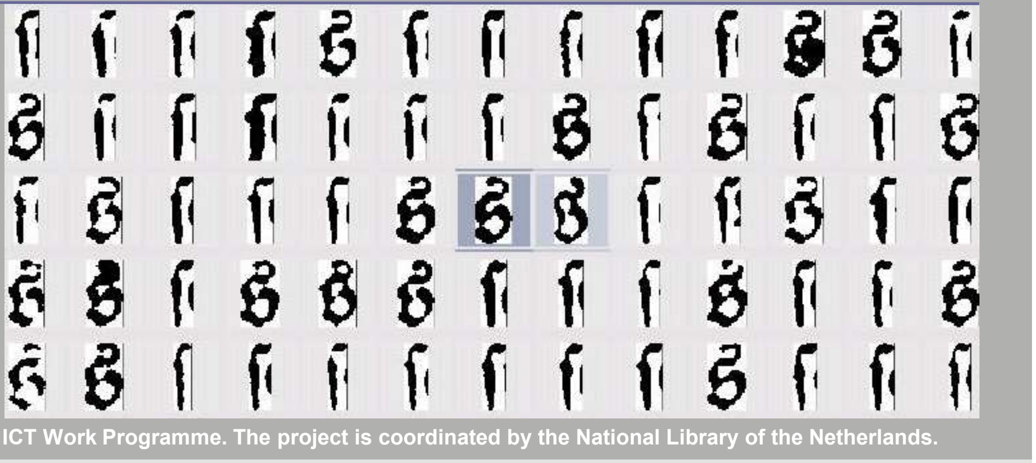
Improving Access to Text

IMPACT



IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands.

The Improving Access to Text (IMPACT) project and other European initiatives

Michael Day
UKOLN, University of Bath
m.day@ukoln.ac.uk
http://www.ukoln.ac.uk/

JISC Workshop: OCR for the Mass Digitisation of Textual Materials, University of Bath 24 September 2009

UKOLN

Presentation outline

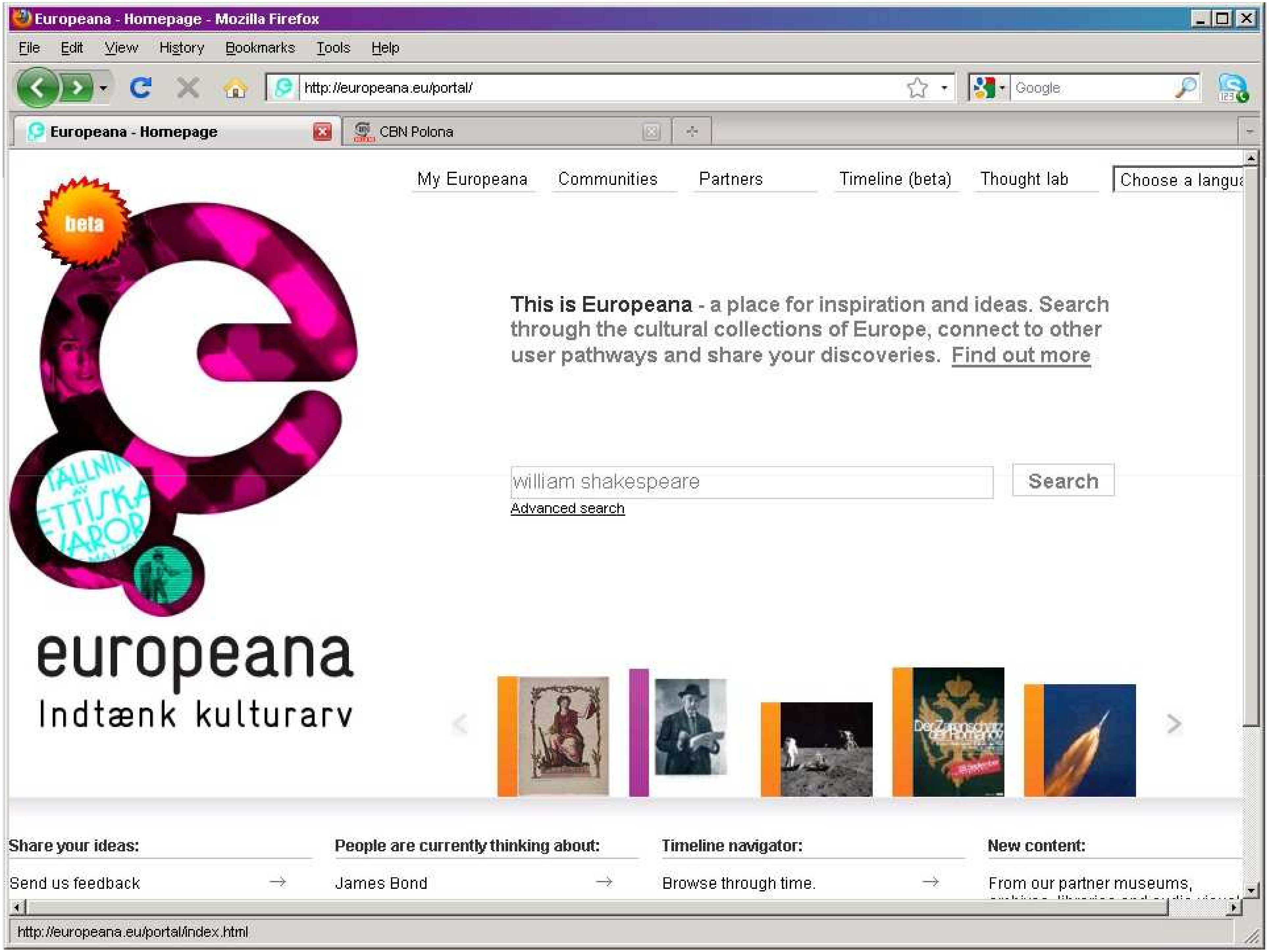
- Contexts
  - Some European digitisation activity
  - Digitisation challenges
- The IMPACT project
  - The consortium and project structure
  - Major project activities

Digitisation activity in Europe (1)

- European Commission
  - i2010 digital libraries initiative
    - Launched September 2005
    - Bringing together European cultural heritage online
      - Europeana portal
- Many projects dealing with the digitisation of texts in Europe
  - Many at large-scale, with selectivity at collection level or higher (industrial-scale mass digitisation)
  - Content holders often work with commercial providers (e.g., outsourcing of conversion processes, partnering with Google Books)
  - However, "Europe is facing a very important cultural and economic challenge: Only some 1% of the books in Europe's national libraries have been digitised so far, leaving an enormous task ahead of us" (Viviane Reding and Charlie McCreevy, EU Commissioners, September 2009)

Digitisation activity in Europe (2)

- Europeana - Europe's digital library
  - Website: http://europeana.eu/portal/
  - Launched in November 2008
  - Hosted by the National Library of the Netherlands; run by the European Digital Library Foundation
  - Part funded by the EU's eContent *plus* programme
  - A portal providing access to ca. 4.6 million items
  - Mixed content:
    - Books, newspapers, photographs, maps, film clips
    - Books included are mainly those in the public domain
  - EC public consultation on Europeana and the digitisation of books, open until 15 November 2009 http://ec.europa.eu/information_society/newsroom/cf/itemlongdetail.cfm?item_id=5181



This is Europeana - a place for inspiration and ideas. Search through the cultural collections of Europe, connect to other user pathways and share your discoveries. Find out more

Share your ideas:

Send us feedback →

People are currently thinking about:

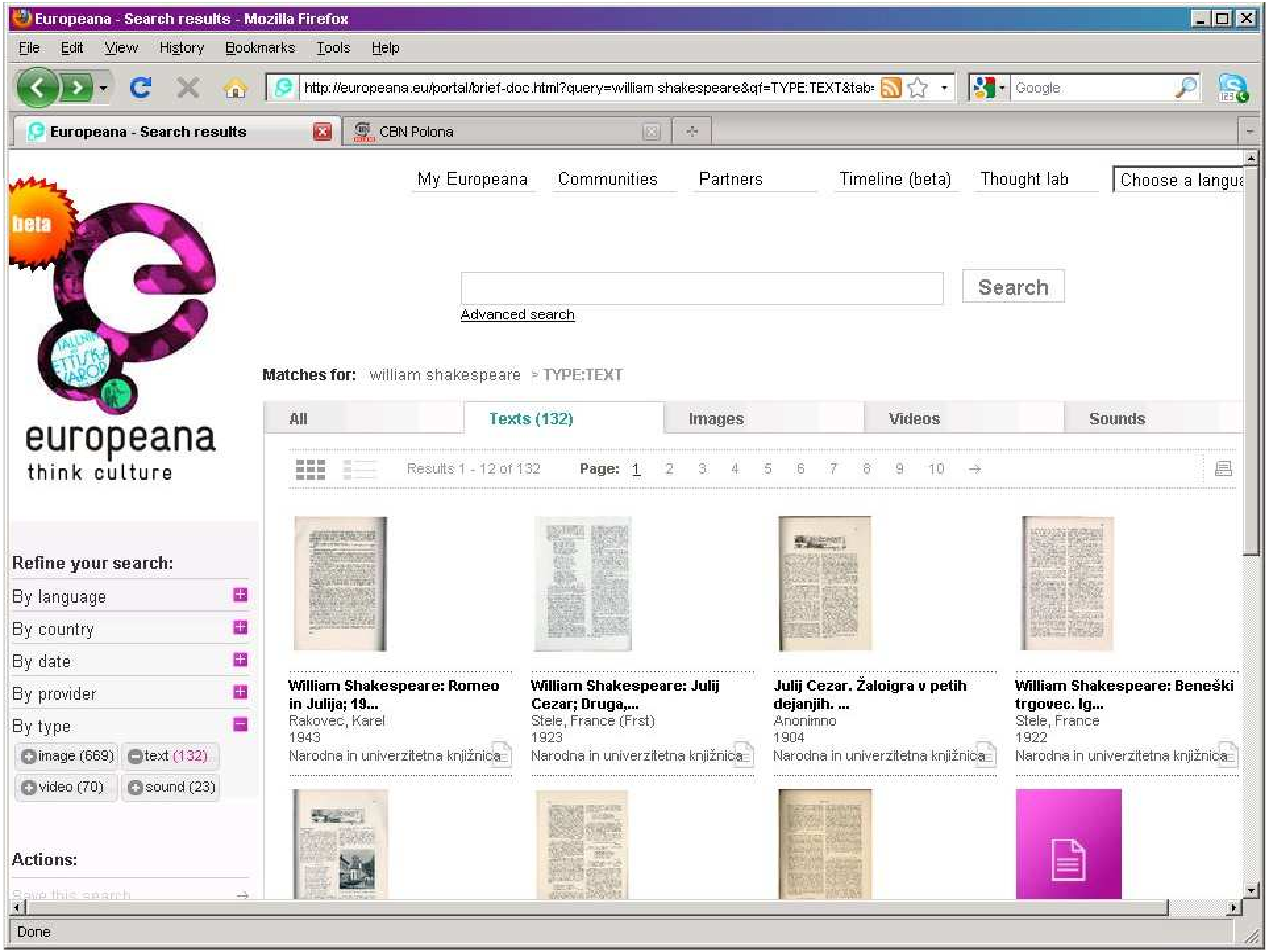
James Bond →

Timeline navigator:

Browse through time. →

New content:

From our partner museums,



Europeana - Search results - Mozilla Firefox

http://europeana.eu/portal/brief-doc.html?query=william shakespeare&qf=TYPE:TEXT&tab:

Europeana - Search results | CBN Polona

Matches for: william shakespeare > TYPE:TEXT

Results 1 - 12 of 132 Page: 1 2 3 4 5 6 7 8 9 10 →

Refine your search:

By language

By country

By date

By provider

By type

image (669) text (132)

video (70) sound (23)

Actions:

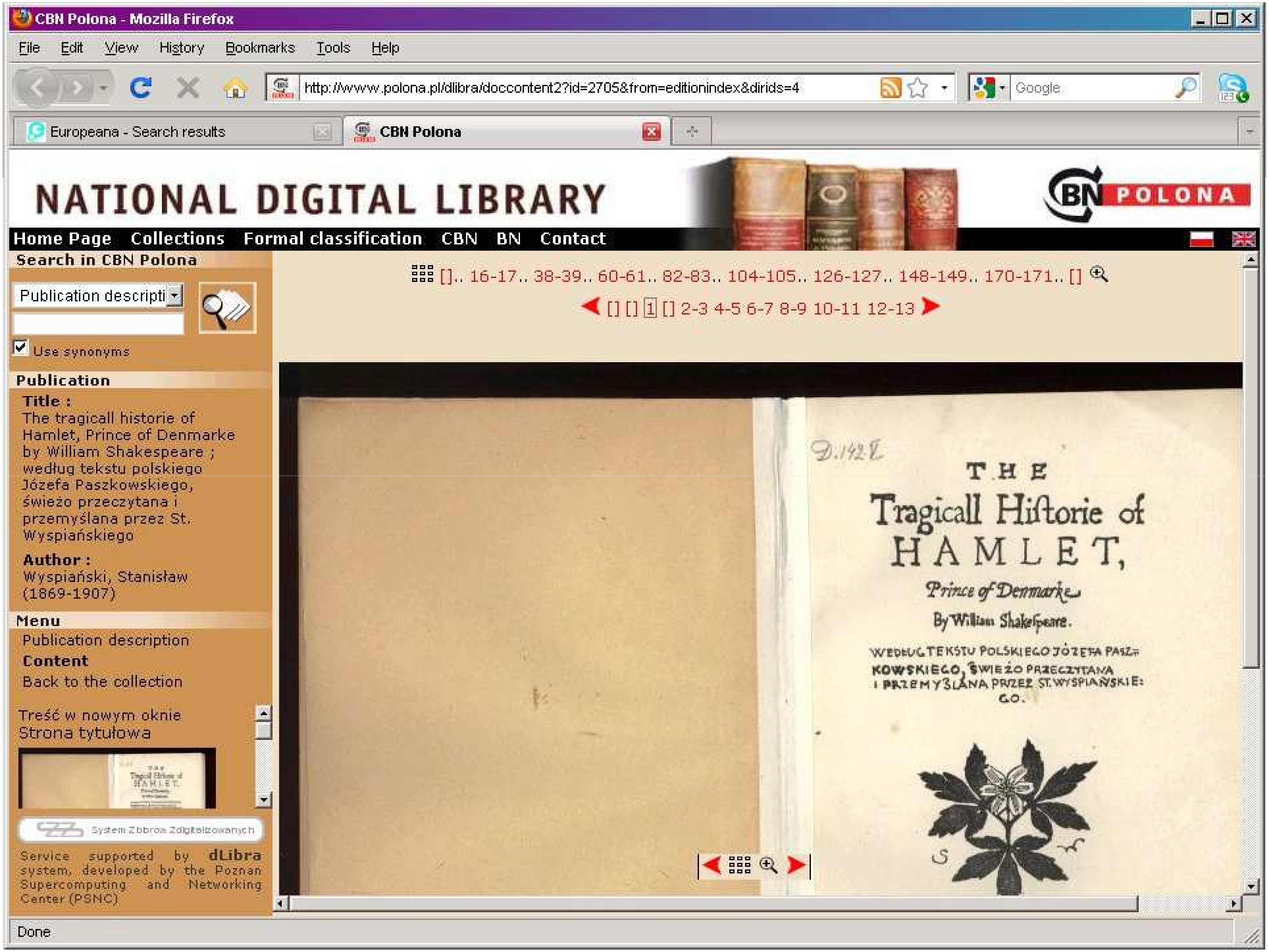
Save this search →

William Shakespeare: Romeo in Julija; 19...
Rakovec, Karel
1943
Narodna in univerzitetna knjižnica

William Shakespeare: Julij Cezar; Druga,...
Stele, France (Frst)
1923
Narodna in univerzitetna knjižnica

Julij Cezar. Žaloigra v petih dejanjih. ...
Anonimno
1904
Narodna in univerzitetna knjižnica

William Shakespeare: Beneški trgovec. Ig...
Stele, France
1922
Narodna in univerzitetna knjižnica



CBN Polona - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.polona.pl/dlibra/doccontent2?id=2705&from=editionindex&dirids=4
Google
Europeana - Search results
CBN Polona

NATIONAL DIGITAL LIBRARY

CBN POLONA
Home Page Collections Formal classification CBN BN Contact
Search in CBN Polona
Publication descripti
☑ Use synonyms

Publication

Title :
The tragicall historie of Hamlet, Prince of Denmarke by William Shakespeare ; według tekstu polskiego Józefa Paszkowskiego, świeżo przeczytana i przemyślana przez St. Wyspiańskiego

Author :
Wyspiański, Stanisław (1869-1907)

Menu
Publication description
Content
Back to the collection
Treść w nowym oknie
Strona tytułowa
System Zbiorów Zdigitalizowanych

Service supported by dLibra system, developed by the Poznan Supercomputing and Networking Center (PSNC)

D.142 I

THE
Tragicall Hiſtorie of
HAMLET,
Prince of Denmarke
By William Shakeſpeare.

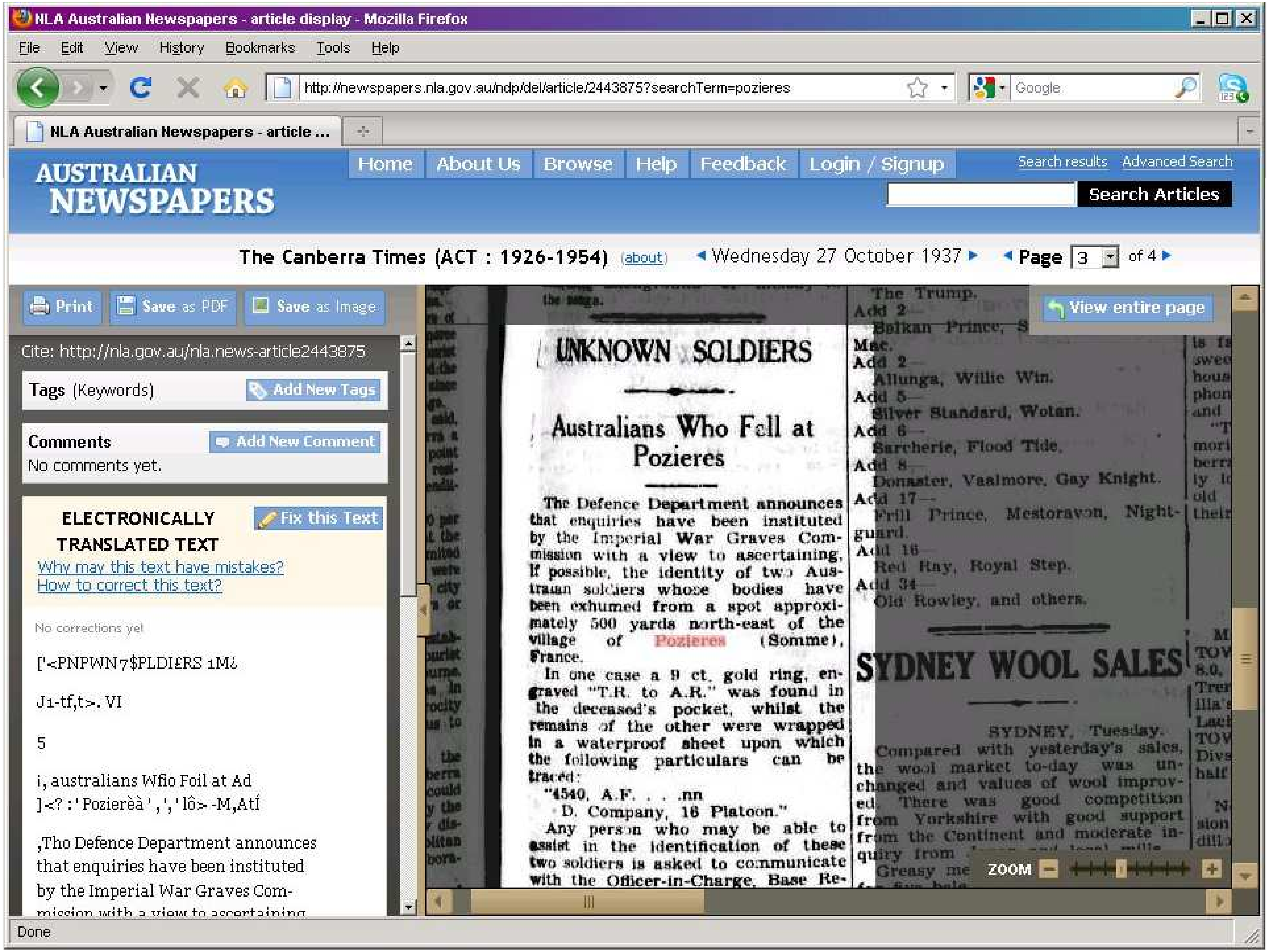
WEDŁUG TEKSTU POLSKIEGO JÓZEFA PASZ=
KOWSKIEGO, ŚWIEŻO PRZECZYTANA
I PRZEMYŚLANA PRZEZ ST. WYSPIAŃSKIE=
GO.

Digitisation challenges (1)

- Large-scale digitisation
  - Mostly based on the image front searching technique (pioneered by projects like JSTOR)
    - Scan physical item to create digital images of pages
    - Subject those pages to OCR
    - Combine OCR output with the images, OCR output considered good enough for searching, but any ambiguous results are able to be compared with page images
    - “The strategy of linking page images with OCR enables us to make effective use of large corpora of relatively cheaply scanned books and was, in large measure, effective because it points backwards to the limitations of print: search gets human readers to the page and leaves them to parse out its meaning” (*Many More than a Million seminar report*, CLIR, November 2007: http://www.clir.org/activities/digitalscholar/Nov28final.pdf)

Digitisation challenges (2)

- Current generations of OCR tools do not always provide satisfactory results for historical documents
  - Main focus of tools is on modern documents
  - Not always fit for historic material with archaic fonts, obsolete characters, complex layouts, warped or degraded pages, language variation, etc.
  - Manual post-correction has a role, but is slow and expensive
- Example of OCR errors
  - From Australian Newspapers (National Library of Australia): http://newspapers.nla.gov.au/
  - "The text in the left panel has been electronically translated by a computer. Computers are not as good at reading as humans, and often make mistakes“
  - This system permits users to correct the OCR output
  - Article by Rose Holley in *D-Lib Magazine*, March/April 2009: http://www.dlib.org/dlib/march09/holley/03holley.html



The Canberra Times (ACT : 1926-1954) (about) Wednesday 27 October 1937 Page 3 of 4

Cite: http://nla.gov.au/nla.news-article2443875

Tags (Keywords)

Comments
No comments yet.

ELECTRONICALLY TRANSLATED TEXT

Why may this text have mistakes?
How to correct this text?

No corrections yet

['<PNPWN7$PLDI£RS 1M¿

J1-tf,t>. VI

5

i, australians Wfio Foil at Ad
]<? :' Pozierèà ' ,',' lô> -M,AtÍ

,Tho Defence Department announces
that enquiries have been instituted
by the Imperial War Graves Com-
mission with a view to ascertaining

UNKNOWN SOLDIERS

Australians Who Fell at Pozieres

The Defence Department announces that enquiries have been instituted by the Imperial War Graves Commission with a view to ascertaining, if possible, the identity of two Austraian soldiers whose bodies have been exhumed from a spot approximately 500 yards north-east of the village of Pozieres (Somme), France.

In one case a 9 ct. gold ring, engraved "T.R. to A.R." was found in the deceased's pocket, whilst the remains of the other were wrapped in a waterproof sheet upon which the following particulars can be traced:

"4540, A.F. . . .nn
D. Company, 16 Platoon."

Any person who may be able to assist in the identification of these two soldiers is asked to communicate with the Officer-in-Charge, Base Re-

The Trump.
Add 2—
Balkan Prince, S Mac.
Add 2—
Allunga, Willie Win.
Add 5—
Silver Standard, Wotan.
Add 6—
Sarcherie, Flood Tide.
Add 8—
Donaster, Vaalmore, Gay Knight.
Add 17—
Frill Prince, Mestoravon, Nightguard.
Add 16—
Red Ray, Royal Step.
Add 34—
Old Rowley, and others.

SYDNEY WOOL SALES

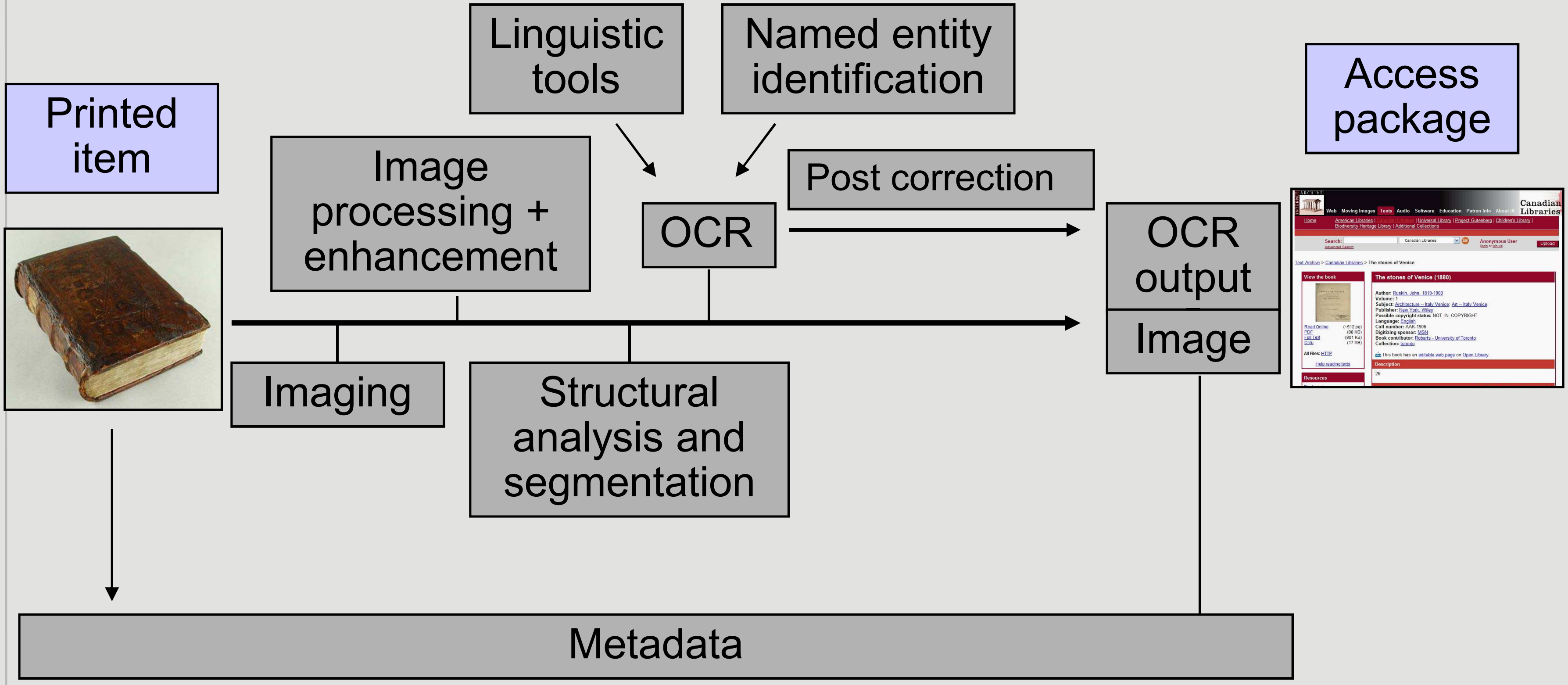
SYDNEY, Tuesday.

Compared with yesterday's sales, the wool market to-day was unchanged and values of wool improved. There was good competition from Yorkshire with good support from the Continent and moderate inquiry from

Greasy me

IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands.

Extremely simplified text digitisation workflow


Linguistic tools
Named entity identification
Access package
Printed item
Image processing + enhancement
Post correction
OCR
OCR output
Image
Imaging
Structural analysis and segmentation
Metadata

INTERNET ARCHIVE
Web | Moving Images | Texts | Audio | Software | Education | Patron Info | About IA
Canadian Libraries
Home
American Libraries | Canadian Libraries | Universal Library | Project Gutenberg | Children's Library | Biodiversity Heritage Library | Additional Collections
Search: Canadian Libraries GO!
Advanced Search
Anonymous User (login or join us)
Upload

Text Archive > Canadian Libraries > The stones of Venice

View the book

Read Online	(~512 pg)
PDF	(88 MB)
Full Text	(901 KB)
DjVu	(17 MB)

All Files: HTTP

Help reading texts

Resources

The stones of Venice (1880)

Author: Ruskin, John, 1819-1900
Volume: 1
Subject: Architecture -- Italy Venice; Art -- Italy Venice
Publisher: New York, Wiley
Possible copyright status: NOT_IN_COPYRIGHT
Language: English
Call number: AAK-1906
Digitizing sponsor: MSN
Book contributor: Robarts - University of Toronto
Collection: toronto

This book has an editable web page on Open Library.

Description

26

The IMPACT project

- Research project funded by the European Commission
  - Large-scale Integrating Project
  - Funded from January 2008, for four years
  - Coordinated by the National Library of the Netherlands (KB)
  - Total budget: EUR 15.5M; EU funding: EUR 11.5M
  - Consortium of 15 partners
    - Libraries
    - Universities and research centres
    - Industrial partners

The IMPACT consortium

- Libraries
  - National Library of the Netherlands (coordinator)
  - The British Library
  - Bibliothèque nationale de France
  - German National Library
  - Bavarian State Library
  - Goettingen State and University Library
  - Austrian National Library
  - University of Innsbruck Library

- Universities and research centres
  - Dutch Institute for Lexicology
  - National Centre for Scientific Research - Demokritos
  - University of Salford
  - University of Munich
  - University of Innsbruck
  - University of Bath (UKOLN)

- Industrial partners
  - ABBYY
  - IBM Haifa Research Lab

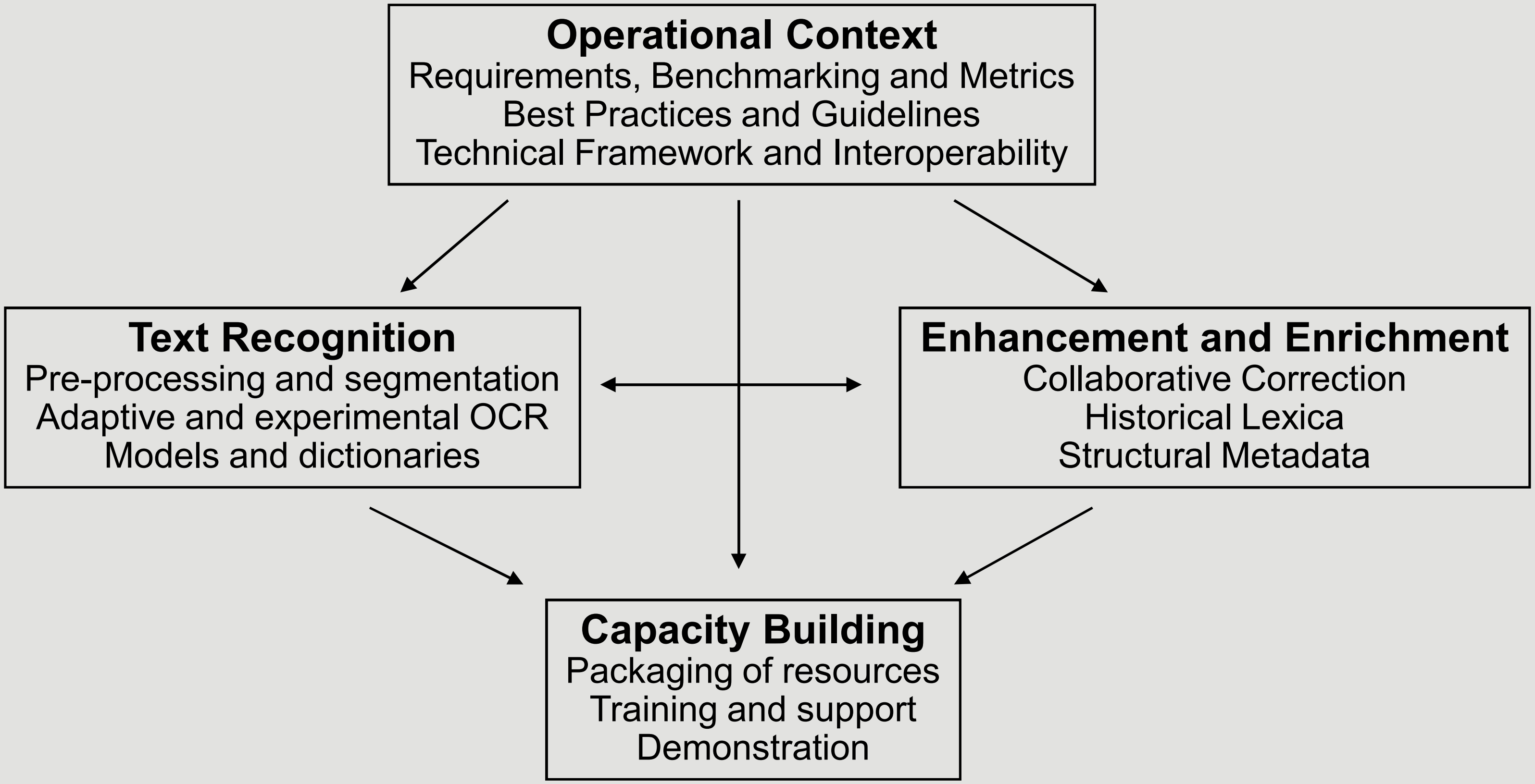
IMPACT project objectives

- Aims to significantly improve the mass digitisation of historical printed text by
  - Innovating OCR software and language technology
  - Sharing expertise and building capacity across Europe
  - Ensuring that tools and services will be sustained after the end of the project
- Specific principles:
  - Reduce effort and enhance speed and results of mass digitisation (speed and scalability)
  - Focus on the whole post scanning workflow: image processing, OCR processing (including dictionaries), OCR correction, and document formatting
  - All research and development to be grounded in the needs of libraries
  - Working with other centres of competence

IMPACT project approach (1)

- Project structure
  - 22 work packages
- Four sub-projects
  - Technical and research based:
    - TR (Text Recognition) focused on the extraction of text in a digital form from an image (OCR)
    - EE (Enhancement and Enrichment) using linguistic technologies to make the results of full-text digitisation more accurate and accessible
  - Strategic:
    - OC (Operational Context) guiding the direction of the project from the libraries' perspective
    - CB (Capacity Building) stimulating the uptake of results in the museums, libraries and archives communities

IMPACT project approach (2)



Operational Context
Requirements, Benchmarking and Metrics
Best Practices and Guidelines
Technical Framework and Interoperability

Text Recognition
Pre-processing and segmentation
Adaptive and experimental OCR
Models and dictionaries

Enhancement and Enrichment
Collaborative Correction
Historical Lexica
Structural Metadata

Capacity Building
Packaging of resources
Training and support
Demonstration

IMPACT tools and services (1)

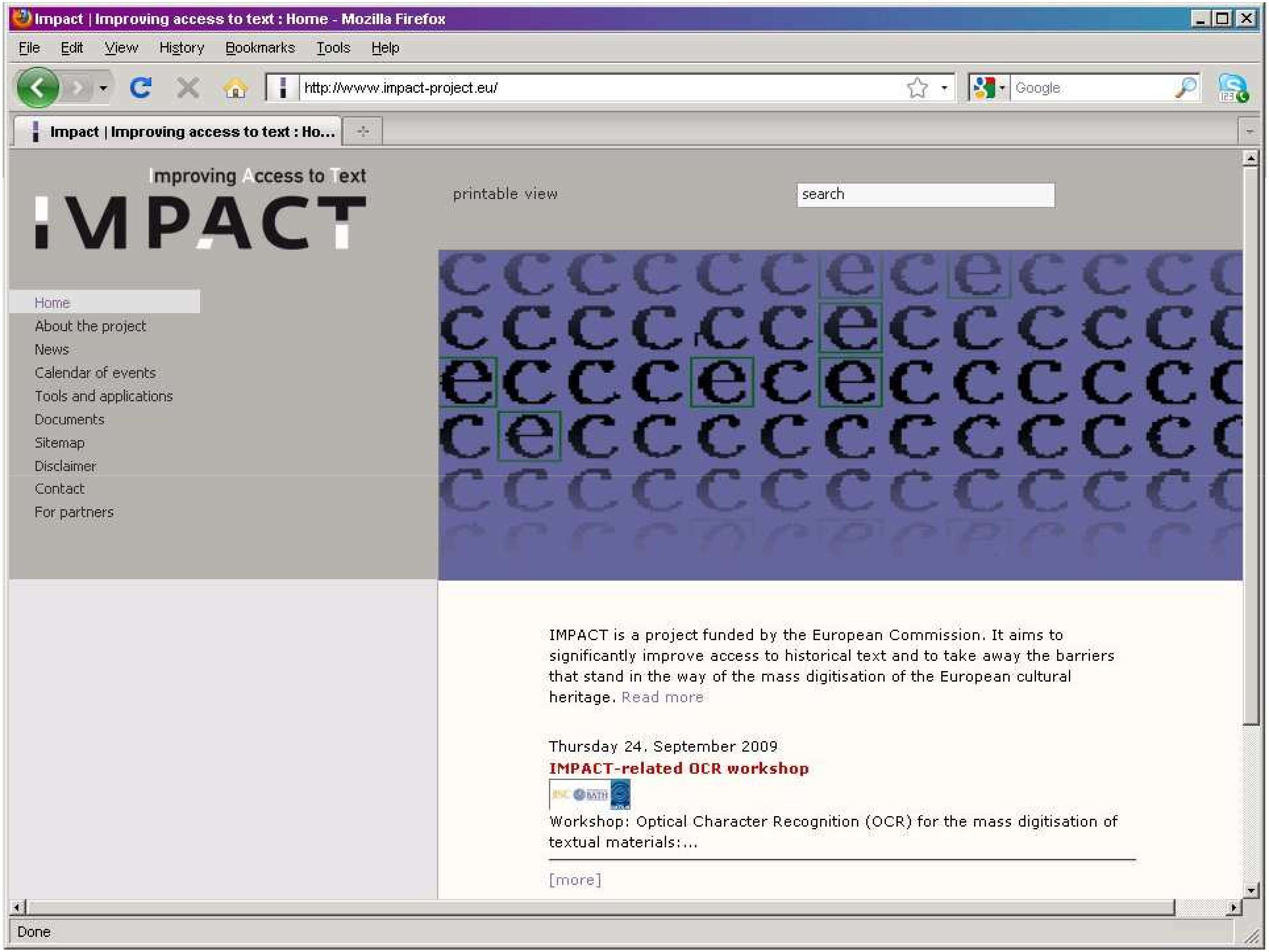
- Text Recognition
  - Technologies for supporting the extraction of text from the page
  - Adaptive OCR engine, integrating:
    - Image enhancement toolkit
    - Segmentation toolkit
    - Post-correction modules
    - Other OCR engines
  - Experimental prototypes
    - Typewritten OCR
    - Wordspotting
    - Inventory extraction

IMPACT tools and services (2)

- Enhancement and enrichment
  - Focus on making OCR results more accurate and accessible
  - Collaborative correction
    - Web based, linked to OCR engine
  - Tools and content
    - General and named entities lexica for Dutch, German and English, general support for lexicon building in other languages
    - Dealing with historical languages
    - Collaborative environments for managing named entities
  - Structural metadata
    - Functional Extension Parser, for the automatic detection and tagging of structural metadata of scanned material

IMPACT tools and services (3)

- Strategic tools and services
  - Website (http://www.impact-project.eu/)
  - Decision support tools, to support the initiation, organisation, management of mass-digitisation projects
  - A set of learning resources providing guidance on the digitisation of texts and the implementation of project tools
  - Training and support
    - Helpdesk
    - Training programme (events)
  - Demonstration of the tools (case studies)
  - IMPACT Centre of Competence



IMPACT is a project funded by the European Commission. It aims to significantly improve access to historical text and to take away the barriers that stand in the way of the mass digitisation of the European cultural heritage. Read more

Thursday 24. September 2009

IMPACT-related OCR workshop

Workshop: Optical Character Recognition (OCR) for the mass digitisation of textual materials:...

[more]

Thank you for your attention!

- Any questions?

- Additional information:
  - The IMPACT project:
    - Website: http://www.impact-project.eu/
    - Project office: impact@kb.nl
  - Europeana: http://www.europeana.eu/portal/

  - Workshop Materials: http://www.ukoln.ac.uk/events/ocr-2009/