ENRICH

European Networking Resources and Information concerning Cultural Heritage

or

Towards a European Digital Library of Manuscripts

ENRICH PROJECT

Basic information

- eContent+ targeted project
- 11/2007 – 11/2009
- 18 full project partners
- 39 associated partners

Main goals

- bring together **substantial part of digitized manuscripts in Europe and make them available through single interface**
- ensure a full interoperability between existing digital libraries, viewing tools and databases **without a need to reallocate source files**
- bring together different types of cultural organizations such as libraries, archives, museums and private collections
- Establish the foundation of **European Digital Library of Manuscripts**

Partners

The workpackage leaders are:

- National Library of the Czech Republic, Prague
- AiP Beroun, s r.o., Beroun, Czech Republic
- Oxford University Computing Services, Oxford, United Kingdom
- Centro per la comunicazione e l'integrazione dei media, Florence, Italy
- SYSTRAN S.A., Paris, France
- Institute of mathematics and informatics, Vilnius, Lithuania
- Biblioteca Nacional de España, Madrid, Spain

Partners

The other partners are:

- Cross Czech, a.s., Prague, Czech Republic
- Københavns Universitet - Nordisk Foskningsinstitut, Copenhagen, Denmark
- Biblioteca Nazionale Centrale di Firenze, Florence, Italy
- University Library Vilnius, Vilnius, Lithuania
- University Library Wroclaw, Wroclaw, Polands
- Stofnun Árna Magnússonar í íslenskum fræðum, Reykjavík, Iceland
- Computer Science for the Humanities - Universität zu Köln, Cologne, Germany
- St. Pölten Diocese Archive, St. Pölten, Austria
- The National and University Library of Iceland, Reykjavík, Iceland
- The Budapest University of Technology and Economics, Budapest, Hungary
- Poznań Supercomputing and Networking Center, Poznań, Poland

ENRICH end-users

- Content owners/holders
- Libraries, museums, archives.
- Researchers & Students
- Policy makers
- General interest users

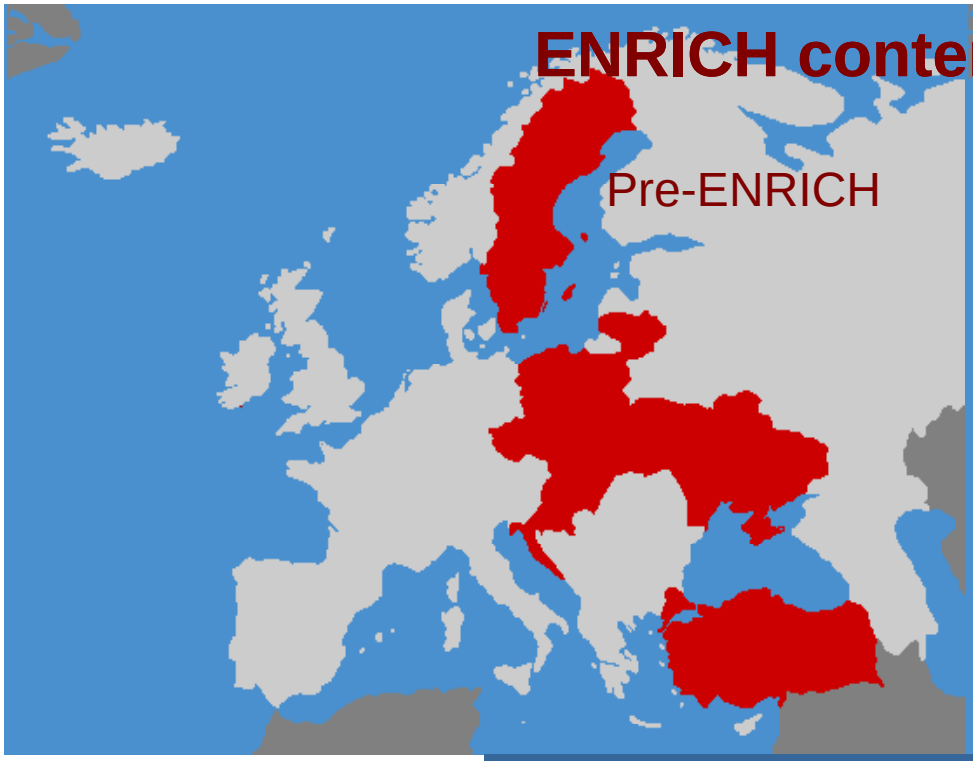
ENRICH introduction: the platform

- Manuscriptorium system
  - online system dedicated to presentation tasks in the area of historical funds
    - catalogue of descriptive records
    - digital library
    - fulltext transcriptions/editions
    - number of end-user tools and features
  - aggregating contents since 2003
    - partners from Czech Republic, Slovakia, Poland, Lithuania, Hungary, Austria, Germany, Romania, Moldova, Iceland, Denmark, Italy, Spain, Turkey...

Aggregated digital content

- ENRICH partners hold around 85% of digitized manuscripts owned by NLs in Europe – these are (together with number of other manuscripts) now available through a single access point: http://www.manuscriptorium.com

ENRICH content development



Aggregated digital content

- Manuscriptorium now provides access to content aggregated from:
  - 12 ENRICH content partners
  - 13 associated content partners

Regular partners

Name of the library	City/Country	Number of image files	Nr. of compound digital documents
Biblioteca Nazionale Centrale di Firenze	Florence, Italy	214 882	1 588
Národní knihovna ČR	Prague, Czech Republic	1 061 782	3 460
Institute of mathematics and informatics	Vilnius, Lithuania	1 223	56
Københavns Universitet, Nordisk Foskningsinstitut	Copenhagen, Denmark	1 572	16
University Library Wroclaw	Wroclaw, Poland	78 404	989
University Library Vilnius	Vilnius, Lithuania	2 510	1 085
Stofnun Árna Magnússonar í íslenskum fræðum	Reykjavík, Iceland	25 607	238
St. Pölten Diocese Archive	St. Pölten, Austria	93 593	51 197
The National and University Library of Iceland	Reykjavík, Iceland	140 059	570
Biblioteca Nacional de Espana	Madrid, Spain	367 997	2 312
The Budapest University of Technology and Economics	Budapest, Hungary	7 903	45
Universität zu Köln	Cologne, Germany	643 434	1 741

Associated partners

Name of the library	City/Country	Number of image files	Nr. of compound digital documents
Heidelberg University Library	Heidelberg, Germany	289 505	882
Holy Trinity-St. Sergius Lavra	Sergiev Posad, Russia	535 026	2 845
National Library of Romania	Bucharest, Romania	72 260	109
The National Library of Belarus	Minsk, Belarus	25 425	51
University Library in Bratislava	Bratislava, Slovakia	49 125	142
Národní pedagogická knihovna komenského	Prague,. Czech Republic	5 590	30
Lund University Library	Lund, Sweden	22 443	69
Knihovna Národního Muzea	Prague, Czech Republic	9 929	3
Moravian Library	Brno, Czech Republic	75 556	173
Toruń University Library	Toruń, Poland	16 641	346
Zielona Góra University Library	Zielona Góra, Poland	24 963	203
Patrimonio Nacional	Palacio Real, Madrid, Spain	18 243	76
Complutense University of Madrid	Madrid, Spain	728 555	1 656

Total aggregated content

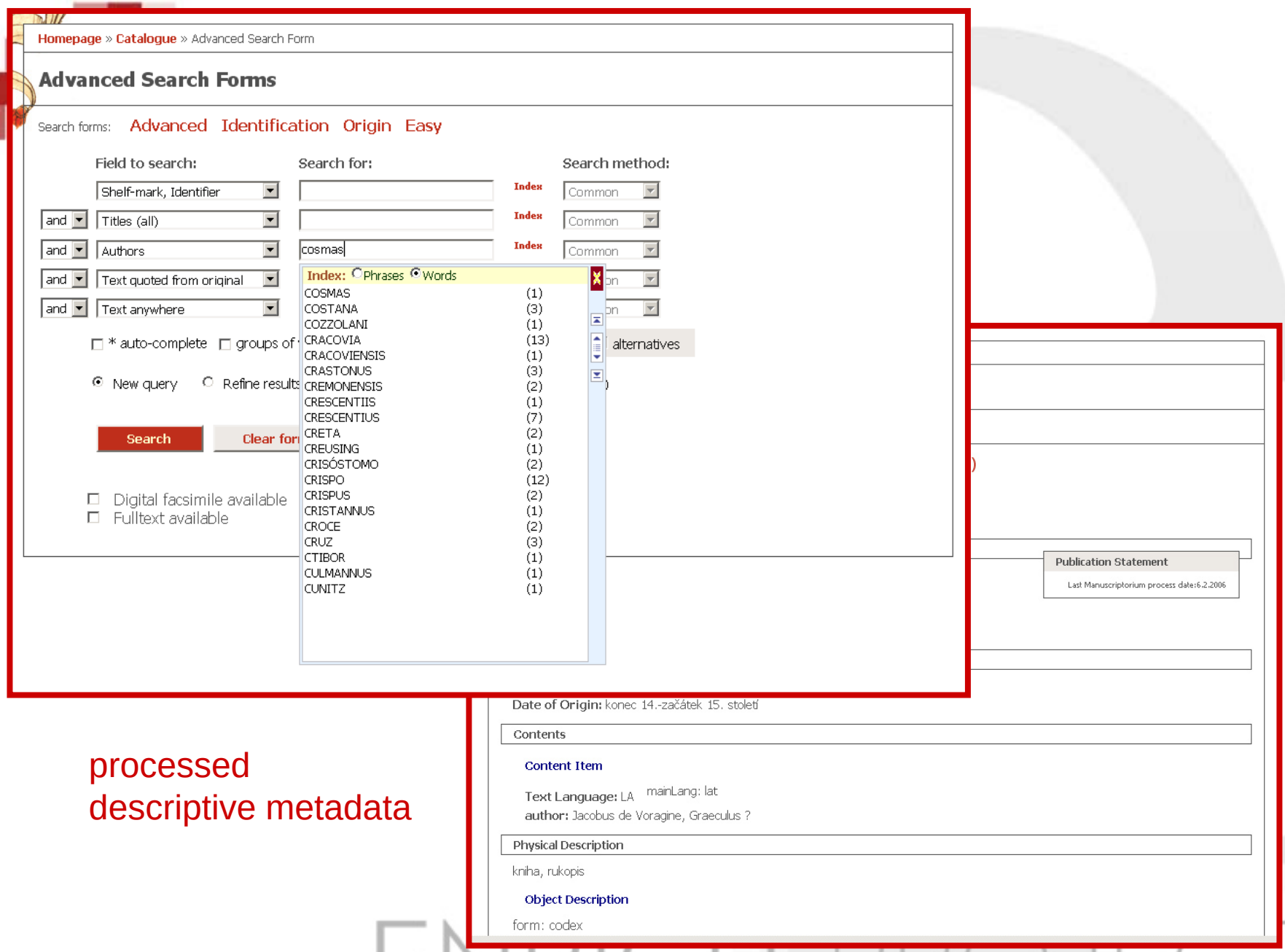
	Number of image files	Nr. of compound digital documents
ENRICH regular partners total	**2 638 966**	**63 245**
ENRICH associated partners total	**1 873 261**	**6 585**
TOTAL	**4 512 227**	**69 830**

Principles of cooperation

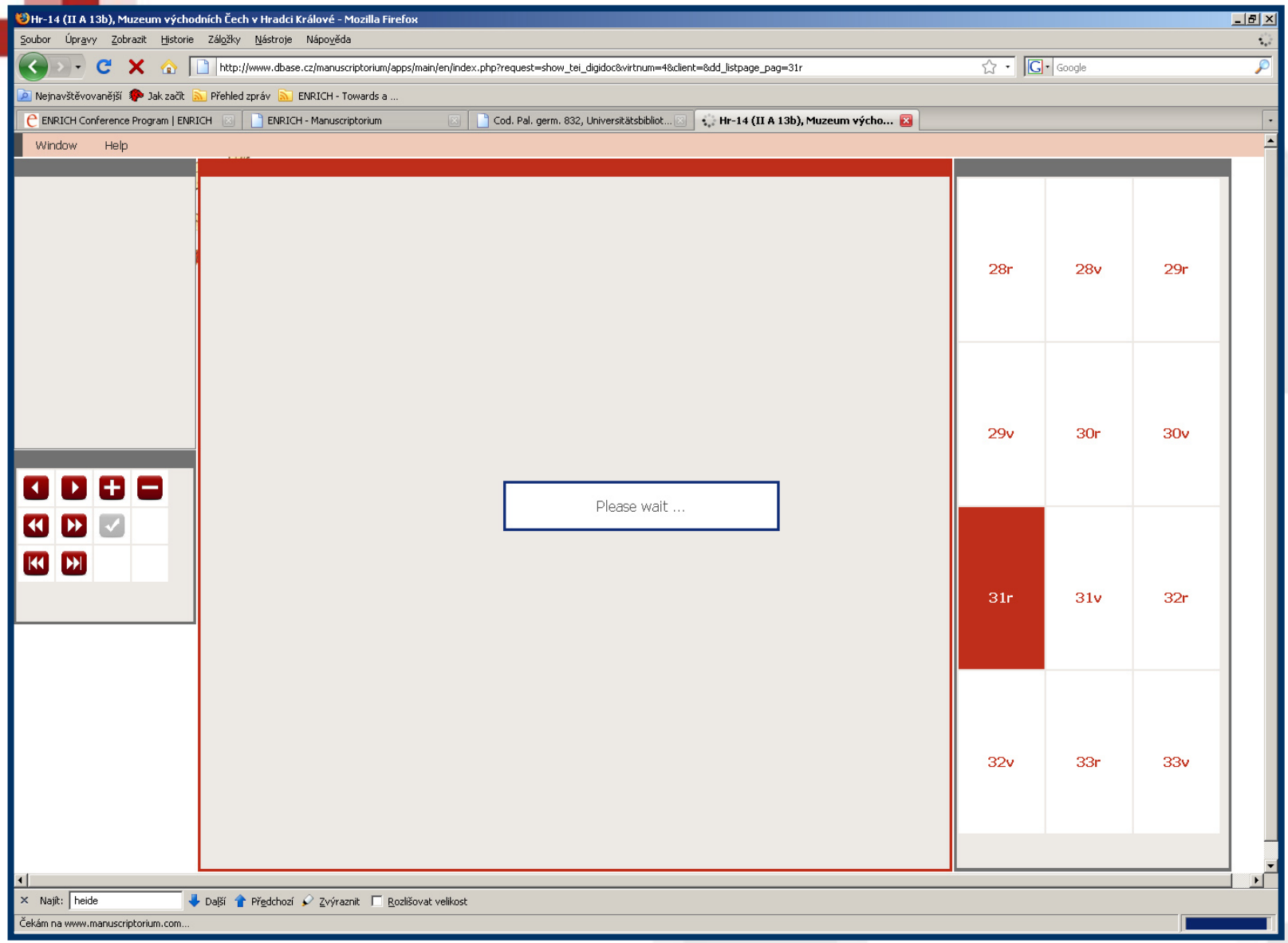
Distributed data storage

- images are managed by its owner/managing institution
- only metadata are imported into the Manuscriptorium system
- final document is completed within end-users browser

- Why:
  - simple technical solution (compared to centralized data storage systems)
  - mutual independency
  - no need to give partner's images to a third party (Manuscriptorium)



processed
descriptive metadata

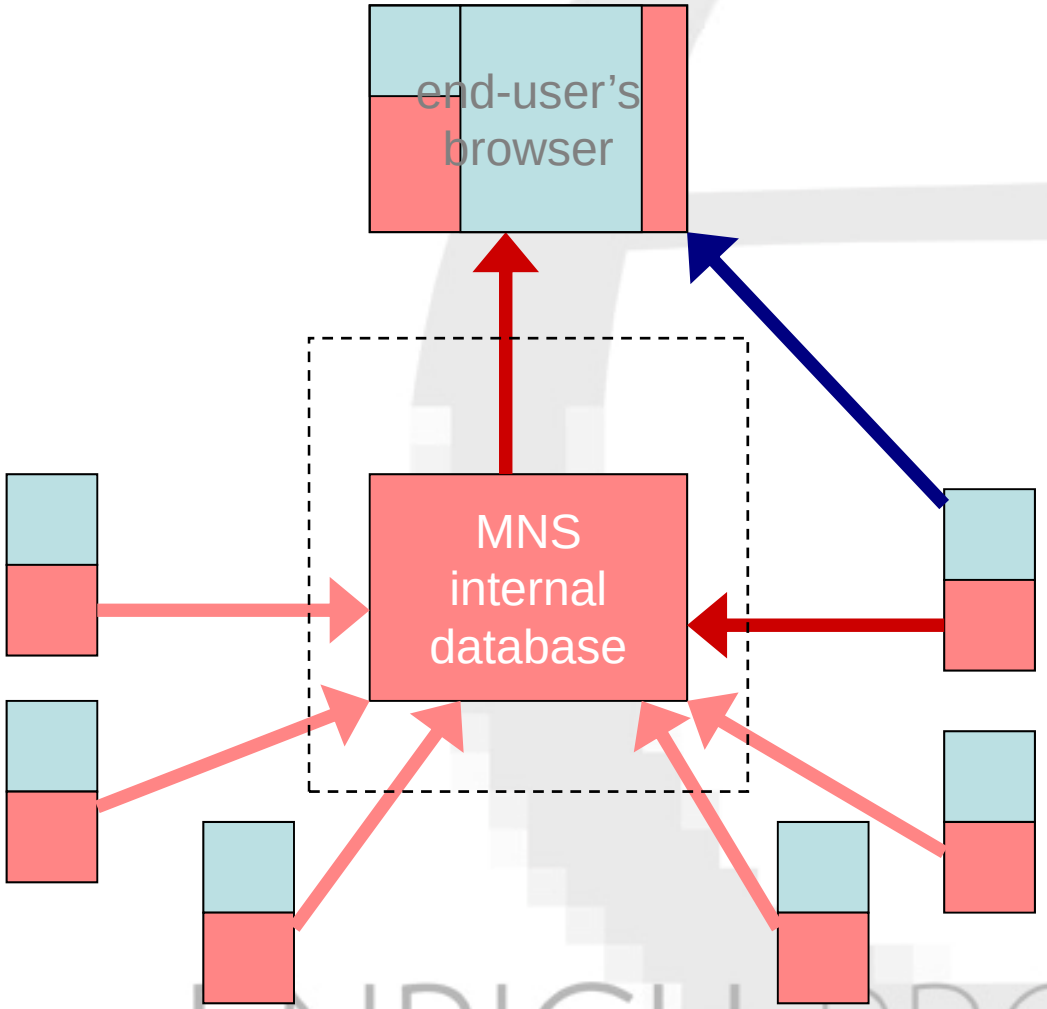


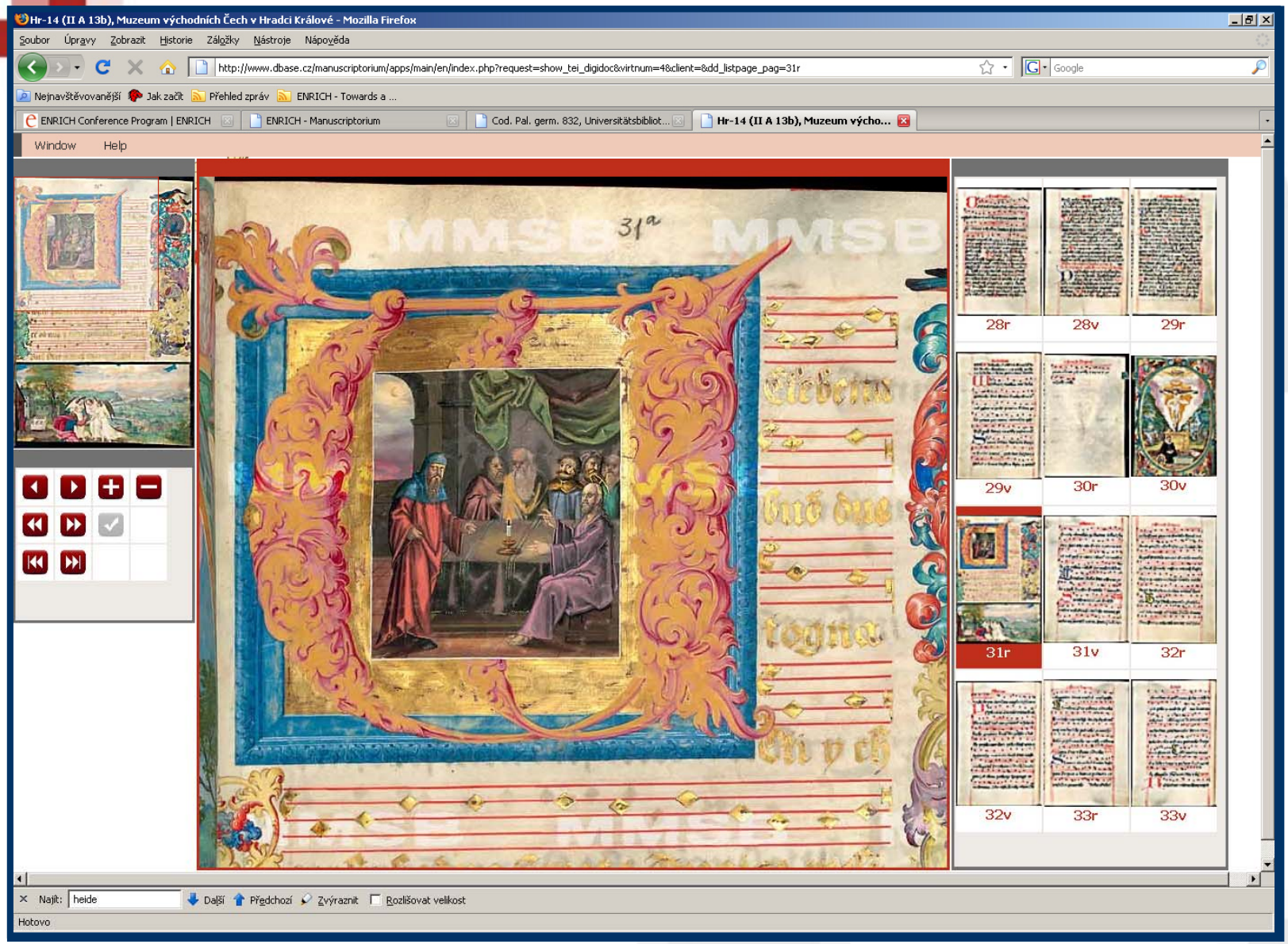
processed
structural metadata

ENRICH PROJECT

Part I: general principles of cooperation

Treatment of data





processed

structural metadata – transfer of images completed

Principles of cooperation

No particular metadata format is required

- TEI P5 is recommended
- Other formats supported
  - MASTER, MARC 21,UNIMARC, MODS ... METS ... generaly any consistently structured metadata (even local and proprietary formats)

– Why:
  - at present many various aproaches to digitisation
  - respect to the various individual conditions within individual institutions
  - no need to interferre with processes within the partner institutions

Standardization of shared metadata

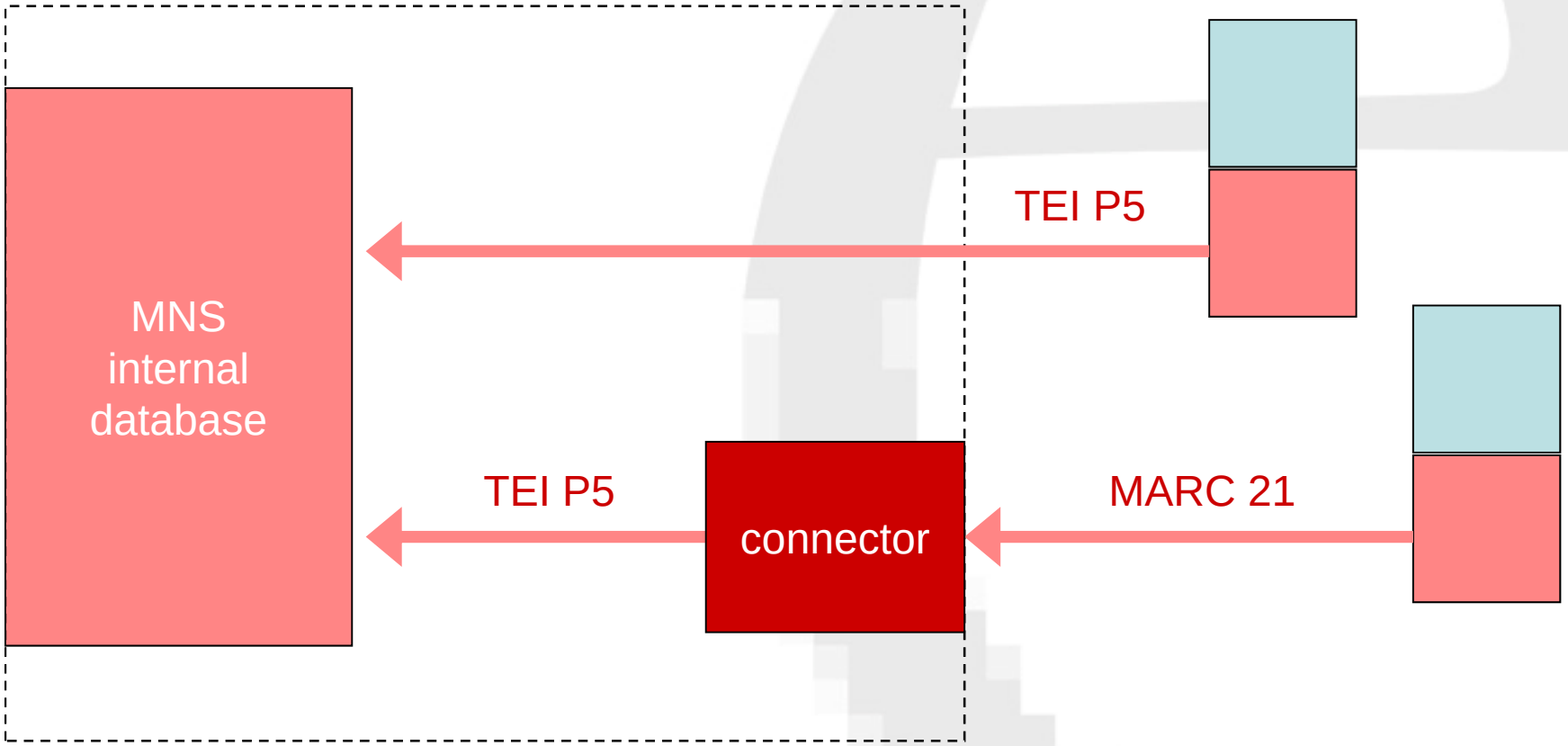
- TEI P5 ENRICH schema was developed by OUCS and KU
- Currenlty used as metadata standard format by several ENRICH partners and within Manuscriptorium, officially recommended by TEI P5 as a format for description of Manuscripts
- Specifications can be found at http://enrich.manuscriptorium.com/index.php?q=node/50

Standardization of shared metadata

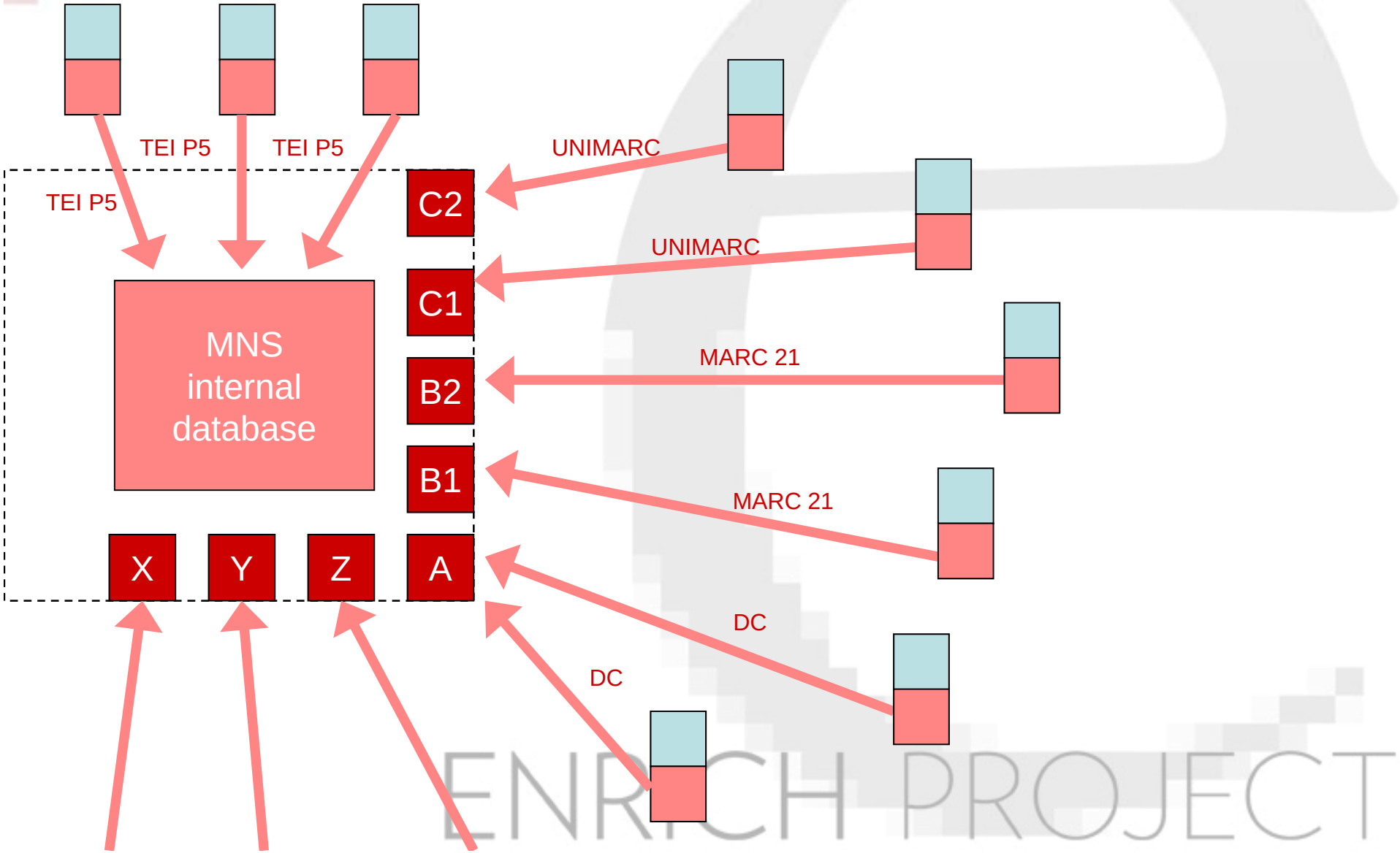
ENRICH GARAGE

- New online tool, based on Java and libraries such as Jing, onvdl, Saxon 9, allowing online conversions of metadata from MASTER, TEI P5 and EAD (and possibly some other) formats to ENRICH TEI P5

Principles of cooperation



Principles of cooperation



Cooperation requirements (technical minimum)

I. Both data (images) and metadata should exist
   - if metadata were not available they should be created for each aggregated document

II. Data (images) files should be accessible via HTTP protocol

III. Data (images) formats should be supported by common browsers (without the need to use plugins)

IV. Basic extent of information should be covered by the metadata:
   - descriptions: identification of the original
   - structural metadata: a list of images

(ENRICH also will recommend a good practice)

Methods of cooperation

1. Creating individual TEI P5 based documents
   a) manual markup using general XML editor
   b) using dedicated tools (M-Tool)
2. Conversions of existing documents metadata
   a) on-line connectors (OAI-PMH harvesting)
   b) off-line connectors (processing of offline exports of metadata)
3. Automated generating of metadata
   – to avoid manual work where it is possible
4. Combination of the above

1 a) Creating new TEI P5 metadata

- Typicaly used by:
  - those who have data (images) but don't have appropriate metadata (descriptions and/or structural metadata have to be newly created)
  - those who start their digitization projects and decided to use TEI P5
  - those who want to (re)create their documents and use TEI P5 (and where the number of documents is lower)

1 b) M-Tool On-line

- Key features:
  - on-line service accessible from Manuscriptorium homepage
  - enables to create TEI P5 metadata (without manual XML markup)
    - bibliographic description
    - structural metadata
  - flexible generator of structural metadata
  - integrated checking mechanism
    - review of descriptive record represenation
    - automated check of image accessibility
  - manual structure modification
  - customisable forms

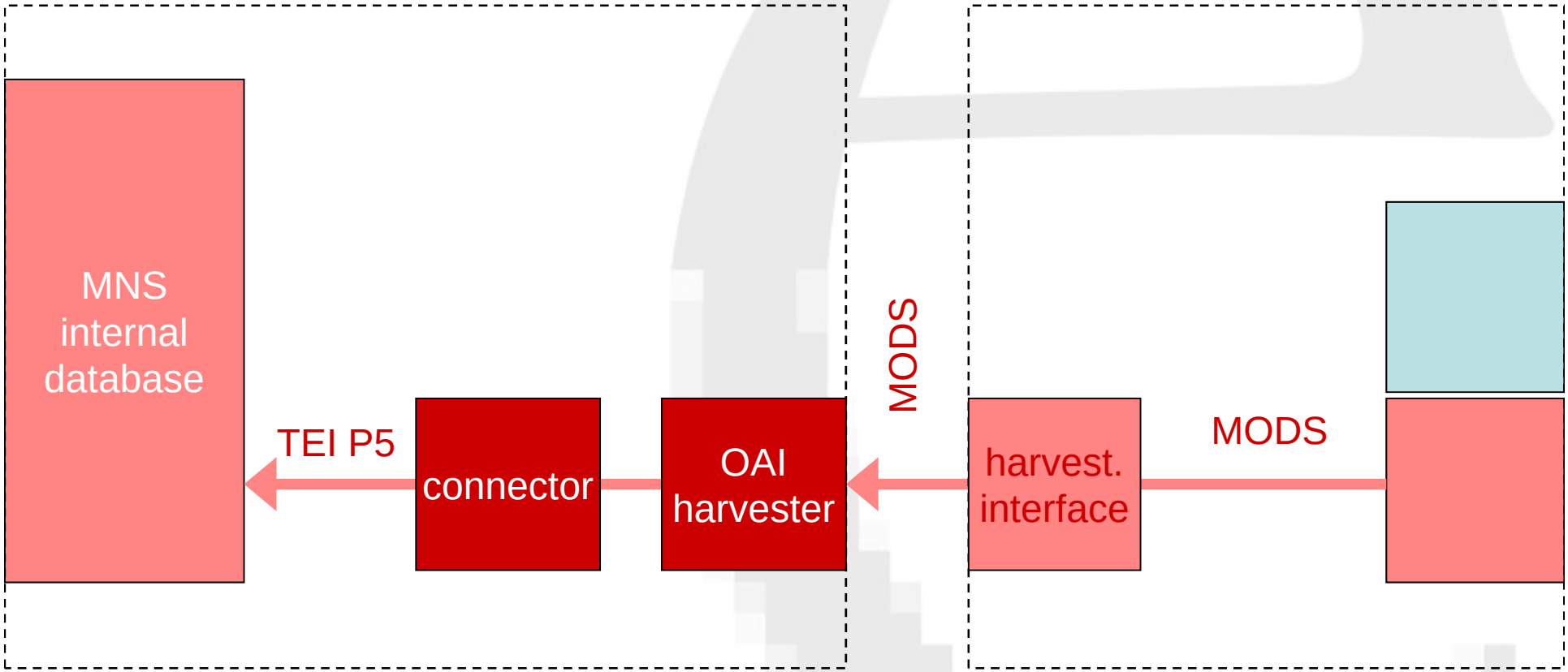
M-Tool On-line

- customisable forms
  - set of common and individualised forms will be available
  - individualised forms will be created for new partners
    - according to properties of documents in partners collection
    - according to partner's cataloguing practice

2 a) Harvesting of documents

- Typicaly used with:
  - OAI-PMH
  - for partners who have their digital library AND who run their harvesting interface
  - partners who provide their metadata in TEI P5 OR partners who contribute relevant amount of documents (to rationalize creation of document)

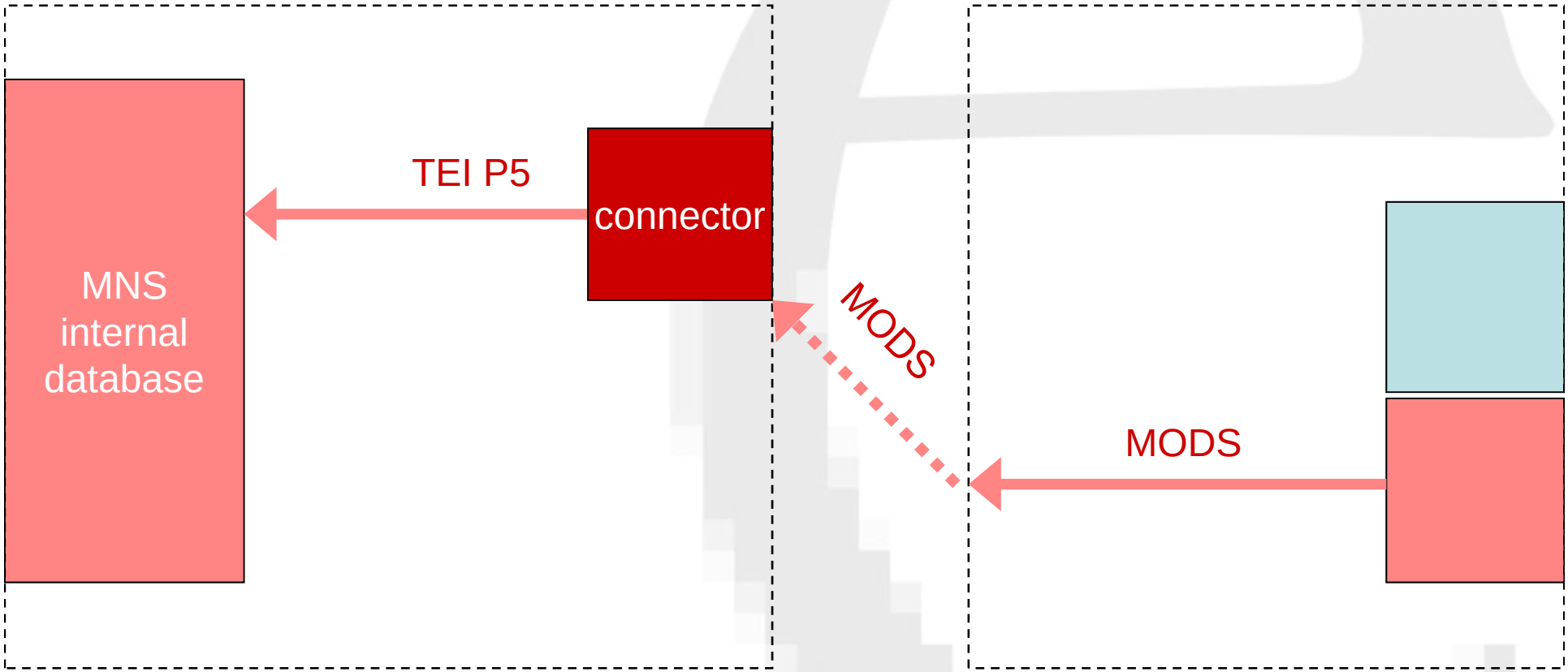
Harvesting of non-TEI metadata



2 b) Off-line export processing

- Typicaly used with:
  - partners who do NOT have harvesting interface
  - partners who provide their metadata in TEI P5 OR partners who contribute relevant amount of document using different formats (to rationalize creation of document)

Processing of non-TEI P5 metadata



3) Generating of structural metadata

- For partners who have large document collection where manual metadata creation would be too time demanding
- Often the basic metadata content carries the image path/filename
  - once the information was created no need to re-create it manualy again
- Tools which examine and record the structure of directory contents
  - output processed and converted to TEI P5

User personalization

- Aiming at individual collections and tools for virtual documents creation
- Two main tools were developed and implemented: **static and dynamic collections**
- Login (free) required for creating and saving collections

Static collection

- simple and most natural way of how to create a smaller set of particular documents. Implementation of a static collection:
  - “Add document into selection“ button is enabled within the user interface.
  - the selection will be available for review and update
  - it will be possible to add the selection to a new or existing collection;
  - basic tools to manage collections and documents within a particular collection are included

Dynamic collection

- Content of the dynamic collection matches with a search query applied above the up-to-date content catalogue.
- The feature allows creating specifically focused thematic collections. Their content will be continuously updated according to the growing content of the source.
- Larger thematic collections can be easily maintained.
- Creation according to either **one simple query** or even to a **sequence of more subsequent queries**
- enable to create more precisely focused collections

Multilingual access

- Translation Stylesheet based on TEIP P5 specifications was implemented into platform
- optimization of historical funds specific dictionary carried out
- VICODI ontologies used for improvement of translations
- multilingual querying - a quick fulltext search above the multilingual contents on a basis of query string translation
- Czech, Danish, Dutch, English, French, German, Greek, Italian, Polish, Portuguese, Serbo-Croatian, Slovak, Spanish and Swedish languages included

Multilingual access

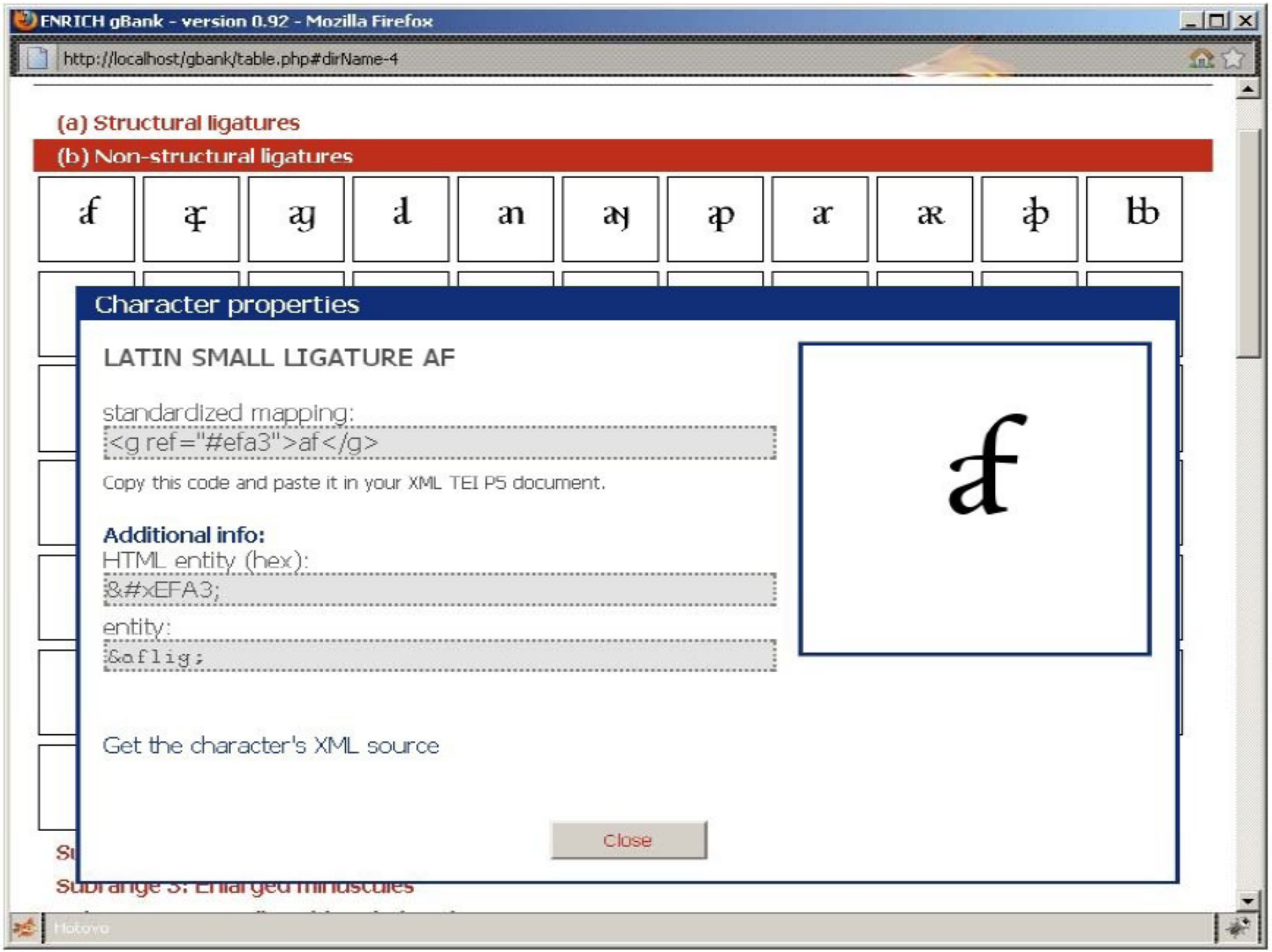
Works done

- multilingual access and searching tools implemented
- XSL based Systran Translation Stylesheet created
- dictionaries, translation memories, normalization dictionaries managed and improved directly by the partners via dedicated interface

- All driven by SYSTRAN machine translation technologies

Gaiji Bank

- Application designed for the improvement of UNICODE treatment within the platform (but also for external users)
- The user friendly interface available at Manuscriptorium homepage displays characters ordered into sets, which can be further searched in order to find individual characters and their equivalents
- valid XML code is displayed - this code can be copied and pasted directly in the XML metadata of the particular digital document


ENRICH gBank - version 0.92 - Mozilla Firefox
http://localhost/gbank/table.php#dirName-4
(a) Structural ligatures
(b) Non-structural ligatures
Character properties
LATIN SMALL LIGATURE AF
standardized mapping:
<g ref="#efa3">af</g>
Copy this code and paste it in your XML TEI P5 document.
Additional info:
HTML entity (hex):

entity:
&aflig;
Get the character's XML source
Close
Subrange 3: Enlarged minuscules
Hotovo

ENRICH Garage Engine (EGE)

- conversion tool within a generic web-service based application
- allows online transformation between other XML formats and ENRICH TEI P5, and also transformations between Microsoft Word and TEI P5.

EGE plug-ins

- **Recognizer** - recognition of the Internet Media Type (MIME type) of the given input data.
- **Validator** - validation of the input data against given standard
- **Converter** - conversion from XML to Word, conversion from Word to PDF, conversion of the XML from one form to another (e.g. MASTER -> ENRICH TEI P5) or even cleaning the input data (e.g. removing redundant information).

To know more:

- Manuscriptorium Digital Library
  - http://www.manuscriptorium.com
- ENRICH project website:
  - http://enrich.manuscriptorium.com

Contacts

ENRICH and Manuscriptorium coordinator:

Zdeněk Uhlíř, National Library of Czech Republic,
zdenek.uhlir@nkp.cz

Technological issues:

Tomáš Psohlavec, AIP Beroun, tp@aipberoun.cz

Administrative issues:

Jakub Heller, Cross Czech a.s., jakub.heller@crossczech.cz