

IFLA Satellite Pre-Conference of the Classification and Indexing Section
“Looking at the Past and Preparing for the Future”
Florence, Italy, 20-21 August 2009



Język Haseł Przedmiotowych Biblioteki Narodowej (National Library of Poland Subject Headings) – from card catalogs to digital library.

Some questions about the future of a local subject headings system in the changing world of information retrieval

Wanda Klenczon

National Library of Poland, Warsaw

JHP BN (National Library of Poland Subject Headings) is an indexing language developed and maintained by the National Library of Poland, used in the Polish current bibliography, the NL catalogs, in the OPAC of majority of Polish public and educational libraries and also in many other libraries (research, academic, school libraries), it is also used as a tool in numerous bibliographic databases available online such as regional bibliographies and indexes of periodicals¹.

The subject cataloging has a long tradition in Poland. For the first time theoretical backgrounds and guidelines for subject cataloging (they are consistent with international standards and instructions accepted today) were established in the 1920's and also a first subject catalogue in Poland was created then². These guidelines were adopted in the early 1950s when the staff of National Library of Poland made a decision to built its own subject

headings system with the aim of indexing the publications registered in the Polish national bibliography and in NL's catalogues. The subject catalog – card-based, and later automated – became the National Library's central descriptive catalog in 1969. Two decades later, in 1989, the National Library published the first edition of *Słownik języka haseł przedmiotowych Biblioteki Narodowej* [Subject Headings Thesaurus of the National Library – referred to below as *Thesaurus*], that contained ca 9 000 headings (topics and subdivisions)

JHP BN is a poly-hierarchical system based on main headings and subdivisions organized in the network of semantic relationships among terms (hierarchical and associative references). The vocabulary covers all subject areas, and the pre-coordinated indexing enables to express complex subjects. The printed *Thesaurus* (published every four years) contains topical headings. Names of persons, corporate body names, uniform titles and geographic names are controlled through an authority file in the MARC21 format, which has been successively enlarged since the mid-1990s. All types of headings are created in compliance with the relevant international or national standards and instructions. In 1995 the list of subject headings was converted into machine-readable form and was transferred into the integrated library system (INNOPAC/Millennium). Since then JHP BN authority file is available online, free of charge and without any restrictions³.

The authority file is updated weekly. Currently it contains over 123 000 terms (including ca 40 000 topics, 25 000 geographical names and 1500 topic and formal subdivisions). About 14 000 new records are created annually (including ca 3 500 topics), authority file of JHP BN strings contains more than 550 000 records and is rising at an annual rate of ca 55 000.

The JHP BN subject headings system was primarily designed as a controlled vocabulary for indexing the books from the National Library of Poland collection, gradually its usage was getting wider and covered other types of documents: periodicals and other

continuing resources, articles from serial publications (newspapers and scientific magazines), sound and videorecordings, motion pictures, printed music, cartographic material, two-dimensional graphic representations, artifacts, electronic documents and ephemera.

At the moment National Library of Poland OPAC⁴ provides access to 1,3 million of bibliographic records related to various types of documents (as of June 25, 2009), this entire collection is cataloged using JHP BN. Last year more than 100 000 documents in our NL were indexed with JHP BN headings (including systematic retroconversion of card catalogues).

What is the future of subject cataloging in National Library of Poland?

Although JHP BN is well-known and used by Polish librarians, we must be aware of changing users needs, of an impact of “googleization” on their search strategies and of a diminishing role of subject indexes.

Subject searching is the most problematic and controversial of all research types in the OPAC. Good subject analysis is difficult and expensive, subject headings that reflect individual knowledge, attitudes and preferences of indexers can be subjective, the rules are complicated and sometimes ambiguous. Subject strings are often incomprehensible for end-users and the complex rules are troublesome also for reference librarians. Thanks the popularity of search engines such as Google, keyword searching - easy, quickly and effective (although its effectiveness is illusory) - is on the rise. As indexing specialists we must try to recognize users' needs and establish subject access points that can be achieved by the For indexers, who have no public contact, especially useful would be analysis of transaction logs in library's OPAC. An examination of indexes use and search results in NL of Poland OPAC shows that an estimated 20-25% of library catalog searches are subject searches⁵.

Indexes Used and Search Results		
From Tuesday 16 June 03:01AM, to Thursday 25 June 03:00AM		
	Number Done	Percent Done
AUTHORS	68,583	23.50%
CALL NOS	2,093	0.72%
SUBJECTS	63,730	21.83%
PUBLISHERS	512	0.18%
ISN'S	10,031	3.44%
TITLES	101,229	34.68%

Search Results for User Keyed Searches - SUBJECTS		
From Tuesday 16 June 03:01AM, to Thursday 25 June 03:00AM		
	Number Done	Percent Done
Searches Retrieving 1 Record	32,142	50.43%
Searches Retrieving 2 to 8 Records	5,287	8.30%
Searches Retrieving 9 to 30 Records	3,926	6.16%
Searches Retrieving 31 to 99 Records	3,891	6.11%
Searches Retrieving 100 to 499 Records	4,888	7.67%
Searches Retrieving 500 to 4999 Records	3,209	5.04%
Searches Retrieving 5000 or More Records	498	0.78%
Total Searches Retrieving Records	53,841	84.48%
Total Records Retrieved/Average per Search	10,957,707	203
Searches With No Retrievals	9,889	15.52%
Total Searches	63,730	100.00%

The end-user retrieve zero result in 15-20% of their searches. The high rate of successful searching is probably caused by a long tradition of subject indexing in our NL: most librarians in Poland know this indexing tool and are aware that entire NL collection available via OPAC is indexed using JHP BN. Unfortunately, transaction logs cannot answer the questions about OPAC users' satisfaction or dissatisfaction with subject headings, subject indexes and information retrieval....

The JHP BN team is examining transaction logs with the aim of identification subject searches that yielded no hits. What else can we observe?

In Order by Popularity - SUBJECTS	
From Monday 22 June 03:06AM, to Wednesday 01 July 03:00AM	
Page 21 of 2428 pages	
Search	Total Searches
hokej	19
historyzm	19
godzina wychowawcza	19
film	19
film dokumentalny	19
dziecko w wieku	19
domeyko	19
bizancjum	19
bazylianie	19

The users rarely utilize complete subject terms, they are unaware of available tools (alphabetical list of suggested subject terms, references, advanced searching) and usually use separate words or combination of two words. The subject index is being used as a simple keyword research tool. This conclusion is not very original, but in this context we must discuss the usefulness of pre-coordinated languages.

The future of indexing, based on of pre-coordinated subject headings strings, is at present a matter of a discussion among LIS specialists, also in Poland. We followed the discussion about LCSH future in context of pros and cons of pre- versus post-coordination⁶. JHP BN team observes with an interest the project FAST (based on the existing vocabulary LCSH, but adapting them in a faceted schema with a simplified syntax). In many opinions subject headings are unable to satisfy information needs of our contemporaries: the syntax is too complex and not enough flexible, users don't understand subject headings and prefer keyword searching, although many of separate concept combinations could be incorrect or irrelevant. Moreover, the management of subject headings authority files is hard and expensive.

We are aware of all these common problems and we analyze obstacles and disadvantages of subject searching our NL OPAC. A lot of JHP BN subject strings rises at a

tremendous rate and very often a new string is used just once for a single work and never again. It is disadvantageous for indexes users. Nevertheless in our opinion the pre-coordinated strings ensure better relevance, tend to low precision and can be useful for both browsing and keyword searching. JHP BN staff and Development Consulting Team decided to continue the current policy, and making our indexation tool easier to use and maintain seems to be a key issue.

What we intend to do?

- in 2009-2010 make a general revision of JHP BN vocabulary: systematic review and evaluation of all topics and subdivisions, and – if necessary - modification. The terminology should be more current and consistent, and - for the inexperienced users- more intuitive
- complete the alphabetical list of terms with the hierarchical list of preferred terms. It will facilitate the control of subject vocabulary and management of relationships (until now we prepared the hierarchical list of religious, laws and educational terms)
- simplify syntax and application rules. Review of rules of attribution of chronological and formal subdivisions seems to be useful in order to make the syntax easier

The time aspect in subject cataloging is an important element of subject analysis.

In order to include time aspect JHP BN uses years and periods: headings and subdivisions consist of either a single date or a starting and ending date, chronological subdivisions for 20th and 21th centuries reflect mostly political and economical reality (1914-1918, 1918-1939, 1939-1945, 1945-1989)

Examples:

651 |a Polska |y 16-18 w

651 |a Warszawa |y od 1989 r.

651 |a Polska |x kultura |y 1918-1939 r.

650 |a Bitwa 1561 r. na Równinie Kawanakajima

In the order to provide greater accuracy and make subject retrieval more consistent, it seems to be useful to introduce the relevant dates. We discuss the use of free-form chronological terms and subdivision, that may reflect with accuracy time period covered by the content of the item. JHP BN team decided to continue the current practice, but the use of free-form chronological terms as supplementary subject access point in 648 field (that may consist of only a single date or a date range) will be analyzed.

For example, statistic of Warsaw in 1993 could be indexed as:

651 |a Warszawa |y od 1989 r.|v statystyka

648 |a 1993

and statistic of Warsaw from 1993 to 1996 could be indexed as

651 |a Warszawa |y od 1989 r.|v statystyka

648 |a 1993-1996

Other solution, that is also being discussed, is a complementary use of standardized code in the 045 field – Time Period of Content, in which specific time period may be recorded in the pattern year-month-day (month and day are optional) and preceded by a code for the era. For the present this solution cannot be used, because the library system doesn't support the retrieval of information from coded field yet.

The JHPBN staff also decided, that the indexing terms *Forme/Genre* can be subdivided chronologically by decade

655 |a Fotografia gabinetowa |z Polska |y 1870-1880

655 |a Miedzioryt |y 1630-1650

Regarding form subdivision, we are discussing their usefulness in subject strings – are they really needed and exploited in information retrieval?

As example, a fragment of NL subject headings index:

Muzyka polska – historia – 19-20 w.

Muzyka polska – historia – 19-20 w. - **bibliografia**

Muzyka polska – historia – 19-20 w. – **ćwiczenia i zadania dla szkół muzycznych**

Muzyka polska – historia – 19-20 w. – **dokument elektroniczny**

Muzyka polska – historia – 19-20 w. – **katalog wystawy**

Muzyka polska – historia – 19-20 w.- **materiały konferencyjne**

Muzyka polska – historia – 19-20 w. - **pamiętniki**

Muzyka polska – historia – 19-20 w. - **publicystyka**

Muzyka polska – historia – 19-20 w.- **szkice**

Muzyka polska – historia – 19-20 w.- **wydawnictwa dla młodzieży**

Muzyka polska – historia – 19-20 w.- **wydawnictwa popularne**

Muzyka polska – historia – 19-20 w.- **źródła**

Are all these information about specific kind or genre of material important and useful for anyone searching for documents about Polish music in 19th-20th centuries? In my opinion they distract bibliographic information in OPAC. We intend to identify the form headings and form subdivisions that currently exist in JHP BN authority file and move them to form/genre indexing term in the 655 field (authority record with MARC21 tag 155). It will simplify the subject headings strings, reduce their number and simplify the subject index. All important information about form/type of document or its genre will be accessible through the keyword searching. These projects will be discussed with librarians representing different types of libraries using JHP BN.

Digital library - subject headings as source of keywords

The syntax simplification is very important in the context of web indexing and the creation of keyword metadata based on the controlled vocabulary. NL of Poland collects and archives some domestic public online documents in the library's internet archive, but at the moment our experience is still modest. We are more advanced in indexing publications in a digital library. National Digital Library Polona⁷ was created in 2007, its mission is to enhance wide and easy access to the digital collections of the NL of Poland, including the most important editions of literature and scientific materials, historical documents, journals, graphics, photography, score and maps.. All the documents presented in CBN Polona are being described primarily in NL catalogue database and are linked (through field 856 in MARC21 bibliographic record) to a catalogue description, that includes the complete information about publication, as well as the full subject access points.

The retrieval of publications in CBN is a simple search, based on keywords used as metadata in Dublin Core Element Set. These keywords are extracting from JHP BN subject headings strings used in bibliographic record in NL OPAC, automatically removed to DC Element: Subject and selected by librarians (the most significant and unique words are chosen). We are going to implement Dublin Core Metadata Set, lately translated into Polish. It will replace DC1.1 used in our digital library so far. DCMS offers more possibilities to describe digital documents, and also to index and create cross-references.

Subject headings and subdivision will move into following DCMS elements:

Topic → Subject

Geographic name or subdivision → Spatial Coverage

Date from chronological topic or subdivision → Temporal Coverage

Genre/Forme → Type

Other aspects may be included to elements such as Audience Education Level or Medium.

In the field of digitalization the principal aim of JHP BN team is to propose terms to describe objects of cultural heritage, such as manuscripts, rare books, graphics or photographs. Until now some NL special collection were neither indexed or classified. Primarily we will provide – in cooperation with potential end-users such as historians, art historians, students - terms for indexing of works of art and related images. We reviewed and analyzed the most important thesauri and classification used as indexing tools of cultural heritage. It turned out that all these tools were too complex for indexing of our collections and couldn't be easily adapted. We decided to add new terms to JHP BN authority file and complete them with equivalent terms from others indexing tools (Art&Architecture Thesaurus, Thesaurus for Use in Rare Book and Special Collection, Thesaurus of Graphic Materials, Iconclass). Thanks to this activities more varied, multilingual access to Polish digitalized cultural resources will be gradually provided.

150	%a Zmartwychwstanie
450	%a Wskrzeszenie zmarłych
550	%a <u>Eschatologia</u>
550	%a <u>Misterium paschalne</u>
550	%a <u>Paruzja</u>
550	%a <u>Zstąpienie do otchłani (ikon.)</u>
750	%a Resurrection %2 tgm

155	%a <u>Drzeworyt</u>
550	%a <u>Druk wypukły</u>
555	%w h %a <u>Drzeworyt chiaroscuro</u>
555	%w g %a <u>Grafika</u>
555	%w h %a <u>Ksylograf</u>
755	%a <u>Wood engravings tgm</u>
755	%a <u>Wood engravings (prints) %2 aat</u>

155	%a Godzinki (księgi)
555	%a <u>Brewiarze</u>
550	%a <u>Godzinki (modlitwa)</u>
550	%w h %a <u>Godzinki Sobieskich</u>
550	%w g %a <u>Księgi liturgiczne</u> %x katolicyzm
550	%w g %a <u>Modlitewnik</u> %x katolicyzm
555	%a <u>Psałterze</u>
550	%w h %a <u>Spitz hours</u>
550	%a <u>Śpiew</u>
550	%w h %a <u>Tres riches heures du duc de Berry</u>
755	%a Books of hours %2 rbgenr

Multilingual access has never been our principal claim. JHPBN has been created mostly for indexing the documents acquired as legal deposit and written or spoken in Polish, which isn't a popular language. Digital library changes our point of view on this matter: non-textual works or rare books, that can be discovered without language obstacles, need universal, multilingual access points.

We are interested in potential value of social tagging. Tags can complete controlled subject headings and may be additional subject access points.. The potential of "common knowledge" may be enormous and useful in many cases, especially in indexing the collection of images, sound recording or moving pictures, that are sometimes extremely difficult to be indexed with use of subject headings. We don't believe that social tagging in real future can replace indexing conducted by librarians (too many tags are of very low semantic value and in the absence of context are quite useless), however our attitude toward a folksonomy in our digital library is open.

JHP BN is a local indexing tool, used only in Polish libraries. The team is small (for many years it consisted of 3 persons only, lately it has 4-5 librarians), the budget is tight, the library system isn't perfect. Moreover, we must take into account technical and financial conditions not only in NL, but also in other libraries (sometimes very small and

underinvested) that use JHPBN. In these not very favourable conditions we are trying to make our subject headings system more flexible, more efficient and adaptable to changing information needs of users.

¹ The second most popular subject headings system in Poland is the KABA Subject Headings system, based on RAMEAU, developed since the early 1990s and used in 40 academic libraries.

² Łysakowski, Adam, *Katalog przedmiotowy* [The Subject catalog], vol. 1, Teoria, Wilno, 1928

³ <http://mak.bn.org.pl/w5.htm>. Since 2005, all JHP BN headings and authorized subject headings strings are available in the National Union Catalogue NUKAT www.nukat.edu.pl

⁴ http://alpha.bn.org.pl/screens/opacmenu_pol.html

⁵ The most popular index in NL OPAC is title index, two others – author and subject index -are used with comparable frequency – up to 25-30 % depending on the time of year

⁶ *Library of Congress Subject Headings. Pre- vs. Post-Coordination and Related Issues* [on-line]. http://www.loc.gov/catdir/cpsd/pre_vs_post.pdf. Seen May 11, 2009.

⁷ <http://www.polona.pl/dlibra>