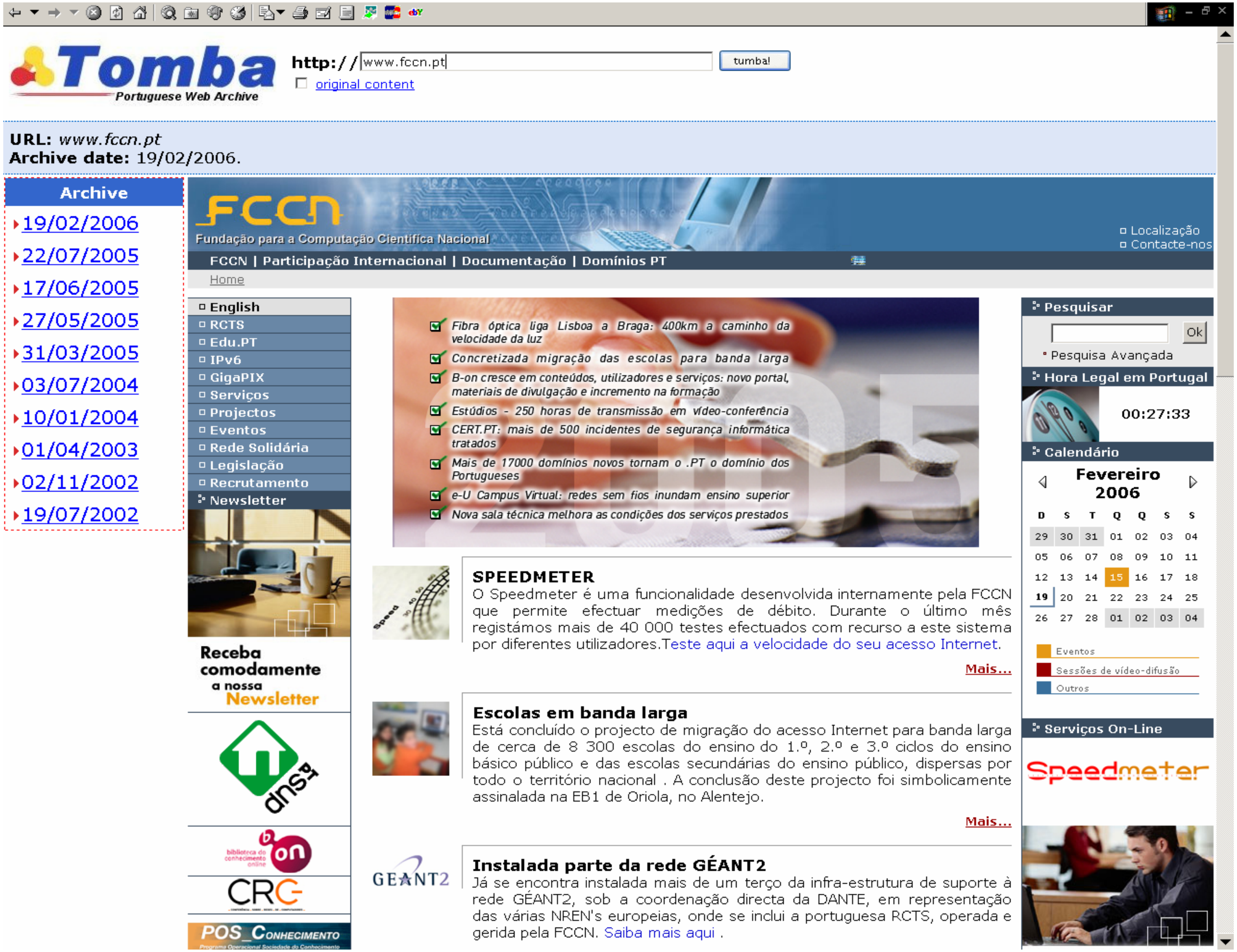
Managing Duplicates in a Web Archive

Daniel Gomes
André L. Santos
Mário J. Silva
University of Lisbon

Tomba Web Archive (tomba.tumba.pt)



Tomba Portuguese Web Archive

http:// www.fccn.pt [tumba!]

□ original content

URL: *www.fccn.pt*
Archive date: 19/02/2006.

Archive

- 19/02/2006
- 22/07/2005
- 17/06/2005
- 27/05/2005
- 31/03/2005
- 03/07/2004
- 10/01/2004
- 01/04/2003
- 02/11/2002
- 19/07/2002

FCCN
Fundação para a Computação Científica Nacional

Localização
Contacte-nos

FCCN | Participação Internacional | Documentação | Domínios PT

Home

- English
- RCTS
- Edu.PT
- IPv6
- GigaPIX
- Serviços
- Projectos
- Eventos
- Rede Solidária
- Legislação
- Recrutamento
- Newsletter

Receba comodamente a nossa Newsletter

- ☑ *Fibra óptica liga Lisboa a Braga: 400km a caminho da velocidade da luz*
- ☑ *Concretizada migração das escolas para banda larga*
- ☑ *B-on cresce em conteúdos, utilizadores e serviços: novo portal, materiais de divulgação e incremento na formação*
- ☑ *Estúdios - 250 horas de transmissão em vídeo-conferência*
- ☑ *CERT.PT: mais de 500 incidentes de segurança informática tratados*
- ☑ *Mais de 17000 domínios novos tornam o .PT o domínio dos Portugueses*
- ☑ *e-U Campus Virtual: redes sem fios inundam ensino superior*
- ☑ *Nova sala técnica melhora as condições dos serviços prestados*

SPEEDMETER
O Speedmeter é uma funcionalidade desenvolvida internamente pela FCCN que permite efectuar medições de débito. Durante o último mês registámos mais de 40 000 testes efectuados com recurso a este sistema por diferentes utilizadores.Teste aqui a velocidade do seu acesso Internet.

Mais...

Escolas em banda larga
Está concluído o projecto de migração do acesso Internet para banda larga de cerca de 8 300 escolas do ensino do 1.º, 2.º e 3.º ciclos do ensino básico público e das escolas secundárias do ensino público, dispersas por todo o território nacional . A conclusão deste projecto foi simbolicamente assinalada na EB1 de Oriola, no Alentejo.

Mais...

Instalada parte da rede GÉANT2
Já se encontra instalada mais de um terço da infra-estrutura de suporte à rede GÉANT2, sob a coordenação directa da DANTE, em representação das várias NREN's europeias, onde se inclui a portuguesa RCTS, operada e gerida pela FCCN. Saiba mais aqui .

Pesquisar
[Ok]
Pesquisa Avançada

Hora Legal em Portugal
00:27:33

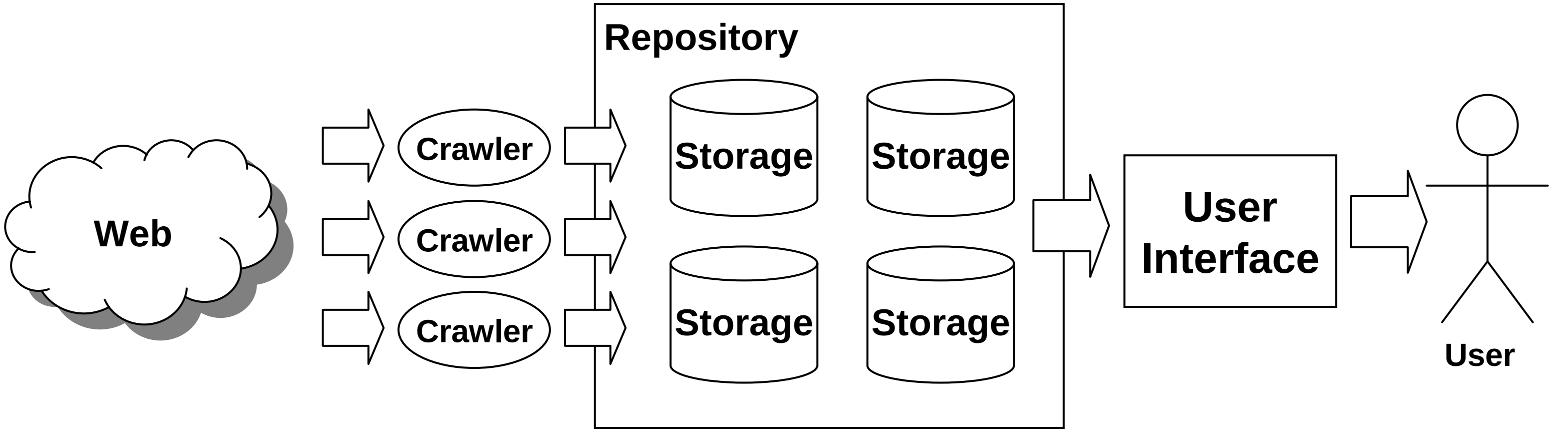
Calendário
Fevereiro 2006

D	S	T	Q	Q	S	S
29	30	31	01	02	03	04
05	06	07	08	09	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	01	02	03	04

- Eventos
- Sessões de vídeo-difusão
- Outros

Serviços On-Line
Speedmeter

Tomba architecture



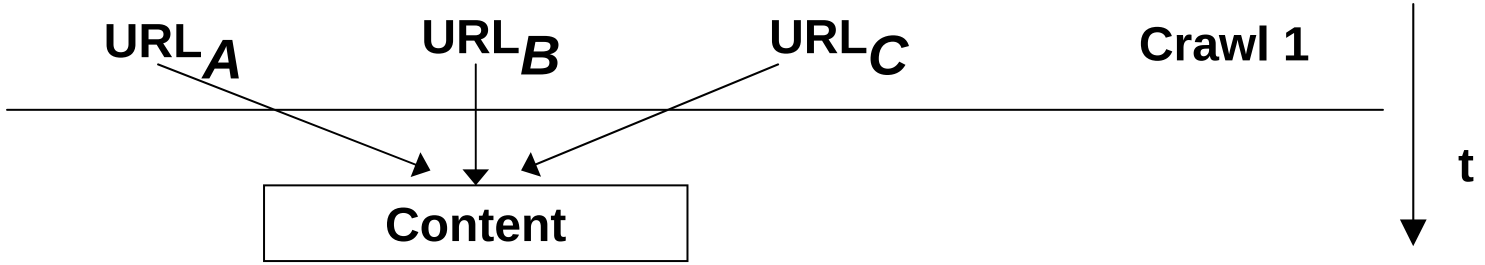
- Periodical crawls of the Portuguese web
- Repository:
    - Distributed and extensible storage system
    - Support accesses while new contents are being loaded

Duplication within the archive

Crawl	Date	#URLs	%duplicates within the crawl (horizontal)	%duplicates from the last crawl (vertical)
1	07-2002	1.6M	23%	
2	10-2002	1.2M	21%	7%
3	03-2003	3.5M	15%	10%
4	10-2003	3.3M	11%	19%
5	06-2004	4.4M	7%	18%

- **On average 25% of the archived contents were duplicates: waste of storage space**
- **Disks are cheap but the web is enormous.**
  - **Storage space must be spared**

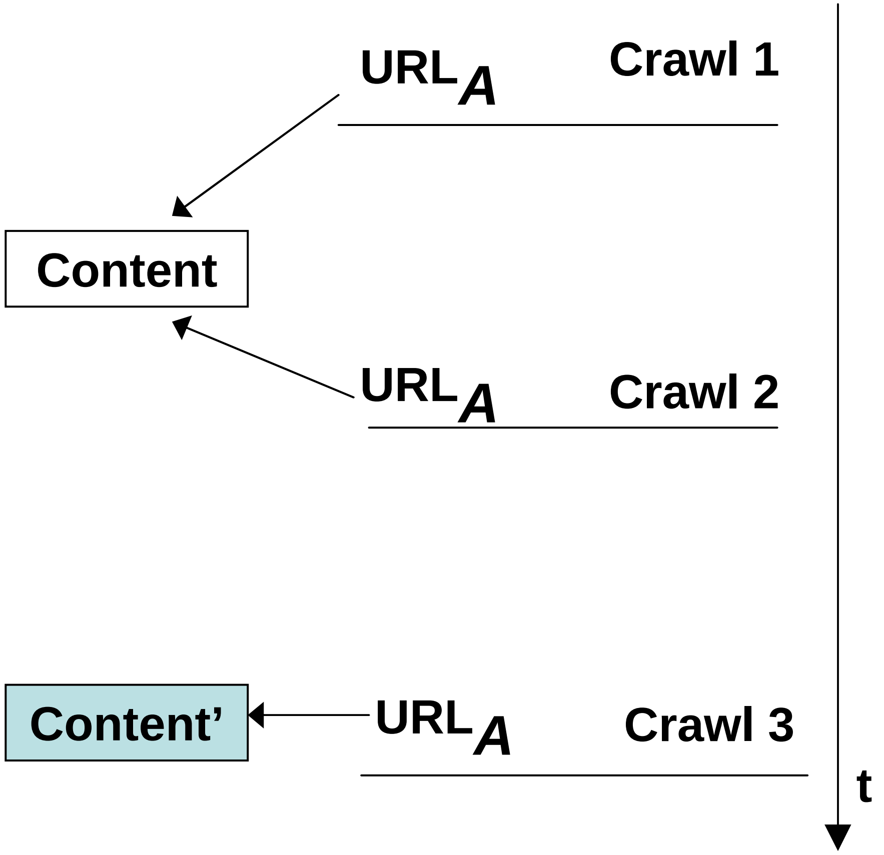
Horizontal duplication



- Mirrors
  - tucows.ip.pt == www.tucows.com
- URL aliases
  - www.yahoo.com/index.html == www.yahoo.com/
- Default error messages
  - “This web page uses frames, but your browser doesn't support them”

Vertical duplication

- Web collections built incrementally
- Exact duplicates
    - The page remains unchanged
- Partial duplicates
    - The page suffers changes with time



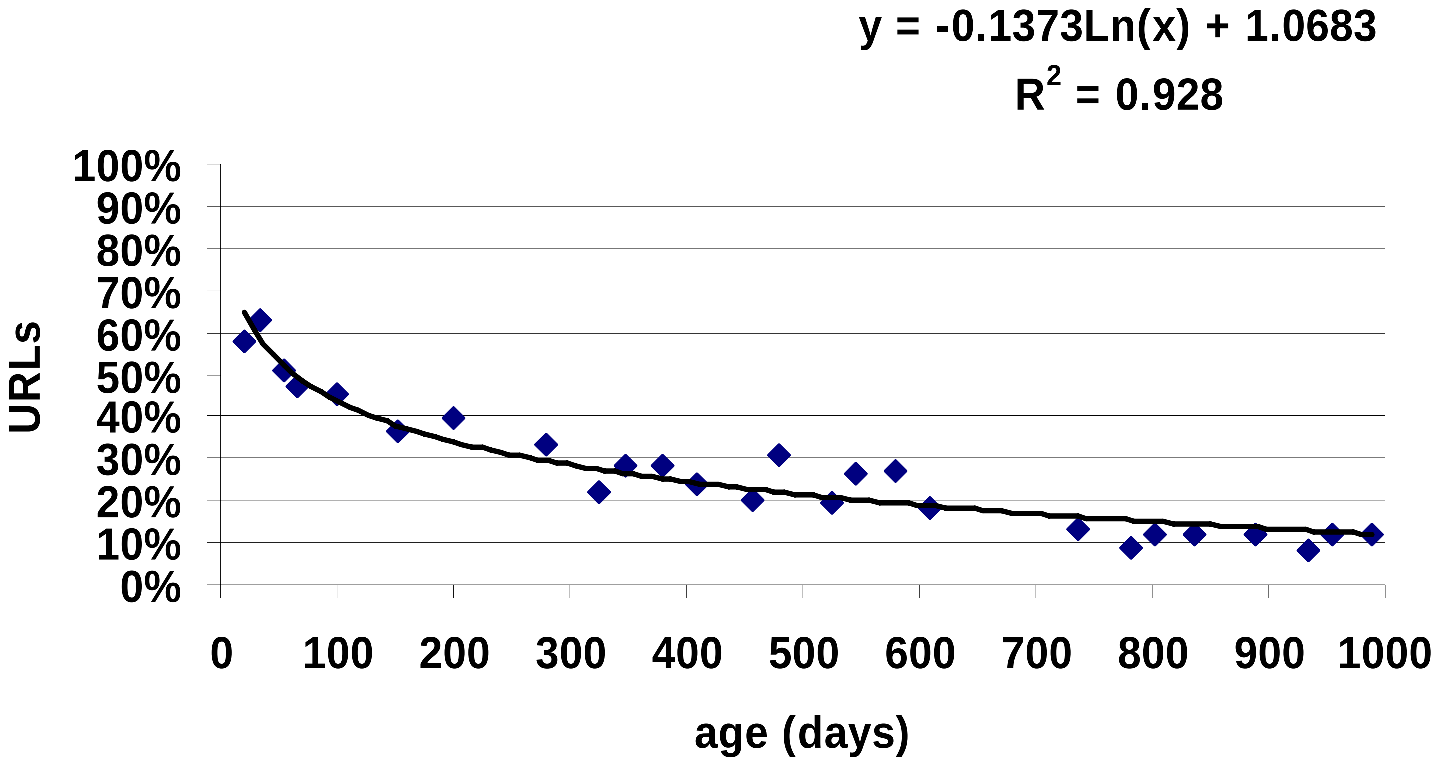
Objective: save on storage space by eliminating duplicates

- Can not jeopardize the system’s performance
- Consider a distributed and extensible storage space
- Consider preservation issues

1. Avoid crawling duplicates: estimate frequency of change (Cho,03)

- Advantages
  - Saves bandwidth
  - Reduces duration of the crawl
- Disadvantages
  - Does not eliminate horizontal duplication
  - Requires historical data to create model
  - Assumes URL persistence

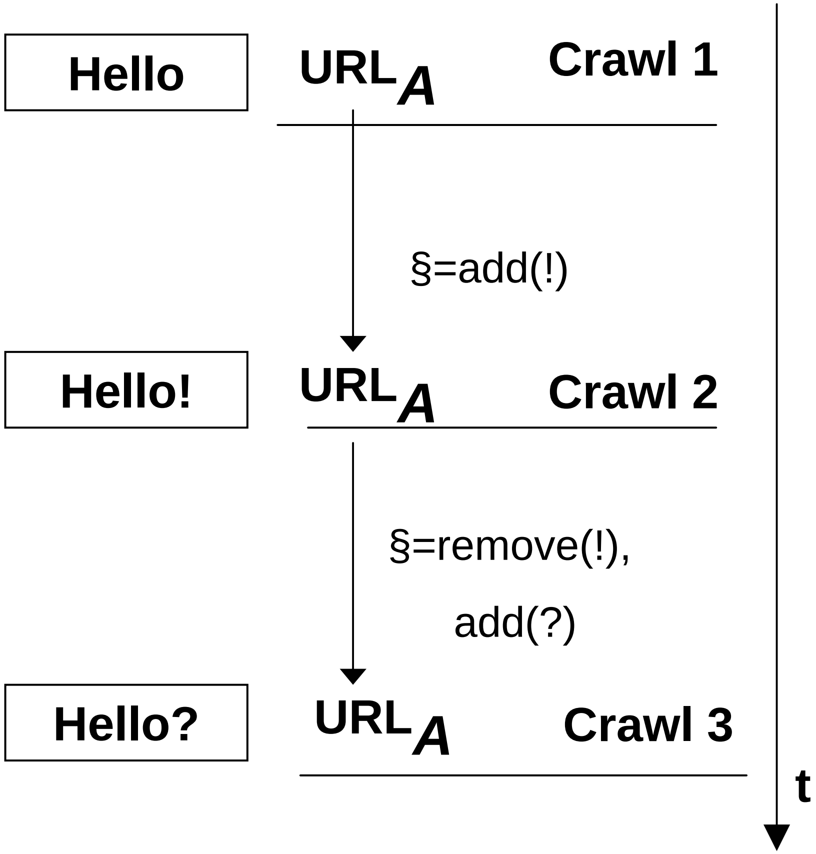
URLs are not persistent



- **Analysis of the persistence of URLs among 10 crawls of the Portuguese web gathered for 3 years**
- **Half of the URLs took less than 2 months to disappear**

2. Delta storage mechanisms

- **Store only the part that has changed**
- **Advantages**
  - **Eliminates partial duplicates**
  - **Available software (CVS)**
- **Disadvantages**
  - **Assumes URL persistence**
  - **Does not eliminate horizontal duplication**
  - **Documents are rebuild: preservation problem**



3. Distributed file systems

- Advantages
  - Available software: NFS, AFS
  - Extensible, distributed
- Disadvantages
  - General usage, not designed to fulfill the requirements of web archives
    - POSIX interface: support permissions, caching of files
  - Difficult management: requires admin. privileges
  - Do not have a built-in mechanism to eliminate duplicates
    - Store everything.
    - Find and delete duplicates in batch
    - Disk IO is expensive

4. Compression

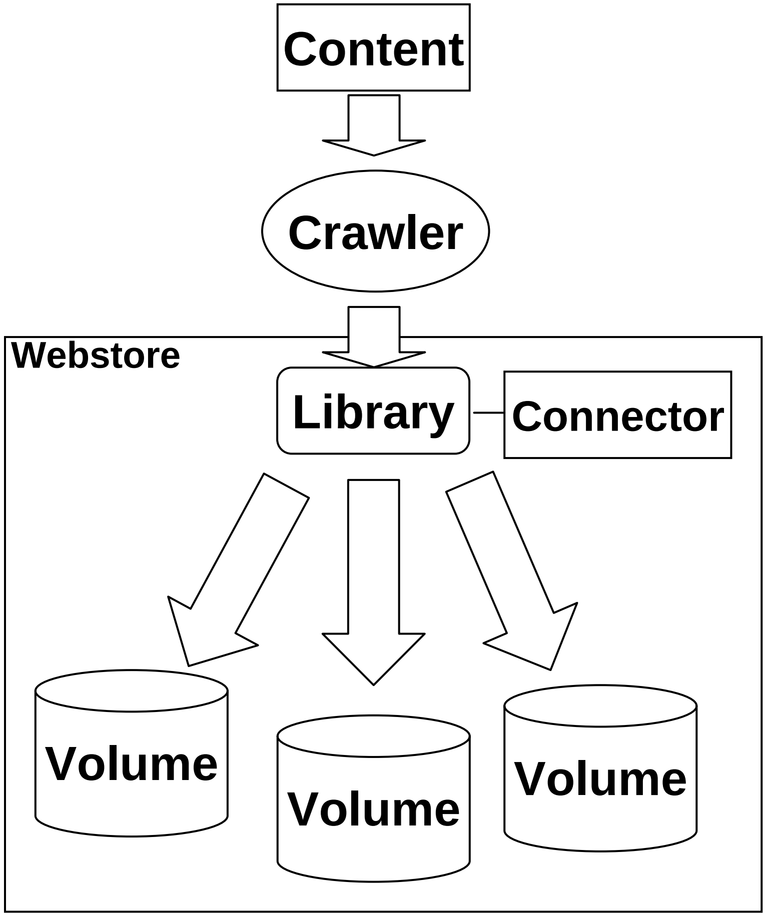
- Compression algorithms save space by eliminating duplication between contents
- Advantages
  - Many compression algorithms available (zip, rar, zlib)
  - Independent from URLs
    - Eliminates vertical duplicates
    - Eliminates horizontal duplicates
- Disadvantages
  - Duplicates are eliminated only within each compressed file
    - Scalability problem: files get too big
  - Corruption of 1 file jeopardizes several contents

Our solution: Elimination of exact duplicates based on content signatures

- Advantages
  - Independent from URLs
  - Eliminates vertical and horizontal duplication
  - Lightweight algorithm
  - Improve storage throughput
    - Duplicates are not written to disk
- Disadvantages
  - Does not eliminate partial duplicates

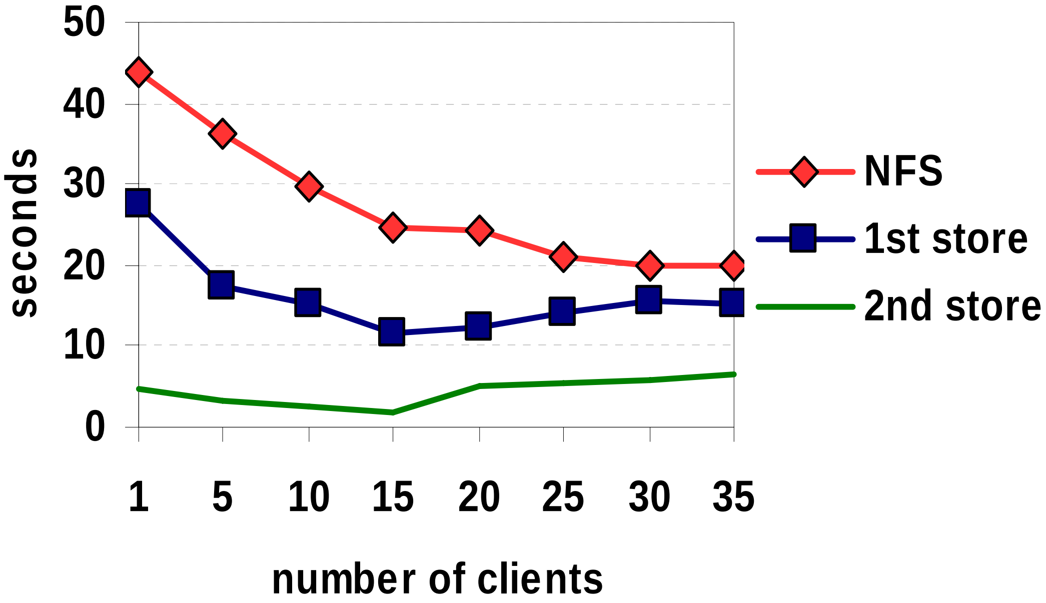
Validation: Webstore

1. **Generate content signature**
2. **Find it among the volumes**
3. **If it is already stored, discard content**
4. **If it is a new content, store content**



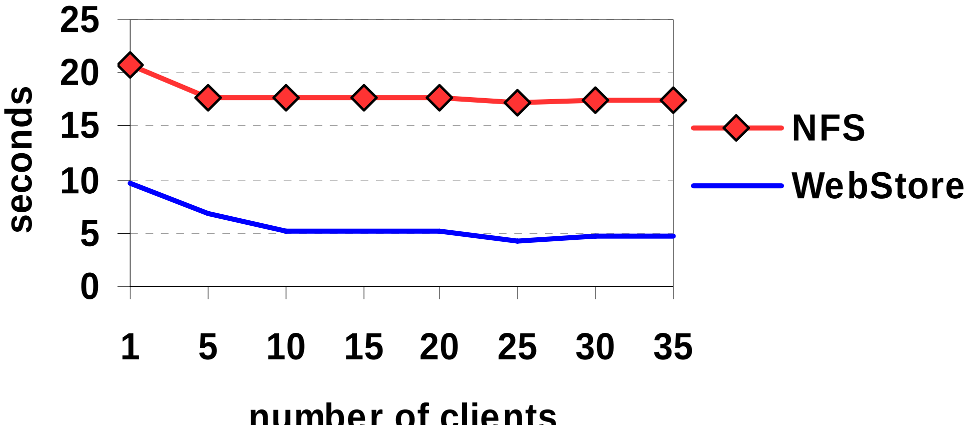
Write

- **Elimination of dups. must not jeopardize performance**
- **NFS as baseline**
  - **Widely available**
  - **Reproducibility**
- **Time took to store 1000 pages**
- **Increasing number of clients storing on 1 volume**
- **Writes on average 50% faster than NFS**

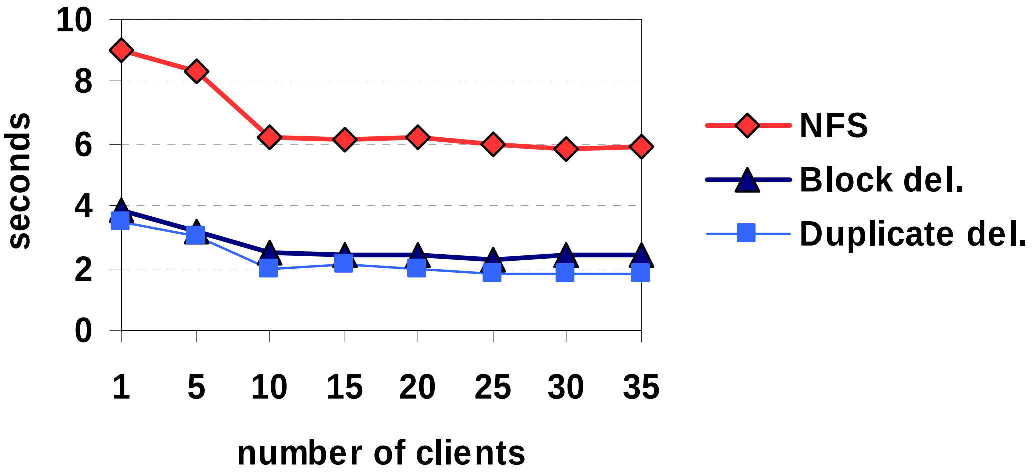


Read and Delete

- **Reads on average 68% faster than NFS**



- **Deletes on average 60% faster than NFS**



Conclusions

- There is vertical and horizontal duplication on the web
  - Duplicates are frequent within a web archive
- URLs transience is a problem
  - Estimation of frequency of change
  - Delta storage
- Webstore eliminates vertical and horizontal exact duplicates without jeopardizing performance
  - 57 million web contents in Webstore (1.5 TB).

“Future” Work

- Study URL and Content Persistence
  - Daniel Gomes and Mário J. Silva, “Modelling Information Persistence on the Web”, ICWE 2006 (to appear).

http://webstore.sourceforge.net (Webstore)

http://tomba.tumba.pt (web archive)

http://xldb.fc.ul.pt (research group)

dcg@di.fc.ul.pt

Thank you for your attention.