Exchange of scanned documentation between social scientists and data archives: establishing an image file format and method of transfer

by Repke de Vries and Cor van der Meer [1]
Steinmetz Data Archive for the Social Sciences
Amsterdam, Holland

Introduction

Social science research uses as its raw material not only datasets but also the accompanying documentation: codebooks, questionnaires and so on. Sometimes these "guidebooks" are machine readable and available as text files. But older studies and questionnaires in their original form are all paper - only documentation. Other examples are handwritten comments on computer print out, sketches and black and white pictures as used in psychological research. Needing this kind of documentation means repeated photocopying by archive or library and mail delivery whereas the actual data may travel by networks like the Internet or be put on tape or any other computer medium. It's a situation disadvantageous to both the archiving world and the researcher in need of complementary documentation - especially if both are geographically wide apart.

Wasn't the Fax machine invented to do just that - to get any sketch, image or piece of text instantaneously from A to B ? To an extent yes - but the resolution is poor and it is still repeated "photocopying" sending and lots of paper again upon receiving. The image can't be pasted in a research paper, nor can it be stored in a database, viewed on screen or read by OCR packages. Fax boards in a personal computer don't change that really: for one thing there can't be constant polling for incoming Faxes or a direct telephone connection is not available to the researcher. And though a Fax board gives you the image (whatever it is) for the first time as a file on the PC, the resolution is still not good enough.

Networking on the other hand is mature now: the integration of local area networks with interconnecting nets like the Internet, often gives the desktop computer global networking facilities whereas the one Fax machine for the department is down the corridor.

Obviously transferring codebook pages etc. as images has to follow a different scenario, avoiding the repetition and manual labour in Fax and taking advantage of network capabilities:

The **scanning** of the document has to be separate from transfer. Scanning should be a one time operation with adequate resolution. Storage has to involve compression techniques. The collection of image files could be handled by a specialised database that also holds descriptive and administrative information. Or the files might be the result of just scanning a few questionnaire pages with hand written comments. The advantage over Fax is that once scanned and stored, sending out an image - like any other computer file - is easily repeated and initiated. And such scanning can be done at a much higher resolution.

Storage formats for scanned images can be the own policy of archive or library but an **exchange format** (and the necessary conversion) should be accepted and adhered to by anyone offering documentation as image files.

The **transfer** comes next and can be done in a number of ways, even as ordinary mail by reprinting the image on paper with a laser printer. Network transfer though is easiest and fastest. The researcher needing the pages receives it as a series of small files on his or her own computer or personal file area in a local network.

The last step involves a tool for the end user to decompress and actually use the images. Ideally the images received can be handled as such by the usual word processing software available to social science researchers. But a free software program will otherwise **translate back from the exchange format** to a "common denominator" format, if need be.

Establishing an Exchange Standard for images.
TIFF as the format of choice for the Exchange of Images.

The "tagged information file format" was launched by Aldus Corporation and Microsoft in 1986 and Revision 6.0 is now (April 1992) in Draft 2 and finalizing.

All this time careful attention has been paid to keep the skeleton of the TIFF header and the mechanism of the format (a pointer structure) the same. Older TIFF readers or writers therefore can exit gracefully if confronted with a TIFF file holding a state of the art colour image. Another feature is the use of tags holding vital information about the kind of image, the compression type used for the image block inside the TIFF file, but also texts of possibly any length describing the image.

If software can't read "TIFF" though it promises clearly to do so, it is just because of this versatility. Often simpler compression types possible in TIFF together with black and white images are handled but grey scale or a more complicated compression method are not. Reading appropriate tags in the TIFF file these packages could have given you helpful hints why it was decided that your TIFF variation can't be imported, but most of the time a misleading message on the screen mutters about "incompatible format". If one knows how to read the information, similarly a TIFF header dumper program tells you straight away how the image in the TIFF file is built up.

For data archives and libraries starting the service of making documentation available as images, it is of paramount importance to choose a standard that:

- has wide acceptance,
- is not in any way patented or licensed (with concern to the compression schemes),
- is not computer type or operating system dependent,
- has features to make it self-explaining (documentation tags)
- and is open to new developments in the imaging field but will never be changed in its basic format.

An indication of the acceptance of TIFF as standard for an image file format is the publication last January of the Memo "A file format for the exchange of images in the Internet" by the Network Fax Working Group of the Internet Engineering Task Force. Authors Alan Katz and Danny Cohen from USC Information Sciences Institute, define "the standard file format for the exchange of bitmapped images within the Internet" as a particular TIFF variation. (TIFF-B, preferably with compression type 4).

TIFF is the format read without any problem by the major OCR programs. Format stability is an issue close to the heart of archives. For the storage of images the long term perspective is carefully planned for in the development of the TIFF standard. On the other hand the TIFF 6.0 Revision draft also shows how flexible the standard really is: if libraries or archives take an interest in offering photographic information as images, the same TIFF format can act as wrapper but this time with JPEG compression that is now accepted as one of the TIFF compacting schemes.

In choosing the right kind of TIFF format for the Exchange standard, the following is presumed:

- foremost is the need for scanning and transferring of text together with some line drawings, as in questionnaires. These are called black and white or, bilevel images.
- the scanning resolution should be 300 dpi. This gives adequate detail and matches best with the printing resolution of today's average laserprinter. Mismatches complicate the software needed to either convert to the exchange standard or use the images afterwards.
- the compression chosen should be optimized for bilevel images and pack as tightly as possible
- each original page of information is kept as separate image and separate TIFF file; TIFF has a multi-page feature (one resulting file, holding a number of compressed images) but this option is for the moment not used.
- there is a need for adding descriptive information to the image; TIFF has tags that can be used for that purpose but this option is for the moment not used.

This leads to the choice of TIFF compression type 4. Well described in the TIFF 5.0 paper, still present in the TIFF 6.0 Revision draft (draft 1, February 1992) as one of the compression schemes. This compression type follows Fax Group 4. (The two numbers "four" are a coincidence). And Fax Group 4 is yet another standard and already fully described in the CCITT Recommendation T.6. The compression and decompression techniques described in the Recommendation are open to anybody for use in own programming. Fax Group 4 is optimized for bilevel images that hold a mix of text and lines: a lot of white with interspersed black dots.

The tags used are (referring to TIFF revision 6.0, February 1992): the Architectural fields and the Resolution fields, both Baseline TIFF fields. TIFF 6.0 has paragraphs in "Section 4" (another coincidence) that further define these fields given bi-level images. Note that in the text mentioned, compression type 2 is used as a working example whereas the Exchange standard employs type 4.

In the future the TIFF Informational fields and Document Storage and Retrieval tags could be exploited to make images self explaining. Contrary to the (unused) multi-page feature of TIFF, these fields or tags can be handled and inspected by the user with any common file viewer: the information stands out as readable text among garbish (though this garbish is a Sleeping Beauty: it is the scanned and compressed image). Either direct at the beginning of the TIFF file or at the very end.

The Steinmetz Archive will help with all necessary

documentation and expertise if a data archive or library wishes to implement it's own TIFF compression type 4 writer or reader. The Archive will also provide a testbank service to judge if one starts using the Exchange Standard with indeed the right TIFF 4 format.

Further reading, commercial conversion packages, shareware TIFF viewers/printers and the anonymous FTP availability of the excellent "Sam Leffler" toolkit to start programming for TIFF - are mentioned at the end of the paper.

The Katz and Cohen proposal, also defines for bilevel images but is less strict in TIFF compression type and resolution of the scanned images. The Working Group allows even uncompressed TIFF for example, though TIFF type 4 is to be preferred. Multi-page files are supported. The perspective however of the proposal seems different from ours: Katz and Cohen have a strong emphasis on the actual transfer of images and leave it to the sending and receiving parties to negotiate a variation of their standard that both can handle. Our emphasis is on establishing an Exchange standard that ensures the researcher that he or she can always use the images received. Hence one compression type and so on.

Producing images by document scanning.

Given the nature of printed text and line art, scanning black and white at 300 dpi or more is adequate. Issues of preserving grays in the original or even colour are not involved. Each separate page scans into one compressed file and these files are kept together by proper file names and subdirectories or folders to mimic the original chapters and separate volumes.

Especially in a closed system scanning station with proprietary software it is not always made clear by the supplier how the images are stored in terms of image format and compression type. The compression (decompression) more over is often done by additional, separate hardware. In order to exchange it is imperative that the system has exporting facilities so that images can be converted to an established standard. Next these converted images should be available in a general file area, open to networking and further handling.

Scanning and storing in a Dos environment without specialized image bank software is more or less open by definition and can produce accessible images in a TIFF format straight away, though often the less compressing TIFF type 2 or even TIFF Packbits is used.

Both closed and open approach don't necessarily produce the TIFF type 4 chosen as exchange format straight away. The following steps have to be taken:

First case: a scanning station with own image database software and hardware compression and decompression. Used for systematically scanning all paperwork of a number of studies. If there is a choice at all and if one only scans bilevel (black and white) printed source, TIFF type 4 is a very good choice for an internal storage format as well. If the software is custom made, even the TIFF documentation tags can be filled in to make the image files self explaining.

Second case: if the scanning station comes as is, caveat emperor:

- given your computing environment, the image files should still be open for access by other software

- if internal format "type X" is used, this format should be convertible by both software and hardware to the required TIFF compression type 4, Exchange format. "Hardware" means that the scanning station software asks its compression/decompression board to do the conversion. But can it ? "Software" means that a separate tool is available or can be written to convert to TIFF type 4. If necessary both approaches can be split in successive steps: if only TIFF 2 or 3 can be managed than an off the shelf graphics format conversion package, can do the TIFF 2 or 3 to TIFF 4 step. Note that a scanning setup that uses a storage format that depends entirely on separate hardware, is a timebomb for a data archive or library. At some point in time the hardware board will fail and if a replacement is no longer available, the whole collection of scanned images is rendered useless.

- if TIFF is used as internal format (and given bilevel images) it is very likely to be TIFF type 4, because it is the most suitable compression scheme. If not than probably a standard package can do the conversion to the TIFF 4 Exchange format.

Third case: a simple scanner setup with some software for viewing and image manipulation, attached to a PC and used for per request scanning of documentary information.

All involved packages in this case handle TIFF but only aiming at import into desktop publishing software, of the wrong, simpler compression types. Standard conversion packages can change into the required TIFF 4.

Note: it is desirable to use this most compact TIFF 4 format also for storage to accommodate future similar requests.

Transferring a group of images.

One of the features of the TIFF format is storing several different images (pages of text) in one file. Scanning and storing however is done on a one page, one image basis. Therefore extra processing would be needed to use this feature and it does not really improve the transfer. Many smaller files travel easier over a net than one big chunk and dissecting and decompressing would be more complicated for the end user. Consequently this feature is not yet part of the Exchange standard.

Compressed images are binary files so in sending over a net, care should be taken to use the transfer protocols accordingly. If the facility is Text oriented - like Send File in BITNET, extra steps are necessary, (uuencode and uudecode) to nevertheless preserve the binary nature of images. The FTP protocol used in Internet, offers both binary transfer and multiple put to handle a stream of images with one command.

Name giving of the image files can be a problem: different operating systems have different conventions and too long a name gives trouble if the receiving end has a simpler scheme. The best seems the DOS convention (8 characters, dot and a three character extension) just because it is the most restricted one and will fit into any other notation. Unfortunately this means that if TIFF type 4 is used for internal storage as well and the platform is Unix with rich name giving possibilities, one still needs a conversion of the file names. This can be done while copying from the image storage area to the transfer area.

The user side: transforming back the images to screen, paper or OCR file.

Implementing an Exchange standard, the focus of attention should of course be the ease of operation for the user to wave the magic stick and have the requested documentation on screen or on laserjet printout. If the data archives or libraries offering the service take the trouble of converting to one and the same exchange format (TIFF 4, 300 dpi, single page per file), the steps to be taken are well defined. If text processing software handles graphics, it can import TIFF (but only the simpler compression types) and print it out. Drawing or imaging software does the same and offers viewing. Disadvantage of this approach are the required expertise to import a received image into one's word processor and above all: get it printed. Drawing software is specialized in importing, viewing and printing images but certainly not everybody masters that kind of software. For various platforms good shareware software is available that reads TIFF (again: the simpler compression types) and lets you both view and print. All software requires a setup of - for the DOS environment - 286 or 386 PC with VGA and a laserprinter available. This printer should be equipped to also print larger chunks of "graphics": an image holding a full page of text is a bit too much for laserprinters with limited graphics capabilities.

A few pages of "how to do" information for some software common to most researchers, could ease this Do It Yourself approach to use tools already available. Such a document will be made available by the Steinmetz Archive and will also point out the usefulness of some shareware already specialized in doing the job.

Remains the demand by popular software to be on a "simple compression TIFF diet". Clearly a conversion aid is needed to change TIFF compression type 4 into one of the simpler schemes mentioned. The Steinmetz Archive has written a tool to do just that and will make it available to data archives to bundle it with requested images. Ideally this tool will also have the option to print the decompressed image directly to a HP LaserJet printer. Printing is much easier accomplished than viewing because of the wide variety in display hardware and the limited resolution or viewing area of most PC screens. The HP printing option will most certainly be included in a future update by the Steinmetz Archive. (Postscript printing would be desirable).

With a growing user base for TIFF type 4 bilevel images, software makers can be urged to implement Importing and handling this TIFF compression scheme as well. Then the separate conversion step is no longer necessary. In the area of OCR programs this already is the case: tests showed that the market leading OCR programs for DOS read the Exchange standard format without need for conversion. (Recognita, OmniPage, Prolector, Liocr, K5200)

Summing up, these are the steps for the user to transfer back:

1. use the free conversion tool to simplify the TIFF compression type

2. with the help of the Cookbook print or view the images by applying existing software: either commercial packages or shareware. The choice should be software commonly available to social science researchers. (The shareware can be redistributed together with the Cookbook - both as files through networking and electronic mail)

3. use the Exchange format without further ado (for example with OCR software)

Further reading, availability of software and source

code.

"A file format for the exchange of images in the Internet"
by the Network Fax Working group of the Internet Engineering Taskforce. Authors: Alan Katz (Katz at ISI.Edu) and Danny Cohen (Cohen at ISI.Edu). Phone: 310-822-1511.

The Sam Leffler TIFF toolkit:
by anonymous FTP:
sgi.com: /graphics/tiff/v3.0.tar.Z (192.48.153.1)
email : sam at sgi.com
(This toolkit also includes the TIFF 6.0 specification)

Aldus can be reached at CompuServe, again the TIFF spec's and a much simpler TIFFRead toolkit

Commercial conversion packages to and from TIFF type 4:
(DOS) HiJaak
(DOS and SUN OS): Image Alchemy
(UUCP: hsi at netcom.COM or:
apple!netcom!hsi)

(DOS) Shareware graphics viewers and printers:
among others:
Graphics Workshop
Optiks (all three don't handle TIFF 4 so need the free conversion tool first)
Pixfolio (runs in a Dos Windows environment)

[1] Presented at the IASSIST 92 Conference held in Madison, Wisconsin, U.S.A. May 26 - 29, 1992.

This paper was also presented at CSS92, May 1992, Ann Arbor, Michigan.