White Paper Report

Report ID: 103312

Application Number: PX5002609

Project Director: David Kohn (dkohn@amnh.org)

Institution: American Museum of Natural History

Reporting Period: 8/1/2009-1/31/2011

Report Due: 7/31/2011

Date Submitted: 8/2/2011

Digitizing Darwin's Library
Phase 1

National Endowment for the Humanities

Transatlantic Digitization Collaboration Grant

Award ID Number PX-50026-09
08/01/2009 – 01/31/2011

White Paper

2 August 2011

White Paper

Charles Darwin's Library

Surrogates & Originals

The working and simplifying premises of this project to digitise Charles Darwin's Library were (1) that the primary objective should be to digitise the annotated half of the 1480 books Darwin owned and (2) to use Darwin's original copies at Cambridge University Library exclusively to digitise heavily annotated works (cover to cover), while using surrogates for the lightly annotated and indeed for the unannotated volumes. These surrogates are bibliographically identical works digitised by the Biodiversity Heritage Library (BHL) and the Internet Archive (IA).

- Challenge of identifying good surrogates. The project began with an exercise in due diligence performed at the Natural History Museum in London (NHM), where that library's holdings were checked against the project's bibliography, which was in turn based on Di Gregorio and Gills' *Darwin's Marginalia* 1992. However, that exercise could only go so far because, obviously, the books in London could not be brought to Cambridge to be physically compared with those in the Darwin Library in Cambridge. The same was of course true for apparent surrogates scanned by other BHL and IA libraries. We were caught in something of a chicken and egg verification problem. In the end this was readily resolved by verifying the digitised NHM, BHL, and IA potential surrogates against the physical Darwin Library volumes. A number of surrogates had to be rejected, but most proved appropriate surrogates. The lesson learned was that we ought to have scheduled this post-scanning effort into our work plan.

Challenge of using surrogates (Frankenbooks). During the original pre-award planning, it readily became apparent that a JISC/NEH award would not cover the cost of digitising all Darwin's books from the original. At that point two apparently incompatible approaches were debated: (1) cover-to-cover digitization of some originals and cover-to-cover digitisation of some surrogates versus (2) digitisation of all individual original pages of all books. Approach (2) might have permitted us to 'cover' all original Darwin pages bearing his writing or marks. But it was rejected on financial and scholarly grounds: cherry picking individual pages costs much more per page than cover-to-cover digitisation and the result would still leave scholars having to go to a variety of libraries to get to relevant editions. So we opted for Approach (1). However, after the grant was already awarded, we were still tangling with this problem and by the time of our first all party meetings in Cambridge and London, we thought that we had found a way to resolve this internal debate and thus to improve on the plan as submitted. The solution would be a mixed strategy, namely: to scan the individual original Darwin pages, but only for lightly annotated works and to combine these 'loose pages' with surrogates. The heavily annotated works would be scanned from Cambridge originals cover to cover, as originally planned. But, when we approached our BHL partners about adopting this mixed strategy, we learned that such hybrids of two physical bibliographic entities produces an bibliographically unacceptable 'Frankenbook'—shades of Mary Shelly and

Gene Wilder in the digital age. So we retreated to our original plan. Lessons learned: 'planning' may have to be a continuous, flexible process. But also, some 'good' ideas don't work.

Jumping Technical Hurdles (Hoops)

- Another example of the need for flexibility. In our original discussions, we considered using the JPEG2000 format for digital image files. Then it was agreed that Cambridge would follow its standard practice of using very high resolution tiffs. However, we were well into scanning when it became clear that the ingestion technology used by Internet Archive—through which all BHL digitisation flows—couldn't handle the large files being generated by Cambridge. The solution was to convert the files to JPEG2000, which provided excellent resolution but smaller file size.

- Developing the interface. BHL endorsed the project because of its intrinsic merit and the excellent fit of the Darwin books within a library devoted to natural history. But it also expressed the hope that in developing the Darwin's Library interface, it could learn lessons directly applicable to other situations where BHL works are annotated or where specimen labels are annotated with bibliographic references. But Darwin's level of annotation proved far more difficult to execute than anticipated. When completed it was also far too complex than to lend itself, at least immediately, to any direct generalization. The particular requirements of the project proved very challenging. But perhaps more fundamental, it took time for the two core groups—those familiar with editing Darwin online and those familiar with the BHL's structure to truly communicate. In the end BHL did succeed in building a superb user interface. Once that was done their mastery of the process of book ingestion and mounting, as well as their experience in developing search tools, came to the fore. Similarly, the Darwin Manuscripts Project and Cambridge University Library teams somewhat underestimated the difficulties the developers would encounter and thus had to scramble at the end to complete the process of checking and correcting the results once we had a functioning interface. One important lesson learned: beware of technological optimism. More generally, the 18 month time frame of the JISC/NEH grant is probably too tight by six months.

Human Factors

- Working across institutions and time zones. At anyone time up to five institutions in three time zones needed to be cooperating and in regular communication. Initially, we were somewhat adrift until NHM and BHL showed us the power of keeping everyone informed simultaneously through group email messages and regularly scheduled conference calls. Once this got going, we often reverted to smaller communication groups. Lesson learned: democracy takes work but pays off.

- Different understandings of open access. Although several members of the team are well versed in licensing procedures and have a strong commitment to the 'open access' concept. Some ways into the project we realized that a mutually agreeable licensing agreement was going to be difficult to achieve. The problem was that there are multiple understanding of 'open access', which in fact proved to be in some

conflict in our case. Ultimately this matter was resolved. Lesson learned: we are left wondering whether the lesson is ‘clarify the licensing assumptions up front’ or ‘wait until the project is so far advanced that all parties are vested enough to make necessary compromises’.

- Finally, perhaps the most important lesson to be learned from such a short term, yet multiparty collaboration is that your project must have the commitment of absolutely qualified partners. What drove this project to completion was everyone’s belief that here was an opportunity to make an original and important contribution. Every person involved in this effort proved to be ‘the’ right person for the job. Lesson learned: success depends on how well you pick both your partners and your project.