CASE STUDY

A Digital Curation Centre Case Study
November 2013

Planning for the future: developing and preserving information resources in the Arts and Humanities.

Section of the How to guide that this supports
Data Management Planning

Jonathan Rans, Digital Curation Centre

Introduction

The Arts and Humanities Research Council (AHRC) requires the production of a technical plan for any project it funds in which digital technologies play a significant part. The plan should give a summary of digital outputs, explain the technical methodology, describe the expected technical support that will be accessed and discuss plans for the preservation, sustainability and future use of resources produced.

In this case study, we examine the development of an AHRC technical plan by Open University (OU) researcher, Francesca Benatti, and discuss the benefits to her project that were realised with focussed forward planning. We examine some of the potential pitfalls when addressing the requirements of a technical plan and suggest methods for mitigating the risks.

Project background

The project is led by Dr Shafquat Towheed, Dr Sara Haslam and Dr Mathieu D'Aquin, and includes Dr Edmund King and Dr Francesca Benatti (all Open University). It follows on from a considerable body of previous work and aims to produce a sister database to the OU Reading Experience Database, developed over the course of the last 10 years and chosen by the British Library as one of their Curators' top 100 sites (http://www.bl.uk/100websites/top100.html). The project will draw primarily on a wide variety of existing digitised source materials, and some hard copy collections digitised as part of the project. These will be collated to produce, among its intended outputs, a searchable database containing around 20,000 entries.

A significant proportion of the project will involve the development, population and hosting of this linked database. Associated with this resource will be a suite of tools enhancing the utility of the data it contains; for example, visualisation tools will be provided which allow data to be overlaid on geographical maps, enabling selected results to be presented in a simple, accessible way.

Beginning the project planning process

At the outset of the project planning phase a two-pronged approach was adopted in which the technical plan was addressed at the same time that an analysis of available resources was conducted. The information gathering exercise looked at the sources from which data could be drawn to populate the database; much of the necessary research was performed by PhD students employed as consultants. Initially, potential institutional partners were approached and then latterly the search was widened to include national and international resources.

The decision to tackle the technical plan at such an early stage was informed by experience that the project team had gained from speaking and working with colleagues on similar projects, most notably the OU-hosted Listening Experience Database (http://led.kmi.open.ac.uk/) an AHRC funded research project currently being developed by Faculty colleagues that bears many similarities to the proposed database. In this case, developing the technical appendix for the AHRC bid required input from a wide range of stakeholders, necessitating multiple rounds of discussion and negotiation and consequently the timeframes for completion were surprisingly long. Although the AHRC had subsequently changed its requirements, with the technical appendix changing to a technical plan, there were still many useful, applicable lessons that could be drawn from the old document and the process of its development.

Defining the digital resource

Initially, the group expected to produce the framework of the database themselves, as Dr King and Dr Benatti have experience of Extensible Mark-up Language (XML) and the Text Encoding Initiative (TEI) standard. At this stage, they planned to host the resource through central IT Services, so they scoped what they wanted to produce and made contact with IT Services. However, IT were unable to provide hosting for this database at that point in time, so an alternative approach was required.

One possibility was to go to external providers who could offer hosting and support solutions; this option was rejected as it was considered essential that the resource be hosted within the bounds of the OU. This decision was primarily reached because of concerns surrounding data ownership but practical considerations around linking the database with existing resources hosted by the OU were also a factor.

Again, the experience of Listening Experience Database colleagues was useful in suggesting potential solutions. They had produced their resource in collaboration with research partners at the Knowledge Media Institute (KMi)[1] , suggesting a means of addressing the issue of live database hosting. At this point, colleagues from the Listening Experience Database team were able to provide two, named contacts from KMi with an interest and experience in this area, thereby saving time and effort tracking down relevant individuals.

Key Points

1) Start early – technical considerations can have a profound effect on how the research is conducted; it is easier to accommodate them during planning rather than trying to unpick a finished proposal.

2) Drawing on the experience of colleagues who have developed successful funding bids in similar areas can realise considerable time savings outlining:

- **Timeframes for robust planning**
- **Questions to address**
- **Specific contacts in central services**
- **Potential technical collaborators**

Entering collaboration

Now the project became a collaboration, in which colleagues from KMi took on responsibility for developing and hosting the database, and its associated tools, accessible via a public-facing website. In return for this technical input, KMi colleagues have incorporated their own research questions into the project proposal, leading to the inclusion of a KMi researcher as Co-Investigator (Dr Mathieu D'Aquin). Reaching an agreement on the terms of collaboration required a series of discussions to ensure that the project's research goals were framed in a mutually acceptable way.

In fact, the involvement of a technical partner introduced new possibilities to the project by boosting the capability of the original team. New ways in which the research data could be manipulated and interrogated opened up, enhancing the utility of the dataset. For example, KMi colleagues suggested the inclusion of a named entity identification system as part of the database. This will enable the unambiguous identification of individuals within the database allowing cross-linking between entries. This fundamentally changes the functionality of the resource and opens the door to many complex queries which would have been impossible to run on a simpler database. This provides an excellent example of the way in which including a technical partner can significantly alter the course that a research project can take and highlights the importance of allowing time in the planning stage to accommodate changes in research direction.

Key Points

1) Technical partners may wish to be Co-Investigators, with input into the research over and above that of paid support.

2) Collaboration with a technical partner can open up new possibilities for exploiting data resources, fundamentally changing the research questions that can be asked.

3) Negotiating with partners and reshaping research questions can take time, so begin planning early.

[1]*The Knowledge Media Institute (KMi) is an R & D lab at the Open University, undertaking research into knowledge and media technologies.*

Creating a sustainable resource

After the end of the funded project phase the AHRC requires that significant digital outputs are preserved in an accessible repository for no less than three years. Of course, for many projects the continued utility of their outputs may mean that it is sensible to build much longer-term preservation into the project plan; in some cases, it will be appropriate to expect to keep the resources useable in perpetuity.

For this case study's database, the question of sustainability has been the primary challenge of the planning phase. After the end of the grant-funded project period, KMi will have to pass the management and hosting of the database over to OU central services. Theoretically, if this were a static resource, the Library Services would be able to hold a preservation copy of the database but the intention is to keep the database live and growing, potentially being used and added to by future projects.

After discussing potential options with the Library Services, the group concluded that IT services were best able to support the continued hosting of the database. As sustaining these digital tools and materials falls outside of normal operations for IT Services and requires investment in specific software, further funding from some source must be identified.

Developing a case for sustainability

Building a robust case for the sustainability of a digital resource in the post-project phase must begin by scoping, as accurately as possible, the scale of technical input required for ongoing support. This detailed requirements list forms the basis of a financial plan, which should be developed in conjunction with appropriate members of central services support.

Initially, the group took advice from the Research Support team based within the Faculty of Arts and Humanities to help frame the scoping questions they needed to ask KMi and IT services. These covered the resourcing costs associated with activities such as data and metadata capture, data storage, and curation. Once the technical requirements had been defined, the group worked with IT services to develop an investment case outlining the expected resource requirements and their associated costs. It took time to finalise the financial plan, with discussions going back and forth between the research group, Research Support, KMi and IT. As is to be expected with so many stakeholders, the process is a complex one, with much iteration.

Once a concrete financial projection has been achieved, the Research Office will take it to the Faculty, with the project team, and make the case for the release of funds. Approval will need to be included in the grant bid as evidence of sustainability.

Web resources vs. preservation datasets

Many projects will produce a website as a technical output of their project; for some this may constitute the main, or even only, technical resource of the project. It may be that the website is a fairly straightforward resource providing a publicly accessible description of the work undertaken. In these cases, maintaining the resource is rarely an issue as preservation actions are relatively trivial and can usually be covered by institutional IT services. However, when websites are used to host research findings and data, a more structured approach may be called for. Depending on the data type, the limitations of a website may make it inappropriate as the sole source of a research dataset. Take a collection of images as an example, for web-hosting it is practical to use low-resolution versions of the image to reduce page loading times. While these may be of sufficient quality to illustrate any points made in accompanying text, it is unlikely that they will be of use to researchers wishing to reanalyse or manipulate the images. When data is presented through a website it may give the impression that all source materials are held in a single location when, in fact, it is a melange, populated from various locations, potentially hard to track down and hard to preserve in the long term. In this case, a sensible strategy is to hold a preservation dataset of the images in a high-resolution, raw format. This would be a static collection of files, held in a suitable, accessible repository, thereby meeting both open data and preservation requirements.