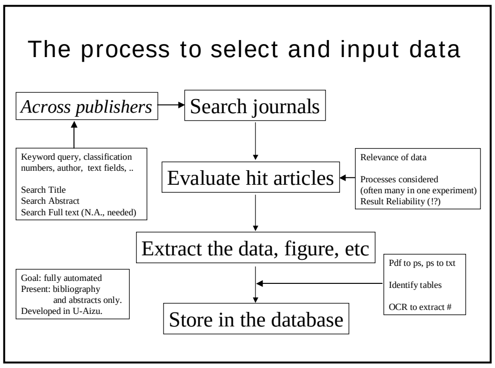
Acquisition of Data for Plasma Simulation by Automated Extraction of Terminology from Article Abstracts

L. Pichl, M. Suzuki[1], M. Murata[2], A. Sasaki[3], D. Kato[4] and I. Murakami[4]

International Christian University, Osawa 3-10-2, Mitaka, Tokyo 181-8585, Japan
1) School of Systems Science, Arkansas Tech University, Russellville, Arkansas 72801, USA
2) National Institute of Information and Communications Technology, Kyoto 619-0289, Japan
3) Japan Atomic Energy Agency, 8-1 Umemidai, Kizu-cho, Kyoto 619-0215, Japan
4) National Institute for Fusion Science, Oroshi-cho, Toki, Gifu 509-5292, Japan

Motivation

- ❑ Maintaining data center DBs costs time and money
- ❑ Data-searching and data-input are very low qualified, manual, stereotype activities → should be automated
- ❑ Commercial DB system solutions: costly & rigid, if a design change is need later.



The process to select and input data

Across publishers → Search journals → Evaluate hit articles → Extract the data, figure, etc → Store in the database

Keyword query, classification numbers, author, text fields, ..
Search Title
Search Abstract
Search Full text (N.A., needed)

Relevance of data
Processes considered (often many in one experiment)
Result Reliability (!?)

Pdf to ps, ps to txt
Identify tables
OCR to extract #

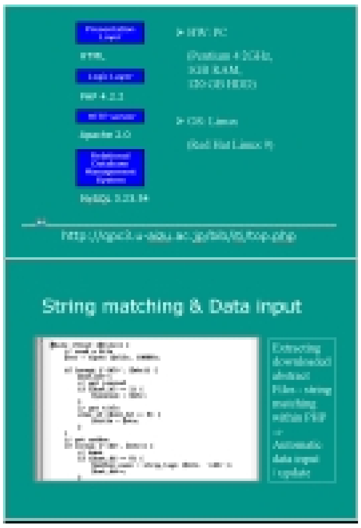
Goal: fully automated
Present: bibliography and abstracts only.
Developed in U-Aizu.

- ❑ <u>Present work automates abstract download & processing</u>

Database System Design

- ❑ FCII linux OS, MySQL DB management system

HTML and logic layer with PHP scripting.


Presentation Layer
HTML
Logic Layer
PHP 4.2.2
WWW server
Apache 2.0
Relational Database Management System
MySQL 3.23.54
➢ HW: PC
(Pentium 4 2GHz, 1GB RAM, 120 GB HDD)
➢ OS: Linux
(Red Hat Linux 9)
String matching & Data input
Extracting downloaded abstract Files : string matching within PHP
→
Automatic data input / update

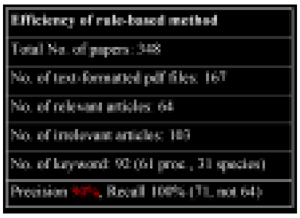
- ➢ Pre-designed queries are sent to the multiple publisher databases by using unix wget in command line mode.
- ➢ HTML output is analyzed for relevant data input fields using string matching search in PHP for each provider.
- ➢ Access rate is interrupted by intermissions following N(T,S) (or time histogram).
- ➢ DB input format is HTML; figures for special characters are also input the DB.

Abstract search and text classification (Dr. Murata, NIICT)

Information extraction

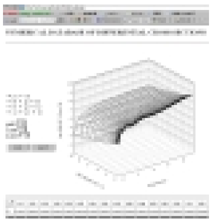
- ✓ Relevance judgment is done based on the extraction of abstracts
- ✓ Special terminology of A+M physics is automatically recognized (species, states, elements, processes)
- ✓ Satisfactory recall and precision are obtained based on
- ✓ simultaneous process/species matching
- ✓ Further improvement is possible by the analysis of table and figure captions

<u>Evaluation</u>:



Efficiency of rule-based method
Total No. of papers: 348
No. of text-formatted pdf files: 167
No. of relevant articles: 64
No. of irrelevant articles: 103
No. of keyword: 92 (61 proc., 31 species)
Precision 91%, Recall 100% (71, not 64)

Work in development



- ▪ Numerical database of differential cross sections
  - prototype on the left, also crdb.nifs.ac.jp
- ▪ Online database builder

Conclusion

Created an automated-input bibliography database for fusion plasma at NIFS.
Free-software open-source abstract and fulltext DB system development is finished.
A differential cross-section database with GUI features is tested on the same footing.
Interfaced with online article search engine, which includes a module for recognition of A+M terminology (Dr. Murata) in a joint project with Dr. Sasaki.