NTK

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

210 mm

# Integration of an Automatic Indexing System within the Document Flow of a Grey Literature Repository

Jindřich Mynarz, Ctibor Škuta
National Technical Library

Grey Literature 12 Conference, 7.12. 2010

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

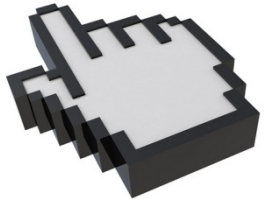
DIY
DO IT YOURSELF

210 mm

# Indexing of Grey Literature

- self-publishing, **self-indexing**
- the **Web** made publishing easier, can it make **indexing** easier as well?
- **make non-professional indexing better** through technology
- increase grey literature **visibility** and support **navigation** interfaces

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



210 mm

# Automatic Indexing

- conditional on **full-text** availability
- **machine learning** based on analysis of language corpora
- automatic **term assignment**
- automatic **suggestions of indexing terms** lessen the cognitive overhead involved in indexing
- **human feedback** to correct the obvious mistakes

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

DIY

210 mm

# Implementation

- **re-use** of existing components
  - combination and extension
- open source, open formats

subject headings system + digital repository + automatic indexer + text corpus + glue code

=

**automatic indexing system**

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

210 mm

# Subject Heading System

- **Polythematic Structured Subject Headings System**
  - universal Czech-English controlled vocabulary managed and used at the National Technical Library
  - expressed in RDF data format via SKOS vocabulary

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

INVENIO

210 mm

# Digital Repository

- **CDS Invenio**
  - open source, modular architecture
  - extensions to the **interface for entering new documents** and the **search interface**



# Automatic Indexer

- **Maui Indexer**
  - automatic **term assignment** with a controlled vocabulary
  - extensions for **Czech** language (stemmer, stopwords)
  - indexing model for **Czech** language with usage of **PSH**

# Text Corpus

- **National Repository of Grey Literature**
  - maintained by the **National Technical Library**
  - aggregates documents from **partner institutions**
  - in some cases, **metadata** are created by the users

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

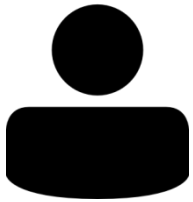

210 mm

# Glue Code

- code to **tie all pieces together**
- **web services**
  - loose coupling
  - re-use of existing code

NTK

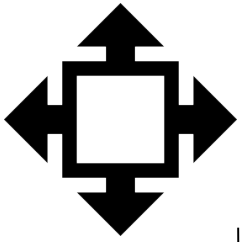
50°6'14.083"N, 14°23'26.365"E
Národní technická knihovna
National Technical Library



210 mm

# User Interface Design Considerations

- **opt-in** indexing procedure
- **suggest** indexing headings
- **autocomplete** headings' fragments
- **learn by example** — show example documents indexed with the heading in question
- **extending search** interface





# Further Possibilities and Challenges

- indexing must be **reflected in end-user interfaces**
- continuous **enhancements** of the individual parts of the document processing pipeline
- **user-generated** indexing
- **feeding back** into the development of the subject headings system

NTK

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

210 mm

# Thank you for your attention!

<mailto:jindrich.mynarz@techlib.cz>
<mailto:ctibor.skuta@techlib.cz>
<http://www.techlib.cz/en/>