Identifying candidates for the new subject headings based on the web behaviour of end-users

Busch, Kristýna
2012

Identifying candidates for the new subject headings based on the web behaviour of end-users

Author: Kristýna Busch
National Technical Library
Prague, Czech Republic

Contact: psh@techlib.cz

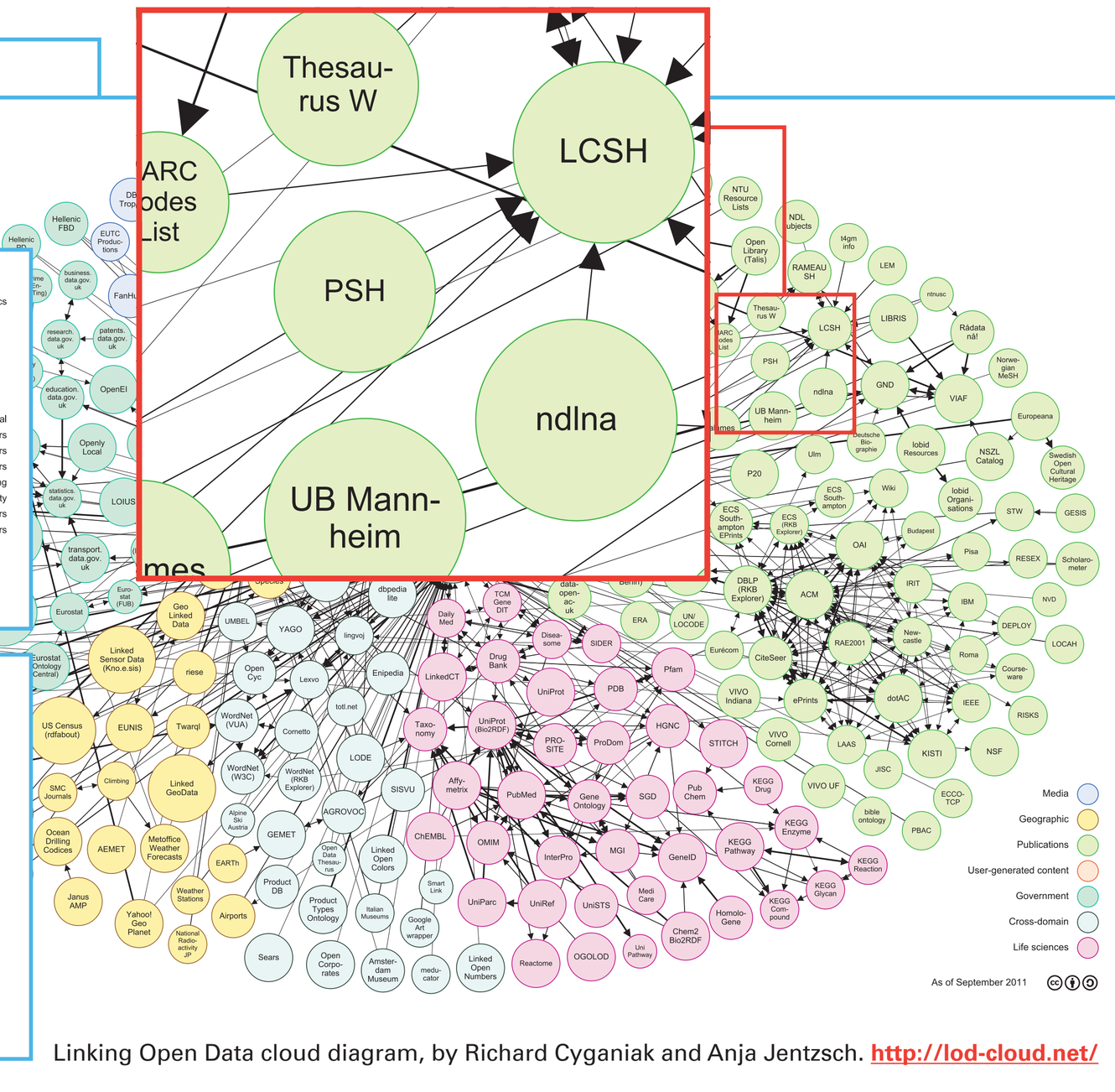
NTK
50°6'14.083"N, 14°23'26.365"E
Národní technická knihovna
National Technical Library

PSH

Subject heading **nuclear reactors** displayed at http://psh.ntkcz.cz/skos/

```
– <rdf:RDF>
    – <skos:Concept rdf:about="http://psh.ntkcz.cz/skos/PSH2459">
        <skos:prefLabel xml:lang="cs">jaderné elektrárny</skos:prefLabel>
        <skos:prefLabel xml:lang="en">nuclear power stations</skos:prefLabel>
        <skos:altLabel xml:lang="cs">atomové elektrárny</skos:altLabel>
        <skos:altLabel xml:lang="en">nuclear power plants</skos:altLabel>
        <skos:related rdf:resource="http://psh.ntkcz.cz/skos/PSH2397"/>
        <skos:related rdf:resource="http://psh.ntkcz.cz/skos/PSH2543"/>
        <skos:related rdf:resource="http://psh.ntkcz.cz/skos/PSH2560"/>
        <skos:related rdf:resource="http://psh.ntkcz.cz/skos/PSH6385"/>
        <skos:broader rdf:resource="http://psh.ntkcz.cz/skos/PSH2450"/>
        <foaf:page rdf:resource="https://cs.wikipedia.org/wiki/Jadern%C3%A9_elektr%C3%A1rny"/>
        <foaf:page rdf:resource="https://cs.wikipedia.org/wiki/Kategorie%3AJadern%C3%A9_elektr%C3%A1rny"/>
        <foaf:page rdf:resource="http://en.wikipedia.org/wiki/nuclear_power_stations"/>
        <skos:inScheme rdf:resource="http://psh.ntkcz.cz/skos"/>
      </skos:Concept>
  </rdf:RDF>
```

Subject heading **nuclear power stations** displayed in SKOS format.



Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/

Polythematic Structured Subject Heading System (PSH) is a bilingual Czech–English controlled vocabulary of subject headings which is used for organizing and searching the documents by subject. It is developed and maintained at the National Technical Library of Prague. PSH is a tree structure with 44 thematic sections. It contains more than 13,900 subject headings, which cover the main fields of human knowledge. Subject headings are included in a hierarchy of six (or seven) levels according to their semantic content and specificity. There are hierarchical, associative and equivalence relations in PSH.

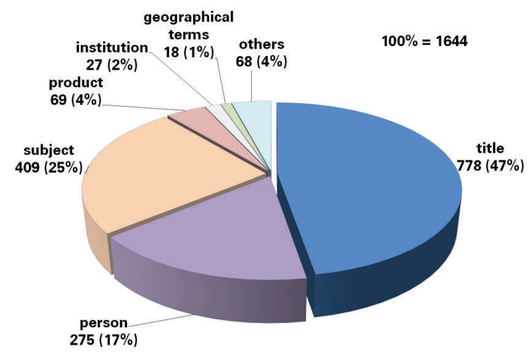
Objectives

National Technical Library of Prague carried out a Web server transaction log analysis in order to idetify possible candidates for the new subject headings in the Polythematic Structured Subject Heading System. The problem is that the language used in the subject heading system created by professionals may differ from the one used by end-users when searching various documents. Bringing both of the languages into the content of the subject heading system may enhance subject access to the National Technical Library´s materials and help users reach better results.

Methods

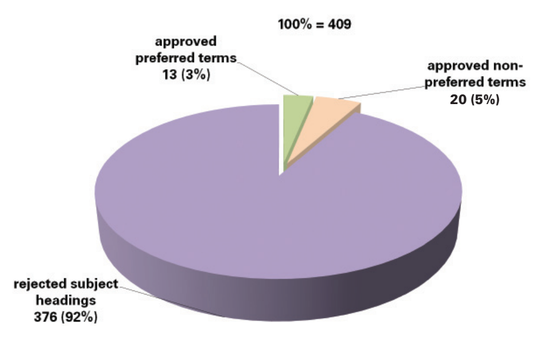
The first yet experimental set of transaction data was extracted from Aleph log files, while the other collection of data is now obtained through a Google Analytics service. Transaction log analysis helped pass through a large scale of data. On the other hand, Google Analytics method is now used for analysing data regularly which reduces the neccessity to process huge data sets. Moreover, Google Analytics provides easy access to an overview of users activity on the web site. This includes both quantitative and qualitative statistics of terms applied by users when searching for the documents. Despite the different methodological approaches, the transaction log analysis as well as Google Analytics method consist of three basic steps – (1) data collection, (2) data preparation (including cleaning up the Aleph data, removing duplicates, and previously processed terms, etc.), (3) data analysis. Within the data analysis, users queries are divided into seven categories that contain the title of the document, person, subject, product (e.g. Microsoft Windows), institution, geographical terms and others (document types, numbers, etc.). Then the candidates for new preferred terms and non-preferred terms are identified in the subject category.

Results



Search terms layout

This pie chart shows all search terms (1644) that were gathered from September 2011 to February 2012. These search terms are divided into seven categories.



The number of approved and rejected subject headings

This pie chart shows the number of approved preferred terms, non-preferred terms, and rejected subject headings. Reasons for rejecting the subject headings may include the vague category or category of too specific terms.

Conclusion

The initial purpose of this analysis was to identify potential candidates for the new subject headings. However, other findings were revealed during the analysis process. Since queries reflect users´ topics of interest, collected searched terms serve as an inspiration for the document acquisition. In addition, detailed observations indicate zero results which may be eliminated by extending the subject heading system and acquiring new documents. There were seventy eight suggestions for the new titles of the books and four titles of the journals for the acquisition department. These suggestions were gathered within the six month period (from September 2011 to February 2012). Hence, the analysis has an impact on both subject access to the documents and the document acquisition.

Next steps

The National Technical Library would like to focus on cooperation with other libraries using PSH in the future. That is why the log analysis in these institutions will be considered, in order to extend PSH in the areas that are not according to the profile of NTK´s library collection.