



Digital Historical Maps

Report from WP4

Storage and Archiving

Final version 2.3

March 30, 2001

Deliverable 23

Stefan Gustafsson, Lars Sävmarker

National Land Survey of Sweden

Table of Contents

1	STORAGE AND ARCHIVING – WP 4	1
1.1	SCOPE	1
1.2	ACCOMPLISHMENT	1
2	DESCRIPTION OF TECHNICAL RESOURCES	1
2.1	STORING OF ORIGINAL FILES	1
2.1.1	National Land Survey of Sweden	2
2.1.2	National Survey and Cadastre of Denmark	4
2.1.3	University of Greifswald	5
2.2	STORING IN THE WEB APPLICATION	7
2.2.1	Volumes	7
2.2.2	Storage media	7
2.2.3	Systems Overview	8
2.2.4	Database	8
2.2.5	Security and Backup	8
2.2.6	Availability	8
3	PROVISION OF EQUIPMENT	9
4	MANAGEMENT COSTS	9
4.1	COSTS RELATED TO TECHNIQUE USED	9
4.1.1	Storage of Original TIFF-files	9
4.1.2	Storage of Compressed MrSid-files	10
4.2	YEARLY MAINTENANCE COSTS	10

Enclosures

- Appendix 1: WP 4 project description
- Appendix 2: TIFF files storage
- Appendix 3: Web Application Storage
- Appendix 4: Checklist for purchasing of equipment

1 Storage and Archiving – WP 4

1.1 Scope

The aim of work package 4 (WP 4) is to create the preconditions for a functioning pilot environment for the storage and archiving of raster data of varying quality and for various user needs. This will include allowing for each institution's resources for the storage of large volumes of raster data, and analyses of suitable file and media formats for long-term viability. The work will utilise the progress report on existing systems from WP1.

Together with WP3, the work will result in a functioning pilot system for the distribution of a limited map set and act texts via the Internet. It will further result in a written report specifying management costs and suitable standards for storage/archiving with regard to quality and media/file format.

Tasks:

- Analyse and report on the technical resources of each archive institution for storing large volumes of data
- Draw up specifications and documentation for purchasing process
- Purchase server capacity
- Purchase storage media
- Analyse and specify management costs (backup procedures, conversion costs etc.) relating to the material produced

1.2 Accomplishment

The time schedule for WP 4 was month 9 - 23 of the project. As the production of image files in the NLS started already in month 7 the work with implementing of a storage environment for original (TIFF) files was initiated in May 1999 (month 5).

2 Description of Technical Resources

2.1 Storing of original files

In this project TIFF 6.0 has been chosen for storage of original files. This standardised format is commonly used on the market for images and can be handled in almost all software on the market. This guarantees the information value for a wide range of users, and makes it easy to convert to other formats when needed.

The only practical problem is the size of the files. For the large amounts of data that is created when digitising entire archives, it is therefore necessary to limit the file size as much as possible, and to find a storage system that is cost efficient.

2.1.1 National Land Survey of Sweden

In the National Land Survey of Sweden the standard system for storing of environmental data is used also for the purpose of this project. This environment is also planned to be used when escalating the system with further more information. In NLS decision is made to digitise the main part of all used archives, an activity that will be finished by 2005, as plans are today.

2.1.1.1 Volumes

Prototype System

TIFF files (Swedish archives) 1 500 GB

Full scale solution in Sweden

TIFF files 65 000 GB

2.1.1.2 Storage media

Storage technology is still an area of great development. Especially if one looks on costs for storage; the trend of halving the cost for storage every 18 months is still applicable.

For storing, DLT IV tapes are used, storing 70 gigabyte of compressed data each. The price for one tape is today about 70 EURO.

In the beginning of the project also storing on optical disks (WORM) were considered, but the construction with a hard disk system on the webserver and the low research frequency in the original storage made this solution not realistic regarding the costs. The cost for one unit storing 5,4 megabyte was almost 120 EURO.

DLT-tapes are considered to have enough good constancy regarding time, at least 10 years without information damage. If the tapes are reconditioned the constancy is even longer.

It is very likely that the storage technology has changed by that, and new technique will be used. In the cost analyses later on, no costs therefore have been allocated for reconditioning.

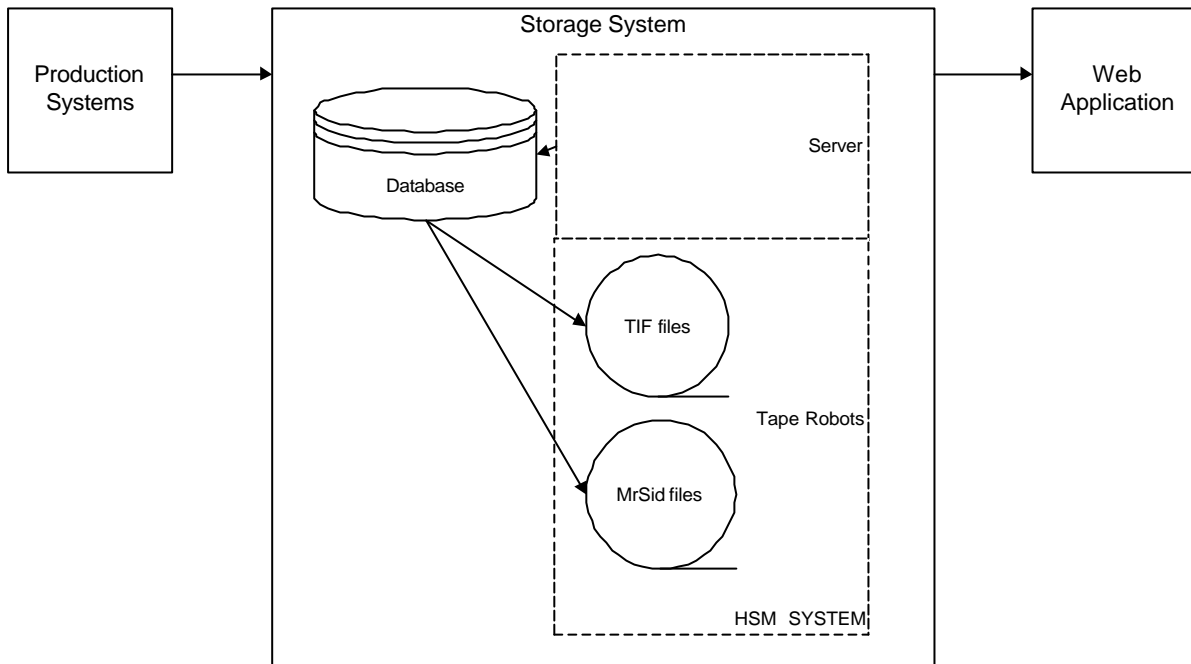
2.1.1.3 Usage of the HSM System in the application

In this project only one part of the HSM system is used, namely the tape storage.

If this changes in the future it is possible to set-up the HSM system so that files in great demand can be stored on optical disks or even on harddisk.

For the Swedish information, the converted files (MrSid), also are stored on DLT-tapes in the HSM-system, due to security reasons.

2.1.1.4 Systems Overview



For a more detailed description of the system, see Appendix 2.

2.1.1.5 Database

Meta data about the image files and paths to the files themselves are managed using Oracle, the standard DBMS of the NLS.

For management of information in the database internal NLS, applications are developed using Oracle Forms and PL/SQL.

Questions from the web application are handled via defined views (OracleView).

In a future system, Oracle might be exchanged for BRS/Search, a database suited for this type of application and often used in corresponding systems. Reason for exchange is that BRS/Search is faster, especially in free-text search.

2.1.1.6 Security

The web application has read-only rights to data inside the firewall. The firewall is designed to fit the standards in the NLS.

The system uses Oracles login routines to the database.

2.1.1.7 Backup

Database: Total backup every night.

2.1.2.4 Database

At present, the database is in the Swedish prototype. Copies of the tables (in Excel) are kept on separate harddisks and shall when the project is finished also be put on a CD-ROM.

2.1.2.5 Security

Copies of the map files and the tables are on CD-ROMs. The original files are in one place, the handled files are at two separate places.

2.1.2.6 Availability

The search for the right file can be done through the prototype in Sweden. For delivering a copy the system is manual, so it is only accessible through work hours. When a file has to be copied, it must be found on the CD-ROM and copied to another.

Finding a file (when being at the storage area) 2 minutes

Copying the file to the pc 2 minutes

Burning the file on CD-ROM + control of product 2 minutes

For one file the netto time is 6 minutes

For each expedition shall then be added time for going to collect the CD-ROMs, inserting them into the pc, setting up the program, putting the CD-ROMS back, etc: 20 minutes

That means that for

1 file, it takes 26 minutes

2 files, it takes 32 minutes

3 files, it takes 38 minutes

and so on.

The copying is made on a semi-modern equipment (PC 550 MHz, 1024 MB RAM, burning takes place at 6x speed), but the hardware has not much to do with the time spent.

The cost will be diminished when the files are put on a server. It would presumably reduce the net time for a file to 4 minutes and the overhead time to <10 minutes

2.1.3 University of Greifswald

At present no sophisticated archive-system for the project exists. Data are stored as usual, in a system consisting of

- original hardware copy (on CD-ROM)
- daily working copy (on CD-ROM)
- back up on Tape
- and time limited storage on workstation.

This is the standard procedure in the Institute and common for all valuable data stored there.

At the moment most parts of the delivery system are manual and occupy much capacity of technical staff. But there will be no problems to handle 140 digital maps and deliver them. In the cause of high frequency of delivery system it will be necessary to develop a semi-automatic system.

For the future it is planned to investigate the results of the project and evaluate especially the system on aspect of storage and response from customers. When the results are satisfying, it is planned to adapt the system of the project and continue the work. It is the aim to establish in co-operation with the Landesarchiv a digital archive for all Swedish Matrikel Maps and additional texts in Greifswald. There are more than 2500 maps stored in the archive, including original texts.

2.1.3.1 Volumes

The prototype system includes the compressed MrSid-files and metadata belonging to them. The MrSid-files occupy on hard disk a space of approximately 340 MB overall, including MrSid-text files. The pure metadata take not more than 1 MB.

The original scans of maps are stored in TIFF-format and occupy 9,5 GB on disk.

2.1.3.2 Storage media

All information is stored on CD-ROM primarily. The original scans occupy 14 volumes and are placed in the data archive of the Institute. Sid-files are stored also on hard disk (PC).

2.1.3.3 Database

The database and overview maps are stored on CD-ROM. In addition, all staff involved to delivering system have copies on hard disk (PC). The database for searching certain digital maps is handled with MS Excel, and ArcView GIS (ESRI) is used for overview maps.

2.1.3.4 Security and Backup

There are 3 separate duplicates of the digital maps:

- Original scans on CD-ROM, in the data archive
- Working copies for the staff involved in the delivering system
- Backup on tape, in the backup archive.

All duplicates are stored in different rooms in the institute which are included in the main security system of the Institute of Geography.

2.1.3.5 Availability

Data archives and backup archives are well known by staff and accessible for them. Responsible for the digital data archive and backup is Jörg Hartleib.

The delivering flow:

- Choose the files in the prototype and order a copy by e-mail to Greifswald

- Collecting CD-ROM, setting programs etc. (15min)
- Find the place on CD-ROM by checking the database (1min/file)
- Select and copy the files on a new CD-ROM (5min/file)
- Send the data to the central mail delivery system of the University Greifswald

2.2 Storing in the web application

The web application only uses the compressed (MrSid) files stored on a hard disk system. The disk system is updated with new MrSid files once a week from the DLT tapes stored in the HSM system. This solution was originally designed on basis of how the security solution in the NLS were then.

The material from Denmark and Germany is delivered on CD-ROM:s and only copied to the hard disk.

2.2.1 Volumes

Prototype System

MrSid files (including Danish, German and Swedish material)	50 GB
---	-------

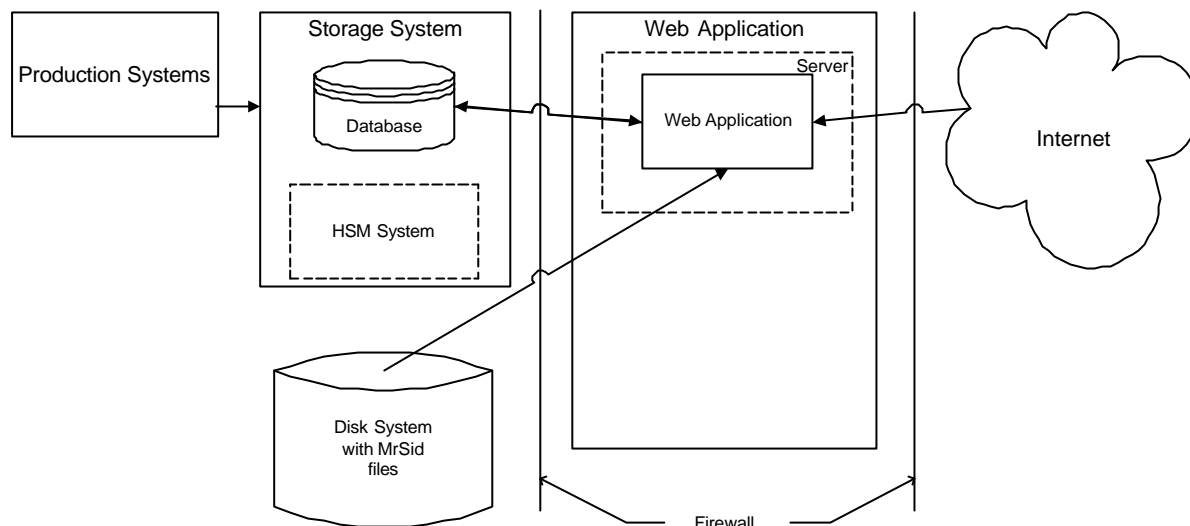
Full scale solution, only Swedish information

MrSid files	4 000 GB
-------------	----------

2.2.2 Storage media

The MrSid files are stored on a harddisk system enabling performance to satisfy a user via Internet. Initially optical disks were considered, but rejected due to time for retrieval.

2.2.3 Systems Overview



For a more detailed description, see Appendix 3.

2.2.4 Database

The web application uses the same database as the original storage. The database can be reached only through the application.

Communication with the database is implemented as scripts (PHP4) towards the views mentioned above

2.2.5 Security and Backup

There is no backup on the file storage on the harddisk. The storage on DLT-tapes in the HSM system for the Swedish material is considered enough in this part. For the Danish and the German material the CD-ROM's are considered sufficient, due to the limited amount of information and the small effort to reconstruct the information if necessary.

The web application itself is protected by a firewall towards the Internet. If the application in spite of this outer firewall is hacked and destroyed a new version always can be replicated from inside the main firewall of the NLS. The information is also located inside this firewall.

2.2.6 Availability

Specified Access times

Prototype

Search in database	
–predefined search	< 10 sec
–free text search	< 20 sec

Presentation of image file on Internet (10 MB)

connection)

Presentation of image file on Internet (56 K modem connection)

The system is available 24 hours a day.

Allowed interruptions:

08.00-16.00 < 1 hour, 99% of time

16.00-08.00 < 1 hour, 90% of time

Performance test, see appendix 4.

3 Provision of equipment

To limit the costs in the project it was natural to investigate current equipment, hardware as well as software, in the NLS, where the system was to be implemented. This environment showed to fit the demands of the project, with small completions.

A checklist for purchasing a storage system is enclosed as Appendix 5.

4 Management Costs

This analyses is based on experiences gained from the first year running the Prototype System and expectations of future development of storage systems and costs related to those. The large part of the future maintenance cost will actually be related to storing of data.

The calculation is also based on a fictive archives system containing all archives of historic maps and documents in the head office of National Land Survey of Sweden in Gävle.

Costs related to order and delivery systems are not included. The reason for this is that each organisation has its own routines and that these systems also are used for other purposes. It is therefore difficult to estimate a realistic share of the total costs for these systems.

4.1 Costs Related to Technique Used

4.1.1 Storage of Original TIFF-files

The original files in the Prototype system are stored on DLT-tapes. Reasons for this choice are the proportionately low price and the security level achieved. The DLT-tapes are considered to be durable for about 10 years.

Costs in EURO:

<u>Cost per GB</u>	<u>Prototype (1,5TB)</u>	<u>Total Archives (65 TB)</u>
0,12	1 765	76 500

4.1.2 Storage of Compressed MrSid-files

For the compressed files, possible to access from the web-application at specified access time, a hard disk system is necessary.

Costs in EURO:

<u>Cost per GB</u>	<u>Prototype (50 GB)</u>	<u>Total Archives (4 TB)</u>
75	3 750	300 000

Comparison with DjVu software

If another software is chosen for compressing and presentation on the Internet, the volumes will be affected. The new software DjVu (Lizardtech), which allows presentation at almost the same quality level via the Internet, will decrease the demands for storage, and the costs, as follows:

Costs in EURO:

<u>Cost per GB</u>	<u>Total Archives (~750 GB)</u>
75	56 250

Following are not included in calculated costs:

- Costs for development within this project
- Software bought during the project time
- Production of digital images
- Future costs for converting image files for security reasons. Probably the storage system will be changed within 10 years and those costs are included in the price for storage
- Costs for storing TIFF-files in another system

4.2 Yearly Maintenance Costs

Calculated annual cost for personnel:

Internal NLS	76 500 EURO
Consultants (included travels, etc.)	176 500 EURO

Digital Historical Maps

Report

Status:

Version:

Date

WP 4

Final

2.3

2001-03-30

Storage and Archiving

Page 11

Costs in EURO

Licences

(included support and maintenance from software provider)

Oracle Site licence

(175 EURO/employee)

Hardware (yearly)

HSM

76 500

Disk-system, MrSid solution

306 000

(DjVu solution)

58 800

CPU, depreciation (investment 58 800)

14 700

Personnel (yearly)

Specifications, test, verifying, training

7 650 (0,1)

Application (adjustments, correcting, development)

15 300 (0,2)

Running of system (supervision, maintenance)

7 650 (0,1)

Helpdesk (user support)

7 650 (0,1)

Maintenance costs a year

435 450

(DjVu solution)

188 250

WORKPACKAGE DESCRIPTION

for WORKPACKAGE N°: 4

Title: Technology B1 — Storage and archiving

Lead partners for this WP: NLS (KMS and Greifswald) **Start month:** 9 **End month:** 23

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

The purpose of this workpackage is to create the preconditions for a functioning pilot environment for the storage and archiving of raster data of varying quality and for various user needs. This will include allowing for each institution's resources for the storage of large volumes of raster data, and analyses of suitable file and media formats for long-term viability. The work will utilise the progress report on existing systems from WP1.

Together with WP3, the work will result in a functioning pilot system for the distribution of a limited map set and acttexts via the Internet. It will further result in a written report specifying management costs and suitable standards for storage/archiving with regard to quality and media/file format.

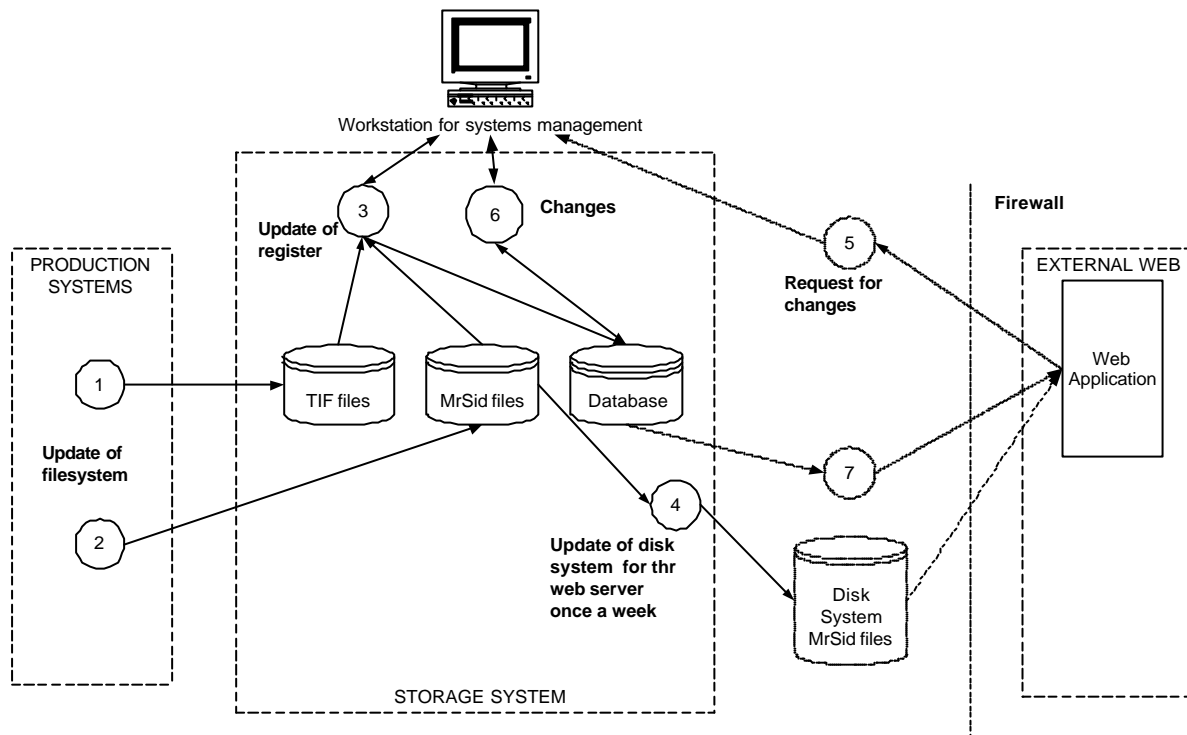
Tasks:

- Analyse and report on the technical resources of each archive institution for storing large volumes of data
- Draw up specifications and documentation for purchasing process
- Purchase server capacity
- Purchase storage media
- Analyse and specify management costs (backup procedures, conversion costs etc.) relating to the material produced
- Devliv. no 13

	Management 1,6 mm,	Technical 5,5 mm,	Other 1 mm	ToT 8,1	
NLS	1,0	5,0	1,0		7,0
KMS	0,3	0,5	-		0,8
Greifsw	0,3	-	-		0,3

TIFF Files Storage

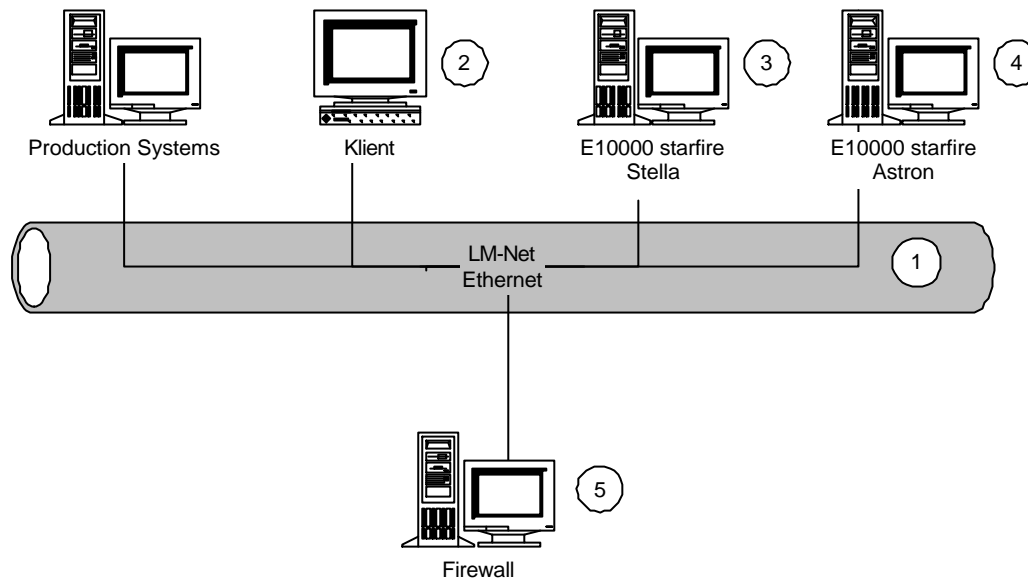
Functions overview:



Functions

- 1 Transfer of original TIFF files to the HSM system (RAKLMS) using an index file containing information of file structure included in the file name as described in WP 2
- 2 Transfer of original TIFF files to the HSM system (RAKLMS_KOMP) using an index file containing information of file structure included in the file name as described in WP 2. The file structure is the same in RAKLMS and RAKLMS_KOMP
- 3 Once a week a log file showing the changes in the storage is captured. After processing the database is updated with the new information. This routine can be automated in a future system.
- 4 Once a week new MrSid files are copied from RAKLMS_KOMP to the disk system on the web server.
- 5 Requests for changes can for example concern register information or quality aspects on the image files.
- 6 In the function requests for changes are investigated, analysed and checked. The function also gives priority and decides whether changes are to be carried through.
- 7 Not primary functionality for storing. Information and presentation of image files.

Technical overview:

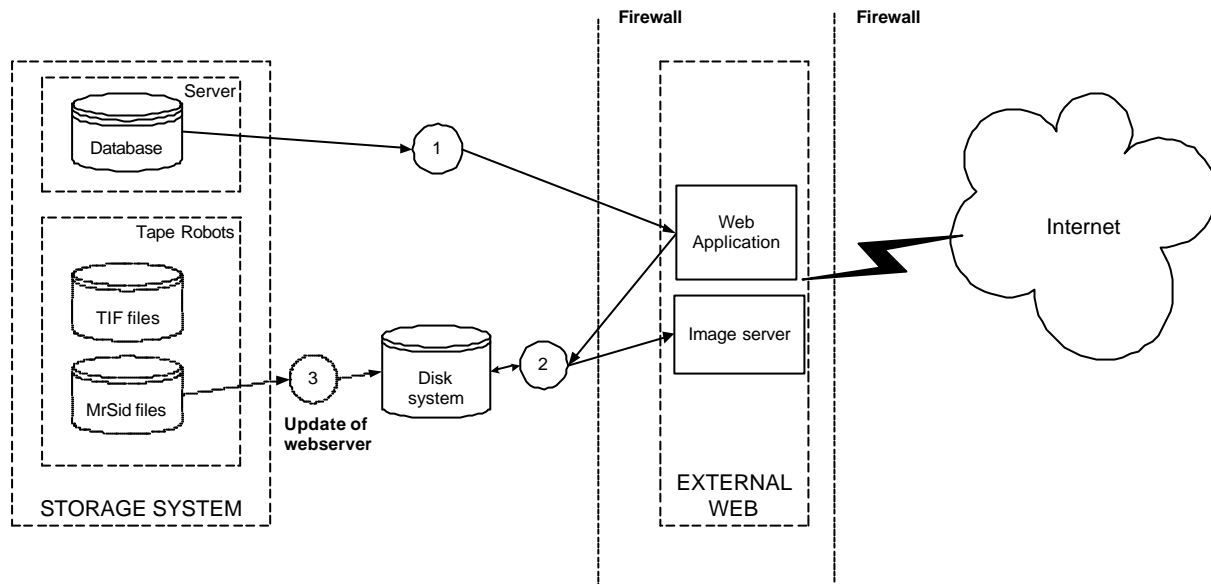


Description

- 1 **Communication:**
 Tcp/Ip over LM-net (the LAN of NLS)
 Tcp/Ip over Internet
 Net Bios (Samba) from NT-systems (some production systems)
- 2 **Client software:**
 Browser: Internet Explorer 5
 Oracle Forms application
- 3 **Structure/access:** Sun E10000 Starfire, HSM och DLT-band (TIFF- och Sid-filer)
System software: Unix O/S Solaris >V5.2.1
Security: Data protected by firewall. Only read-only rights from the web-application. Oracle login to the database. Validation in Samba.
- 4 **Structure/access:** Oracle 8i
System software: Unix O/S Solaris >V5.2.1, Oracle 8i (the same physical server as above)
Security: Data protected by firewall. Only read-only rights from the web-application. ODBC access in SQL (UNIX) to the database from the application. Oracle login to the database. Validation in Samba.
- 5 **Security:** Firewall in accordance with standards in the NLS

Web Application Storage

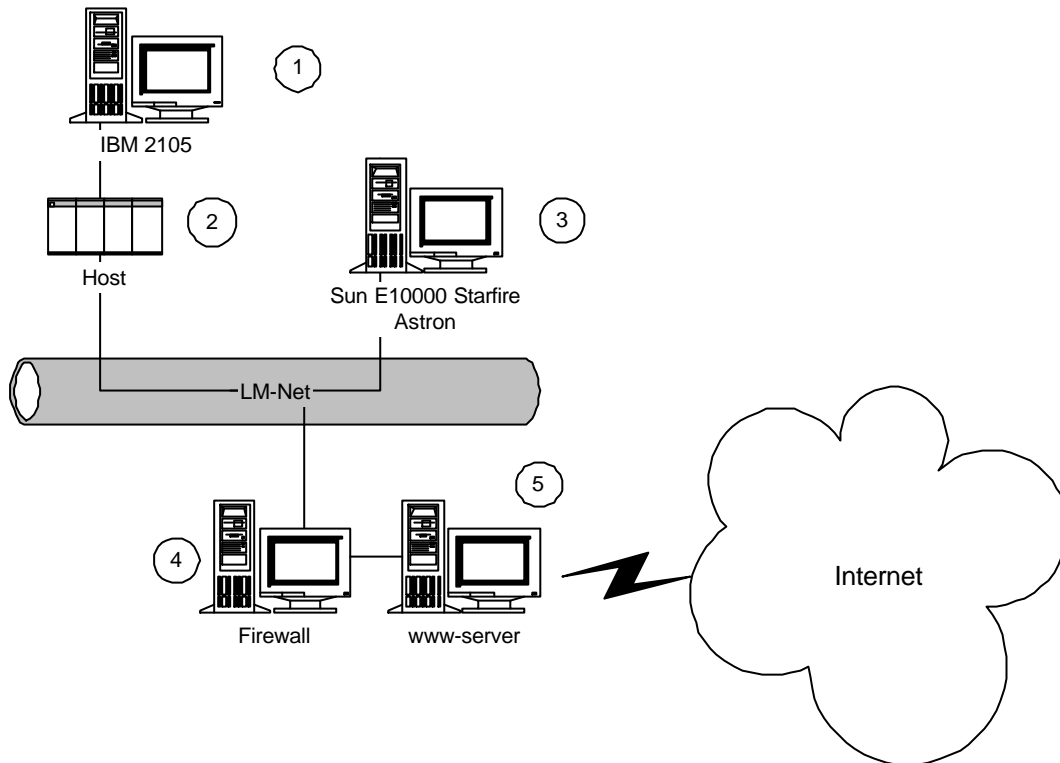
Functions overview:



Functions:

- 1 Information retrieval is based on information from the user of the web application and call for a defined view. A predefined answer is sent to the application to enable presentation of metadata concerning the image files and the path to corresponding image file on the disk system.
- 2 The web application asks for and gets the demanded image for presentation
- 3 Once a week new MrSid files are copied from RAKLMS_KOMP to the disk system on the web server.

Technical overview:



Description:

- 1 **Structure/access:**
 IBM 2105, cache disk (Sid-filer)
 IBM 2105, disk storage (Sid-filer)
- 2 **Security:** Login to disk system via host, O/S: AIX
- 3 **Structure/access:** Oracle 8i
System software: Unix O/S Solaris >V5.2.1, Oracle 8i
Security: Data protected by firewall. Only read-only rights from the web-application. ODBC access in SQL (UNIX) to the database from the application. Oracle login to the database. Validation in Samba.
- 4 **Security:** Firewall in accordance with standards in the NLS
- 5 **Structure/access:** HTTP-files
System software: Unix O/S Solaris >V5.2.1, Oracle 8i, MrSid Image Server, Web server Apache
Security: ODBC access in SQL (UNIX) to the database from the application.

Document name

Status:

Version:

Date

WP 4

Final

2.3

2001-03-30

Storage and Archiving

Page 1

Checklist for Purchasing of a Storage System

General description

-Describe the goal and purpose of the system

IT strategies

-Your internal strategies

Computer environment

-Describe the existing systems and LAN.

Volumes

-Today's volumes and the expected growth for the next 5 years.

Access times

-Declare the demands - how long time a user can wait for data

Availability

-In terms of 99.xyz depending on redundancy

Security

Installation

-When and how long time it will take

Training

-For system administrator, operators.

Service and support

-Describe the wanted maintenance agreement

Plans for future development of the system

Hardware and software