Dejan Pajić
Faculty of Philosophy, Novi Sad
Centre for Evaluation in Education and Science, Beograd

Scriptor
Bibliographic Information
Parsing Tool
Background

- Development of NCDb (SCIndeks) requires efficient, accurate, and thorough processing of citation data

- More powerful database searching capabilities

  - Evaluation of all subjects involved in scientific production

- Primarily journals and authors, but also affiliated institutions, publishers, proceedings, projects, and research teams

- Everything and everybody who gained financial support of the Ministry of Science

- Locating articles that cited particular author, paper or journal

- Locating related records (shared references)

- Linking with external and internal documents (DOI)
Requirements for parsing tool

- cannot be based purely on information extraction
  - there is no unnecessary data: every information is useful
- cannot relay on simple template mining
  - lack of standards in reference citing
- there is no need for completely autonomous system
  - relatively small number of journals are available online
    (some even cannot provide a preprint electronic form)
  - SCIndeks is intended to be an “error free” evaluation information resource
- has to provide standard text manipulation functions
  - manual editing, copying, converting from Cyrillic to Latin text, etc.
- has to be intuitive and user friendly
  - operators are not highly qualified computer experts, but well educated professionals in their own fields of interest
Implementation

- Scriptor was designed as Microsoft Word add-in entirely programmed in VBA
- all functions are available through:
  1. main menu (appearing on the menu bar)
  2. shortcut menu (displaying on right-click)
  3. toolbar


Algorithm (simplified)

- Template searching
  [12]### or ##/[90]#

- Template searching
  XXX+, X. or similar

- List of authors’ names
- List of titles’ first words

- List of journal titles
- Publishers list
- Towns list
- Common terms (tokens)
  journal, vol., no., pp., In:...

- Year is near the beginning of the reference
- Punctuation
  The most unreliable, and yet the most useful type of information
Difficulties

▪ extraction of citation information from footnote and endnote text is not supported

▪ some reference types (e.g. historical archives) are difficult to parse (semi) automatically

▪ sometimes authors use different styles within the same block of references
• Scriptor parses highly structured references (e.g. medicine - Vancouver) with minimal error rate
• expanding the lists of authors, journals, publishers, etc. is usually sufficient for Scriptor use with citation data from other scientific fields
• thorough list of tokens is easily translatable so to provide the support for parsing references in languages other than currently supported (Serbian and English)

• however, it is necessary to suggest detailed revision of some journals’ editorial processes