



Big (meta)data metadane jako źródło badań w wielkiej skali

dr Piotr Malak, Uniwersytet Wrocławski

**XII Konferencja z cyklu „Automatyzacja bibliotek”
Biblioteka w cyberprzestrzeni
Inspiracje Światowego Kongresu IFLA we Wrocławiu 2017**

Warszawa, 13-14 listopada 2018

Big Data – metody i narzędzia analizy rozległych zbiorów oraz strumieni danych wykorzystywane do pozyskania informacji i wiedzy z danych, które nie są ze sobą powiązane wprost.

Model podstawowy: 3 V (*Volume, Velocity, Variety*)

- ❑ **Wielkość** zbioru danych (*volume*) – ilość danych nie podlegająca przetwarzaniu i analizie za pomocą tradycyjnych metod,
- ❑ **Dynamika** przyrastania oraz wykorzystania danych (*velocity*) – metadane, rozpatrywane w skali wszystkich dostępnych ich źródeł, przyrastają w sposób ciągły,
- ❑ **Różnorodność** danych (*variety*) – dostępne i używane różnorodne formaty metadanych

Analiza danych, pozyskiwanie wiedzy, wnioskowanie:

- bankowość,
- media społecznościowe,
- dane klimatyczne,
- dane medyczne i zarządzanie zdrowiem,

- przetwarzanie języka naturalnego,
- analizy stylometryczne,
- itp.

Kluczowe narzędzia:

- Hadoop:** <https://hadoop.apache.org/> - platforma statycznego przetwarzania Big Data
- Spark:** <https://spark.apache.org/> - platforma przetwarzania strumieni Big Data
- SPARQL:** <https://www.w3.org/TR/rdf-sparql-query/> - język zapytań dla RDF
- Wizualizacja**
- API** (ang. *Application Programming Interface*) – narzędzia programistyczne umożliwiające zautomatyzowane pobieranie zestawów lub strumieni danych lub informacji



METADANE W BIBLIOTEKACH

❑ Wielkość

- ❑ **Biblioteka Narodowa:** ponad 4 678 549 rekordów bibliograficznych, 2 175 147 r. wzorcowych¹, 4,4 GB danych²
- ❑ **Federacja Bibliotek Cyfrowych:** dane ponad 5 500 000 obiektów³
- ❑ **Google Books Dataset:** 3 000 000 tomów, 2,9 TB danych, 11 GB metadanych³
- ❑ **British Library. Collection metadata:** tematyczne zestawy metadanych

1. Aktualności BN, 21.12.2007: <https://www.bn.org.pl/aktualnosci/3345-biblioteka-narodowa-otwiera-najwieksza-polska-baze-danych-bibliograficznych.html>
2. BN, *Bazy do pobrania*, 4.11.2018, <https://data.bn.org.pl/databases>
3. *Federacja Bibliotek Cyfrowych*, 4.11.2018, <https://fbc.pionier.net.pl/>
4. Google Books Dataset, 4.11.2018, <https://lib.msu.edu/gds/>
5. *British Library. Collection metadata*, 4.11.2018, <http://www.bl.uk/bibliographic/download.html>

- Dynamika** – przyrost ciągły, lecz nie strumieniowy
 - Repozytoria,
 - Dane dotyczące badań naukowych,
 - ...
- Różnorodność:**
 - Dublin Core,
 - XML,
 - JSON,
 - MARCXML,
 - MARC,
 - RDF,
 - CSV,
 -



METADANE — PRZYKŁADY ANALIZ

Określenie typu publikacji w bibliotekach cyfrowych

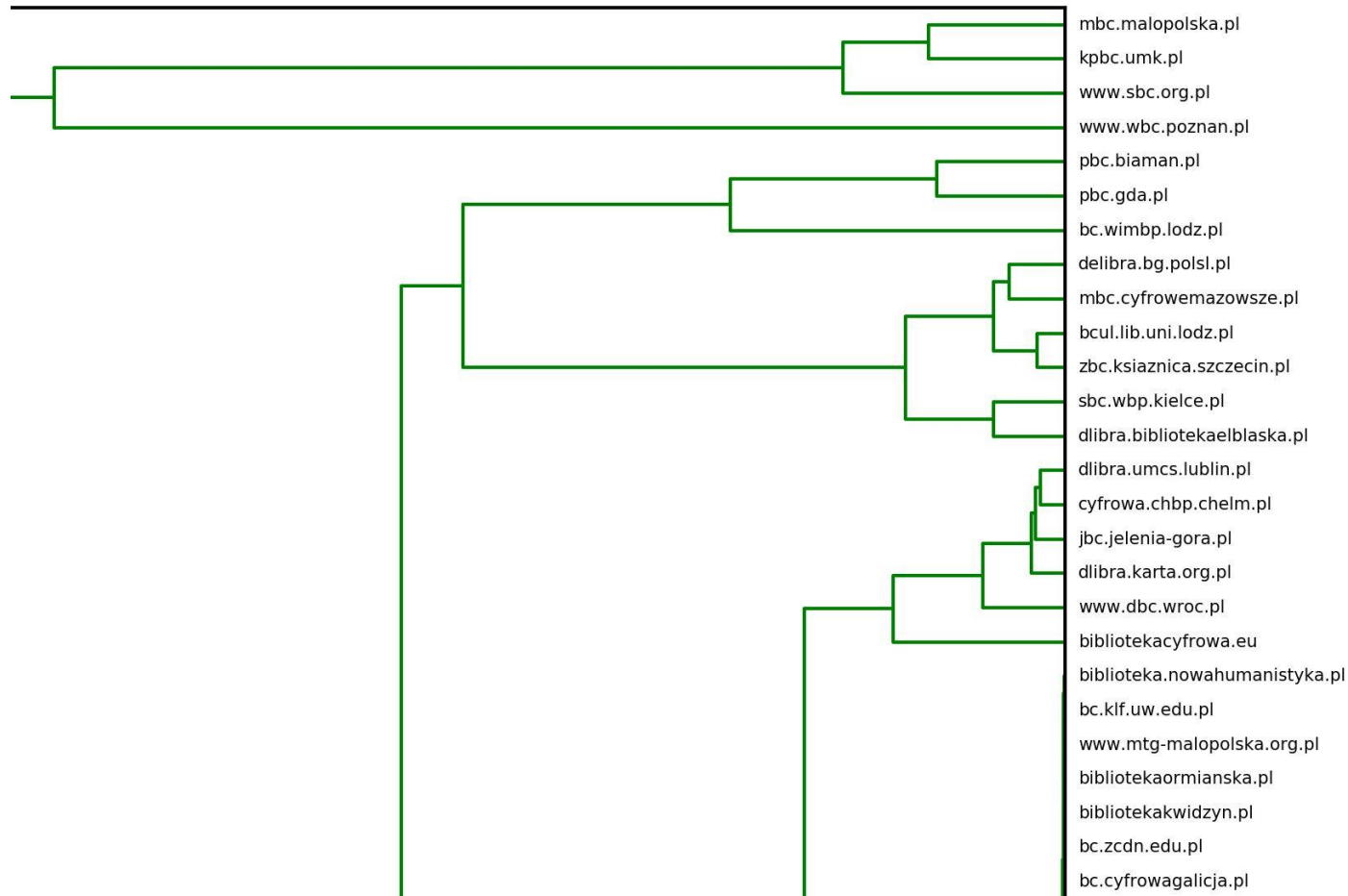


The image shows a screenshot of a metadata table, likely from a digital library system. The table is oriented vertically on the page. It contains numerous columns, each representing a different metadata field, and several rows of data. The text is small and difficult to read, but the structure is clearly that of a database table. The columns are separated by vertical lines, and the rows are separated by horizontal lines. The data appears to be organized in a structured manner, typical of a metadata catalog.

Określenie typu publikacji w bibliotekach cyfrowych

www.wbc.poznan.pl	1594	gazety	61588	czasopismo	40687	gazeta	38227	czasopisma	33403	archiwalia
dlibra.karta.org.pl	1135	czasopismo	9541	list	1892	karta	1249	wykaz	678	ulotka
cyfrowe.mnw.art.pl	331	obraz	3651	rysunek	2608	grafika, odbitka graficzna	1771	moneta	1274	fotografia, fotografia z natury
www.sbc.org.pl	244	czasopismo	90432	opracowanie statystyczne	3934	fotografia	2875	druk ulotny	1934	pocztówka
bibliotekacyfrowa.eu	185	magazyn, czasopismo, journal	9589	zeitschrift, czasopismo, journal	6454	zeitschriften, czasopismo, journal	5144	czasopismo	4804	fotografia
www.pbc.zesow.pl	166	czasopismo	2364	gazeta	2339	sprawozdanie szkolne	2219	książka	1091	fotografia
cyfrowa.chbp.chelm.pl	158	czasopismo	7969	książka	444	afisz	413	zaproszenie	322	pocztówka
bcpw.bg.pw.edu.pl	136	czasopismo	2914	książka	329	artykuł naukowy	272	fotografia	215	artykuł
jbc.bj.uj.edu.pl	126	czasopismo	243654	starodruk	1513	rozprawa doktorska	1510	druk muzyczny	1293	książka
dlibra.bibliotekaelblaska.pl	124	czasopismo	28239	zeitschrift, czasopismo	7070	starodruk	1300	afisz/plakat	897	zeitschrift
cybra.lodz.pl	116	dokument piśmienniczy, czasopismo, .	3783	czasopismo - artykuł, dokument piśmienniczy	994	dokument piśmienniczy, mikrofisz, czas	766	dokument piśmienniczy, czasopismo - a	754	dokument elektroniczny, czasopismo
bbc.mbp.org.pl	99	czasopismo	4396	fotografia	1108	druk ulotny	903	artykuł z czasopisma	620	książka
www.bibliotekacyfrowa.pl	88	czasopisma	22332	wydawnictwa ciągłe	7807	książka	2121	stary druk	1810	rękopis
zbc.uz.zgora.pl	84	czasopismo	5667	gazeta	1768	grafika	1438	informator	934	książka
kpbc.umk.pl	79	czasopismo	64194	książka	3691	gazeta	1846	grafika	1418	ekslibris
pbc.gda.pl	78	gazeta	23948	czasopismo	4582	książka	3633	stary druk	2462	druk muzyczny
mbc.cyfrowemazowsze.pl	72	czasopismo	16415	wydawnictwo urzędowe, czasopismo	3084	książka	1221	dokument życia społecznego	979	fotografia
zbc.ksiaznica.szczecin.pl	70	czasopismo	19775	stary druk	2421	akt prawny	1372	karta pocztowa	1091	afisz
pbc.biaman.pl	61	gazeta	12232	czasopismo	9309	dziennik urzędowy	3931	książka	2897	źródła historyczne
www.tnp.org.pl	57	broszura	122	rękopis	35	ustawa, statut	20	stary druk	18	czasopismo
muzeumleszno.pl	52	medalierstwo, medal	189	portret mieszczański, obraz, portret	83	portret, grafika	74	obraz, scena rodzajowa	64	grafika
ebuw.uw.edu.pl	49	czasopismo, periodical	159153	książka, book	1856	stary druk, early book	183	pocztówka, postcard	105	czasopismo
www.dbc.wroc.pl	42	czasopismo	9114	stary druk	5357	książka	3732	artykuł	3079	rękopis
jbc.jelenia-gora.pl	41	czasopismo	10331	pocztówka	1990	artykuł	460	fotografia	417	książka
delibra.bg.polsl.pl	37	czasopismo	15650	rozprawa doktorska	4712	książka	1038	praca habilitacyjna	430	artykuł
dlibra.umcs.lublin.pl	37	czasopismo	9306	artykuł	1819	książka	1566	plakat	167	stary druk
dlibra.karta.org.pl-catl	36	fotografia	698	karika pocztowa	22	czasopismo	15	zaświadczenie	9	legitymacja
bc.mbradom.pl	33	czasopisma i gazety	26344	książka	2678	starodruk	387	katalog wystawy	115	fotografia
mbc.fundacjamorska.org	32	biuletyn	1054	czasopismo	232	audycja radiowa	162	zarządzenie	117	ustawy
bc.pollub.pl	31	branzowa norma	6730	opis patentowy	604	książka, book	420	czasopismo, periodical	376	rozprawa doktorska, doctoral thesis
mbc.malopolska.pl	29	czasopismo	77645	mapa	3132	dokument archiwalny	2819	książka	1146	rękopis
bc.wbp.lublin.pl	27	czasopismo	3613	książka	1280	fotografia	832	pocztówka	658	brozura
biblioteka.wejherowo.pl	26	artykuł z czasopisma	8919	artykuły z czasopism ogólnopolskich	366	książka	224	afisze i plakaty	170	zaproszenia
sanockabibliotekacyfrowa.pl	26	czasopismo	495	książka	99	rękopis	99	dzienniki urzędowe	83	sprawozdania szkolne
www.bc.ore.edu.pl	26	artykuł	204	czasopismo	66	czasopismo internetowe	60	program nauczania	59	materiały edukacyjne
obc.opole.pl	22	czasopismo	3046	książka	611	karta pocztowa	587	grafik, grafika, graphic	432	druk ulotny
sbc.wbp.kielce.pl	22	czasopismo	26698	czasopismo, mikrofilm	688	książka	266	czasopismo, mikrofilm	159	dokument elektroniczny
sbc.bdsandomierz.pl	21	stary druk	400	książka	96	czasopismo	93	rękopis	62	fotografia

Określenie typu publikacji w bibliotekach cyfrowych



Analiza współpracy naukowej

WSPÓŁPRACA
MIĘDZYNARODOWA

DLA BIZNESU

Strona główna > Badania naukowe > Konferencje > Lista konferencji

LISTA KONFERENCJI

Szukaj:

Nazwa	Termin	Miejsce	Organizator	więcej»
XIV zjazd Polskiego Stowarzyszenia Psychologii Społecznej	2017-09-15 - 2017-09-17	Toruń	Katedra Psychologii, Wydział Humanistyczny	więcej»
V MIĘDZYNARODOWY KONGRES RELIGIOZNAWCZY: RELIGIE W DIALOGU KULTUR. 500 LAT REFORMACJI	2017-09-14 - 2017-09-16	Toruń	Wydział Politologii i Studiów Międzynarodowych	więcej»
2nd International scientific conference "Positive management and leadership in socially responsible organisations"	2017-09-14 - 2017-09-14	Toruń	Katedra Doskonałości Biznesowej, WNEiZ UMK	więcej»
Forensically important Diptera. Identification workshop	2017-09-11 - 2017-09-15	Toruń	Wydział Biologii i Ochrony Środowiska	więcej»
VII konferencja z cyklu „Synchronia i diachronia w językach słowiańskich – zbliżenia i dialogi”, pt. „Języki słowiańskie w kontekstach kultur dawnych i współczesnych”	2017-09-07 - 2017-09-08	Toruń	Instytut Języka Polskiego	więcej»
Perspektywy rozwoju farmakognozji w XXI wieku	2017-09-07 - 2017-09-08	Bydgoszcz	Katedra Farmakognozji, Wydział Farmacji, Collegium Medicum	więcej»
XV Ogólnopolskie Seminarium Naukowe Profesora Zygmunta Zielińskiego Dynamiczne Modele Ekonometryczne 2017	2017-09-05 - 2017-09-07	Toruń	Wydział Nauk Ekonomicznych i Zarządzania	więcej»
Mikrobiologia środowiskowa szansą bezpiecznego życia i postępu biotechnologicznego	2017-09-05 - 2017-09-08	Ciechocinek	Zakład Mikrobiologii	więcej»
Pollen Monitoring Programme, 11th Meeting and Workshop	2017-08-28 - 2017-09-05	Toruń, Brodnica, Białowieża	Instytut Archeologii, Wydział Nauk Historycznych	więcej»
Niejednorodne modele kosmologiczne	2017-07-02 - 2017-07-07	Piwnice k. Torunia	Centrum Astronomii	więcej»
Krajowy raport Banku Światowego „Systematyczna diagnoza dla Polski”	2017-06-29 - 2017-06-29	Toruń	Ośrodek Studiów Fiskalnych UMK	więcej»

Analiza współpracy naukowej

WSPÓŁPRACA
MIĘDZYNARODOWA
DLA BIZNESU

Strona główna > Badania naukowe > Konferencje > Konferencja

KONFERENCJA



XIV zjazd Polskiego Stowarzyszenia Psychologii Społecznej

15.09.2017 - 17.09.2017 - Toruń

Katedra Psychologii, Wydział Humanistyczny
87-100 Toruń, ul. Gagarina 39
(056) 611-36-55

Współorganizator:

Polskie Stowarzyszenie Psychologii Społecznej

Kierownik naukowy:

prof. dr hab. Maria Lewicka
e-mail: marlew@umk.pl

Sekretarz naukowy:

mgr Lilianna Jarmakowska-Kostrzanowska
e-mail: lkostrzanowska@umk.pl
telefon: (056) 611-36-55

Podczas zjazdu odbędą się prezentacje badań oraz dyskusje panelowe dotyczące różnych dziedzin psychologii społecznej - takich jak psychologia wartości, pamięć społeczna, relacje międzygrupowe, agresja, wpływ i spostrzeganie społeczne, zachowania prospołeczne oraz badania międzykulturowe. Głównym wątkiem zjazdu będzie psychologia środowiskowa i prośrodowiskowa, w tym dyskusja z udziałem praktyków. Wykłady wygłoszą zaproszeni goście zagraniczni: Prof. Shalom Schwartz i Prof. Robert Gifford.

[«powrót](#)

Analiza współpracy naukowej

1 "Cykl wykładów otwartych ""W szkole miłosierdzia chrześcijańskiego - Uczynki miłosierdzia wobec ciała Kościół wobec wyzwań XXI wieku (styczeń - maj). W szkole miłosierdzia chrześcijańskiego - Uczynki miłosierdzia wobec ciała (październik - grudzień)."

2 Miejsce Józefa Chałasińskiego w socjologii polskiej. W stulecie urodzin Celem konferencji jest ocena miejsca Józefa Chałasińskiego w socjologii polskiej oraz przegląd badań nad ewolucją wsi i zmianami w perspektywach życiowych młodzieży.

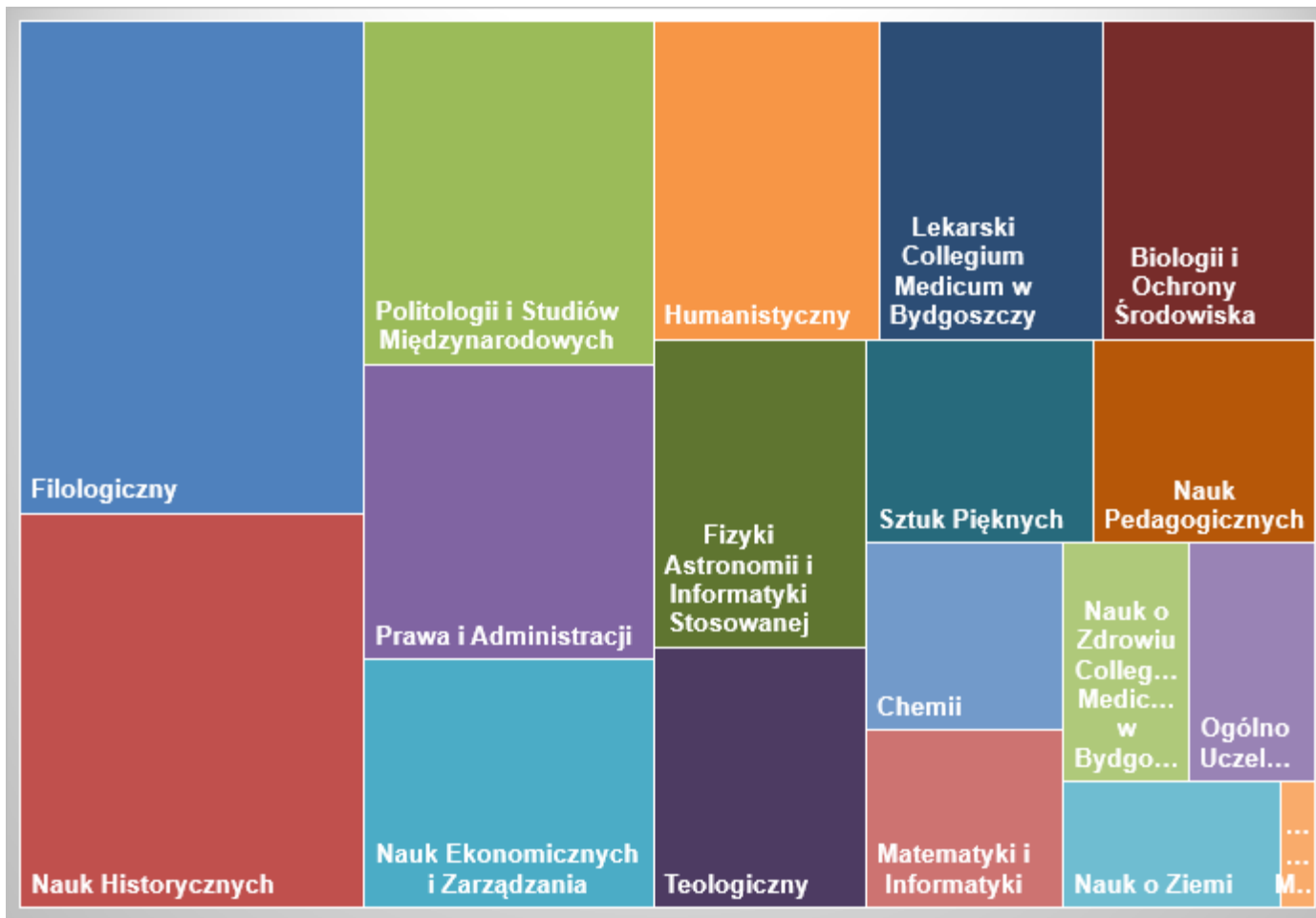
3 Społeczeństwo inwigilowane w państwie prawa? Granice ingerencji państwa w sferę praw jednostki

4 "Ogólnopolska Konferencja Naukowa Studentów i Absolwentów Konserwacji Zabytków ""KONSERWACJA ZABYTKÓW - STUDIA I PRAKTYKA"" Konferencja naukowa stanowi interdyscyplinarną platformę wymiany doświadczeń konserwatorów zabytków. Celem imprezy jest przedstawienie akademickiego modelu konserwacji zabytków wraz z ukazaniem realizacji będących przykładem codziennej praktyki. W trakcie spotkania prezentowane są prace, wykonane zarówno przez studentów i pracowników naukowych uczelni i instytucji związanych z ochroną dziedzictwa kulturowego, jak i przez aktywnych zawodowo konserwatorów."

5 Prezentacja dotychczasowych działań związanych z tworzeniem programu konserwatorsko-restauratorskiego dla Katedry świętych Jana Chrzciciela i Jana Ewangelisty w Toruniu Seminarium służy przedstawieniu dotychczasowych działań zmierzających do sformułowania całościowego programu prac konserwatorskich i restauratorskich dla Katedry. Uczestnicy zamierzają podjąć próbę znalezienia odpowiedzi na dwa ważne pytania: 1) Jakie jeszcze działania należy podjąć, aby zgromadzić wiedzę niezbędną do opracowania spójnego, dalekosiężnego programu konserwatorsko-restauratorskiego w Katedrze?, 2) Jakie są nasze wizje Katedry?

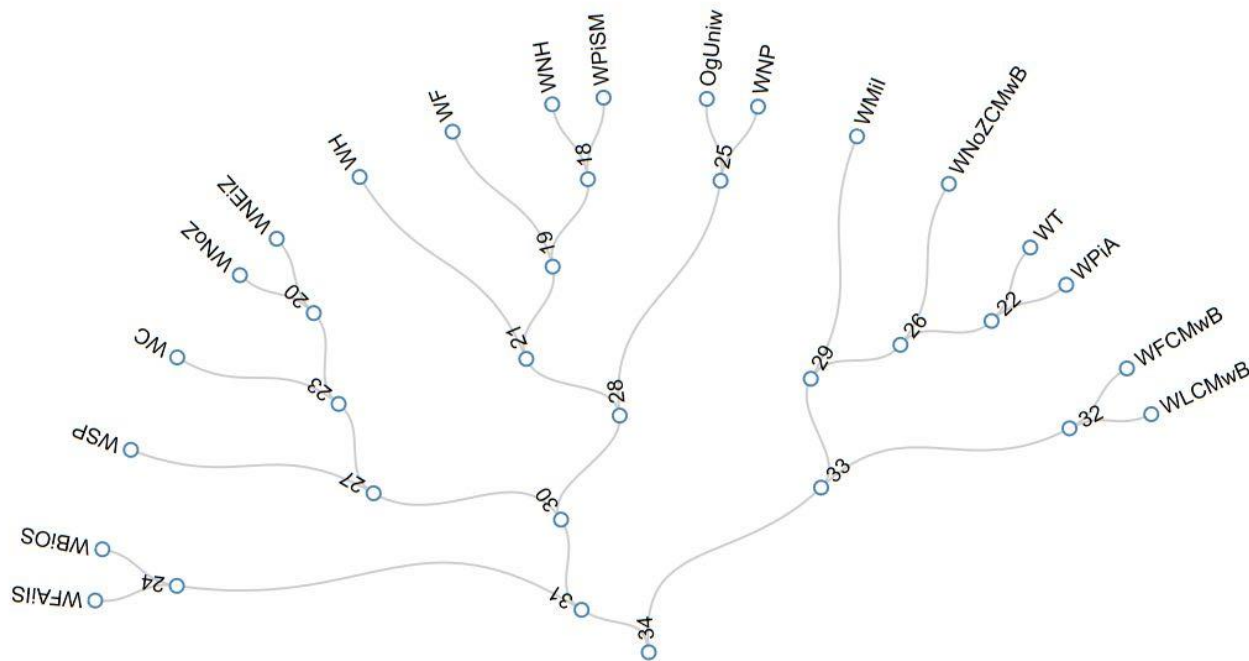
6 Harmonizacja wybranych dziedzin polskiego prawa ze standardami Unii Europejskiej Celem konferencji jest przedstawienie wybranych zagadnień prawa polskiego w kontekście porównawczym ze standardami Unii Europejskiej. Referentami będą osoby studiujące na studiach doktoranckich na WPiA UMK, którzy włączą się w ten sposób do ogólnej dyskusji na temat stanu przygotowania Polski do przystąpienia do Unii ...

Analiza współpracy naukowej



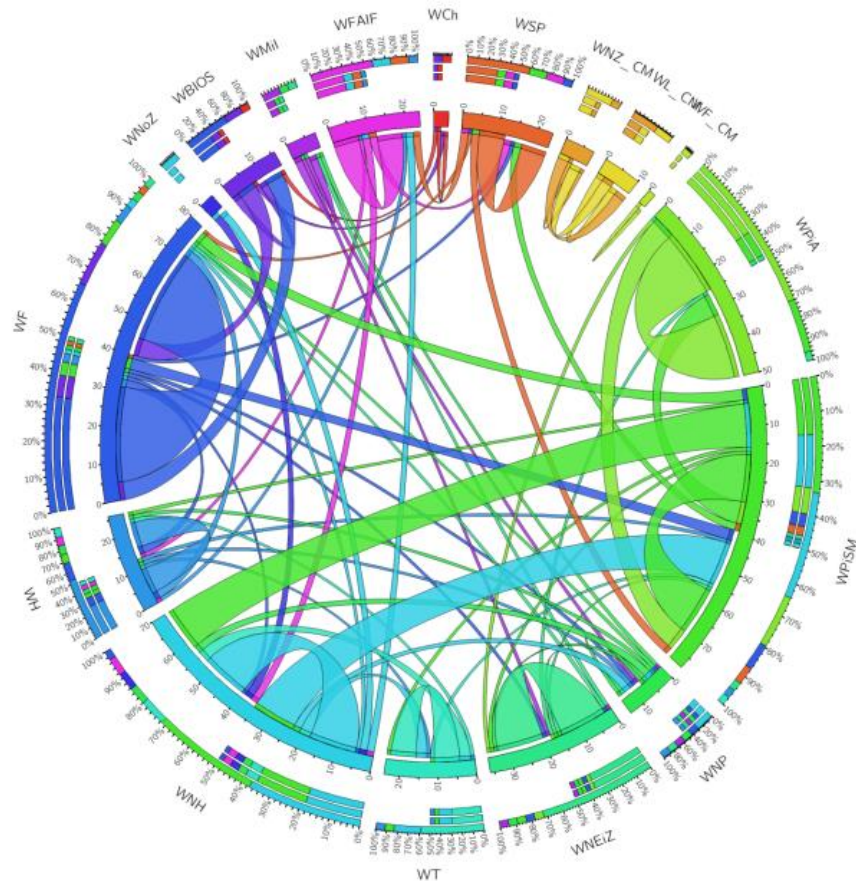
Tree Map liczby konferencji organizowanych przez Wydziały. Opracowanie własne

Analiza współpracy naukowej



Dendrogram – podobieństwo opisów konferencji organizowanych przez Wydziały. Opracowanie własne

Analiza współpracy naukowej



Graf współpracy pomiędzy Wydziałami. Opracowanie V. Osińska



METADANE - PROBLEMY

- Problemy z jakością metadanych występują w całym spektrum metainformacji o publikacjach i dokumentach:
 - w polu typ,
 - w polu prawa,
 - w polu data,
 -
- Problemy z formą prezentacji metadanych

```
<dc:title xml:lang="pl"><![CDATA[Brudnopisy listów Eugonii Brończykowej do Czerwonego Krzyża w Moskwie i Ministra Spraw Zagranicznych, Wiaczesława Mołotowa w sprawie jej syna [nazwa red.].]]></dc:title>
```

```
<dc:type xml:lang="pl"><![CDATA[Brudnopisy listów Eugonii Brończykowej do Czerwonego Krzyża w Moskwie i Ministra Spraw Zagranicznych, Wiaczesława Mołotowa w sprawie jej syna [nazwa red.].]]></dc:type>
```

```
<dc:type xml:lang="pl"><![CDATA[33 x 21 cm]]></dc:type>
```

```
<dc:type><![CDATA[Rybicki, Leszek]]></dc:type>
```

```
<dc:type><![CDATA[23.07.1989]]></dc:type>
```

- domena publiczna - Bogurodzica powstała we wczesnym średniowieczu, chociaż dokładna data jest nieznana. przeważają hipotezy badaczy wskazujące na xi lub xii w. jako czas jej powstania. pierwszy zapis tekstu jest późny, z początku xv w. (ok. 1407), wcześniejsze zapisy mogły zagiąć, ale też tekst mógł krążyć w obiegu ustnym.

- domena publiczna - Bolesław Leśmian zm. 1937
- domena publiczna - Bolesław Leśmian zm. 1938
- domena publiczna - tłum. Leon Ulrich zm. 1885
- domena publiczna - tłumacz Adam Mickiewicz zm. 1855
- domena publiczna - tłumacz Leon Ulrich zm. 1885

Nazwy wydawców wyodrębnione z rekordów bibliograficznych BN za lata 1997-2017

Nazwa wydawcy	Liczba zarejestrowanych książek
Wydawnictwo C. H. Beck	4325
C. H. Beck	1186
Wydawnictwo Amber	2365
"Amber"	2111
Wydawnictwo Helion	3886
"Helion"	955
Wydawnictwo Adam Marszałek	2212
Wydaw. Adam Marszałek	1086

► Indeks: Typ zasobu Wyników: 80

- film
- fotografia
- fotografia
- fotografia
- Fotografia
- fotografoa

► Indeks: **Temat i słowa kluczowe** Wyników: 31

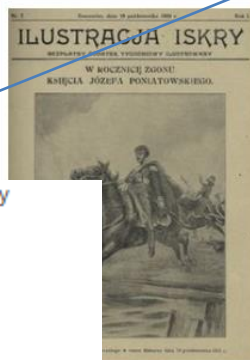
- 1813
- 1814
- 1815
- 1816
- 1817
- 1818
- 1819
- 1820
- 1821

Wybierz pozycje z listy

- * almanach
- * almanachy
- * analiza
- * anegdoty
- * antologia
- * antologie
- * apele
- * apologetyka
- * archiwalia
- * argument
- * article
- * artykuł
- * artykuły
- * artykuły historyczno-prawne
- * atlas
- * atlasy
- * atlasy historyczne
- * bajka
- * bajki
- * ballady

► Szukana fraza: **[Temat i słowa kluczowe = 1813]** Wyników: 31

FILTRUJ WYNIKI ►



OPIS

INFORMACJE

Tytuł:

Ilustracja Iskry. Bezplatny dodatek tygodniowy ilustrowany 1924, nr 7

Data wydania:

1924

Typ zasobu:

czasopismo

Mniej ^

Temat i słowa kluczowe:

Poniatowski, Józef (1763-1813)

trzy rozwiązania standaryzacyjne z zakresu tworzenia bibliotek cyfrowych z lat 2002-2006:

- NISO 2004,
- Strategia Europejska I2010,
- Strategia Zespołu ds. Standardów dla Bibliotek Naukowych

A. Domagalska - w żadnym z nich nie poświęcono wcale uwagi kwestiom oceny jakości bibliotek cyfrowych, a tym samym pominięto również kwestie jakości metadanych¹

1. Domagalska A. (2006). Problemy jakości metaopisów w bibliotekach cyfrowych – II Krajowa Konferencja Naukowa Technologie Przetwarzania Danych [online] [05.01.2017], http://www.cs.put.poznan.pl/kkntpd/tpd_pliki/publikacja/pub/55.pdf

M. Werla - postulat stosowania słowników lub kartotek haseł wzorcowych przy tworzeniu metadanych w celu osiągnięcia odpowiedniego poziomu ich interoperacyjności. A w przypadku danych typu data, gdzie trudno wykorzystać słowniki zamknięte - normalizacja ich zapisu, np. w notacji RRRR-MM-DD.

- postulat zaprojektowania schematu metadanych tak, aby możliwe było automatyczne wyodrębnienie określeń przestrzennych czy czasowych¹

1. Werla M. (2009). Wykorzystanie metadanych z polskich bibliotek cyfrowych

Metadane z bibliotek – mapowanie MARCxx na współczesne formaty:

```
{"fieldTag":"r","marcTag":"300","ind1":"","ind2":"","subfields":[{"tag":"a","content":"270, [2] s."},{tag":"b","content":"il. (w tym kolor.), faks., fot., 1 mapa, portr. ;},{tag":"c","content":"20 cm."}]}
```

```
{"fieldTag":"a","marcTag":"100","ind1":"1","ind2":"","subfields":[{"tag":"a","content":"Urbanek, Mariusz"}]}
```

```
{"700":{"ind1":"1","ind2":"","subfields":[{"a":"Dudek-Bujarek, Teresa."}]}}
```

```
{"700":{"ind1":"1","ind2":"","subfields":[{"a":"Filip, Elżbieta Teresa."}]}}
```

```
{"700":{"ind1":"1","ind2":"","subfields":[{"a":"Kenig, Piotr."}]}}
```

```
{"700":{"ind1":"1","ind2":"","subfields":[{"a":"Haftarczyk, Patrycja."},{e:"Tł."}]}}
```



Uniwersytet
Wrocławski

LITERATURA

1. Idzik P., *Analiza Big Data. Badania niereaktywne w erze Internetu 2.0*, w: Radmoski A., Bomba R. (red.), *Zwrot cyfrowy w humanistyce*. Lublin 2013, e-naukowiec, s. 153-168. [dostępne online]: http://e-naukowiec.eu/wp-content/uploads/2013/05/Zwrot_cyfrowy_w_humanistyce.pdf
2. Domagalska A. (2006). Problemy jakości metaopisów w bibliotekach cyfrowych – II Krajowa Konferencja Naukowa Technologie Przetwarzania Danych [online] [04.11.2018], http://www.cs.put.poznan.pl/kkntpd/tpd_pliki/publikacja/pub/55.pdf
3. Werla M. (2009). Wykorzystanie metadanych z polskich bibliotek cyfrowych W: C. Mazurek, M. Stroiński, J. Węglarz (red.). *Polskie Biblioteki Cyfrowe 2010*. Materiały z konferencji zorganizowanej w dniach 20-21 października 2010 roku przez: Bibliotekę Kórnicką PAN, Poznańską Fundację Bibliotek Naukowych, Poznańskie Centrum Superkomputerowo-Sieciowe [online] Poznańskie Centrum Superkomputerowo-Sieciowe, Poznań 2011, s. 125-129 [04.11.2018] <http://lib.psnc.pl/Content/376/BC-22-Werla.pdf>

PODZIĘKOWANIA

1. Badania przeprowadzono w ramach grantu NCN 2013/11/B/HS2/03048 *Badanie struktury i dynamiki cyfrowych zasobów wiedzy za pomocą metod wizualizacji*, UMK
2. Badania przeprowadzono w ramach grantu NCN 2016/23/B/HS2/01323: *Metody i narzędzia lingwistyki korpusowej w badaniach bibliografii polskich wydawnictw zwartych z lat 1997-2017*, UW r