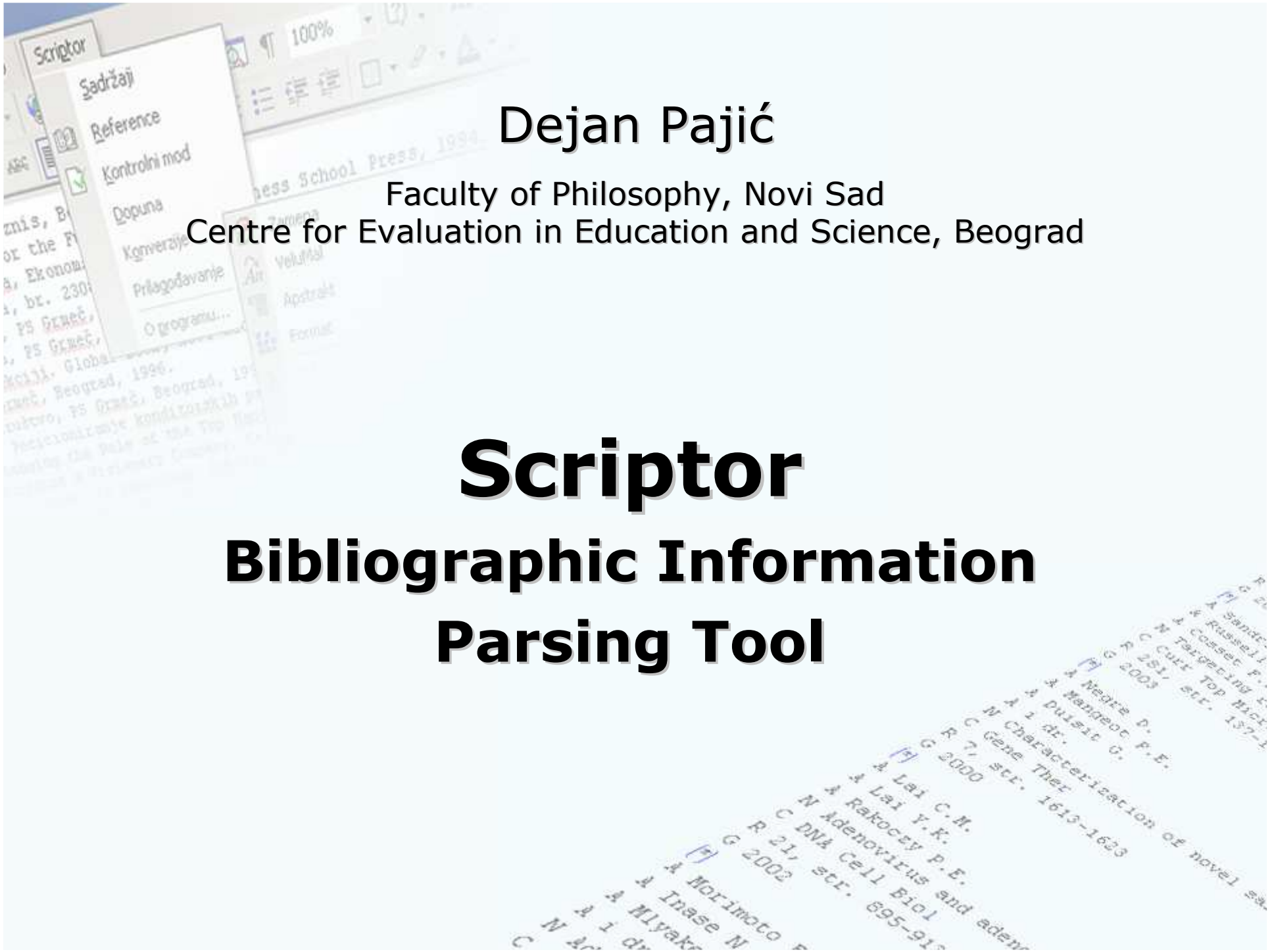



Dejan Pajić

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
- ✓ affiliated research
- ✓ everythin the Minis

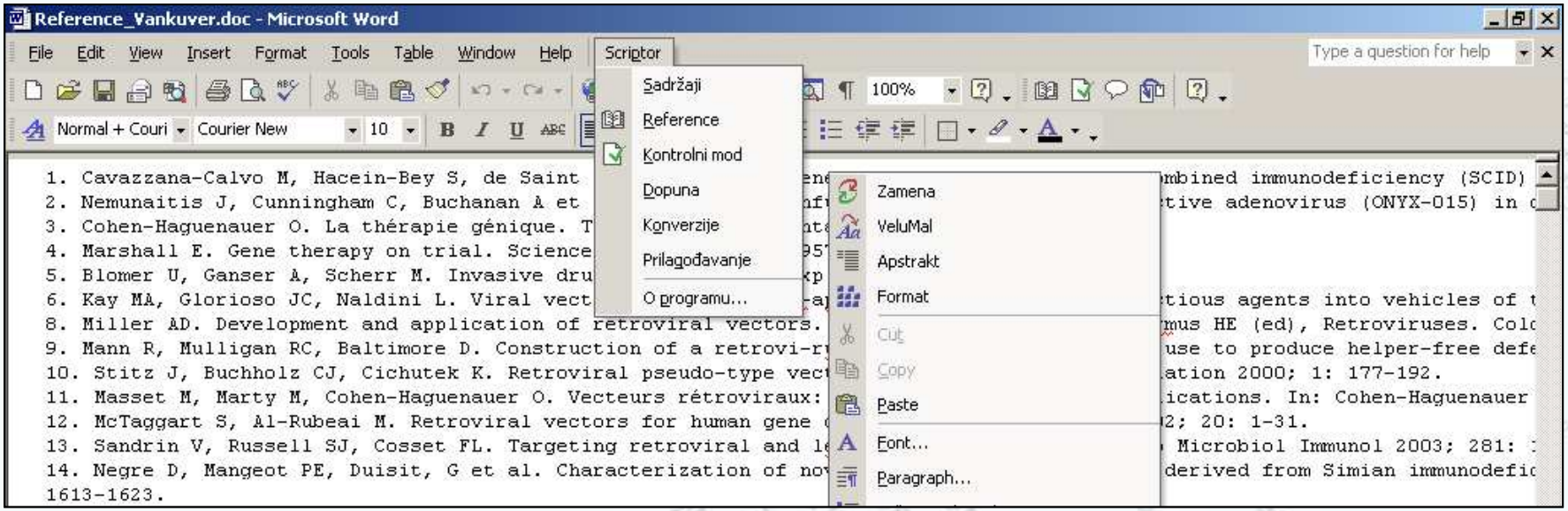
- ✓ locating articles that cited particular author, paper or
- ✓ locating related records (shared references)
- ✓ linking with external and internal documents (DOI)

# 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



# Example (block of references)


Stitz J, Buchholz CJ, Cichutek K. Retroviral pseudo-type vectors. Gene Therapy and Regulation 2000; 1: 177-192.

Selden RF. Transfection using DEAE-dextran. In Ausubel FM (ed), Current protocols in molecular biology. John Wiley & Sons, Inc., New York, 1996: 9.2.1-9.2.6.

McTaggart S, Al-Rubeai M. Retroviral vectors for human gene delivery. Biotechnol Adv 2002; 20: 1-31.

Sandrin V, Russell SJ, Cosset FL. Targeting retroviral and len-tiviral vectors. Curr Top Microbiol Immunol 2003; 281: 137-178.

Negre D, Mangeot PE, Duisit, G et al. Characterization of novel safe lentiviral vectors derived from Simian immunodeficiency virus (SIVmac251) that efficiently tranduce mature human dendritic cells. Gene Ther 2000; 7: 1613-1623.

Lai CM, Lai YK, Rakoczy PE. Adenovirus and adeno-associated virus vectors. DNA Cell Biol 2002 ; 21 : 895-913.

Morimoto E, Inase N, Mlyake S et al. Adenovirus-mediated suicide gene transfer to small cell lung carcinoma using a tumor-specific promoter. Anticancer Res 2001; 21: 329-331.

Nishikawa M, Huang L. Nonviral vectors in the new millennium: delivery barriers in gene transfer. Hum Gene Ther 2001; 12: 861-870.

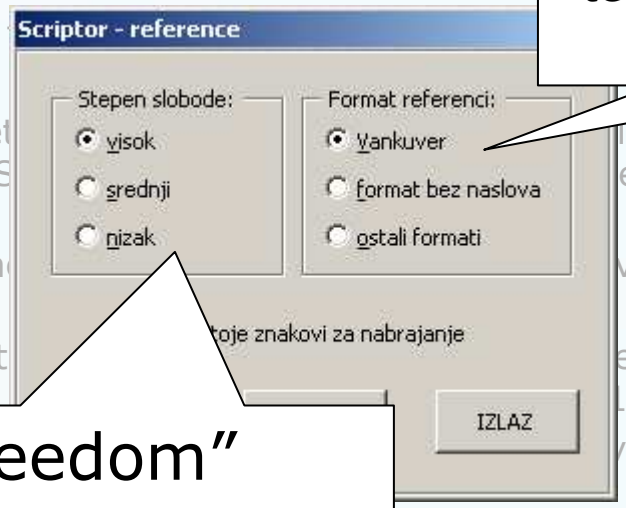
Nishitani MA, Sakai T, Kanayama HO, Himeno K, Kagawa S. Cytokine gene therapy for cancer with naked DNA. Mol Urol 2000; 4: 47-50.

Vacheri A, Pagano, JS. Infectious poliovirus RNA: A sensitive method of assay. Science 1965; 175: 434.

Masset M, Marty M, Cohen-Haguenauer O. Vecteurs rétroviraux: technologie, sécurité, applications. In: Cohen-Haguenauer O (ed), La thérapie génique. Technique & Documentation, Paris, 2001: 35-72.

# Example (options dialog)


Stitz J, Buchholz CJ, Cichutek K. Retroviral pseudo-type vectors
2000; 1: 177-192.
Selden RF. Transfection using DEAE-dextran. In Ausubel FM (ed
biology. John Wiley & Sons, Inc., New York, 1996: 9.2.1-9.2.6.
McTaggart S, Al-Rubeai M. Retroviral vectors for human gene d
31.
Sandrin V, Russell SJ, Cosset FL.
Immunol 2003; 281: 137-178.
Negre D, Mangeot PE, Duisit, G et
Simian immunodeficiency virus (S
Gene Ther 2000; 7: 1613-1623.
Lai CM, Lai YK, Rakoczy PE. Aden
: 895-913.
Morimoto E, Inase N, Miyake S et
carcinoma using a tumor-specific
Nish
Hum
Nish
nake
Vach
434.
Masset M, Marty M, Cohen-Haguenauer O. Vecteurs rétroviraux: technologie, sécurité, applications. In: Cohen-Haguenauer O (ed), La thérapie génique. Technique & Documentation, Paris, 2001: 35-72.

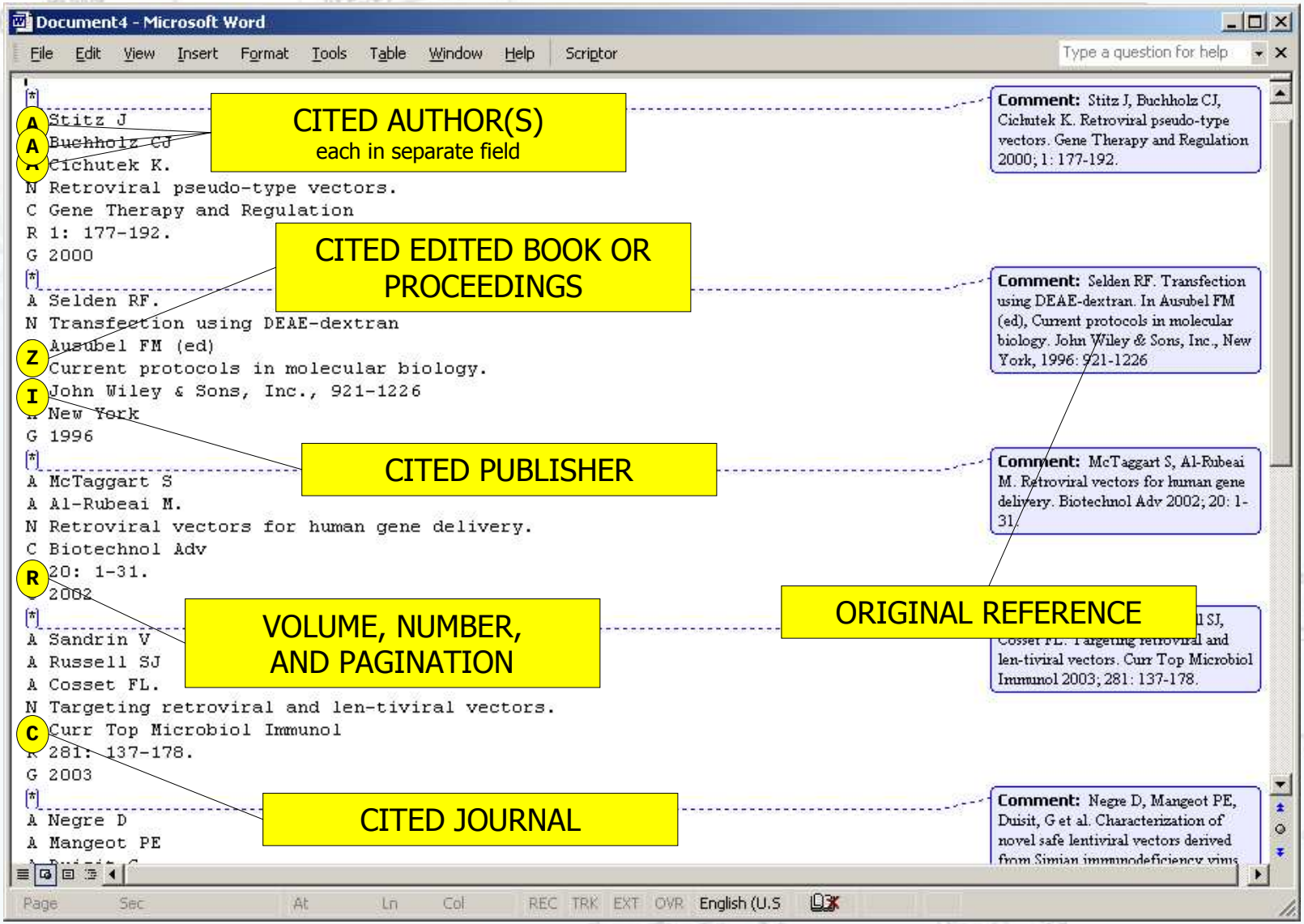

## reference format

by choosing one particular references format, less possible templates will be tested, thus parsing is faster and more accurate

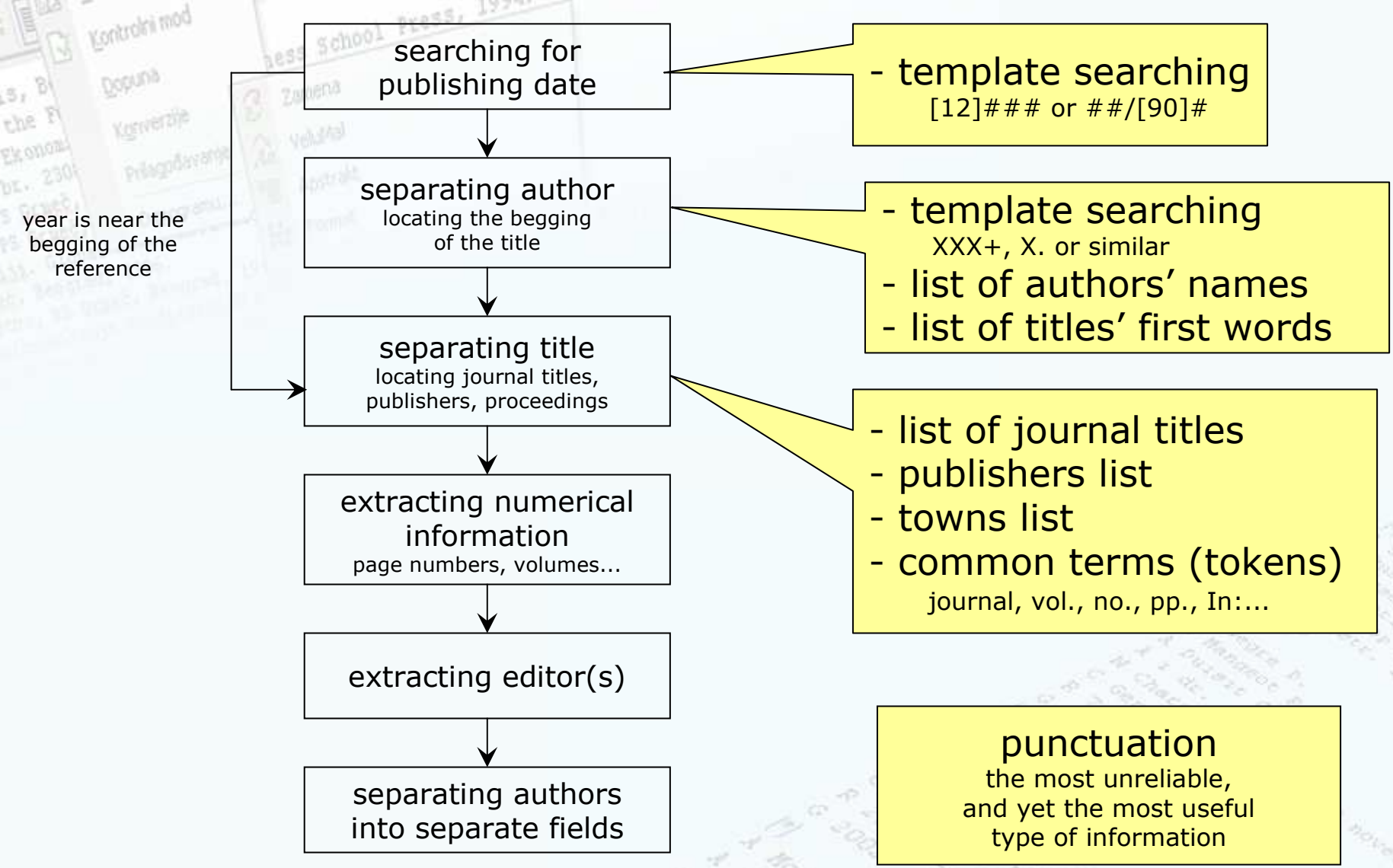
## level of "freedom"

more structured material means less afterwards interventions and corrections

# Example (parsed references)



# Algorithm (simplified)



# 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

# Perspectives

- 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