

EVEN THOUGH many have realised that the task of digitisation is not so innocently simple, the creation of electronic resources continues to be popular. Funding bodies may be suspicious of injecting large sums of cash to provide access to a small number of collections; and those managing projects may be wary of the problems that can accompany even the most straightforward of plans. Nevertheless, in their ability to inform and educate (and entertain) at a global level, digitised images, datasets and texts remain an incredibly powerful resource. As a result, there continues to be a large number of interested parties requiring advice on various topics related to digitisation projects.

The AHDS aims to provide an extended advice service for those developing such resources, and examples of the types of advice are noted in this edition of the newsletter. Mark Merry highlights the issues involved in putting historical datasets on the internet and Alan Morrison surveys some of the issues related to the process of OCR, Optical Character Recognition. Meanwhile, Catherine Hardman looks at three new Guides to Good Practice that have become available in recent months. Various workshops run by the AHDS are also examined, including a set of general digitisation workshops organised for autumn this year. These are featured on page 7.

The newsletter also looks at developments in other areas of AHDS work. The Performing Arts Data Service is updating its website and the result is a much more impressive web service, with new collections added (*see right*). The newsletter also has details of a project undertaken with the Library School at University College, London. Exploiting the intriguing acronym FAT, the project aims to improve mechanisms for classifying digital resources in the arts and humanities.

Contents of this AHDS Newsletter ...

Designing Shakespeare	1 + 8
Putting Resources on the Web	2
Looking at Optical Character Recognition	3
Good Practice Makes Perfect	4
PICTIVA: Recycling Project Experience	5
AHDS Gets to Grips with FAT	6
AHDS Digitisation Workshops	7

Designing Shakespeare

IN OCTOBER, the Performing Arts Data Service (PADS) unveils a new website with enhanced features that allow users to search and browse seamlessly across datasets for the first time. As well as a unified and attractive design, the new system offers much greater functionality for data depositors and users, as data can be presented in a variety of different views and managed in a flexible environment to ensure long-term access to key research resources.



“Poison I see hath bin his timelesse end:
O churle, drunke all, and left no friendly drop”

Photograph from Greenwich Theatre production of Romeo and Juliet, February 1998, and featured in the Designing Shakespeare resource. Image copyright Donald Cooper. Reproduced by kind permission.

One of the first new collections to be delivered online using this system is Designing Shakespeare, an AHRB-funded project developed by Dr Christie Carson and her colleagues from the Department of Drama at Royal Holloway College, London. The aim of the collection is to broaden the range of materials available for the study of Shakespeare in performance, and to support teaching and research by students, scholars and interested theatre audiences and practitioners.

This exciting collection offers over two thousand images of performances of Shakespeare's plays, drawn from three quite

continued on back page /

Putting Resources on the Web

PROJECTS concerned with capturing historical sources are increasingly keen to disseminate their digital results via the World Wide Web. Scholarly resources made available on a website enjoy a number of benefits, not least of which is the potential for wide and convenient access among the academic community. But there are also a number of issues which such projects should bear in mind from the very beginning if they ultimately intend to employ the internet to make their resources available.

The first and most important thing to bear in mind is that while webpages may be a suitable format

While webpages may be a suitable format for disseminating historical resources, they're not at all suitable for either capturing or preserving historical data.

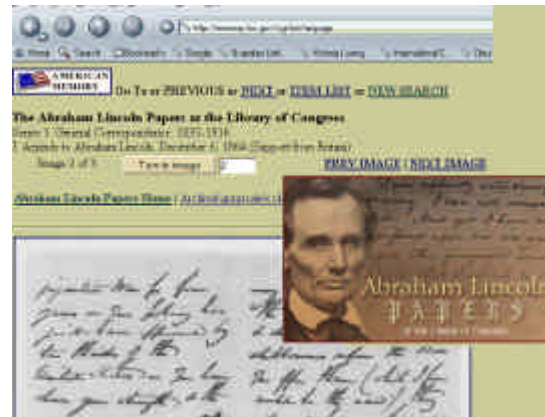
for disseminating historical resources, they're not at all suitable for either capturing or preserving historical data. HTML is a format designed to fit text and images into a structured layout so that the content of a webpage will appear in a certain way when that webpage is displayed in a browser. Unfortunately for most historians wishing to

use digital resources, a web browser is a very limited analytical tool, even taking into account recent advances in web technologies. In terms of preserving historical resources, websites are notoriously ephemeral even when hosted by university departments, and can often fall victim to changes in standards and the way that web browsers interpret HTML.

The solution to both issues, the lack of analytical capabilities and preservation, is to capture and store data in another format initially, and then translate it to HTML later. Building the resource as a database or a text archive, both areas where the AHDS can provide advice, allows the project the flexibility to prepare data for paper, CD-ROM and web-based publication later.

Of course the translation of data from database to webpage isn't a simple matter either. The simplest type of website to build is one that consists entirely of static pages, that is, pages with fixed content. In some situations this is quite appropri-

ate, for example, a critically edited collection of letters from a 19th century politician could be usefully organised on a website according to pre-defined criteria. But often of much more use to scholars is a site where the content is generated dynamically according to the preferences of the user. With the example above, if the collection of letters is large, it may be very helpful if users can locate particular letters by date, correspondent and topic. A dynamic website is a way of enhancing the resource created by a project, but it should be noted that linking a website to an underlying database requires specialist technical knowledge of various kinds of software and hardware, and this is something that a project should consider in its early stages. And of course, whether the website is static or dynamic, it should conform to web accessibility standards.



The Abraham Lincoln Papers on the American Memory website are a good example of a dynamically driven database - users are presented not with static html pages but dynamic pages that draw content from the underlying database.

<http://memory.loc.gov/ammem/alhtml/malhome.html>

The other issue that a project should bear in mind before setting up a website is that of rights. When publishing any kind of digitised version of a source on the internet, the project should be as confident as possible that they have the permission of the relevant rights holders to do so. Clearly this is something that should be obtained right at the start of the project. The flipside to this is that some form of control might need to be exerted over who can access the website, as once the resource is launched, it's in the public domain to a greater or lesser extent. Some projects may not feel the need to protect their resource by limiting access to certain groups, but this is a decision that needs to be made.

Mark Merry, History Data Service

Looking at Optical Character Recognition

Producing a digital copy or surrogate from an original paper copy is a process common to many digitisation projects, but the methodology and technology which underlie this process have remained largely unchanged in the past decade, and fall within two broad areas: Optical Character Recognition (OCR) and re-keying the material.



Projects often combine image scanning with OCR techniques. The Internet Library of Early Journals shows users images of the journal, but the search mechanism has been enabled by OCRing the originals to provide a full-text index

<http://www.bodley.ox.ac.uk/ilej/>

Optical Character Recognition can be a powerful technique for the rapid digitisation of printed materials. As personal PCs have become more and more powerful, OCR has become a realistic and affordable desktop application. But although it may seem like a simple matter to take a group of pixels and work out what shape they are, it is in fact an extremely intensive computational task, with vast potential for error. As a result, OCR packages will often only be guaranteed to work efficiently with a small subset of character fonts, and will be confused into making errors by defects in the original print quality. It also means that recognition of handwriting from images is still well beyond the capabilities of current personal computers. Modern Personal Digital Assistants (PDAs) which recognise handwritten input can do so because they specify how the user should form their letters with the stylus, and observe the creation of the letter as it takes place. Without these clues, human handwriting is simply too irregular to be analysed accurately using current systems. Handwritten materials, therefore, must still be digitised manually, by keying them in to the computer. OCR techniques are also extremely poor

at differentiating the 'layers' of text which make up many documents. For example, an administrative form may contain printed text comprising the form's original field descriptions, a typewritten layer of the information entered on the form, and an official stamp indicating that the form has been processed. Although some OCR packages, such as Omnipage, will attempt to unravel these layers, they are still extremely error-prone. In the case of complex layered documents, it may again be more efficient to transcribe the text by hand.

Another problematic area is that of non-standard character sets. Commercial OCR packages recognise only a handful of modern languages and character variations. If your material falls outside of this subset, you may wish to look into trainable OCR software. Training is commonplace in the world of software speech recognition, where individual voices vary. A user reads a set of words aloud to the program in order to 'train' it to recognise one's voice. Certain OCR projects have adapted this approach. An open source trainable OCR product called Clara <http://www.claraocr.org> is currently being developed. It promises to allow recognition of arbitrary characters, given some training input. Whether the time spent training the software would represent an efficient use of time overall depends very much on the quantity and uniformity of the material to be digitised.

The decision to use OCR must always be taken carefully, after appropriate pilot studies have been undertaken. If the work proves to be too difficult or too time-consuming, one option would be to contract it out. Whatever software you use, ensure that proper proof reading is done at all stages. Print that appears uniform to the human eye can vary greatly when it is analysed by the OCR algorithms. And finally, retain all the scanned images where possible. They are a resource in themselves, and there are many reasons why these images may prove useful in the future, for your own project or for other researchers. If nothing else, they can be extremely useful when testing new OCR software!

More information on OCR can be found in the Guide to Good Practice, *Creating and Documenting Electronic Texts*, <http://ota.ahds.ac.uk/documents/creating/chap3.htm> and also on the website of the Higher Education Digitisation Service, <http://www.heds.herts.ac.uk>

Alan Morrison, Oxford Text Archive

Good Practice Makes Perfect

While the AHDS is happy to receive telephone calls and emails to discuss aspects of a digitisation project, many of the principals that those undertaking digitisation need to be aware of are laid down in our series of Guides to Good Practice, available from <http://ahds.ac.uk/guides/htm>. The AHDS has recently published electronic versions of three new guides.

The first, *CAD: A Guide to Good Practice* written by Harrison Eiteljorg II, Kate Fernie, Jeremy Huggett and Damian Robinson and produced by the Archaeology Data Service (ADS), is based on the premise that many organisations create and hold digital project archives involving computer aided design (CAD) files and 3-D CAD models, digital objects that one cannot reproduce on paper. For this reason the CAD Guide to Good Practice is aimed at the creators of CAD files, primarily those produced in research. The guide provides a basic description of CAD software; a discussion of the use of CAD in a variety of situations; descriptions of data acquisition and capture methods; and information about archival practices. As well as providing a source of useful generic information, the guide emphasises the processes of long-term preservation, archiving and effective data re-use. As a result, the importance of adhering to recognised standards and documenting essential pieces of information about a given resource are recurring themes. The guide will be of use not only to those in Higher Education, but also national heritage agencies and local authorities; and those managing museums, National Monuments Records and Sites and Monuments Records. CAD is a foundation on which many mapping applications are built so this guide will also be of use to any scholars working with GIS.

The second guide, *A Place in History: a Guide to using GIS in Historical Research* by Ian Gregory, has been commissioned by the History Data Service (HDS). The guide is intended for historians who want to use Geographical Information Systems (GIS), describing how to create GIS databases and how to use GIS for historical research. The author defines GIS; evaluates the way GIS models the world; describes how to get data into a GIS; demonstrates the basic operations

that GIS offers to explore a database; reviews how time is handled; explains how GIS can be used for simple mapping and more advanced forms of visualisation; discusses quantitative data analysis within GIS; illustrates the use of GIS for qualitative analysis; and provides help on how to document and preserve GIS datasets. It also includes case studies from a variety of historical projects and an extensive reading list of GIS texts relevant to historians. A minimum of jargon has been used throughout, and no prior knowledge of GIS has been assumed. The aim is to highlight the main themes in using GIS and provide references that allow the reader to follow them up in more detail. There is an extensive literature on GIS in general but in terms of historical research this literature is limited and widely scattered. This guide attempts to bring this literature together to illustrate how historians have used GIS.



The two guides most recently produced by the Performing Arts Data Service have also been published in print by Oxbow Books. Interested customers can visit the Oxbow website <http://www.oxbow-books.com>, from where a range of AHDS Guides to Good Practice can be purchased.

The final guide is *Creating Digital Audio Resources*, authored by Nick Fells, Pauline Donachy and Catherine Owen, and intended as a basic 'how to' guide for those wishing to use audio materials in the creation of digital resources. It deals with such issues as copyright, choosing equipment, playing audio media, delivery of audio to users, data management and preservation. While not intended as an exhaustive guide and while it will almost certainly be necessary for resource creators to consult additional, more specialised sources, the guide provides a vital opening to the issues in developing digital sound.

Catherine Hardman, Archaeology Data Service

Recycling Project Experience: VADS & the PICTIVA Project

To complement the AHDS' core services of collecting digital resources and advising on related issues, a significant research and development programme of projects exists across all its service providers. The primary aim of such projects is often to produce particular end products, such as improved systems or resources for research, learning and teaching. However, the undertaking of such projects by AHDS service providers also provides additional benefits to the communities we serve. One such case in point is the Visual Arts Data Service's (VADS) JISC-funded Learning and Teaching project: PICTIVA - Promoting Image Collections for Learning and Teaching in the Visual Arts, the fruits of which are now becoming available.

The project has produced tools and resources to enhance the delivery of VADS visual arts image collections and these are going through final evaluation stages before being made available via the VADS website. New features on VADS image catalogue, produced through the project will allow users to:

- ♦ Save, manage, annotate and disseminate personal image-sets (Lightbox)
- ♦ Find visually similar images by colour, contrast and shape ('Content Based Image Retrieval' - CBIR)
- ♦ Engage with exemplar learning resources based on VADS collections, to illustrate the applicability of digital images to learning and teaching.

These have been developed to specifically benefit learning and teaching in the visual arts. However, a number of broader benefits to other activities and communities will result.

An example of this is in the applicability of the tools beyond the learning and teaching community to those engaged in research. The ability to save images into personally managed image-sets will allow researchers to gather, retain, annotate and distribute images according to particular interests and needs. This will also increase the potential for VADS systems to be utilised as a communication tool, rather than solely as an informational one, with the potential to stimulate debate and research.

The research community will also benefit from VADS' development of its online delivery systems to cater for the production and publication of learning materials. The same technology can be readily applied to produce other types of publications based

around material deposited with VADS, such as scholarly papers, and to facilitate the analysis of individual images, sets of resources or whole collections.

Similarly, the Content Based Image Retrieval tool will increase access methods for researchers as well as learners and teachers, enabling all users to better locate material of relevance to their interests, without having to rely on textual descriptions. Additionally, the feature will also provide an exemplary implementation of this cutting edge technology for researchers in the field to analyse. The visual search has been implemented by the Institute for Image Data Research, University of Northumbria and feedback will inform their continuing work in this area, as well as others doing similar projects.

This indicates another broad benefit of PICTIVA and project work generally. End products and findings can often be applied across subject boundaries and also assist interdisciplinary activities. This will principally be across other arts and humanities subjects, through the application of the tools to locate and utilise materials for various subject-specific needs, but can also extend to the technologies themselves being the subject of enquiry for a broader range of disciplines.

The final area of broad benefits of the project results from the investigation of new technologies and methodologies that VADS has undertaken in developing its systems. The PICTIVA project has seen VADS engaging in Human Computer Interfacing issues, such as good practice for navigating online resources intended for direct-user interaction; Learning Object Metadata for appropriate resource discovery and project management issues more generally. The lessons VADS have learnt add to its knowledge base and this expertise can be passed on through VADS advisory activities.

To summarise, much is learnt through project work that can inform an organisation's own future practices. But more importantly for an organisation such as the AHDS, expertise gained can be passed on to a range of users and interested communities.

Phill Purdy, Visual Arts Data Service

AHDS gets to grips with FAT

FATKS (Facet Analytical Theory in managing Knowledge Structures) aims to explore the potentials of facet analytical theory in building a flexible and modern controlled vocabulary for subject access to pools of distributed digital resources in the humanities.

FACET ANALYTICAL THEORY IN MANAGING KNOWLEDGE STRUCTURE *for humanities*

FATKS is a one-year project (2002), funded under the AHRB Innovations Awards scheme, and jointly undertaken by the School of Library, Archive and Information Studies at University College London, the AHDS and the Humbul Humanities Hub.

What is Facet Analytical Theory (FAT) ?

FAT is a method for building knowledge organisation tools such as classifications and thesauri. The method core is the organisation of concepts into categories (facets), providing a unified logical framework from which the development of structure and vocabulary is raised. The FAT logical framework provides general concept categories (facets such as entities, properties, processes, and operations, etc.) usable in almost every field of knowledge. When applied to a specific field of knowledge, the same logical framework can expand into facets specific to each field. For example, in medicine this would produce facets such as organs, diseases, therapy, persons etc.

Advantages of FAT

A controlled vocabulary built on facet analysis functions like a set of building blocks. Single concepts from different facets can be combined to express more complex subjects, where each building block is used in a clearly defined context, i.e. with an explicit 'role'. Providing subject indexing and access with logically structured context is a major potential of FAT, because it enhances better control over the filing, access, retrieval and presentation of concepts.

Beyond the flexibility in representations of the intellectual content itself, facet analysis also allows the definition of other distinct complementary facets such as type of audience (e.g. for children, for given professional groups), type

of content (dictionary, learning material, map, fiction etc.), type of information object (text, image, sound etc.) or carrier (paper, digital files), which are also important for end-user subject-based selections. All these elements are important search criteria and the flexibility provided by faceted vocabularies is particularly valuable in meeting the needs of subject indexing/retrieving of distributed sources in the digital environment, especially when different information fields are encompassed.

For the humanities, where interdisciplinarity is a major issue, building subject access is particularly pertinent. Resources in the humanities tend to show strong interactions among many different fields (e.g. the influence of philosophy on literature, computing on the humanities, building materials on architecture, and industry on design). Therefore, closed, simply enumerative or in any way restrictive vocabularies are very limiting. With a faceted structure one is free to choose how to organise, browse or display, for instance, historical resources: by place and then by time and then by form (maps or texts) and then by type of resource, or in some other sorting criteria that may prove useful for a given collection. Furthermore, a faceted classification structure can be used to generate default logical and hierarchical presentations.

FAT allows for a categorisation of arts and humanities subjects that reflects their strong interdisciplinarity.

Therefore, closed, simply enumerative or in any way restrictive vocabularies are very limiting. With a faceted structure one is free to choose how to organise, browse or display, for instance, historical resources: by place and then by time and then by form (maps or texts) and then by type of resource, or in some other sorting criteria that may prove useful for a given collection. Furthermore, a faceted classification structure can be used to generate default logical and hierarchical presentations.

Subject portal requirements

The project's goals are linked to the development of a future joint Humbul / AHDS Art & Humanities portal that may use powerful subject browsing and advanced searching across distributed digital collections. In order to meet these goals, the controlled vocabulary will be backed by a classification structure, allowing:

- ♦ Structured and flexible browsing
- ♦ Unlimited combination of concepts while indexing (e.g. pre-coordination)
- ♦ Expression of the relationships/roles of different concepts in compound and complex index entries
- ♦ Links between concepts in the classification structure and the natural language terms that are expected to be used in retrieval

Main foci of the Project

In order to make full use of facet analysis features, a vocabulary has to be adequately modelled and implemented for testing. Therefore, the project has three equally important foci:

1) *Building a knowledge structure*

Basic classification vocabulary (compiled from existing faceted classification schemes):

- ♦ A broad classification structure, universal in scope
- ♦ A faceted knowledge structure suitable for all the humanities fields
- ♦ A set of generally applicable facets, e.g. Place, Time, Materials, Person, Form, Languages, Properties, Processes

Controlled vocabulary for the humanities:

- ♦ A newly developed and full faceted subject vocabulary; target areas for the prototype will be religious studies, the visual arts and history

2) *Building a data model to hold the knowledge structure and vocabulary*

3) *Model implementation, and prototype testing and evaluation*

FATKS will provide a subject browsing structure for the humanities portal and underlying local collections. The structure is envisaged as a general knowledge representation 'umbrella' that may be later linked or mapped to vocabularies particular to each of the source databases encompassed by the portal, or to any other subject organisation tools relevant for the humanities (e.g. Dewey Decimal Classification or the Art and Architecture Thesaurus).

For more information please visit the FATKS website at <http://www.ucl.ac.uk/fatks/>

Aida Slavic, University College London

AHRB Research Centres

The last edition of the AHDS Newsletter featured an article (entitled *'Is a Picture Worth a Thousand Words?'*) that mistakenly attributed the Central St Martins Study Collection to the AHRB British Cinema and Television Study Centre. The correct appellation is the AHRB Centre for British Film and Television. Apologies to all concerned.

AHDS Digitisation Workshops

The Arts and Humanities Data Service regularly run workshops dealing with the key issues in creating or exploiting digital resources.

The flagship series consists of AHDS Digitisation Workshops, organised for those who are planning, beginning or running digitisation projects. Presentations deal with the issues that need to be addressed, including project management, data capture, metadata, depositing and delivering collections, and copyright. There will also be case studies from a wide range of subject areas, and the opportunity to discuss individual projects with AHDS staff. The workshops are open to any institution or individual who is planning or running a digitisation project, but they may be of special interest to those who are applying to one of the funding schemes organised by the Arts and Humanities Research Board or the Joint Information Systems Committee

Locations and Dates for **AHDS Digitisation Workshops** (the London event is already fully booked)

- ♦ Glasgow - 30th October
- ♦ York - 11th November

See <http://ahds.ac.uk/workshops.htm> for more details and a booking form. More digitisation workshops will be organised for spring next year.

The AHDS is also planning to run other workshops. In January 2003, there is due to be a workshop on **copyright issues in the arts and humanities digital resources**, and a second workshop on **Geographical Information Systems**. Each of the AHDS service providers also run subject-specific workshops at different periods throughout the year. A full list of AHDS workshops can be found at <http://ahds.ac.uk/programme.htm>.

Materials and presentations from our last workshop on GIS in the Arts and Humanities are now available from the AHDS website. These provide a thorough introduction to the various technical and conceptual issues involved in GIS. See http://ahds.ac.uk/gis_workshop.htm to download the materials.

Designing Shakespeare (Part Two)

/ ... continued from front page
different theatre archives.

All of the images have been selected to illustrate aspects of the production design of each show. Audio and video interviews with theatre designers in Britain, and virtual reality models of the most commonly used theatres have also been made available, to add further richness to the collection. These multimedia materials complement a textual database of production details and review extracts of performances of the plays. Together these resources constitute a substantial digital archive of Shakespearean performance at the principal theatres in London and Stratford from 1960 to 2000.



*Photograph from Old Vic production of Macbeth,
September 1980.*

Image copyright Donald Cooper. Reproduced by kind permission.

Although the new PADS website offers a generic search with full interoperability across selected collections, the design of the search mechanisms for individual datasets supports specialised access to subject-specific resources. For the Designing Shakespeare collection, users may search according to criteria such as director, designer, theatre, company and individual actors, plays or dates.

Users are also encouraged to browse through the PADS collections using a hierarchical navigation system. From the most basic selection of e.g. theatre resources, users can navigate through every relevant collection and access a variety of different objects and records. Within Designing Shakespeare, at performance level,

production details and reviews appear in full, together with a series of thumbnails or icons, which link to the relevant images, video and VRML models.

The Shakespeare materials made available online through PADS are either new or formerly restricted in their use, so it is hoped that their free availability will inspire new forms of teaching and research. Users are encouraged to copy the materials into their own courseware and lecture notes, to use them in seminars, presentations and online discussions and debates. However, commercial use of these materials or alteration of the original materials is subject to copyright restrictions.

The new PADS web site is now undergoing final testing and will go publicly online at <http://www.pads.ahds.ac.uk> by the end of October 2002.

Iain Wallace, Performing Arts Data Service

New Staff at the AHDS

The AHDS has recently begun a process of restructuring and various new members of staff have been appointed, both at the Executive and at the Service Providers. At the Executive, **Hamish James**, previously of the History Data Service, has been appointed Collections Manager, while **Andrew Speakman** becomes the new Technical Services Manager. **Emma Beer** is in charge of the Resource Guide for the Arts and Humanities, **Gill Veldon** is Events and Communications Assistant, while **Lize Blom** provides administrative support. **Martha Brundin** is the AHDS' new Project Manager.

At PADS, **Iain Wallace** is now Collection Development Officer, while at the Archaeology Data Service **Donna Page** has replaced the long-serving **Maureen Poulton** as Office Administrator. And at the Visual Arts Data Service, **Kay Barrett** is the new Administration Officer.

**Joint Information
Systems Committee**

The AHDS
is funded
by the

A · H · R · B
arts and humanities research board