AHRC ICT Methods Network Expert Seminar

SUSTAINABILITY OF DIGITAL RESOURCES IN THE ARTS AND HUMANITIES

Franklin-Wilkins Building, King's College, London, Wednesday 29 November 2006

Sustainability of Digital Resources in the Arts and Humanities: the AHRC Resource Enhancement Scheme

David Robey, AHRC ICT Programm, UK

1. Since its establishment the AHRB/AHRC has funded a substantial number of digital resource creation projects in the arts and humanities.This was done initially through the Major Research Grant scheme, then through the Resource Enhancement scheme, in both cases in responsive mode. The old Resource Enhancement scheme has now come to an end, and will be replaced by a new Strategic Resource Enhancement scheme; however Major Research Grants will continue to be available for the responsive-mode funding of digital resources. This paper discusses some of the sustainability issues that arise from the Resource Enhancement Scheme outputs, and suggests some solutions for the future. While some of these issues are peculiar to the arts and humanities and to the way in which the projects have been funded, they echo those that arise in other domains, and can therefore, I hope, be usefully compared and contrasted with them.

2. A survey of all Major Research Grant and Resource Enhancement projects funded during the period of the AHRB has showed that about half produced some form of digital output. The proportion is very much higher if Resource Enhancement projects are taken on their own. In the latter case the digital resource was also almost always the primary output of the project and, given the nature of the scheme, intended to be a lasting creation; in contrast digital outputs under the Research Grants scheme were often secondary to print publications, in the sense of supporting data not necessarily intended for general use. In the present discussion of sustainability, therefore, the Resource Enhancement scheme constitutes a significant body of evidence. At the last time of counting, of all awards made under the scheme up to the end of 2004 well over a hundred, to a total value of some £24m, involved the production of digital resources.

3. The typical digital resource produced under the scheme is an on-line database system of a specialist nature. This does not mean that some are not widely consulted: the Electronic Old Bailey Sessions Proceedings (http://www.oldbaileyonline.org/) have shown how much general interest a specialist resource can generate. But many are likely to be of use to only a restricted group of researchers. The content of the database is, in the majority of cases, either an archive of primary materials, often multi-media, for instance Ronnie Mulryne's 'Digitization of Renaissance Festival Books in the Collections of the British Library' (http://www.bl.uk/treasures/festivalbooks/homepage.html); or an electronic register of items, for instance David Walker's 'Dictionary of Scottish Architects' (http://www.scottisharchitects.org.uk/); or a catalogue, electronic edition or bibliography.

4. It should be underlined that the digital resources funded by the AHRC have been content resources like those above, not, with very few exceptions, digital tools. If the use of ICT for research in the UK arts and humanities is to continue to develop as vigorously as it has, there is a real need to ensure better provision for tools development. A major opportunity for tools funding currently exists through the joint AHRC-EPSRC-JISC Arts and Humanities e- Science Initiative, but apart from this the prospects of obtaining tools funding from the AHRC in the future seem quite problematic. The present paper, at all events, is only concerned with content resources of the type listed in the previous paragraph, and not with the sustainability of digital tools.

5. Along with the ESRC, the AHRB/C has been ahead of most of the other Research Councils in requiring archival deposit of the digital outputs it funds, in our case in the Arts and Humanities Data Service (AHDS). Applicants have been required to contact the AHDS before submitting their pplication,

and to complete a Technical Appendix detailing the technical and managerial procedures proposed for the digital output; the AHDS then reports on the Technical Appendix to the relevant AHRC panel. Until recently there was no formal requirement to remain in contact with the AHDS after the proposal was submitted, but steps are now being taken to ensure that such contact does take place in future, and that material is submitted to the AHDS in a form suitable for ingestion. Deposit in the AHDS is, however, only part of the story.

6. It is important for the present discussion that Resource Enhancement outputs tend to be quite complex, often multi-media, systems, including a user interface or front end where much of the information about the nature and purpose of the resource is presented, and in which a good part of its intellectual content resides. These have generally been developed by the grant-holder for use on the host institution's web-site, and are not easily transferable to other locations. The typical AHRC-funded resource is thus different from those funded by the ESRC, and normally deposited in the UK Data Archive. The latter are mostly tables of (generally numerical) data, which are downloaded by researchers and used with tools that the researchers themselves supply. To gain full benefit from the typical Resource Enhancement project the researcher needs to access it through its own on-line interface.

7. We have surveyed the main outputs from all Resource Enhancement projects that have completed before the end of 2005, a total of 74. For the moment any results that we can report are approximative, since the outputs have often been quite difficult to find. Only a handful have been printed. Most, as I have said, are on-line database systems of one kind or another. In a few cases (10, as far as we can see) the full resource is hosted by the AHDS; the great majority are hosted by a wide range of university or library sites across the country. Some of these have deposited copies of their data with the AHDS, in many cases the AHDS is still negotiating for deposit, and in a few cases there seems have been no contact with the AHDS at all. The onus, it should be said, is on the grant-holder to make the contact.

8. The AHDS is in the process of developing a collections management database, and this, together with the application of a degree of pressure by the AHRC, should in due course be sufficient to ensure that all, or almost all, grant-holders comply with the conditions of their grant and give the AHDS copies of their data. However the variation in technologies and standards used, combined with developer preferences and quirks, means that the AHDS cannot routinely take on-line systems in their entirety, including the user interface or front end. We thus have in place, or soon will have, a reasonably effective system of data preservation; what we do not yet have is a system for ensuring the sustainability of the entire digital resource.

9. We need to distinguish between at least two aspects of sustainability, which we can call academic and technical: it involves keeping a resource (i) up-to-date in terms of its academic content, and (ii) available and fully functional within the technical environment in which it has been created and presented.

10. Academic sustainability is an issue whose importance varies from case to case: some resources are static or 'frozen', as far as their content is concerned, at the point of delivery, for instance an electronic edition of a text. Others are essentially dynamic, and quickly lose value if their content is not updated; an example is The Royal Historical Society Bibliography (http://www.rhs.ac.uk/bibl/), which is updated three times per year and requires the employment of one salaried full-time person. Many, possibly most, resources are somewhere in between, and require only occasional updating to keep up with scholarly developments. These present a widespread if not very grave sustainability problem, since there is at present no regular funding stream to cover the occasional costs involved; updating depends on the willingness of those responsible to carry it out unfunded. The likely consequence is that most such resources will lose value rather slowly, but lose it nonetheless.

11. The problem of academic sustainability could in principle be solved quite easily, in many or most cases, if a funding stream could be found to cover it. I shall be concentrating mainly on technical sustainability in this paper, since the solution to the problem is rather less obvious; as it happens I also believe it can in principle be dealt with within the existing resource framework. Technical, like cademic, sustainability is a problem for some but not all digital resources. Where the data is simple in form, for instance a single electronic text, the current preservation system may be sufficient for purposes of technical sustainability: for as long as it remains deposited with the AHDS, a simple electronic text can

generally be made available for downloading. In the case of the, more typical, complex on-line systems, if these have been co-developed and are maintained by the AHDS, then as long as the AHDS continues to be funded to do so it will maintain them in their full functionality. The problem arises where they have been developed and are maintained outside the AHDS, on university or library sites.

12. How realistic is it to expect such host institutions to guarantee full availability and functionality of these systems long after the end of the grant period? Perhaps three years might be reasonable, but the normal processes of hardware and software development will inevitably mean that some degree of technical updating will quite soon be necessary. Estimates are that some form of upgrading is likely to be required after three years, major upgrades after five years, and the resource is likely to be unusable after ten years unless significant work is undertaken to upgrade and migrate systems. To undertake this work properly requires a serious long-term commitment, and one that many institutions are unlikely to wish to take on. Moreover, the expertise and knowledge available is likely to diminish as staff - particularly research assistants involved in the resource's creation - move on. One grant-holder we approached for technical information about his AHRC-funded database claimed to be unable to provide it, on the grounds that his technical assistant had moved elsewhere. On the whole I think the realistic solution can only be to develop procedures that do not rely on grant-holders and their institutions to sustain AHRC-funded digital resources.

13. Looking at the range of hosts of existing Resource Enhancement outputs, one can distinguish between different levels of sustainability risk, depending on the interest and competence that the host is likely to have in maintaining the resource. The British Library ispresumably pretty safe; a university department's web-site much less so, because of the likelihood of a change of personnel. We shall soon be surveying as many hosts as possible to find out more about their perceptions of the risk to the resource's sustainability.

14. When the host institution fails to maintain an on-line resource, the underlying content can usually be made available through the AHDS, where it should have been deposited. As we have seen, however, it is generally not feasible for the AHDS to take over the whole system, including the user interface or front end, in order to provide the same access to and functionality of the content for which the original system was designed. To do this would usually involve redoing at considerable expense work already carried out by the original developer, a task for which the AHDS is not currently funded.

15. It might be thought that this problem could be solved by imposing tight technical specifications on grant holders, so as to ensure that everything they develop could if necessary be taken over by the AHDS in its entirely. But this could on the one hand be unduly restrictive, and on the other would in practice be difficult to implement effectively. Nor would the problem be solved by a blanket requirement to use open-source solutions, as some have suggested. The AHDS currently accepts material in both proprietary and open-source formats, an obvious reason for adopting open source being that it is free and thus reduces costs. But the AHDS is not funded to preserve every kind of open-source database on AHDS systems. It would also be unduly restrictive in terms of technical solutions to exclude the use of proprietary systems altogether.

16. The solution, I believe, cannot lie in mere specifications, but in the institution of new collaborative processes for the creation of digital resources. Examples of these already exist in the form of database systems hosted, for instance, by AHDS Archaeology, that have been jointly developed by the AHDS Centre and an external grant-holder. But before considering how this model might be extended, there are other points that need to be kept in mind and provided for if the sustainability issue is to be properly addressed. These concern quality assurance and reusability, and harmonization or interoperability. Issues of intellectual property rights are also critical to the general sustainability question, but lie beyond the scope of this paper. Nor will I have anything to say here about payment or other funding models beyond the present arrangements for the AHDS and the AHRC's present grant-awarding procedures. As will be seen, the solution I propose for the future can be provided for within the present funding system.

17. The issues of quality assurance and reusability need to be addressed for the simple reason that if we are to invest in making digital resources sustainable, we need to know whether they are worth sustaining. The present system of AHRC end-of-award assessment provides some guarantee of the academic quality of the digital outputs of research grants, but is not rigorous from the technical point of view, and may even leave something to be desired from the academic point of view as well. We need

to develop a more rigorous process for the publication of digital resources, as opposed to simply making them available on the web: what procedures can we develop for digital resources that could provide guarantees of quality comparable to those that a reputable publisher provides for printed books? Among other things, digital data projects seem to vary greatly in the extent to which they take the needs of users into account, other than those of the users involved in the creation process. Many may only be concerned to cater for their own academic interests, without thinking much about the issue of reusability, that of the other potential uses to which the data could be put. To cater properly for reusability the data creator needs to be aware not only of the data's potential academic significance, but also of the full range of technical methods that can usefully be applied to it: digital output projects need to be informed by methodologies of use as well as of creation and preservation. We need to ensure that this area of expertise, currently represented by the AHRC's ICT Methods Network, remains available to data creators in the future.

18. There is a further aspect of reusability which is not central to the AHRC's mission, but to which thought ought surely to be given in the resource creation process: the potential use of the resource for teaching and learning. Given the cost of producing the typical AHRC on-line databases, and the general desirability of promoting links between research and teaching, there is a serious case for seeking to facilitate their usability as resources for education as well as research. This is not to say that the AHRC should allow significant funding to be allocated to this purpose; it merely makes sense to try to ensure that front-end design is informed by an awareness of teaching and learning needs.

19. Harmonization and interoperability are important because, in addition to ensuring that the resources the AHRC funds are sustainable on their own, we should also aim to maximize their value by providing the means to connect them together. Excellent though many existing resources are, their value is significantly limited by their technical separation from one another: the more related resources can be connected, the more useful they are likely to be. Considerable intellectual and technological effort is currently being spent on the solution to this problem, mainly in science, technology and medicine, but increasingly in the social sciences and the arts and humanities as well. This effort is a large part of what the UK e-Science programme is about, a programme to which the AHRC is now contributing through the AHRC-EPSRC-JISC Arts and Humanities e-Science Initiative. But the technology developed through this and related initiatives will not provide instant solutions to the problem of data dispersion. The goals of interoperability and harmonization will need to be pursued from both above and below: from above through the development of software solutions that facilitate the interlinking of dispersed data; from below by developing the data in formats, and on the basis of standards, that makes it more capable of being linked, alongside the provision of metadata that supports integration and harmonization and facilitates rich, deep querying of distributed resources.

20. Many of the considerations put forward above are developed and reinforced by five survey projects recently funded by the ICT Programme as part of its ICT Strategy Projects scheme. These have now reported, and we shall shortly be making their findings more widely available. All contain a great deal of material relevant to the topic of this paper, and also show a striking measure of agreement on a number of issues. Their findings are based on a mix of surveys of academic users, interviews, focus groups, seminars, and in two cases web-log analyses.

The projects and their principal investigators are as follows:

- Sheila Anderson, 'Scoping e-science and e-social science developments and their value to the arts and humanities'
- David Bates, 'Peer Review and evaluation of digital resources for the arts and humanities'
- Mark Greengrass, 'RePAH: Research in Portals in the Arts and Humanities'
- Lesly Huxley, 'Gathering Evidence: Current ICT Use and Future Needs for Arts and
- Humanities Research'
- Claire Warwick, 'LAIRAH: Log Analysis of Internet Resources in the Arts and Humanities'

21. David Bates and his colleague Jane Winters will be presenting their project to the seminar, so I shall say no more about it other than to underline its importance for the present discussion, because it

proposes precisely the sort of model that I suggested above was needed for the quality assurance of data outputs. Sheila Anderson's scoping survey has helped to inform the Arts and Humanities e-Science Initiative, and has much to say on issues of data integration; it is however mainly for discussion in another context. Mark Greengrass's study of portals also mainly relates to a somewhat different issue from that of sustainability, though it has much to say about ways in which researchers currently use, and might in the future use, data resources. On the other hand Lesly Huxley's report on uses and needs is very relevant to the present context because of key concerns it identifies on the part of the researchers they surveyed. Exactly half of their respondents cited the quality and reliability of resources as the source of most difficulty in their use of ICT in research; respondents also identified sustainability of research resources is as a problem, particularly when material is deposited locally rather than nationally. It is helpful, if not reassuring, to see that some of the concerns I have expressed above seem to be shared by a significant proportion of the relevant research community.

22. Claire Warwick's study has even more to say in relation to the present topic. Its principal purpose of conducting a web-log analysis of internet resources in the arts and humanities was limited, as it turned out by practical problems, but was supplemented by a user questionnaire, interviews and workshops focusing on a selection of arts and humanities data projects. Among other things, its findings suggest that 30-35% of digital resources remain unused, that the title of the resource affects whether it is used or neglected, that users tend to abandon a resource if in any doubt about a its quality or authority, that non-expert users found it difficult to understand the purpose of several resources. Yet few projects kept formal documentation or made it easily available from the project's own website, few carried out formal user testing, few realized the importance of ensuring their resource remained sustainable and that both content and interfaces must be maintained and updated.

23. This report in particular bears out the points made above about the need for more attention to sustainability, quality and reusability issues. One obvious conclusion is that more control needs to be exercised over the data development process, in order to ensure that the outputs are as well designed, effective and useful as they can be. Another conclusion is that the problem of long-term sustainability is most easily resolved by ensuring that the AHDS can preserve not just the underlying data of resources, but the full system including the user interface. This can only be achieved if the AHDS has a much larger hand in the resource development process.

24. A possible solution, therefore, but not I think the best one, would be to establish a new, joint-development, relationship between future grant-holders and the AHDS. Following some examples in AHDS Archaeology, the AHDS could provide a technical infrastructure, into which content could be placed, for which an adapted and customized interface could be created in conjunction with AHDS staff for each grant-holder's needs. The specifications of the infrastructure need not be particularly restrictive. It could include a variety of solutions, including proprietary systems, so long as the AHDS can develop, support and migrate them. The cost of the underlying technical infrastructure would be included as part of the core grant for the AHDS, whilst the cost of creating and customizing the front-end, and developing the functionality specified by the grant holder, could be built into the grant application. Different levels of support could be offered within this model, ranging from basic programming to create and customize the front end through to a full partnership. The Stormont Papers (http://ahds.ac.uk/about/projects/stormont-papers) project is an example of the latter, with three members of AHDS staff in London aiding this major digitization project based in Belfast. The development could be based either at the project's institution with help from AHDS staff, or at the AHDS.

25. When completed, the resource would either be kept at the AHDS, or mounted on the grant-holder's website and connected to the AHDS technical infrastructure to ensure cross-searching and full interoperability with other AHDS collections. The AHDS would also monitor the need for updates and migrations across systems and hardware. These could either be done in collaboration with the grant-holder's institution or the resource could transfer to the AHDS in the event of the host institution being unable or unwilling to continue to maintain and upgrade it. Resources could be personalized for the grant-holder or his/her institution, and, even if kept only at the AHDS, could look to the world as if they resided on the grant-holder's site. The grant-holder could retain ownership of the resource, with the AHDS having the right to disseminate and update the technical infrastructure in which it sits. The key consideration is that the technical sustainability of the resource would be assured by the AHDS, with the minimum updating required to ensure continued accessibility and functionality covered by the AHDS's core funding. The AHRC could consider making funds available against further grant

applications after the end of the project period for more extensive technical upgrading and academic updating.

26. I think a solution along these lines is needed, but not quite in this form. While the key requirement must be that the whole resource should be capable of being maintained by the AHDS, Insisting that the AHDS should carry out all technical development work would 0unnecessarily curb the creativity or flexibility of researchers. Just as importantly, there are other centres in the UK that have expertise in developing data systems comparable to that of the AHDS, and it will surely be desirable to take advantage of their expertise in future AHRC-funded development work. The best solution therefore seems to be the one that Sheila Anderson will propose later in this seminar, in which the key technical development role would be shared by a network of existing expert centres including the AHDS and coordinated by it. Grant applicants would designate their preferred partner in their grant application and provide in their applications for the necessary funds to support the partner's involvement in their project. At the same the coordinating role of the AHDS should ensure that projects are developed in a form that the AHDS is capable of maintaining, if necessary, in its full functionality.

27. I leave it to Sheila to explain how this network would operate in practice, and will simply conclude by detailing the reasons why this seems the best solution to the problems outlined above. It is obvious that having a network as a required development partner, rather than the AHDS alone, greatly increases the flexibility of solutions available to data projects. Equally importantly, the network organization can ensure that a common body of expertise, tools and resources is built up over time, so that each project can benefit from those that have gone before it. In particular the network would

a) be cost-effective, since technical development costs would be kept low through the use of standard tools and a network experienced staff, avoiding reinventions of the wheel;

b) help close the circle between creators and users of data, since AHDS could channel expertise on uses of data (e.g. from the ICT Methods Network) into the data creation process;

c) liaise and share expertise with other agencies in the data development field, notably JISC and the Data CurationCentre;

d) take care of technical quality assurance--though a parallel system of academic quality assurance would need to be implemented;

e) help to ensure the resource's wide dissemination, through the AHDS catalogue system, and by appropriate exchange of information with organizations such as Intute Arts and Humanities;

f) increase the scope for harmonization or interoperability between different resources through the use of a consistent technical envelope;

g) give the AHRC added value from a service which it already funds.

28. Let me make it clear, finally, that these ideas have been developed by the AHRC ICT Programme with substantial input from a number of participants, especially in the AHDS, and are proposed here for further discussion. They are not AHRC policy, though one of the results of the present proposals and the discussions that follow will be a set of recommendations to the AHRC for a new sustainability policy for the forthcoming Strategic Resource Enhancement scheme, and for future Research Grant Awards where data resources are a major output. We are also still left with the sustainability problems that are likely to arise with a large number of existing outputs. If these are to be resolved, a new, though not very large, funding stream will be required both for content updating, and to provide, at least in many cases, for eventual ingestion into and maintenance by the AHDS.