Ecole du Louvre
ICHIM
Paris 03

www.ichim.org

Les institutions culturelles et le numérique
Cultural institutions and digital technology

École du Louvre
8 - 12 septembre 2003

MANAGING, DELIVERING, AND ENHANCING CULTURAL DIGITAL COLLECTIONS

Nic P. Sheen, Ph.D. - iBase Image Systems Ltd., UK

Annette A. Ward, Ph.D. - University of New Mexico, USA

« Acte publié avec le soutien de la Mission Recherche et Technologies
du Ministère de la Culture et de la Communication »

Abstract

Museums, galleries, and other cultural heritage institutions have collections in their care that face increasingly complicated problems in management, conservation, and dissemination. The right combination of services can raise the profile of an institution, increase visitor numbers, and generate revenue essential to sustain heritage activities. Digitally accessible collections have the potential to enhance and complement the physical museum while facilitating conservation and increasing accessibility through carefully-developed user interfaces. This paper provides a framework for delivering digital access to a wide variety of collections. Practical, technical, and innovative considerations are discussed with a focus on defining objectives and priorities and creating an infrastructure that can grow and change over time. Digitisation, delivery systems, user interfaces, and search tools are emphasised. Common techniques, such as free-text searching and thesauri are covered alongside a discussion of content-based image retrieval as applied as an enhancement to the Corporation of London Guildhall Library and Art Gallery site, Collage.

Keywords: Digital Images, Cultural Heritage, Technical Requirements, Content-Based Image Retrieval

Introduction

Museums, galleries, and other cultural heritage institutions have collections in their care that face increasingly complicated problems in management, conservation, and dissemination. Many are growing and have the potential to attract numerous visitors. However, some of these collections are inaccessible because of space constraints (objects are stored rather than displayed), geographical distance (someone in London may not be able to visit Paris to view an object), and financial restrictions (funding to cover costs of acquisition, preservation, maintenance, cataloguing, and exhibitions is scarce). All of these constraints reduce the number of objects that may be viewed by visitors. Digitally accessible collections have the potential to enhance and complement the physical museum while facilitating preservation and conservation, and increasing accessibility through carefully-developed user interfaces. They may also generate revenue for the institution, a bonus when financial resources are limited.

Objectives

This paper presents a framework for delivering digital access to a wide variety of collections; from paintings, sculpture, and installations to textiles, trains, and local history. The framework describes the practical, technical, and innovative considerations when digitising collections. Digitisation strategies are discussed within an infrastructure that can grow and change over time, allowing for integration with existing sources of information, while delivering them in various ways for general access, education, and revenue generation.

Experience in Digital Applications

The framework described in this paper is based on practical experience with heritage organisations in the UK involving a variety of applications, from digitisation and digitisation workflow processes to Internet, kiosk, and CD delivery systems. iBase Image Systems Limited; developers of information management and retrieval software with specialization in storage, organisation and retrieval of media files, still and moving images, and sound and graphic displays; has extensive experience in delivery of digitization services. Clients include The British Library, Corporation of London Guildhall Library and Art Gallery, National Library of Wales, Science Museum, the Royal Shakespeare Company, Wedgwood, National Museum of Photography Film and Television, and the National Railway Museum. iBase experience in digitisation projects is diverse, ranging from a small number of very high resolution (750MB) images to mass production, digitising 400,000 images at 1MB in six months; and projects that start from nothing but physical objects and a printed catalogue to ones that supplement a large existing digital collection.

Introduction to the Framework

The framework for managing and delivering projects is based around digitisation, gathering existing metadata, creating new information, indexing the digital collection, and then delivering and using these resources in a variety of ways. Like any good framework, it is comprised of interrelated components, each with a defined input and output. Planning and design of information is as crucial as systems integration in order to achieve a dynamic digital collection that makes best use of resources and that can evolve effectively.

The framework has four key components: sources, resources, relationships, and targets that together comprise a Digital Asset Management (DAM) system. These are defined as follows:

A Source is any repository of information and generally falls into two categories: “text”, such as collection records, narratives, or biographies; and “digital assets” (such as images, video clips, sound files, and PDF documents). Each source provides information according to a pre-defined specification. For example:

Images must be of a specified resolution and colour depth, and contain attached information about the capture device, operator, and colour profile, with a standard file naming convention.

Text sources provide information in a format that can be validated. Schemas in XML (eXtensible Markup Language), a language designed especially for Web documents, are an effective way to achieve this and are discussed briefly in a following section. Legacy systems (an existing application that doesn’t understand XML) that can only provide data in CSV, MARC, or spreadsheet formats are handled using transformation algorithms to convert these data into XML.

A Resource is a single unit of information, such as a description of an object in the physical collection, a digital image, a narrative describing a group of related physical objects or an artist and/or his/her biography.

A Relationship is a link between two resources. A relationship has a starting point and an ending point, and a description linking the two. For example, a link between a painting and an artist would have the painting at one end of the relationship and the artist at the other. This relationship would have a description "created by/creator of". The description is bi-directional in the sense that if one is linking an image to its creator, the description is "created by". Conversely, if one is connecting from the creator to the images they created, the description is "creator of" or "created".

A Target is any consumer of information; for example, a content-management system that drives a Web front end or a multimedia CD based on a particular topic. A target is defined by a set of resources (text and images) and a protocol (i.e., a method of transferring those resources such as XML).

The DAM system and the processes for integrating information from sources and delivering information to targets are shown in Figure 1.

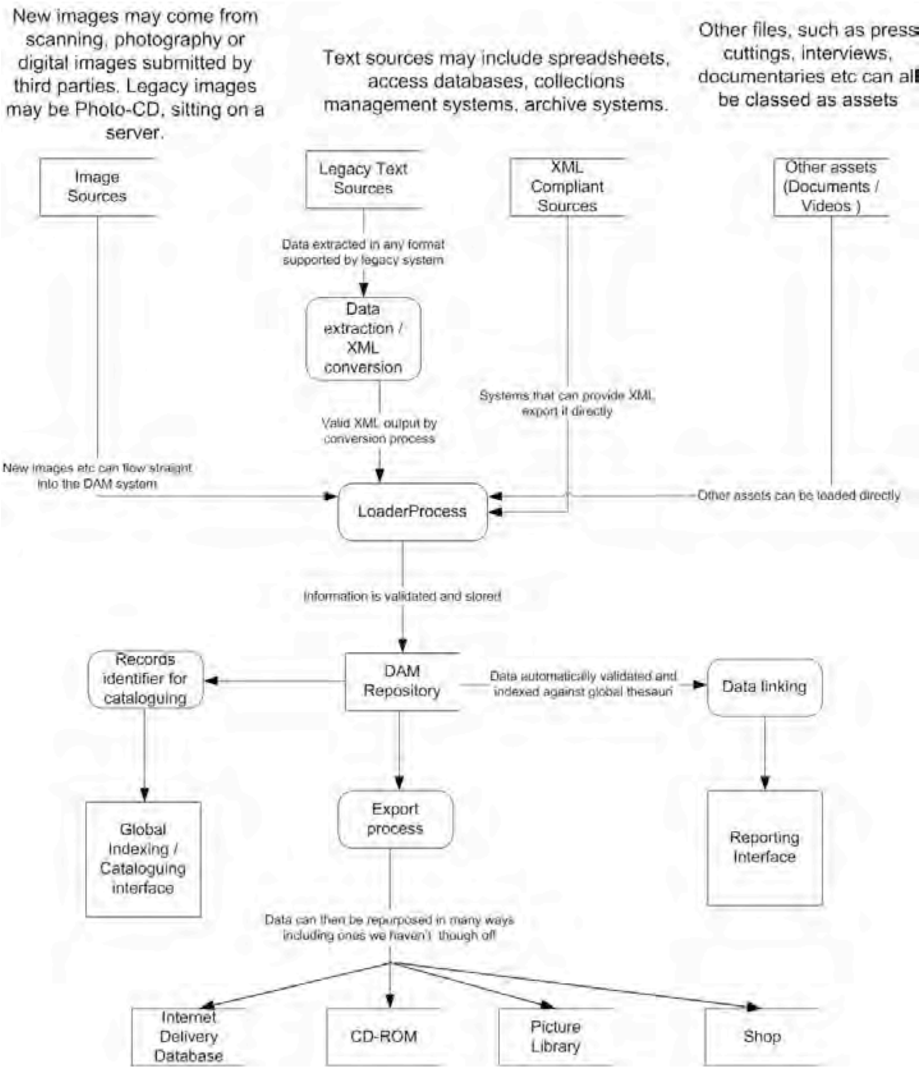

New images may come from scanning, photography or digital images submitted by third parties. Legacy images may be Photo-CD, sitting on a server.
Text sources may include spreadsheets, access databases, collections management systems, archive systems.
Other files, such as press cuttings, interviews, documentaries etc can all be classed as assets
Image Sources
Legacy Text Sources
XML Compliant Sources
Other assets (Documents / Videos)
Data extracted in any format supported by legacy system
Data extraction / XML conversion
Systems that can provide XML export it directly
Valid XML output by conversion process
New images etc can flow straight into the DAM system
LoaderProcess
Other assets can be loaded directly
Information is validated and stored
Records identifier for cataloguing
DAM Repository
Data automaticilly validated and indexed against global thesauri
Data linking
Global Indexing / Cataloguing interface
Export process
Reporting Interface
Data can then be repurposed in many ways including ones we haven't though of!
Internet Delivery Database
CD-ROM
Picture Library
Shop

Figure 1. Overview of the DAM framework illustrating the overall process flow for gathering, integrating, and delivering information.

The following sections discuss the key processes to ensure that sources produce appropriate output and the data integration processes that take place in the DAM to bring together these sources. Digitisation is discussed in some detail. From experience, digitization is the main area where customers have difficulty. There is also a brief discussion of Web delivery.

Digital Sources of Data

The most common source of image data is a digitisation (scanning/photography) process. The purpose of digitising an item, regardless whether the source material is photographs, paintings, drawings, text, videos, or 3-D objects, is to create a digital asset at a resolution and cost appropriate to its end-use. If a set of objects is only displayed on the Web, there is no point spending €20 per object creating a 100MB file when a 1MB file, costing €2, would be just as functional.

Quality and Resolution

Digital image quality is directly related to the "size" of the digital image and is affected by several factors including resolution and colour depth. This quality can either be expressed as the number of pixels, or as the number of dots per inch (DPI). Designers often use DPI, whereas computer technologists tend to use pixels to describe the size of an image.

Image file sizes do not always give an indication of quality because some formats store images in a compressed format requiring less space, but sometimes reducing quality since information is lost in the compression process. For example, an image that is 100K on a hard disk may be 3000 pixels square but would be very low quality because of the compression used. When opened, this image would actually take up about 20MB of memory in the computer.

An image that is 300 DPI and fills an 8 x 11.5 in.-paper would be 11.5 in. x 300 pixels by 8 in. x 300 pixels or roughly 3500 by 2400 pixels. If this image was scanned at 300 DPI, it would be approximately (11.5 in x 300) x (8 in x 300) x 3 bytes, or about 20 MB. About 3 bytes are required to store each pixel; one byte for each of the red, green, and blue (RGB) components in the image. Each pixel is often represented as a triple number. For example, red is 255, 0, 0; blue is 0, 0, 255; and grey is 128, 128, 128.

Image colour can also be stored in different ways. For images that use the CMYK model, (often used for full-colour offset printing), the image is composed of a Cyan, Magenta, Yellow, and a Black (K) element.

Note: Images that are often called JPEG or JPG are strictly stored in the JFIFF file format, using a JPEG compression algorithm. Other file formats also use the JPEG algorithm and the JFIFF format can use other compression schemes as well. However, this paper uses the term JPG to refer to images stored in the JFIFF format using JPEG compression.

Image Capture and Colour Management

Different image capture devices (scanners, cameras, etc.) react to colour in different ways. For example an image stored in CMYK format will need to be converted prior to onscreen display, since monitors generate colours as a combination of red, green, and blue elements. The converse also applies and images stored in an RGB colour space have to be converted before printing on a CMYK printer. Lack of appropriate conversion is the reason that printed literature often looks different to its screen image. To compound this problem, different devices have different concepts of red, green, and blue and whilst the RGB values (248, 113, 30) may be orange-like on one device, it may be redder on another.

Fortunately, there is a solution to this problem: colour spaces & colour profiles. A colour profiles is information that can be embedded into an image that provides characteristics of the device on which it was captured or indicate that colours in the image are stored in known colour space (e.g., “sRGB” for standard RGB). Similarly, output devices (such as monitors and printers) also have colour profiles. By correctly calibrating devices, one can ensure that colours always look the same wherever they are viewed.

It is essential when undertaking digitisation that all devices are configured for appropriate colour management. The images created as a result of a digitisation process must be a true representation of the original which usually can be achieved by comparing onscreen images with originals. If monitors are not correctly calibrated, the master image may be ruined when correcting the colour.

File Naming

How images are named is also important. It is pointless to capture an image if one cannot locate it at a later date. Whilst some systems will generate unique identifiers for images, it is recommended that the master file always has a name that is meaningful.

There are three key parts to a filename:

[Meaningful unique identifier] [View indicator] [Version indicator]

as well as the extension that indicates the format of the file.

The unique identifier may be an accession number. However, numbering systems often contain characters that are not allowed in filenames (e.g., / or *). A unique translation for these characters can be defined to ones not used within the numbering system. For example, if there is an accession number: 114/CXYZ*, the first part of the filename could be 114_CXYZ$. It can often be useful to pad the various parts of a filename with zeros so they sort sequentially in file lists (e.g., 11.jpg appears before 2.jpg in Microsoft® Windows Explorer because 11 sorts alphabetically before 2). However, using filenames 000011.jpg and 000002.jpg resolves this issue.

Often multiple images will be taken of an object, either for different purposes (e.g., conservation or public access), or to highlight particular areas, or different 3-D views. A set of short codes should be created to indicate the type of view contained within a file. The same view may be captured repeatedly, particularly for conservation; consequently, it is important to be able to differentiate the version. Rather than simply using a numeric indicator, it is recommended to use the date in reverse order, again so images sort correctly (e.g., 20030408 for 8 April 2003).

Archiving and File Formats

Master image files (i.e., those created by a scanner or camera) should be archived in an unprocessed form. It is important to use a file format that retains as much information as possible. The TIFF format is most popular not only because it stores images in a form where no data are lost, but because it also allows for the inclusion of metadata, such as colour profiles, caption, copyright, and information about the manufacturer of the source device (e.g., Sony and Fuji), model of the device (e.g., Cybershot), date captured, and particular settings.

Master files should be stored on a medium that can be located off-site in a secure location, for example, on a DVD. Magnetic tape is a poor format for master files because of degradation over time. Still, no one is sure how long DVDs and CDs will last, and the prudent approach would be to remaster the archive every few years to utilize technological advances.

Processing and Quality Control

Ideally, no image processing should be performed before archiving an image. The settings on the scanner should be checked for each image to produce the best quality digital image. In practice, this is too time consuming and one adjusts the image levels in software after scanning and prior to archiving.

It is essential that each image is quality checked for colour and tonal range prior to archiving. Although this adds cost to a process, it is cheaper than learning six months later that 10% of the images are unusable because of scanning errors six months later. If digitisation is being done from transparencies, screen images should be compared to the original using a high quality light box to ensure colour correctness, orientation (i.e., the correct side of the transparency has been scanned), and tonal range.

Surrogate Generation

Often it is not necessary to take the extremely expensive step of storing master images online (i.e., on hard disks), so smaller images are generated for working purposes. From experience, it is recommended to generate image sizes as shown in Table 1. These surrogate images can be generated during scanning (i.e. provided to the DAM by the source), but it is preferable to create these surrogates as images are imported into the DAM. This ensures consistency across all sources of images.

Table 1.   Recommended Sizes of Images for Various End-Uses.

Image Type	Size (Pixels)	Format	Approximate Size on Disk	End-Use
Thumbnail	100	JPEG	2KB	Display with search results in lists
Reference Image	400	JPEG	10KB	Display alongside detailed data
Screen Image	1000	JPEG	100KB	Full screen on typical monitor and for Microsoft□ Power Point□

Online Master	2000	PNG	5MB	Used for small high-quality prints and generating new surrogates
Online High-Resolution	2000	JPEG	500K	Suitable for printing in newspapers and at small sizes in magazines; appropriate for delivery over the Internet

The reason for recommending that a PNG file is stored online is that regenerating other images from a JPG file causes loss of quality. Furthermore, opening a JPG file, applying a process such as sharpening, and then saving the file will cause increasingly more data loss during subsequent processes. Simply opening and resaving a JPG file causes loss of data.

Metadata

Modern capture devices automatically store certain information with an image when it is digitised. This information is usually stored in a format called EXIFF inside an image. The JPG and TIFF formats both support the inclusion of EXIFF headers. EXIFF covers key device information such as: resolution, date, colour space, colour depth, device make, and device model. EXIFF information should be “extracted” into the database when images are imported so it can easily be searched.

In addition, manual data about the following should be captured:

Location of digitisation

Digitisation operator

Lighting

Capture software

Device settings not embedded into the EXIFF header

Additional information, such as a digital signature, a digital code that identifies the image content, can also be captured. This is especially important if one needs to know that an image has not been modified. For example, if objects are photographed before transportation and later it is necessary to file compensation or insurance claims, the digital image may be required as proof that damage occurred during transport. In this instance it is often necessary to demonstrate that the image has not been modified. Ideally, copyright and credit information should be embedded into the image header so it always travels with the image. This can happen at capture time, export time, or both.

Watermarking and Protecting Copyright

Watermarks can be used to indicate ownership and track images to avoid breaches of copyright. There are still significant legal hurdles in pursuing breaches of copyright overseas; however, one assumes that at some point the legal system will catch up with the Internet. In reality, the problem of copyright violation is far worse in traditional media. One can purchase a high quality art book for a few euros and generate far better digital images than can be downloaded off the Internet.

There are two types of watermarks: visible and embedded. Visible “watermarks” that appear on the image are usually the logo of the institution holding the rights to the image. They may appear very prominent on the image (Figure 2) or be more obscure. Embedded watermarks hide a code inside an image file that subtly alters the image data in a way that the code can be retrieved even if the image is printed and rescanned. Embedded watermarks can sometimes be removed using image processing techniques, but only if details of the way the watermark was embedded are known.

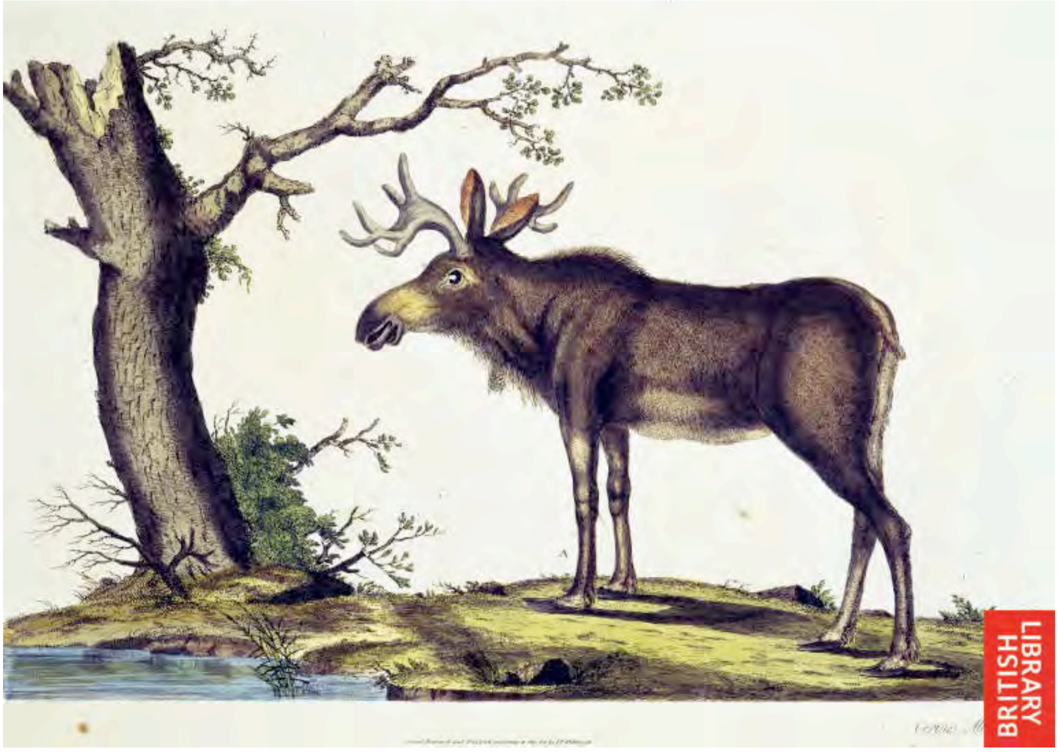


Figure 2. Sample image from a visible watermark from The British Library Images Online Web site (www. bl.uk/images). Image courtesy of The British Library, shelfmark 38.i.11.

Embedded watermarks come in two forms; a "one-time" signature that can be applied to all of the images that allows one to determine whether an image originated from a particular collection, or "unique" signatures that are embedded into images each time they are used. Unique signatures are most valuable when delivering high-resolution images for licensing since one can link the signature to a transaction database containing the details of who purchased the image and for what purpose. One-time signatures are most applicable to Web sites when it is important to know that an image was taken from the site. Watermarks should not be applied to master images, but should be applied when sets of images are exported, because once a watermark has been embedded it cannot be removed without affecting the quality of the original image.

DAM Layer: Data Structures, Indexing, and Data Integration

Data Structures

Data falls into two forms; data about the asset itself, such as digitisation workflow and EXIFF information, and additional metadata about or related to the asset, such as bibliographies and descriptions of historical events. All data coming into the system should be validated against a specified standard, regardless whether this is a public standard or one specifically derived for a project. Not all data necessarily need be validated against the same standard, but it is beneficial when possible and it also makes linking elements of the resultant data set simpler.

Although standards are debatable, we have found that simple standards where it is relatively easy to gather all information to a consistent level is better than gathering monumental amounts of data from one system and incomplete data from others. When the data are integrated, the disparity will be even more apparent. A Dublin Core-like structure (www.dublincore.org) can be an appropriate choice since it can be used to describe any type of resource from objects to people and events. Table 2 provides a listing of the Dublin Core fields that are key to building an integrated repository.

Table 2. Fundamental Dublin Core Fields for the DAM

Element	Description
Identifier	Uniquely identifies this resource within the collection; may be an accession number, a fully-qualified person’s name, or an image file name
Relation	Contains multiple elements, either internal identifiers so this object can be linked to other resources in the system, or external references, such as URLs
Format	The “format” of the resource (JPEG, XML, etc.)

Type	What the resource represents (image, physical object, etc.)

“Identifier” is defined in different ways for different types of records. For example, for an image-record, the identifier may be the filename; for a collection-object, it may be the accession number; or for a person, it may be a fully-qualified version of their name (surname and first name along with their date of birth and/or death).

Identifier and relation are both critical to data structures. Without a good identifier, it is impossible to find a single object effectively, and without relation, it is difficult to amalgamate collections (link collection records to digital images and book references). Other elements may also link to identifier. For example, Figure 3 illustrates four Dublin Core records and their relationships.

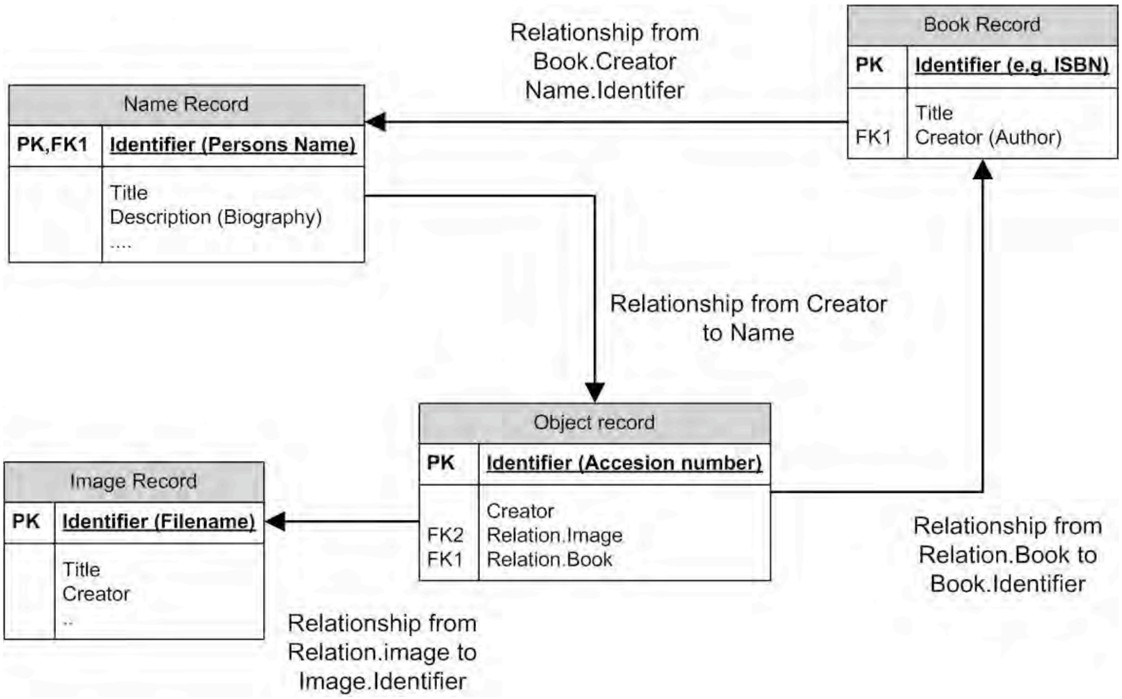


Figure 3 Dublin Core records showing how relationships are built and the homogeneity between the various data structures.

The homogeneity of the data structures mean it is trivial to add-in additional types (sources) of information and connect them to other resources in the system. For example, if exhibition information were included at a later date, a Dublin Core record for an exhibition would resemble the information in Table 3. Since this data is in exactly the same format as all the other resources in the system, it will load directly into the DAM and processing that is already in place will link the exhibition to the objects (and images) that are on display. This ability significantly reduces ongoing development costs and simplifies addition of new services.

Table 3. Possible Dublin Core Record for Exhibition Information

Column	Description
Identifier	Name of exhibition + date
Title	Name of exhibition
Creator	Who is organising the exhibition
Coverage Spatial	Location of the exhibition
Description	Detailed description of what is featured at the exhibition
Relation	List of object identifiers that will be on display

Data Integration

Validation is the key to successfully integrating disparate datasets. Historically, this was done with SQL scripts, or by enforcing constraints in a relational database. Since XML has emerged as a standard for exchanging information, a better mechanism has emerged, namely XML Schemas.

An XML document, which is simply text, contains structure and data. The structure describes the data. An XML schema describes the structure. An example of the definition of a creator element for a Dublin Core record, followed by an instance of a Dublin Core record using this schema is shown below. In this case, the XML fragment defines a name/role type that contains two elements; a name and a role and an element definition for a creator element that specifies that a record must have at least one creator, but can have as many additional occurrences as desired. XML schemas can also restrict elements to a specific length, specific characters, or a specific list of items.

Name element definition

```
<xsd:complexType name="nameRole">
<xsd:sequence>
<xsd:element name="name" type="xsd:string"/>
<xsd:element name="role" type="xsd:string"/>
</xsd:sequence>
</xsd:complexType>
<xsd:element name="dc_creator" type="nameRole"
minOccurs="1" maxOccurs="unbounded" />
```

Description of photographer “Watson, Harry J, 1850 - 1918” using the creator element defined above

```
<dc_creator>
<name>Watson, Harry J., 1850-1918</name>
<role>photographer</role>
</dc_creator>
```

The use of a schema such as XML for validating data prior to loading into the DAM is important since it isolates and forces data issues to be resolved. Also, an XML schema, unlike a traditional database schema, is highly portable. For example, it can be passed on to others who may also provide data, and, as long as it conforms to the schema, it will integrate into the DAM. However, many legacy systems don't support XML. The advantage of this framework is that it allows for pre-processing of legacy data into XML. It is relatively straightforward to produce routines that convert database tables, text files, and spreadsheets into XML in a format that can be validated.

Resolving Data Differences

Use of the XML format will resolve many data issues before they occur and before data arrives in the DAM. However, some processing is still necessary in the DAM, specifically, integration of resources; that is, the building and checking of relationships between resources and linking to a common indexing system.

In essence, when a resource is loaded into the system, for example, a resource with creator Watson, Harry J., as previously shown, it must be linked to the resource for Harry Watson. Different source systems may use different naming conventions, but these issues can usually be resolved by applying mapping to the source names or using punctuation-insensitive comparisons. For example, if there were two resources, such as Watson, Harry J. and Watson Harry J. (no comma and extra space in the second Watson data), comparing the fields without punctuation (or case) yields:

WATSONHARRYJ = WATSONHARRYJ

Discrepancies such as the one above are easily preventable at the DAM level.

Indexing

Source systems that supply text-type records will usually have some form of "key-wording" or subject classification system, appropriate to the specific aim of that system. Resolving and integrating keyword systems is probably the most complex area of bringing together separate data sources for various reasons. The indexing system used in the DAM must satisfy the current output requirements (a public Web site, shop, or an online representation of a catalogue). The system needs to respect academic correctness from the collections-management perspective, but be generally accessible and applicable to a wide range of applications. The term, "Engine", regarding a railway line, presents a good example. The majority of people who want to see "Engines" will likely search on the word "Train". Whilst not academically correct, it is commonly used.

The difficulties here are not really technical, but have to do with balancing the demands of time while satisfying most of the users, most of the time. Technology can help by providing automatic mappings between different systems so all object resources classified as "Engines" are also marked as "Trains" in an area of the index for public access.

Keyword mapping can also be used in the grouping sense. If one wanted to create a selection of assets related to "Understanding the Universe", one might map keywords from astronomy, areas of religion, and the history of science onto this keyword. It is fundamentally important to be prepared to change the way that collections are indexed. It is best to test initial ideas on a sample of records, ensuring that records are drawn from all facets of the collections since what works for one type of record will not necessarily work for another. As indexing progresses, it will be necessary to refine, restructure, and adapt the indexing. Finally, it will be necessary to reflect upon the changing interpretation and understanding of collections over time and add additional information as it becomes available.

Delivery

Once the process for integrating resources into the DAM is operating, delivery can commence. This section describes a process for delivering a collection on the Internet although the principles apply to most types of resource usage including CD-ROM and public access kiosks.

Get Something Running

Most commercial solutions provide some sort of standard Web interface. This should be used to get the content on the Web, even if initially it is only used internally. From this, one can assess what is really important and deliver services in a component-based fashion over time.

Technology and market needs change very rapidly and if there is a period of 12 months between project inception and delivery, expectation and standards will have changed.

Get Feedback

What the developer thinks is important and useful to users may be incorrect, particularly when putting collections on line for the first time. One benefit of delivering a simple system early is that it allows engagement with and understanding of potential users. However, once the system is put up, additions and modifications must be delivered in a timely fashion, otherwise users will become disenchanted and lose interest.

Refine, Improve and Extend

Some enhancements will be simple, some complex. Critical assessment must be made of any changes to the system from the standpoint of the end users. Additional services may be added in a structured fashion and each should be run as a separate project. This ensures that small benefits are delivered in an efficient manner, rather than having large complex developments that tend to be difficult to manage and, as a consequence, overrun in time and budget.

Furthermore, delivering additional elements in small parts enables the benefits of these additions to be more easily measured.

I Want to Find

Users often come to Web sites in two different ways; because they have a specific enquiry and want to find something or they know the site has the content they are interested in, but want to be shown and/or guided in their search. These two types of users call for different approaches and search mechanisms.

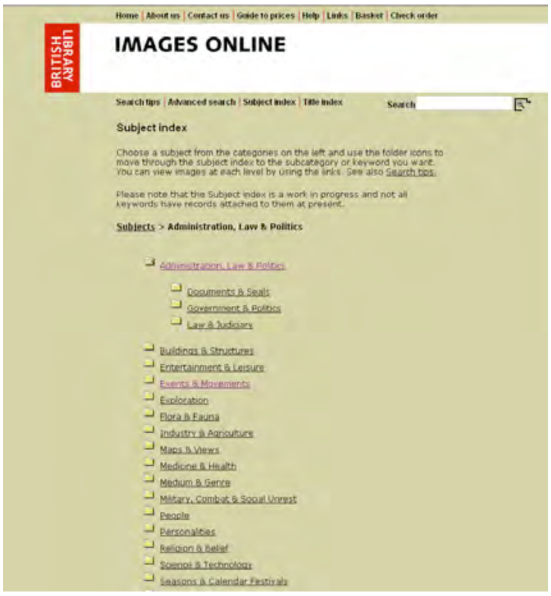
Text-Based Searching

The de-facto standard for the Web is to have a box in which you type a term to initiate a search. Although on the surface this is easy to use, results can be confusing. Most users will not know how a collection has been indexed or how the search is structured to handle such things as multiple words, hyphenated words, punctuation, or other syntax.

Two solutions may be helpful. A description of how the search engine works, but is unlikely to be read, and even less-likely to be understood by many users. Or, it is probably best to provide a more structured search where users may make specific requests such as “Show me oil paintings created between 1820 and 1865 in Italy.” From our experience, it is best to allow users to make a specific query, being careful to avoid complicated, specialized language.

Keyword/Subject Searching

This type of search can satisfy both the specific query and the general browsing user. Typically, collections have a hierarchical index of terms that can be presented in two ways; as a tree structure, where they can narrow down to the terms upon which the users want to search (Figure 4), or as an alphabetical list (Figure 5).



Home | About us | Contact us | Guide to prices | Help | Links | Basket | Check order

BRITISH LIBRARY

IMAGES ONLINE

Search tips | Advanced search | Subject index | Title index

Search

Subject index

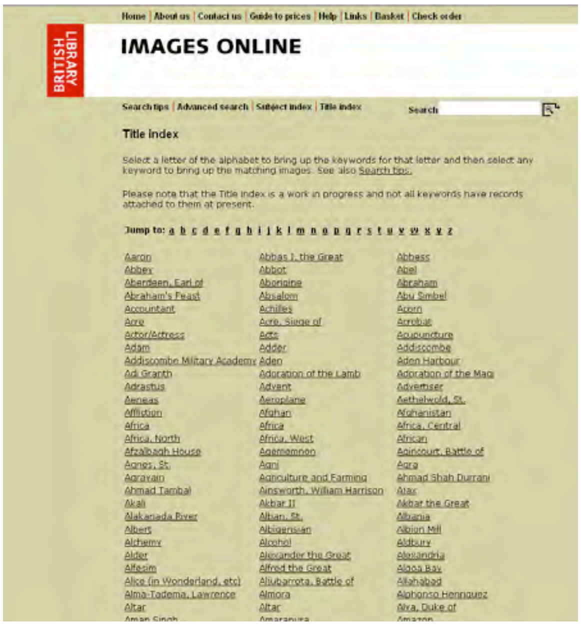
Choose a subject from the categories on the left and use the folder icons to move through the subject index to the subcategory or keyword you want. You can view images at each level by using the links. See also Search tips.

Please note that the Subject index is a work in progress and not all keywords have records attached to them at present.

Subjects > Administration, Law & Politics

- Administration, Law & Politics
  - Documents & Seals
  - Government & Politics
  - Law & Judiciary
- Buildings & Structures
- Entertainment & Leisure
- Events & Movements
- Exploration
- Flora & Fauna
- Industry & Agriculture
- Maps & Views
- Medicine & Health
- Medium & Genre
- Military, Combat & Social Unrest
- People
- Personalities
- Religion & Belief
- Science & Technology
- Seasons & Calendar Festivals

Figure 4. Tree-like "subject" search from The British Library Images Online Web site (www. bl.uk/images). Image courtesy of The British Library.



Home | About us | Contact us | Guide to prices | Help | Links | Basket | Check order

BRITISH LIBRARY

IMAGES ONLINE

Search tips | Advanced search | Subject index | Title index

Search

Title index

Select a letter of the alphabet to bring up the keywords for that letter and then select any keyword to bring up the matching images. See also Search tips.

Please note that the Title index is a work in progress and not all keywords have records attached to them at present.

Jump to: a b c d e f g h i j k l m n o p q r s t u v w x y z

Aaron	Abbas I, the Great	Abbess
Abbey	Abbot	Abel
Aberdeen, Earl of	Aborigine	Abraham
Abraham's Feast	Absalom	Abu Simbel
Accountant	Achilles	Acorn
Acre	Acre, Siege of	Acrobat
Actor/Actress	Acts	Acupuncture
Adam	Adder	Addiscombe
Addiscombe Military Academy	Aden	Aden Harbour
Adi Granth	Adoration of the Lamb	Adoration of the Magi
Adrastus	Advent	Advertiser
Aeneas	Aeroplane	Aethelwold, St.
Affliction	Afghan	Afghanistan
Africa	Africa	Africa, Central
Africa, North	Africa, West	African
Afzalbagh House	Agamemnon	Agincourt, Battle of
Agnes, St.	Agni	Agra
Agravain	Agriculture and Farming	Ahmad Shah Durrani
Ahmad Tambal	Ainsworth, William Harrison	Ajax
Akali	Akbar II	Akbar the Great
Alakanada River	Alban, St.	Albania
Albert	Albigensian	Albion Mill
Alchemy	Alcohol	Aldbury
Alder	Alexander the Great	Alexandria
Alfesim	Alfred the Great	Algoa Bay
Alice (in Wonderland, etc)	Aljubarrota, Battle of	Allahabad
Alma-Tadema, Lawrence	Almora	Alphonso Henriquez
Altar	Altar	Alva, Duke of
Aman Singh	Amarapura	Amazon

Figure 5. Example of an index from the British Library Images Online Web site (www.bl.uk/images). Image courtesy of The British Library.

I Want to Be Shown

In addition to using the keyword/subject techniques described in the previous section, narratives, highlights, and popular searches can all be used to bring users into an online collection. Narratives bringing together a set of related resources may be gathered together for illustrating a particular topic, period, or event. More detailed explanation on the context of the images and information about the common theme may encourage users to simply explore the collection by following links to related resources. The narrative can also be another type of Dublin Core resource in the database that is stored and reused, just like any other asset.

Similar in concept to narratives, "highlights" feature famous objects or groups for which the collection is known. This too provides contextual information and initiation into the collection.

Instead of deciding which images and terms are most important, the software can determine it automatically, based on searches performed previously and then present users with a list of the most popular ones. Within one "click" users can then begin exploring the collection.

Exploiting and Repurposing Information

Marketing

All institutions must market their collections. The DAM provides a key source of material for press releases, brochures, and other marketing material. By enabling marketing departments to reuse the content stored in the DAM, cost savings should be realised. In addition, event photography and other event-based imagery, typically created by marketing departments, should be input back into the DAM as part of the overall institutional resource.

Picture Libraries and Shops

The most common exploitation of heritage collections are picture libraries that license images from the collection for use in publications and sell prints to the general public. A Web front-end to the DAM can open these channels to new markets and increased customer bases. Our experience indicates that picture library revenues can increase 10-15% by having an online facility. The online service can simply be a shop front whereby customers select images for their own use, but still go through a traditional rights negotiation offline. Increasingly, images are being provided online for a limited set of rights. Users can purchase images online using a credit card and then immediately download images. This is essential to opening new markets that may currently "ignore" heritage institutions because they believe that images will not be provided promptly.

Links can be built from images in the collection to related products in an institution's shop, encouraging and guiding users to purchase items. Someone exploring the collection may not have the intention of making a purchase, but if they find an object or topic in which they are interested, making them aware of merchandise that is available undoubtedly increases sales.

Education

Portions of the database can be published to CD either as an HTML-based interface or into a FLASH application, creating themed packages and educational aids that provide additional sources of revenue.

Search Engines

There is no point digitising a collection if it is not indexed. There is little point in putting a collection online if it is not indexed in the search engines. Most sites are listed at a basic level; institution name, for instance. Ideally all objects in the collection should be registered, as well. This is often a problem with database-driven Web sites because there are no physical pages for a search engine to index since pages are created "on the fly", in response to user queries. This leads to what has been called "the hidden Web". That is, content is available but cannot be found directly. This problem can be resolved by creating index pages for the search engine to pick up and using techniques to steer the search engine to dynamic pages

Portals

Often, the cost of implementing a project of the nature described in this paper is prohibitive. Furthermore, some institutions will not generate sufficient revenue from their collections to justify the expenditure necessary to create a new service.

The framework discussed in this paper can be equally and effectively applied to portals where many institutions collaborate to deliver a rich online resource and associated commercial services.

This approach can provide many benefits, including:

Initial development costs may be spread across many institutions

Larger number of visitors will use the expanded collection, therefore providing larger overall revenue opportunities

One set of staff are required to maintain the system

Costs can be spread, providing a greater chance the portal will adapt as the customers change

However, there are disadvantages:

It can be difficult to reach decisions about style and content unless all parties are committed to achieving the objectives of the portal in a timely manner

Since more institutions are involved it is likely that a larger number of source systems, data formats, and indexing systems will need to be integrated, inevitably leading to an increased number of issues to be resolved

To effectively compete in today’s rapidly changing Internet world, institutions have to embrace the portal concept by collaborating with partners and other content providers to develop shared services.

Enhancement: Collage, Corporation of London Guildhall Library and Art Gallery

Although experimentation and innovation come with risks, the inclusion of enhancements within the framework allows an institution to promote a service that sets them apart from other organizations, establishing their identity while strengthening their visibility. One enhancement made to Collage was adding innovative software that allows users to explore the collection without using metadata.

Collage, the online collection of the London Guildhall Library and Art Gallery, presents a comprehensive view of London through paintings, engravings, maps, photographs, prints, and drawings from the 17th to 20th centuries. iBase, providers of the software for Collage since its inception in 1995, collaborated with the Institute of Image Data Research (IIDR) at Northumbria University in Newcastle upon Tyne, England, to incorporate content-based image retrieval (CBIR) onto the site.

Content-based image retrieval is a computer-derived technique for retrieving images based on elements such as colour, texture, and shape. It uses features of a selected painting, print, photograph, or other object to find visually similar images and locate matches regardless whether they share key words with the original image. Although Collage provides three traditional search mechanisms, a simple text search, advanced search, and a subject search, as well as links to highlights in the collection, the addition of CBIR provides unique visual access to the collection, allowing users to browse the collection in a more self-directed way.

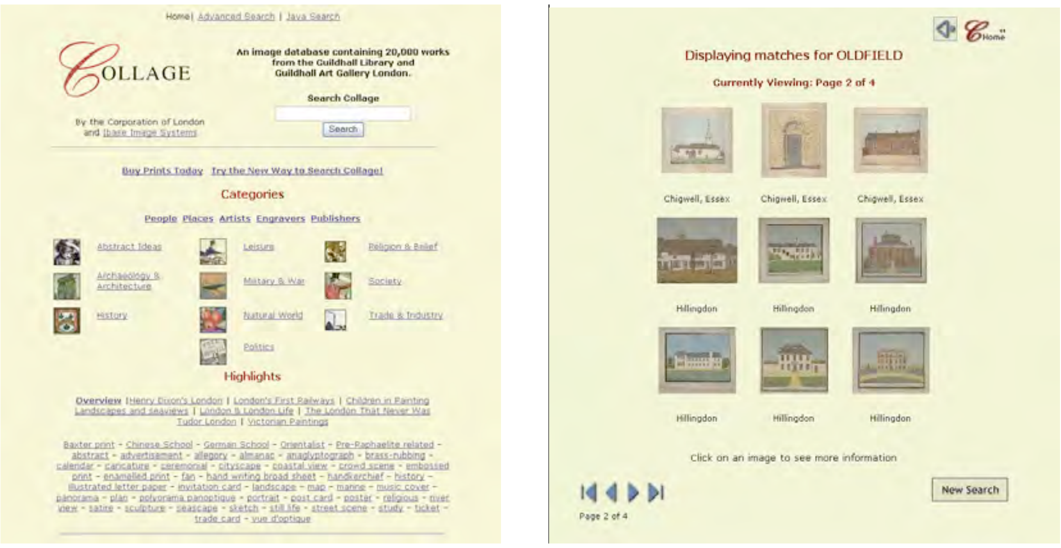
Technical Application of Content-based Image Retrieval Software to Collage

In the study by Ward, Graham, and Riley (2002), funded by Resource: The Council for Museums, Archives and Libraries (CMAL/RE/103), CBIR software was applied to the Collage database. Application of software to Collage required a smooth transition with no disruption to services and no confusion to users. The simple mechanism for trying the software system was designed to allow users the option to specify preferences on search parameters, return results in a familiar format, and work with common browsers.

System Design

General testing of the site was simplified because of the modular architecture of the computer system. Components could be developed and tested before integrating into the complete system. Once the database was created, CBIR software from Virage® was linked to Collage to perform image analysis and image comparison, resulting in a mathematical representation of the visual content that would be used to compare image characteristics

In order to initiate a CBIR search, the user first starts with a traditional Collage search (Figure 6). Once a specific image has been selected, the user may conduct a "Standard Visual Search," with default search parameters for colour, visual texture, and shape (Figure 7). When an "Advanced Visual Search" is conducted, the user ranks colour, texture, and shape on a five-point Likert-type scale indicating the importance of the characteristic (Figure 8). Virage® software sorts the images and returns the 8 nearest matches in descending order of similarity along with the original image (Figure 9).


Home| Advanced Search | Java Search
COLLAGE
Search Collage
Search
Buy Prints Today Try the New Way to Search Collage!
Categories
People Places Artists Engravers Publishers
Abstract Ideas
Leisure
Religion & Belief
Archaeology & Architecture
Military & War
Society
History
Natural World
Trade & Industry
Politics
Highlights
Overview |Henry Dixon's London | London's First Railways | Children in Painting
Landscapes and seaviews | London & London Life | The London That Never Was
Tudor London | Victorian Paintings
Baxter print - Chinese School - German School - Orientalist - Pre-Raphaelite related - abstract - advertisement - allegory - almanac - anaglyptograph - brass-rubbing - calendar - caricature - ceremonial - cityscape - coastal view - crowd scene - embossed print - enamelled print - fan - hand writing broad sheet - handkerchief - history - illustrated letter paper - invitation card - landscape - map - marine - music cover - panorama - plan - polyorama panoptique - portrait - post card - poster - religious - river view - satire - sculpture - seascape - sketch - still life - street scene - study - ticket - trade card - vue d'optique
Home
Chigwell, Essex
Chigwell, Essex
Chigwell, Essex
Hillingdon
Hillingdon
Hillingdon
Hillingdon
Hillingdon
Hillingdon
New Search
Page 2 of 4

An image database containing 20,000 works from the Guildhall Library and Guildhall Art Gallery London.

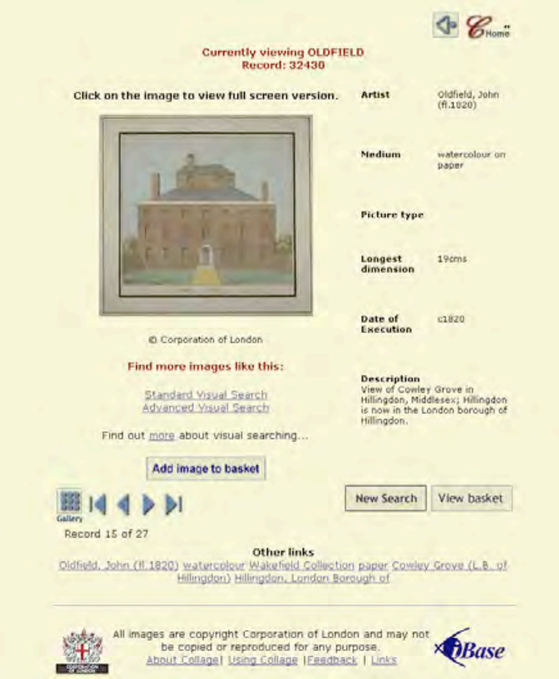
By the Corporation of London and Ibase Image Systems

Displaying matches for OLDFIELD

Currently Viewing: Page 2 of 4

Click on an image to see more information

Figure 6. A traditional Collage search must be conducted (a) and images returned (b) prior to conducting a CBIR search. Image courtesy of Corporation of London, Library and Guildhall Art Gallery Department.



Currently viewing OLDFIELD
Record: 32430

Click on the image to view full screen version.

Artist	Oldfield, John (fl.1820)
Medium	watercolour on paper
Picture type	
Longest dimension	19cms
Date of Execution	c1820

© Corporation of London

Find more images like this:

Standard Visual Search
Advanced Visual Search

Find out more about visual searching...

Description
View of Cowley Grove in Hillingdon, Middlesex; Hillingdon is now in the London borough of Hillingdon.

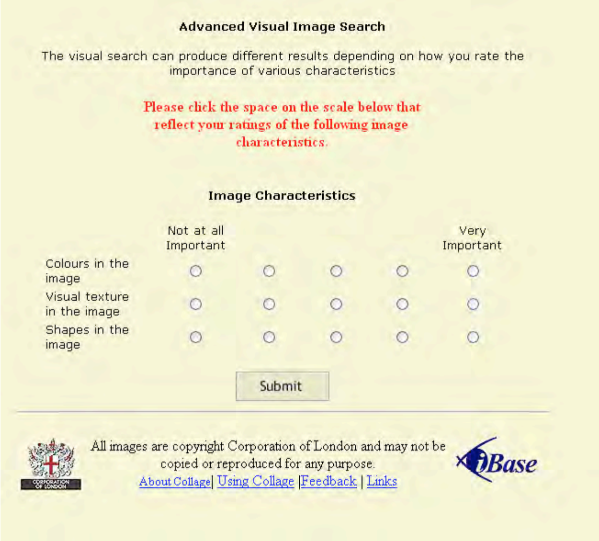
Record 15 of 27

Other links
Oldfield, John (fl 1820) watercolour Wakefield Collection paper Cowley Grove (L.B. of Hillingdon) Hillingdon, London Borough of

All images are copyright Corporation of London and may not be copied or reproduced for any purpose.
About Collage | Using Collage | Feedback | Links

Figure 7. A Standard Visual Search may be conducted once an image is selected from the gallery.

Image courtesy of Corporation of London, Library and Guildhall Art Gallery Department.



Advanced Visual Image Search

The visual search can produce different results depending on how you rate the importance of various characteristics

Please click the space on the scale below that reflect your ratings of the following image characteristics.

Image Characteristics

	Not at all Important				Very Important
Colours in the image	○	○	○	○	○
Visual texture in the image	○	○	○	○	○
Shapes in the image	○	○	○	○	○

All images are copyright Corporation of London and may not be copied or reproduced for any purpose.
About Collage | Using Collage | Feedback | Links

Figure 8. An Advanced Visual Search allows the user to rank image colour, texture, and shape on a five-point Likert-type scale indicating the characteristics as "Not at all Important" to "Very Important".

Image courtesy of Corporation of London, Library and Guildhall Art Gallery Department.

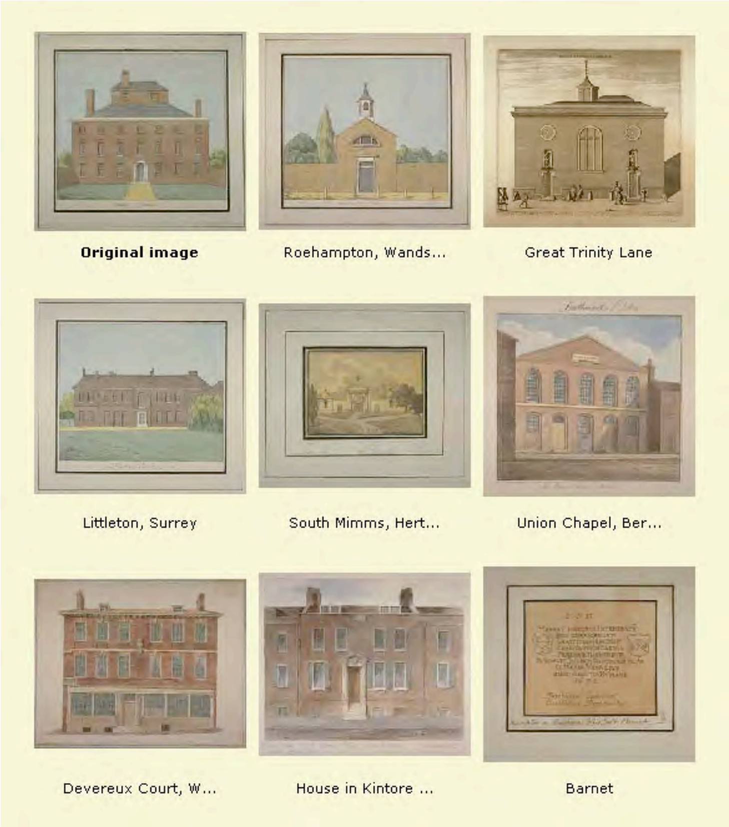


Figure 9. Results of the CBIR Advanced Visual Search are displayed using a gallery format similar to the traditional Collage search.
Image courtesy of Corporation of London, Library and Guildhall Art Gallery Department.

When CBIR is initiated, information about the request is sent to the IIDR server (Figure 10). Images are compared, then sorted to create a list of image ID's for display. In order not to affect the general performance of Collage, the CBIR engine was placed on a second server, removing processing away from the main Collage database maintained by iBase, and shifting maintenance and monitoring of the CBIR segment to IIDR.

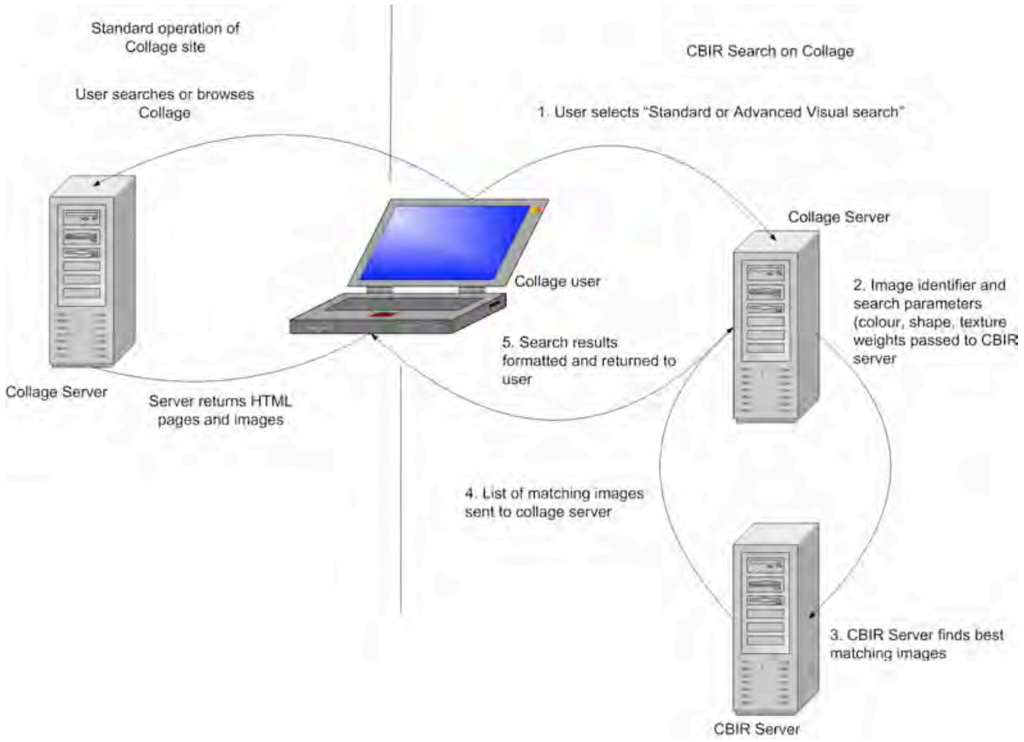

Standard operation of Collage site
User searches or browses Collage
Collage Server
Server returns HTML pages and images
CBIR Search on Collage
1. User selects "Standard or Advanced Visual search"
Collage Server
Collage user
2. Image identifier and search parameters (colour, shape, texture weights passed to CBIR server
5. Search results formatted and returned to user
4. List of matching images sent to collage server
3. CBIR Server finds best matching images
CBIR Server

Figure 10. The standard operation of the Collage site illustrates a text-based search and communications between Collage and the Institute for Image Data Research when processing a CBIR search.

Testing and Evaluation

During project development, Collage was made available on the Institute's network to facilitate testing prior to Web introduction. At the same time, an on-line questionnaire was developed to assess CBIR functionality and satisfaction, the Collage service, and demographic information about the users. Format of the questionnaire was designed to match the original Collage site and was pilot tested on browsers to ensure compatibility on alternate systems.

Evaluation data collected online from 181 subjects provide a user's perspective regarding the enhanced services (Ward et al., 2002). Although respondents indicated that only about one-half (48%) of all the images retrieved by the CBIR software were good matches to the original and 40% of the users "agreed" or "strongly agreed" they found what they wanted by using the visual search, over 65% still indicated results were useful. Of particular importance was that nearly 80% indicated they would like to use the visual image search again and 73% of the users indicated that the visual image search was a good method to retrieve images. This is strong evidence to support the addition of CBIR to enhance the Collage Web site.

Conclusions

Digital image collections become increasingly difficult to manage as they and their users continue to grow. The dynamic nature of these collections require that practical, technical, and innovative considerations be well-defined so that digitisation, delivery systems, user interfaces, and search tools are applied in the most effective ways. Planning and staged delivery are crucial in all aspects of delivering a digital collection.

The objective of this paper was to highlight the key aspects necessary for digitising and delivering an online collection, making best use of practical experience gained from working with many collections from high-profile institutions. The framework presented in this paper, should enable digitisation projects to create a reusable resource that can be built upon effectively over time and include enhancements that will enrich the users' experiences on the site.

Finally, there is a need for increased collaboration in the sector to deliver shared portals, reducing the cost to individual institutions, spreading the work of ongoing content and service development, and thereby reaping the benefits of a larger resource attracting more visitors.

References

Ward, A. A., Graham, M. E., and Riley, K. J. (2002). Final Report: Evaluation of Content-based Image Retrieval in Operational Settings: The Corporation of London Guildhall Library and Art Gallery, The British Library, and The British Broadcasting Corporation. Submitted to Resource: The Council for Museums, Archives and Libraries (CMAL/RE/103), UK.