Vienna University of Technology
Institute of Information Systems
Database and Artificial Intelligence Group

DBAI

Creating Permanent Test Collections of Web Pages for Information Extraction Research

Bernhard Pollak & Wolfgang Gatterbauer

presentation by Bernhard Pollak

for the SOFSEM 07

January 2007

It´s about ...

Creating Permanent Test Collections of Web Pages for Information Extraction Research

Saving Web Pages

Why saving web pages?

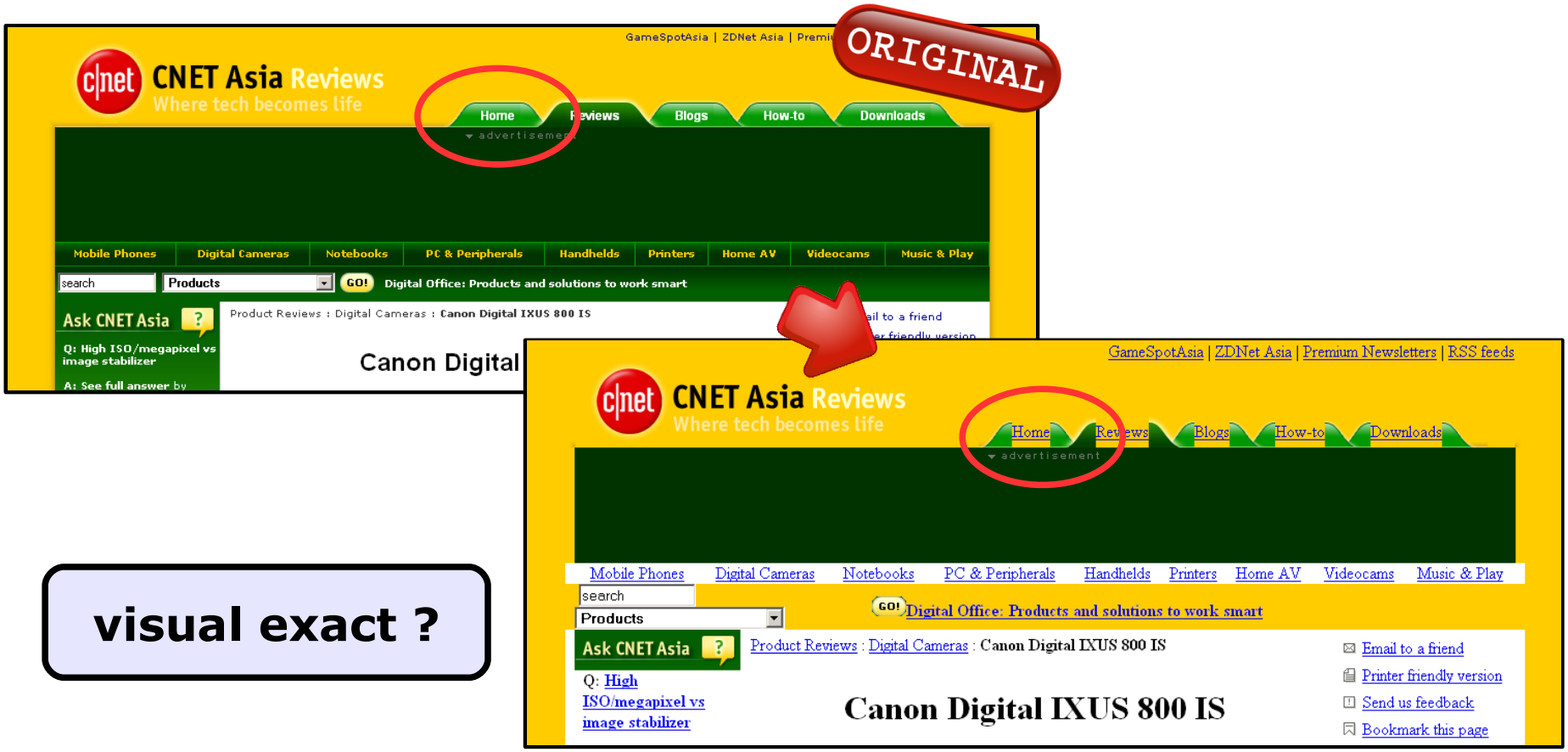
⇒ We **want to**:

- *reproduce*
- *share*
- *compare*

⇒ *AND* because of **visual** web page IE:

- **visual exact** reproduction

Firefox 2.0 "Save Complete"



visual exact ?

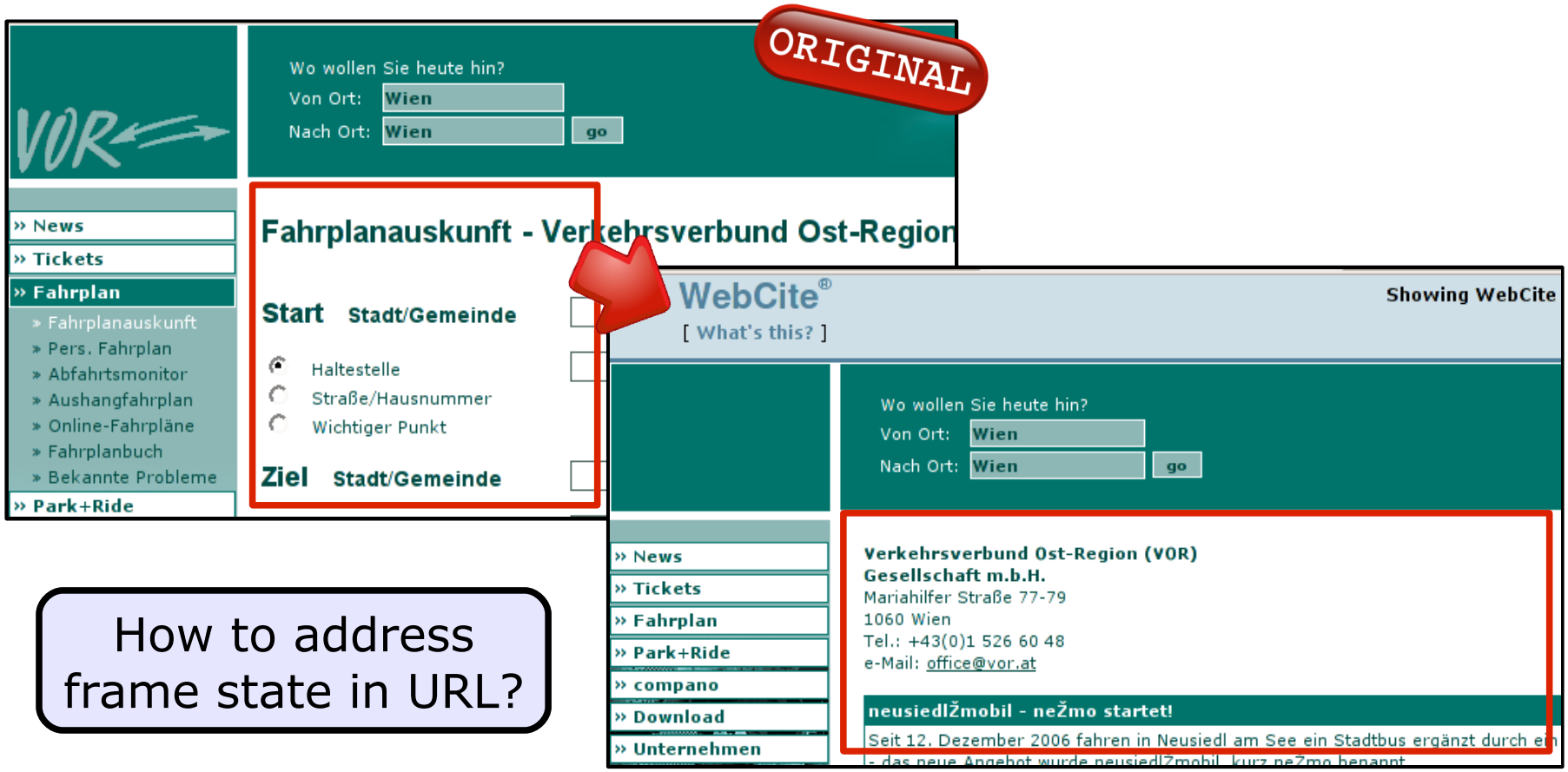
Source: http://asia.cnet.com/reviews/digitalcameras/0,39005890,39116525p,00.htm

Web Site Downloader: HTTrack 3.33

The menu is completely lost!

Source: http://www.booking.expedia.de/booking/exp/ibe/index.php?KID=360010&tab=...

Personal Web Archive: WebCite



How to address frame state in URL?

Source: http://www.booking.expedia.de/booking/exp/ibe/index.php?KID=360010&tab=...

Personal Proxy: WWWOFFLE 2.8e

HTTPS problem!

Source: https://shop.sacher.com/sacher/SacherShop/index_en.jsp

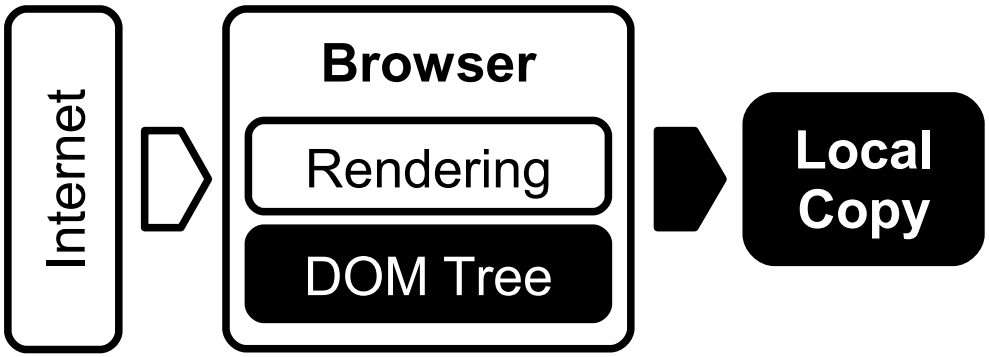
The Solution: Saving based on DOM Tree

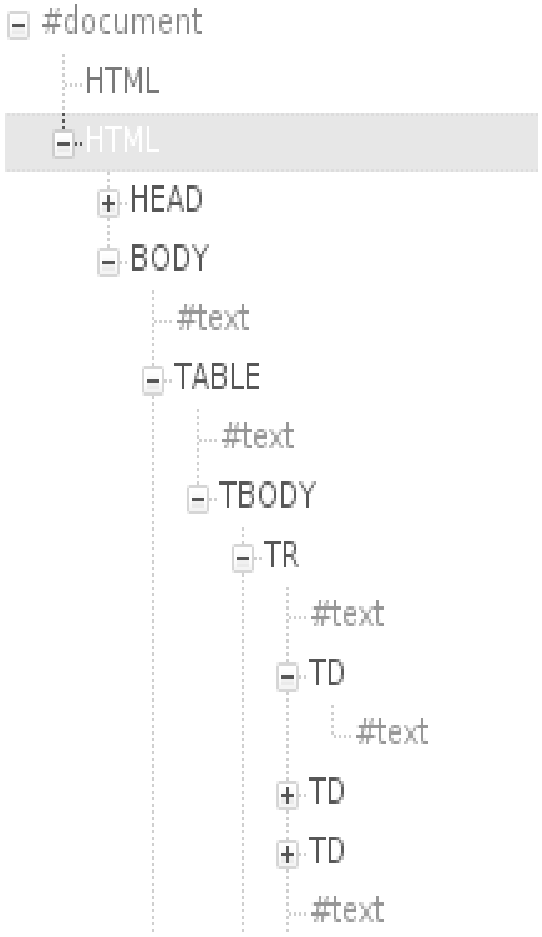
⇒ **no conceptual problems**

⇒ a *client centric* approach

⇒ *dynamic* changes are considered

⇒ *decision on visual relevant objects*





Our Solution: WebPageDump

⇒ based on the *Scrapbook DOM Saver* extension

⇒ adapted for IE researchers needs

⇒ AND: ***improved quality of local copy***

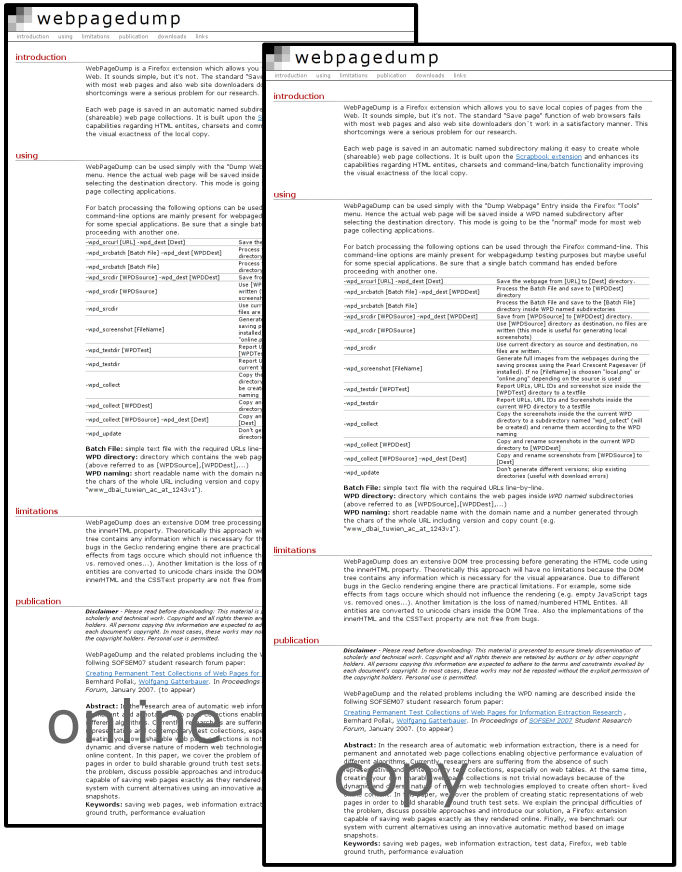
But: **How** to measure the **visual exactness** of web pages?

For download at http://www.dbai.tuwien.ac.at/user/pollak/webpagedump

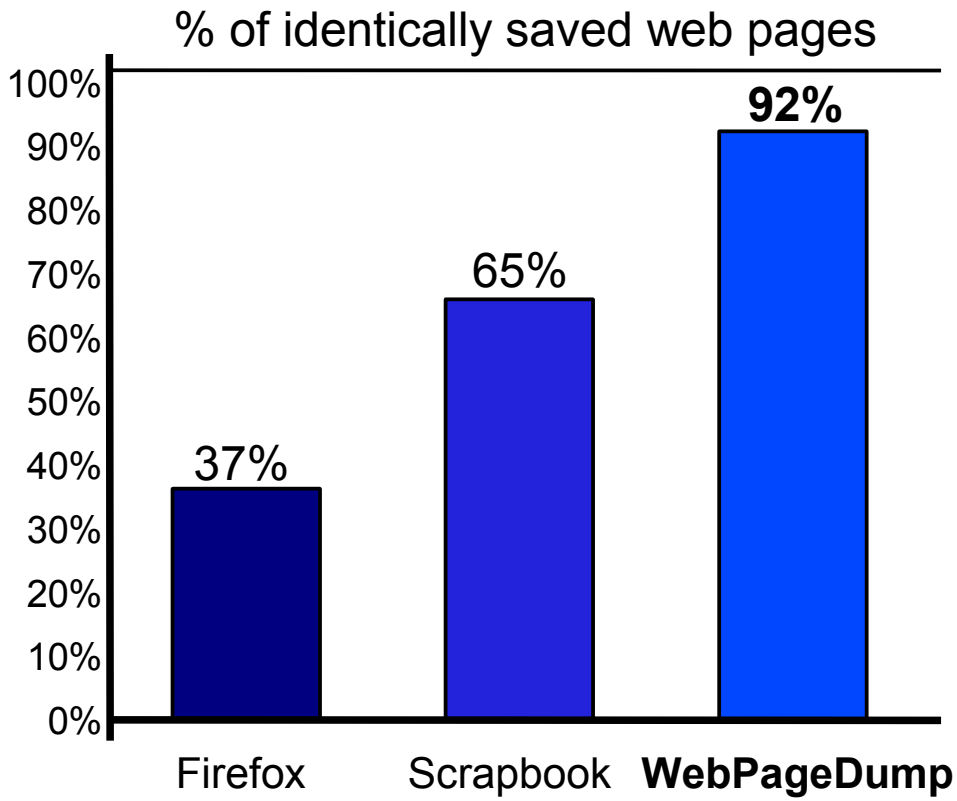
Experimental Setup

⇒ Image Comparison
- using *web page image*
- PNG Format (compressed)

⇒ Strong correlation
- between *pixel differences* and *compressed file size*



Experimental Results



⇒ 450 web pages

⇒ Tested domains:

- arabic
- chinese
- various languages
- random
- digital cameras

reduction of faults from 35% to 8%

Thank you for your attention

Paper:

Bernhard Pollak, Wolfgang Gatterbauer. Creating Permanent Test Collections of Web Pages for Information Extraction Research. In Proceedings of SOFSEM 2007 Student Research Forum, January 2007.

WebPageDump:

http://www.dbai.tuwien.ac.at/user/pollak/webpagedump

Additional References:

HTTrack: http://www.httrack.com

WWWOFFL: http://www.gedanken.demon.co.uk/wwwoffle

WebCite: http://www.webcitation.org

Scrapbook: http://amb.vis.ne.jp/mozilla/scrapbook