Large-scale aggregation of digital content from distributed digital libraries in Poland

Adam Dudczak, Agnieszka Lewandowska, Marcin Werla
Poznań Supercomputing and Networking Center
Polish Academy of Sciences
Poznań, Poland
Overall number of digital objects
✓ 280 thousands

Number of active digital libraries:
✓ 19 regional
✓ 22 institutional

+ several other digital libraries in the phase of planning, configuration or initial content uploading

Number of cooperating institutions:
✓ over 200 libraries, museums and archives
• Main aims
  – To facilitate the use of resources from Polish digital libraries
  – To increase the visibility and popularity of resources from Polish digital libraries in the Internet
  – To enable new advanced network services based on the resources from Polish digital libraries to Internet users and digital libraries creators
• PIONIER Digital Libraries Federation
  – Network service harvesting the metadata from all OAI-PMH-enabled digital libraries in Poland
  – Gives access to new services based on harvested metadata
  – Created, developed and maintained by Poznań Supercomputing and Networking Center (PSNC)
  – Publicly available since June 2007
  – http://fbc.pionier.net.pl/
Available publications: 288,421
Planned publications: 4,363

The Database of Polish Digital Libraries at the PIONIER DLF

From today on The Database of Polish Digital Libraries (an abbreviation: DPDL) is available at the PIONIER Network Digital Libraries Federation. Its purpose is to collect and share data on digital libraries in Poland - not just those already existing, but also the ones in the process of formation.
• Basic assumptions
  – Neither need nor requirement to deposit digital objects from digital libraries into the PDLF
  – No fees for using the PDLF or participating in it
  – Open standards as the basis for communication and interoperability for all PDLF features and mechanisms
• Functionality
  – Search in the metadata of available digital objects
    • Simple
    • Advanced
  – Digitisation plans
    • Searching
    • Report
  – Duplicated digitisation detection and prevention
    • Based only on the metadata (title, creator, publication date, ...)
    • Full usage requires compatible digital library software and on-line digitisation plans
  – OAI identifiers resolving (also ISBN etc.)
  – Networked user profile
  – Statistics
  – Add-ons for the promotion of the PDLF and its resources
## Summary of the number of duplicates

Generated on: May 6, 2009

The summary below contains only those digital libraries, which publish any potential duplicates. Currently, the total number of potential duplicates is 237.

Image ??? means that the logo of a digital library is temporarily unavailable.

| Library | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | Sum | % |
|---------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| GREEK  | 155 | 65% |
| KOPC   | 79  | 33% |
| PRPC   | 72  | 30% |
| JBC    | 63  | 27% |
| PBC    | 52  | 22% |
| MBPC   | 21  | 9%  |
| LESC   | 10  | 4%  |
| SBC    | 9   | 4%  |
| MBC    | 7   | 3%  |
| DBC    | 1   | 0%  |
| HBC    | 1   | 0%  |
| LBC    | 1   | 0%  |
| BHP    | 4   | 2%  |
| BIPB   | 3   | 1%  |
| BLCN   | 3   | 1%  |
| BUCN   | 3   | 1%  |
| BUCN   | 3   | 1%  |
| FIBES  | 2   | 1%  |
| SBC    | 2   | 1%  |
| BBC    | 2   | 1%  |
| EPC    | 2   | 1%  |
| BSC    | 2   | 1%  |

2009-10-05
Promotion of Polish Digital Resources

• Basic tools of a Polish Internet user:
  – Web browsers
    • Firefox (46,0%)
    • MSIE (43,7,5 %)
    • Other...
  – Web search engines
    • Google (96,4%)
    • Other...

• And digital libraries?
  – To use resources from Polish digital libraries, the Internet user must... know about their existence...

• How to get with this knowledge to a typical Internet user?
  – It must be visible in tools, which he/she uses:
    • Take care about the visibility and possibly high ranking in Google
      – PageRank: 6 – not too bad 😊
    • Be visible in the web browser?
Total number of visits in the PDLF each month

- **July 2007**: 1206 visits
- **August 2007**: 2583 visits
- **September 2007**: 6403 visits
- **October 2007**: 8219 visits
- **November 2007**: 11893 visits
- **December 2007**: 12350 visits
- **January 2008**: 15100 visits
- **February 2008**: 11006 visits
- **March 2008**: 18176 visits
- **April 2008**: 18412 visits
- **May 2008**: 20837 visits
- **June 2008**: 26251 visits
- **July 2008**: 42635 visits
- **August 2008**: 36767 visits
- **September 2008**: 47200 visits

**Note:** PDLF included in Firefox 2 and 3.
AGGREGATING DIGITAL LIBRARIES IN POLAND... FOR EUROPEANA
• The coordinated work of many institutions during recent years (since 2002) delivered the efficient network of distributed digital libraries

• The metadata of the majority of digital objects from Polish digital libraries is indexed and searchable via
  – search engines like Google
  – several OAI service providers, like OAIster or ScientificCommons
  – services developed in European projects like ENRICH or CACAO

• The next step is to make the PDLF resources accessible via the Europeana interface and usable with (future) Europeana tools
• According to the public draft version of Europeana Outline Functional Specification the role of aggregator is:

1. to collect information about providers and their delivery systems
2. to collect data about content being provided as a surrogate
3. to de-duplicate, disambiguate, clean, enrich the data with meaningful attributes, possibly associate content in collections
4. to verify the accessibility of content
5. to make data ready for Europeana data collection using the OAI-PMH protocol
PIONIER Digital Libraries Federation as a Metadata Aggregator for Europeana

Metadata aggregation service

Digital libraries
- Institutional
- Regional
- National
- Other

Institutions
- Libraries
- Archives
- Museums
- ....
• To collect information about providers and their delivery systems
  – Name and logo of a digital library, its website URL and the address of the OAI-PMH interface for digitized objects and objects planned for digitization
  – Detailed description with list of participating institutions
  – Sample objects
  – Basic statistics
### Polish digital libraries listing

<table>
<thead>
<tr>
<th>No.</th>
<th>Logo</th>
<th>Name</th>
<th>State</th>
<th>ID</th>
<th>ID</th>
<th>EID</th>
<th>EID</th>
<th>EID</th>
<th>EID</th>
<th>EID</th>
<th>EID</th>
<th>EID</th>
<th>EID</th>
<th>EID</th>
<th>EID</th>
<th>EID</th>
<th>EID</th>
<th>EID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>📚</td>
<td>Wielkopolska Biblioteka Cyfrowa</td>
<td>Online</td>
<td>89,889</td>
<td>568</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>🌐</td>
<td>Polska Biblioteka Internetowa</td>
<td>Online</td>
<td>32,071</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>🍁</td>
<td>Kujawsko-Pomorska Biblioteka Cyfrowa</td>
<td>Online</td>
<td>28,710</td>
<td>205</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>🇵🇱</td>
<td>Cyfrowa Biblioteka Narodowa</td>
<td>Online</td>
<td>20,620</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>🇵🇱</td>
<td>Biblioteka Cyfrowa Uniwersytetu Wrocławskiego</td>
<td>Online</td>
<td>20,528</td>
<td>1,408</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>🌹</td>
<td>Małopolska Biblioteka Cyfrowa</td>
<td>Online</td>
<td>18,451</td>
<td>19</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>🌹</td>
<td>Śląska Biblioteka Cyfrowa</td>
<td>Online</td>
<td>13,474</td>
<td>726</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>🌹</td>
<td>Podlaska Biblioteka Cyfrowa</td>
<td>Online</td>
<td>7,419</td>
<td>21</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>🌻</td>
<td>Świętokrzyska Biblioteka Cyfrowa</td>
<td>Online</td>
<td>5,841</td>
<td>31</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>🌻</td>
<td>Zachodniopomorska Biblioteka Cyfrowa &quot;Pomerania&quot;</td>
<td>Online</td>
<td>9,474</td>
<td>8</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>🌻</td>
<td>Zielonogórska Biblioteka Cyfrowa</td>
<td>Online</td>
<td>5,335</td>
<td>568</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note:** The table contains information about various digital libraries in Poland, including their names, states, and identification numbers.
# Digital libraries description

<table>
<thead>
<tr>
<th>Chinese digital library:</th>
<th>Kujawsko-Pomorska Biblioteka Cyfrowa (KPBC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>General information</td>
<td></td>
</tr>
<tr>
<td>State:</td>
<td>On-line</td>
</tr>
<tr>
<td>Type:</td>
<td>Regional</td>
</tr>
<tr>
<td>Web page:</td>
<td><a href="http://kpbcd.umk.pl/">http://kpbcd.umk.pl/</a></td>
</tr>
<tr>
<td>Contact email:</td>
<td><a href="mailto:kpbc@umk.pl">kpbc@umk.pl</a></td>
</tr>
<tr>
<td>Start date:</td>
<td>Sep 1, 2005</td>
</tr>
</tbody>
</table>

Biblioteka Główna Uniwersytetu Mikołaja Kopernika w Toruniu
(coordinating)
http://www.bu.umk.pl/

Dr Miroslaw A. Supruniuk
Dyrektor

Biblioteka Główna Uniwersytetu Kazimierza Wielkiego w Bydgoszczy
(cooperating)
http://biblioteka.ukw.edu.pl/

Dr Aldona Chlewicka
Dyrektor

lic. Ewa Wójcik
Referent
Kujawsko-Pomorska Biblioteka Cyfrowa recommend:

**Apokalypse - Heinrich von Hesler**


**Ryciny Erika Dahlberga z dzieła Samuela Pufendorfa**

Ryciny Erika Dahlberga z dzieła Samuela Pufendorfa "De rebus a Carolo Gustavo Sueciae Rege ...", pochodzącego ze zbioru Biblioteki Uniwersytetu Kazimierza Wielkiego w Bydgoszczy.

**Stemmata genealogica praecipuarum in Prussia Familiarum Nobilium - Hennenberger, Johann**

Rękopis z końca XVI w.
To collect data about content being provided as a surrogate

- Done with the OAI-PMH protocol
  - Strict compliance with the protocol specification is required
- At this moment the metadata is harvested only in Dublin Core
  - Extension to ESE planned for the nearest future
- In some extraordinary cases the additional work is required
  - The Polish Internet Library (http://www.pbi.edu.pl/)
• To **de-duplicate**, disambiguate, clean, enrich the data with meaningful attributes, possibly associate content in collections
  – Makes sense only in the context of libraries
    • In museums and archives each object is unique
  – Based on the comparison of metadata
    • Small differences in metadata considered
  – 0.2% of aggregated objects on the list of potential duplicates
  – Duplication prevention on the stage of digitisation is important
• To de-duplicate, **disambiguate, clean**, enrich the data with meaningful attributes, possibly associate content in collections
  
  - From the contents of aggregated metadata the PDLF builds vocabularies
    - Separately for each DC element
    - Separately for each language of description
  
  - The differences in metadata from different digital libraries significantly influence the searching possibilities
  
  - It is crucial to clean and disambiguate the metadata both for internal use on the level of aggregation and for external use in Europeana
<table>
<thead>
<tr>
<th>DC element</th>
<th>No. of unique values</th>
<th>Number of associations</th>
<th>Average no. of occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>format</td>
<td>39</td>
<td>209 789</td>
<td>5 379,2</td>
</tr>
<tr>
<td>language</td>
<td>195</td>
<td>210 529</td>
<td>1 079,6</td>
</tr>
<tr>
<td>type</td>
<td>822</td>
<td>211 816</td>
<td>257,7</td>
</tr>
<tr>
<td>rights</td>
<td>1 192</td>
<td>246 093</td>
<td>206,5</td>
</tr>
<tr>
<td>coverage</td>
<td>66</td>
<td>2 390</td>
<td>36,2</td>
</tr>
<tr>
<td>publisher</td>
<td>18 002</td>
<td>310 764</td>
<td>17,3</td>
</tr>
<tr>
<td>contributor</td>
<td>12 979</td>
<td>83 464</td>
<td>6,4</td>
</tr>
<tr>
<td>subject</td>
<td>78 440</td>
<td>438 871</td>
<td>5,6</td>
</tr>
<tr>
<td>relation</td>
<td>9 292</td>
<td>48 319</td>
<td>5,2</td>
</tr>
<tr>
<td>date</td>
<td>47 581</td>
<td>209 589</td>
<td>4,4</td>
</tr>
<tr>
<td>identifier</td>
<td>6 426</td>
<td>27 666</td>
<td>4,3</td>
</tr>
<tr>
<td>description</td>
<td>43 657</td>
<td>180 391</td>
<td>4,1</td>
</tr>
<tr>
<td>source</td>
<td>16 996</td>
<td>52 506</td>
<td>3,1</td>
</tr>
<tr>
<td>creator</td>
<td>21 908</td>
<td>67 503</td>
<td>3,1</td>
</tr>
<tr>
<td>title</td>
<td>210 745</td>
<td>227 039</td>
<td>1,1</td>
</tr>
</tbody>
</table>
• Format
  – In 99% of descriptions: MIME type (e.g. text/html, image/x.djvu)

• Language
  – In most cases: ISO 639-2 (pol, ger, lat, fre etc.)
  – Sometimes one value „pol, ger” instead of „pol”, „ger”

• Rights
  – Name of the institution which holds the original object

• Type
  – ...
<table>
<thead>
<tr>
<th>Values for „Type” (top 20)</th>
<th>Number of objects with the value</th>
<th>% of aggregated objects</th>
<th>% of aggr. obj. (after clean-up)</th>
</tr>
</thead>
<tbody>
<tr>
<td>czasopismo</td>
<td>44 709</td>
<td>20,9%</td>
<td>33,8%</td>
</tr>
<tr>
<td>gazeta</td>
<td>32 921</td>
<td>15,4%</td>
<td>31,3%</td>
</tr>
<tr>
<td>gazety</td>
<td>23 119</td>
<td>10,8%</td>
<td></td>
</tr>
<tr>
<td>Czasopismo</td>
<td>20 965</td>
<td>9,8%</td>
<td></td>
</tr>
<tr>
<td>książka</td>
<td>12 503</td>
<td>5,8%</td>
<td></td>
</tr>
<tr>
<td>Gazeta</td>
<td>11 098</td>
<td>5,2%</td>
<td></td>
</tr>
<tr>
<td>pocztówka</td>
<td>5 768</td>
<td>2,7%</td>
<td></td>
</tr>
<tr>
<td>czasopisma</td>
<td>4 962</td>
<td>2,3%</td>
<td></td>
</tr>
<tr>
<td>text</td>
<td>4 452</td>
<td>2,1%</td>
<td></td>
</tr>
<tr>
<td>grafika</td>
<td>3 863</td>
<td>1,8%</td>
<td></td>
</tr>
<tr>
<td>fotografia</td>
<td>3 596</td>
<td>1,7%</td>
<td></td>
</tr>
<tr>
<td>artykuł z czasopisma</td>
<td>3 164</td>
<td>1,5%</td>
<td>2,6%</td>
</tr>
<tr>
<td>artykuł</td>
<td>2 455</td>
<td>1,1%</td>
<td></td>
</tr>
<tr>
<td>Czasopisma</td>
<td>1 710</td>
<td>0,8%</td>
<td></td>
</tr>
<tr>
<td>dzienniki urzędowe</td>
<td>1 516</td>
<td>0,7%</td>
<td></td>
</tr>
<tr>
<td>stary druk</td>
<td>1 222</td>
<td>0,6%</td>
<td>1,1%</td>
</tr>
<tr>
<td>starodruk</td>
<td>1 221</td>
<td>0,6%</td>
<td></td>
</tr>
<tr>
<td>rysunek</td>
<td>1 094</td>
<td>0,5%</td>
<td></td>
</tr>
<tr>
<td>rękopis</td>
<td>1 062</td>
<td>0,5%</td>
<td></td>
</tr>
<tr>
<td>mapa</td>
<td>1 028</td>
<td>0,5%</td>
<td></td>
</tr>
<tr>
<td><strong>Sum</strong></td>
<td></td>
<td><strong>85,1%</strong></td>
<td><strong>68,9%</strong></td>
</tr>
</tbody>
</table>
PIONIER Digital Libraries Federation as a Metadata Aggregator for Europeana

• To de-duplicate, disambiguate, clean, **enrich the data with meaningful attributes**, possibly associate content in collections

  – ESE ver. 3.1 consists of:
    
    A. 15 Dublin Core elements
    + 21 Dublin Core element refinements
    
    B. 1 Dublin Core terms element
    
    C. 11 Europeana specific elements

  – Majority of elements from A and B should be obtained from aggregated digital library

  – Some of A and B elements can be extracted from other elements

  • It strongly depends on the rules of metadata creation used in particular digital library
11 Europeana specific elements

- `isShownBy`, `isShownAt`
  - links to objects used in Europeana interface
- `unstored`
  - placeholder for everything that cannot be mapped to DC
- `object` – URL to the object which will be used for thumbnail/sample generation
  - Creation of this element may be automated on the basis of digital library interface URLs
    - http://www.wbc.poznan.pl/Content/2752
- `hasObject`
  - true or false – indicates if ‘object’ element is available
• 11 Europeana specific elements
  – provider
    • name of the Europeana content provider (the one who sends data to Europeana – eg. aggregator)
  – language
    • ISO 639-1 official language of the content provider
  – country
    • ISO 3166 code of the content provider’s country
  – uri
    • unique identifier of the aggregated object
11 Europeana specific elements

- **type**
  - one of TEXT, IMAGE, SOUND, VIDEO
  - automated mapping from the vocabulary of the aggregator (on the basis of DC:type and DC:format)

- **userTag**
  - tags describing the object, created by (Europeana??) users

- **year**
  - 4 digit year in Gregorian calendar, for the timeline
  - in many cases it can be extracted from the DC:date
• To de-duplicate, disambiguate, clean, enrich the data with meaningful attributes, possibly associate content in collections
  • Not right now...

• To verify the accessibility of content

• To make data ready for Europeana data collection using the OAI-PMH protocol
  • The OAI-PMH interface is now available
  • Polish National Library is visible in the PDLF, but wants to be connected to Europeana directly – it will not be visible in the PDLF OAI-PMH interface
• The software on which the PDLF is based will be released in next few months as an open-source package.

• First version of package was presented at the ECDL 2009 conference during the tutorial "Aggregation and reuse of digital objects’ metadata from distributed digital libraries".
Thank you!

Adam Dudczak (maneo@man.poznan.pl)

Digital Libraries Federation
http://fbc.pionier.net.pl/
PSNC Digital Libraries Team
http://dl.psnec.pl/
– looking forward to cooperate with you!