An Arts and Humanities Data Service

Report of a Feasibility Study
commissioned by the
INFORMATION SYSTEMS SUB-COMMITTEE of the
JOINT INFORMATION SYSTEMS COMMITTEE of the
HIGHER EDUCATION FUNDING COUNCILS

Lou Burnard
Oxford University Computing Services
Harold Short
King's College London

November 29, 1994

This page left intentionally blank

Contents

Contents

1 Introduction

This document is the Report of a Feasibility Study funded by the Information Systems Sub-Committee (ISSC) of the Joint Information Services Committee (JISC) of the Higher Education Funding Councils. The Study arose out of a Proposal submitted to the Committee by the authors of the present report in February 1994.[1] Support for a feasibility study on this topic was also recommended in the *Report of the Joint Funding Councils Libraries Review Group* (the "Follett Report").[2]

The Proposal was also in part prompted by the report of a Humanities Information Review Panel, published jointly by the British Academy and the British Library,[3] which documented the growth of electronic resources in the humanities, and recommended the development of a national strategy for "the holding, maintenance and provision of access" to such materials.

1.1 Summary

The objective of the Feasibility Study was to consider how a broad and flexible *Arts and Humanities Data Service* (AHDS) might be established, within which the development and use of electronic resources would be both widened and facilitated, while remaining both responsive to change and cost-effective.

The recommendations of the Feasibility Study may be summarized as follows:

- the role of an Arts and Humanities Data Service should be to promote effective, low-cost access to the widest range of relevant digital resources by UK researchers and teachers from all Arts and Humanities disciplines;
- the functions to be performed include: cataloguing and identification of resources; development and application of standards; documentation and training; archival preservation of existing resources; identification and creation of new resources; protection of intellectual property rights;
- the Arts and Humanities Data Service should have a distributed structure, with some functions being performed by a central infrastructure, and others by co-operating specialist data service providers;
- the Service should define standard policies with respect to the description and encoding of data resources, their preservation, their dissemination, and a range of ancillary support facilities and services;
- the Service should invite bids from suitably qualified institutions wishing to provide specific resources within the AHDS infrastructure and define ways in which such service providers may co-operate;
- initial funding is required for an executive to oversee the process of establishing an organizational framework for the Service, and for a small group of four or five collaborating service providers. These should be chosen by open tender, with a view to covering the areas of textual data, numeric and historical data sets, records of images and artefacts, time-based resources such as music and video; the total budget for the Service envisaged here is likely to fall between 500,000 and one million pounds per annum during the start-up period.

[1] L. Burnard and H. Short, *Feasibility study into establishing an Arts and Humanities Datacentre: a proposal to the Joint Information Systems Committee*, February 1994.
[2] See para 299 of the Report (hereafter *Follett, 1993*).
[3] *Information Technology in Humanities Scholarship,* British Library R & D Report 6097, London 1993.

1.2 Outline of the Report

The remainder of the current section summarizes the principal issues addressed by the Feasibility Study, focusing on its timeliness, the problems it addresses and the solutions proposed. A brief account is also given (in section 1.8) of the consultative process by which the Study was carried out.

Section 2 contains a summary overview of institutions, services, and patterns of provision currently obtaining in this area, some of which are described in more detail in Appendix 8. This section also includes a discussion of the user needs identified by this study and the way in which those needs are currently met, which is largely derived from the consultative process carried out during the Feasibility Study.

The Recommendations of the Study are presented under four major headings. In section 3 the scope of the Service is discussed in terms of disciplines and data types, and the community which is to be served. This section also describes the organizational structure proposed for the Service, together with the functional requirements for its various proposals. The proposals are intended to combine the virtues of both distributed and centralized approaches to the management of resources.

Section 4 discusses the application of standards within the AHDS, which form a key element in the scheme proposed here. Section 5 outlines a possible implementation strategy for setting up such a service and section 6 makes some preliminary estimation of the likely costs of the proposed service.

1.3 Timeliness and relevance

Recent major developments in information technology are beginning to have a significant impact on humanities research and teaching. This is due partly to the appearance of more sophisticated hardware and more appropriate software, but it is above all the rapid spread of local, national and international networks which is now presenting the higher education community with both problems and opportunities.

The opportunities arise in particular from the rapid increase in the number, variety and accessibility of electronic data sets now being created for research and teaching. Electronic data sets are being produced across the whole range of humanities disciplines, covering both primary and secondary materials in a wide variety of formats – text, data, image, sound, and multimedia. At the same time, as a result of institutional information technology strategies, the number of teachers and researchers who are equipped to take advantage of these resources is steadily increasing, as is the recognition of their value in exploratory, resource-based, learning.

1.4 Current problems

Unfortunately, however, there are a number of problems associated with these developments. These include:

dependence on specific hardware or software: this makes it hard both to archive data sets effectively and to ensure their long term *re-usability*, without significant additional costs;

inadequate or non-existent documentation: electronic resources are largely unusable without some form of documentation, which is all too often not readily available, either as a result of lack of awareness of the need for it, or because in many areas there are no agreed documentation standards;

inadequate resource discovery tools: in the absence of both widely accepted standards for cataloguing such resources and any simple mechanism for searching such catalogues as exist, it is difficult or impossible for individual researchers or teachers to locate existing data sets, or to carry out reliably complete searches through available resources, even with skilled bibliographic assistance;

a variety of incompatible complex access tools: because of the wide variety of software mechanisms now used for accessing and manipulating data sets, potential users must acquire a daunting range of information and skills, largely irrelevant to their immediate concerns;

insufficient attention to data preservation: in many cases there are no procedures or plans to ensure the long term preservation of the data sets, so that the investment in creating the data set will be wasted;

institutional inflexibility: the rapidity with which information technology changes has made it very difficult for existing institutional structures to adapt to new opportunities. Despite widespread recognition of the gradual convergence of library and computing services functions, it remains difficult to integrate the complementary skills of library and computing staff in most institutions. Such integrated expertise is, however, essential for the support of the growing numbers of IT-aware teachers and researchers.

1.5 Proposed solutions

This report proposes the establishment of a distributed Arts and Humanities Data Service as a means of both taking full advantage of the opportunities and also addressing the problems.

Such a Service will need to concern itself with a wide range of activities. We list here a few of the most immediately obvious requirements:

- identification of user needs;
- creation and acquisition of data sets;
- development and promotion of standard data formats;
- documentation of data sets;
- development of a comprehensive catalogue of resources;
- development and promotion of suitable tools for resource discovery and manipulation, both for UK and overseas data sets;
- archiving of resources;
- protection of the rights of data providers;
- provision of user training and support;
- promotion of the use of data sets in both research and teaching.

Some of these activities might be most effectively provided in a distributed environment, notably those requiring extensive specialist knowledge; others, notably the establishment of standards and the provision of comprehensive resource discovery tools, seem better suited to a centralized infrastructure.

The present proposal therefore proposes a framework in which a network of autonomous *service providers* can co-ordinate activities under the direction of a *service infrastructure*.

This distributed approach provides the flexibility needed to adjust to a rapidly changing landscape of information provision. The Humanities comprise many widely-divergent disciplines, each with its own highly specialized requirements, but each also highly inter-dependent. The approach proposed here encourages the emergence of a consensus on such issues as standardization of resource description and provision, without imposing a rigid structure. It promotes the development and application of relevant specialist expertise and facilitates the evolution of a service which accurately reflects the real needs of a large community with widely-ranging and diffuse needs. At the same time, the existence of an effective infrastructure allows for significant cost-sharing and the pooling of resources wherever appropriate.

The proposed Arts and Humanities Data Service will comprise:

- a *Management Committee*, acting as liaison between the AHDS and its funders, which has final responsibility for the Service;
- an *Executive*, with responsibilities for co-ordination of AHDS-funded services, policy-making and day to day running of the Service as a whole;
- a number of *Service Providers* providing specialist services of various kinds, according to AHDS-defined standards.

This report proposes that a small number of such Service Providers be established initially, covering as wide a spread of Arts and Humanities data as possible. Service Providers will typically provide a range of expertise and support appropriate to particular disciplines and data types, building on existing areas of expertise, such as the Computers in Teaching Initiative (see section 8.1.2) and the Teaching and Learning Technology Programme TLTP projects, and liaising with relevant existing bodies, such as the JISC Datasets Steering Group.

In the long term, additional AHDS Service Providers will be established, and other independently-funded projects or institutions may wish to become *Affiliated Service Provider*s; mechanisms to facilitate this are also proposed.

A key service provided by the AHDS will be its *Catalogue* – a register of all the resources maintained by the Service Providers. The development of this Catalogue should be one of the first priorities in establishing an Arts and Humanities Data Service. The Catalogue will also specify the extent to which the resource concerned conforms to defined AHDS standards.

Information about resources available from institutions outside the scope of the AHDS will also be included within its Catalogue, where possible. One or more "Gateways" may be provided to such external services.

AHDS Standards will be defined in a *Reference Guide*, production and maintenance of which will also be the responsibility of the Executive, acting in consultation with Resource Providers and experts within the community. Layered standards will be defined not only for data formats and data description but also for data provision and preservation, as described in section 4.

In all aspects of its work, especially policy formulation, the Executive will work closely with a *Service Providers Forum*, comprising representatives of all Service Providers.

1.6 Benefits

The report identifies the following benefits resulting from the establishment of a distributed Arts and Humanities Data Service:

- transferability and re-usability of expensive data resources;
- a framework for the creation of complementary and linkable data sets;
- co-ordinated and coherent procedures for the investigation and adoption of relevant standards in data preparation, documentation, preservation and dissemination, and for involvement in European or international initiatives to create new standards;
- a framework and procedures to ensure the long-term availability of electronic data sets;

- faster and better-informed access to remote data sets by individual researchers, teachers and students;
- wider awareness and understanding of new access mechanisms and manipulation tools, thereby ensuring more efficient use of national and international resources.

The establishment of the AHDS should therefore make a significant contribution to improved efficiency in the conduct of research and the quality of research results. It should also greatly improve the efficiency and quality of teaching and learning facilities within the Higher Education community.

The proposal is timely, in that it comes at a moment when there is widespread recognition of the scope of the problems which it seeks to address. Its proposals are well-grounded in that they result from a broad-based consultation with a range of expert groups, as well as the specialist knowledge and experience of the Study's authors and Steering Group.

1.7 Financial considerations

The costs of providing a Service of the type envisaged here are non-trivial. We estimate that service providers will typically need at least two full-time employees, with secretarial support. Assuming adequate institutional support for computing facilities, and for most other aspects of the necessary technical infrastructure, this gives a minimum cost of 80K - 100K per service provider per annum. Where significant institutional support is not forthcoming, the costs of providing a service would, of course, be much higher.

For the infrastructure itself, we estimate an annual cost of the order of 150K, to cover at least three full time appointments, and funding for meetings, workshops, publications, etc. Again, it is assumed that the central AHDS office will be located at some existing institution which will underwrite establishment and similar costs.

1.8 Conduct of the Feasibility Study

This sub-section briefly summarizes how the Study was carried out, and the present report produced.

1.8.1 Workshop

The Feasibility Study began with a Workshop at which several of the ideas now proposed were first discussed. The Workshop was held at the British Academy on 29-30 March 1994. More than fifty invitations were issued, and the Workshop was attended by a total of 28 experts, drawn from a wide range of academic backgrounds and with particular interest and expertise in most of the relevant academic, professional and technical areas. Following the Workshop, a summary report was published which identified the key issues raised and points of view expressed.[4]

The Workshop participants took the unanimous view that what was needed was some kind of infrastructure which could effectively manage the provision of a range of distributed services, rather than a single centre to hold and archive data. This notion, presented tentatively at the Workshop, has become a key issue in the present proposals. The Workshop also strongly recommended that the term "data service" should replace "data centre" in the original project title.

[4] *Towards an Arts and Humanities Data Service*: Report of a Workshop held at the British Academy on 29 - 30 March 1994.

1.8.2 Consultation

The intention of the project team and steering group was to consult as widely as possible within the relevant communities. To this end the Workshop Report was disseminated widely by both paper and electronic means. Groups, institutions and individuals known to have an interest were identified as specific targets, but the distribution was deliberately wider than this, and comments were invited from any interested party. Full details of the process are given in Appendix 7.2.

Despite the very short time available for public discussion and comment, distribution of these documents generated a large and almost entirely enthusiastic response from the community. Many useful comments and ideas were provided, which have been incorporated in the present report wherever possible. Appendix 7.2.2 lists the institutions and individuals who responded during this consultation process. Full details of their responses are available on request from the authors.

1.8.3 Fact finding

It was hoped to investigate in some detail the provision of relevant data sets from a number of existing institutions in the UK, Europe and North America. To this end, a set of questions (reproduced in Appendix 7.3.5) was devised for use as a basis for gathering information by interview or other means.

Although time did not permit all the relevant institutions to be pursued, those approached were unfailingly constructive and helpful and provided much useful information. Where possible this information gathering was carried out by means of visits and interviews, and otherwise by electronic mail. There were also a number of conferences which covered issues relevant to the study. The institutions consulted and events attended are listed in Appendix 7.2.

In a number of cases, detailed visit reports were produced by project staff or colleagues; these are not included in this report, but are for the most part available electronically from the authors.

1.8.4 Project Group and Acknowledgments

This Feasibility Study was carried out by Lou Burnard of Oxford University Computing Services and Harold Short of King's College London Computing Centre. The project received support and guidance from a Steering Group appointed by the ISSC. Members of this group were:

- Chair: Dr Richard Heseltine, Librarian of the University of Hull, and member of the ISSC;
- Shirleen Craig, ISSC Co-ordinator;
- Dr Marilyn Deegan, Director, Office for Humanities Communication;
- Professor Denise Lievesley, Director, ESRC Data Archive, University of Essex;
- Dr Seamus Ross, Assistant Secretary (IT), British Academy.

Members of the above group provided helpful criticism and greatly assisted during the drafting of the report. Assistance in drafting and researching several sections of the report was provided by the Marc Fresko Consultancy. The authors of the report also gratefully acknowledge the assistance provided by colleagues and management at their own institutions, by many correspondents, and by Lorcan Dempsey of the UKOLN Office at the University of Bath. Likewise the authors gratefully acknowledge the hospitality and help provided by the institutions and individuals visited during the course of the study. The bulk of the report is, however, the work of the authors, who are responsible for any errors or inconsistencies found in it.

2 Perceived needs and the current situation

In this section we review the general HE framework within which data sets and networked information services are being developed, identify the needs of the Arts and Humanities within this framework, and discuss the current situation in relation to each of the needs.

2.1 Background

2.1.1 Data sets and services today

In some disciplines, and in some countries, deposit of digital data sets in "approved" archives is a common and accepted condition of grant. This is frequently not the case in humanities disciplines in the United Kingdom, as a brief informal survey of four funding institutions has shown.

Each year the British Academy funds a significant number of projects which result in the creation of arts or humanities data sets. Since 1992 it has required applicants for grants to provide evidence that they have made adequate provision for the long-term access and preservation of the resources generated by the project.

The British Library Research and Development Department funds several projects every year, some of which result in the creation of relevant data sets, though these often are small and with limited scope for re-use (some are specifically precluded from widespread use due to copyright restrictions). It typically has "encouraged" creators of data sets to deposit them at the ESRC Data Archive (see 8.1.3), without making this a condition of grant; not all data sets so created have been deposited. The Library is considering a policy which requires deposit as a condition of grant.

The Getty Art History Information Program has funded two projects in the UK explicitly concerned with the creation of data sets.[5] In both cases, the projects were required to deposit copies with the Institute (in California), but not elsewhere.

The Leverhulme Trust funds about 50 humanities projects each year, many of which include data set creation. It does not impose conditions concerning access, deposit or preservation.

The Economic and Social Sciences Research Council (ESRC) funds a variety of projects relevant to the humanities, particularly in history and linguistics; deposit of resultant data sets is a condition of grant. However, where researchers in the humanities have been successful in obtaining funding for major projects from more scientifically-oriented research councils, these have not always included any provision for the long-term availability of any resources created. For example, several language engineering projects were funded by the Science and Engineering Research Council (SERC), under the Speech and Language Technology initiative, but their availability to the humanities community at large remains unclear. Often individual resource creation projects have been expected to manage distribution for themselves, with consequent dangers of invisibility outside the immediate research communities served.

There is an expectation that data set provision can be made economically viable and self-sustaining in the long term. This may be true for some kinds of resource where the demand is clear and sustained; it is less self-evident in the general case, particularly for the wide range of materials likely to be of interest to the Humanities.

The provision of services is varied and and uneven. There are two formally constituted data archives – the History Data Unit of the ESRC Data Archive at the University of Essex,[6] and the Oxford Text Archive at Oxford University. Otherwise the

[5] These are the *Witt Computer Index*, at the Witt Library, Courtauld Institute of Art, and the *Census of Antique Art and Architecture Known to the Renaissance* developed jointly by the Warburg Institute in London and the Biblioteca Herziana in Rome.

[6] The History Data Unit was established by a grant from ESRC ending in December 1994.

management of data sets, including access, documentation, support and preservation, depends very much on the awareness and resources of the individual projects and their host institution.

2.1.2 Commercial data sets

In addition to the academic projects and institutions, several commercial organizations are now becoming involved in producing and publishing digital documents for scholarly use. Frequently, these are traditional paper-based publishers which are diversifying into electronic media. Their products range from novels sold on diskette for a few pounds to corpora sold on CD-ROM for tens of thousands of pounds.[7]

Recently the British Academy carried out a *Networked Data Information Survey* of all Arts and Humanities departments in the UK. This suggested that there is a demand for commercial data sets, many of which are too expensive for smaller institutions to obtain individually.[8]

While it is in any single publisher's interest to maintain archival copies of its materials, there is no commercial imperative to do so. On the contrary, it is easily possible to imagine commercial pressures which would result in data sets for which there is little demand being discarded. We are not aware of any unified source or catalogue for these materials; nor are we aware of any widely-used standard for works published in electronic form.

2.1.3 Resources for teaching

There have been three major HEFC-funded UK initiatives in the past few years, all aimed at promoting the use of electronic resources in teaching, and at creating a range of transferable materials. First was the Computers in Teaching Initiative (CTI) (see 8.1.2), followed more recently by the Information Technology Training Initiative (ITTI), and the Teaching and Learning Technology Programme (TLTP).

2.1.4 Networked information in the UK

The creation and management of networked electronic resources is currently a major area of activity within the Higher Education community, with a number of bodies and institutions involved, and a number of projects and initiatives.

Within the UK the academic communications infrastructure – JANET and SuperJANET – has largely, up to now, been centrally funded by the HEFCs. The funding councils are advised by an Advisory Council on Networking (ACN), and the work of creating and maintaining the networks is carried out by UKERNA (UK Educational Research Network Association)[9].

There have also been a number of recent policy reviews, including the SuperJANET Project on Information Resources (SPIRS),[10] and more recently the *Joint Funding Councils' Libraries Review*, [11] The Follett recommendations were accepted by the Councils and are now in process of implementation. A large number of the recommendations are concerned with the creation and delivery of electronic resources, and their implementation is being planned and carried out by the Follett Implementation Group

[7] For example, Chadwyck-Healey's *English Poetry Database* has a price in excess of 20,000 pounds per institution.

[8] The results of this survey have not yet been published.

[9] Formerly the *Joint Network Team* (JNT)

[10] This was a study group set up by the JNT and chaired by Michael Breaks. An edited version of its report is included in *Libraries and IT*.

[11] Follett, 1993.

on Information Technology (FIGIT).[12]

A number of projects and groups are active in developing network provision and access. Notable among these are the UK Office for Library and Information Networking (UKOLN), based at Bath University, and the recently formed special interest group Information Networking Alliance (INA), which represents librarians (through SCONUL), academic computing professionals (UCISA) and publishers (BIC).

2.1.5 The international dimension

Although, for practical reasons, the focus in our discussion is largely within the UK, there are many relevant institutions and projects outside the UK. There has always been a strong international tradition among scholars, which has been greatly facilitated by the ease with which students and researchers can now cross international frontiers electronically in their pursuit of digital resources. Indeed, there is a clearly emerging tendency for a small number of academic experts working within a comparatively narrow specialization but spread across several countries, to gain "critical mass" by working together using the Internet, thereby achieving results greater than the sum of the results they might achieve individually.

Alongside this the rapid and chaotic growth of the Internet has to be taken into account, comprising as it does a certain number of stable well-ordered data sets and services and very large volume of unregulated activity and information. This presents problems of its own; for it has become difficult to distinguish useful high-quality information from the low-grade material available side by side. ("On the one hand everything is available. But on the other hand, *everything* is available." - Alfred Glossbrenner). The difficulty of locating appropriate resources has become a common complaint.

2.2 Needs

The following sections present a number of specific issues in the provision of networked information. They formed a major component of the AHDS Workshop at the British Academy, and continued to be important throughout the Feasibility Study. Discussion is presented under broad headings chosen to reflect the "life cycle" of an electronic resource:

- resource creation;
- resource management;
- resource preservation;
- resource dissemination and discovery.

Within this framework, we identify the specific *needs* of the Arts and Humanities community, and try to summarize the *current situation* in each case.

2.3 Resource creation

The resources covered here include:

- Those created from scratch, and also those created by digitising existing paper resources, or by re-purposing existing electronic resources;
- New types of resource, such as electronic journals, or interactive (modifiable) materials;

[12] FIGIT's planned modus operandi is set out in JISC Circular 4/94 - *Follett Implementation Group on Information Technology: Framework for Progressing the Initiative.*

- Both research and teaching materials. In any case the distinction between the two may be far from clear, as when "research" resources are used in "resource-based learning" programmes.

All these resource types are covered within the FIGIT framework, in particular Programme Areas 2 - Electronic Journals, and 3 - Digitization.[13]

2.3.1 Academic quality.

Need Peer review procedures are needed which provide external validation for AHDS data sets. Without these the creation of such data sets, whether as resources for research or teaching, will not receive recognition from HE institutions or from the academic community as a whole.

Current situation Funding bodies such as the British Academy and the Leverhulme Trust carry out refereeing procedures when bids are assessed for projects which will create electronic data sets, as do a number of HE institutions when allocating funds internally for such projects. However, there are large numbers of data sets which are conceived and initiated without any such procedure. Even more important, perhaps, peer review of *completed* projects is relatively rare. It may be that such procedures should be external to the AHDS, but it is nevertheless a matter which the AHDS must take very seriously.

2.3.2 Copyright

Need First, it is important that copyright issues are addressed at a very early stage in the planning of individual projects. More generally, agreements and guidelines are needed at HE community level to provide a coherent and well-understood framework within which resource creation projects can operate.

Current situation There is discussion going on amongst publishers, and between publishers and HE authors and institutions. The British Library has a working party on Legal Deposit of electronic material which is of relevance here. The development of "a model agreement with publishers" is being addressed by FIGIT.[14] The WATCH project, with its aim of creating a database of copyright holders, is of interest in this area.[15] It is also likely that the Copyright Licensing Agency will make their database of copyright holders available on-line in the relatively near future.[16]

2.3.3 Design standards

Need There is a need for more widespread awareness and use of good design practice, both for research data sets and for teaching modules.

Current situation A number of data analysis and design methodologies are available in the commercial world which can be brought to bear on the development of research data sets.[17] There is also now an extensive literature concerning the design

[13] JISC 4/94.
[14] JISC Circular 4/94, Introduction, paragraph 18.
[15] See Appendix 8.3.1
[16] Source: private conversations with CLA staff.
[17] One example, of many, is *Structured Systems Analysis and Design Methodology* (SSADM). This widely used methodology was originally developed by Learmont and Burchett, and was subsequently adopted as a standard within the Civil Service. Most of these methodologies are aimed at administrative systems involving the dynamic flow of data, but there are elements which can be applied to data design in academic research projects.

of *Computer-Based Learning* (CBL) packages.[18] However there are many projects in both research and teaching contexts which do not use specialist designers or specific design standards.

2.3.4 Re-usability and access

Need Platform-independent standards for data encoding are essential to ensure transferability and re-usability of data. Such re-usability has academic benefits in terms of increasing the resources available for research and teaching, of allowing arguments and conclusions to be challenged by manipulation of a particular data set, and also, where copyright conditions of use permit it, of enabling resources to be "re-purposed" – i.e. modified or adapted to serve a new teaching or research purpose. Re-usability also has economic importance: high-quality electronic resources are expensive to produce, and it is important to ensure that as far as possible their re-use is a matter of academic judgment rather than of technical or financial constraint.

Current situation Currently platform-independence remains a goal rather than a reality for the most part. However, data encoding is an area of great activity and promise for all data types, with the work of the Text Encoding Initiative (TEI) offering particular promise (see Section 4). On the other hand, for certain types of resources, such as databases and CBL packages, platform-independence is somewhat distant, which means that for the foreseeable future AHDS Service Providers would need to cope with a number of proprietary software packages. Even so, it is often possible to ensure platform-independent *access* to the data, as for example in constructing relational databases so they can be manipulated by means of *Structured Query Language* (SQL).[19] In hypermedia work, for CBL and other purposes, the *HyperText Mark-up Language* (HTML) used by the World Wide Web is very popular, in part because of the platform independence consequent upon its use of the Standard Generalized Markup Language (SGML). [20]

Following on from the above, an increasingly important issue is whether the resource is to be accessed and manipulated on-line. If so, consideration has to be given to which access tools users will be able to employ. Among the most common network access mechanisms are *Gopher*, *World Wide Web* (WWW), and *Wide Area Information Server* (WAIS). Use of any of these systems involves additional encoding issues which should be considered when the resource is created.

2.3.5 Resource identification

This is the first in a range of documentation which is required if electronic resources are to be both accessible and usable, and which is best produced by the resource creator.

Need Standards are urgently required which enable resources in the rapidly growing Internet world to be uniquely identified and located. The need is far from simple, as for example in identifying resources which are replicated in a number of places.

Current situation The Internet Engineering Task Force (IETF) has a work group on *Uniform Resource Identifiers* (URIS) which aims to provide a standard for the unique identification of networked resources, by means of a *Uniform Resource Name* (URN)

[18] Jonassen, 1988 indicates the range of material available by 1988, and much more has appeared since then.
[19] ISO/IEC 9075 (1989).
[20] ISO/IEC 8879 (1986)

and a *Uniform Resource Locator* (URL). Where libraries and archives have on-line catalogues, the individual entries will follow the standards used by the institution.[21]

2.3.6 Resource description

Need Standards are needed for the electronic equivalent of full catalogue entries, to enable prospective users to make a judgment on the suitability of the resource to meet their needs without having to retrieve the whole data set. A considerable amount of technical information is needed, beyond the standard catalogue entries, from the details of the encoding scheme used to the software and hardware required in order to manipulate the resource.

Current situation Within the Library community there has been a great deal of work on resource description, mostly involving variations on the MARC standard, e.g. USMARC, UKMARC, UNIMARC. The "TEI Header" provides another possible approach, and the *TEI Guidelines* describe how the two approaches may be linked.[22]

2.3.7 Technical description

Need Standards are required so that all the technical information needed to understand and manipulate the data set is available. Documentation of this type is particularly important for certain types of data set, e.g. for complex relational databases where schemas showing the tables and their relationships are essential. It is also essential whenever proprietary software or hardware is used.

Current situation It is far from clear that technical descriptions are produced as a matter of course along with every new data set. A number of system design methodologies, e.g. SSADM, include documentation standards, such as "data structure diagrams", which can be very helpful. These could form the basis of guidelines for use by academic projects. A number of software packages enable relevant technical information to be embedded within the data set, as do mark-up systems such as SGML.

2.3.8 User Guidelines

Need Guidelines are needed to ensure first that users have available the necessary information on how to use the resource, and second that the way in which this information is presented is consistent from one data set to another.

Current situation It is widely accepted that while a paper version may be appropriate in some circumstances, help for users of on-line resources must be provided on-line. There are no standards as yet for the structure and presentation of such material, but hypertext approaches – often using HTML – are increasingly used.

[21] For example, the BIRON system provides an on-line catalogue of the data sets held by the ESRC Data Archive. However, since the catalogue is itself a networked resource, it too will require a unique network identifier.

[22] Part 4 of C.M. Sperberg-McQueen and Lou Burnard, Editors, *Guidelines for Electronic Text Encoding and Interchange (TEI P3)*, Chicago and Oxford, 1994. For a more detailed discussion of these and other approaches, see L. Dempsey, "Network Resource Discovery: a European library perspective", in N. Smith, Editor, *Libraries, networks and Europe: A European Networking Study*, London, 1994 (hereafter *Dempsey, 1994*).

2.4 Resource management

2.4.1 Basic procedures

Need Policies, standards and procedures are needed for accessioning data sets, documenting them where necessary, adding entries to the local catalogue(s) and indexes, and where appropriate to central catalogues.

Current situation All the established data archives have such procedures as a matter of course (as, of course, do all libraries).

2.4.2 Information content

Need Procedures are needed to assure the integrity of the resource against accidental or deliberate corruption. Where a resource is subject to amendment, *version control* procedures are required to ensure that users know which version of the resource they are using, and which versions are available for use.

Current situation Version control procedures are well established in existing data archives and university computing services. The *TEI Guidelines* include detailed proposals for embedding version control information within an electronic resource.

2.4.3 Intellectual history

Need Documentation procedures are required which will allow the usage of the resource to be recorded, along with comments by the users, in particular any reports of "errors" in the data. Clearly it would be improper for an archive to "correct" such errors silently, but equally it has a duty to report them to future prospective users. Over time, networked resources will inevitably accrue critical commentary (often in electronic form) in exactly the same way as printed resources. It would be desirable for at least the bibliography for this commentary to be included with the documentation for the data set.

Current situation It seems that most data archives, e.g. the ESRC Data Archive (EDA) and the Oxford Text Archive (OTA), have mechanisms to allow users to comment on the data sets they use (and the service provided by the archive). In the past these have been paper based, but electronic forms, bulletin boards and discussion groups are increasingly being used for this purpose.

2.4.4 Physical access

Need Procedures are needed which ensure that there is no unauthorized access to networked resources.

Current situation For some data sets and services there is no restriction on access. Where access has to be restricted, electronic "registration" is widely used, typically involving the allocation to a user of a "user identifier" and "password". Where necessary, passwords – and even complete data sets – may be encrypted. Usually, registration will entitle the user to the basic services offered by the institution. However, additional procedures are required where particular conditions are attached to the use of specific data sets, as described below.

2.4.5 Conditions of use

Need Procedures are needed which ensure that conditions of use imposed by creators, copyright holders, and the institution managing the resource are agreed and respected by users.

Current situation Typically, as part of the registration process, the prospective user signs a "Conditions of Use" form in relation to the archive generally. However, further conditions may be imposed in relation to specific data sets.

2.4.6 Intellectual Property Rights (IPR)

Need Procedures are needed which ensure that the rights of authors and other rights-holders are protected, as far as is within the power of the service. (See 4.4.2.)

Current situation This can be a complex area. At present it is covered typically within the registration and "conditions of use" procedures, with the user undertaking to cite all materials used for further publication in the appropriate manner.

2.4.7 Charging

Need Procedures are needed which can cope, potentially, with a variety of charging mechanisms. (See 4.4.1.)

Current situation There are few electronic charging mechanisms in current use. Typically charges are levied as part of the registration or "conditions of use" procedures.

2.4.8 Platform-specific data

Need Until such time as platform independence is uniformly enforceable, individual Service Providers will need to be able to manage data sets produced in a number of formats, and to handle both the variety of software common among their constituency of creators and users, and the hardware on which it runs. It is desirable if conversion between formats can be carried out where necessary. It would also be desirable for old non-conformant data sets to be "upgraded" to comply with defined standards, particularly where they are likely to be widely used.

Current situation Typically, existing data archives accept and manage data sets in a variety of formats. In certain circumstances data conversion may be carried out at the request of a user. The Oxford Text Archive has undertaken a phased programme of marking up its non-TEI conformant texts in SGML to the TEI standards, but its policy remains to accept and distribute material in any format.

2.5 Resource preservation

This is an area of particular importance to the funding bodies, who are anxious to ensure that expensively created resources will remain available for use after the project funding runs out.

2.5.1 Preservation procedures

Need Procedures are needed which can ensure the preservation of electronic resources indefinitely, until a decision is taken to destroy them or to allow them to lapse. Associated documentation procedures are also needed, e.g. to record the versions and formats in which a resource is held.

Current situation For all too many data sets there are no measures to ensure their preservation. For others, advantage is taken of the routine back-up and archive facilities offered by the local computing service. Where data is deposited with one of the established archives, there are usually more elaborate procedures in operation, with data sets held in several formats, and new copies made periodically.

2.5.2 Preservation media

Need Effective preservation procedures need to take account of the likely longevity of the media on which the data sets are to be stored, both in terms of physical degeneration and of the availability of the necessary hardware and software to decode the data.

Current situation There is considerable activity in this area among the Archive community. Until recently, the major preservation medium was magnetic tape, with new copies made periodically to protect against physical deterioration. Optical disc technologies are now increasingly widely used, but there are no universally accepted standards. The Council of European Social Sciences Data Archives (CESSDA) has initiated a project on preservation methodology, based at the ESRC Data Archive.[23]

2.6 Resource dissemination and discovery

"Discovery" and "dissemination" are very closely related. The emphasis in "dissemination" is on the steps taken by resource publishers to make materials available and usable; in "discovery" it is on the range of tools needed by users to access and manipulate the resources. However, the latter depends very directly on the former.

2.6.1 Delivery platforms

Need For the foreseeable future it is likely that a variety of delivery methods will be required. Guidelines are needed covering dissemination of resources in parallel: over the networks, on other electronic media such as CD-ROM, and on paper. Authors and funders will be reluctant to limit themselves to any single delivery platform while the capabilities of users and their equipment are so varied.

Current situation Mixed platform delivery is well established in existing archives, where data sets are available over the Internet, on CD-ROM, on floppy discs, and even on magnetic tape if required. Commercial publishers are increasingly involved, at least in parallel publishing on paper and CD-ROM; a number are also very interested in publication over the Internet.[24] FIGIT Programme Areas 1 - Electronic Document and

[23] Where the scale of copying required from an old medium to a new one makes the task too costly, it may be necessary for old hardware and/or software to be retained. The ESRC Archive, for example, retains a punched card reader in order to maintain the availability of certain of its older data sets.

[24] For a recent summary of electronic publishing in the UK, see David Brown: "Electronic publishing in the UK publishing industry" in *Libraries and IT*. Publishers of course have a particular interest in journal publication. Inter alia, Brown describes the work of the SPIRS Publishing Group and the intention to establish an *Electronic Journal Testbed*.

Article Delivery, and 4 - On-Demand Publishing are particularly significant in this context.[25]

2.6.2 Communications infrastructure

Need The networking of resources is possible only if a suitable communications infrastructure is in place, and valuable only if institutional infrastructures provide access to the resources for a critical mass of researchers, teachers, and students.

Current situation *SuperJANET* will provide a nation-wide high-speed, high-bandwidth platform, with links to other national and international networks, and capable of carrying the whole range of arts and humanities data types, including high resolution still and moving colour images. There are a number of SuperJANET demonstrator projects, including some from the Arts and Humanities, such as the *Remote Access to Museums and Archives* (RAMA) and *Van Eyck* projects.[26] UKERNA in the UK, and the IETF, are concerned with creating the basic platform, and also with the whole range of related issues, such as resource identification and resource discovery tools. It will be important for the AHDS to stay in close touch with developments in this area.

2.6.3 Catalogues and Indexes

Need The creation of comprehensive *metadata* resources is crucial if the networks are to be a successful method of dissemination. This involves not only on-line catalogues of resources, and catalogues of catalogues, but also searchable indexes of various types, including by subject, by author and by data type.

Current situation This has long been an area of considerable activity in the library community, and *On-line Public Access Catalogues* are well established. Many existing data archives have on-line catalogues and indexes. On a larger scale, in the US, the Research Libraries Group's Research Libraries Information Network (RLIN) holds a database of information on more than 56 million items held in over 200 member institutions, accessible by a search service called *Eureka*.

2.6.4 Gateways

Need Gateways are needed to provide simplified access to resources, saving users a great deal of time and effort. Busy academics need "one stop shops" which will provide access to all the resources they require, using a single interface or at worst a small number of consistent interfaces.

Current situation There is a great deal of activity in this area. One approach, originally intended for Gopher-based services only, has been to develop a *National Entry Point* (NEP). The UK NEP is currently based at the UK Office for Library and Information Networking, and several other European countries have followed this model. The *TopNode* project of the Coalition for Networked Information (CNI) is based on a similar concept.

Another approach has been to develop interfaces between gateways of different types. The *Common Gateway Interface* (CGI), for example, specifies interfaces between WWW and other services.[27] Within the last year the Working Group on Access to Networked Information Resources recommended to the ISSC that a *Network Information Development Agency* should be established to develop a range of National

[25] JISC 4/94

[26] See Appendix 8.3.2.

[27] A number of issues and projects in this area are described in Dempsey, 1994.

Entry Point services. On a more modest scale, the Humanities Bulletin Board (HUMBUL) has been transformed into the "HUMBUL Gateway" on the World Wide Web.[28]

2.6.5 Service reliability

Need If networked resources are to be utilized to their full potential, it is important that they form part of a well-supported service rather than being just a data set. Further, unless users can rely on the resources to be available, the services will soon lose their credibility.

Current situation A great many of the resources currently available on the Internet cannot be relied upon. Many exist only on an individual's personal computer, and are therefore unavailable if that machine is switched off or if it has a hardware, software or communications fault. Among the established data archives, the situation is a great deal better, of course. In many cases the file servers are mirrored, so that a fault with one does not bring the service down. It is also common to design communications systems with a degree of redundancy. Another benefit of redundancy can be performance. The DANTE InfoFlow project is based on a design of interlinked (national) data servers, with data replicated on two or more of these, to avoid access bottlenecks as well as to ensure a stable and reliable service (see Section 8.2.5).

2.6.6 Documentation, training and support for users

Need These are crucial if a service is to be provided rather than just a data set. Standards are needed to define the form which the documentation, training and support are to take.

Current situation Without exception, the Data Archives we contacted all stressed documentation, training and user support as being a particularly important aspect of the service they provide. Publicity material includes leaflets sent to Higher Education institutions or distributed at conferences. Other documentation may include bulletins and newsletters, made available on paper and on-line. Training includes self-help guides and scheduled courses and workshops, offered in-house, and at conferences and in HE institutions. Other aspects of user support are typically centred on a help-desk, but increasingly include electronic activities, such as an email query service, an electronic bulletin board, and electronic discussion groups open to all registered users. This group of activities is included in FIGIT Programme Area 5 - Training and Awareness.[29]

2.6.7 Resource discovery tools

Need Standard tools are needed for:

- resource location: identifying what resources are available and where;
- browsing: looking through the resource documentation, or in some cases the resource itself, in order to decide whether it is of further interest;
- retrieval: down-loading the documentation, or part or all of a resource, for local use;

[28] Something of this type could provide the starting point for a national Humanities Gateway, but it would be important for this to be developed within the framework of UK and international developments. Its *Universal Resource Locator* (URL) is `http://www.ox.ac.uk/depts/humanities/`.

[29] JISC 4/94.

- manipulation: in some cases it may be sensible to manipulate the data in some way before down-loading the result. This might be of particular importance with very large data sets, or with databases, or where other specialized manipulation software is available on the host equipment.

Current situation The current work on resource location (URIS) has been discussed briefly above. Browsing has also been discussed, in relation to *Gopher*, WAIS and WWW. The most common means of document retrieval over the Internet is *file transfer protocol* (FTP). There is considerable work being done based on the *Search and Retrieve* (SR) and Z39.50 protocols.[30] This work promises to provide tools for location and retrieval of complex multimedia documents. There are also an increasing number of SGML-aware browsers, so that cross-platform access tools will give access to platform-independent resources. In the area of databases, work continues on developing tools based on SQL, including *natural language interfaces* to SQL. This area of development is covered within FIGIT Programme Area 6 - Access to Network Resources.[31]

2.6.8 Documentation and training in resource discovery

Need Resource packs are needed which give new users basic guidance on what tools are available, and how they can be used to explore the world of electronic resources. On-line documentation also has an important role to play, to provide help once the user has succeeded in making a connection to the network.

Current situation There are a large number of books about the Internet, how to get started, and how to navigate around Gopher-space and the World Wide Web. The AHDS would not need to re-invent this material, but a simple introduction to electronic resources in the context of the AHDS would be invaluable in getting busy academics started.

[30] The SR protocol is ISO 10162/3. For more detailed discussion of access protocols and related issues, see Dempsey, 1994.

[31] JISC 4/94.

3 Scope and Structure of the AHDS

This section discusses the scope of an Arts and Humanities Data Service, in terms of the disciplines and data types to be supported, and the community of users served. This is followed by a description of the proposed constitution and organization of the AHDS, and a detailed discussion of its functions.

3.1 Disciplines and data types

The subject areas constituting the Humanities are potentially very large and diverse. They include the study of all languages, both individually and as linguistic systems, from all periods (modern, medieval and classical); they also include the study of an equally wide spread of literatures, as well as many other art forms, visual and plastic, performing and creative. The historical disciplines are at least as wide-ranging in their scope as the literary or linguistic ones, including all forms of historical text, data set, artefact and social structure. Any workable definition of the Humanities must also include such topics as theology, philosophy, religious studies, archaeology, art history, anthropology, social and political studies. Eventually, even purely scientific data sets become of interest to the historian of science.

There are many areas of overlap both within the Humanities, and between the Humanities and other disciplines, particularly but not exclusively the Social Sciences. The disciplines most involved would include History, Political Studies, Linguistics, Anthropology, Archaeology, and Geography. In terms of the kind of services under discussion therefore, the question of service overlap (and collaboration) must be addressed, especially with respect to services already funded by the various scientific research councils. For example, ESRC funds the ESRC Data Archive at Essex and NOMIS at Durham. Similar services in other disciplines are funded by the Engineering and Physical Sciences Research Council (EPRC), the National Environmental Research Council (NERC), the Royal Commission for Historical Monuments (RCHME), English Heritage etc.

As well as a richness and variety of disciplines, resources of interest to the Humanities are characterized by richness and variety of data types. At least the following may be identified:

- texts in all languages and orthographies, of all times;
- still images (bi-level bitmap, colour bitmap, three-dimensional);
- sounds (both naturally-occurring recorded sound and synthesized digital information intended to be played as sound);
- moving images (with or without sound, two- or three-dimensional, colour or monochrome);
- numerical data, both scalar (e.g. survey results) and vector (e.g. pictorial);
- symbolic data, for example describing musical scores;
- multimedia, which could include all of the above.

Some disciplines are primarily text-based, and others artefact- image-, or sound-based; all, however, typically combine resources of all kinds. The need to share expertise in the management of different data types across disciplines is correspondingly important.

3.2 The AHDS Community

The primary community served by the AHDS may be defined as comprising all institutions and individuals engaged in research or teaching in higher education in the UK, within any of the humanities disciplines as described above, and producing or using electronic materials of any type (as described above).

Although the emphasis must be on creating an infrastructure to serve the needs of this community, this can be done effectively only by collaborating actively with projects and institutions in wider Europe, North America and beyond. At a practical level this will have implications for the Catalogue and Gateway services of AHDS, and for the initial Service Providers. The opportunity to pursue formal organizational relationships and joint projects is of considerable importance.

In the same way, it is important to ensure that the community served by the AHDS can readily be extended to the wider UK community. Contacts with Local History associations, for example, have been found a productive source of expertise for the History Data Unit at the Essex Archive.[32] As free access to networked resources becomes less and less a prerogative of the academic world, the AHDS must be ready to provide its services to a wider public. The funding implications of this potentially enormous explosion in demand have yet to be thought through but it is clear that the AHDS will have a central role to play in establishing policy and procedures by which Arts and Humanities data services may be provided to other education sectors (in particular secondary education) and to the general public, for example in public libraries.[33]

3.3 Organization of the AHDS

The proposed Arts and Humanities Data Service comprises a centralized infrastructure, which acts as a framework within which a number of *Service Providers* support a wide range of *Services*.

The infrastructure will consist of a *Management Committee*, an *Executive* and a *Service Providers Forum*. The AHDS Management Committee will have the overall responsibility for the service. The AHDS Executive will be responsible for implementing the decisions of the Management Committee and, together with the Service Providers Forum, for formulating policy proposals for its consideration.

It is envisaged that initially the infrastructure will support a small number of *Service Providers*, each offering a wide range of services, and managing a number of *data sets*. The flexibility of the proposed structure means that in due course these may be joined by further full Service Providers and/or more specialized "Affiliated" Service Providers. The latter will typically provide services in relation to a single activity or data set, e.g. a *specialist collection*. It is anticipated that central funding will be made available to establish the service initially, but the proposed structure does not depend on any particular funding model.

The AHDS infrastructure has both strategic and operational functions, and the Executive will be responsible to the Management Committee for the proper performance of both function types. For convenience, the strategic functions are discussed in section 3.5, separately from operational functions, which are discussed in section 3.6.

[32] In Denmark, the Danish Data Archive is co-ordinating a very large *Data Entry Project* in which a small army of volunteers is being organized to encode in electronic form all extant Danish census, parish, land and military registers, for research use.

[33] As an example of what is possible we may cite the outreach work of the Norwegian Social Sciences Data Archive (NSD). This Archive co-ordinates mock elections in Norwegian secondary schools before national elections and referenda, analyzing and publishing the results nationally. Their *NSDStat* statistical software package is now used in over 80 percent of Norwegian schools, and in most entry-level University social science courses.

3.4 AHDS Constitution and procedures

This section proposes an outline Constitution and set of Working Procedures for the AHDS.

As outlined in Section 3.3 above, the AHDS will comprise:

- a *Management Committee*;
- an *Executive*;
- a number of *Service Providers* contracted to supply specified *Services*;
- a number of additional *Services*, each provided by an *Affiliated Service Provider*;
- a *Service Providers Forum*.

The AHDS Management Committee will be responsible to the funding councils (via the JISC) for the creation and operation of AHDS. In consultation with the Service Providers Forum, the AHDS Executive will also formulate policy recommendations for the Management Committee. The AHDS Executive and the Service Providers will be responsible for implementation of the approved policies.

The basis of funding proposed for the AHDS is a rolling contract, of between 3 and 5 years, with a periodic mid-cycle review.

3.4.1 AHDS Management Committee

The Management Committee will be nominated (by the JISC or one of its sub-committees), with the recommendation that one third of the members are nominated by the British Academy on the advice of its Humanities Research Board (HRB).

The Management Committee will be responsible to JISC for the AHDS, and will be responsible for receipt and disbursement of funds, and for ensuring the provision of the services to the proper standards. It will decide the policies of the AHDS, including the component services to be provided.

The Management Committee will normally meet twice a year, to receive reports and policy proposals from the AHDS Executive.

3.4.2 AHDS Executive

The Executive will consist of a full-time AHDS *Director* and a small number of full- or part-time support staff covering both the central technical and administrative aspects of the Executive's work. The exact composition and structure of the Executive should be the subject of bidding, but a number of important functions are identified below.

For practical reasons, the Executive should be established within some suitable host institution. The relationship of the Executive to its host institution will need to be specified, but the host would be expected to provide considerable institutional support.

The functions of the Executive will be:

- to service the AHDS Management Committee by preparing the papers for its meetings, communicating its decisions, and dealing with its other formal correspondence;
- to administer the service contracts entered into between the Management Committee and the Service Providers;
- to maintain close contact with the Service Providers, to provide technical support where possible, and to act as a focal point for information gathering and exchange within the component services of AHDS;

- to convene and service the meetings of the AHDS Service Providers Forum;
- to maintain contact with international institutions, in particular those in the European Union and in North America, who are engaged in activities relevant to the AHDS;
- to carry out the day-to-day operational functions of the AHDS, including:
  - preparation and distribution of AHDS publications;
  - arranging workshops and conferences and other events to provide information and training within the AHDS community;
  - operation of a help desk, to provide technical and operational advice to Service Providers in relation to AHDS policies, and to provide general information to the AHDS community.

3.4.3 Service Providers

Each Service Provider will be established on the basis of a contract from the AHDS Management Committee to supply one or more services to specified levels. Each Service Provider will have a Director or Manager who will be responsible to the Management Committee for the delivery of the contracted services.

We distinguish between *full service providers* and *specialist service providers*: the former will generally provide the full range of AHDS-standard services; the latter may provide only one such service. The services which may be provided are discussed more fully in section 3.8.

As part of its contract each Service Provider will be required to establish a *Specialist Advisory Committee* to provide independent guidance and advice. This Committee would normally contain representatives from the relevant academic, library and computing fields.

3.4.4 AHDS Service Providers Forum

The Service Providers Forum (SPF) will consist of one representative from each Service Provider. The Forum may co-opt additional members as considered appropriate. Meetings of the Forum will be convened and serviced by the AHDS Executive.

The primary function of the SPF will be to provide a regular context in which Service Providers can meet to discuss matters of common concern, and to collaborate with the AHDS Executive in the formulation of AHDS policy.[34]

To this end the SPF may commission reports, discussion papers and proposals from among its members, or from sub-committees or working parties.

The SPF will normally meet no more than three times a year.

3.4.5 Additional resources

There are likely to be some non-affiliated institutions wishing to make resources available to the AHDS Community, but unable to meet any level of the standards required of AHDS Service Providers. Where such resources are considered sufficiently valuable to the AHDS community, they may still be included in AHDS with a minimal level of support. Typically this will mean inclusion in the AHDS Catalogue with associated documentation which shows clearly the respects in which the resources fail to meet the normal AHDS standards. Where such resources are particularly valuable, it is likely that high priority would be given to requests for assistance in bringing them up to the appropriate level.

[34] If the AHDS is set up as a JISC-funded service, the AHDS Director will become eligible to participate in a forum for all JISC-funded academic service providers. This forum is convened three times per year to discuss matters of mutual interest, and to report to and receive reports from ISSC and JISC.

3.5 Strategic functions of the Infrastructure

The strategic functions of the AHDS will be carried out by the Executive, at the direction of the Management Committee. However, it will be important for the Executive to consult closely with the Service Providers Forum, and for the latter to play a significant role in the development of policy.

One of the primary functions will be to formulate long-term strategic policy with respect to the following fundamental requirements:

- initiation and reception of proposals for service provision;
- maintenance of appropriate consultation and debate with the AHDS community;
- methods of making and implementing policy decisions (for example, on the kinds of service to be supported);
- methods of making funding or other recommendations to external funding bodies;
- policies on service charges and cost recovery;
- policies on raising revenue from external sources, public and commercial.

An essential component of the strategic role will be the development of a framework within which standardized codes of practice may be developed and articulated. Such codes of practice are needed in all the following areas:

- data creation, preparation and encoding;
- data documentation and description;
- creation of catalogue entries;
- data preservation;[35]
- delivery of services;
- access and discovery tools;
- service level agreements, including suitable performance indicators;
- extension of standards.

At this strategic level the concern will be with the developments of *guidelines* within which specific standards will be adopted or implemented at the operational level, as further discussed in section 4. These guidelines should incorporate a considerable degree of flexibility to allow for different types of data or service, and differing local institutional structures. A "layered" approach to the issue of conformance will be taken, for at least the short term, in which suppliers can elect to conform at a specific level. Again, the definition of the exact nature of conformance at a specific level is an operational matter; the principle of layered conformance, and the extent to which it will be supported, is a strategic one.

In line with JISC policy for all the services which it funds, the AHDS will be required to establish evaluation criteria and procedural mechanisms to set in place both initial and continuation funding of Service Providers. It will also need to establish credible performance evaluation and review procedures and indicators.

[35] These will include not only standards on the number of security copies, frequency of update, and the location and security of the "vaults" in which they are held, but also the technical aspects relating to media type.

AHDS policy will be to attempt to establish a framework within which appropriate academic credit can be given for the creation of data sets, both for teaching and research, which conform to AHDS-specified standards, and which will place appropriate emphasis on their documentation, accessibility, and re-usability. Peer review procedures, perhaps external to the AHDS, will be required to assure the academic quality of the data sets. Creating this framework will involve both a close monitoring of the usage of AHDS data services, (inter alia, to ensure proper citation) and promoting greater awareness of the use and potential of electronic resources within the Higher Education community, in co-operation with the British Academy and its Humanities Research Board, as well as the Computers in Teaching Initiative and related initiatives.

In the development of its strategic policy the AHDS will need to work closely with relevant national initiatives and institutions. For example all of the programme areas of the Follett Implementation Group on Information Technology (FIGIT) will be relevant to the AHDS. Liaison with the JISC's Datasets Steering Group will be needed in the development of data acquisition strategies. The British Library has a number of current and planned projects which are pertinent to an AHDS,[36] and it will continue to be major partner in the development of networked information resources and services in the UK.

European and international dimensions will also be important in strategy development. In Europe, one particularly relevant project is *InfoServices*, a joint initiative in the Netherlands between the National Research Network Service and the National Library to develop a national information service. This project will be addressing many of the same strategic issues, in relation to documentation, standards and resource discovery which are being addressed by FIGIT, and within that framework, by the AHDS.

From the USA, the progress of the *Humanities and Arts on the Information Highways: A National Initiative*[37] is likely to have a significant impact on the way humanities resources can be interchanged between European and North American institutions.

3.6 Operational functions of the Infrastructure

The operational functions of the AHDS infrastructure will normally be carried out through the agency of the Executive. However in appropriate cases, some of these functions may be delegated by the Executive to an appropriate sub-committee or contracted out to a Service Provider.

All aspects of AHDS's operational role are covered either specifically or tangentially within FIGIT programme areas, or by other national projects or institutions. Once again, close liaison will be of vital importance if the AHDS is to contribute to and benefit from the ongoing development of a coherent national information structure.

We list below a number of key operational functions for the AHDS infrastructure.

- Monitoring of services provided:
  - assess proposals for new services (ensure appropriate specialist input or peer review);
  - allocate and review contracts for centrally-funded services (monitor conformance to standards, and quality of service delivery);
  - agree and review status of locally-funded services;
  - ensure that unnecessary or inappropriate duplication of services is avoided.
- Monitoring of standards applied:

[36] David White, *A list of Information Technology projects within the British Library* and T H Cannon, *IT research supported by the British Library R &D Department*, both in *Libraries and IT.*

[37] This initiative is sponsored by the Getty Art History Information Program, the Coalition for Networked Information, and the American Council of Learned Societies.

- monitor development of international standards in relevant fields;
- obtain and disseminate information about relevant standards to potential service providers or data suppliers;
- maintain the AHDS *Standards Reference Guide* (see section 4).

• Liaison with funding bodies.

• Collaboration with other data services, Humanities and non-Humanities, in the UK, Europe, N.America, world-wide.

• Liaison with network providers (e.g. UKERNA and IETF), and with network service providers, both academic (e.g. BUBL and NISS) and commercial (such as, for example, Bunyip Information Systems[38]).

• Negotiation of agreements on Intellectual Property Rights on behalf of the AHDS Community:

  - Collect and maintain information on current law and practice on copyrights and IPR;
  - Liaison with relevant bodies - e.g. Copyright Licensing Agency, Patents Office;
  - Liaison with other interested HE agencies, in order to develop common approaches, and share community resources;
  - Liaison with publishers;
  - Where relevant, undertaking negotiations with providers of copyright materials;
  - Provision of information and advice about IPR to resource creators.

• Negotiation of central licensing and charging policies on behalf of the AHDS Community, within the national HE framework:

  - Creation and maintenance of template licence agreements covering institutional and third party copyrights, and IPR of data set creators;
  - Liaison with other interested HE agencies, including European and N.American bodies;
  - Liaison with publishers and other commercial distributors;
  - Where relevant, undertaking negotiations with commercial distributors.

• Acquisition of core data sets:

  - negotiation, purchase or licensing;
  - inviting tenders for inclusion in existing service or creation of a new one;

• Data set "adoption agency":

  - identify homeless data sets worthy of preservation;
  - invite tenders/award contract for inclusion in existing service or creation of a new one;

• Support of AHDS service providers (current and potential):

  - preparation and maintenance of on-line documentation on the strategic framework and the operational rules and guidelines of AHDS;

[38] Current and planned Bunyip services are described in Dempsey, 1994, pp 16-17

- preparation and maintenance of on-line training materials where appropriate, for example in such areas as codes of practice, copyright, IPR, licensing;
- provision of mechanisms, including on-line services, for responding to questions, dealing with problems, addressing issues.

- Promotion of AHDS services:
  - preparation of publicity material about the AHDS;
  - arranging workshops and seminars to promote awareness and use of electronic resources by humanities scholars and students.
- Monitoring of AHDS usage:
  - Collection and maintenance of information on AHDS usage;
  - Collection and maintenance of information on AHDS users.

3.7 Centrally-provided services

Certain services are strategically important to the AHDS as a whole, and should be centrally managed. Chief among these are the *Catalogue* and the *Gateway*. There might also be benefit in having a single central *Data Vault* for the long-term preservation of resources.

3.7.1 AHDS Catalogue

A major AHDS priority should be the creation of a comprehensive on-line catalogue of electronic resources in the Arts and Humanities – an *Arts and Humanities Electronic Resource Catalogue*, with extensive and consistent indexes. Careful research is needed to identify the most appropriate mechanisms for gathering the information (a non-trivial task), and for maintaining it and making it available (a little less problematic perhaps once the AHDS is established). Some of the information necessary for this has been acquired, e.g. in recent surveys carried out by the British Academy and the Office for Humanities Communication. However, a great deal of systematic development work will be required.

It is likely that the central catalogue would replicate the catalogues at Service Provider or local institutional level, or would be regularly updated from these sources. However, it would be important to have a single agency within the AHDS charged with the responsibility of maintaining this co-ordinated service.

The Catalogue service provider would naturally have the role of liaison with the library and networking bodies involved in developing standards for resource identification and resource description, and of co-ordinating the AHDS approach to this.

3.7.2 AHDS Gateway

The functions of the Gateway would be to provide a convenient point of access to registered AHDS resources, and also to international and non-humanities resources. In relation to the first of these functions, it would make a lot of sense for the Gateway to be linked to the Catalogue, and there would be strong arguments for the development of the two as a single integrated service.

Going outwards, the primary links from the Gateway would be to other gateway services. In this context the Gateway service provider would have the role of liaison with other gateway providers in the UK, Europe, and elsewhere, and of monitoring developments in this rapidly developing field.

The Gateway service provider should also be very involved in the national and international work to develop resource discovery tools and common user interfaces, and should co-ordinate AHDS activity in this area.

An AHDS Gateway could also, if so desired, be used as a means to control access to AHDS resources, and could even provide a basis for central access charges.

3.7.3 AHDS Data Vault

Consideration should be given to the question of whether it would be more beneficial to have a single "Arts and Humanities Data Vault", rather than distributed preservation activities, carried out by each of the Service Providers.

It may be that the expertise required in the area of long-term archiving of data has more to do with technical and preservation issues than with data type or discipline, and that there would be significant economies of scale in establishing a single archival operation for the purposes of the long-term preservation of data sets. The project which the ESRC Data Archive has with the University of London Computing Centre (ULCC) is of interest, with EDA piping data to ULCC's Mass Storage system. The Research Libraries Group's *Archival Server Project* is also of interest.[39]

3.8 Distributed Data Services

In this section, we discuss in more detail the functions common to AHDS Service Providers. As noted above, not all service providers will necessarily perform all of these *data services* directly. They will, however, be expected to demonstrate an ability to support the majority of these functions, either directly or by subcontract, as a prerequisite when bidding for a service provision contract. (See further section 3.9.)

3.8.1 Accession of data sets

Under this heading we group such functions as the creation of new data sets and the acquisition of existing but unsupported data sets. In each case this may involve negotiation with IPR holders and other contractual arrangements with data depositors, which should be carried out according to norms defined by current AHDS policies. Where unsupported data sets, possibly of unknown provenance, are acquired, care should be taken that rights are not unwittingly infringed.

Where new data sets are being created, the principles on which they are selected and processed should be specified, and some indication given of the methods of quality assurance to be applied. In particular, individual Service Providers will seek to liaise closely with resource creators to ensure that relevant AHDS standards are applied *ab initio*. It may also be desirable to provide some degree of AHDS funding for such purposes, and for the AHDS to act pro-actively in facilitating the creation of specific resources where a clear demand has been identified.

Once acquired, data sets must be catalogued locally, and then incorporated within the central catalogue. Again, this should be carried out according to AHDS-defined norms. Cataloguing is a highly specialized (and expensive) task, and it would therefore be reasonable for individual service providers to seek the advice and support of subject specialists, provided that these are aware of AHDS standards. Some aspects of cataloguing require detailed subject expertise which may be more readily available locally. A staged approach may also be appropriate, in which a minimal level of cataloguing

[39] In February 1994, RLG announced a project "to specify the requirements for an Archival Server – a repository of digital images and full-text files – and to demonstrate the validity of the concept with a prototype implementation." The announcement is contained in a paper: *Preliminary Thoughts on an Archival Server*, 18 February 1994. The paper was included in a pack of RLG information sent to authors by John Haeger of RLG.

detail is provided for use by the central cataloguers, with more detailed information being added later from the local store.

3.8.2 Maintenance and development of data sets

Under this heading we class a number of functions applicable to electronic resources in general, but of particular importance to shared data sets:

- maintenance and error correction;
- long term security and integrity;
- enhancement and conversion;
- validation.

These functions apply equally to data resources and their documentation. Service providers will need to specify the level at which the resources provided will conform to AHDS standards and policies with respect to archival storage (for example, how many copies will be kept, the mode of storage, etc.) as well as to the correction of notified errors, the degree of encoding applied, the checking applied to ensure conformance to specific standards, and so forth.

As with cataloguing, long-term storage may be regarded as a specialized task contracted out to specialist agencies. All service providers must, however, take full responsibility for ensuring that the relevant standards are implemented to AHDS norms.

3.8.3 Data access facilities

Service providers may elect to provide one or more of a range of access facilities, ranging from simple file transfer only to fully interactive on-line browsing and manipulation. In the latter case, adequate documentation and support facilities (discussed in the next section) are particularly important. AHDS standards relating to "look and feel" of service provision should be applied, where these are defined.

Monitoring and control of access is also essential, for contractual and performance reasons alike. Service providers will be required to produce regular statistics on the use made of their facilities which will feed into the AHDS's own reports, and should form a part of any case made for continued support.

3.8.4 Support facilities

Under this heading, Service Providers will be expected to specify the degree of user training and support provided. This is likely to range widely for different kinds of data and different user communities. Telephone or email queries should be promptly and helpfully answered within defined norms. Introductory guides and tutorials, and specialist workshops or conferences as appropriate, should be organized. In each case, minimal service level agreements must be observed, and quality assurance measures enforced. Evaluation of take-up of the services offered is also a necessary part of service provision.

Because of the variation in user communities, and the corresponding importance of subject-specific knowledge, it is probable that these facilities will be best organized in close collaboration with specialist subject centres such as those offered by existing CTI and TLTP projects. Affiliation with such centres is not, however, a prerequisite for service providers.

3.9 Service Provision contracts

Service provision should be defined by contracts between the AHDS Executive acting on behalf of its Management Committee, and individual Service Providers. We anticipate only a small number (three or four) of such contracts in the short to medium term, but propose that a mechanism should be defined now to enable a graceful scaling-up of the operation.

In the long term, we suggest that there should be calls for proposals for service provision in specified subject areas, to enable the service to respond easily to changes within the field. These calls should invite bids which will be assessed by peer review, paying particular attention to the following criteria:

- existing or demonstrable demand for the resource being offered;
- existing expertise and experience or the capacity to develop and manage it;
- significant institutional commitment and support;
- demonstrable ability to support the full range of data service functions defined in section 3.8.

Bids may request funding for additional equipment or support staff required to provide access to the service from outside the host institution; in principle the host institution will be expected to bear all costs of providing internal access. Funding may also be requested for additional work needed to bring an existing service into line with specified AHDS standards.

Because it is assumed that infrastructure costs will be largely borne by the host institution, bids should also demonstrate that an appropriate level of computing, library, and other relevant support will be available from their host institutions.

We propose that contracts should be awarded for a minimum period of not less than three years, but should be subject to regular review. Full details of the funding model are, however, beyond the scope of the present report.

4 Standardization

We identify the following crucial areas for standardization: data format, data documentation, data provision, and data preservation.

data format concerns all aspects of physical storage, transmission and representation of information in electronic form. Examples include SGML for textual data, SPSS export or save sets for numerical data sets, CGM for graphical data, etc. We also include under this heading standards specific to data transmission (such as TCP/IP), user interfaces, etc.;

data documentation concerns all aspects of the provision of resource-discovery tools, cataloguing information, etc. Examples include the *International Standard Bibliographic Description* (ISBD) for bibliographic data, the *Standard Study Description* for social science data, the TEI Header, IAFA templates etc.;

data provision concerns aspects of service provision such as codes of practice with respect to copyright issues, availability and other licensing issues, fitness for use, level of advisory support and teaching provided, etc.;

data preservation concerns all aspects of the specific services and functions needed to ensure the preservation of resources without loss or degradation in their function or usability, effectively for ever.

In each case, a balance must be found between the need to ensure that resources remain usable over long periods of time and in many different contexts, and the need to adapt rapidly to changing technology (and fashions) in a rapidly developing field. Fortunately, this is a problem by no means unique to arts and humanities data users or providers.

4.1 Standards Reference Guide

The AHDS will need to identify and specify the standards considered relevant in each of the above mentioned areas. We propose that an early work item should be the production of a technical report, which will form the basis of a *Standards Reference Guide* (SRG) defining levels of standardization mandatory on all AHDS-affiliated data services and data sets. Important input to this process will be the FIGIT Framework Program for Standards activities. The SRG will of necessity be an extensible and changing document, but its core provisions should be relatively easy to define.

Some basic principles underlining the SRG can be stated at the outset:

- proprietary standards are in general to be avoided;
- unenforceable or unachievable standards are in general to be avoided;
- standards based on existing practice are to be preferred, where this does not conflict with other goals;
- defined levels of conformance to particular standards are useful;
- different standards may obtain for data formats in different subject specializations;
- uniform standards for data provision and data description should be definable across different subject specializations.

The existence of the SRG will help individual resource providers position their services with respect to overall AHDS policies, and hence to define meaningful service level agreements. It will also provide agencies concerned with the funding of such services with a set of appropriate performance metrics. Its definition is one of the most important of the AHDS strategic functions. Its continued maintenance will also be an essential task, requiring non-trivial effort for the foreseeable future.

4.2 Data formats

Standardization is currently well advanced in this domain, with many competing or overlapping standards in a number of areas. Fortunately, considerable expertise exists within the academic and industrial data processing communities with respect to data conversion and inter-operability of systems. Service providers and other appropriate groups of experts will be asked to provide short lists of recommendations, perhaps with differing emphases in different subject areas, as to the commonly-observed standards for data format.

4.2.1 Textual data

For the AHDS purposes, the most appropriate standard is likely to be that provided by the Text Encoding Initiative's "Guidelines" (TEI P3, 1994). This provides an internationally-accepted set of recommendations for the standardized encoding of a very wide range of textual and other data formats, based on SGML which is itself a widely implemented ISO standard (ISO 8879), of increasing importance in both commercial and academic publishing worlds. We know of no other equally relevant standard for the widest variety of arts and humanities data. The TEI recommendations have the unusual virtue of combining in one integrated framework a set of recommendations for data representation, data interpretation or analysis and data documentation, which is designed to be independent of any particular application area, software environment or hardware platform.

Although under development since 1989, the TEI "Guidelines" were published only in May 1994. Expertise and software capable of exploiting them is rapidly becoming available, in part because of an increased awareness of SGML itself within the academic community, but is by no means widespread. It is therefore important to ensure that large existing collections of data do not remain inaccessible to the AHDS community until they can be converted to use it.

We anticipate that there will be a considerable start-up cost associated with conversion of "legacy data", but that this cost will rapidly decrease.

4.2.2 Tabular data

For the rectangular data sets useful in certain types of research in the social and historical sciences, data in delimited or fixed format columns is widely accepted. Where information has been stored in databases based on the relational model, the export and transfer of data is also generally easy, although the transfer of associated descriptive meta-information may be more difficult.[40] For data sets stored in databases using other models, for example with multi-valued fields, data export and transfer generally pose serious problems, because each such database system will have its own proprietary internal format, and there may be no conversion or import/export facility.

[40] It has to be said that some of the earlier and some less well-featured "relational" software packages do **not** have good data export facilities.

4.2.3 Graphical data

Many different proprietary formats are used, each of which has its advocates. There is considerable inter-operability, because of a number of proprietary converters and because it is common for software vendors to support more than one format. This is, however, a rapidly changing and developing field.

4.2.4 Audio and other time-based data

Time-based data sets, both in the form of compressed audio or video and in symbolic form (such as an encoded representation of a musical score) will be of increasing importance as equipment capable of managing them becomes more readily available. At present, handling such data sets appears to be a highly specialized activity, although audio and video "clips" are becoming a normal part of the material presented in on-line services such as the World Wide Web, and have obvious academic uses.

The AHDS will need to consult with appropriate specialist groups, at the British Library National Sound Archive, the Center for Computer-Assisted Research in the Humanities (CCARH) at Menlo Park, California, and elsewhere, to identify de facto standards in this area and integrate them into its framework. At the time of writing, for example, the CCARH is about to commence distribution of a non-commercial basis of large music data sets comprising a significant proportion of the works of classical composers such as Bach, Handel, Mozart, and others, derived from out-of-copyright scholarly editions. However, the music will be issued in a proprietary binary format (*MuseData*) together with software tools for data-retrieval and processing.[41]

The commercial protocol Musical Instrument Digital Interface (MIDI) is a very widely-used binary system for control of musical devices, which has an associated *Standard MIDI File* format (SMF). While this is a de facto standard for musical performance, it has well-recognized drawbacks for the symbolic representation of scores.[42]

SGML-based initiatives may be of more significance in this connection, in particular those based on HyTime, which is an ISO standard (ISO/IEC 10744:1992) for the encoding of time-based data, specifically for multimedia purposes.[43]

The Standard Music Description Language (SMDL) is a specialized subset of HyTime concerned with the encoding of symbolic representations of music.[44]

Standards for data compression are relevant both to image (e.g. JPEG) and time-based data; the AHDS will have little choice but to follow industry trends.

4.3 Data documentation

We use this term to refer to the general topic of standards and codes of practice relating to meta-information, that is (in the most mature case) bibliographic standardization such as that embodied by the International Standard Bibliographic Description, well known in library circles, or the Standard Study Description, equally well known in social science. We also include under this heading consideration of the indexing techniques used to facilitate access to image and sound data.

Conventions for the documentation of networked electronic resources are still in a state of ferment, despite a number of years of discussion both within the library and information science community and outside it. It is only a slight exaggeration to say that at present the quality of information available in "gopher space" is indirectly proportional to its quantity: never before have so many known so little about so much.

The most promising basis for an improvement in this situation will be a careful examination of existing standards for bibliographic description, and their relevance to

[41] Selfridge-Field, 1993-4
[42] Sitter, forthcoming.
[43] Newcomb, Kipp and Newcomb, 1991.
[44] SMDL is as yet still in draft form (ISO/IEC DIS 10743); see Sloan, 1992.

the description of non-book materials. Relevant standards include AACR2 (1988), and ISO 690 (1987). Monitoring of the development of more electronically-relevant standards such as MARC will be of great importance, and is likely to have considerable impact within the library community. The recommendations of the TEI for formal documentation of electronic texts are also of great relevance here, since these are explicitly modelled on the ISBD.

The identification and approval of any pre-existing descriptive thesauri appropriate within given subject areas will be of considerable assistance to subject-specific resource providers. The high cost of building and maintaining indexes based on such thesauri should be set against the enormous increase in data accessibility which they typically facilitate.

This is particularly true for the description of visual images in verbal or encoded form, without which large-scale image databases are likely to be entirely unmanageable. Natural language descriptive thesauri (similar to those used for other data sets) are widely used, for example by the Getty Foundation's projects. For representational art in the Western European tradition there also exist abstract descriptive schemes such as *Iconclass*.

Wide consultation involving creators, users and subject specialists will be necessary to ensure that data documentation supports both informed and responsible data usage. The AHDS will need to formulate clear guidelines as to the level and scope of data documentation which is essential, desirable, or useful, for each of a variety of resources. It should also provide helpful guidelines for those seeking to create documentation tools, such as database proformas, or could initiate and manage projects to develop such tools.

4.4 Data provision

There are few pre-existing formal standards of relevance in this area. We expect therefore that the AHDS will begin by tabulating and codifying existing codes of practice with a view to agreeing levels of conformity amongst all AHDS service providers. We propose that an effective way of carrying this out might be to set up small task groups to report on specific areas in which individual practice tends to diverge. These task groups should be composed of a small number (five or less) of experts, not necessarily involved with the AHDS, who might be required to report back with detailed proposals for standardization within a reasonable time scale (say nine months).

Most of the areas of concern common to all AHDS service providers were discussed above in section 2. Specific topics on which we anticipate that such reports would be particularly useful are:

- Charging and licensing policies;
- IPR Policies;
- Service level agreements;
- Maintenance and error correction.

Each of these is discussed in more detail below; others may be added as the need emerges.

4.4.1 Charging and licensing

It has to be recognized that there are always costs associated with the creation and maintenance of electronic resources. It is not unreasonable to seek to relate the value of such resources for subsequent users directly to these costs. However, some ambivalence exists within the academic community on this point. There is a widely-held view amongst information consumers that all services should be free, just as there is an equally widely-held expectation (often located in the same individuals) that information provision should be remunerated in some way. A variety of mechanisms have been developed to resolve this contradiction, of which the three most popular appear to be: *top-slicing* – where access to a resource is centrally-funded once for all, but appears to be free to the individual; *differential charging* – where free access to a resource by one group (academics) is effectively subsidized by the high charges made for access to the same resource by another group (commercial users); and *subscription* – where payment of a single fee provides unlimited access to members only. In the network environment other modes of charging have also been proposed, such as the notion of "pay by use" – where an automatic charge is levied for each access to a given network resource or group of resources.

It would be premature for the AHDS to **begin** by proposing any **particular** charging policy. However, it is an area of policy-making which should be tackled as a matter of urgency. Longer term, the AHDS will need to remain capable of responding flexibly to changes in public attitude, or even legislation, which are the ultimate determiners of such matters. All of that said, it is clearly in the interests of the AHDS community that as few barriers as possible be introduced between users and providers of data resources.

Service providers may in any case need to adopt different cost-recovery procedures. Some may need to recover only media costs; others may need to recover a proportion of the cost of creating and maintaining the resources within a defined period. In some cases, costs may be beyond the control of service providers, being determined by external agencies who make resources available to service providers on certain terms only.

In the case of commercially-funded resource provision, the existing expertise of bodies such as CHEST should be employed to negotiate satisfactory contracts for the AHDS community as a whole, rather than solely for the benefit of the richer members of that community.

Where resources are prepared specifically for the AHDS, clear policies regarding their possible commercial exploitation need to be set in place. There is growing evidence, for example in the TLTP programme, of the commercial viability of certain kinds of Arts and Humanities Data Resources: the expertise of bodies such as the Association for Learning Technology (ALT), of commercial publishers, and of academic industrial liaison officers should be sought to ensure that appropriate royalties are derived from such resources.

Where charges are made, for whatever reason, it is likely that differential rates may be applied to different users. One common model is to provide commercial access at a premium rate, royalties from which can then be used to subsidize academic access. However, it is not always easy to distinguish "academic" and "commercial" institutions or usage. Equally, if institutions not funded by the Funding Councils wish to access the service, it is likely that differential charging rates would be applicable.

Cost recovery for the service as a whole is greatly complicated by the fact that very large economies of scale may be obtained by localization of some key strategic functions such as data preservation and data description.

We recommend that, as a first step, existing charging policies of all AHDS service providers be collated, and if possible made consistent. Where no financial charge is made, but certain conditions are imposed (for example, as with the GNU public licence), the AHDS should attempt to define a legally acceptable formulation of the conditions. Where charging at some level is essential to the provision of a resource, guidelines for an acceptable level of charging, and mechanisms for its recovery from the community or elsewhere need to be defined. At this stage, it is probably only possible to identify commonly occurring practices and highlight the strengths or weaknesses of each. The proposed structure allows for cost-recovery and administration at data set, service provider or AHDS level.

4.4.2 Intellectual Property Rights

IPR considerations have always been closely related to charging issues: ownership of an intellectual property right is one way of justifying making a charge for it. However, the AHDS has a wider role to play in defining acceptable agreements with IPR holders than simply setting a fair price. It is widely recognized that this is a highly complex issue: current legislation in this area is both confused and confusing, with entirely different principles applying to information in different formats, or distributed to different countries. As with pricing, it is something which will need to be addressed as a matter of urgency by AHDS, working in close collaboration with FIGIT and other relevant initiatives, and monitoring closely the developments in European legislation.

It is clearly essential that the AHDS be seen to be behaving in a responsible way with respect to the rights of IPR holders. This implies, for example, that service providers should be required to maintain records of all those with IPR in their holdings, and that appropriate legal instruments be drawn up defining the nature of those rights and the licence under which the holdings are being made available. Definition and provision of such legal instruments should be charged to a small task group of appropriate experts.

4.4.3 User support and Consultancy

It is anticipated that the level of advisory support available for different data sets may vary considerably. At one extreme, it may be impossible to do more than supply documented data resources on a "caveat emptor" basis; others may readily provide detailed instruction in their use, complete with tutorials, self-help systems or casebook studies, perhaps even without actually supplying the data resources themselves. Most will fall between these two extremes. In line with other AHDS policies, Service Providers will be required to define the level of service which they can provide, enhancing it to minimal levels defined by the SRG where possible.

From the end-user's point of view, what is important is a clear statement of the level of service a given provider can reasonably be expected to provide. This might include simple "performance indicators" such as time taken to turn round a request for a resource, or to answer a query, or more imponderable qualitative matters such as the extent to which the service offered actually improved the effectiveness of a researcher's work or a teacher's productivity. In attempting to provide such information, the AHDS will simply be fitting in with the general pattern of performance monitoring now being required throughout academia.

4.4.4 Maintenance, error correction, quality assessment

Electronic data sets are not necessarily static objects. A few data sets are highly dynamic, changing from day to day, but even those which are supposedly authoritative may need maintenance or error correction. In the software context, users are accustomed to the notion of regular updates – whereby manufacturers bring out new versions in which existing errors are removed (and replaced by new ones). Similar considerations apply to data sets.

It is the users of data sets who are most likely to identify discrepancies or inconsistencies and they should be encouraged to report these for the benefit of all. Different service providers may have different policies as to whose responsibility it should be to act on such error reports, but we recommend that redistribution of such information should be a pre-requisite for AHDS service providers. Electronic discussion groups and mailing lists provide a natural and effective mechanism for this.

Thought should be given to ways of providing *qualitative* information about data resources (that is, some indication of the accuracy or reliability of particular resources) in ways compatible with the general principles of detailed data description discussed in section 4.3. The definition of acceptable levels of accuracy is an important area for AHDS service providers, given both the very variable quality of many present-day networked resources and also the desirability of allocating academic credit for the creation of such resources.

4.5 Data preservation

All AHDS service providers will need to arrange for the long term preservation and storage of their resources, either locally, or by delegating this function to some other service provider, within the AHDS or elsewhere. In either case, it is clearly essential that this archival function itself be responsibly and adequately carried out. The objective is to ensure that resources can be recovered, without loss of functionality, and made available on computing or other information processing equipment at any time in the foreseeable future.

In general the term *archive* implies "for ever"; all archivists know, however, that policies allowing for some culling (or removal) of unused or duplicated resources at some time are necessary. There is no reason to believe that electronic archives will be any different from others in this respect, except perhaps quantitatively. To some extent, the existence of properly defined and agreed quality assurance procedures for the accession of AHDS resources may reduce the urgency of the need to define such policies, but their definition, and a clear statement of the criteria underlying them, will still be necessary at some point. In this respect, the AHDS will naturally look to the wider library and academic communities for guidance.

Unlike the archiving of documents or other artefacts, the archiving of electronically-held information resources requires a separation between the medium and its content which may not always be easy to carry out. Hardware changes probably make impossible the indefinite preservation of software. However, there is ample evidence that data sets can be preserved for almost indefinite lengths of time simply by establishing routines for the migration of the data from one carrier medium to another, without loss of information.

Because this task of archival storage is likely to be delegated in many cases to specialist agencies, it is correspondingly important for the AHDS to specify minimal operational requirements for such agencies. The nature of these specifications will also need constant review as technology changes – a standard based on the use of magnetic tape might no longer be appropriate in five years time.

Some aspects are unlikely to change, however: these include considerations such as the number of copies held, their locations, the nature and degree of integrity checking carried out both across the resources synchronically and periodically, to check for degradation.

In general, economies of scale argue strongly for provision of archival facilities as a strategic AHDS service, which may be contracted out to specialist agencies offering a guaranteed level of service. Individual AHDS service providers may also be able to offer equivalent facilities to others less well endowed. The cost-sharing implications of this are considerable.

5 Implementation Strategy

This section provides a strategic overview of the way forward, on the assumption that the case for something like the model outlined in section 3 has been accepted.

It assumes that an Executive has been appointed, and a small nucleus of Service Providers has been selected, whether by open tender, by nomination, or by some other means.

The first step for this nucleus will be to review the constitutional and procedural issues discussed in section 3.4 and to agree on any necessary revisions or modifications to them with the Executive.

The next step will be to set in motion the process of defining the standards and procedures which will constitute a functioning Service. Since the success of the AHDS will be largely contingent on close attention to appropriate international and *de facto* standards, we discuss this process in some detail in section 4 above. Since the service levels and performance indicators included in the Service provision contracts will be related to the standards and procedures defined in this stage, it is anticipated that the initial contracts may be provisional, and that the final contracts with the initial set of Service Providers will follow this stage.

In addition to defining standards for data format, data description, data preservation and data provision, the AHDS Executive has to set in motion the creation of a unified Catalogue. One simple way of beginning this process might be to create a World Wide Web server linking information provided by each of the pilot service providers. However, in the long run the AHDS catalogue can only be provided using the expertise of the library community. Existing services, such as NISS and BUBL, and pilot projects such as CATRIONA, and other relevant initiatives will be consulted for assistance in defining a development path for this purpose.

5.1 Pilot Data Service Providers

As noted in section 3.1, the Arts and Humanities Data Service will address a very large and eclectic community, with very widely varying interests. It is correspondingly difficult to identify the minimum set of services which the AHDS should aim to provide, or to define a “critical mass” of resources and services which the initial implementation might aim to achieve within two years of start-up.

Identifying a core collection of data sets is a worse problem than the task of selecting a core collection for a new library, since its scope is not one but many institutions with varying policies for teaching and research. The present report therefore makes no attempt to identify core resources by discipline. Instead, we propose that the initial nucleus of service providers should be chosen in such a way as to maximize the spread of data types supported and disciplines covered.

At least the following subject areas and data types should be supported in the initial nucleus:

- historical documents and derived data sets;
- literary and linguistic textual materials;
- art-historical and archaeological images and artefacts;
- music, film, and other time-based materials.

In selecting service providers capable of setting up this nucleus, we expect that the AHDS Management Committee and funding bodies will choose to build on existing centres of expertise. However, such decisions form no part of the remit of the present report.

6 Financial and Costing Considerations

The following costs of the service can be identified:

6.1 AHDS Executive

It is envisaged that the staff will comprise a Director, a technical/research officer and an administrative assistant. All of these staff are required on a full-time and permanent basis. It is vital that the Director be someone of standing in the Arts and Humanities community and it is recommended therefore that the post is at professorial level. The costs of these three posts at current prices including national insurance and superannuation but excluding overheads will be in the order of 90 - 95K annually (about 45K, 27K and 20K respectively). In addition to this approximately 50K will be required annually for equipment, publications, travel and subsistence.

6.2 Consultancies

It is anticipated that during the first two years there will be the need for an additional budget to enable contracts to be provided to specialists to assist in the development of the methodology and procedures to be used by the data service. For example studies might be conducted on the implications of Intellectual Property Rights, or on charging systems. A budget of 50K a year for two years will be required to meet these start-up costs.

6.3 Service Providers

One of the Service Providers will have responsibility for the development, compilation and maintenance of the catalogue and for the operation of the gateway, whilst the others will have responsibility for the delivery of specified services. Service Providers might be very variable in their scale of operations but for the purpose of costing the service it should be assumed that they will each require two staff and some equipment and travel costs. The costs of this including national insurance and superannuation but excluding overheads will be of the order of 80 - 100K annually per Service Provider. The AHDS Executive may bid to be a Service Provider and may be a particularly appropriate candidate to provide the catalogue and gateway, provided it has the relevant expertise.

6.4 Other Issues

Three issues need special consideration in relation to the funding of the Arts and Humanities Data Service:

- whether the host institutions will be expected to make a contribution to the costs of hosting the Executive or a Service Provider and whether overheads will be provided. Although there will be benefits to a host institution in running a service it is essential that a national service is provided and that lack of adequate finance does not distort it towards local needs, nor that small institutions with specialist expertise are discouraged from bidding by the need to make a contribution. Local institutional support may be more important for the initial Service Providers than for services added once the initial service is established.

- whether non-Higher Education institutions will be eligible to receive funding to be a Service Provider. This is a critical issue since some of the organizations with the most expertise of data on different media are not in the HE sector.

- the development and implementation of a charging policy must be a high priority, and it must take into consideration the fact that some Service Providers will have greater potential for raising all or part of their costs in the long term than others.

7 Appendixes

7.1 Feasibility Study Proposal and Initial Workshop

7.1.1 Proposal

The text of *Feasibility study into establishing an Arts and Humanities Datacentre: a proposal to the Joint Information Systems Committee* is available by anonymous ftp from the Oxford Text Archive. Its Universal Resource Locator (URL) is `ftp://ota.ox.ac.uk/pub/ota/AHD/original.proposal`.

7.1.2 British Academy Workshop

The Feasibility Study was initiated with a workshop, held at the British Academy on 29-30 March 1994. It was attended by 28 people with expert interests in the subject, drawn from a wide variety of academic and professional backgrounds. Over 50 invitations were issued, so that just over half of those invited were able to attend.

The text of the Workshop report, *Towards an Arts and Humanities Data Service: Report of a Workshop held at the British Academy on 29 and 30 March 1994* is available by anonymous ftp from the Oxford Text Archive. Its Universal Resource Locator (URL) is `ftp://ota.ox.ac.uk/pub/ota/AHD/workshop.report`.

The Workshop participants were:

Mr Michael Breaks Library, Heriot-Watt University;

Mr Lou Burnard Oxford Text Archive;

Prof Averil Cameron Director, Centre for Hellenic Studies, King's College, London;

Mr Terry Cannon British Library, Research and Development Dept;

Dr Thomas Corns Dept of English and Linguistics, University of Wales;

Ms Shirleen Craig ISSC Co-ordinator, c/o Heriot-Watt University Library;

Dr Marilyn Deegan Office for Humanities Communication, Oxford University Computing Services;

Prof Roderick Floud Provost, London Guildhall University;

Mr Marc Fresko Management Consultant;

Miss Catherine Harbor Computer Centre, Royal Holloway, University of London;

Dr Richard Heseltine (Chairman) Librarian, University of Hull;

Dr Eddy Higgs Wellcome Unit for the History of Medicine, University of Oxford;

Prof Barry Ife Cervantes Professor of Spanish, King's College London;

Prof Howard Jackson School of English, University of Central England;

Dr Donna Kurtz Beazley Archive, Ashmolean Museum, University of Oxford;

Prof John Laver Chairman, Humanities Research Board, University of Edinburgh;

Prof Denise Lievesley Director, ESRC Data Archive, University of Essex;

Dr Cliff McKnight HUSAT Research Institute, Loughborough University of Technology;

Dr George Munroe Computing Centre, The Queen's University of Belfast;

Dr Andrew Prescott Manuscript Collections, British Library;

Dr Sebastian Rahtz Archaeo Informatica;

Dr Dominic Rathbone Dept of Classics, King's College London;

Mr Hugh Robertson Department of Humanities, University of Huddersfield;

Dr Seamus Ross British Academy;

Dr Kevin Schurer ESRC Data Archive, University of Essex;

Mr Harold Short Computing Centre, King's College London;

Prof Kathryn Sutherland Professor of English Literature, University of Nottingham;

Mr Andrew Wright University of North London.

7.2 Consultation Exercise

The Consultation phase of the Feasibility Study was initiated by publishing an announcement, on paper to a large but selected audience, and electronically as widely as possible. The announcement contained information about the Feasibility Study and invited responses from anyone who wished to comment.

7.2.1 Announcement of Feasibility Study

Paper Distribution Over 500 copies of the announcement and the Workshop report were sent on paper to:

- all University Librarians;
- selected Fellows of the British Academy;
- the Chairmen of the British Academy Research Projects;
- all those who attended the Feasibility Study Workshop plus those who had been invited but had been unable to attend;
- all those who attended the Elvetham Hall Conference (13-16 April 1994), both UK and Overseas delegates;
- all those who attended the Conference on *Telecommunications and the Visual Arts* in Ravello, 16-17 April (55 humanists and policy makers from Europe, the US and UK);
- all those who attended the summer 1993 workshop at the British Academy on *Electronic Information Resources and Historians*;
- all those who attended the British Academy seminar in June 1994 on *The Problems and Potentials of Electronic Information for Archaeology*;
- all those who requested copies of the British Academy/British Library report *Information Technology in Humanities Scholarship*;
- all the projects that contributed information to the above report.

Electronic Distribution Electronic versions of the announcement, the Workshop report and the original Feasibility Study proposal were made available for access by anonymous ftp from the Oxford Text Archive.

The announcement and the text of the Workshop report were sent to the following electronic bulletins boards:

- BUBL;
- HUMBUL;
- NISS.

The announcement was sent to the following electronic discussion groups, along with information on how to obtain copies of the Workshop report and the original proposal:

- Humanist;
- Lis-Link;
- Unite;
- Etextctr;
- DBFT;
- IR-L;
- PACS-L.

7.2.2 Responses

The following is a list of those who responded by sending comments and information (in Institution order):

Donna Kurtz and Thaddeus Lipinski Beazley Archive, Ashmolean Museum, Oxford;

Alison Northover Reader Services, Bodleian Library;

Peter Allan British Atmospheric Data Centre;

Alice Prochaska Director of Special Collections, The British Library;

Dennis Nicholson BUBL;

Patricia Crosett Dartmouth College;

Brian Allison Allison Research Index of Art and Design (ARIAD), De Montfort University;

Jeremy Fox School of Modern Languages and European Studies, University of East Anglia;

Nigel Birch Engineering and Physical Sciences Research Council;

T. Sorrell Dept of Philosophy, University of Essex;

Paul Ellison The SGML Project, University of Exeter;

Donald Spaeth CTI Centre for History with Archaeology and Art History, University of Glasgow;

Rick Trainor CTI Centre for History with Archaeology and Art History, University of Glasgow;

Gillian Barnes Humanities Division, Computing Centre, King's College London;

Roy Wisbey School of Humanities, King's College London;

Barry Ife School of Humanities, King's College London;

Martin Frost Department of Geography, King's College London;

John Lindsay School of Information Systems, Kingston University;

Michael Cook Archival Description Project, University of Liverpool;

Stuart Lee CTI Centre for Textual Studies, Oxford University;

Jeremy Black Oxford University;

Toby Bainton Library, Reading University;

Alasdair Paterson University of Sheffield;

Frank Colson University of Southampton;

T.M. Devine Department of History, University of Strathclyde;

Ian Budden Information Services, University of Sussex;

Lorcan Dempsey UKOLN, University of Bath;

Robin Alston University College London;

J.B. Trapp Warburg Institute.

7.3 Fact Finding

A number of institutions and individuals were approached by the authors during the course of the Feasibility Study, in some cases in person and in others by electronic mail. In nearly all cases a questionnaire was used as the basis for information gathering. In a few cases the questionnaire was not considered appropriate, but useful information about the institution or service was sent to the authors. There was also a range of additional material which the authors acquired along the way and found useful.

7.3.1 Visits

The following institutions and individuals (shown in Institution order) were kind enough to receive visits by one or both of the authors.

Susan Hockey Center for Electronic Texts in the Humanities (CETH), Princeton and Rutgers Universities;

John Sunderland Witt Library, Courtauld Institute of Art;

Cathy Gordon Witt Library, Courtauld Institute of Art;

Per Nielsen Dansk Data Arkiv (Danish Data Archives);

Hans Joergen Marker History Archive, Dansk Data Arkiv (Danish Data Archive);

Prof Denise Lievesley History Data Unit, ESRC Data Archive, University of Essex;

Michael Neuman Centre for Text Technology, Georgetown University, Washington DC;[45]

Peter Doorn Netherlands Historical Data Archive, Leiden;

Arne Svindland Norwegian Centre for Computing in the Humanities (NCCH);

Espen Ore Norwegian Centre for Computing in the Humanities (NCCH);

Per Vestbøstad Norwegian Centre for Computing in the Humanities (NCCH);

Bjorn Henrichson Norwegian Social Sciences Data Archive (NSD);

Vigdis Kvalheim Norwegian Social Sciences Data Archive (NSD);

David Seaman Electronic Text Center, University of Virginia;[46]

Claus Huitfield Wittgenstein Archives, Bergen, Norway.

7.3.2 Questionnaires

Completed questionnaires were received from the following individuals and institutions:

Donna Kurtz Beazley Archive, Ashmolean Museum, Oxford;

Donald Spaeth CTI Centre for History with Archaeology and Art History, University of Glasgow;

Sarah Davnall Manchester Information Data Sets and Associated Services (MIDAS), University of Manchester;

Stuart Lee CTI Centre for Textual Studies, University of Oxford.

7.3.3 Documentation

Extensive documentation was received from the following individuals and institutions:

Patricia Young Canadian Heritage Information Network;

Ann Okersen Coalition for Networked Information;

Josefien Bersee DANTE;

Susan Siegfried Getty Art History Information Project;

John Haeger Research Libraries Group.

7.3.4 Additional Material

Additional material was received from or in relation to the following:

- SURFnet, The Netherlands;
- Ashmole Archive, King's College London (from Hafed Walda);
- Association of Digital Publishers, Dayton, Ohio;

[45] Dr Neuman agreed to be interviewed during a visit to the UK.
[46] This visit was carried out on the authors' behalf by Alan Morrison, Oxford Text Archive.

- Abstracts for First International Conference, ELVIRA (Electronic Library and Visual Information Research);
- The GENIE project, Loughborough;
- GIS Digital Data Directory, Oxford Brookes University (from Paul O'Carroll);
- Special Collections, University of London Library (from Julia Walworth);
- INFOLINGUA, University of Montreal;
- Consortium for Lexical Research, Computing Research Library, New Mexico State University;
- Norwegian Documentation Project, The Faculty of Arts, University of Oslo, Norway (from Christian-Emil Ore, via Espen Ore);
- Corpus of Insular and Anglo-Saxon Illuminated Manuscripts, Purdue University (from Tom Ohlgren);
- WATCH project, Library, University of Reading (see Appendix 8.3.1);
- NetEc (Networking Interaction in Academic Economics), University of Surrey;
- Electronic Antiquity, University of Tasmania.

7.3.5 Consultation Document

The following set of questions was used as a basis for enquiry during interviews with particular projects and was also circulated electronically for comment.

1. Mission What is your centre's mission?

Does it support teaching, research or both?

In what respects do you think it excels?

With what other centres would you compare it?

What does it do which a conventional library does not?

2. Staffing and Establishment How many staff are employed at your centre (full/part time)?

What functions do the staff perform?

Is there a clear career pattern for the staff?

What is the relationship between the centre and its host institution (if any)?

To whom is the centre accountable?

Where, in broad terms, does funding for the centre come from?

What are the chief areas of expenditure?

3. Holdings What kinds of material do you hold?

What is your acquisitions policy?

Do you ever refuse material?

Do you purchase material?

Do you generate materials?

How are accessioning decisions made?

Who is responsible for documenting the materials you acquire?

Who is responsible for cataloguing the materials you acquire?

Who is responsible for archival storage of the materials you acquire?

4. Services What kind of user services do you currently provide or plan to provide?

What charges (if any) are attached to your services?

Who uses your services ?

Do you collaborate with other centres or projects?

5. Planning What major problem areas do you foresee over the next five years?

What major opportunities do you foresee?

7.3.6 Conferences

Project team members were also able to make informal contact with several relevant individuals and projects by attending some key conferences and symposia which took place during the Feasibility Study. Several of the ideas in the present report were first discussed, both formally and informally, with delegates at the following events:

- Second Elvetham Hall Conference on Scholarship and Technology in the Humanities: Networking and the European Cultural Heritage, Elvetham Hall, Hampshire, April 1994;
- *Consensus ex Machina*: the annual joint conference of ALLC/ACH, Paris, April 1994;
- SGML Europe, Montreux, April 1994;
- World Wide Web, Geneva, May 1994;
- CETH Workshop on Documenting Electronic Texts, Rutgers University, May 1994;
- SIGIR 94, Dublin City University, June 1994;
- Association for History and Computing International Conference, University of Nijmegen, August 1994;
- Computers and Teaching in the Humanities, University of Glasgow, September 1994.

8 Institutions and Projects

This section describes several institutions and projects around the world which provide services or which hold digital resources which are relevant to the AHDS. In section 8.1 we list some existing UK-based Institutions which already offer facilities similar to those proposed for AHDS Service Providers, to a greater or lesser extent. In section 8.2, we list some comparable Institutions in Europe and North America. Finally, in section 8.3, we list a few examples of current Arts and Humanities projects likely to be of interest to the AHDS community.

It should be emphasized that these lists are all intended for **illustrative purposes only**, and are **not** intended to be comprehensive. The services discussed are not all exclusively concerned with Humanities resources, and range from major well-established Services on the one hand to small-scale pilot projects on the other. Information here was provided by the Services concerned before or during the Feasibility Study, and has not been checked in detail.

8.1 UK Services

8.1.1 British Atmospheric Data Centre (BADC)

Overview This Centre provides data on the atmosphere and its boundaries to researchers working in the area of atmospheric sciences, global climate change and earth observation. Like others which hold data on other physical sciences (e.g. oceanography, solar-terrestrial physics, etc.), the Centre provides access to experimental data sets, many of which can be very large, and very quickly generated (for example, the digital data from satellite observations). Such data sets are thus rather different from those typically used by humanities disciplines, but other aspects of the service provided by such Centres are very similar.

The emphasis is entirely on research rather than teaching.

Key services:

- collect data, convert as necessary;
- catalogue data;
- archive data;
- answer queries about the data and about software used to process it;
- provide access to the computer holding the data via JANET.

Extent Data relevant to atmospheric science in its widest sense.

Source of funding National Environmental Research Council.

8.1.2 Computers in Teaching Initiative (CTI) Centres

Overview There are 20 CTI Centres in the UK, of which six are immediately relevant to the Humanities disciplines. The Centres, which are based in universities, promote and support the use of computers in the teaching of higher education subjects.

In some ways the CTI structure may be seen as a model for the proposed AHDS structure, with a number of autonomous discipline-related Centres, and a central support organisation – the Computers in Teaching Initiative Support Service (CTISS), based at Oxford.

Key services:

- production of resource guides (both in print and using network formats), newsletters etc.;
- workshops, seminars;
- visits to universities, demonstrations of software;
- provision of high level advisory services (e.g. on matters of policy);
- holding software and data set resources;
- contributing to standards;
- providing an information network.

Extent Each centre typically consists of two or three persons; their holdings of software and data sets can amount to some hundreds of items. The humanities-related centres cover the following subjects:

- History, Archaeology and Art History (Glasgow);
- Library and Information Studies (Loughborough);
- Modern languages (Hull);
- Music (Lancaster);
- Textual Studies(Oxford);[47]
- Geography (Leicester).

Source of funding Higher Education Funding Councils.

8.1.3 ESRC Data Archive (EDA)

Overview The Archive's primary mission is to promote wider and more informed use of data for teaching and research. As well as encouraging current use, there is a strong emphasis on long-term preservation. In January 1993 the Archive set up a History Data Unit, with funding from the ESRC for an initial two years. Some continuation funding will be provided by the British Academy and is being sought from other sources.

A particular feature is the BIRON catalogue. This catalogue covers most of the Archive's holdings, and has been built up over a number of years by trained librarians using a controlled vocabulary. The Archive has relationships with several other institutions, such as:

- MIDAS (see 8.1.5 below), which provides on-line access to the census and other large data sets on the Archive's behalf;
- Edinburgh Data Library for the distribution of certain data sets;
- Northern Ireland Regional Research Library, which acts as a conduit to users and producers of data in Northern Ireland;

[47]The Oxford CTI Centre also covers the humanities disciplines not covered by any other centres, e.g. Philosophy and Theology.

- NOMIS (Durham University), an agency providing local authority data sets, and more recently in a joint European Resource Centre.

Much data documentation is paper based, but work is under way to introduce electronic formats, and so far this has been completed for about 850 data sets.

The Archive currently holds the secretariat of the Council of European Social Sciences Data Archives (CESSDA), and the UK membership of the US-based Inter-University Consortium for Political Science Research (ICPSR). It is due to become a European specialist centre for data archiving methodology, under the CESSDA umbrella.

Key services

- Archiving and provision of access to data sets;
- cataloguing and documenting data;
- user support;
- training.

Extent Disciplines include social sciences, economics, history, geography, and law.

The Archive is one of the largest social science archives in the world. It has over 7,000 data sets, of which nearly 4,000 are fully catalogued and the rest are opinion polls. The number of data sets is increasing at a rate of about 250 per year.[48] Each year more than 2,000 data sets are accessed, and over 1,500 data catalogues are distributed.

Source of funding: ESRC, Essex University, JISC.

8.1.4 Global Environmental Change Data Network Facility (GENIE)

Overview This is a project initiated by the ESRC, which is funding and managing the facility on behalf of the Inter-Agency Committee on Global Environment Change (IACGEC). It is a consortium consisting of the University of Loughborough, University of Nottingham and Genasys II Ltd; it started work in April 1992. It consists of a centralised facility and a decentralised network. Of particular interest in this project is the use of knowledge engineering approaches to provide intelligent concept-based searches of materials held at all the participating institutions.

The primary tasks of the project are:

- the creation of a master directory and knowledge base serving enquiries from the natural and social sciences wishing to access UK-held data sets pertaining to global environmental change;
- the development and establishment, through bi-lateral collaboration and agreement, of a federal network which links existing HE institutions holding global environmental change data and provides a mechanism for researchers to locate data.

Key services:

- generate metadata at central level;
- allow advanced forms of access to data held throughout network.

[48] The Archive has an arrangement with the Economic and Social Sciences Research Council whereby all contracts which involving data collection specify that the data must be deposited with the Archive.

Extent The project is still in its initial pilot phase. The aim is to cover all the relevant materials held at all the participating institutions.

Source of funding Inter-Agency Committee on Global Environment Change and ESRC.

8.1.5 Manchester Information Datasets and Associated Services (MIDAS)

Overview This organisation exists for the installation and management of strategic research and teaching data sets for access and analysis by the UK academic community, and the provision of the services required to facilitate and support this. It specializes in the provision of on-line access to large-scale data sets and the software needed to process them for research and teaching purposes. Principal examples of the data are the 1981 and 1991 census data, and government survey data sets. Data sets are currently stored in software-specific formats. Software supported include SIR, Ingres, BRS/Search and SASPAC.

Access is available without charge to all *bona fide* requesters, subject to the copyright restrictions imposed by the depositors.

Key services:

- provision of data sets and software;
- advisory services and consultancy;
- training courses.

Extent A large number of the data sets managed by MIDAS are obtained via the ESRC Data Archive (see 8.1.3), and support services relating to them are run in conjunction with the Data Archive.

The data sets are relevant to the social sciences and some humanities disciplines.

There are currently some 800 users, largely staff and postgraduates, from about 100 sites. However, MIDAS is a new service, and the number of registrations increases daily. For its predecessor, there were several thousand users registered, from over 150 sites.

Source of funding JISC.

8.1.6 NetEc

Overview "NetEc" is a term that unites a number of projects for networked interaction in academic Economics. At the time of writing, there are three projects, called *BibEc*, *WoPEc* and *CodEc*. They aim to improve the communication of new research results in Economics via electronic media. Traditionally, new research results have been published in paper documents called "working papers" or "discussion papers". The aim of *BibEc* is to enhance the awareness of these papers in the academic community by establishing means to announce the publication of new papers via electronic media. The aim of *WoPEc* is to build a collection of electronic working papers in PostScript format, for free retrieval by anybody with Internet access. *CodEc*, which opened in June 1994, aims to build a collection of software routines which are useful in the study of Economics.

The effort is multinational in scope, with involved people in six countries, but no formal management structure. The project uses an internal bibliographic format referred to as BibEc format. Papers are stored as PostScript files, usually compressed.

Key services:

- bibliography;
- storage of documents,
- storage of software routines.

Extent Over 20,000 bibliographic references (late 1993), growing rapidly; small number of papers.

Covers academic economics only.

Source of funding Managed on a voluntary basis.

8.1.7 Office for Humanities Communication (OHC)

Overview This organisation aims to influence, and to provide a centre of expertise on, academics' use of computers in all aspects of their work (compare the CTI Centres discussed in section 8.1.2 below).

Key services:

- provision of consultancy and advisory services;
- research, publication of reports;
- maintenance of a humanities resource guide and gateway on World Wide Web;
- workshops, seminars, conferences;
- conducting surveys.

Extent The office, which consists of three people, covers all the Humanities disciplines in Britain.

Source of funding British Library Research and Development Department.

8.1.8 Oxford Text Archive (OTA)

Overview The OTA exists to serve the interests of the academic community by providing archival and dissemination facilities for electronic texts at low cost.

It offers scholars long term storage and maintenance of their electronic texts free of charge. It manages non-commercial distribution of electronic texts and information about them on behalf of its depositors.

The texts vary in format, quality and degree of mark-up. There is a gradual process of converting to standard formats. The texts are available for research and teaching, subject to the copyright limitations imposed by their depositors.

Key services:

- archiving and making available electronic texts;
- re-formatting texts (with TEI conformant tagging).

Extent More than 1,300 texts, mainly in the domains of literature and linguistics, in many languages.

Source of funding Funded by Oxford University Computing Services and a grant from the Text Encoding Initiative (TEI). It also makes a small charge to cover media services when copies are sent other than by Internet.

8.1.9 UK Office for Library and Information Networking (UKOLN)

Overview UKOLN provides services to the library and information community. It conducts, manages and publishes research, and also maintains a number of information resources (on networking issues, the Internet, bulletin boards, discussion groups, "gopher- space", the World Wide Web, WAIS, etc.). UKOLN also plans to set up a digital archive of library literature in the near future.

Key services:

- conduct and management of research on networked information;
- dissemination of research findings and other information;
- raising awareness of relevant issues, both for those in the community of interest and within funding bodies;
- maintaining the UK National Entry Point gopher and the Bulletin Board for Libraries (BUBL).

Extent Consists of seven full-time staff; areas of interest are:

- the use of networks by the library and information community;
- the management of bibliographic data.

Source of Funding British Library Research and Development Department and JISC.

8.2 Comparable services based outside the UK

8.2.1 Bibliothèque Nationale de France (BNF)

Overview This library is currently building a data service which will be one of the largest, if not the largest, in Europe. It will consist of a public library, which will succeed the Bibliothèque Nationale as France's major repository of printed books. However, it will also include a vast collection of books, images, sound and motion pictures in digital form. The digital resources will include Frantext and several other substantial collections. It is due to open in 1997.

Key services The library will provide facilities at its site for members of the public and accredited researchers to refer to digital documents. Large numbers of workstations, some specially developed, will be available for consultation of these resources. The number of workstations envisaged by the year 2000 is:

- 254 audio-visual workstations (some of these may be analogue rather than digital);
- 8 CD-ROM playback workstations;
- 12 multimedia workstations;

- 86 academic research workstations.[49]

The academic research workstations are particularly interesting. Named the Computer-Aided Reading Environment (CARE) or in French: Poste de Lecture Assistée par Ordinateur (PLAO)), these will offer sophisticated facilities for accredited researchers to use text and image documents. The facilities will include features such as navigation, temporary annotation, linking, searching, printing, extracting and temporary saving.

The library has negotiated with French publishers agreements to permit electronic storage and use of works still in copyright, thought the initial agreements are for a limited period. Perhaps for this reason, there are no plans to allow remote access to the digital collection; all access will be from the main site in Paris.

Extent The digital collection is due to include books, journals, and secondary literature, extending to:

- 100,000 books (average 250 pages each) stored in image form;
- 10,000 books stored in text form, possibly with some form of mark-up;
- 60,000 photographs;
- sound, motion picture and multimedia documents.[50]

The intention is to build a strong corpus of French literature and reference texts. The main domains identified for digital holdings are French literature, anthropology, philosophy, history, ethnology, economics, law and the history of science. The corpus is expected to contain both original texts and modern editions which include commentary on the originals.

Source of funding French Government.

8.2.2 Center for Electronic Texts in the Humanities (CETH)

Overview This center was established by Rutgers and Princeton Universities in 1991 and is funded by the National Endowment for the Humanities to act as a national focus within the USA for the creation, dissemination and use of electronic texts in the humanities.

Key services:

- Inventory of Machine-Readable Texts in the Humanities, currently held on RLIN – other formats are planned, and research on the documentation and cataloguing of humanities primary source material in electronic form;
- access to collections of SGML-encoded electronic text, the emphasis here being research and development of new access methodologies;
- educational programs, including an international summer school on Methods and Tools for Electronic Texts in the Humanities (co-sponsored by the Centre for Computing in the Humanities, University of Toronto);
- information services, newsletter, directory of electronic text centers. CETH was also responsible for establishing an electronic text center discussion group within the American Library Association;

[49] These figures are from February 1993.
[50] 1993 figures.

- Humanities Electronic Resources Center for Rutgers University (CD-ROMS etc) and electronic text resources in the Firestone Library at Princeton.

Extent Bibliographic database; other material is mostly primary source texts in the humanities.

Source of funding Rutgers and Princeton Universities, the National Endowment for the Humanities, and the Andrew W Mellon Foundation. A subscription-paying "consortium" of member institutions will start in 1995, and there are also plans for associated grant-funded projects.

8.2.3 Canadian Heritage Information Network (CHIN)

Overview This agency has 21 years of experience acting as an information service for the museums and heritage communities both in Canada and elsewhere. It creates national collection databases for humanities, natural sciences and for archaeological sites from data contributed by museums and heritage agencies across Canada. Beside these national databases, CHIN hosts more specialised databases related to the heritage community.

Holdings include bibliographic databases, directories, controlled vocabularies (for example the Getty Art History Information Program's *Art and Architecture Thesaurus*), indexes to artists, etc. These databases are created by other organisations and distributed and managed by CHIN.

Key services:

- provision of on-line databases;
- a communications network;
- consulting services in information management.

Extent The range of materials covered includes heritage information, archaeology, humanities, and the natural sciences.

Source of funding Communications Canada, an agency of the Canadian Government.

8.2.4 Dansk Data Arkiv (DDA)

Overview This Archive preserves and disseminates data sets in the domains of social sciences, history and medical research, almost exclusively from questionnaire-based surveys. Early holdings were in social sciences, the other domains having been added since 1985. On an exceptional basis, the archive may accept electronic texts outside the domain of history. It supports both research and teaching functions. The current level of use is rising "rapidly", but is currently approximately 250 requests, representing 750 data sets, per year.

Data is obtained from the following sources:

- private polling organisations;
- university researchers;
- government research agencies.

Note that deposit of data with the DDA is often a condition of grant from Danish research councils.

Key services:

- archiving of electronic data;
- provision of data sets for research and teaching;
- processing incoming information to transform it into one of the "standard" formats such as SPSS, SAS, comma- or tab-separated text;
- preserving data beyond the lifetime of its original storage medium;
- generating new materials (to a limited extent);
- providing advice to researchers on which holdings might be relevant to particular research. This is the most resource-intensive service, but is regarded as one of the most valuable;
- data analysis (fee-based);
- data documentation;
- training courses.

Extent Archives data from the following disciplines:

- social sciences;
- history;
- medical research.

About 1,000 data sets are held with complete documentation.

Source of funding Danish Ministry of Culture.

8.2.5 DANTE InfoFLOW

Overview DANTE was set up in 1993 by several European research networks. Its purpose is to provide advanced international computer network services for the European research community. One of its major activities is the management of EuropaNET. It is today a small organisation employing some ten people (but it is actively recruiting more).

In June 1994, DANTE prepared a proposal to set up an information server called *InfoFLOW* for the European research community, the main purpose of which would be to assist people to locate valid information sets. It is now recognised that further work is required before such a proposal can be taken forward. Work on the proposal is expected to continue, though there is no firm timetable for it.

If InfoFLOW goes ahead in the shape currently envisaged, it would have, potentially at least, a large area of overlap with AHDS. Specifically, the AHDS and InfoFLOW would each wish to hold metadata about the same data sets.

Key services InfoFLOW would hold metadata in a series of servers in different countries, with the actual data sets being held mainly by the data provider (i.e. on systems not under the control of DANTE). The proposal describes indexing by subject and location.

Extent The proposal envisages a service to serve all disciplines (not just Arts and Humanities). Users would refer to the metadata by WWW and Gopher unless and until other approaches displace these as de facto standards.

It is suggested that accession decisions will be taken by external subject experts contracted for this purpose.

Source of funding Not determined yet; likely to be drawn from infividual information providers, national networks and the European Union.

8.2.6 Netherlands Historical Data Archive (NHDA)

Overview This institution, which is located at the University of Leiden, provides services and resources for research and (to a lesser extent) for education. Its central aim is to collect, document, store and disseminate research files created by historians working within academic and official institutions.

It is also a centre of excellence for scanning and *OCR* of "difficult" materials (such as carbon copies); and it administers part of a postgraduate programme at the university.

Key services:

- holding and making data sets available;
- scanning and data conversion;
- documenting and cataloguing data sets;
- providing advice and chargeable consultancy.

Holds about 100 data sets, mainly structured social and economic history files.

Source of funding Approximately 50 per cent from projects of the Dutch Ministry of Education or National Science Association; the balance from other external projects.

8.2.7 Getty Art History Information Program (AHIP)

Overview This is one of seven operating programs of the J. Paul Getty Trust, based in Santa Monica, California. Its purpose is to make computer-based art-historical information accessible to scholars. It collaborates with a number of US and international organisations in a wide range of research and development projects.

Among AHIP's priorities are the development of standards relevant to art historians – e.g. of terminology, the extension of bibliographic resources, and the development of tools to aid art historical research – such as *The Authority Reference Tool*.

Key services Initiating and supporting programs in four major areas:

- Standards and Resources;
- Collaborative Projects;
- Bibliographic and Data Collection Projects;
- Technical Development and Research.

Extent Amongst major research resources developed by AHIP or with its support are:

- *The Union List of Artist Names*;
- *The Art and Architecture Thesaurus (AAT)*;
- *Guide for the Description of Architectural Drawings and Archives*;
- *Anthology of Museum Automation*;
- *Bibliography of the History of Art (*BHA*)*;
- *The Provenance Index*;
- *The Avery Index to Architectural Periodicals*;
- *Witt Computer Index*;
- *Census of Antique Art and Architecture Known to the Renaissance*;

Source of funding The J. Paul Getty Trust.

8.2.8 Norwegian Computing Centre for the Humanities (NCCH)

Overview This centre exists to provide information and consultancy services for Norwegian research projects involving the use of computers in the humanities. NCCH is also involved in research activities; it is particularly known for its expertise in corpus building.

Most work at present is in support of research; but it is felt that support for teaching activities will grow.

In the furtherance of its research activities, it has assembled a collection of digital resources. However, the Centre does not get involved with the commercial aspects of marketing and distribution.

Key services:

- consultancy on humanistic research projects;
- consultancy to museums;
- secretariat to the International Computer Archive of Modern English (ICAME);
- work on CBL projects.

Extent All holdings and activities relate to the humanities.

Source of funding Until 1992, the NCCH was funded directly by the Research Council of Norway. In 1992, it became a part of the University of Bergen, and is now jointly funded 50 per cent by the Research Council and 50 per cent by the University. Funding is guaranteed for the next five years.

8.2.9 Norwegian Social Sciences Data Archive (NSD)

Overview This is the primary archive service in Norway for the preservation of electronic research data. It started as an archive for social sciences, but is now active in other fields.

It supports both research and teaching functions.

Some data sets receive a high level of use; up to 400 accesses per month in one case.

Key services:

- archiving of electronic data;
- provision of data sets for research and teaching;
- services to the university research community and the research councils;
- acting in the role of "broker" between the scholarly community and the Data Protection Board in Oslo;
- documenting ongoing research and maintaining a database of Norwegian research projects;
- producing teaching materials, including software;
- pro-actively seeking data (based in part on monitoring the progress of research projects where deposit is a condition of grant);
- training courses, seminars and workshops; teaching methodology classes in universities;
- development of documentation standards.

Extent Archives data from the following disciplines:

- social sciences;
- medicine;
- history;
- the humanities;
- psychology.

Source of funding 60 per cent from the Research Council of Norway, balance from special projects and the provision of services.

8.2.10 Research Libraries Group (RLG)

Overview RLG was founded in 1974, as a non-profit alliance of Higher Education institutions, devoted to improving access to information that supports research and learning. It sees its major role as supporting and co-ordinating the development of co-operative solutions to the problems associated with the acquisition, delivery and preservation of research information.

Its structure is relevant to the proposed AHDS structure, in that has a co-ordination and support role among its constituent organisations, and centrally manages certain key services. One of its new projects is also significant – the *Archival Server Project*, which seeks to address most of the issues of resource management, preservation and access which will be fundamental to AHDS.

Key services:

- the *Research Libraries Information Network* (RLIN), an internationally available information management and retrieval system which is a focal point for the sharing of information resources between member institutions;
- a focus for more specific collaborative projects between member institutions – e.g. the shared resources program *ShaRes* for inter-library loans;
- information delivery – e.g. *CitaDel*, a service offering access to citation data, and *Ariel*, software for image transmission;
- information management tools for archives and museums – e.g. *Archives and Museum Information System* (AMIS);
- access tools – e.g. *Eureka*, RLG's "patron-oriented search service".

Extent RLG has more than 120 member institutions, including universities, archives, historical societies and museums.

The RLIN database holds more than 56 million items drawn from over 200 institutions.

Source of funding There are three revenue streams:

- service fees for use of RLIN (68-74 per cent);
- membership dues and fees for collaborative activities (12-16 per cent);
- grants and development contracts (12-16 per cent).

8.2.11 University of Virginia Electronic Text Center (UVA)

Overview An "Electronic Text Center" was founded at the University of Virginia in 1992. It exists to create and support a new broadly-based user community within the humanities at the University, and to establish the use of electronic texts as a mainstream resource for pedagogy and research in the University.

It has a diverse collection of humanities data sets (see below). Most are in text form, with a few in image form. The texts are encoded with TEI-conformant SGML. A key feature of the Center is that the same access software can be used across the collection, even though some of the documents come from CD-ROMs which include proprietary software.

Because of contractual obligations with the vendors who supply the texts and the search software, access to the on-line text service is restricted to University of Virginia students and staff.

Key services:

- development of common access methods and interfaces;
- access to a collection of humanities data sets;
- training.

Extent Originally the Center started up with commonly available humanities resources in English and Latin, such as the *Oxford English Dictionary*, the *Patrologia Latina* and a collection of texts from the Oxford Text Archive. This corpus has since been added to by data sets generated by the University's staff and students, and it is expected to grow continually in this way. English is the main language, but there are also texts in many other languages including French, German, Latin, Greek and Hebrew.

Source of funding University of Virginia, Alderman Library.

8.3 Example projects

In this section we list a number of projects in the Arts and Humanities which are concerned with the creation or dissemination of resources of potential relevance to the AHDS community. No attempt has been made here to undertake a systematic overview; instead we have tried only to indicate the range and variety of data sets currently or potentially available. The entries are grouped by broad project type.

8.3.1 Network and metadata projects

Georgetown Catalogue of Projects in Electronic Text (CPET) From 1989 to 1992, the CPET project at Georgetown University Center for Text and Technology compiled a catalogue of projects that create and analyse electronic text in the humanities. The result is a powerful database which includes information on electronic text projects throughout the world, though it is not wholly up-to-date. It includes a variety of information on many collections of literary works, historical documents, and linguistic data which are available from commercial vendors and scholarly sources.

Norwegian Documentation Project This project at the University of Oslo started in 1991, and is due to last six years. Its objective is the conversion of paper-based archives of the so-called collection departments at universities in Norway ranging from the Viking ship museum to the lexicographic departments.

The final product of the project is to be a networked national information system for the humanities, called "The Norwegian universities' databases for language and culture", which will integrate these diverse resources to allow (inter alia) multidisciplinary research. It will also allow for improved public access. An article by C. Ore, describing the project, is to appear in the Norwegian special issue of *Computers and the Humanities*, in 1995.

Writers And Their Copyright Holders (WATCH) This project is a collaborative venture between the University of Reading library and the Harry Ransom Humanities Research Center at the University of Texas, Austin. It started in April 1994.

The project aims to build a database of copyright holders, initially for literary works in the English language whose papers are housed in archives and manuscript repositories. The database is intended to be freely accessible via the Internet. Early work suggests that 8,000 authors should be included in the first phase, and that this may be achieved by 1995.

InfoServices This a co-operative project in the Netherlands between the National Research Network (SURFNET) and the Netherlands Royal Library. Its purpose is to manage a national information server, and in so doing to establish standards for good practice. It aims to develop organised access to Internet resources, taking over the Dutch *National Entry Point*, and experimenting with a subject-based approach to search hierarchies. It has links with the *SURFDOC* project, which will make technical and other literature available in a consistent way over linked servers.[51]

The Nordic WAIS/World Wide Web Project This project was funded by NordInfo and involved the Danish Technical Library and the Lund University Library. Its primary goals are to develop tools for automated processing of WAIS source files to produce classified indexes available over World Wide Web, and to develop better gateways between WAIS and World Wide Web.[52]

The Allison Research Index of Art and Design (ARIAD) ARIAD is a comprehensive bibliographic database covering all aspects of research in art and design. It is based at De Montfort University, and its early development was funded by a grant from the DTI. Currently the database is available electronically on discs in MS-DOS and Macintosh formats. However, there is great interest in making it available over the networks.

8.3.2 Art and Art History projects

This group is represented because of the increasing importance, to art historians and other scholars, of networked information, including high quality images.

Remote Access to Museums and Archives (RAMA) This is an experimental project which demonstrates network links between participating museums in six European countries. It allows images and text (primarily) stored in these museums to be found and viewed from terminals in the other locations. One of the museums also contributes video clips with sounds. At present, the project is at demonstration stage. It is hoped that permanent connections will be established at a later stage. It is funded by the European Union under its *Esprit III* initiative.

Beazley Archive This is an archival collection of photographs, notes and drawings relating to ancient Greek art located in the Ashmolean Museum, Oxford.

Currently based on text, the 50,000-record database is being enhanced to include images and video clips. Additionally, work is in progress to allow remote access to these enhancements. The text portion is already available to over 25 remote sites around the world via the Internet.

Corpus of Romanesque Sculpture This is a project at the Courtauld Institute of Art, sponsored by the British Academy, to create a complete record of the surviving heritage of sculpture produced in Britain and Ireland between c.1066 and 1200, by means of scanned monochrome photographs, and textual descriptions. It is intended to make available the images to visitors and wider distribution may also be considered.

Daidalos: Database of Ancient Greek Sculpture This project, funded by the Leverhulme Trust, is based at the Ashmole Archive, King's College London. Its purpose is to research and process the complete evidence for all known ancient Greek sculptors, including ancient texts, epigraphy, sculptures, coins, etc., and to link this mass of data together in a systematic way in order to create a major new research resource. It will include a large number of images, based largely on the photographs of the Ashmole Archive, but including also material obtained from a number of museums and other institutions.

---

[51] This information is taken from Dempsey (1994), Sidebar 6. A more detailed account will be contained in T. Van der Werf, "InfoServices: co-operation between the National Research Network Service and the National Library in the Netherlands", *Journal of Information Networking*, 2(1), 1994 (forthcoming).

[52] This information is taken from Dempsey (1994), Sidebar 5.

Witt Computer Index This is a major art history project funded by the Getty Art History Information Program and the Courtauld Institute of Art. It is based on the holdings of the Courtauld's Witt Library – photographs and illustrations of American and 18th-century British paintings and drawings. It is creating a comprehensive database of information about these works, using the *Iconclass* system to classify and describe each work. This makes it possible for information about the works to be accessed under a wide range of headings – by artist, date, and subject, and in ways not previously possible. Currently the database is entirely textual, and no images of the works are included.

8.3.3 Text and image projects

This group is represented because of the increasing number of projects which are "text"-based but where the inclusion of images, in particular of manuscripts, significantly enhances the usefulness of the resource as a research tool.

Archivo Digital de Manuscritos y Textos Españoles (ADMYTE) Based at the Biblioteca Nacional of Madrid, this project has the objective of capturing in digital form a large collection of manuscripts and incunabula. The result will include monochrome and colour images, and a modern interpreted version of the documents in full text, tagged with SGML. The project is running in parallel with the development of a dictionary of 15th-century Spanish.

Canterbury Tales Project This is a project, with major funding from the British Academy, to digitize in text and image form all 83 manuscripts and four pre-1500 printed editions of Chaucer's *Canterbury Tales*, together with some analyses. The project is in its early stages.

The project is taking place at the University of Sheffield and Oxford University.

Electronic Peirce Consortium This is a project to scan and transcribe the manuscript works of Charles Saunders Peirce, the 19th-century American philosopher. The project may grow to include up to 100,000 pages (depending on funding), though only 500 pages are available initially. The project is based at a number of institutions in the US, including Georgetown University, Brown University and Texas Tech.

Wittgenstein Archive This project is based at the University of Bergen, Norway. It is preparing all Wittgenstein's unpublished manuscripts (over 20,000 pages) in machine-readable and facsimile form, for which purpose it has also developed a sophisticated mark-up scheme, and software to process it.

Hartlib Papers The Hartlib Papers project, sponsored by the British Academy and the Leverhulme Trust, is based at Sheffield University. It is preparing a text and page-image database of the manuscripts of Samuel Hartlib, who lived in the 17th-century. The database will include some 20,000 leaves of manuscript, plus printed material in four languages.

Hartlib had the self-appointed task of collecting and disseminating knowledge; his surviving papers cover every aspect of 17th century intellectual life, including education, language and literature, philosophy, science, religion, politics and agriculture. The database will include hypertextual links between the page images and text transcriptions.

8.3.4 Language corpus projects

International Computer Archive of Modern English (ICAME) This is an informal international organization of linguists and scientists working with English machine-readable texts. Its aims are to collect and distribute information about English language materials available for computer-aided linguistic research, to compile an archive of English text corpora in machine-readable form, and to make material available to research institutions. The Norwegian Computing Centre for the Humanities acts as a distribution centre for ICAME materials.

British National Corpus (BNC) This is a collaborative project jointly funded by the DTI and SERC. It has created a 100 million word representative corpus of modern English of all kinds, spoken as well as written, tagged in SGML with word-class annotations. The corpus will be made freely available for research. The partners in the project are Oxford University Press, Longman UK, Chambers-Larousse, Oxford University Computing Services, the University of Lancaster, and the British Library Research and Development Department.

Corpus of Contemporary Spanish This project, based at King's College London, aims to create a corpus of 20 million words drawn from a wide range of sources and subject areas, and covering both Iberian peninsula and Latin-American Spanish. The first phase of the project – 5 million words of Iberian Spanish – is nearing completion.

8.3.5 Historical, musical and specialist interest projects

There is a very large number of historical projects, a growing number of music and music history projects, and a wide variety of projects concerned with highly specialist interests. One project from each area is described briefly below. The AHDS would help to ensure the wide availability of such resources to their academic audiences, both nationally and internationally.

Prosopography of the Byzantine Empire This is a major project of the British Academy, housed at King's College London. Its goal is to record all surviving information about every individual mentioned in Byzantine sources during the period from 641 to 1261, and every individual mentioned in non-Byzantine sources during the same period who is "relevant" (on a generous interpretation) to Byzantine affairs. It is intended to make the material available, with sophisticated search tools, both on CD-ROM and on-line.

Thesaurum Musicarum Latinarum (TML) This is an evolving database that will eventually contain the entire corpus of Latin music theory written during the Middle Ages and the early Renaissance. The project is run by a consortium of universities; the Project Office is centered at Indiana University, Bloomington. Special provision is made in the database for musical notation, both by encoding in a standard manner and by the inclusion of scanned images compressed in GIF format. The database is freely accessible via the Internet.

Syriac Computing Institute This Institute, at Cambridge University holds a number of texts about, and in, Syriac. These include the *Syriac Electronic Data Retrieval Archive*(SEDRA) and four New Testaments in Syriac (versions according to the Siniaticus and Curetonianus manuscripts, and Peshitta and Harklean versions), all of which are available for research use.

References

[1] British Library Research and Development Department and the British Academy: *Information Technology in Humanities Scholarship*, British Library R & D Report 6097, London, 1993.

[2] David Brown: "Electronic publishing in the UK publishing industry" in *Libraries and IT* ([26]).

[3] L. Burnard and H. Short: *Feasibility study into establishing an Arts and Humanities Datacentre: a proposal to the Joint Information Systems Committee*, 1994 (Available at URL `ftp://ota.ox.ac.uk/pub/ota/AHD/proposal.text`).

[4] [L. Burnard and H. Short]: *Towards an Arts and Humanities Data Service: Report of a Workshop held at the British Academy on 29 - 30 March 1994* (Available at URL `ftp://ota.ox.ac.uk/ota/pub/AHD/wkshop.report`).

[5] Special Issue on "Intelligent Agents", *Communications of the ACM*, Vol 37 No 7, July 1994.

[6] T H Cannon: "IT research supported by the British Library R & D Department" in *Libraries and IT* ([26]).

[7] Special Issue on "Information Retrieval", *The Computer Journal*, Vol 35 No 3, June 1992.

[8] L. Dempsey "Network Resource Discovery: a European library perspective", in N. Smith, Editor, *Libraries, networks and Europe: A European Networking Study*, London, 1994. (Available in RTF format at URL `gopher://ukoln.bath.ac.uk/00/Publications /ResDes/europe.rtf`).

[9] J. Foster (ed): *A status report on networked information retrieval: tools and groups*, work in progress, available at URL `ftp://nic.nordu.net/internet-drafts/draft-ietf-nir-status-report-03.txt`.

[10] The Getty Art History Information Program, The American Council of Learned Societies, The Coalition for Networked Information: *Humanities and Arts on the Information Highways: A Profile*, Getty Art History Information Program, 1994.

[11] *Joint Funding Councils' Libraries Review Group: Report*, HEFCE, 1993 ("Follett, 1993").

[12] E. Higgs and S. Ross (eds): *Electronic Information Resources and Historians: European Perspectives*, St Katharinen: Scripta Mercaturae Verlag, 1993.

[13] L. Hughes and S. Lee, (eds): *CTI Centre for Textual Studies: Resources Guide 1994*, Oxford, 1994 (ISBN 0 9523301 0 5).

[14] *Follett Implementation Group on Information Technology: Framework for Progressing the Initiative* (JISC Circular 4/94), Joint Information Systems Committee, 1994.

[15] D. Jonassen (ed): *Instructional Design for Microcomputer Courseware*, Lawrence Erlbaum Associates, 1988.

[16] S. Kenna and S. Ross (eds): *Networking in the Humanities: Proceedings of the Second Conference on Scholarship and Technology in the Humanities held at Elvetham Hall, Hampshire, UK, 13-16 April, 1994*, London: Bowker Saur, 1994. ISBN:1 85739 064 4.

[17] I. Lancashire (ed): *The Humanities Computing Yearbook 1989-90: a comprehensive guide to software and other Resources*, Oxford University Press, 1991.

[18] D. Law: "The development of a national policy for dataset provision in the UK: a historical perspective" in *Journal of Information Networking*, 1(2), 1994.

[19] D. Lievesley: "Increasing the value of data" in *Electronic Information Resources and Historians*, 1993 ([12]).

[20] *Standard MIDI Files 1.0* , The International MIDI Association, 1988.

[21] S. Newcomb, N. Kipp, V. Newcomb: "The "HyTime" Hypermedia/Time-based Document Structuring Language", in *Communications of the Association for Computing Machinery*, 34/11, November 1991.

[22] E. Selfridge-Field: "The MuseData Universe: A System of Musical Information", in *Computing in Musicology*, Vol 9, 1993-4.

[23] P. Sitter: "MIDI", in *Handbook of Musical Codes*, The International Musicological Society, forthcoming.

[24] D. Sloan: "HyTime/Standard Music Description Language and Computer-Aided Music Research", in *Computing in Musicology*, Vol 8, 1992.

[25] C.M. Sperberg-McQueen and Lou Burnard, (eds): *Guidelines for Electronic Text Encoding and Interchange (TEI P3)*, Chicago and Oxford, 1994.

[26] *Libraries and IT: Working papers of the Information Technology Sub-committee of the HEFCs' Libraries Review*, UKOLN, 1993.

[27] T. van der Werf: "InfoServices: cooperation between the National Research Network Service and the National Library in the Netherlands" in *Journal of Information Networking*, 2(1), 1994 (forthcoming).

[28] David White: "A list of Information Technology projects within the British Library" in *Libraries and IT* ([26]).