Information that Come as Images: Overview of Issues

by *Repke de Vries* *

Abstract
Increasingly information is available as images: colour, black and white - from satellites, photography, scanning in the biomedical sciences or scanning of printed sources, like codebooks accompanying numeric data and historic records. These images have very specific computer formats but need to be easily embedded, identified, searched, transferred and browsed or reproduced to paper to make them useful as information carriers. Principal media for distribution and access are CDROM and Internet. Briefly formats, standards and applications are discussed to indicate how much actual progress is being made in putting image information into the hands of researchers and interested users.

For the last five years and not only in the social sciences but in many areas of research, in CD-ROM and Internet publishing and in library services, information is increasingly available as images. There are several reasons for this increase:

1. more applications that create images : data visualisation, analytic mapping, flow charting to represent questionnaire routing, GIS systems, convenient document delivery

2. more applications that need images as content matter: computer assisted learning, distance learning, multimedia

3. increased transformation feasibility from other media (like printed or hand written sources) to images: including image enhancement and manipulation to correct imperfect originals

4. new types of hardware that can produce images: digitising video sources , digital photocameras, digitising to images in biomedical research and health care, remote sensing, forensic applications

5. new possibilities for distribution of series of images: improving Internet transfer capacity, new multi-page image formats that fit better with Internet client - server approaches, easy creation of single or small series of CD-ROM's that can hold large numbers of images

6. new approaches that explore the next step from preservation to presentation by bringing images into structures like SGML and HTML designs or by betting on new formats like PDF

7. better availability of applications that - also in the public domain - view, browse series, and print images, including reproducing to colour on a low end desktop with colour printing ; new (PDF) reader software that can "search" images of textual information with help of shadow pages that hold text approximations of the original after "dirty OCR"

8. recognition of images as another electronic source of information that needs appropriate bibliographic description and identification

In summary: more and different types of information are provided as images and at the same time there has been a shift in emphasis from long term storage with exact replication of the original and efficient compression towards giving access : finding ways to locate, search, browse and navigate the information content of images.

In particular a format like TIFF represents that longer established practice of preservation : it handles virtually any image original - level of detail, colour schemes (including black and white) and the like - has different compression choices (including a lossy one), is computer platform independent, is recognised as a standard and by consequence a format of choice for scanning stations and almost any application that one way or the other converts, transfers, shows, does OCR or reproduces images to print. In addition to long term storage it is efficient in simple distribution: like FTP, CD-ROM's with series of TIFF images and document delivery of scanned journals. Internet and TIFF though proved difficult to match: the format has the multi-page feature but this allows by it self only a linear forward and backward browsing and TIFF cannot be transferred to Internet client software with features like an increasing level of detail or with "page on demand" . Always a complete (multi-page) TIFF with all original detail, has to arrive over the Net before the image information can be viewed or printed. A possibly very time consuming exercise. [1]

In terms of file formats and compression schemes did the Internet and the relatively slow speed of parts of the electronic connection between user and provider, sparse improvements to bring down transfer time : JPEG for example reduces original image file size drastically - but is a lossy technique: after decompressing there are differences with the original ; Progressive JPEG and Interleaved GIF are Web developments that help transfer by first presenting a rough impression to the user and gradually completing the image. [2]

This development forces archival organisations that both have to preserve an image collection and are information provider with that very same collection, to keep their images in two formats and face a dilemma. A preservation choice should be TIFF, probably off-line on CD-ROM's. But a different format would be needed to suit Internet presentation and electronic publishing.

Though bringing down transfer time is needed where Internet speed is not ideal , the real issue of access is a different one:

1. finding approaches to bring images into a structure together with other information that can have arbitrary other formats, like ASCII text, raw data files or even sound samples when biological research has habitat photos of birds together with recordings of their calls and field notes. Such a structure would also allow browsing and navigating with hyper-links and not just linear as in the single multi-page image file

2. finding ways to examine the information contained in images: scanned text can have a rough searchable real text duplicate after "dirty OCR" , other images (like photographs or historical sources) can have meta-data attached that allow for searching and locating. This meta-data would take the usual form of keywording, applying subject classifications or standard bibliographic reference.

The two approaches go particularly well together when they complement each other : access to an image collection or single image after searching meta-data, with the images subsequently structured in a way that puts them into context with other available information and permits easy walk-abouts and retrieval.

The main choices to realise this kind of access seem to be the following:

***For structure*:**

1. SGML and HTML as SGML applied to the Internet

Pro's:
- both meta-data and any further information type besides images, can be brought into this structure and at the same time keep their own specific (technical) format;
- it has hyper-linking;
- it is a general standard with good availability of software to code into this structure and to decode it;

Con's:
- HTML Web browsers which are themselves in the public domain may need additional commercial software to view and manipulate particular image-formats ; this forces the provider to choose a format that is either supported natively by the common Web browsers or for which the additionally needed image-viewer comes free;
- The same software arguing holds for SGML
- The Data Documentation Initiative is an SGML structure initiative that explicitely defines the possibility to bring needed information into SGML in wahtever format it comes: also questionnaire pages scanned to images (URL: http://www.icpsr.umich.edu/DDI/ and the May 1996 paper by K Rasmussen "Convergence of Meta Data. The Development of Standards for Social Science Data" (contact the author at boye@get2net.dk).

2. PDF with its additional hyper-linking and annotation features

Pro's:
- public domain availability of PDF reader software;
- the reader software has all navigating, browsing, searching and image-viewing together in one application ;
- it has hyper-linking

Con's:
- the structure is limited to images and text information (no raw data for example) that both first have to loose their original format and have to be imported into PDF with commercial software;
- while HTML and SGML know many applications that decode this structure and explore its information content, has complete PDF with mixed image/text content, searching and hyper-linking only the (free) Adobe reader software for technical access
- Adobe, the initiator of PDF, can be found at URL: http://www.adobe.com/
- "Internet publishing with Acrobat" by G Kent , published by Adobe Press (1996) discusses "..creating and integrating PDF files with HTML on the Internet .." and marks Adobe's move to adapt PDF to Internet presentation

3. A SGML and Internet - HTML approach supported by PDF, where PDF only holds the images but the (free)

Adobe reader software brings sophisticated viewing and manipulation , certainly after the recent adaptation of PDF to Internet requirements like “image on request” from a multi-page file.

- the PSID Internet site can serve as an example: URL: http://www.isr.umich.edu/src/psid/pdf.html It has to be noted that contrary to the above, the PSID site and many other providers use PDF only as a distribution format or document delivery mechanism, which was made attractive by Adobe putting PDF reader (and printing) software in the public domain. None of the expanded features of PDF are utilised.

For searching information hidden in images:

1. Database approaches

Several scenario’s exist:
- descriptive information is indexed and searched, with the images as search result;
- WAIS makes this possible for the Internet outcomes of “dirty OCR” on images from scanned documentation, are indexed and searched; a further linking scheme has to connect with the right image free-text indexing and searching, where the retrieved text has graphic-annotation links to images;
- ISYS database software and the International Social Survey Program CDROM are an example. URL: http://www.za.uni-koeln.de/data/en/issp/index.htm

Pro’s :
-the database or search engine approach quickly brings results, like any indexed free-text search

Con’s :
for the same reason is precision in search outcome not always high ;
additional effort is needed to create descriptive and image linking information, or with text from some source available at least the links to images (like in codebook text and scanned questionnaire examples);

2. Subject classifications and keywording

- The CASS Question Bank on the World Wide Web illustrates this:
- URL: http://kennedy.soc.surrey.ac.uk/qb/Welcome.html
- It has in fact both different subject trees and a search engine.

Pro’s :
a very precise search that takes time when browsing subject trees but can still be fast when assisted by a thesaurus approach

Con’s:
Even more than in rough free-text indexing, is human effort needed to apply keywords or classify each set of images (like a questionnaire in the CASS example)

3. Applying the commercial Adobe Acrobat software possibilities to expand the PDF format with searchable text that can be produced with internal “dirty OCR” . This search is linear and images that have a non-text content would obviously need other Acrobat means to pinpoint particular images.: annotations and hyper-linking. The hyper-linking to relevant images within the single PDF-file can interestingly enough be realised with the idea of a “clickable map” : particular clickable areas of an image, like a scanned Table of Contents or an overview picture of the human body, can be hot-linked to where further image information starts. This can create a very intuitive guidance in searching.

- A very recent CDROM produced by the German Zentralarchiv and the Dutch Steinmetz Archive holding part of the Eurobarometer questionnaires in the original languages, uses this type of clickable hyper-linking
- ICPSR produces CD-ROM’s that have the search facility in PDF after internal “dirty OCR” : for example the CD0013 “Health and Well-Being of Older Adults” prepared by NACDA (URL: http://www.icpsr.umich.edu/nacda) This way both the original codebook page is available in PDF as image and for searching the same format holds a text equivalent where possible.

Pro’s :
- creating the descriptive information to search on is part of the Adobe software
- the idea of “click and go”

Con’s:
- the search is slow ;
- non-text images cannot do with automated “dirty OCR” and need complicated, manually added hyper-linking and annotations ;
- clickable links need human effort to organise and implement ;
- the extended PDF format with the above mentioned features, probably has to be regarded proprietary and would always need the hitherto free Adobe reader software

Conclusion.

The newer, extended features of PDF are too recent to have had much evaluation . Many services and products have been realised in PDF but predominantly as carrier for document delivery and distribution - following the bringing in the public domain by Adobe of the PDF reader software.

SGML (HTML) seems to offer the better, more general structure for access, which structure can also hold the descriptive information , keywords or meta-data that a search engine could take for indexing to help locate particular images. HTML Web applications have brought considerable illustration of giving access to information as images but not in any complicated way, other SGML examples like the DDI still have to make their point.

This overview makes a distinction between preservation and presentation and has focused on the latter. It is an exciting new area of services, still open ended in direction but moving forward quickly.

[*] Paper presented at IASSIST/IFDO ‘97, Odense, Denmark, May 6-9,1997. R de Vries, Steinmetz Archive, Netherlands Email: repke.de.vries@niwi.knaw.nl

[1] The Netscape Web site “Inline Plug Ins: Image Viewers” nevertheless feature several TIFF applications

[2] Also the PNG image format is an interesting new development: URL: http://www.w3.org/pub/Graphics/PNG/