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
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
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
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
Subject Heading System

• Polythematic Structured Subject Headings System
  o universal Czech-English controlled vocabulary managed and used at the National Technical Library
  o expressed in RDF data format via SKOS vocabulary
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**
  o maintained by the National Technical Library
  o aggregates documents from partner institutions
  o in some cases, metadata are created by the users
Glue Code

- code to *tie all pieces together*
- web services
  - loose coupling
  - re-use of existing code
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
Thank you for your attention!

<mailto:jindrich.mynarz@techlib.cz>
<mailto:ctibor.skuta@techlib.cz>