Webarchiv

Budoucnost českého webového archivu

Jsme Webarchiv

digitální knihovna,
která uchovává webové zdroje
pro budoucí generace.

WWW

Pokud je nebudeme průběžně
archivovat, zmizí významná
součást národního
kulturního dědictví.

Jak archivujeme?

Provádíme kompletní archivaci
“celého” českého webu.

Souběžně probíhá výběrová
a tematická archivace.

Bohužel!

Ne všechna data jsou
dostupná online.

Může za to současná podoba
autorského zákona, která byla vytvořena
pro knihy. Pro přístup k celému archivu
musíte prozatím až k nám.

Budoucnost

Webový archiv není jen skladiště URL,
na které usedá prach. Pracujeme na vytvoření
fulltextu celého archivu. Potřebujeme
porozumět tomu, co nesou jednotlivé
digitální objekty a co budou
znamenat historicky.

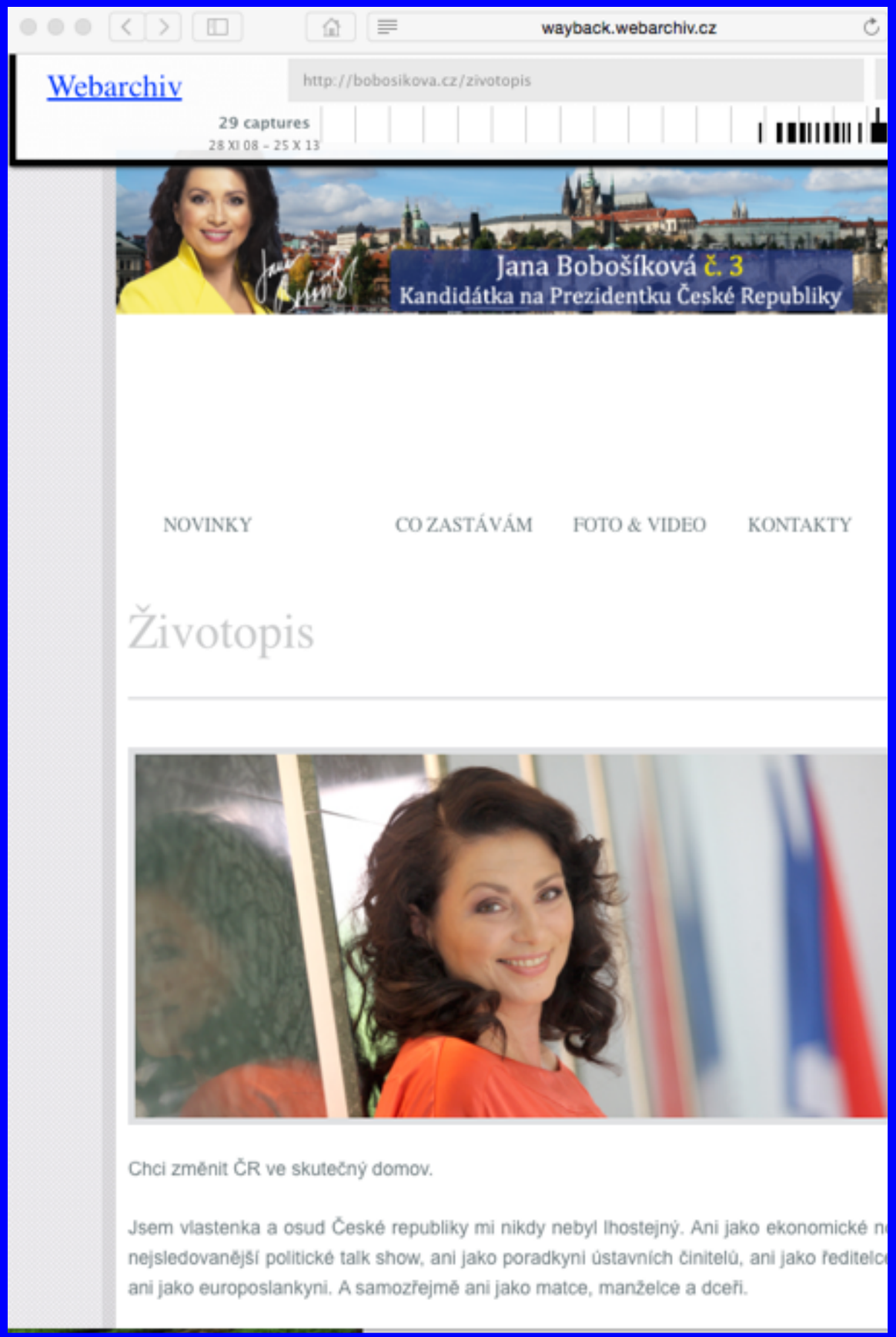
Čeká nás otevření Webarchivu
analytickému výzkumu a
propojení našich dat s jinými archivy.

wwW

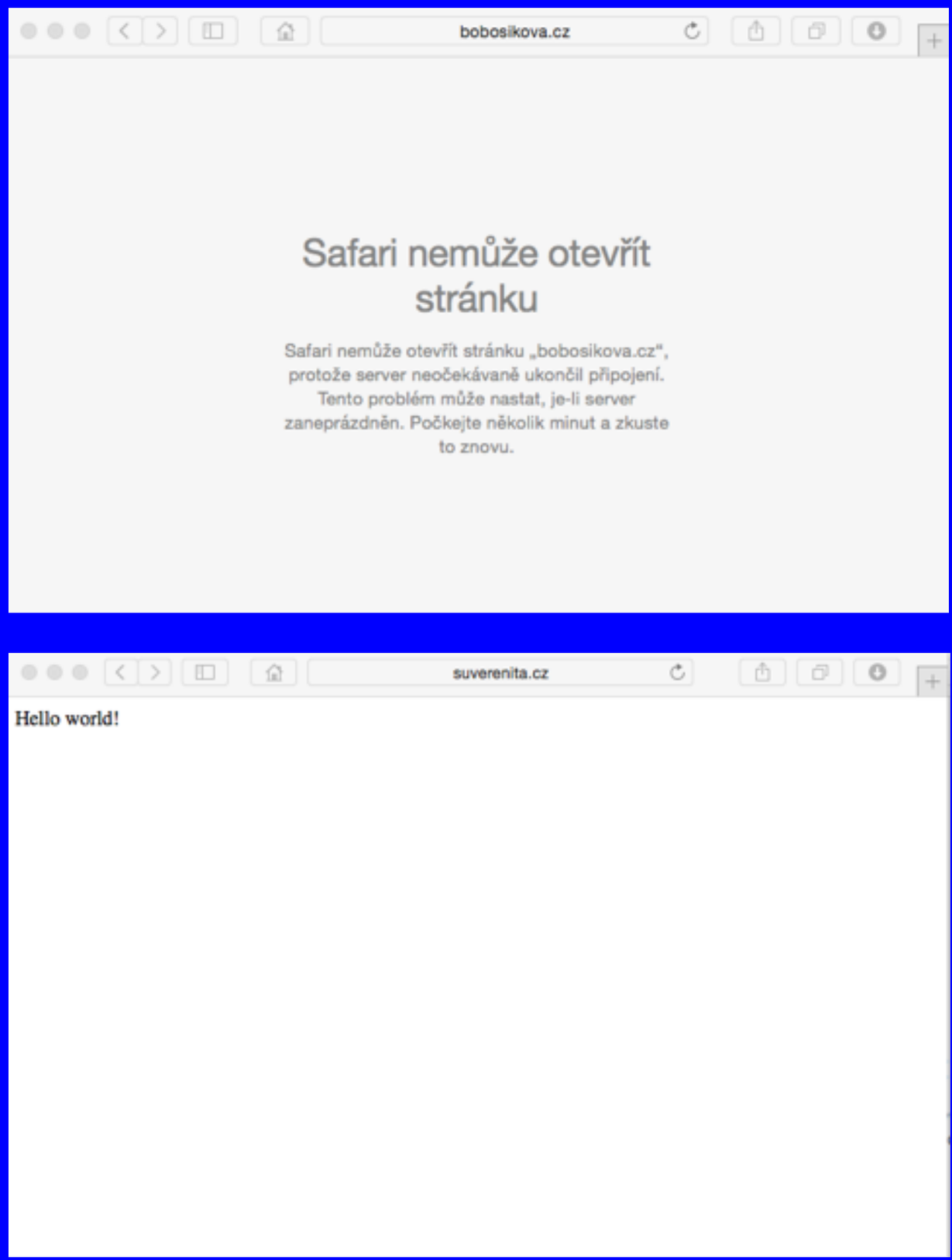
Bude možné studovat 90. léta a dál bez webových archivů?

Ian Milligan

Webový archiv



Živý web



Bude možné studovat 90. léta a dál bez webových archivů?

Ne.

Nové otázky, nový přístup

Jak cirkulují obrázky v kolekci v průběhu let?

Profilování domén v čase, souborové formáty, velikost a hloubka.

Které weby již neexistují, které se často mění a které málo?

Jaké se měnili verze HTML v čase?

Mění se témata webů během let?

Na jaké weby nesměřují žádné odkazy?

Atd..

Situace

Do konce roku 2015

350 TB komprimovaných dat

~4 miliardy nejrůznějších digitálních objektů

Jak na to

Výpočetní výkon

Apache Hadoop klastr v Metacentru NGI

Nástroje na zpracování dat

YARN a machine learning

Přání versus **Realita**

Zpracování kolekce Webarchivu

Identifikace formátu jednotlivých dig. objektů

verze PDF, HTML, MS Word apod.

Extrakce plného textu

z HTML, PDF, DOC apod.

Extrakce metadat

Geotagy aj. info z EXIF u obrázků, autoři z dokumentů apod.

Analýza textových dokumentů Webarchivu

Textový hash dokumentů: pro hledání podobných textů

Rozponání žánru: např. recenze, rozhovor, článek apod.

Identifikace entit: např. místa, osoby, události apod.

Identifikace témat a klíčových slov

Rozpoznání jazyka

Analýza obrazových dokumentů Webarchivu

Obrazový hash: hledání podobných obrázků

Slovní popis obrázků včetně klíčových slov

černé a ryšavé koťátko si hrají na zelené trávě

Rozpoznávání tváří

Intepreter pro historické formáty

Software Library: MS-DOS Games

Search

Share

Favorite

Software for MS-DOS machines that represent entertainment and games. The collection includes action, strategy, adventure and other unique genres of game and entertainment software. Through the use of the EM-DOSBOX in-browser
MORE

About | Collection

SORT BY VIEWS · TITLE · DATE ARCHIVED · CREATOR

2,603 RESULTS

Oregon Trail, The
by MECC
1.1M | 1,000 | 104

Prince of Persia
by Jul 29, 2014
491,926 | 534 | 25

Oregon Trail Deluxe, The
321,478 | 561 | 67

Wolfenstein 3D
by id Software, Inc.
320,598 | 452 | 9

SimCity
by Maxis Software Inc.
237,059 | 287 | 10

Search this Collection

software 2,603

PART OF

The Software Library: MS-DOS Software Library

Co můžeme realizovat nyní

WAT: Web Archive Transformation
Metadatový výcuc z každého ARC/WARC v kolekci
Struktura souboru: JSON
Obsahuje: HTML hlavičky, MIME type, velikost souborů, URL odkazy apod.

Wayback CDX API
Vystaví obsah indexu všech URL v kolekci, včetně záznamu o datu.

Co realizujeme na Hadoop

Datasety
výsledky formátové analýzy kolekce
hashe dokumentů a obrázků
prolinkování domén v čase

Služby
3CPO - webový explorační nástroj na procházení výsledky
FITS formátové analýzy

A co zajímá vás?

Děkujeme za pozornost!

Jaroslav Kvasnica

Rudolf Kreibich