How to find, read, understand and trust digital information in a 50 year perspective

Jon Ølnes, Olga Cerrato, Inger-Mette Gustavsen, Thomas Mestl
DNV Research & Innovation

NOKIOS, Trondheim, 17th October 2008

Long-Term Records Management

The digital disease

- Symptoms
- Development of the disease
- The infectious agents
- The patients

- Wrap-up

DNV – an independent foundation

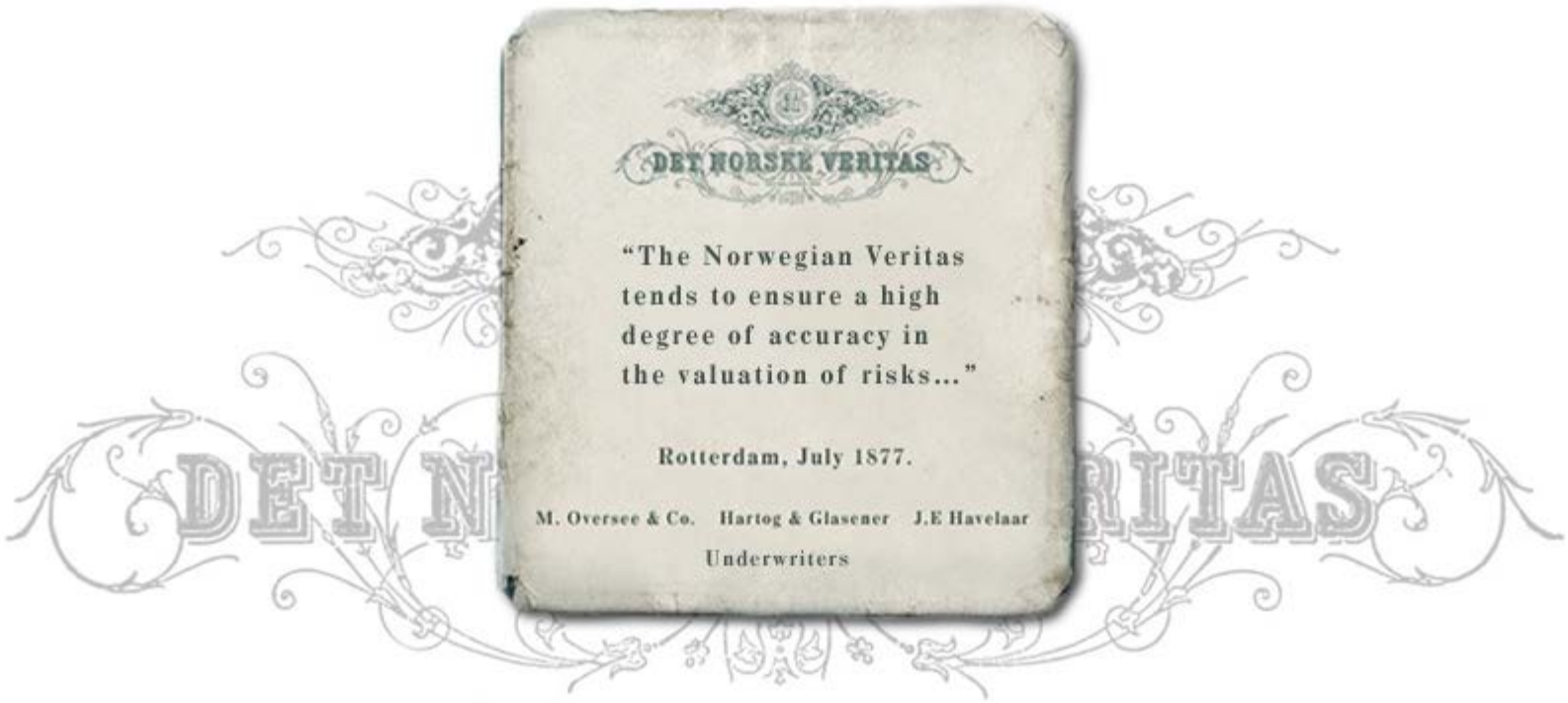
MANAGING RISK

"Safeguarding life, property, and the environment"

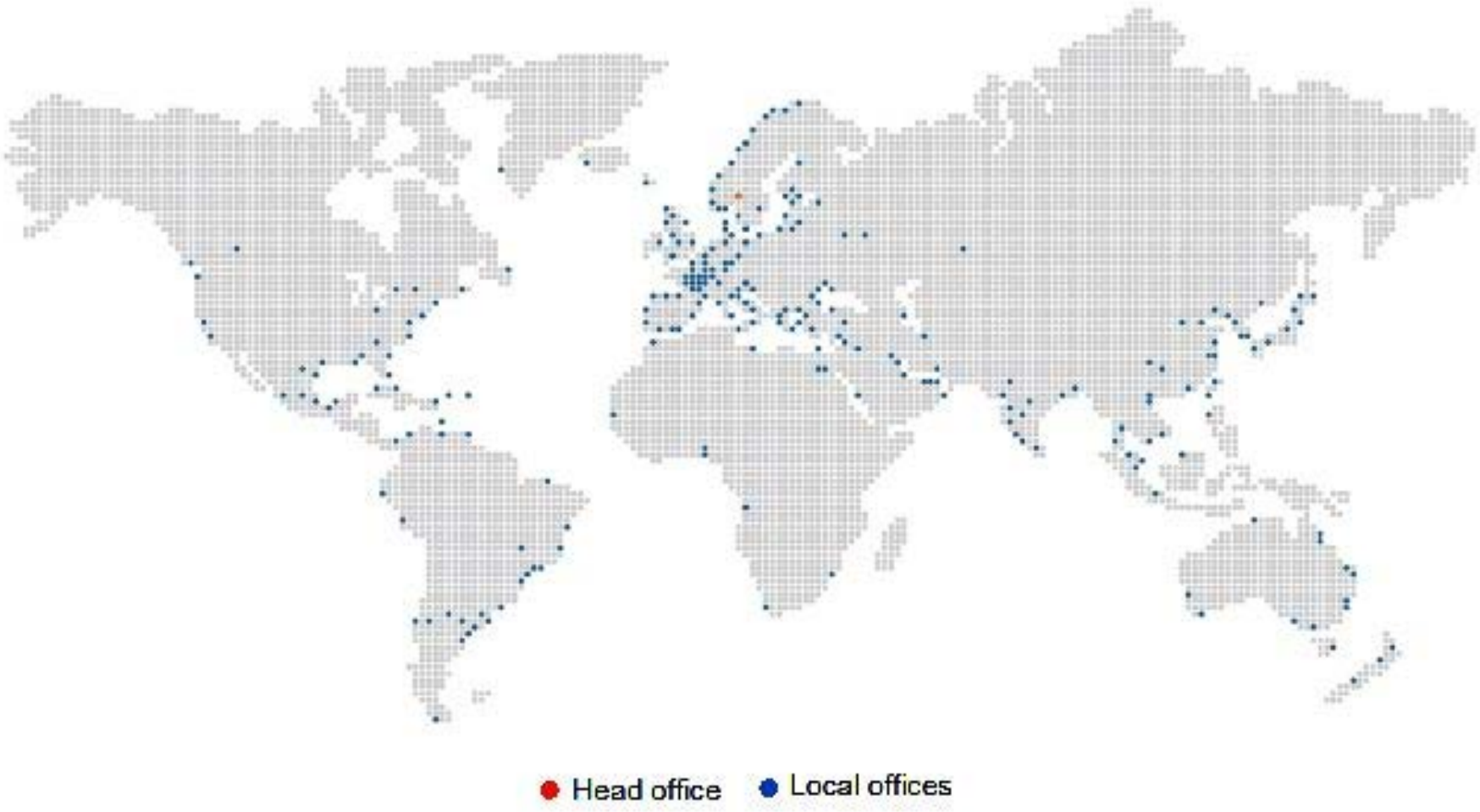
More than 140 years of managing risk

MANAGING RISK

- Det Norske Veritas (DNV) was established in 1864 in Norway
- The main scope of work was to identify, assess and manage risk – initially for maritime insurance companies

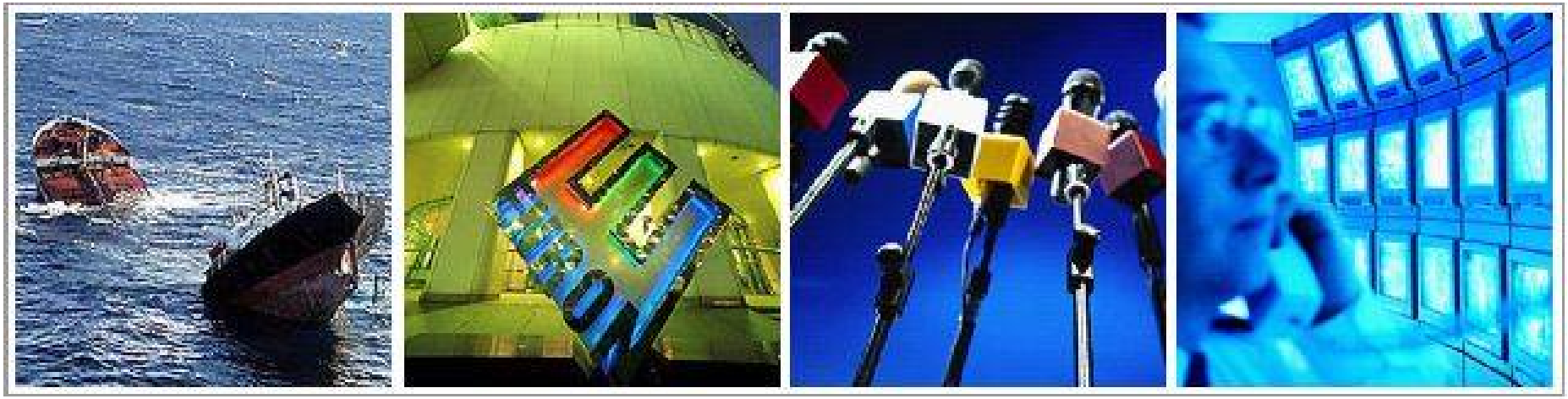


300 offices in 100 countries



New risk reality

- Companies today are operating in an increasingly more global, complex and demanding risk environment



- Society at large is gradually adopting a “zero tolerance” for failure
- Increased demands for transparency and business sustainability
- Stricter regulatory requirements
- Increasing IT vulnerability

Target industries

Managing risk

Maritime

- Ship classification
- Certification of materials and components
- Assessments and solutions
- Fuel testing
- Training

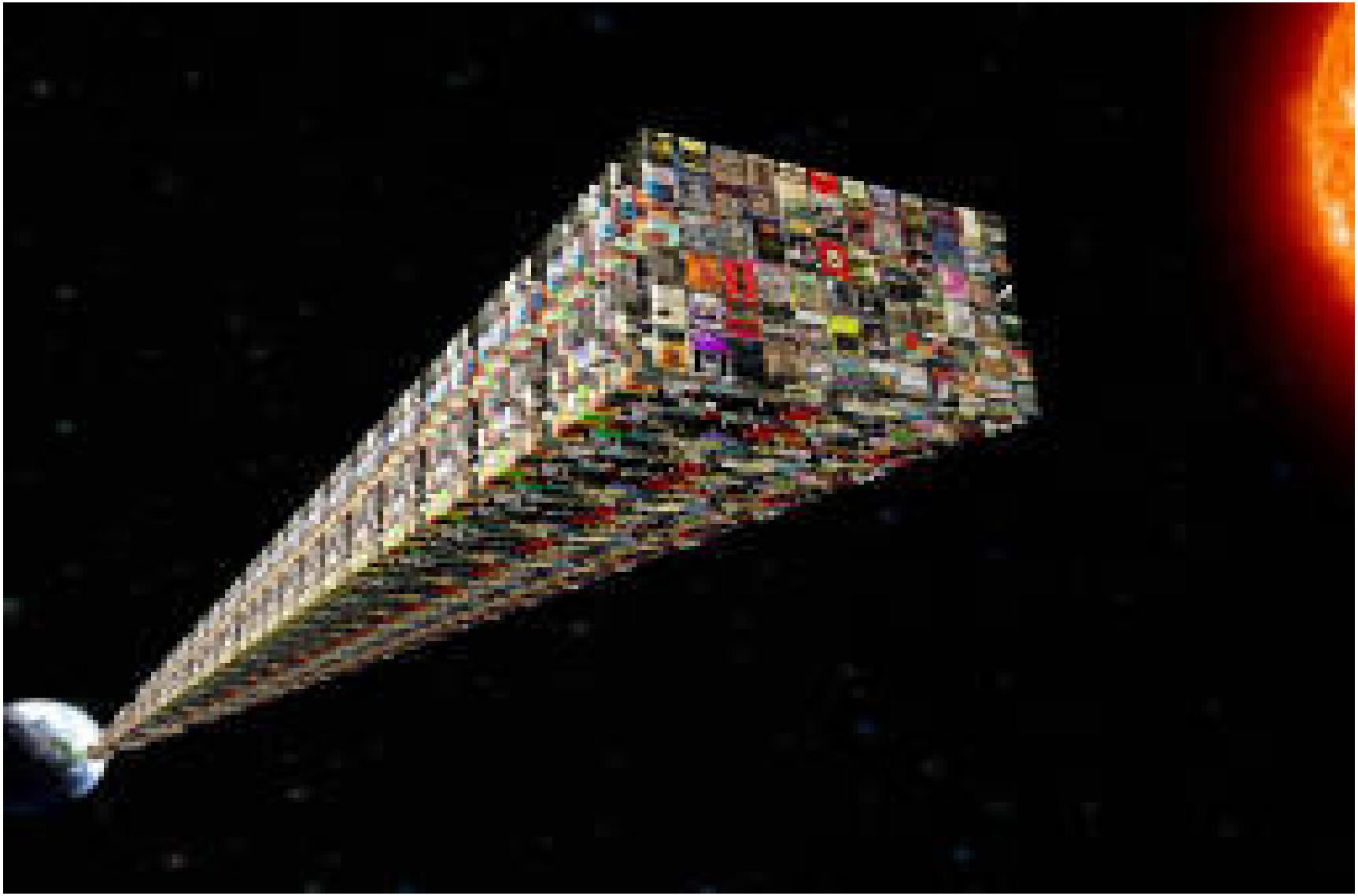
Energy

- Risk management consulting
- Qualification and verification
- Offshore classification
- Laboratory services
- Training

Other prioritised industries

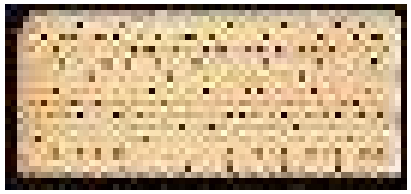
- Automotive 
- Defence 
- Finance
- Food and Beverage 
- Health care 
- ICT and telecom 
- Public sector 
- Transportation 

Digital Information Production in 2010 in the World



Does anyone remember how to use this?

Robotron 1370



Or these?

Or…very soon these?

Symptoms?

Accessibility of content

Rosetta stone

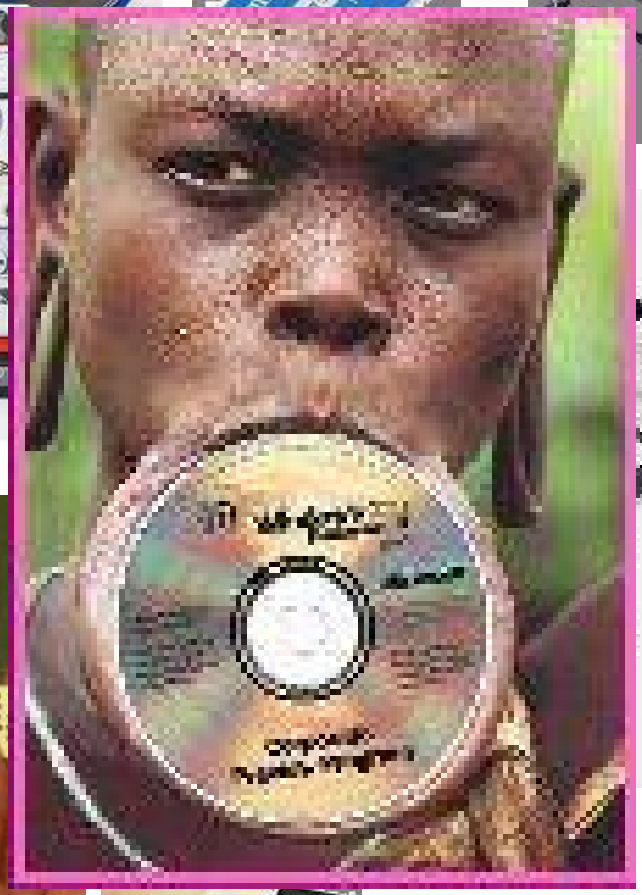
- Text understood today

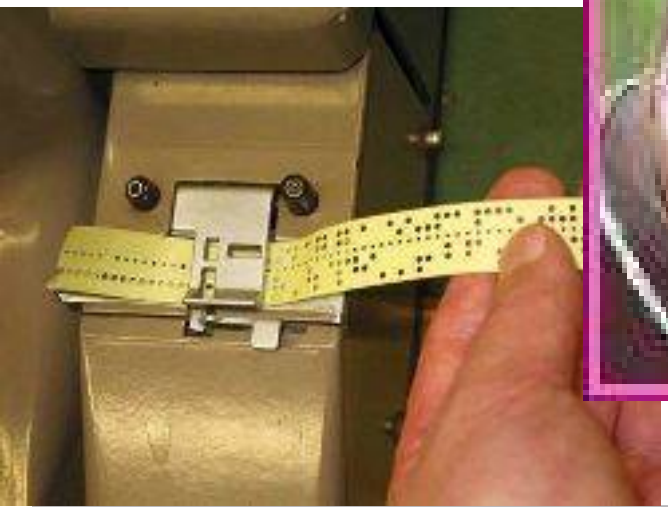
Robotron

- East German computer
- No-one knows how to use it

BUT INFORMATION PRODUCED AFTER 1990 will be lost if we don't do anything!

Symptom: Hardware Obsolescence





Symptom: File Format Obsolescence

- **Proprietary, closed specifications, e.g. Word.doc.** Evolve quickly, exist in many different versions for different platforms, with only limited backward compatibility
- **Proprietary, open specifications**, e.g. Adobe.pdf. Vulnerable to market forces as they can be abandoned for commercial reasons.
- **Non-proprietary, open specifications**, e.g. JPG. Guaranteed long-term availability, specifications published by international standards bodies. BUT these standards must be widely adopted by both user and developer.

Other Symptoms: Traceability over time

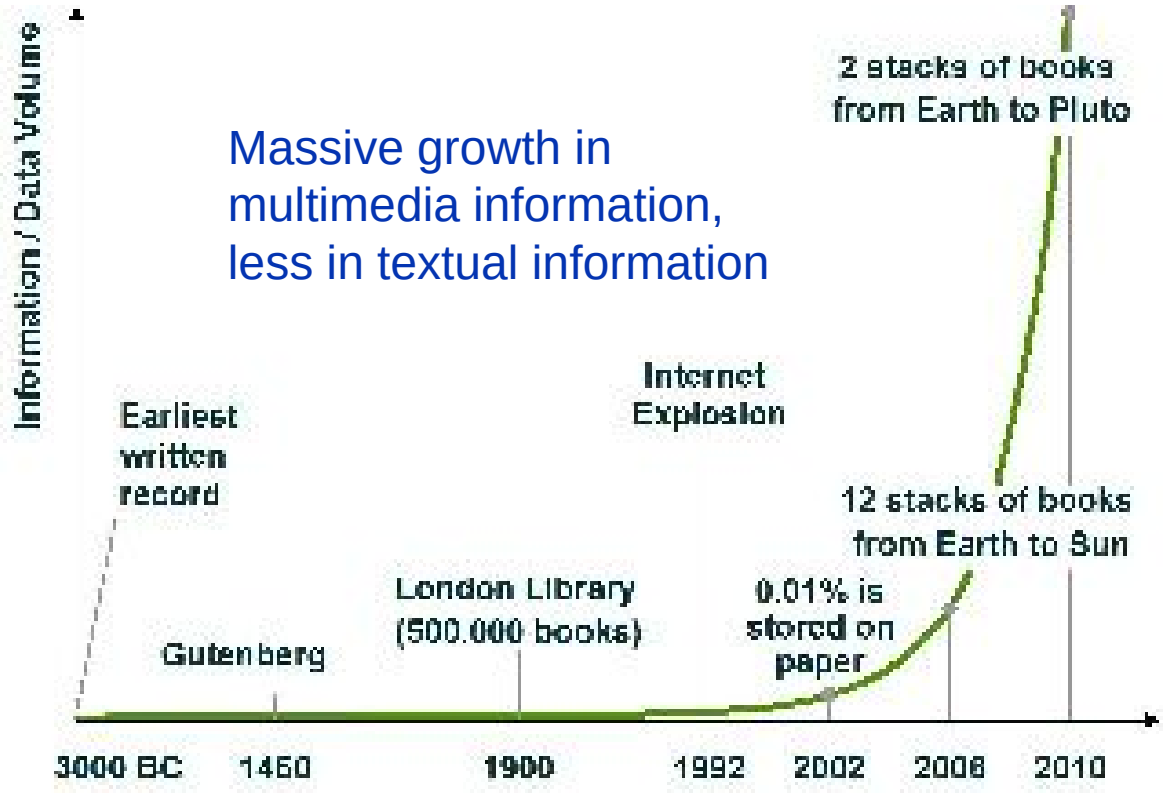
The Norwegian branch of Nordic bank Nordea vows a full investigation into how bank account statements for Princess Märtha Louise and other celebrities wound up in the hands of reporters at magazine *Se og Hør.* (Aftenposten, 12/2-2007)

The bank, regulators and other media are crying foul after newspaper *Dagens Næringsliv* reported over the weekend that the royal bank account statements were leaked to the magazine. It's the latest in a string of revelations about reporting techniques at *Se og Hør,* most of which have been revealed in a new book by a former staff writer at the magazine.

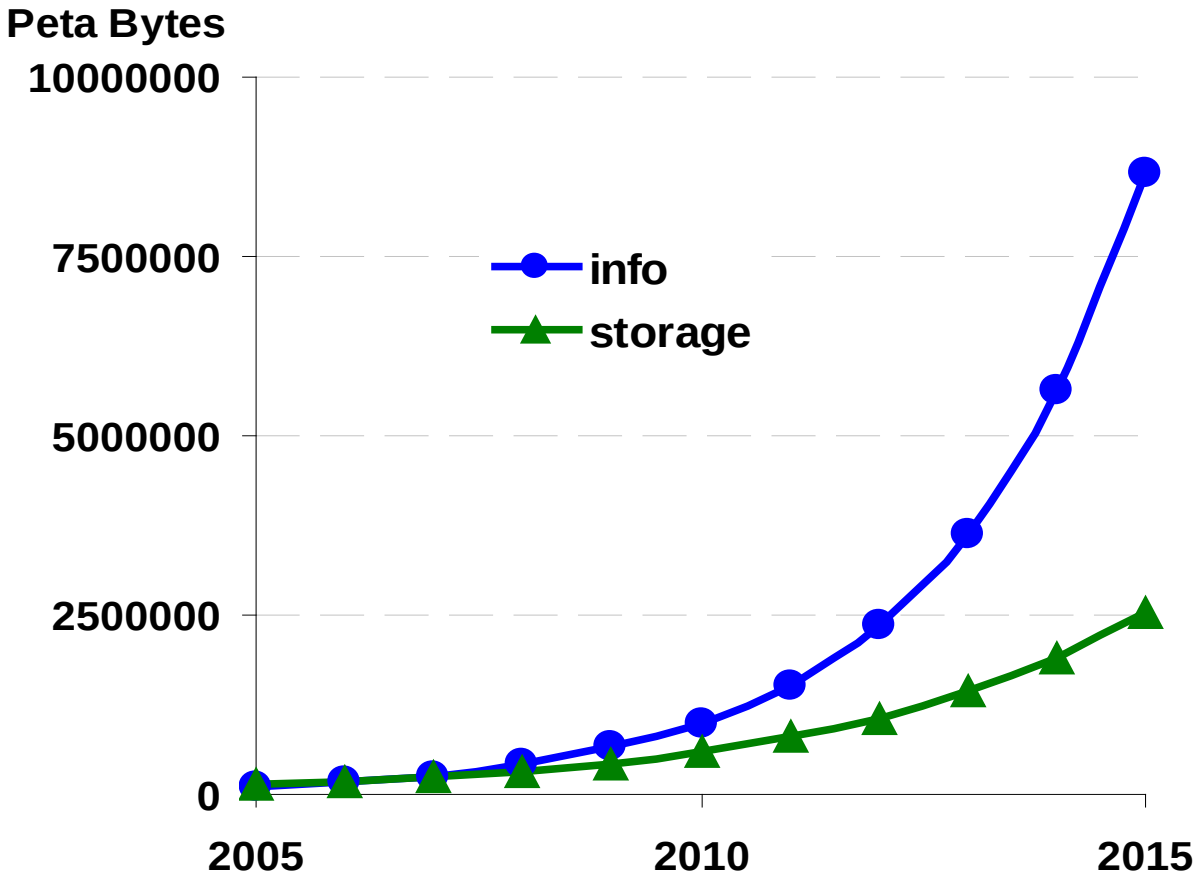
- **No special security**
  Account information for members of the royal family or other public officials or celebrities isn't subject to any stricter security controls, meaning that anyone dealing in customer service at the bank can have access to the accounts. Nordea has nearly 4,000 employees in Norway. …**it's possible to track who may have accessed the accounts, but it may be difficult to track such information if the access occurred many years ago.** Other banks in Norway have much the same practice as Nordea, meanwhile, with all customer service employees able to access all accounts.

Development of disease: Volume explosion

- 90% of all data is unstructured (pictures, video, e-mails, blogs, …)
  - no data model, no meta data
- 70% of all data belongs to individuals and are stored de-centralized
  - Video, Photos, web pages, etc.



Development of disease: Storage Shortage



- 2015 annual growth:

  8 Zettabytes (= 1 million Petabytes),

  = ca 20 x higher than 2008.

- **2015: Data created will be three times amount of available storage.**
- **Lots of data will be for immediate consumption only**

The infectious agents

Disease also known as:

- technological advances
- new developments
- new products

Challenges:

- Technology/systems life-time
- Software lifetime
- Formats' lifetime
  - Conversion, migration
- Processes' lifetime
  - ➢ Roles, authorisations, people
- Organisations' lifetime
  - Merger, split, re-organisation, close down
- Volume
  - ❖ Search and retrieval
- Trust
  - Compliance (laws and regulations)
  - Authenticity, integrity, confidentiality

In 2015 **80%** of today's employees will still be working but **80%** today's technology will be replaced

INFORMATION outlives most of us and much of it should live forever!

The ticking digital bomb…

- Volume produced in 2010 six times the data produced in 2006
- In 3 years we will produce the same amount of data as previously produced in the history of mankind
- Hidden information cost
  - Massive volumes
  - Unstructured information
- More rules and regulations
- More integrated tools
- Increased organized internet crimes

The information outlives the information carrier !

Are we prepared for this?

We need to find routines, procedures, and supporting technology to ensure that digital information can be read and understood into eternity

LongRec – one step closer to the solution…

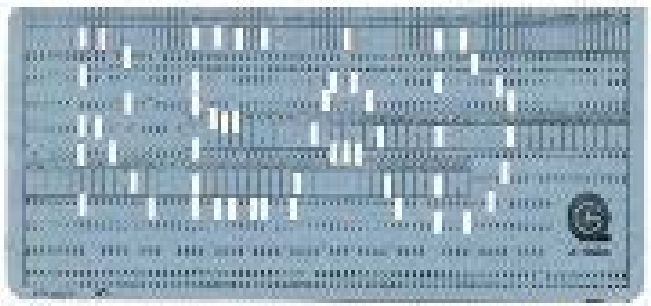
DATA =

DIGITAL ACCESS THROUGH AEONS

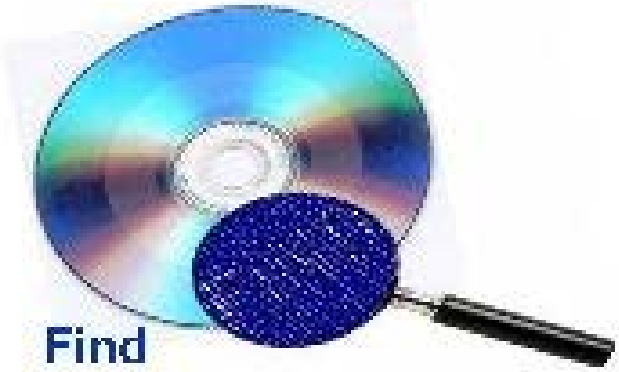
- 3+ year project, research and case studies
  - DNV R&I lead, 10 partners
  - Start October 2006, end November 2010
  - Overall budget 27,6 MNOK, Norwegian Research Council grant 9.2 MNOK
  - 3 PhD theses in work

http://www.longrec.com

DATA = **D**igital **A**ccess **T**hrough **A**eons



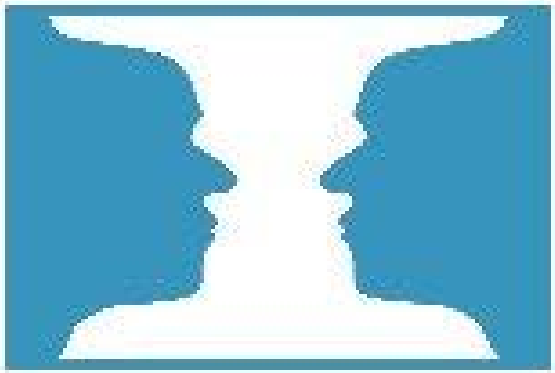
Read



Find

Trust

Understand



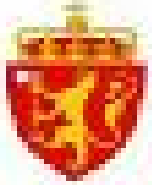
+ COMPLIANCE

Project partners

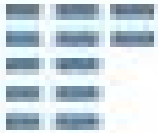
The National Library

Norsk Regnesentral

The Ministry of Foreign Affairs



The National Archival Services of Norway



Brønnøysundregistrene

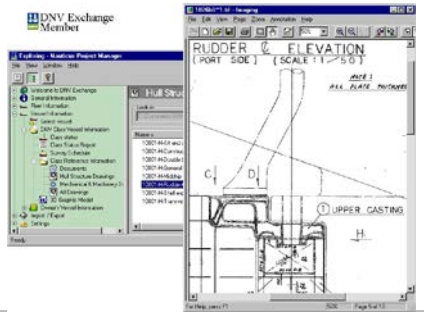
- InterPARES 3: http://www.interpares.org
- ICRI (Interdisciplinary Centre for Law and ICT), Katholieke Universiteit Leuven

The primary objective of LongRec

- Persistent, reliable and trustworthy long-term archival of digital information records, with emphasis on availability and use of information
  - Enable transition to digital (original) information and digital work processes even for information that must be available and in use over decades
  - Explore the potential for commercial products/services in this area
- Digital preservation is the foundation but not enough
  - Frozen records OK for data that shall not be changed
  - But maintenance needed (formats, storage media etc.)
  - And support for long-term work processes may be needed

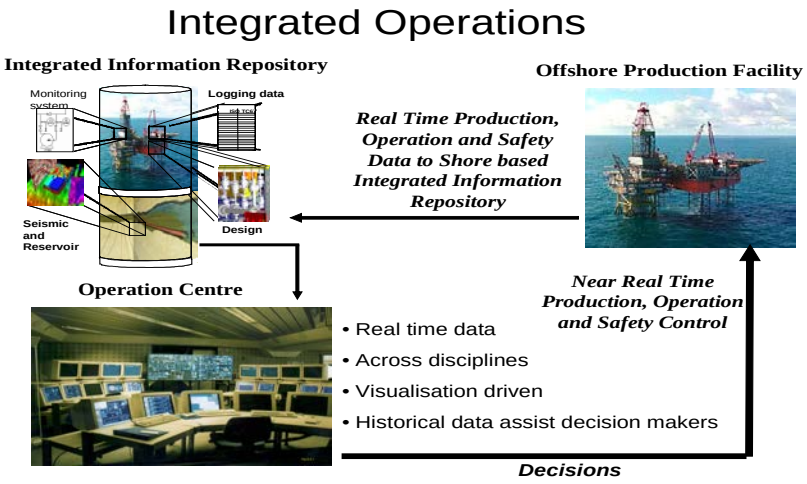
The Patient: DNV (1)

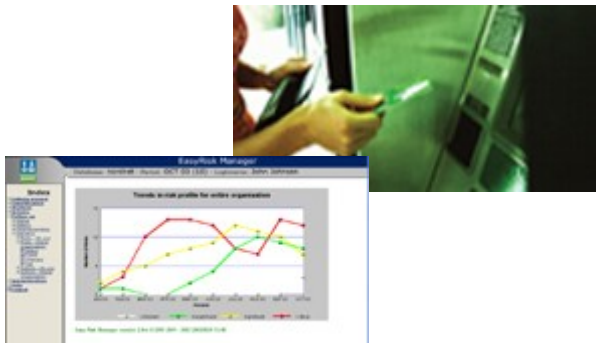
- Transition to digital documents and work processes
  - Not just digital representation of paper originals
  - To gain full benefit from the technology, processes must change
- DNV requirements
  - Documents to be stored for at least 40 years
  - Textual documents, drawings, perhaps photos and multimedia information
  - High demands for availability, integrity, authenticity and confidentiality
  - Digital signatures needed for some documents (DNV certificates)
- DNV interoperability requirements
  - Offices in more than 100 countries
  - Information from/to many actors (wharfs, ship owners, flag states, port states, insurance companies etc.)



The Patient: DNV (2)

- In 40 years, everything will have changed
  - Software, computers, formats, organization, personnel, roles
  - Records management must handle this
- Service development (external services from DNV)
  - Validation and notary services (trusted third-party roles taken by DNV)
  - Information Quality Management
  - Risk management in an information or document life cycle perspective





The patient…

The National Library of Norway

or

How to store the memory of the nation?

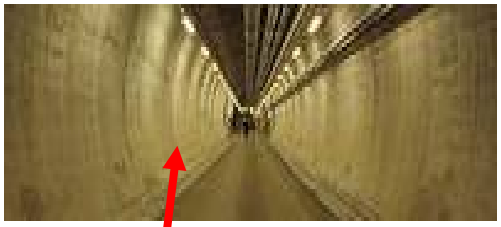
The Legal Deposit Act

The memory…

90 m long

4 floors

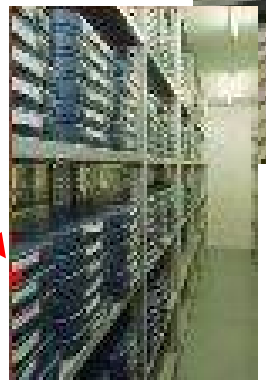
42 km shelves



100 m inside the mountain

Cold storage

Many tons of film



Automatic storage

Place to 1.500.000

documents

The patient: multitude of record types

73 000 music printings,

75 000 hrs/160 000 records

410.000 Norwegian

The national memory and multimedia knowledge centre…

Systematic digitalization of EVERYTHING

Digitalization started a while ago…

Status:

200 000 of 4 700 000 newspapers

365 000 of 1 800 000 pictures

47 000 of 410 000 books

500 of 400 000 hrs film/video/TV

1000 of 75 000 hrs music

5000 of 40 000 posters

300 000 of 1 200 000 hrs radio

0 of 4 000 000 manuscripts

0 of 55 000 maps

0 of 2 500 audio books

….

Digitalization

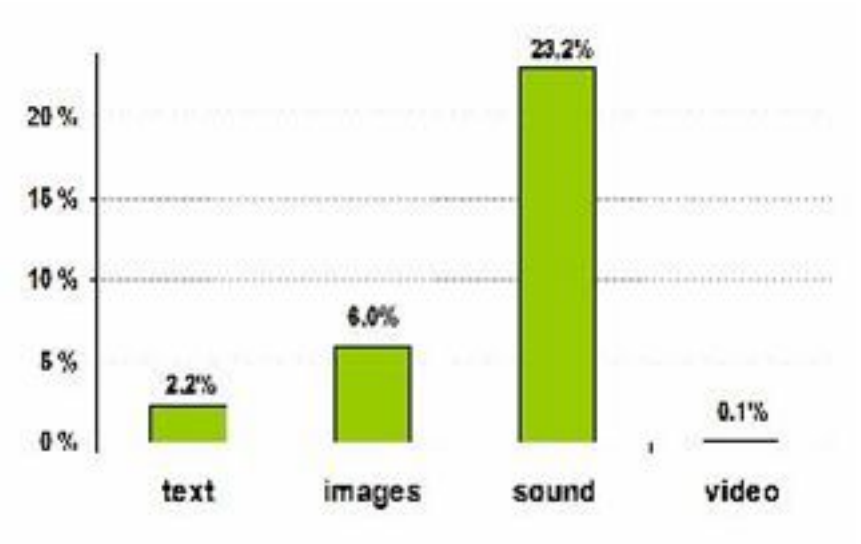
Current state of digitalization: ≈5%

Total volume when today's collections are digitalized (≈ 2018)

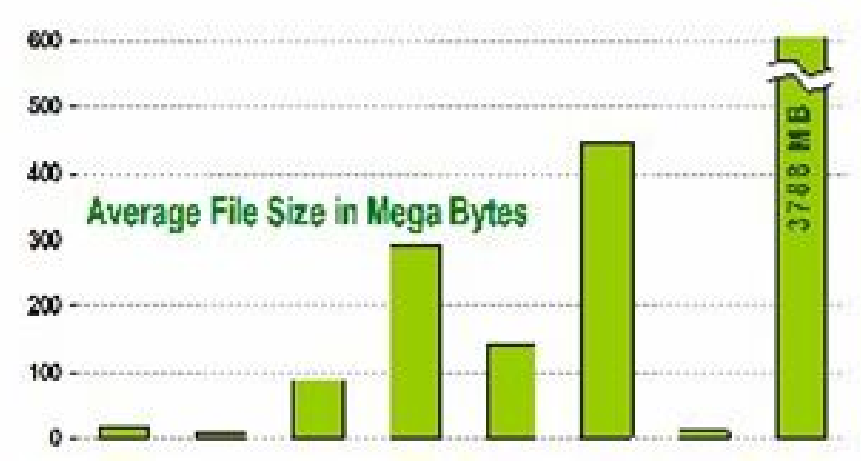
- Estimated total volume: 37 Petabyte
- Estimated number of files: 564.000.000

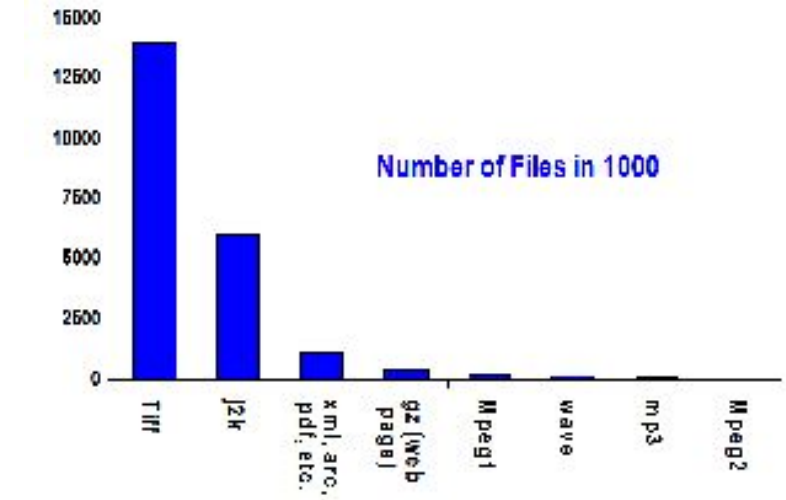
- **In addition:**
  - newly submitted materials
  - TV broadcasts, e.g. digital TV
  - web harvesting (.no domain)

Percentage of completed digitalization



File formats and volumes





File format obsolescence:

not yet an issue

Hardware Support:

3 (4) years only!!

=> copying of ALL files to new storage

(server and 2 tape)

Migration: Moving all the files to a new storage

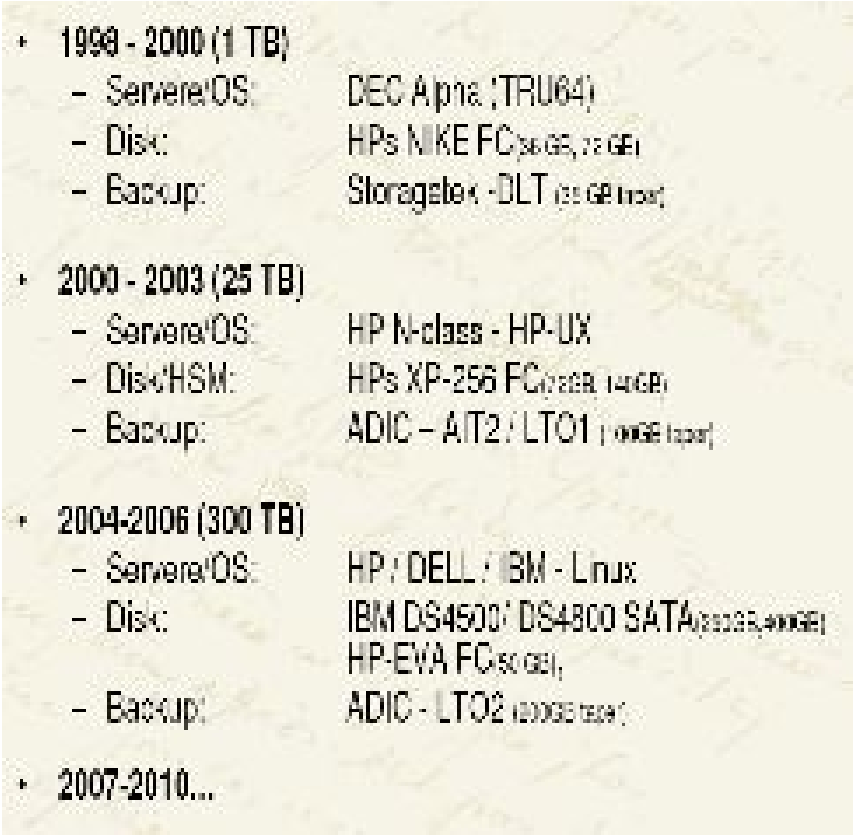
Estimated:

- 40 Petabytes (≈ 1000 TerraB ≈ 1000*1000 GigaB)
- 560 million files

Assume:

- 1 sec per file transfer
- => 17.7 years !!

More than 4 times the hardware support period



Main challenges

Data volume

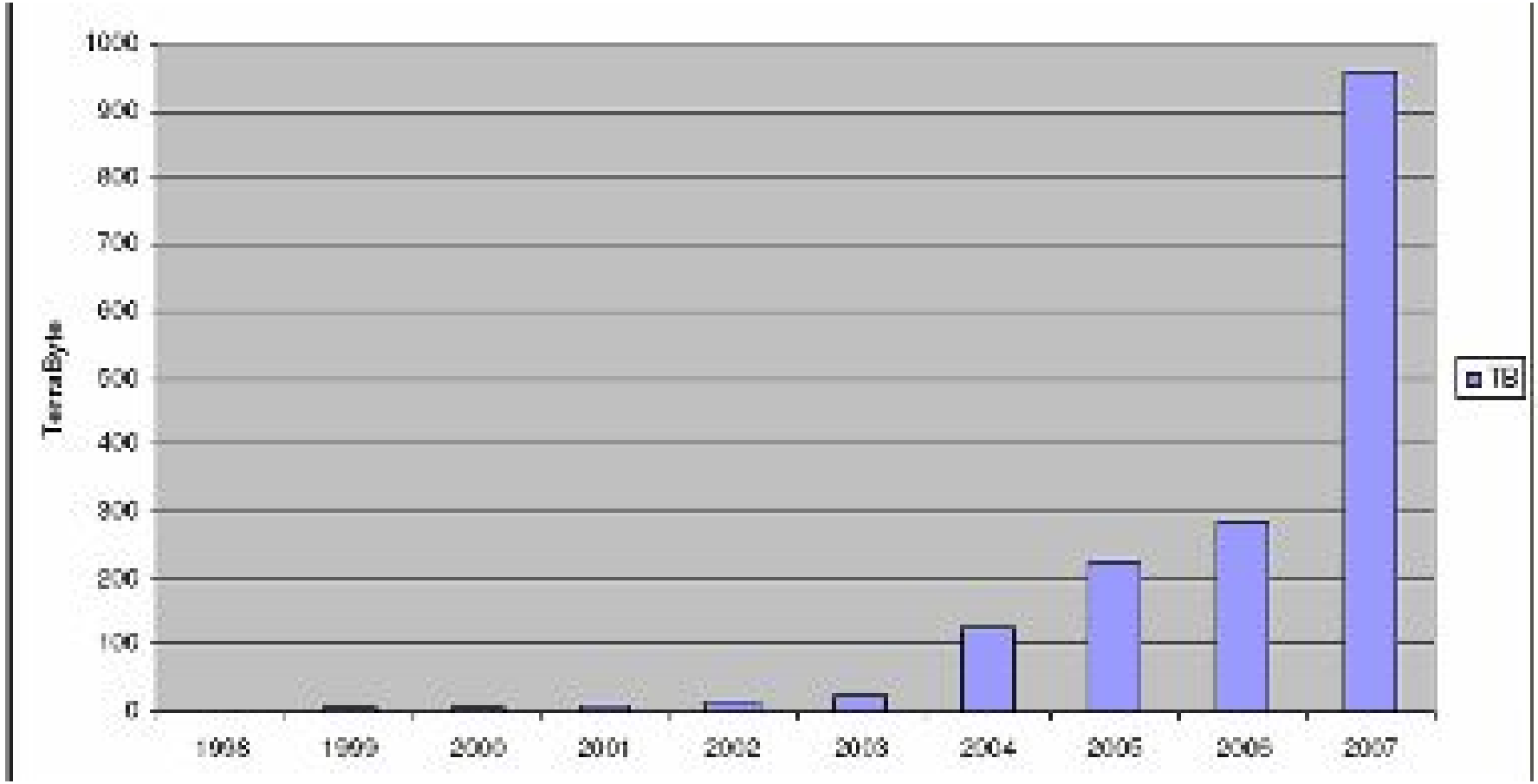
"This year there are TB, the next PB"

- The main principles: 3 copies, to different technologies, 3 places
- 1000 TB (x3) today
- + 750 TB growth annually
- Nothing can be deleted (incl. web-harvesting)

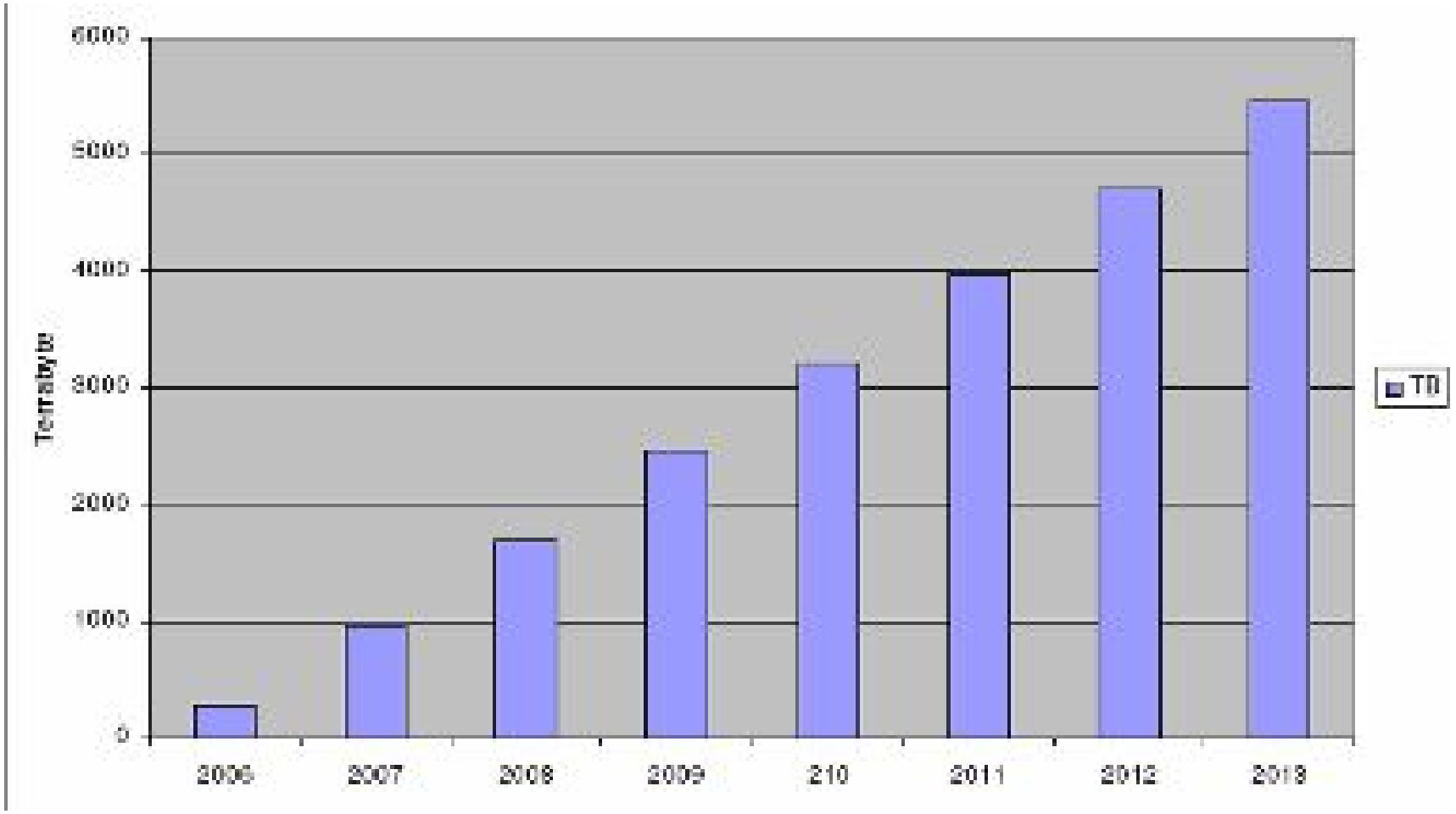
Long-term storage

- All digital content shall be preserved for at least 1000 years:
  - Searched
  - Retrieved
  - Shown
- The item displayed shall be as close to the original as possible
- Data integrity shall be secured

Data volume 1998 -2007

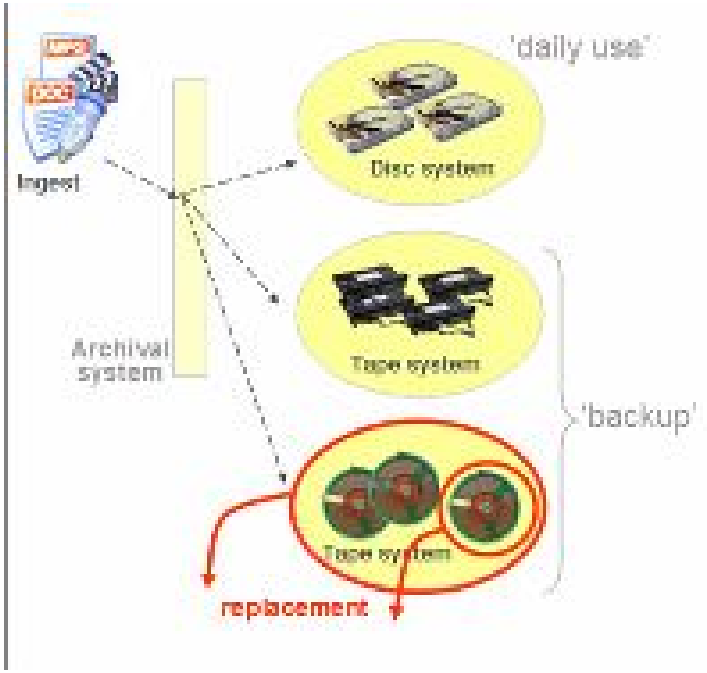


Data volume – prognosis, net



What is being done today?

- The highest quality possible for the storage of digital objects
- Unique ID
- Metadata
- Minimum 3 exemplars, 2 technologies, 2 localities
- Data integrity check
- DSM (Trusted Digital Repository) application (developed in-house): handles preservation MD and physical placement of the objects



Challenges for public agencies

1. Capture information and metadata
   - Preservation metadata (how is this preserved?), content metadata (can I search it?), context metadata (which process created this?)
   - What must be archived?
   - Turn information into archive records
   - Enforce retention policies – forever or just for some years
   - MAIN FOCUS: Work processes with user-friendly IT support
2. Ensure readability over time
   - Open, well-specified formats (but what about conversion from the original format to the archival format?)
   - Enable search and retrieval – metadata and indexing
3. Deposit information as long as agency is responsible
   - Make sure nothing is lost
   - Control of storage media and formats necessary
   - Need for external services – complexity + survival of organisational changes
4. Deliver information to National Archives
   - After 20 years or so ….

- NOARK-5 is current version of the regulations

Additionally …

1. What is digital communication?
   - Plain old documents
   - Web-forms, email, instant messaging
   - Video, pictures, multimedia, broadcasts, …
   - Geographical information
2. What goes in the archive, and why?
   - Formal correspondence, whatever that is ..
   - “Nice to have” background material
   - For historical reasons
   - … you don’t know what will be interesting in the future (old advertisements, old pictures may be of more value than the formal business documents)
3. But you probably cannot preserve everything …

Additionally 2 …

1. Public agencies run my archives
   - Individuals will not store such digital information over time
   - Must have easy access at public agency – even to old information
   - Disputes over content may arise – authenticity and integrity
2. What about signatures?
   - Store with signatures
   - Remove and record as metadata
   - Store and forget
3. Authorisations and access
   - Confidentiality must be ensured
   - Access when owner no longer able to access?
4. Compliance
   - Did I follow the rules that were in force at that time?

Wrap-up

- Challenges: HW, SW, format, processes obsolescence, organizational changes
- Volume explosion and storage shortage
- The Digital Bomb Metaphor and the LongRec project
  - Patient 1: DNV (drawings, at least 40 yrs)
  - Patient 2: the National Library of Norway
- Challenges to public agencies

Contact

- Inger-Mette Gustavsen, DNV Research & Innovation
  inger.mette.gustavsen@dnv.com
  +47 6757 7049 / +47 917 08 230

- Jon Ølnes, DNV Research & Innovation
  jon.olnes@dnv.com
  +47 478 46 094

- Olga Cerrato, DNV Research & Innovation
  olga.cerrato@dnv.com
  +47 957 35 880