Strategies for preserving textual heritage in the digital domain in developing economies

Dibyajyoti Ghosh

PhD student

Jadavpur University

Though CIDOC's primary mandate is documentation, I shall be speaking not so much about documentation but rather about digital preservation. Also, as the title of my presentation specifies, I shall be speaking solely about textual material—a kind of artifact associated more with libraries and archives than with the two other sects of the GLAM quartet, that are museums and galleries. I shall divide my presentation into six categories, such as 'Academic outsourcing', 'Training', 'Data Enhancement', 'Digital Archives', 'Physical Digital' or 'Digital Materiality' and 'Open Access'.

Academic outsourcing

The background paper for this conference discussed the need to forge new partnerships, overcome the paucity of funding for training personnel and make the effort financially sustainable. I shall begin my presentation with a brief narration of my own entry into the world of archives and digital preservation, not merely to put my opinions in perspective but rather to illustrate a point regarding academic outsourcing and training.

While Linked Data and RDF are the latest buzzwords in the field of textual computing, six years back, in 2009, it was the XML-Text Encoding Initiative (XML-TEI). While digital humanities, as 'computing in the humanities' renamed itself in the first decade of the twentieth century, was a thriving field in North American, European and Australian humanities scholarship of the period from 2000-2010, it was a largely untapped field in the world of humanities studies in South Asia. In India, one of the first centres of digital humanities scholarship was an archive created at Jadavpur University in Kolkata, named the School of Cultural Texts and Records. It was set up in 2004. In 2009, a team of Australian humanities teachers were trying to create a digital variorum edition of the works of the 19th century Australian poet, Charles Harpur. With the meagre amounts of funding that they had managed to get, it was difficult to afford more than 2000 person-hours of work involved in transcribing 19th century manuscripts and encoding them in XML-TEI. On mentioning this problem to a fellow scholar in India, the Indian scholar readily agreed to complete the work in India as an 'outsourced' job, knowing that in his university in India, both the knowledge of advanced scholarship in English and basic familiarity with low-level programming in computers was available. That is where I stepped in, as a fresh Master's graduate of English. The Indian scholar I mentioned was my teacher and the founder-director of the archive named the School of Cultural Texts and Records. He employed me in this XML-TEI encoding project, the first such project in India. It is not as if XML-TEI is a skill that was taught to me as part of my BA and MA degree courses in English. However, I had picked up very basic HTML as a hobby. Armed with this basic knowledge, the rest of the skills needed in order to use XML-TEI were picked up on the job, as it were. Given that the project was time and money-bound, I, as a practitioner of the digital humanities and an employee of an archive, had to learn as much TEI as necessary to complete the job satisfactorily.

The point I am trying to make through this personal anecdote is that in developing economies such as India, which in many senses, is trying to catch up with the state of digital scholarship in archiving and preservation in countries with more developed economies, there is usually a dearth of funding for training personnel. However, some aspects of digital scholarship in preservation do not require

a. physical proximity to the material

b. significant interaction with people very familiar with the material, or in other words, people who can put the material 'in context', to borrow a phrase from the background paper of this conference.

Such aspects of digital scholarship can often be outsourced. There are pros and cons of outsourcing of course. However, with limited funds in some cases, in terms of preservation and archiving, even in developed economies, it would not be prudent to not consider the issue of outsourcing at all. India, with its size and history, is perhaps best suited for such academic outsourcing with respect to material in English. Not only is there a large pool of practitioners with advanced skills in the language but also, given the economic conditions of humanities academia in countries with significant primary material in English, it seems a good deal for both sides.

Training

While I have just discussed training people by directly employing them in projects, it is also necessary, in order to enroll large numbers, to have proper university courses to impart such skills. Students best suited to working as practitioners in Galleries, Libraries, Archives and Museums or

GLAMs are perhaps those who combine a sense of history with other managerial and technical skills. Given such a premise, it is perhaps worth it for higher education policy makers in India to encourage Indian universities equipped to do so, to offer courses in what is called 'digital humanities'. While it may seem a trite comment to make to scholars from North America, Europe, Australia and Japan, it is worth pointing out, that in India, there is only a single full-time digital humanities course, and that is again offered by the archive which I mentioned earlier, the School of Cultural Texts and Records at Jadavpur University in Kolkata. It would be worthwhile for higher education policy makers in India to encourage other Indian universities to offer similar courses.

While universities may slowly start offering such courses, the fact remains that the sole such course in India now is not wildly popular. The reasons are many. One of them being that students who graduate from the course are at least 24 years old, have advanced humanities and technical skills, but are usually unemployed! University courses run in tandem with industry demand. GLAM in India usually do not hire digital humanities specialists. They have their own IT teams for sure, if their budgets permit, but such IT teams perform the back-office work. Merging the two skills in one person is not a recipe for the dreaded 'rationalisation of the workforce' but rather a merging of two visions. It is again up to policy makers, this time of GLAMs, to encourage the hiring of such personnel. With this twin process of encouraging both supply and demand, the 'Digital India' that Indians are led to push forward would take shape in a more meaningful way.

The conference background paper also mentioned about the fact that the way funding is secured in both universities and GLAMs is often through proof of published research output and thus the disincentives to collaborate. A method of overcoming this challenge is by GLAMs offering their own resources, such as cameras and computers, to work with their own material, such as the holdings of that particular GLAM, in their own premises, to university students. While it may seem that the university student would gain nothing for her or his labour whereas at the risk of exposing their material and their equipment, the GLAM would get free labour, the equation can be more balanced if the students are offered academic credits towards their courses for their labour. A press report came out that the National Library of India was already working towards such a plan. It is again for education and cultural policy makers to give the gentle nudge and push to such matters to facilitate such collaboration. Whereas most GLAMs which have the budget to do so have some kind of digitisation programme, such unpaid internships where the students are compensated for their labour not through monetary payments but by credit points towards their degree courses, is perhaps an idea that both sides can explore.

To train greater numbers in data archiving and preservation, all universities, in fact, can be encouraged to set up archives of physical and digital material and offer degree-level courses in archiving and digitisation. Creating physical archives at the university level will ensure the creation of archives with a local focus. Digitising that same material will also create a digital database. Whereas, creating a physical archive is a less fund-intensive exercise if one were to not buy the material but rather depend on donation of the material, digitising such material is a fund-intensive process, and funds for such processes, as most of us have experienced, are not enough to go around. Offering degree-level courses in archiving and digitisation is one way however of getting the students to digitise such material at no additional cost. Not only do the students get hands-on training, as I did and which was the point of me recounting my personal experiences, but the digital archive is created as creating it is being taught.

Data enhancement

One of the major differences among nations which use the Latin script is that much of their textual digital material is text-searchable, whereas for a nation like India, the absence of good OCR (Optical Character Recognition) software prevents digital archives from enhancing their digital data through such means. While the history of OCR attempts in India is a long and considerable one, it is also a history of failure. Even Google, which promotes itself as the most innovative corporate behemoth in the twenty-first century, has failed to make much headway with OCR of Indic scripts. Research in OCR of Indic scripts is not as well-funded as it should perhaps be. In order to stimulate such research, a public-private partnership in developing OCR software for Indic scripts is an idea which can be explored. Whereas the private entity can come up with the money for research, it can thereafter make up for its expenditure by being allowed to carry out the task of OCR-enabling already digitised material held by public institutions in India.

Digital archives

Gutenberg is said to have produced around 180 bibles in the 1450s. Of those 180, 135 were printed on paper, and the remaining 45 on vellum. Almost 675 years later, 45 of these 180 bibles are still extant. 4 vellum and 12 paper copies are complete, and the remaining 29 copies are fragments.

Tim Berners-Lee's website was created in August 1991. 24 years later it is no longer extant.[1] While most things are easy to destroy, things that are easy to create, are perhaps even easier to destroy. Digital archives disappear fast, for various reasons. Lack of money once a project submits its 'deliverables' is one of them. In case of individual efforts, the failure to pass on the mantle is often the primary reason. Whatever be the reason, the issue remains that digital artefacts have a surprisingly short life. This has been one of the issues discussed in earlier CIDOC conferences as well. Whereas the Digital Library of India and the various DSpace repositories of various institutions hold digitised versions of print resources, there are no repositories for born-digital resources in India, along the lines of the Internet Archive Wayback machine or the British Library's UK Web Archive. The UK came out with the Non-Print Legal Deposit Act in 2013. In India, the Print Legal Deposit Act is rarely enforced. But failure to enforce existing rules should not be an excuse to not come up with new rules. India too needs to develop a non-print legal deposit act and have repositories for born-digital resources. The Digital Library of India can perhaps have a separate section which houses such born-digital resources and perhaps that is one of the ways to go about creating an Indian Internet Archive.

Physical digital or digital materiality

Along with the lack of resources to maintain digital archives and the failure to transfer managerial responsibility, one of the other major reasons for the destruction of digital resources is the failure to preserve digital data. One of the reasons why extremely expensive LP records are now being re-introduced by major music labels is because LP records last much longer than optical disks. Not only is the high failure rate of hardware resources a major reason for destruction of digital archives but also the absence of working software is another important cause.

Thus, in addition to digitisation programmes, every GLAM having such a programme should also be asked to maintain working-models of devices for accessing data on a certain medium, such as floppy drives, so that already-digitised material stored in such media can be retrieved and transferred onto future media and ensure zero data-loss. Such working-models of devices also need

[1] I have picked up this observation from a slideshare presentation by Andrew Prescott. Andrew Prescott, 'Sustainability: Some Moral Tales', Early Modern Digital Agendas, 29 June 2015, http://www.slideshare.net/burgess1822/sustainability-50056149

to be preserved with ‘old’ and ‘outdated’ software, as the latest updates often make some forms of data inaccessible.

Open access

Open access is a fraught issue. However, instead of discussing that I plan to end my presentation with the idea of making data sets open-access.

The humanities research involved in most digital humanities projects is a laborious task and has usually a longer period of validity as opposed to the technological aspects of DH projects. However, whenever the funds run out, the technological aspect of DH projects is no longer updated. Given the short-term validity of technology, most DH projects seem outdated technology-wise within say a period of 5 years since its completion. Thus, unless the data sets that underlie DH projects are made easily accessible to the public at large, the data of most DH projects is not amenable to re-use. Thus, funding agencies which insist on open-access should also ensure that digital archives make their data sets free to access and easily downloadable in large batches.

These are some of the strategies that India can adopt in its initiative to create a ‘Digital India’. Some of these strategies may also find resonance with GLAM professionals from other countries and other situations, such as participants at a CIDOC conference and hence my presentation.