufficient [Shillingsburg 2006]. Even the more developed features that of a few DL systems are simplistic when compared to the complex relationship between different renditions and editions of a work that critical scholarship requires. Current methods relate entire separate items together – e.g. a chapter to a chapter – but scholarly criticism and annotation do not neatly conform to the clean structural boundaries favoured in computer or library science.

Thirdly, whilst some specific digital library installations do permit individual works to be linked to their author, or even specific words in a text to related material, this is not a standard part of DL software, and in contrast to the advanced facilities available in the best hypertext systems, the current technologies are primitive [Goose *et al.* 2000].

These shortcomings represent only a few of the problems already identified, and while we have developed partial solutions to parts of these, our technologies are not yet comprehensive, and other challenges have yet to be answered at all.

4. Summary

This presentation will articulate the shortcomings and problems raised when collections of critical materials are hosted through current digital library systems, and the contrasting "siloing" problems faced in the field of digital critical editions. We will demonstrate the requirements that will have to be matched to provide a single software system that delivers the needs of critical editions whilst also providing methods to develop collections of critical works. We illustrate how some of these requirements can be met, and prioritise and elucidate the remaining challenges in creating a unified system.

References

Shillingsburg, Peter (2006). *From Gutenburg to Google*. Cambridge: Cambridge University Press.

Buchanan, George (2006). 'FRBR: Enriching and Integrating Digital Libraries'. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. Chapel Hill, NC, USA, June 11-15, 2006, pp. 260-269.

Rimmer, Jon, Warwick, Claire, Blandford, Ann, Gow, Jeremy, Buchanan, George (2008). 'An examination of the physical and the digital qualities of humanities research'. *Information Processing and Management*. Elsevier **3**: 1374-1392.

Buchanan, George, Gow, Jeremy, Blandford, Ann, Rimmer, Jon, Warwick, Claire (2007). 'Representing aggregate works in the digital library'. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. Vancouver, BC, Canada, June 17-23, 2007, pp. 247-256.

Goose, Stuart, Hall, Wendy, Reich, Siegfried (2000). 'Microcosm TNG: A Framework for Distributed Open Hypermedia'. *JIEEE MultiMedia.* **3**: 52-60.