DIGITAL LIBRARIES FEDERATION

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

PIONIER

Digital Libraries in Poland

Overall number of digital objects

✓ 280 thousands

Number of active digital libraries:

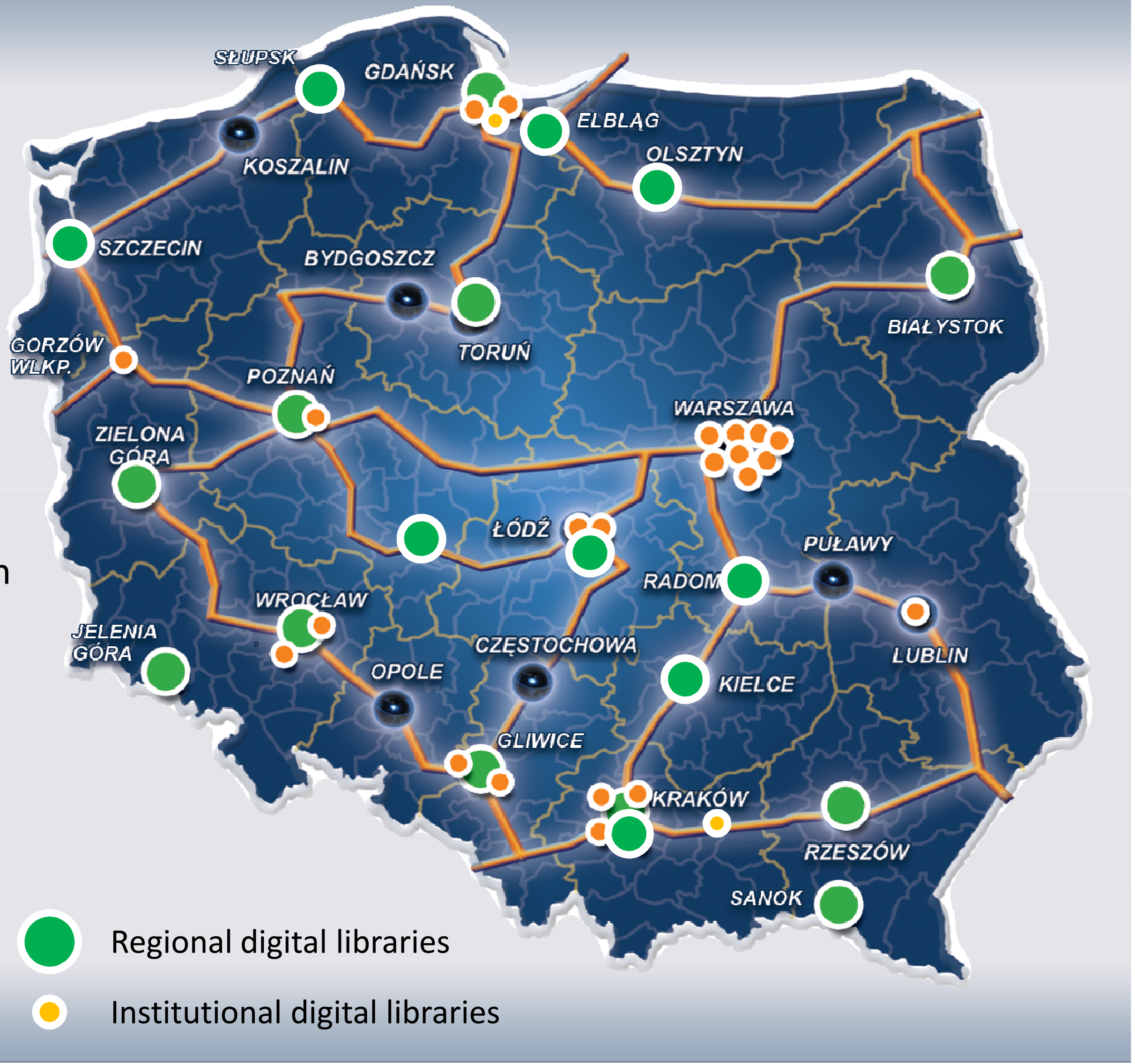
✓ 19 regional

✓ 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

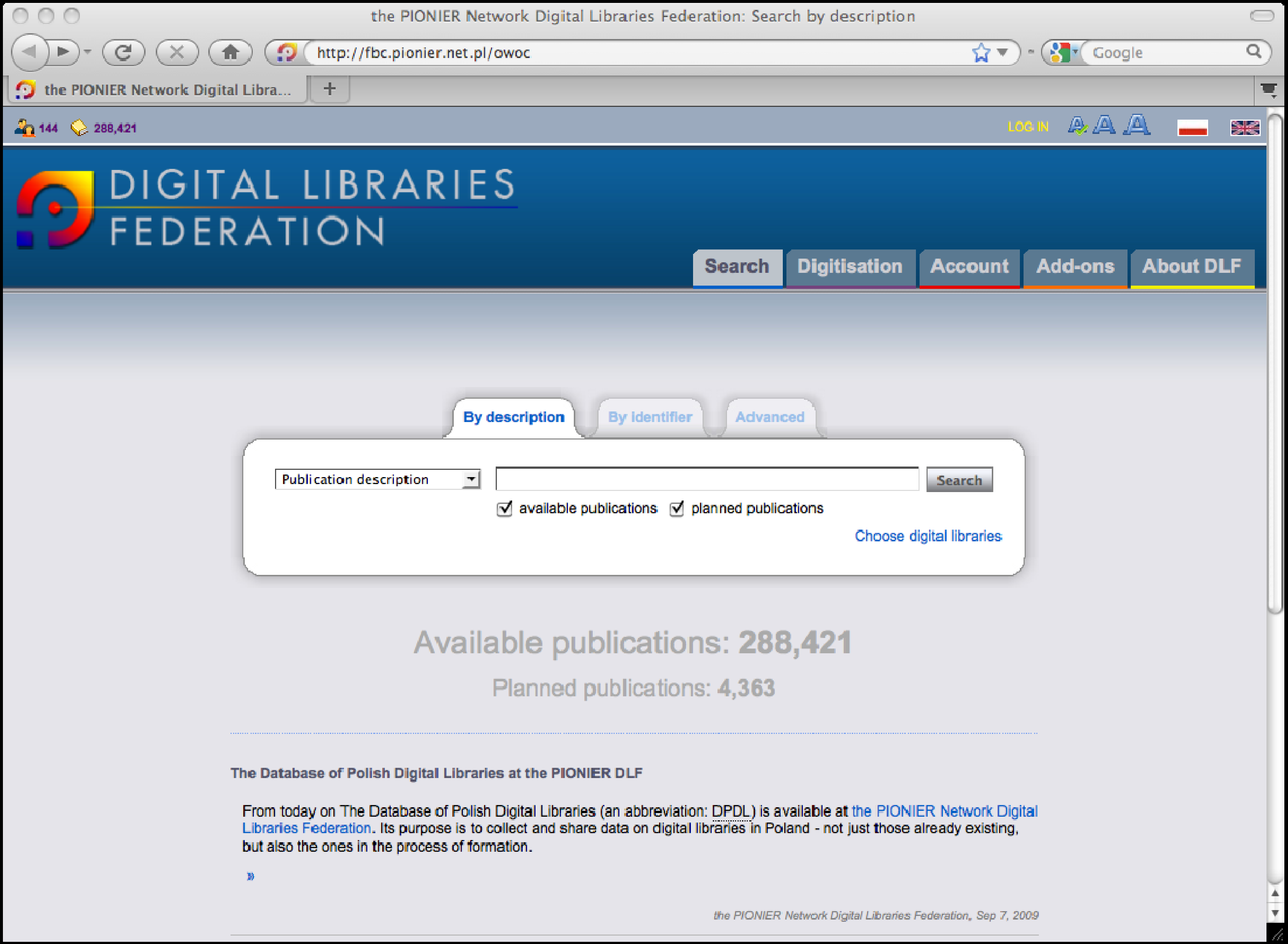


● Regional digital libraries

● Institutional digital libraries

- 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 ☑ planned publications

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.

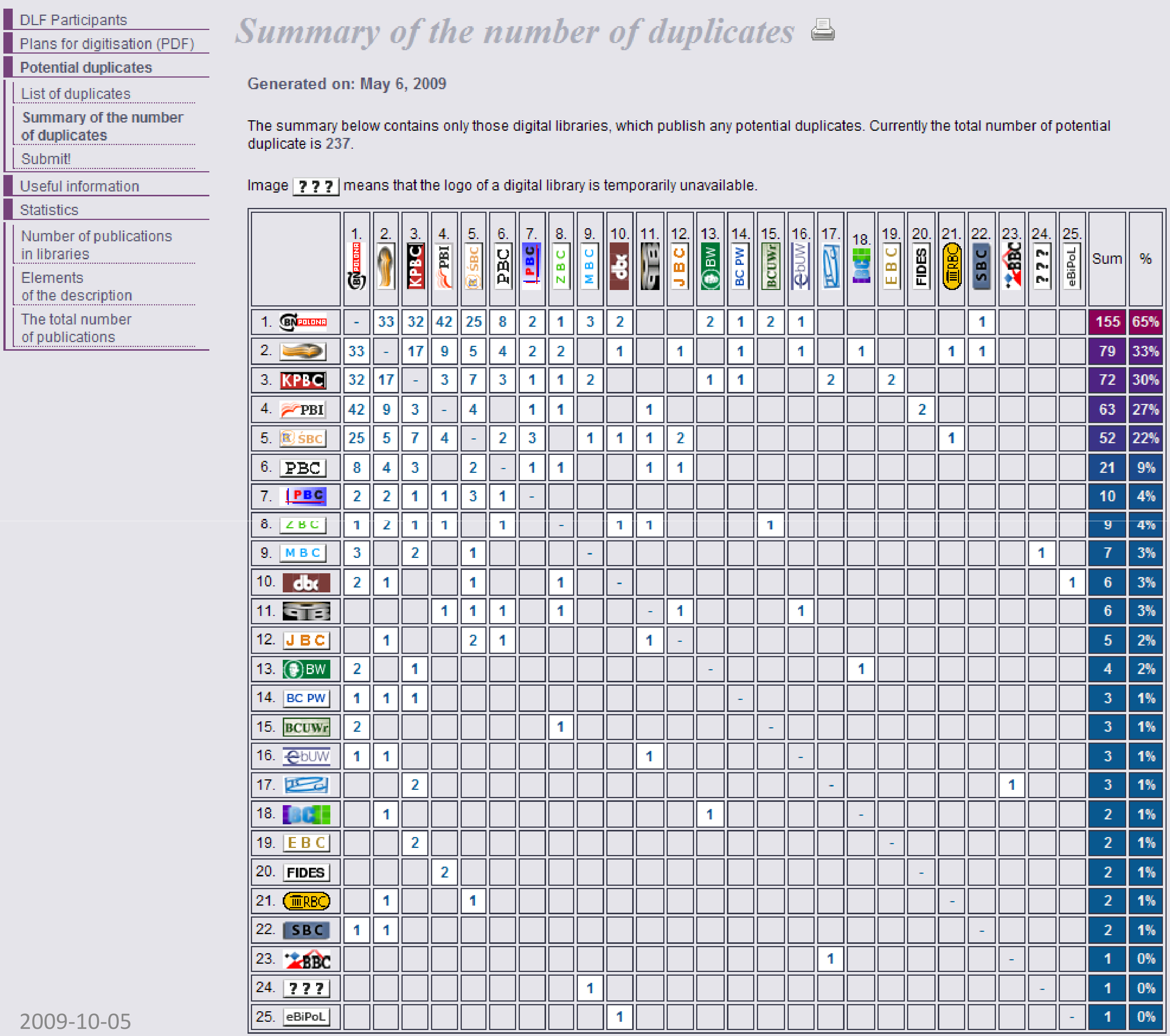
»

the PIONIER Network Digital Libraries Federation, Sep 7, 2009

- 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

PIONIER Digital Libraries Federation

- 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



- DLF Participants
- Plans for digitisation (PDF)
- **Potential duplicates**
  - List of duplicates
  - **Summary of the number of duplicates**
  - Submit!
- Useful information
- Statistics
  - Number of publications in libraries
  - Elements of the description
  - The total number of publications

Summary of the number of duplicates

Generated on: May 6, 2009

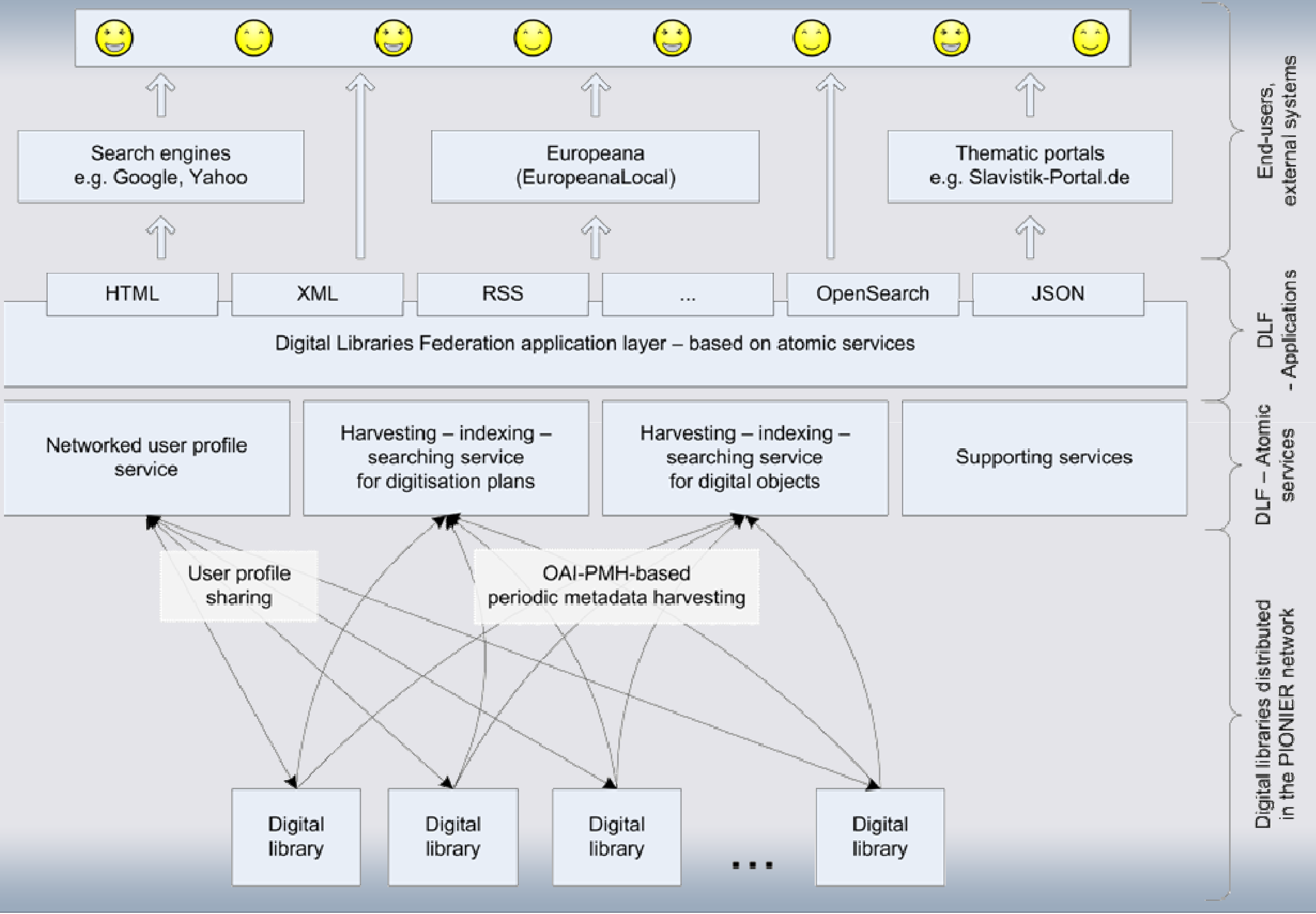
The summary below contains only those digital libraries, which publish any potential duplicates. Currently the total number of potential duplicate is 237.

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

	1. BN POLONA	2.	3. KPBC	4. PBI	5. ŚBC	6. PBC	7. PBC	8. ZBC	9. MBC	10. dbc	11.	12. JBC	13. BW	14. BC PW	15. BCUWr	16. ebUW	17.	18. BC	19. EBC	20. FIDES	21. RBC	22. SBC	23. BBC	24. ???	25. eBiPoL	Sum	%
1. BN POLONA	-	33	32	42	25	8	2	1	3	2			2	1	2	1						1				155	65%
2.	33	-	17	9	5	4	2	2		1		1		1		1		1			1	1				79	33%
3. KPBC	32	17	-	3	7	3	1	1	2				1	1			2		2							72	30%
4. PBI	42	9	3	-	4		1	1			1									2						63	27%
5. ŚBC	25	5	7	4	-	2	3		1	1	1	2									1					52	22%
6. PBC	8	4	3		2	-	1	1			1	1														21	9%
7. PBC	2	2	1	1	3	1	-																			10	4%
8. ZBC	1	2	1	1		1		-		1	1				1											9	4%
9. MBC	3		2		1				-															1		7	3%
10. dbc	2	1			1			1		-															1	6	3%
11.				1	1	1		1			-	1				1										6	3%
12. JBC		1			2	1					1	-														5	2%
13. BW	2		1										-					1								4	2%
14. BC PW	1	1	1											-												3	1%
15. BCUWr	2							1							-											3	1%
16. ebUW	1	1									1					-										3	1%
17.			2														-						1			3	1%
18. BC		1											1					-								2	1%
19. EBC			2																-							2	1%
20. FIDES				2																-						2	1%
21. RBC		1			1																-					2	1%
22. SBC	1	1																				-				2	1%
23. BBC																	1						-			1	0%
24. ???									1															-		1	0%
25. eBiPoL										1															-	1	0%

DIGITAL LIBRARIES FEDERATION

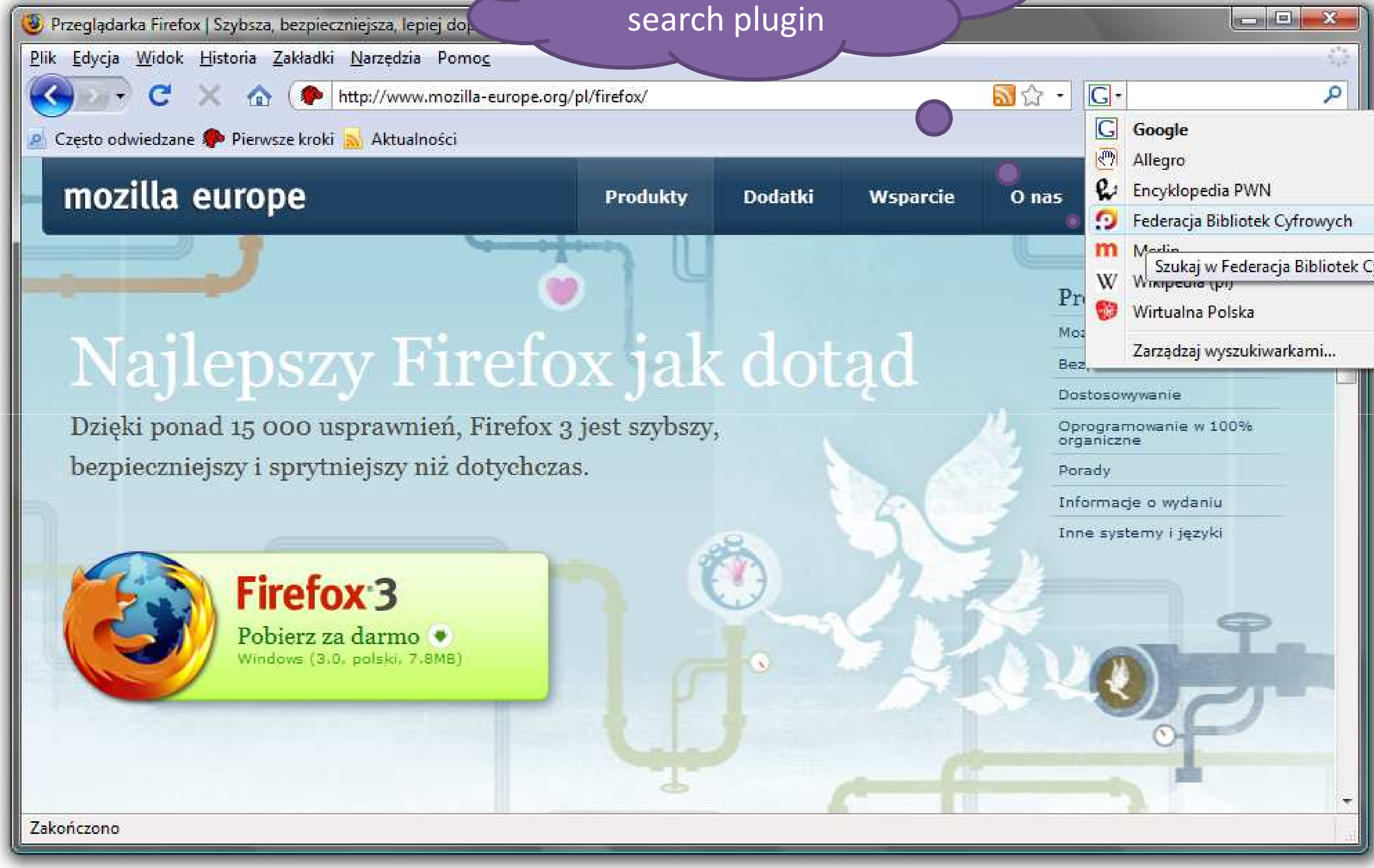
PIONIER Digital Libraries Federation - Architecture


Search engines e.g. Google, Yahoo
Europeana (EuropeanaLocal)
Thematic portals e.g. Slavistik-Portal.de
HTML
XML
RSS
...
OpenSearch
JSON
Digital Libraries Federation application layer – based on atomic services
Networked user profile service
Harvesting – indexing – searching service for digitisation plans
Harvesting – indexing – searching service for digital objects
Supporting services
User profile sharing
OAI-PMH-based periodic metadata harvesting
Digital library
Digital library
Digital library
...
Digital library
End-users, external systems
DLF - Applications
DLF – Atomic services
Digital libraries distributed in the PIONIER network

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?

Digital Libraries Federation search plugin

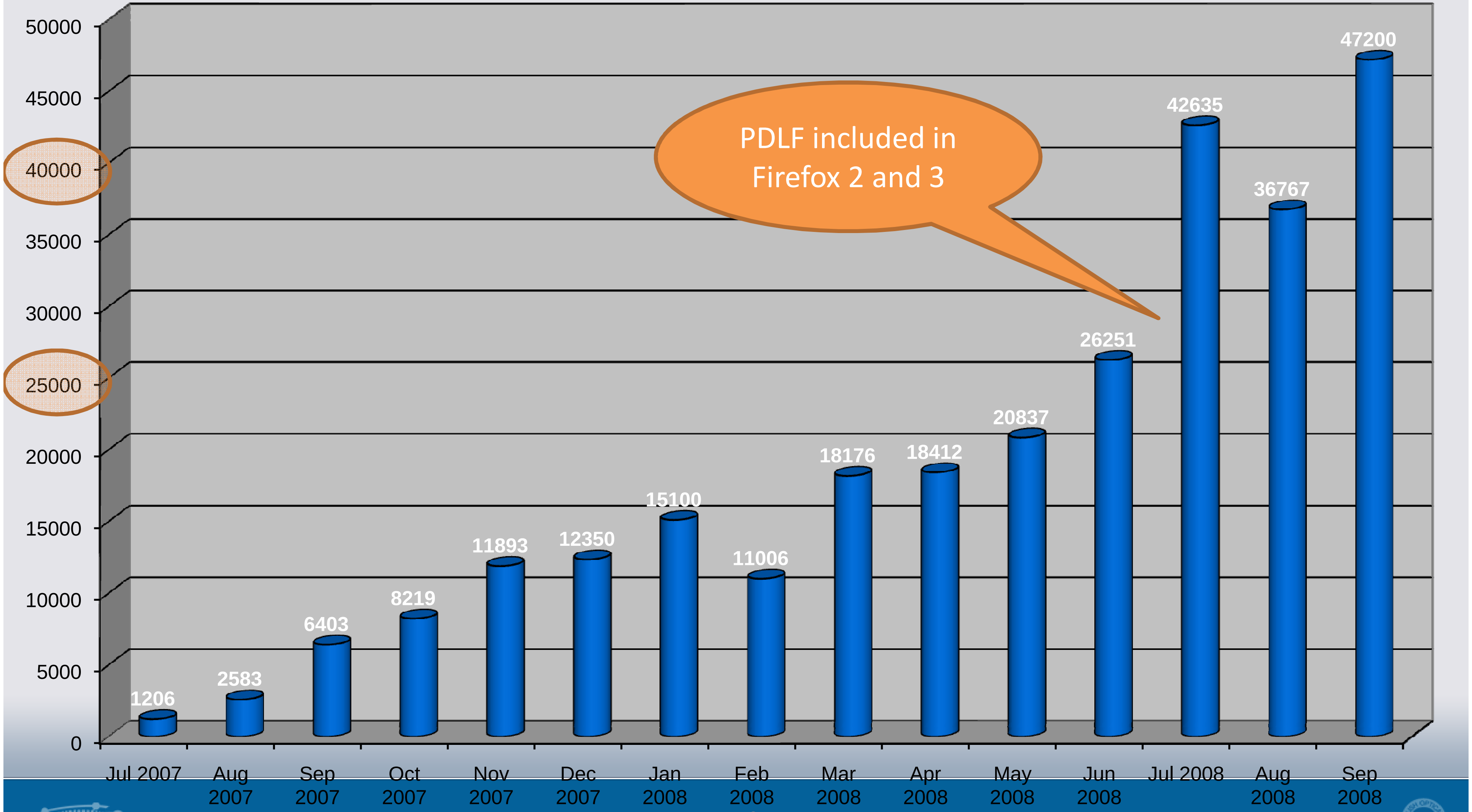

Przeglądarka Firefox | Szybsza, bezpieczniejsza, lepiej do
Plik Edycja Widok Historia Zakładki Narzędzia Pomoc
http://www.mozilla-europe.org/pl/firefox/
Często odwiedzane Pierwsze kroki Aktualności
mozilla europe
Produkty Dodatki Wsparcie O nas
Google
Allegro
Encyklopedia PWN
Federacja Bibliotek Cyfrowych
Merlin
Szukaj w Federacja Bibliotek Cyfrowych
Wikipedia (pl)
Wirtualna Polska
Zarządzaj wyszukiwarkami...
Dostosowywanie
Oprogramowanie w 100% organiczne
Porady
Informacje o wydaniu
Inne systemy i języki
Firefox 3
Pobierz za darmo
Windows (3.0, polski, 7.8MB)
Zakończono

Najlepszy Firefox jak dotąd

Dzięki ponad 15 000 usprawnień, Firefox 3 jest szybszy, bezpieczniejszy i sprytniejszy niż dotychczas.

DIGITAL LIBRARIES FEDERATION

Total number of visits in the PDLF each month



AGGREGATING DIGITAL LIBRARIES IN POLAND… FOR EUROPEANA

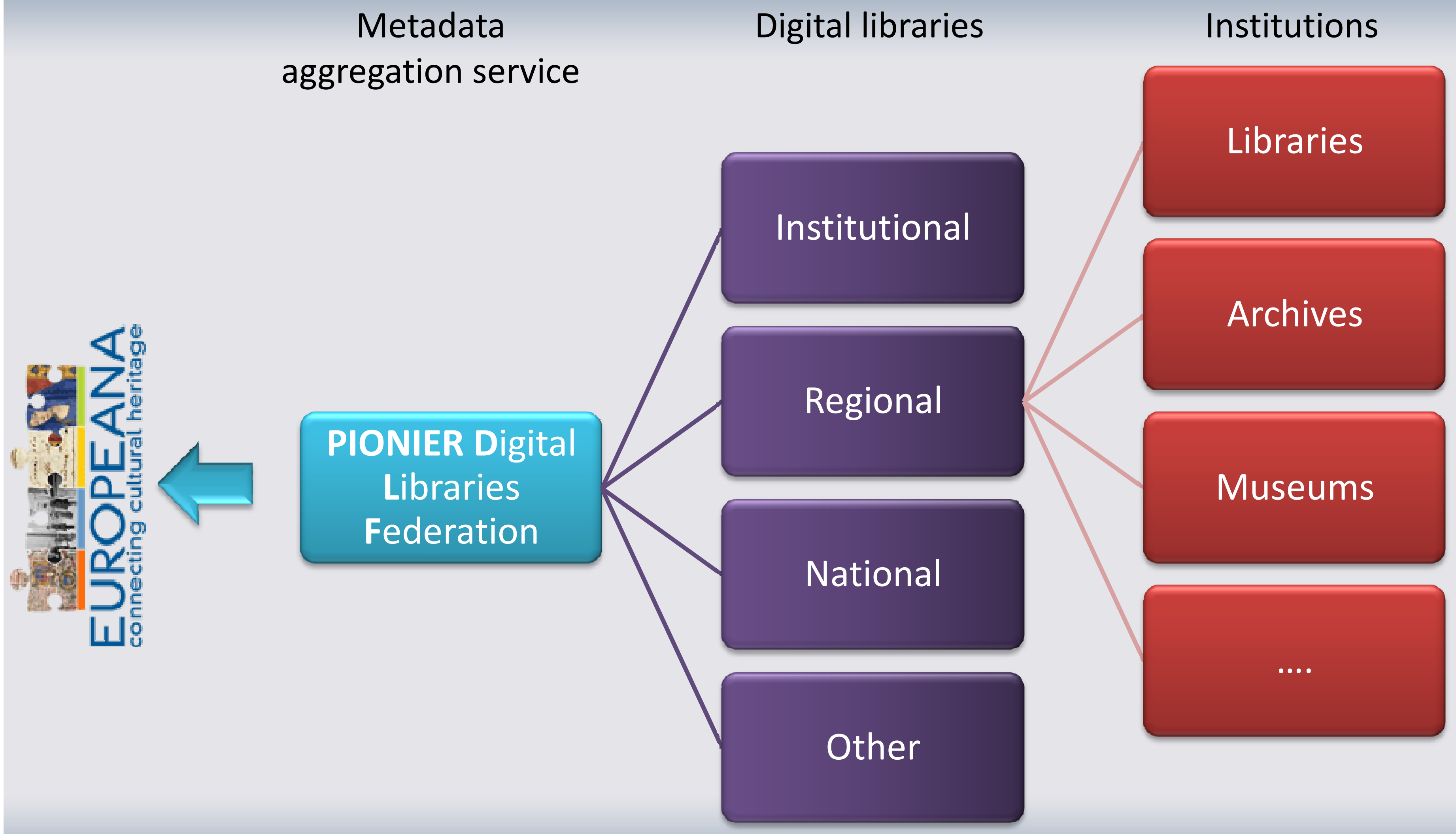
Connection to European and Worldwide Infrastructure

- 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

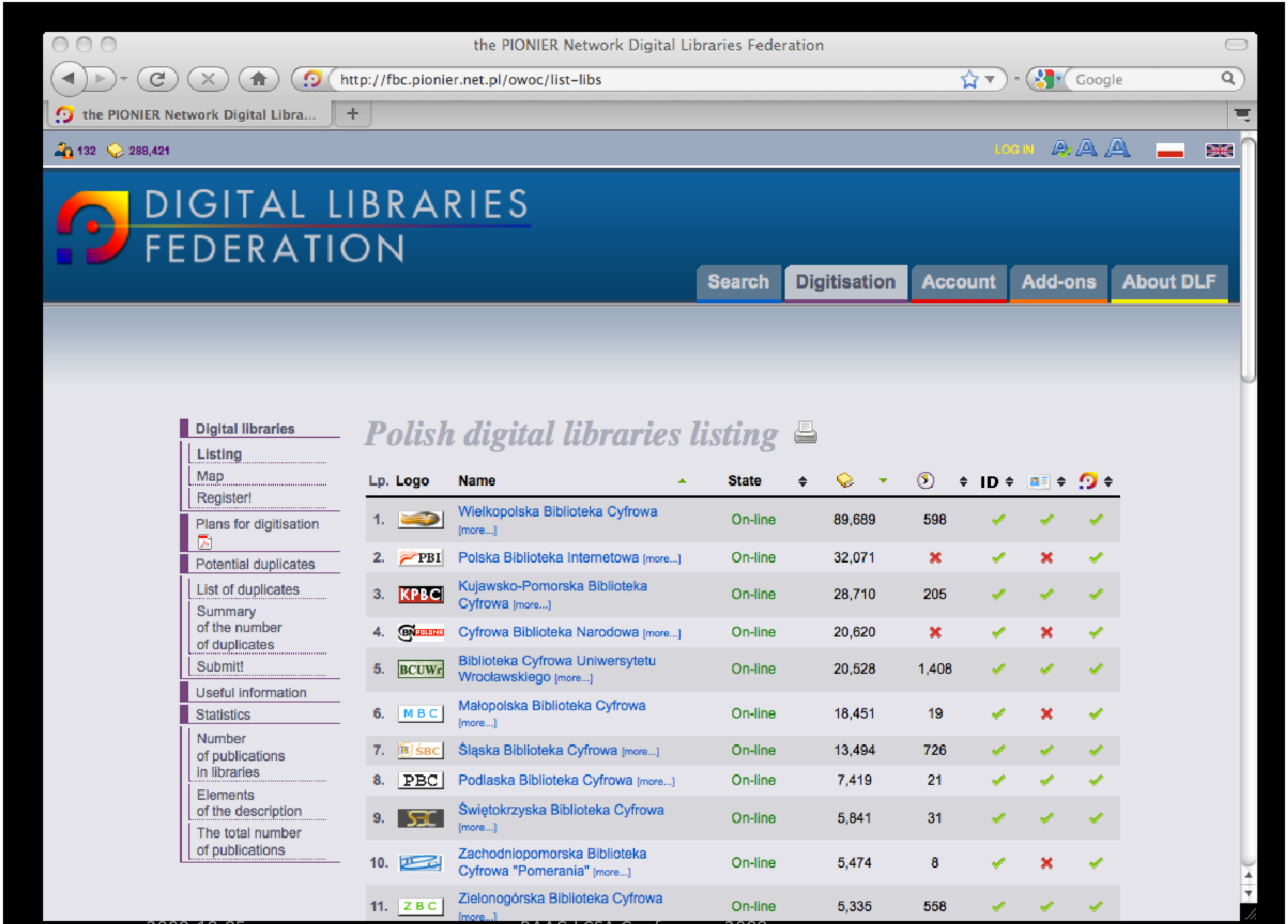
Metadata aggregators for Europeana

- 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
Institutions
EUROPEANA connecting cultural heritage
PIONIER Digital Libraries Federation
Institutional
Regional
National
Other
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 statisitics



the PIONIER Network Digital Libraries Federation

http://fbc.pionier.net.pl/owoc/list-libs

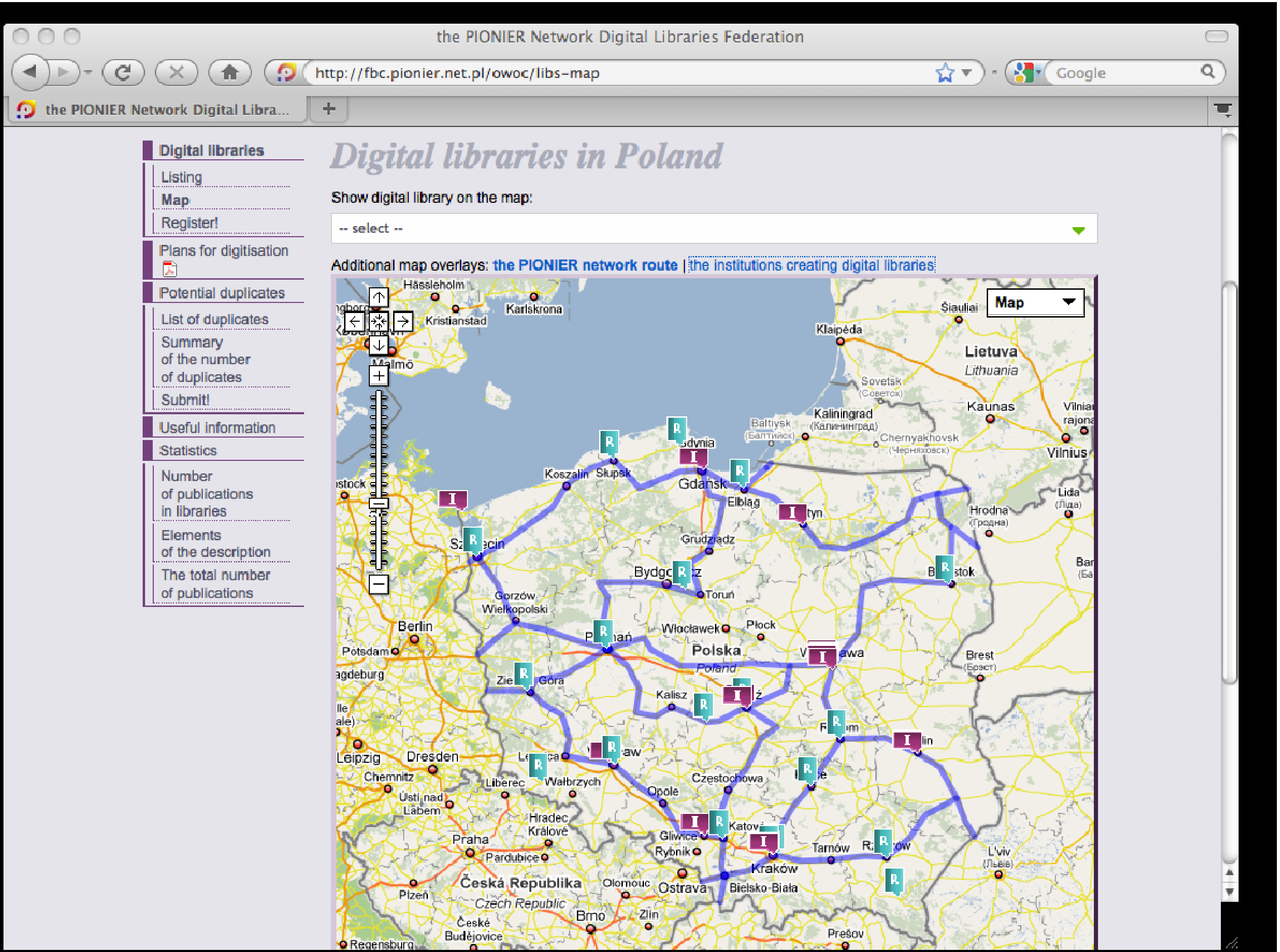
132 288,421 LOG IN

DIGITAL LIBRARIES FEDERATION

- Digital libraries
  - Listing
  - Map
  - Register!
- Plans for digitisation
- Potential duplicates
  - List of duplicates
  - Summary of the number of duplicates
  - Submit!
- Useful information
- Statistics
  - Number of publications in libraries
  - Elements of the description
  - The total number of publications

Polish digital libraries listing

Lp.	Logo	Name	State			ID		
1.		Wielkopolska Biblioteka Cyfrowa [more...]	On-line	89,689	598	✓	✓	✓
2.	PBI	Polska Biblioteka Internetowa [more...]	On-line	32,071	✗	✓	✗	✓
3.	KPBC	Kujawsko-Pomorska Biblioteka Cyfrowa [more...]	On-line	28,710	205	✓	✓	✓
4.	BN POLONA	Cyfrowa Biblioteka Narodowa [more...]	On-line	20,620	✗	✓	✗	✓
5.	BCUWr	Biblioteka Cyfrowa Uniwersytetu Wrocławskiego [more...]	On-line	20,528	1,408	✓	✓	✓
6.	MBC	Małopolska Biblioteka Cyfrowa [more...]	On-line	18,451	19	✓	✗	✓
7.	ŚBC	Śląska Biblioteka Cyfrowa [more...]	On-line	13,494	726	✓	✓	✓
8.	PBC	Podlaska Biblioteka Cyfrowa [more...]	On-line	7,419	21	✓	✓	✓
9.		Świętokrzyska Biblioteka Cyfrowa [more...]	On-line	5,841	31	✓	✓	✓
10.		Zachodniopomorska Biblioteka Cyfrowa "Pomerania" [more...]	On-line	5,474	8	✓	✗	✓
11.	ZBC	Zielonogórska Biblioteka Cyfrowa [more...]	On-line	5,335	558	✓	✓	✓

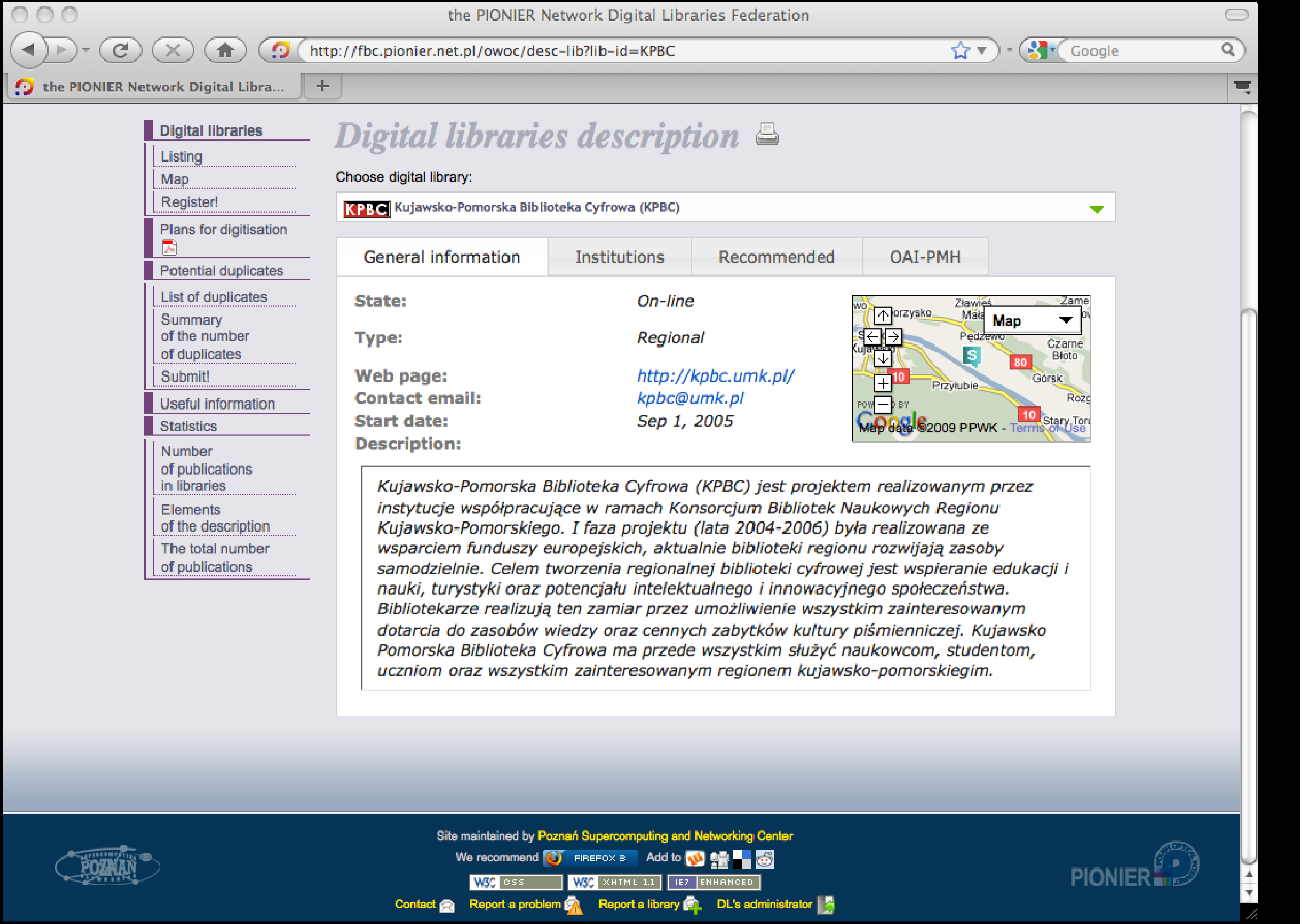


Digital libraries in Poland

Show digital library on the map:

-- select --

Additional map overlays: the PIONIER network route | the institutions creating digital libraries



the PIONIER Network Digital Libraries Federation
http://fbc.pionier.net.pl/owoc/desc-lib?lib-id=KPBC
Google
the PIONIER Network Digital Libra...
Digital libraries
Listing
Map
Register!
Plans for digitisation
Potential duplicates
List of duplicates
Summary of the number of duplicates
Submit!
Useful information
Statistics
Number of publications in libraries
Elements of the description
The total number of publications

Digital libraries description

Choose digital library:

KPBC Kujawsko-Pomorska Biblioteka Cyfrowa (KPBC)

General information
Institutions
Recommended
OAI-PMH

State: *On-line*

Type: *Regional*

Web page: *http://kpbc.umk.pl/*

Contact email: *kpbc@umk.pl*

Start date: *Sep 1, 2005*

Description:

Map
Map data ©2009 PPWK - Terms of Use
Przyłubie
Górsk
Czarne Błoto
Pędzewo

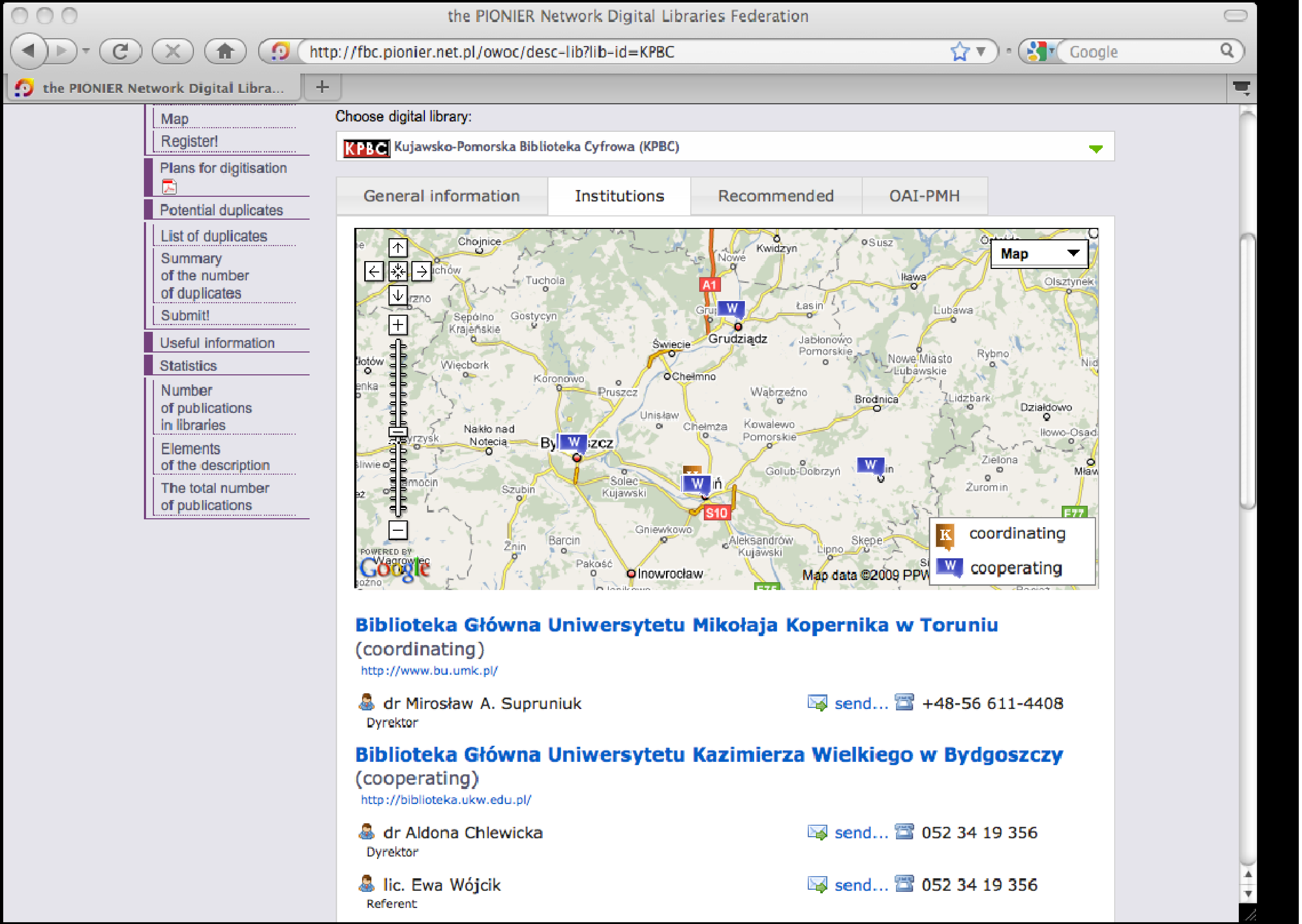
Kujawsko-Pomorska Biblioteka Cyfrowa (KPBC) jest projektem realizowanym przez instytucje współpracujące w ramach Konsorcjum Bibliotek Naukowych Regionu Kujawsko-Pomorskiego. I faza projektu (lata 2004-2006) była realizowana ze wsparciem funduszy europejskich, aktualnie biblioteki regionu rozwijają zasoby samodzielnie. Celem tworzenia regionalnej biblioteki cyfrowej jest wspieranie edukacji i nauki, turystyki oraz potencjału intelektualnego i innowacyjnego społeczeństwa. Bibliotekarze realizują ten zamiar przez umożliwienie wszystkim zainteresowanym dotarcia do zasobów wiedzy oraz cennych zabytków kultury piśmienniczej. Kujawsko Pomorska Biblioteka Cyfrowa ma przede wszystkim służyć naukowcom, studentom, uczniom oraz wszystkim zainteresowanym regionem kujawsko-pomorskiegim.

Site maintained by Poznań Supercomputing and Networking Center

We recommend FIREFOX 3 Add to

W3C CSS W3C XHTML 1.1 IE7 ENHANCED

Contact Report a problem Report a library DL's administrator


the PIONIER Network Digital Libraries Federation
http://fbc.pionier.net.pl/owoc/desc-lib?lib-id=KPBC
Google
the PIONIER Network Digital Libra...
Map
Register!
Plans for digitisation
Potential duplicates
List of duplicates
Summary of the number of duplicates
Submit!
Useful information
Statistics
Number of publications in libraries
Elements of the description
The total number of publications
Map
Chojnice
Kwidzyn
Susz
Tuchola
Grudziądz
Świecie
Chełmno
Brodnica
Nakło nad Notecią
Toruń
Inowrocław
Map data ©2009 PPWK
K coordinating
W cooperating

Choose digital library:

KPBC Kujawsko-Pomorska Biblioteka Cyfrowa (KPBC)

General information | Institutions | Recommended | OAI-PMH

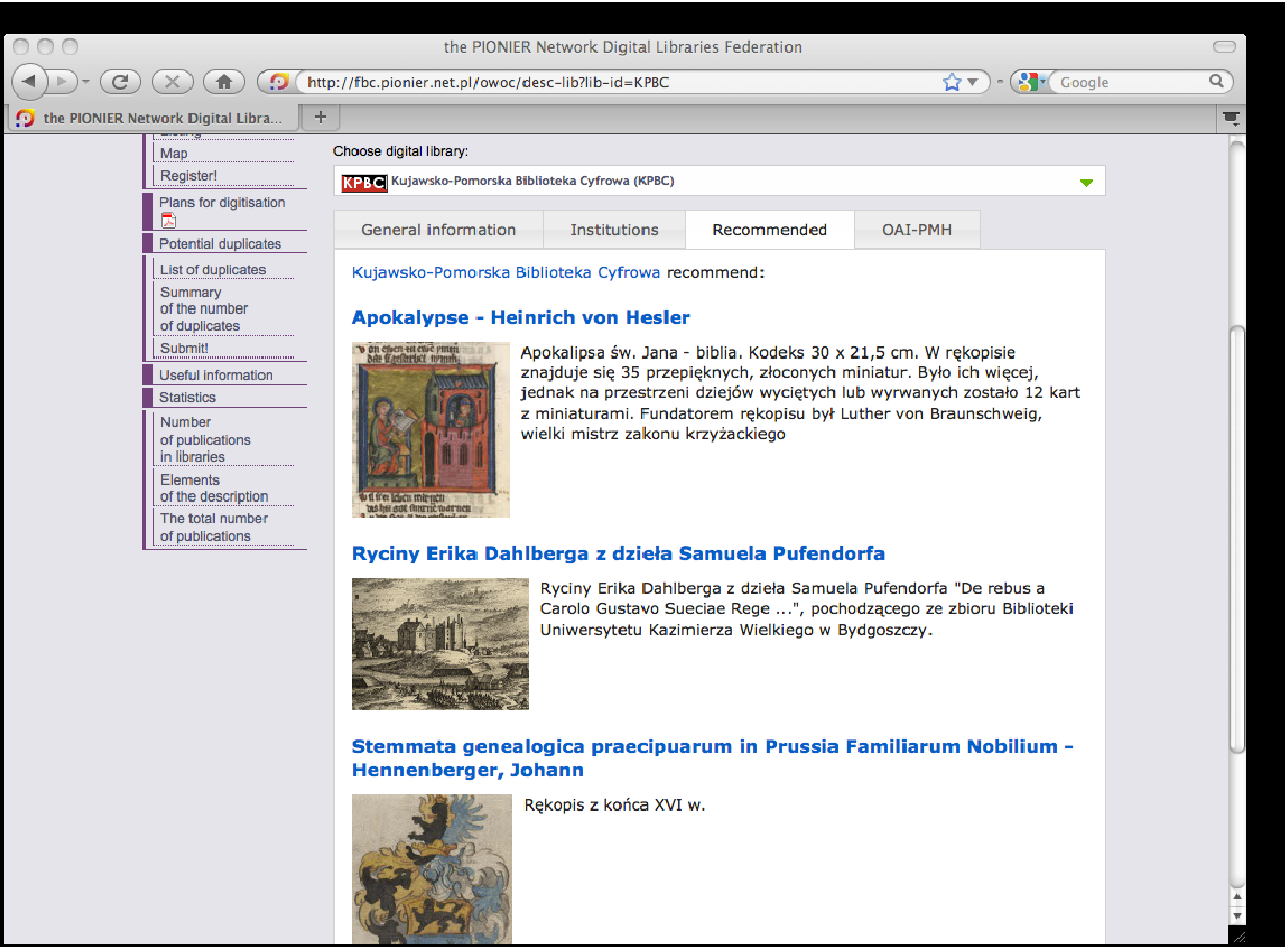
Biblioteka Główna Uniwersytetu Mikołaja Kopernika w Toruniu (coordinating)
http://www.bu.umk.pl/

dr Mirosław A. Supruniuk — send... ☎ +48-56 611-4408
Dyrektor

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

dr Aldona Chlewicka — send... ☎ 052 34 19 356
Dyrektor

lic. Ewa Wójcik — send... ☎ 052 34 19 356
Referent


the PIONIER Network Digital Libraries Federation
http://fbc.pionier.net.pl/owoc/desc-lib?lib-id=KPBC
Google
the PIONIER Network Digital Libra...
Map
Register!
Plans for digitisation
Potential duplicates
List of duplicates
Summary of the number of duplicates
Submit!
Useful information
Statistics
Number of publications in libraries
Elements of the description
The total number of publications
General information
Institutions
Recommended
OAI-PMH

Choose digital library:

KPBC Kujawsko-Pomorska Biblioteka Cyfrowa (KPBC)

Kujawsko-Pomorska Biblioteka Cyfrowa recommend:

Apokalypse - Heinrich von Hesler

Apokalipsa św. Jana - biblia. Kodeks 30 x 21,5 cm. W rękopisie znajduje się 35 przepięknych, złoconych miniatur. Było ich więcej, jednak na przestrzeni dziejów wyciętych lub wyrwanych zostało 12 kart z miniaturami. Fundatorem rękopisu był Luther von Braunschweig, wielki mistrz zakonu krzyżackiego

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 <u>potential</u> 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

DC element	No. of unique values	Number of associations	Average no. of occurrences
format	39	209 789	5 379,2
language	195	210 529	1 079,6
type	822	211 816	257,7
rights	1 192	246 093	206,5
coverage	66	2 390	36,2
publisher	18 002	310 764	17,3
contributor	12 979	83 464	6,4
subject	78 440	438 871	5,6
relation	9 292	48 319	5,2
date	47 581	209 589	4,4
identifier	6 426	27 666	4,3
description	43 657	180 391	4,1
source	16 996	52 506	3,1
creator	21 908	67 503	3,1
title	210 745	227 039	1,1

- Format
  - In 99% of descriptions: MIME type(eg. 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
  - …

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

- 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/dlibra/docmetadata?id=2752
      http://www.wbc.poznan.pl/Content/2752
      http://www.wbc.poznan.pl/image/edition/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
  - **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

Conclusions

- 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.psnc.pl/

– looking forward to **cooperate** with you!