Digital Historical Maps

Report from WP10

Evaluation of Storage and Archiving

Final version 1.2

Deliverable 23

April 9, 2001

Tanja L. Skovsgaard

National Survey and Cadastre of Denmark

Table of Contents

1 Scope

The aim of Work Package 10 (WP 10) is to evaluate the storage and archiving environment developed in WP 4. It is important for the future use of the information that these environments function satisfactorily and that the management of the material is guaranteed. In the course of evaluation each institution has reported on its experience relating to functionality and management. Based on this evaluation the project recommends improvements.

A satisfactory storage and archiving environment does therefore contain data of high quality, it is secure, well maintained, ready for the future and with good access for the professional and private users.

Tasks:

Evaluation of the storage/archiving environment developed in WP 4

Recommendation for improvements

Revision of management costs.

The maps and documents included in this project are of high quality and of high historical value. It must therefore be expected that data will be of interest for civilization in many hundreds of years to come. By making the material accessible over the Internet and giving options or ordering the files the wear and tear on the material is avoided. For historical maps and documents the sole interest is not just to achieve readability but also to secure the material for the future. But from a fund raising point of view the accessibility for the majority is more important

From a general point of view the pilot system placed in the National Land Survey of Sweden must be considered appropriate for its purpose: To capture, store, visualize and deliver digital copies of the historical maps. However there are experiences that cause for reservations. These reservations will be covered in the evaluation of the specific points.

2 Data in the image files

The maps are delicate to scanning and also very costly to scan because most of the work is left to a manual and time consuming process with manual controls of quality included. The right choices should therefore be made not only for the production but also for the storage. And with the right choices for storage could against all odds be achieved that the digital media in time overcome the very long lasting media: good quality paper.

For an archive the choice of format for storage will always be the choice between quality and security on one side and costs of storage media on the other. For scanning has been chosen TIFF 6.0 for maps and TIFF group IV for text documents. The original scanned images are made in a resolution of at least 508 dpi and afterwards saved as files with 254 dpi. The latter is in the project referred to as the original images even though they are copies of original data. Files in TIFF are in general too large to work with over the WEB. Therefore it has been chosen to compress data for web-use. After a lot of consideration it was decided to use wavelet-compressing methods with techniques and software from MrSID. A security copy has been made of the so-called original images and of the compressed images.

The choice is the most ambitious from an archive point of view. It means that it is not only chosen to store the original images, it has also been chosen to store a compressed format. And there is backup on both of the copies: the original images and the compressed files. Together with the original scanning that means that all together five copies of the same image is stored.

The solution is optimal from an archive point of view when it comes to security for the future and safety in the present. But from a storage point of view it is costly in space and financial resources because of the large amount of stored data. It is therefore a solution with large challenges.

2.1 Format of original files

The storage of the original images has been chosen to be in the TIFF format. TIFF is considered to be the most widespread used format (defacto standard) not only for archiving but also in general on the market. Therefore nearly all established software programs supports TIFF. And TIFF is a standard well known to most customers.

TIFF is especially popular in the archive world because the format itself is not adding any further distortions to the image in addition to the distortion from the scanning process it self. Choosing TIFF is therefore best way to secure that original data is stored in a format without loss of quality. An important issue when dealing with archive material such as historical maps and documents. Put on top of that that governments making demands for raster data in public archives require some version of the TIFF format. And that it is possible to geocode TIFF files.

Regarding future conversions to new formats it is clearly an advantage that the original files are stored in the TIFF-format. The format is loss less and very widely used. It can therefore be expected that software for new formats will be able to read and convert the TIFF-files and the distortion caused by the conversion will only be the distortion caused by a later compression.

The only reason for choosing other formats is the large size of the TIFF-files. This consumes a lot of time in several procedures and takes up a lot of space. For the consumer this means that it is difficult to order the wanted files in the TIFF format. And in reality unadvisable to get the TIFF files over the Internet. But in many ways the archives are already used to large amount of data and find that other formats (GIF, JPEG) would lead to too much information loss.

Therefore because of its good qualities for archiving, because of the TIFF-files that are going to be delivered on CD's are probably on of the most widely used loss less raster formats, and because of the fact that the TIFF-format is a de facto standard on the market, then TIFF must still be considered the best choice for both the customers and the archives.

2.2 Resolution and colour depth

It has been chosen to scan in a higher resolution and then save and deliver the image files in 254 dpi. The colour depth of the maps has been 24 bits RGB Colour Space (minimum 24 bits in scanning and 24 bits in saving). Some concern must be raised regarding the resolution and the colour depth. The decided resolution and colour depth meet some of the present demands. And the quality level must be considered sufficient here and now – but also the absolute minimum. A risk when considering the lack of knowledge about future needs.

The resolution of 254 dpi is giving a pixel size at 1/10 of a millimetre. To compare then in National Survey and Cadastre of Denmark it is not uncommon to have lines on the maps with a thickness at 1/10 of a millimetre or even less. With 254 dpi there is left no room during the later compression.

When it comes to the users there will probably be satisfied users. These being primarily people that need some illustration for a purpose, e.g. family history or area planners needing a part of a village to illustrate conservation of hedges in a letter. Others like archaeologists and people working with research within landscape use and development can be expected to have greater demands. So far the Danish users have split into these to groups, were the latter have asked for an improvement of the resolution to 300 dpi. 300 dpi is already the commonly used standard in National Survey and Cadastre, and the original scannings are expected to be converted into this resolution. The difference between a resolution of 254 dpi and 300 dpi is only visible when using high quality paper.

Also the colour depth at 24 bits in the scanning must be considered the absolute minimum. Although 24 bits colour depth is appropriate for the human eye it leaves no room for potential digital adjustment of lightness, saturation and hue. If the maps are scanned with e.g. a 30-bit colour depth the original images could be saved with a 24 bits colour depth after final rectification.

An expansion of the dpi or the colour depth will have impact on file size, thus effecting longer production time, larger files on the net and larger cost for storage media. Whether changes should be made most conclude on the evaluation of the users. The colour depth cannot be improved when first production has finished. However the resolution delivered for the consumer can be changed as long as the original scannings are saved. But naturally it takes time to convert data all over again.

But there is left a disharmony between the high ambition of choosing to store the same files as: original images in duplicate and the compressed files in duplicate. In Germany and Denmark the original scannings are also saved. Four copies of the same data take up a lot of space. Space which perhaps for some parts could have been used to higher the quality of data.

The text files in the Swedish database is a matter of its own, as they are scanned from black/white (1 bit colour depth) microfilms and compressed in TIFF Group IV-format. Sometimes the text files captured from microfilm might be difficult to read due to the quality in the microfilm. This is the case especially concerning maps that originally are coloured or have small measurement numbers. The texts smear and are hard to read. The discolouring of the background has turned into a lot of black pixels, which makes the text even harder to read and interpret. By tests made in the project a better result is achieved when using the DjVu format instead of MrSID when converting.

The future method to be used in the National Land Survey is therefore to make direct scans of the text-files where it is necessary for interpreting the information in the described TIFF Group IV-format. The future archives will therefore have resources for scanning on demand.

If this method does not prove satisfying for the users an natural recommendation would be to scan with higher colour depth, e.g. as it has been done for the German text files and some of the texts from Swedish land books. They are scanned using the same parameters as for the maps.

2.3 Compression

The compression technique in MrSID software, wavelet is well known and used by many compression software producers. Wavelet compression is the compression method that gives the

highest compression rate compared to the loss of quality when dealing with continues tones images. For black and white images as the text files there are better formats, but in the prototype system the same compression technique has been used to ensure that only one kind of software is introduced. This has been chosen to minimise the problems for the users and reduce the development work. In a future production system it could be something to address.

Though the wavelet compression is well known by software producers it is new on the market and the MrSID is not commonly used by the professional customer or by the public. For the customer who wants to use a compressed format the wavelet compression will properly not be the first choice because of lack of support in the common dtp and picture publisher software products without additional plug ins. Actually the necessity to download and install plug-ins for the browser can hamper the use.

Alternative choices would be pointing at the ISO-certificated and international standard JPEG where the pictures can be viewed directly in e.g. Microsoft Internet Explorer. JPEG however does not offer the high compression rate. And in contraire to the MrSID files JPEG files are not geometrically correct.

Because of the high compression rate the wavelet compression used in the MrSID format is well suited for the web. The image size transferred over net is proportionately small. Despite this, processing time might be experienced to be long, especially on modem connections. To improve the processing time (and reduce the consumption of storage) the rate of compression could have been set to a higher level than used in the project. Higher compression would naturally reduce the quality. For presentation on the Internet the quality would might still be satisfying, but for further use, in GIS applications, printouts etc., the quality would not be as good as required. With the used rate of compression the images are nearly as good as the original images and meet the expectations for a solution that distributes images over the Internet. The chosen rate of compression is probably the best approximation between quality and file-size.

An important feature is that the MrSID format allows geocoding. This is important for certain users and in more advanced future product versions. Also it is very important for search functionality in future distribution systems because the only common link for all kinds of maps are their coordinates.

With concern to the compromise between quality and file-sizes this wish for geocoding can explain why the DjVu was not chosen instead of MrSID. DjVu gives hard competition to MrSID. The idea of DjVu compression is that there are two types of objects on a printed page: characters and pictures. The pictures are smooth, do not have to many edges, and have colour content, while the characters have many hard edges, but do not have much colour and texture content. Those two components should be separated and compressed with different methods. In DjVu the background and pictures are coded with a wavelet based technique called IW44, while the characters are coded with a new bi-level compression technique called JB2. The result is that with higher compression (and thereby smaller files) the images appear better than with the MrSID compression. But even though the images appear better, then DjVu images have a tendency to melt the colours together and to leave out some of the grey tones, the latter resulting in lack of information. The problem with the DjVu is that it has no possibilities for geocoding the DjVu images.

The conclusion of the chosen compressing format is that it perhaps was not the only good choice. But it was a good choice. The main problems are the use of plug-in and the fact that we are not dealing with a standard on the market. But these are only the copies of the original files, and when new and better formats appear on the market it is possible to convert the TIFF files into these new formats.

3 Data in the registers

In the project is involved four registers from: University of Greifswald, National Land Survey of Sweden, Rikes Allmänna Kartverk in Sweden and National Survey and Cadastre of Denmark. The register data is stored in the individual countries with back-ups and a copy of data is placed in National Land Survey of Sweden with their security procedures.

Register data are mainly captured from archives and therefore static and not necessarily updated in other aspects than eventual corrections or expansions with new tables. Data concerning the administrative division of today might naturally change, and a relation between the old division, mirrored in the archives, and the actual division therefore might be necessary to maintain.

The information in the registers must be expected to be in accordance with the needs and information available in the three different organizations. However some comments could be added about the openness of the registers.

3.1 University of Greifswald

The register from University of Greifswald appears to contain the tables and fields necessary. There also seems to be room for adding new kind of materials e.g. type of map. The data stored is not as detailed as the Swedish but are more kind of metadata. However it gives the user sufficient basic information. A view in the contents of the register shows that some of the fields for the maps origin year are empty. If a search by year is later implemented this could be considered changed.

3.2 National Land Survey of Sweden

The Swedish registers are divided into two parts: the part in National Land Survey of Sweden and the older maps in Rikets Allmänna Kartverk. For the latter the data seems adequate and able to contain various kind of material. It is very detailed and can therefore be used as basic for research. For the register from the National Land Survey of Sweden much the same positive things can be said. But looking trough the fields the structure a few places seems to need more discipline. As an example then ways to store material takes up four fields where an “X” can be marked if it is the right one. A natural and space saving structure would be to have only one column and here write were the material is stored either in letters or by a code.

Fields are made for coordinates in both registers. But the advantage of this change is probably too small to pay for the cost of the change.

Searching by year in the database of the Swedish material reveals some problems both because of empty fields but also because of inhomogeneous structures for typing in the years. The problems are examined in Work package 2, Appendix 1 and indicate that if a search by year is later to be implemented in the system, then the database should be updated by splitting the field of performed year into two fields and by register the homogeneous.

3.3 National Survey and Cadastre of Denmark

The register from National Survey and Cadastre of Denmark contains only little information. Especially there seems to be lacking information about the quality of the scanning, important information if the standards used in the DHM project are not followed in new scanning.

But more important the Danish register is designed especially for these kinds of maps. When other kinds of maps and perhaps new kinds of material are to be added the register will have to be changed. This applies for information about county, municipality and parish but perhaps also for other information such as lacking information about state of conservation, year of survey, year of revision, name of village etc. There is however fields made for coordinates even though there are no coordinates connected to the used maps. Like the register from University of Greifswald the data is more to be considered metadata, in time more information could be added for the use of research.

The examination of search by year in Work package 2, Appendix 1 points out that there also should be some changes made in the Danish register when it comes to the stated year of the material. Text should be removed from the year fields into the remark fields, and the empty fields in the register should be filled.

All three organisations are saving information about the compressed image (size, data of compression). It could be considered to save information about the software version used for the compression as well for later error tracking.

In general there has been made fields for coordinates. The use of coordinates has been left out of the DHM project because it has been estimated a far too large effort in a first phase. But the fields secure future development.

In general some changes should be made concerning the mentioned years in the registers. The largest changes are needed in the Swedish registers, but this is only naturally since the Swedish registers are containing different kind of material giving problem with inhomogeneous registration.

4 Data structure

4.1 Data model

The use of a relational database should secure the possibilities to expand the database including adding new kind of data, as for example vector data files.

Without an E-R diagram in the WP2 it is difficult to decide if the data model in any way is hampering the ability to expand and add further data.

It could however be an advantage that the register data and the images files are separated. The linkage between database and images files is the file name and it does not matter what kind of data is linked to the file names.

4.2 File name conventions

For all file name conventions used in the project there has been chosen to use the filename as information carrier. Thus information about map type, administrative identity, numbering and for the Swedish material also several other information is build into the filename.

This strategy of making the filename an information carrier can be discussed. Because information about all the files is stored in the database it could be considered to use artificial filenames generated by the software. This will save time for typing a controlling the filenames without reducing security or information in the system.

Also the use of meaningful filenames could make it difficult to add further datasets to the system, because the filenames needed for new data could be conflicting with the filenames for the data already in the system.

Even though the archive doesn't benefit from the filenames as information carriers then the use of meaningful filenames can be of important in the production process as an extra caution against loosing the image files.

If it is later chosen to use artificial filenames without information, it should be warned against distinguishing between uppercase and lowercase in filenames, because not all operation system are case sensitive in file names.

But the choice has been settled on meaningful filenames and from this point of view all the file name conventions seems satisfactory in identifying the unique files. In both the Swedish and the Danish file naming there is however some discrepancy between the older original archive structure and the modern administrative divisions which becomes apparent in the search system.

5 Platform

There can be said to be all together not one but four storage systems. The WEB application with the MrSID files and the register data are all kept together at National Land Survey of Sweden. The storage of TIFF files is stored separately in National Land Survey of Sweden, University of Greifswald and National Survey and Cadastre of Denmark.

5.1 Database

The registers from the three different countries involved in the project are stored at the National Land Survey of Sweden in an Oracle database together with the MrSID files and the Swedish TIFF files. There is a security copy of the Swedish data. At the University of Greifswald a duplicate of the German register data is stored in an Access database. The image files are not stored in this database but separately, which cause some manual work. In National Survey and Cadastre of Denmark a copy of the table are stored in Excel as for the data in University of Greifswald the images files are not stored in the database.

National Land Survey of Sweden so far maintains the Oracle database. This means that the participants from Denmark and Germany have no possibility to make remote changes of the files or the registers in the prototype.

In all three organizations it is standard to use Oracle. MS-Access is also either a standard or a well-known database. Therefore all participants are confident that they have the necessary manpower and knowledge to maintain the database and later to make changes. The databases format enable export to other database formats if later required.

Database search in the prototype (Oracle) is tested fast. The speed and liability in the search process is very much depending of the database. The amount of records generated in the database is so far relatively small and should not cause any troubles. The retrieval of the raster data is depending on the file system including the storage media.

5.2 Storage in the WEB application

For purpose of the prototype the MrSID files, the overview maps and the three databases of the German, Swedish and Danish material are placed in one system in National Land Survey of Sweden. The data are stored on a hard disk in the organization. The demand of the system is that it should have the capacity to the existing data that today occupy 50,5 GB. The system should also be ready for expansions to a full-scale solution. Only estimations for the Swedish compressed material have been made in the WP4, the demands here settle with all together 4000 GB.

Cheaper storage could be achieved by using optical disks instead of hard disk. This will however affect the access time. The choice of hard disk storage must be considered the best for the purpose of a WEB application.

5.3 Storage in National Land Survey of Sweden

The demand to the system is that it should have the capacity to store the compressed files, the original image files, and for back-up a copy of the original image files and the compressed files. Existing data in the system occupy 1500 GB. The system should besides that be ready to contain even more data in the future, all together 65000 GB.

The original data and a copy of the compressed files are stored on DLT IV tapes in a HSM computer system. A method used in the organization since 1990. Using tape instead of magnetic disks or optical disks makes the access to the system slower. The choice has been made in order to keep the costs down from the point of view that the original files are not often used and not expected to be nearly as fast to retrieve. If the choice proves to be a problem it is possible with the HSM system to add other and faster media for instance hard disk for the files often used.

National Land Survey of Sweden has many concerns about the huge amount of space used to store the image files. It could therefore be considered to compress the TIFF-files with the LZW compression that is loss less. Besides the impact on the demand for storage space, the smaller files could improve the search and retrieval times when the files are fetched from the tape archives.

Another issue to consider is to restrict the compressed files to one copy. In the present configuration two copies of the compressed files are saved: one copy inside the firewall and one copy outside the inner firewall (in the prototype). Because there only is read access from the Internet to the files between the two firewalls these files must be considered rather safe.

5.4 Storage in University of Greifswald

The part of the system in University of Greifswald is easy to maintain with a currently low amount of data. The system contains the original scannings, the original images and a copy of the latter. Data is stored on CD-ROM and back up is stored on tape. The German material occupies up to 9,5 GB. This means that an expansion from the present 140 images to the planned 1500 images roughly would end up in 100 GB.

Metadata and MrSID files are stored not only in the WEB application in National Land Survey of Sweden but also at the University of Greifswald. These extra data take up approximately 341 MB stored on a hard disk

In University of Greifswald it is standard to use CD-ROM for storage media together with a disk server. Considering the access time it seems to be the wrong choice to let the working files stored on CD-ROM and the copies stored on the faster media tape. This might not be a problem now, but when there are 1500 image files it should be considered to change this in a future system so the work is done on the tapes.

The choice of not making the storage only on CD-ROM seems advisable considering the risk of these media not being readable in the long term. Tests from 1995 shows that depending on temperature and sun light there can be errors in 3 years[*]. If copies from the tapes are made unto CD-ROM continuously the choice of tape and CD-ROM together is acceptable safe. But considering the work put in copying especially the TIFF files, then it should be considered to choose the more costly media tape. In the end it could be more costly to keep the CD-ROM solution. Add to this the benefits from using tape instead of CD-ROMs when it comes to accessibility.

As in the Swedish storage it seems superfluous to have copies of the MrSID files. In this case however the data are of very small size and it seems natural that the national participant want its own copy.

5.5 Storage in National Survey and Cadastre of Denmark

The system in National Survey and Cadastre of Denmark should in the first phase be able to contain 700 GB. This includes original scannings, original images, a copy of original images, a copy of MrSID files and copy of the tables of the database.

For the storage is only used CD-ROM. This is a risky choice because of the insecure stability of the CD-ROM, as mentioned earlier there can be errors within 3 years. Beside this then the disks limited of 650 MB per disk means that data is difficult to handle when it comes to search, conversions, compressing and distribution. The media also takes up a lot of physical space. The choice is the one of yesterday and even though CD-ROMs are cheap, then another media must take over. Because of an expectation of high access to the data the National Survey and Cadastre is expecting to store the data on disk systems in near future.

As for the copies of the MrSID files the same comments can be made as for copies in University of Greifswald.

[*] SKANNI-project 27.1 – 31.8 1995

5.6 The server and database system in the prototype

The system implemented at National Land Survey of Sweden contains both the WEB-application and the original material from Sweden. In general it seems to be based on mainstream technology and there is no reason to believe that it will not be easy to handle the system. This goes for both the daily operation of the system and future extensions to the system as well as taking the backups. The platform is standard in National Land Survey of Sweden and therefore there is the required information about the platform in the organization. Except for the use of Samba, the pieces of the system could also easily fit into the environment of National Survey and Cadastre of Denmark because of the similarities with the equipment used here.

One of the main servers in the system is based on the SUN Starfire E10000. This server is a high-end UNIX server with build in mainframe capabilities and the stability and security will benefit from this. The server is highly scalable and can have a maximum of 64 CPU. The bus-system as well as the I/O systems has similar high performance. The draw back of using this server configuration is that the price of the server that is relatively high. The new mid range SUN-servers will have some of the same features, and it could be well worth considering one of the servers in the midrange instead if the scalability of the E10000 is not needed.

The system in National Land Survey of Sweden has two firewalls, an inner and an external, around the WEB page. That is enough security towards outside. For the inner security a backup is taken every night. National Land Survey has two large robot systems for backup. The chosen Solaris system is recognized as a stable platform (more than for example NT). The organization also has rules for maintenance of both the WEB server and the firewall. The system is thereby considered secure, stable and well thought through.

5.7 Delivery

It is likely that the bottleneck in the delivery of data to the customers will be the production of the CD-ROMs. Because CD-ROMS is a de facto standard for delivering data too voluminous to be delivered on floppies or through the Internet, there is no real alternative. It will be possible to change the CD-ROM burner to faster technology when appropriate and later on supplement or replace with DVD-burners when the technique has matured and the market is ready. All three organizations with storage seem to have fast burner stations. The problems are therefore not on the hardware side.

6 Applications

6.1 Workflow for access and delivery

There have been some problems in developing the search application based on Oracle. The digitised version of the register was earlier in Access. The search possibilities have not been as developed as originally intended.

Searching is possible only from administrative information (county, parish, municipality) not via geocoded data or coordinates. Also presently it is impossible to add vector data to the prototype. Other additional information is stored in the meta data catalogue, which could also be of interest for customers, but the prototype system does not provide search routines for these (e.g. year of the

making of the maps, map type, map author). All mentioned facts leads to the necessity of further development of the search system.

The user interface from the homepage works quite smoothly and seems acceptable fast so long as the user wants only to view the data on the screen. Even with the mentioned lacking in search and even without geocoding this is a massive improvement of access to historical data for all parts concerned.

Delivery is, however, another matter. The delivery system for whole compressed files over the Internet is not fulfilled, although this was the original intention. Instead it is only possible to download a test file from each country, and thus it is not possible to make an order with direct delivery. The user can instead make an e-mail ordering data to be delivered on a CD-ROM or over the Internet. Some users might loose interest without the possibilities for direct delivery.

The functionality of downloading the test file works without giving the user any trouble and so does the e-mail ordering. But it must be expected that the effort needed to complete the ordering application on the server could be considerable. And until this is done the costs based on manual handling of every order remains.

This takes up a lot of manpower. After the ordering via e-mail is sent to the correct part, then there must be a search in the database for the location on CD-ROM, then this is copied onto a new CD-ROM, it is sent to be billed, and then sent via ordinary mail to the customer. Sometimes the customer's demand to formats and resolution differs from what is stored and there also have to be a conversion before data is burned onto the CD-ROM.

To achieve an automated production system, functionality for delivery from the original file storage on desired media and in desired formats has to be developed. The system must also be completed with automated invoice routines. So long as these activities are manual it is resulting in costs that makes it impossible to supply the market with the material at a reasonable price level.

When it comes to the system to delivery, ordering and billing there has only been achieved a solution on the lowest ambitions. But it is considered a natural place to cut in a prototype where data from three different organization and countries is to be distributed. Distributions systems in the organizations have to development together with at system for other kinds of data.

The security in ordering application is unknown in relation to misuse from the users, but it seems it could be quite easy to make orders using the names of other persons.

6.2 Development tools

The programming or scripting languages used seem to be some of the widely used and there are no objections against them, except that the use of Oracle Forms will hamper the use of the application together with other RDBMS than Oracle. Furthermore, it cannot be expected an organization will have expertise in Oracle Forms. However Oracle combined with Oracle Forms is standard in the National Land Survey of Sweden, and their computer department has enough knowledge for development and running of Oracle systems.

The script language PHP is introduced with this project to enable a pure Solaris platform. The knowledge of PHP is therefore limited. ASP is the most used script language as to the rest of web

applications. The possibility to run ASP on National Land Survey of Sweden's Solaris system has been introduced lately. For building the web pages National Land Survey of Sweden has used Dreamweaver from Macromedia and FrontPage from Microsoft. To fix the objects in the web pages is used Photoshop and Illustrator from ADOBE.

6.3 The MrSID software

The wavelet compression used for the compressed files has not been used for many years, and are still depending of vendor specific software.

The MrSID software comes with a licensed converter that converts (and compresses) image files into its own format. This format can then be handled in different applications. The MrSID Image Server, MrSID Plug-in, MrSID Stand-alone Viewer and extensions for Photoshop are all free for download from the homepage of Lizardtech. The software, it is pointed out from the project members, is very easy to use compared with other similar software. The MrSID software is convertible and can be used on many computers.

Compression, decompression and conversion are important aspects when dealing with digital copies of archive material. The compression and decompression of MrSID files are fast and easily handled. Batch files can be run and thereby work resources reduced. With the MrSID plug in it is easy to convert the SID-file to JPEG or BMP. For conversion to other formats must be used an image editing software e.g. Photoshop.

MrSID has a strong connection to the GIS/Map industry through collaborations with ESRI and USGS and besides the MrSID software the MrSID format can also be used in common GIS environment as ArcView, Arc Info, ERDAS and MapInfo

The MrSID software, although much used in the US, is rather unknown in the three countries involved in the project. Resellers of the software do not have sound knowledge of the software, how it shall be implemented in different technical environments, and what to do if problems occur. That has been a problem during the project, combined with the fact that the support from Lizardtech in Seattle has not always been as easy to establish and as fast as needed.

During the project time there has been experienced problems with presentation of text files saved as one bit TIFF group IV. After converting they cannot be presented in the quality assumed by using the MrSID Image Server. When using the stand-alone viewer the image quality is satisfying. The problem is addressed in the new version, Content Server. To make a solution here and now for the prototype system, the text files have been modified to be of eight-bit origin. This solves the problems with the MrSID Image Server. But the problems were not fast solved and have had a negative effect in the development of the user interface, as no material has been available for tests until recently.

7 Costs

The system costs as described in the WP 4 are what to be expected of a system of this kind. But there seems to be lacking information about the costs for manual delivery of files on CD-ROM. And the estimated costs do not include the costs needed for development of an automated distributions system with automated workflow for production system, delivery and invoice routines.

8 Conclusion

Digitizing of maps and documents is a technical matter. Techniques and tools used might differ over time, especially as conditions change in the fast development within the areas concerned. What is absolutely necessary is to know what sort of material and why it shall be digitised. The costs are high. Therefore a strategy, founded in the organisations ordinary activity, is needed. The archives strategy of the National Land Survey of Sweden is the base for solutions within this project.

The format and resolution in the specification for scannings of maps are considered satisfying. But it is recommended to change the specifications for scanning so that the minimum colour depth is increased and there is room for potential digital adjustment of lightness, saturation and hue. The format and colour depth of the original images are considered satisfying. But it could be considered in the organisations to change the specifications of the resolution to 300 dpi. The format of the compression MrSID is satisfying both in size and quality. But it could hamper the use that the format is not a standard on the market and that the customers have to use a plug-in.

There have been problems with some of the text files in the Swedish database. The quality in some cases seems to be to low and there have been problems with the software used in the prototype. A solution for the software appears to be on its way. But a natural recommendation is to examine further what could be done to higher the quality.

The register from National Survey and Cadastre of Denmark contains only little information and the thought of redesign should be addressed before expansions to new types of maps.

It is recommended that the registers in all countries are made ready for search by year by filling the empty fields, and in Swedish and Danish registers to make the changes concerning more homogeneous registration and concerning leaving words out of the fields for year.

Without an E-R diagram in the WP2 the system is not documented properly. This must be done for the data in all the three registers.

It has been chosen to use filenames as information carriers though this can reduce the flexibility of an electronically archive system but it provides higher security in the production phase.

The chosen platform is satisfactory and well chosen. The security of the data is high. The data is protected by firewalls, access control and backup. There is no indication if the tapes from the backup are located on another physical location, but that is expected. However, this is more an organizational issue than a technical issue. The system maintenance and scalability benefit from the chosen hardware and RDBMS as well as the stability.

The selection of storage media will be a compromise where the user has to choose between speed and price. This choice will be up to the individual organization. But it is recommended giving up CD-ROMs as only base for storage media as it is done in National Survey and Cadastre of Denmark.

To achieve an automated production system, functionality for delivery from the original file storage on desired media and in desired formats has to be developed. The system must also be completed with automated invoice routines. So long as these activities are manual it is resulting in costs that makes it impossible to supply the market with the material at a reasonable price level.

As already mentioned in the beginning of the rapport the pilot system gained by this seems to be well suited for its purpose. Some comments could however be added. And here the most important aspect is that the costs for software and hardware should not be the largest issue. The most costly is the manual work invested in scanning and converting. And the manual work in delivery. Material should be scanned and saved in a quality ensuring that it never will be necessary to repeat the process. And the workflow in the archive and delivery system should be as automated as possible. Saving money on using one media instead of another. Or saving cost on media by reducing the quality of the files is unadvisable.

9 Experiences in the project and plans for the future

9.1 University of Greifswald

It is planned to investigate the results of the project and evaluate especially the system on aspect of storage of different kinds of digital objects (maps, texts, photos, sounds, videos) and of the metadata structure and handling.

Another topic of further evaluation and changes will be the implementation of the geographical search system and routines. Unfortunately the reports of the DHM project do not provide any information on the structure and software technology of the part of the prototype, which was made by ESRI Sweden.

If the results of further evaluation are satisfying, it is planned to adapt the prototype system of the project, or parts of it, and continue the work. In cooperation with the Landesarchiv and the University of Greifswald computer centre it is aimed at establishing a digital archive for all Swedish Matrikel Maps of Pomerania and all of the 70 text volumes. The Landesarchiv Greifswald now already starts with photo scanning of its 1500 Matrikel Maps. The experiences from the DHM project, which in Greifswald serves as a pilot project, now are very valuable.

It also will be evaluated, if another technology (IBM content management – digital library) is appropriate as an alternative. This may become necessary, because the university looks for a homogeneous technical solution for the storage and archiving not only of digital maps and texts, but also of other digital objects (catalogues of archive and library, videos, sounds, photos, etc.), and because this should be done in the same way (the same system) for different applications at University of Greifswald. Therefore, different parties will be involved in the decisions.

The experiences of the DHM- storage and archive system will be used in other parts of the university, above all in projects of University of Greifswalds computer centre concerned with archiving services for the university library and the Landesarchiv Greifswald.

9.2 National Land Survey of Sweden

Future Activities in the National Land Survey of Sweden

In National Land Survey of Sweden by now a project, the Digitala arkiv (Digital Archives Project) is ongoing. This project started as a result of an Archives Strategy established in 1998, implying a digitisation of the often demanded and used archived material, and aimed to lay the foundation for the planned digitalisation.

The archives of the National Land Survey consist physically of two different parts, the central archives at the head office in Gävle and the regional archives, one in each of the twenty-four counties. The central archive contains, except of geographic maps and deeds from land reforms starting in the 1630's and recent property formation activities, also of huge amounts of other material, as geodetic data, aerial photographs and later databases. The regional archives contain mostly deeds from the property formation activities beginning in the 1750's until now. The total amount is estimated to more than 2,5 million acts. When all material is digitized, about 50 million image files will be created.

By now a new project, the ArkIT project is under initiation for production of digital versions of the regional archives. This activity is to be carried through during 2002-2004. The costs are estimated to 128 million Skr. Running parallel to this, the central archives are planned to be digitized to a cost of at least 10 million SKr.

Experiences from the DHM-project to be used in the future system

The DHM project has proved the feasibility in making demanded archived material accessible for internal and public use in digital form via a web interface on Internet. For the public this implies that material, earlier difficult to find and get access to, is possible to search and read at home. For persons doing research the digitized version often is sufficient, this holds for professionals as well as local historians and genealogists.

The principals for acquisition developed in the project will in the future system be implemented as a search and view function free of charge, combined with possibilities to order printouts and image files for further processing. An Extranet solution for contracted users, also including copying and printing, is going to be launched, during 2001 or 2002.

Naturally the information also will be available on Intranet for internal users, mainly working with property formation activities. A new function in the management system TROSSEN already makes it possible to transfer electronic documents from the system directly to the digital archives. The automated transmission process includes creating of Tiff images of the documents, conversion to DjVu format and storing, the Tiff files on DLT tapes in the HSM system and the DjVu files on the web server.

Technically, the image specifications from work package 2 are to be followed in the digitisation work:

The original image files will be saved in 254 dpi, 24 bits colour depth, RGB and Tiff 6.0 format (coloured maps).

Images from black and white originals (for example microfilm) will be saved in Tiff group IV

The systems developed for storage will be used.

Production flows will be automated and improved based on the ones from the project

The MrSID software for converting and viewing the images on Internet will probably be changed to the DjVu software. With a marginal deterioration of the visible quality, this software makes it possible to decrease the disk storage more than 10 times compared to the technique used in the project. This has a considerable impact on the incomparably most expensive part of a running

system, the disk system to the web server. The MrSID software might be an alternative format for deliverance (download or CD-ROM) as this format gives good quality at large compression and still is possible to handle in image processing tools as Photoshop, Erdas Imagine and GIS software as MapInfo and ArcView.

Systems for order, deliverance and invoicing will be further developed and adapted to other similar functions in the National Land Survey.

9.3 National Heritage Board of Sweden

For the National Heritage Board of Sweden the DHM project is part of a larger work, aiming at improved heritage management and increased interest and awareness in heritage for the public. Parts of the scanning specification actually originated from early Swedish work, but have been refined in the DHM project.

They will be used further in a planned co-operation between the National Heritage Board of Sweden and the Swedish National land Survey, focusing on delivering historical map information to the heritage management. This involves large scale scanning and distribution by Extranet from National Land Survey of Sweden servers.

The DHM concept is also a natural starting point when discussing further products for the public. Such discussions now take place within the framework of the National Land Survey of Sweden large-scale project Digital Archives, where also the National Heritage Board of Sweden is involved.

Finally many of the experiences will be used in the ongoing development work at the National Heritage Board of Sweden. There is an apparent growing interest for the material in Swedish heritage management. In this the long-term efforts by the National Heritage Board is supported by the new possibilities of distribution and development, offered by the digital formats.

In the spring of 2001 two days courses will take place at the National Heritage Board of Sweden, on historical maps and their use in heritage management. This will include analysis, rectification and to some extent vectorisation and visualization. Approximately 120 people from different administrations will participate. This means that the DHM concept and possibilities will be introduced to almost every county administration and county museum in Sweden.

9.4 National Survey and Cadastre of Denmark

National Survey and Cadastre of Denmark is in a process where all the data managed by the organization – historical and actual – are subject for possible delivery through the Internet. For the time being the first phase of a Retrieval and Delivery System (in Danish abbreviated LDS) is planned. In this first version only vector data will be incorporated, but it is expected that the system in future versions will handle raster data as well in an integrated environment. Because of these circumstances National Survey and Cadastre are going to establish its own solution. There probably will be some similarities with the system used in the DHM-project because of the HW and SW platform.

National Survey and Cadastre of Denmark has decided to use the Oracle as the main vendor of RDBMS. As servers it is the aim to use SUN with SUN/SOLARIS (UNIX) as the operation system. However, the organisation has some server applications that are only running on Windows

NT and are therefore not able to establish a pure UNIX environment on the server side at the moment.

The planned LDS is based on a common data warehouse containing all vector information. It is not yet decided how the raster data will be handled. Because of the unpredictable demand of the data, National Survey and Cadastre is expecting to store the data on disk systems. In this connection the use of Oracle Internet File System (iFS) will be considered because of the following key benefits:

Universal access.
The same files and folders are accessible through multiple network protocols, from Windows to the web.

Integrated storage.
Rather than storing different types of files in different server applications, iFS lets you store and manage all your files in a single place.

Content management.
Features that help you manage your files better are built into the file system. You can better manage files through their entire life cycle, from authoring to publication.

Ease of development.
Using XML and Java, application developers can customize and tap into the rich capabilities of the file system.

One of the key issues for National Survey and Cadastre of Denmark is to make the data storage and applications independent of each other, to be able to choose the application best suited for a certain purpose and to be able to change or replace the application without any influence on the underlying data structure. National Survey and Cadastre is aware of the influence on the data structure from the chosen compression method. The specification of the new format JPEG2000 compression hopefully is going to be the standard minimizing the problem with the compressed files.

The file formats and resolution is to be reconsidered when a general policy is formulated. It is to be expected that the present policy of scanning in 508 dpi TIFF is to be continued. The down sampling standard of 254 dpi might be reconsidered as a response to consumer wishes.

The historical maps span back more than two centuries. As time went by, administrative boundaries changed and so did the maps. This causes no problems for the topographical maps, but for the cadastral maps it is necessary to have search keys based upon the old administrative structures as well as the present. As the users have different angles it shall be important to have this in mind when making the search paths.

The modern way of searching shall cause few problems. But it shall be necessary to make indexes base upon one or two fixed administrative divisions of Denmark. The two in question could be the original division at the time of creating the cadastre. This has the advantage of an existing map of the division; this map being a possible basis for creating search keys through a map. The second could be a much later one, when the division was (more) stable and relating to the present system. The structure and information content of the tables has to be given a lot of thought.

For the topographical maps geo-coding is the solution, as it takes less time to do and is better for the user. Whether some cadastral maps should be geo-coded is a matter for decision.

The work with file naming shall be restricted to make the files easily identifiable in the work process and still make as little room for mistakes as possible. A policy of deciding which information shall go with each kind of map shall have to be made.

Presently, most of the scanning is done externally, partly paid with funds from external sources. If an internal production on a larger scale shall be started, it shall be on a project basis with a lot of thought given to the work-flow, the making of tests to find and correct errors, the possibility of automating processes, the IT-equipment and the training of people.

DESCRIPTION for WORKPACKAGE N°: 10

Title: Technology B2 — Follow-up and evaluation of Technology B1 (WP4) — storage and archiving

Lead partner for this WP: KMS (NLS) Start month: 15 End month: 26

Initial state, work already done, preconditions for starting tasks, end result expected:

The purpose of this work package is to evaluate the storage/archiving environment developed in WP4. It is important for the future use of the information that these environments function satisfactorily and that the management of the material is guaranteed. In the course of the evaluation each institution will report on its experience relating to functionality and management. Based on this evaluation, the project will recommend improvements and, if need be, revise the management costs estimated in WP4.

If necessary the specifications and management costs (WP4) will be revised. The report from WP4 will also be completed with experience to date with the storage/archiving of the information and recommendations of necessary improvements.

Tasks:

Evaluation of the storage/archiving environment developed in WP4
Recommendations for improvements
Revision of management costs
Deliv. nr 23 and 26

	Management 1 mm	**Technical 3,5 mm**	**Other 1 mm**	**ToT 5,5**
KMS	0,5	0,5	-	1,0
NLS	0,5	3,0	1,0	4,5