IADLC 05
(International Advanced Digital Library Conference)

A UNIFIED FRAMEWORK FOR AUTOMATIC METADATA EXTRACTION FROM ELECTRONIC DOCUMENT

Asanee Kawtrakul, Chaiyakorn Yingsaeree and Team

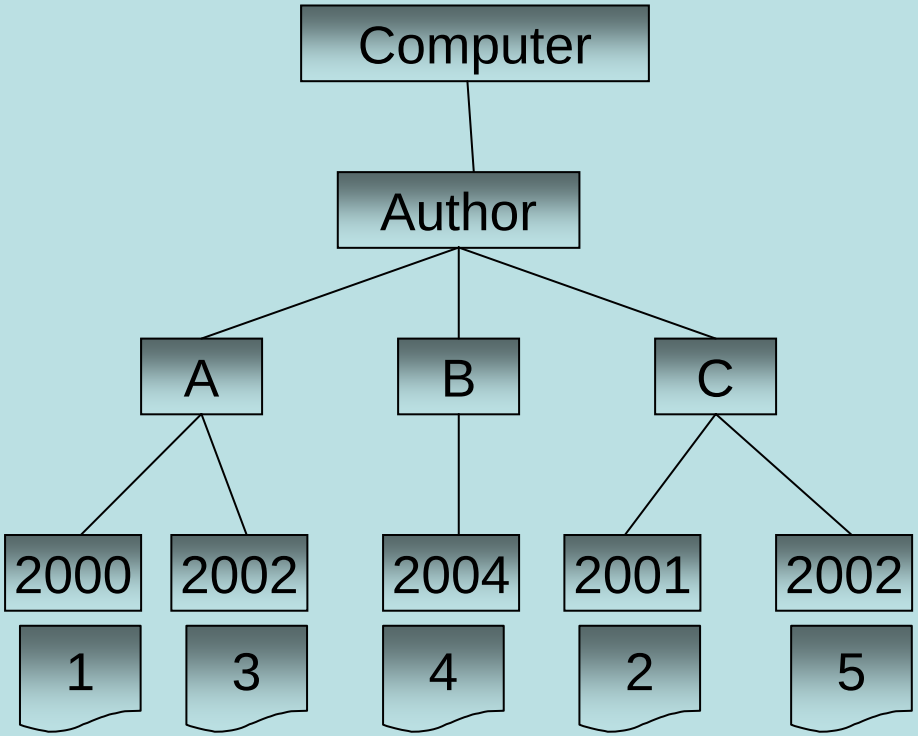
NAiST Research Laboratory
Dept of Computer Engineering, Faculty of Engineering
Kasetsart University, THAILAND

26 August 2005, Nagoya

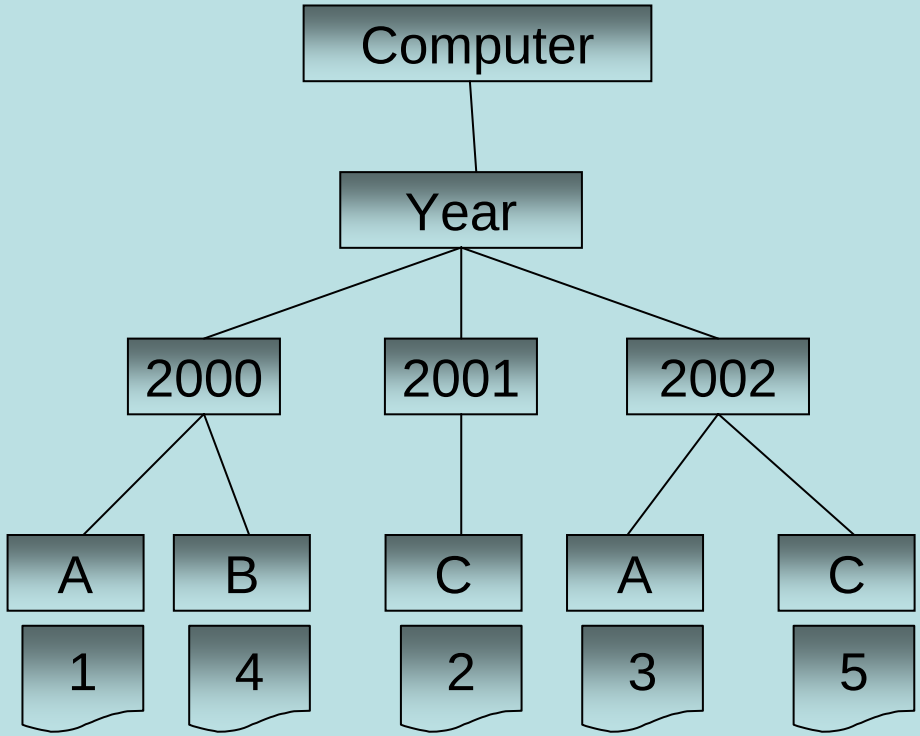
Goal

Multi-viewpoint Knowledge Tracking

Knowledge Tracking : Different Tracking Paths (Same Documents)



Another Knowledge Gain :
-Author B is a new researcher.
-Author C publishes papers continuously
-Author A do not publish in year 2001
-And more...



Another Knowledge Gain :
-Author C is only one who published in year 2001
-Author A and B are pioneer researchers in domain.
-And more ...

Today's Outline

- Introduction
- Problems
- Architecture
- Current Status
- Conclusion
- Ongoing Projects

Introduction

- **What is metadata?**
  - Data about data
  - Ex:
    - About document:Traditional library card catalogue,
    - About content: purpose, problem spaces, methodologies, and results.

- **Why is it important?**
  - Help people distinguish relevant from non-relevant documents,
  - Multi-view point of Knowledge Tracking

Examples of Metadata

Introduction (2)

- Where does it come from?
  - **By Human**
    - Annotating the document manually
  - **By Computer**
    - Metadata Harvesting
    - Metadata Extraction

Introduction (3)

Metadata Harvesting

- Collect metadata from previously defined metadata
- Usually performed by creating a parser to analyze source metadata and transform parsing results into an appropriated format
- Application includes interoperability between metadata of different systems and platforms

Introduction (4)

Metadata Extraction

- Extract metadata from document content
- Usually performed by machine learning, rule-based parser and Regular Expression
- Machine learning approaches are robust and adaptable, but require a large training example
- Rule-based parsers and Regular Expression are dependent on an application domain, and no training example is required

Introduction (5)

- **Objective**
  - Create a framework for automatic metadata extraction from technical and thesis documents which have fixed format.
- **Solution**
  - Use rule-based parser due to simplicity and cost

Problems

- **Variety of electronic document formats**
  - E-Document can be stored in a variety of formats
    - e.g. Microsoft Word, Adobe Acrobat, Image of document, etc.
  - It is necessary to convert such document into text file in order to access document content
- **Quality of extracted metadata**
  - Extracted metadata may contain errors both from original documents and text conversion process,
  - Some mechanisms are required to produce high-quality metadata

Architecture

Tanyaratana Dumkua 2000 : Differential Expression of 1-Aminocyclopropane-1-Carboxylate (ACC) Oxidase Gene in *Carica papaya* (Kaegdum). Master of Science (Genetic Engineering), Major Field Genetic Engineering, Interdisciplinary Graduate Program. Thesis Advisor : Associate Professor Supat Attathom, PhD. 70 pages. ISBN 974-461-265-7

Text Conversion Module

Task-Oriented Parser Module

Author's Name	Tanyaratana Dumka
Thesis's Title	Differential Expression of 1-Amniocyclopropane-1-……
Degree	Master of Science
Major	Field Genetic Engineering

Extracted Metadata

Language Model

Dictionary

Existing Metadata

IDENTIFY&CORRECT ERRORS from
- task-oriented parser
  - -controlled vocabularies
  - -general vocaburaries
- existing metadata repository

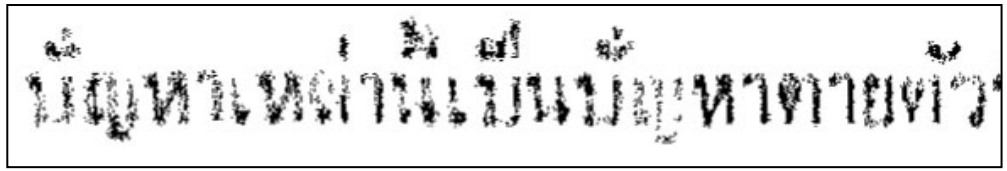
Data Verification Module

Corrected Metadata

Text Conversion Module (1)

- **AFPL Ghostscript (for PS & PDF)**
- **CATDOC (for Microsoft Word & Excel)**
- **OCR (for Image Document)**
  - Document Skew Correction
  - Marginal Noise Removal
  - Salt-and-Pepper Noise Removal
  - Broken Character Management

Text Conversion Module (2)



Broken Character

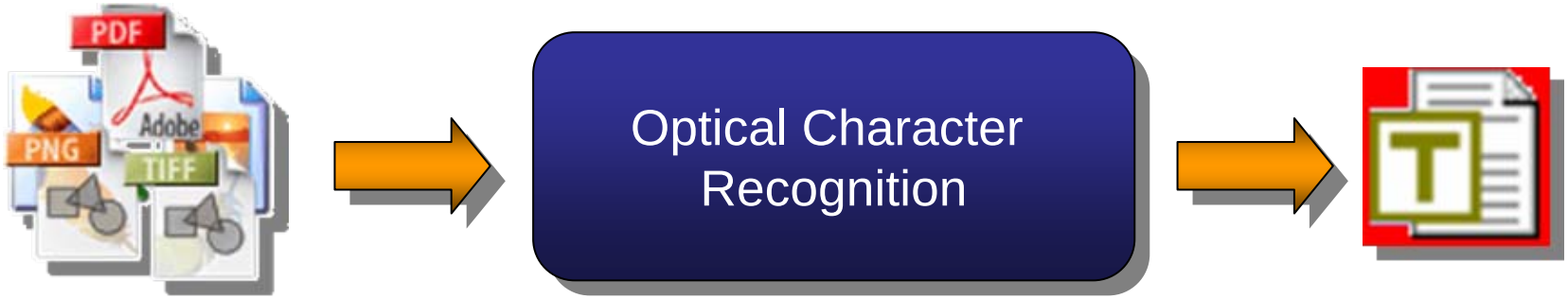
Skew

Salt-and-Pepper Noise

Marginal Noise

Optical Character Recognition

❑ Conversion from Image to Text

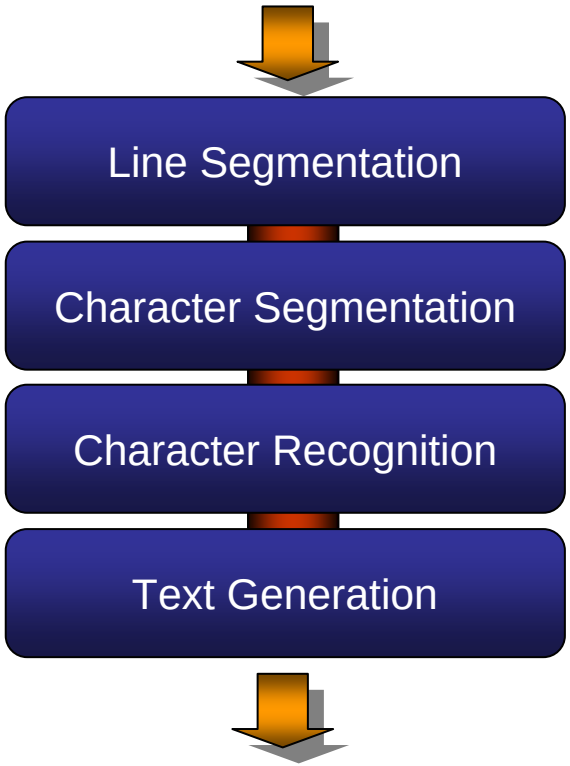


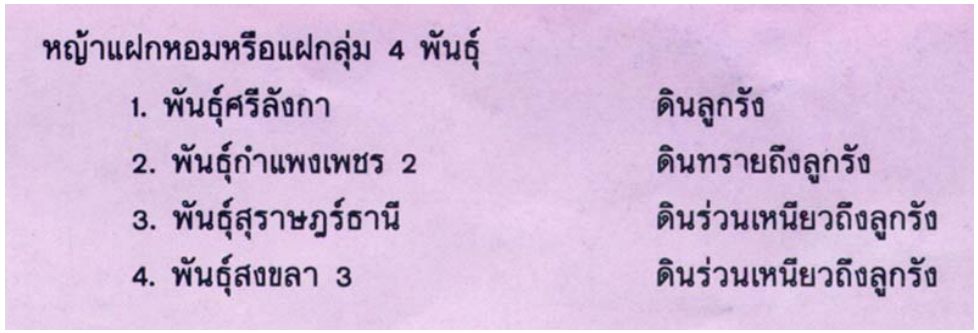
หญ้าแฝกหอมหรือแฝกลุ่ม 4 พันธุ์

1. พันธุ์ศรีลังกา	ดินลูกรัง
2. พันธุ์กำแพงเพชร 2	ดินทรายถึงลูกรัง
3. พันธุ์สุราษฎร์ธานี	ดินร่วนเหนียวถึงลูกรัง
4. พันธุ์สงขลา 3	ดินร่วนเหนียวถึงลูกรัง

หญ้าแฝกหอมหรือแฝกลุ่ม 4 พันธุ์
1. พันธุ์ศรีลังกา ดินลูกรัง
2. พันธุ์กำแพงเพชร 2 ดินทรายถึงลูกรัง
3. พันธุ์สุราษฎร์ธานี ดินร่วนเหนียวถึงลูกรัง
4. พันธุ์สงขลา 3 ดินร่วนเหนียวถึงลูกรัง

Optical Character Recognition



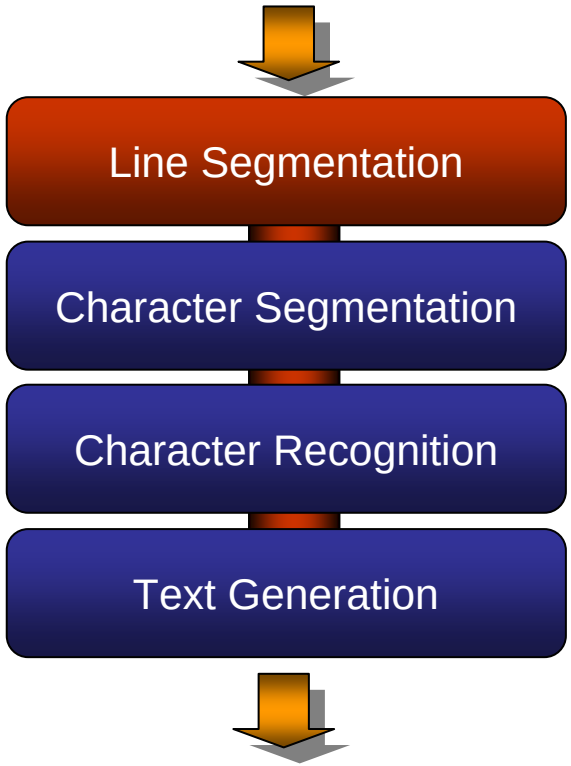

หญ้าแฝกหอมหรือแฝกลุ่ม 4 พันธุ์

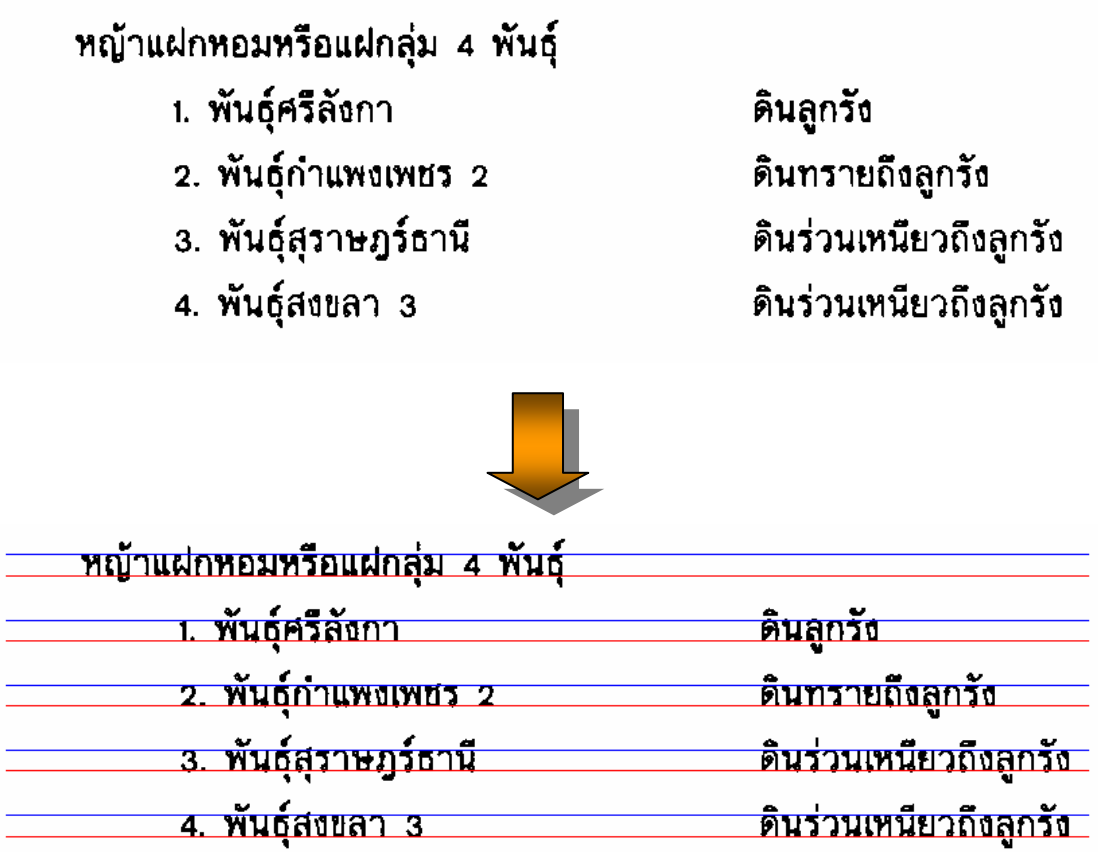
1. พันธุ์ศรีลังกา ดินลูกรัง
2. พันธุ์กำแพงเพชร 2 ดินทรายถึงลูกรัง
3. พันธุ์สุราษฎร์ธานี ดินร่วนเหนียวถึงลูกรัง
4. พันธุ์สงขลา 3 ดินร่วนเหนียวถึงลูกรัง

หญ้าแฝกหอมหรือแฝกลุ่ม 4 พันธุ์

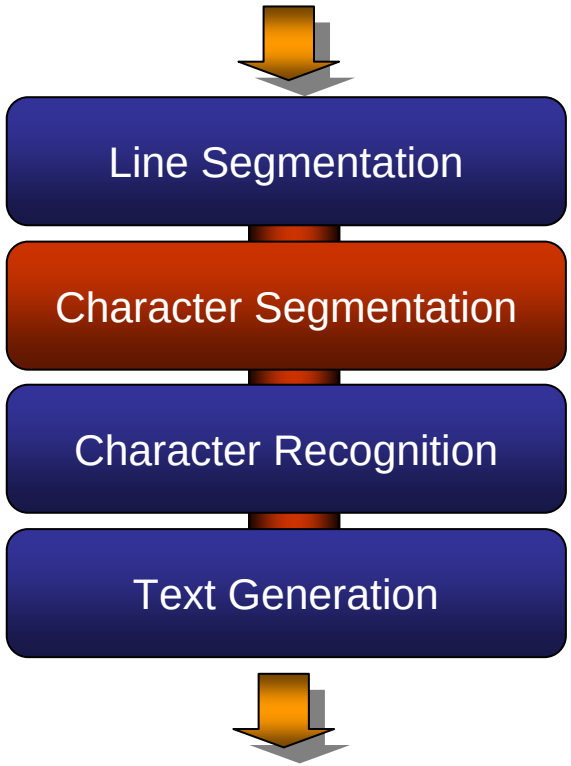
1. พันธุ์ศรีลังกา ดินลูกรัง
2. พันธุ์กำแพงเพชร 2 ดินทรายถึงลูกรัง
3. พันธุ์สุราษฎร์ธานี ดินร่วนเหนียวถึงลูกรัง
4. พันธุ์สงขลา 3 ดิน

Optical Character Recognition

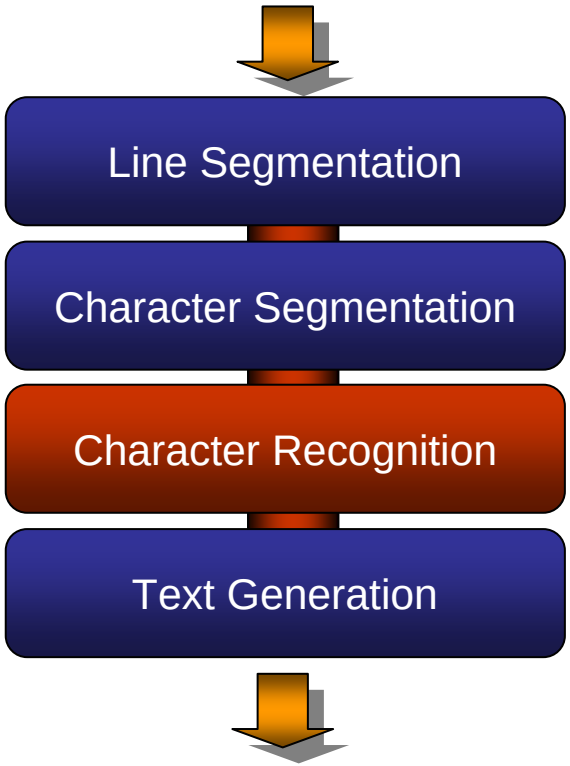


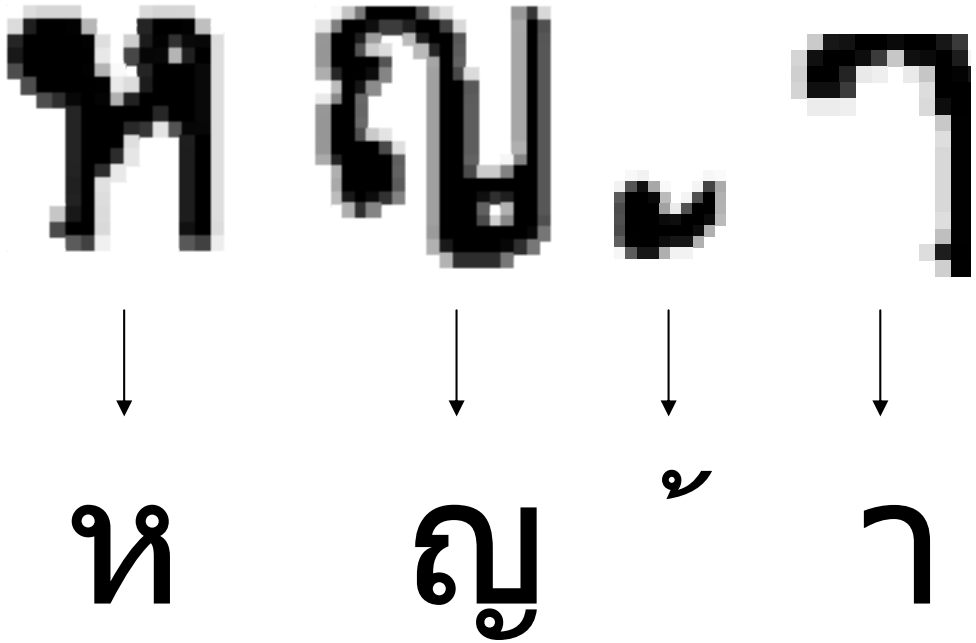


Optical Character Recognition

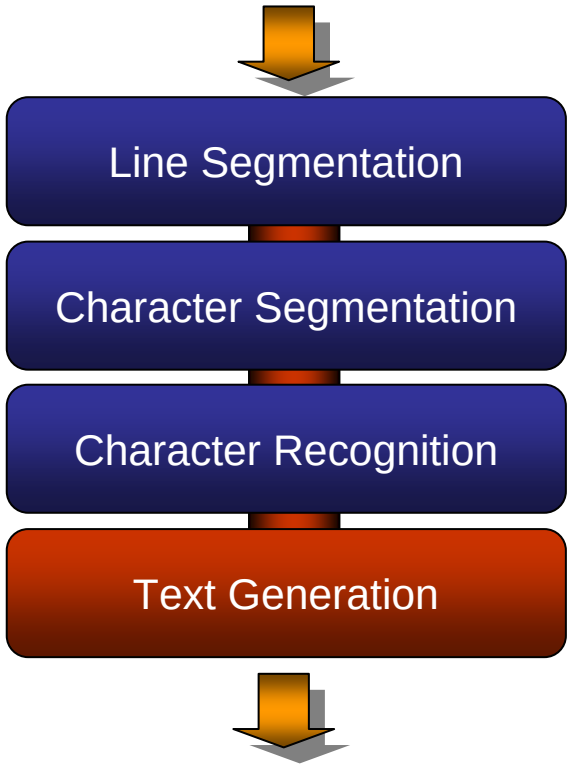


Optical Character Recognition

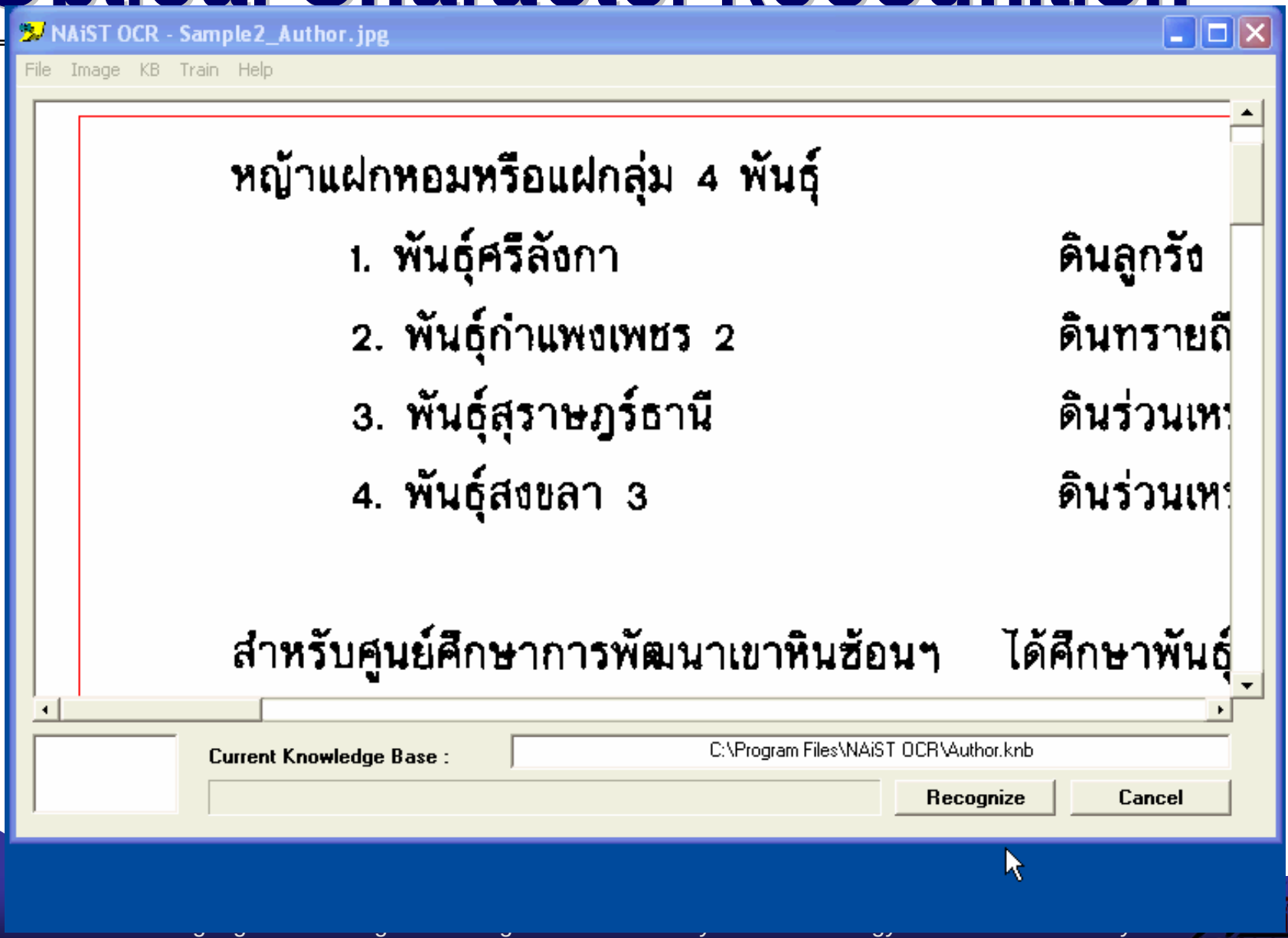




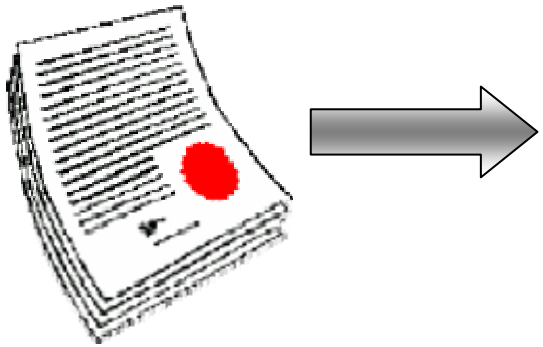
Optical Character Recognition



Optical Character Recognition



Automatic Metadata Extraction

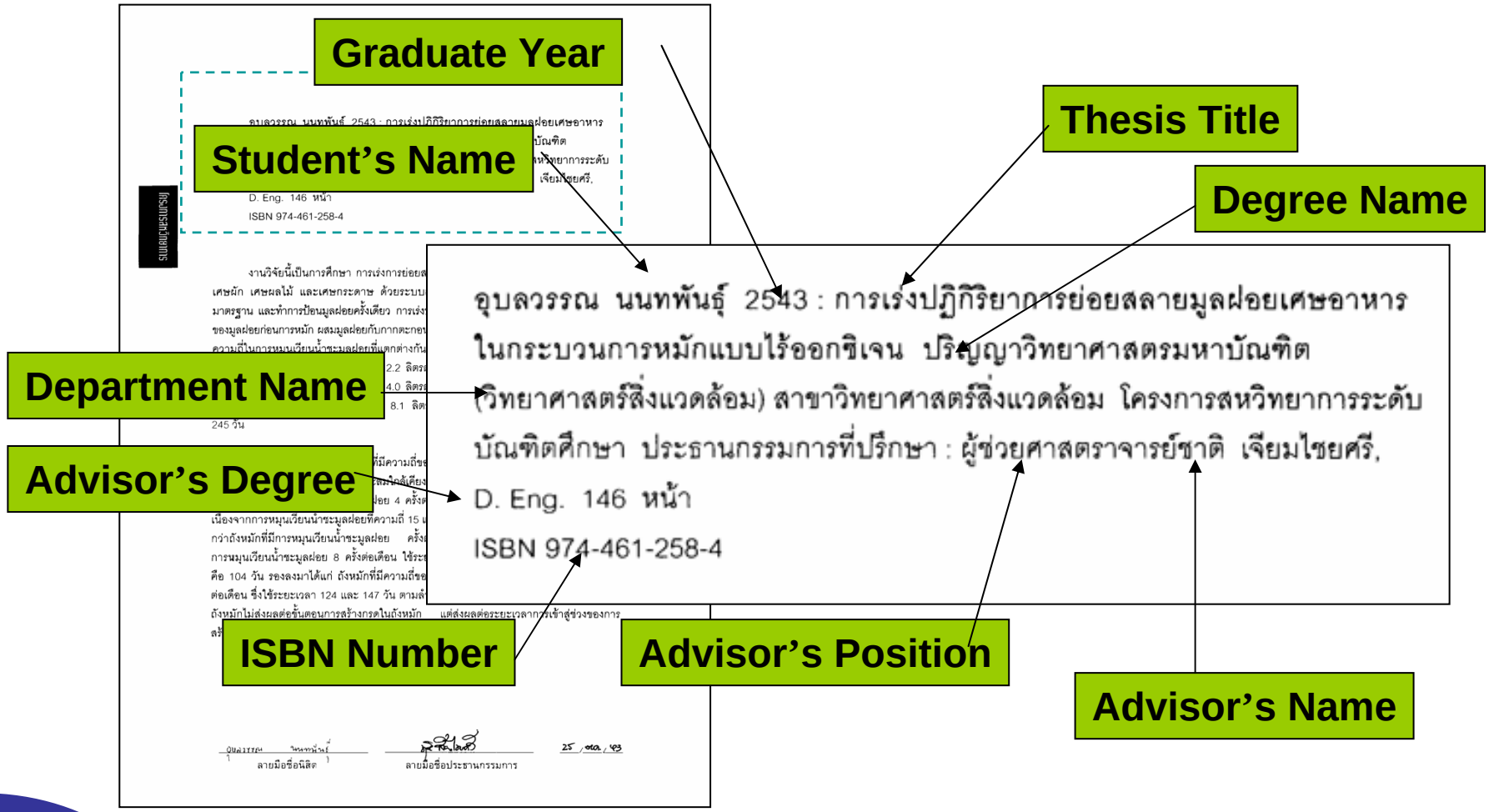


Metadata

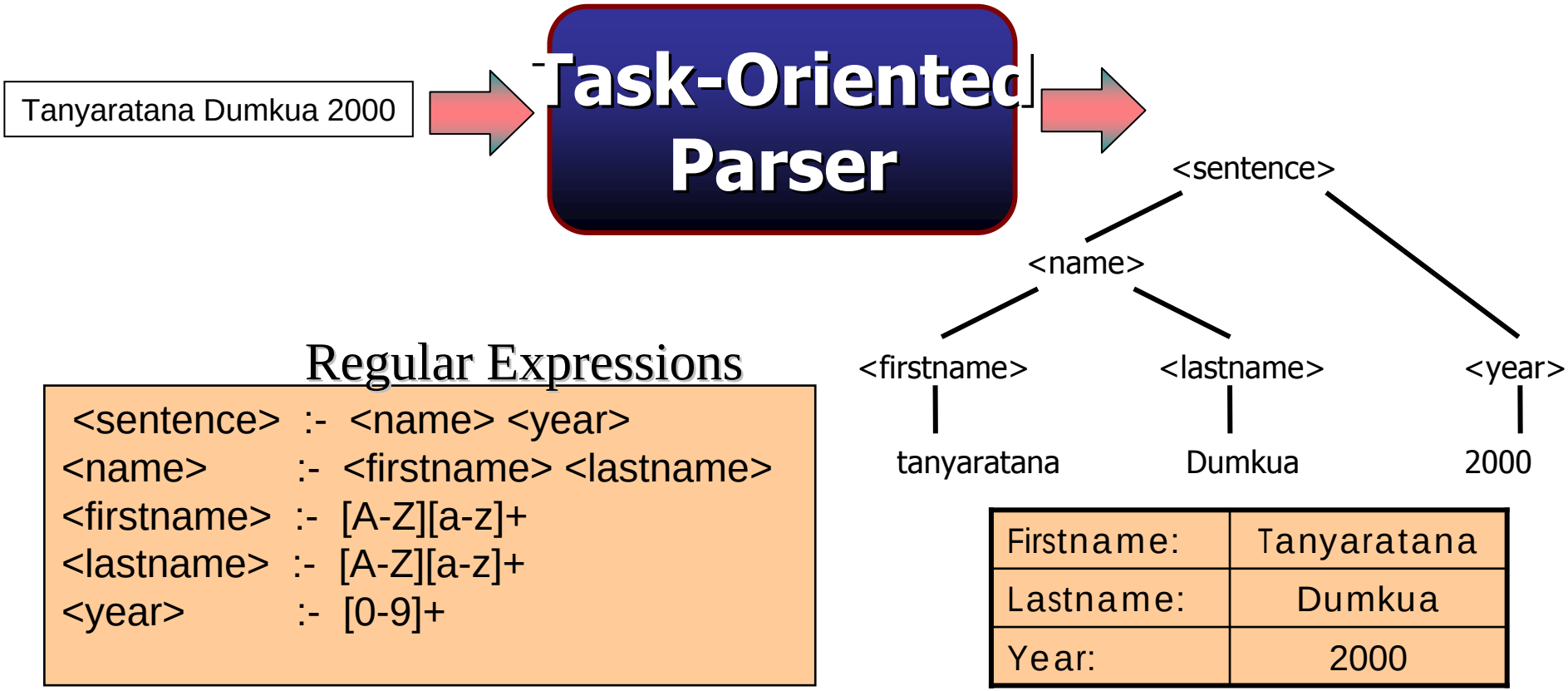
ผู้แต่ง	นายวีร์ สัตยมาศตร์
ชื่อเรื่อง	การสร้างดัชนีหนังสืออัตโนมัติ
สำคัญ	การประมวลผลภาษาธรรมชาติ การสร้างดัชนี
…ฯลฯ…	

Needs Resources and Cost

Extraction Meta Data for e-thesis



Automatic Metadata Extraction

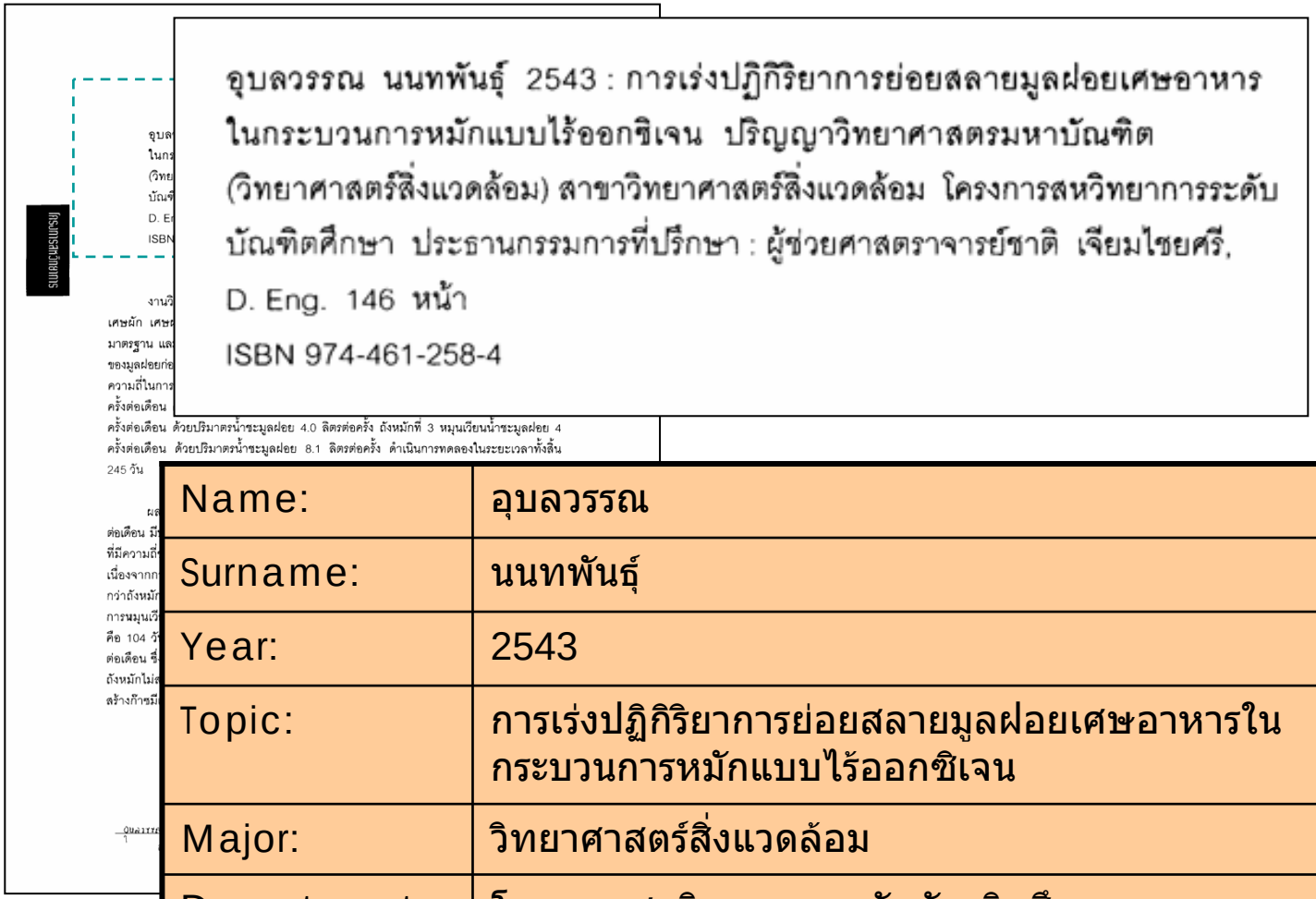


Regular Expressions

```
<sentence>  :-  <name> <year>
<name>      :-  <firstname> <lastname>
<firstname> :-  [A-Z][a-z]+
<lastname>  :-  [A-Z][a-z]+
<year>      :-  [0-9]+
```

Firstname:	Tanyaratana
Lastname:	Dumkua
Year:	2000

Extraction Result for e-thesis



อุบลวรรณ นนทพันธุ์ 2543 : การเร่งปฏิกิริยาการย่อยสลายมูลฝอยเศษอาหารในกระบวนการหมักแบบไร้ออกซิเจน ปริญญาวิทยาศาสตรมหาบัณฑิต (วิทยาศาสตร์สิ่งแวดล้อม) สาขาวิทยาศาสตร์สิ่งแวดล้อม โครงการสหวิทยาการระดับบัณฑิตศึกษา ประธานกรรมการที่ปรึกษา : ผู้ช่วยศาสตราจารย์ชาติ เจียมไชยศรี, D. Eng. 146 หน้า

ISBN 974-461-258-4

Name:	อุบลวรรณ
Surname:	นนทพันธุ์
Year:	2543
Topic:	การเร่งปฏิกิริยาการย่อยสลายมูลฝอยเศษอาหารในกระบวนการหมักแบบไร้ออกซิเจน
Major:	วิทยาศาสตร์สิ่งแวดล้อม
Department:	โครงการสหวิทยาการระดับบัณฑิตศึกษา
...	....

Data Verification Module (1)

- Error from Task-Oriented Parser Module
  - Controlled Vocabularies
  - General Vocabularies
- Error in Existing Metadata Repository

Data Verification Module (2)

- **Error from Task-Oriented Parser Module**
  - The parser might not be able to parse some documents due to incomplete grammar, error from text conversion, or defect in the document itself
  - To solve the problem, either creating new rules or fixing the defect is required

Data Verification Module (3)

- **Error in Controlled Vocabularies**
  - Some metadata fields' value can be only a word(s) in controlled vocabularies
  - Error identification can be achieved by comparing extracted data with a dictionary
  - When error occurs, the correction process simply replace the error word with its closest word in the dictionary by means of Edit Distance

Data Verification Module (4)

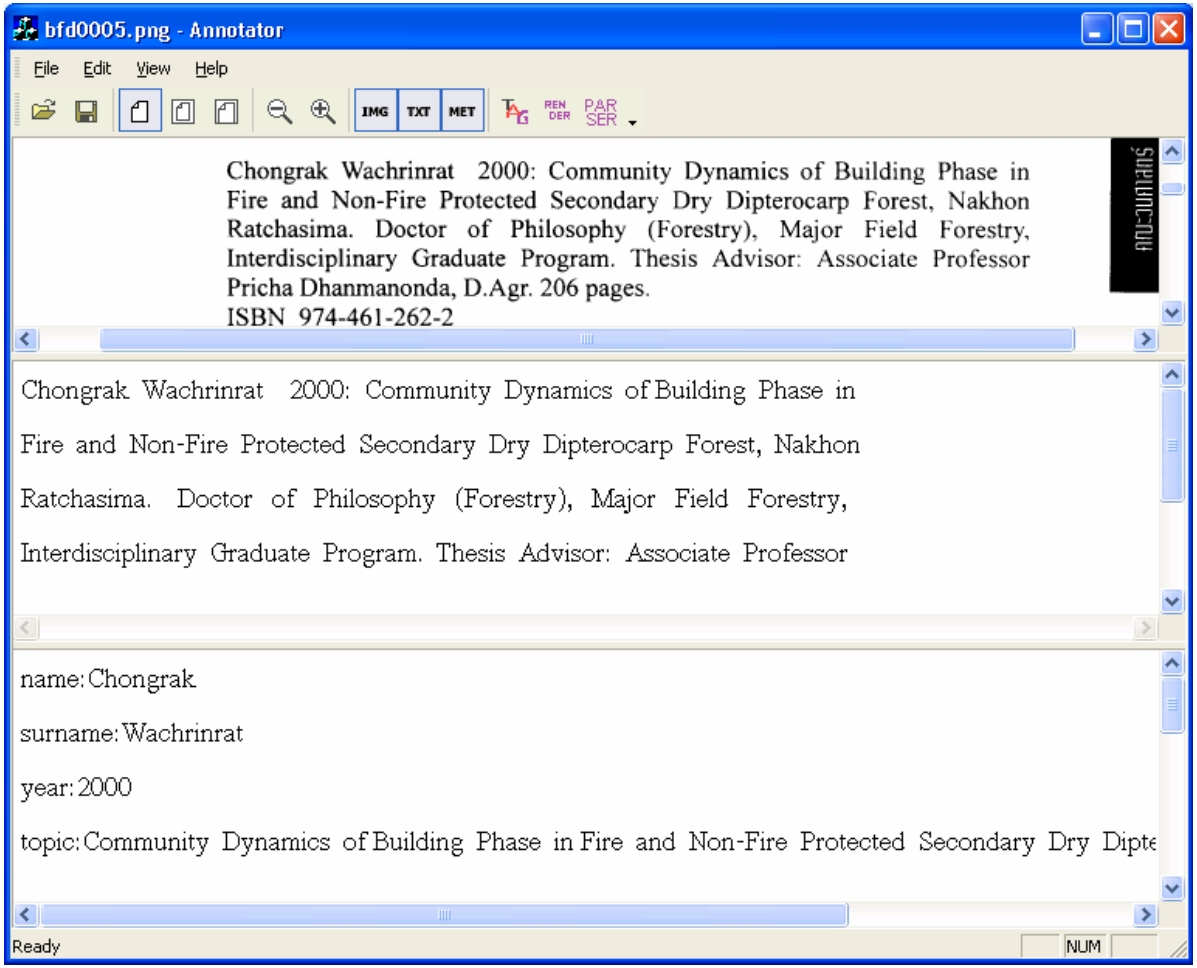
- **Error in General Vocaburaries**
  - **Use spelling correction technique to detect and correct the errors**
    - OCR Error Correction
    - Typing Error Correction
  - **This module is under development**

Data Verification Module (5)

- **Error in Existing Metadata Repository**
  - Hand-made metadata usually contained many errors
  - Instead of manually correcting the error, we can use automatic metadata extraction and alignment tool to ease data correction process

Current Status

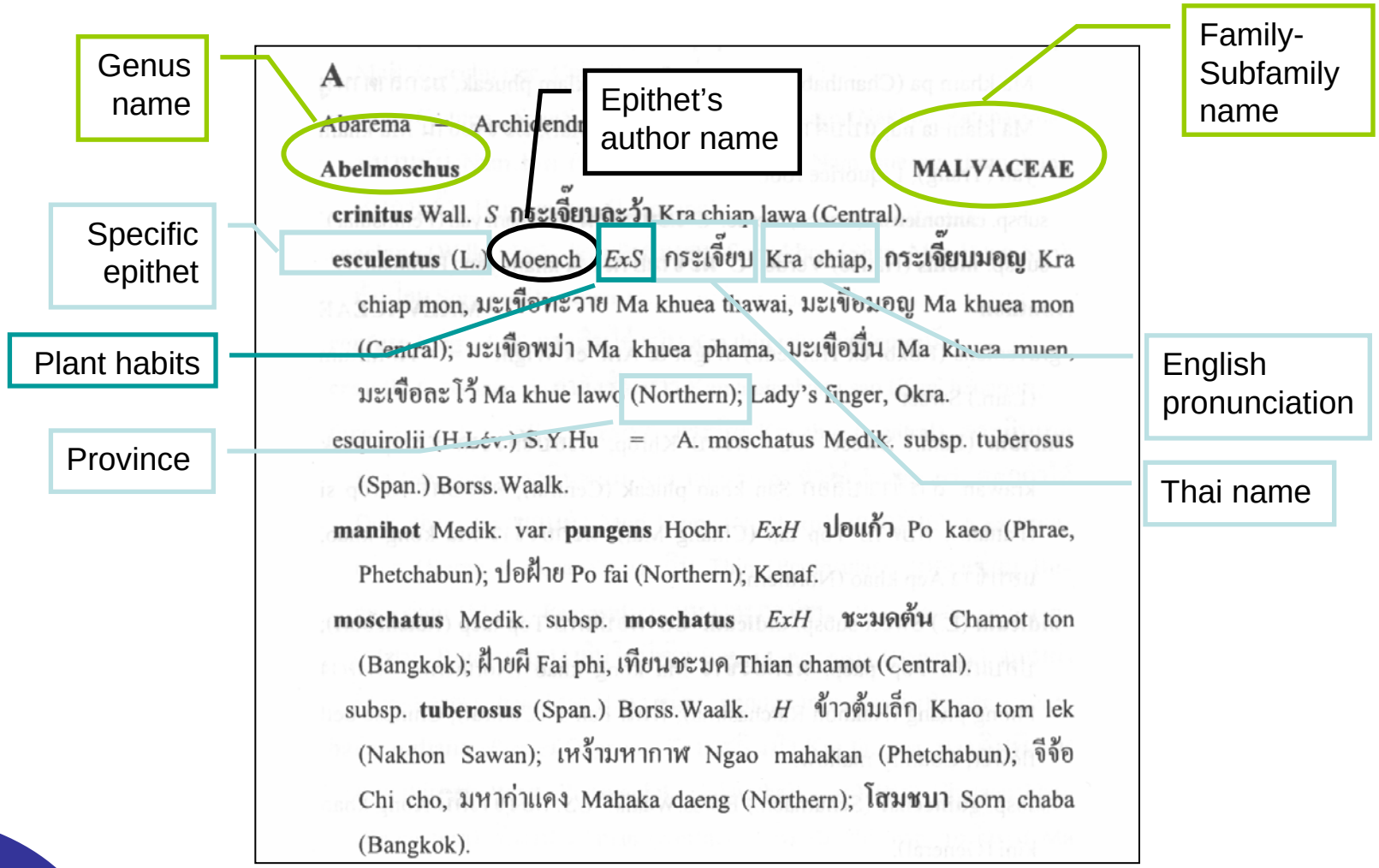
Extracting metadata from students' thesis abstract (1)



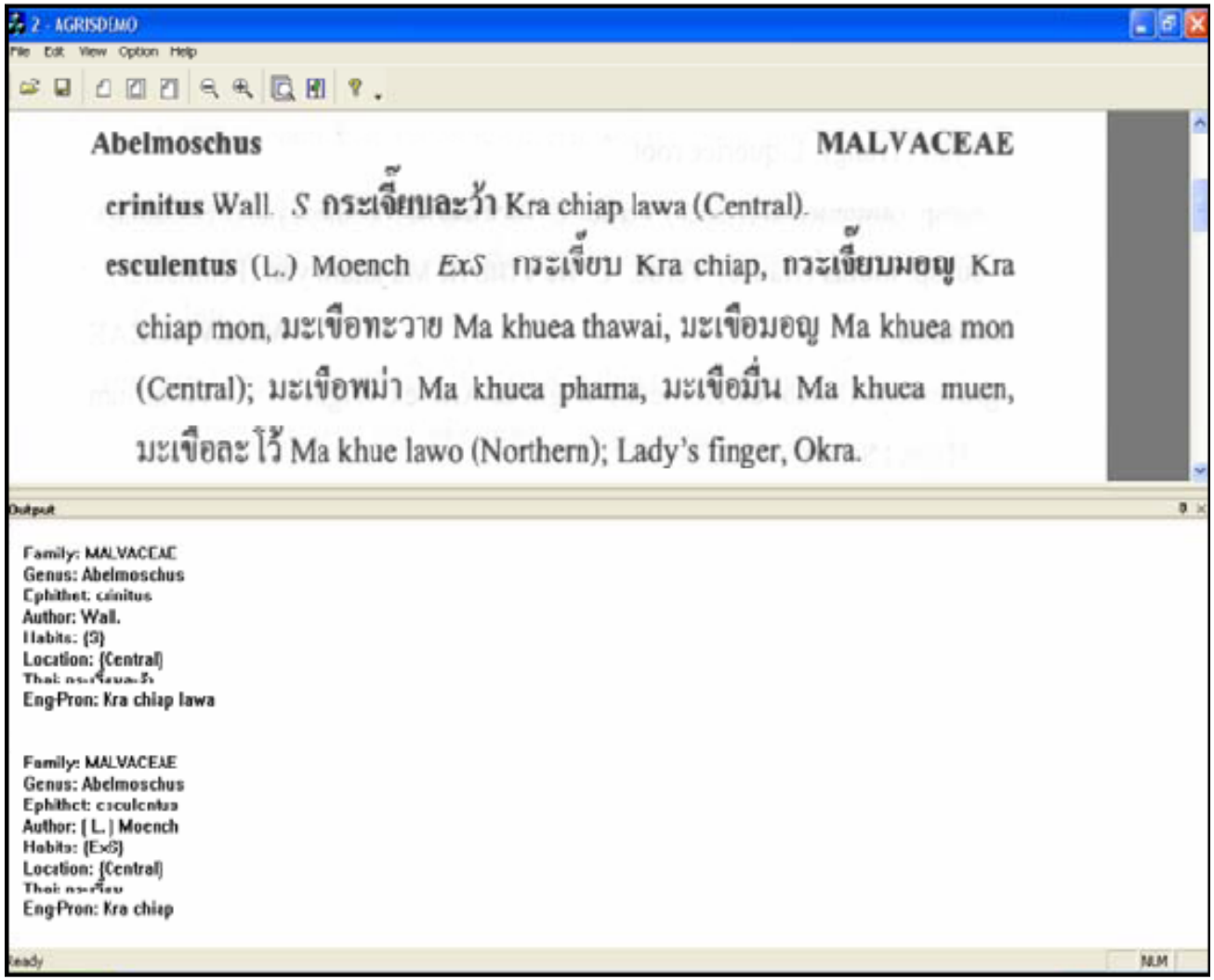
Extracting metadata from students' thesis abstract (2)

The preliminary results with 3,712 thesis show that using this system greatly reduce the labor work of metadata creation process by correctly extracting metadata 91.41% of the documents.

Extracting plant information from image of Thai plant name dictionary(1)



Extracting plant information from image of Thai plant name dictionary (2)

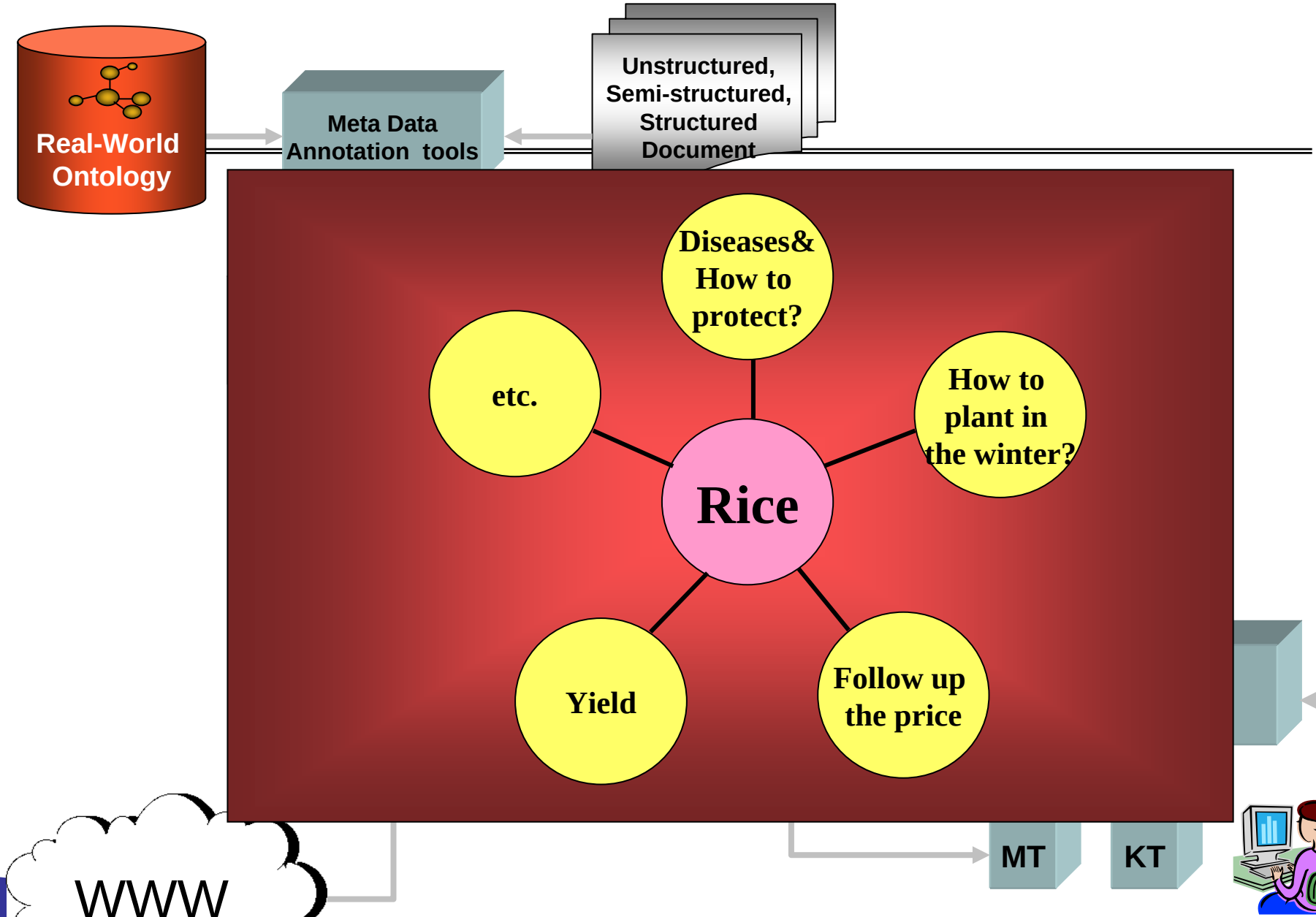


Conclusion

- **A Unified Framework for Automatic Metadata Extraction from Electronic Document**
- **Consists of three main components**
  - text conversion module
  - task-oriented parser module
  - data verification module
- **The experimental result shown that using the framework greatly reduce the labor work of metadata creation process**

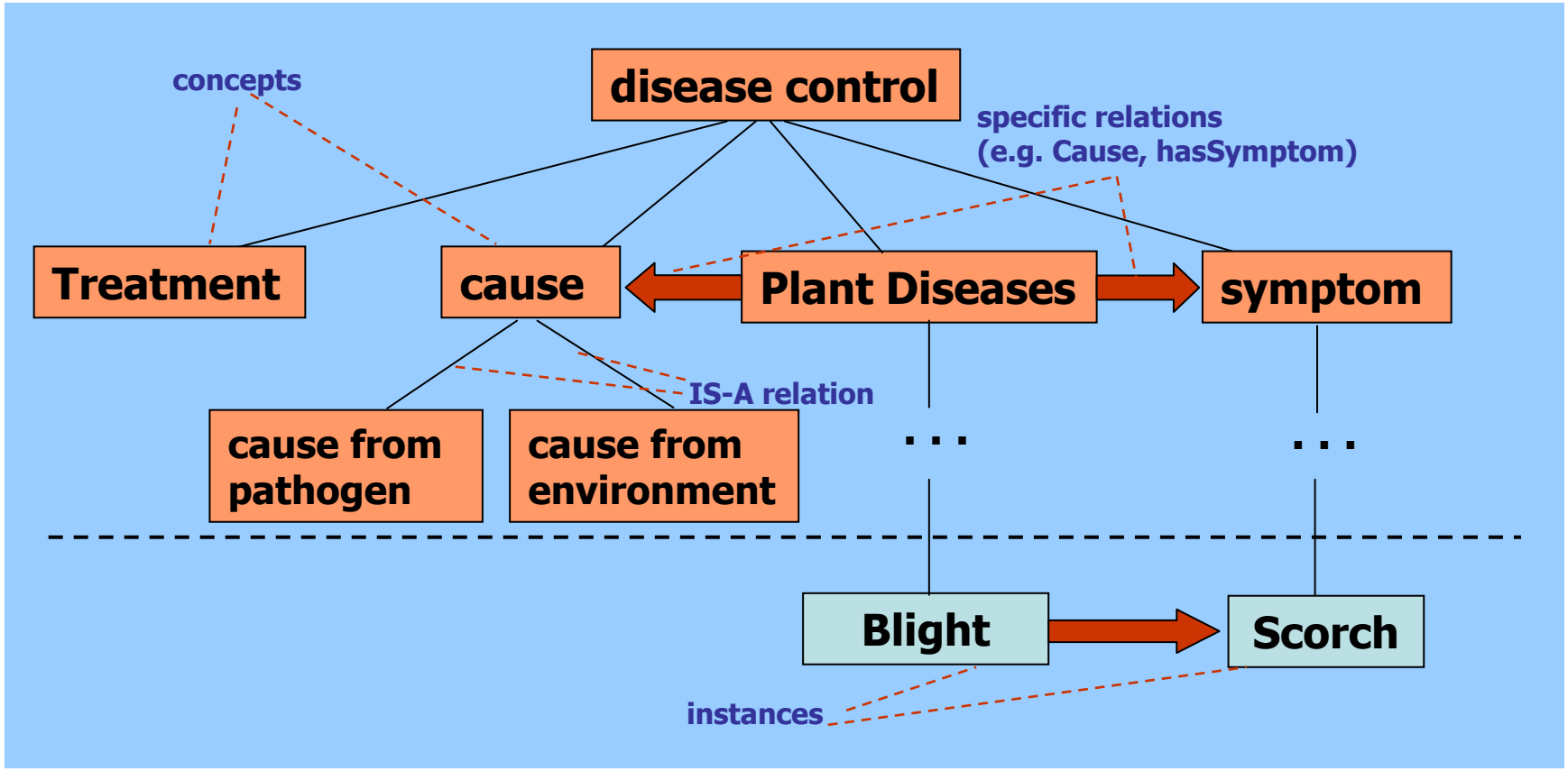
Ongoing Projects

- Agricutural Knowledge Portal
  - Ontology Maintenance
  - Information extraction
  - Knowledge Mining
- Open source Digital Library
  - Knowledge collecting, sharing and Accessing (DSpace)
  - Library System Management (Koha)


Real-World Ontology
Meta Data Annotation tools
Unstructured, Semi-structured, Structured Document
Diseases& How to protect?
etc.
How to plant in the winter?
Rice
Yield
Follow up the price
WWW
MT
KT

Ontology Development for Enhancing Service

Task Oriented ontology



Automatic Ontology Construction

Raw Text Example

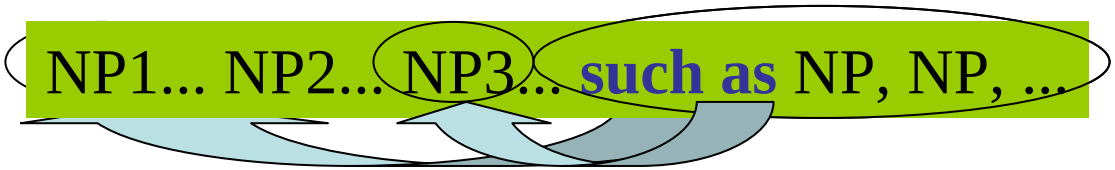
ผักกาดหอม

ผักกาดหอมเป็นผักที่ใช้บริโภคส่วนใบ เป็นผักจำพวกผักสลัดที่มีคุณค่าทางอาหารสูง นิยมบริโภคกันแพร่หลายที่สุดในบรรดาผักสลัดด้วยกัน โดยส่วนใหญ่นิยมรับประทานสดและนำมาประกอบอาหารหลายชนิด คนไทยนิยมใช้ผักกาด[illegible]บอาหารจำพวกยำต่างๆ สาคูหมู หรือข้าวเกรียบปากหม้อ เป็นต้น[illegible]ของผักกาดหอมนอกจากจะใช้กินเป็นผักสดที่มีคุณค่าทางอา[illegible] ยังจัดเป็นอาหารทางตาด้วยโดยการนำมาตกแต่งอาหารให้มีสีสั[illegible]น่ารับประทานมากขึ้น นอกจากนี้ผักกาดหอมยังมีคุณสมบัติในการเป็นยาอีกด้วย ความต้องการผักกาดหอมมีอยู่ตลอดทั้งปี โดยเฉพาะในช่วงเทศกาลต่างๆ จึงนับได้ว่าผักกาดหอมเป็นผักที่มีความสำคัญทางเศรษฐกิจชนิดหนึ่งที่นับวันจะทวีความต้องการเพิ่มขึ้นเรื่อยๆ

ผักกาดหอมมีชื่อเรียกอื่นๆ ได้หลายชื่อเช่น ภาคเหนือเรียกว่า ผักกาดยี ภาคกลางเรียกว่าผักสลัด เป็นต้น ผักกาดหอมเป็นพืชที่จัดอยู่ในตระกูล Compositae มีชื่อวิทยาศาสตร์ว่า Lactuca sataiva มีถิ่นกำเนิดในทวีปเอเชียและยุโรป มีปลูกในประเทศไทยมาช้านานแล้ว

Corpus based Ontology Construction

- **Problems in this process:**
  - Many Candidate Terms



Ex1. Many *herbs* can be used as *medicine* and some of them are manufactured in the *industry* level, **such as garlic, ginkgo biloba** and **green bean**.

Ex2. *Sun flower* is rather enduring with dry season while comparing to other *field crops* **such as corn, soy bean**.

Candidate Terms => *herbs*, *medicine*, *industry*

Candidate Terms => *Sun flower*, *field crop*

Ontological Term Selection

- Statistical Technique
  - Mutual Information, the measure of word association

$$MI(w_1, w_2) = \log_2 \frac{P(w_1, w_2)}{P(w_1)\,P(w_2)}$$

Where

w_1 is a candidate term w_2 is a related term

$P(w_i)$ is probability of term w_i

$P(w_i, w_j)$ is probability of co-occurrence of term w_i and w_j

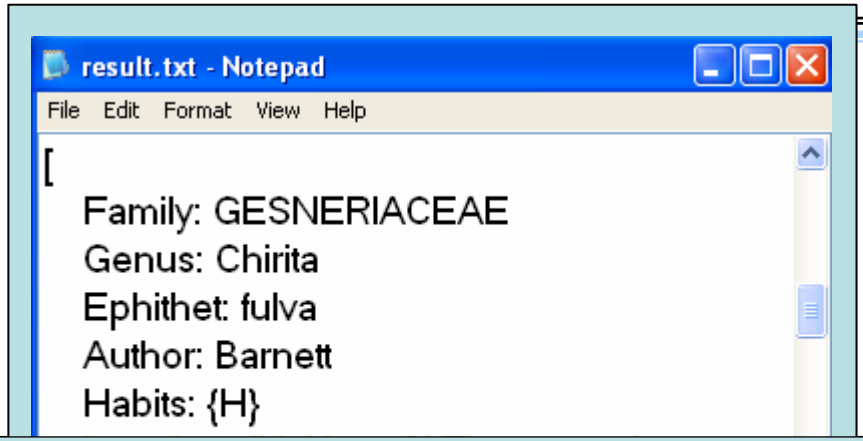
  - Example

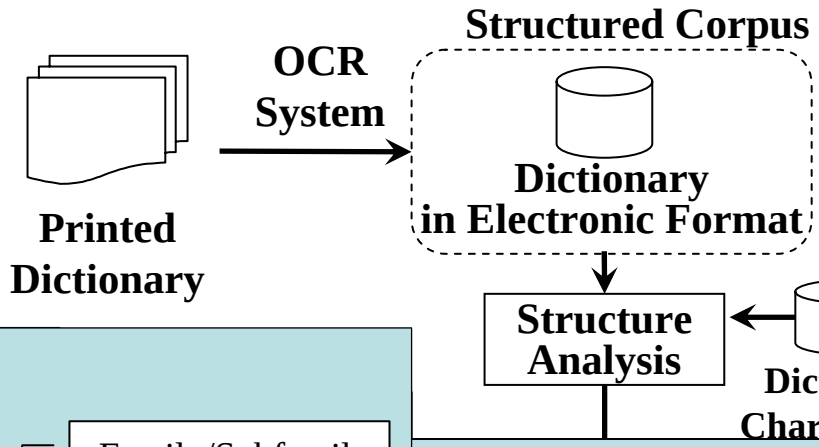
Many **herbs** can be used as **medicine** and some of them are manufactured in the **industry** level, ***such as*** **garlic, ginkgo biloba**

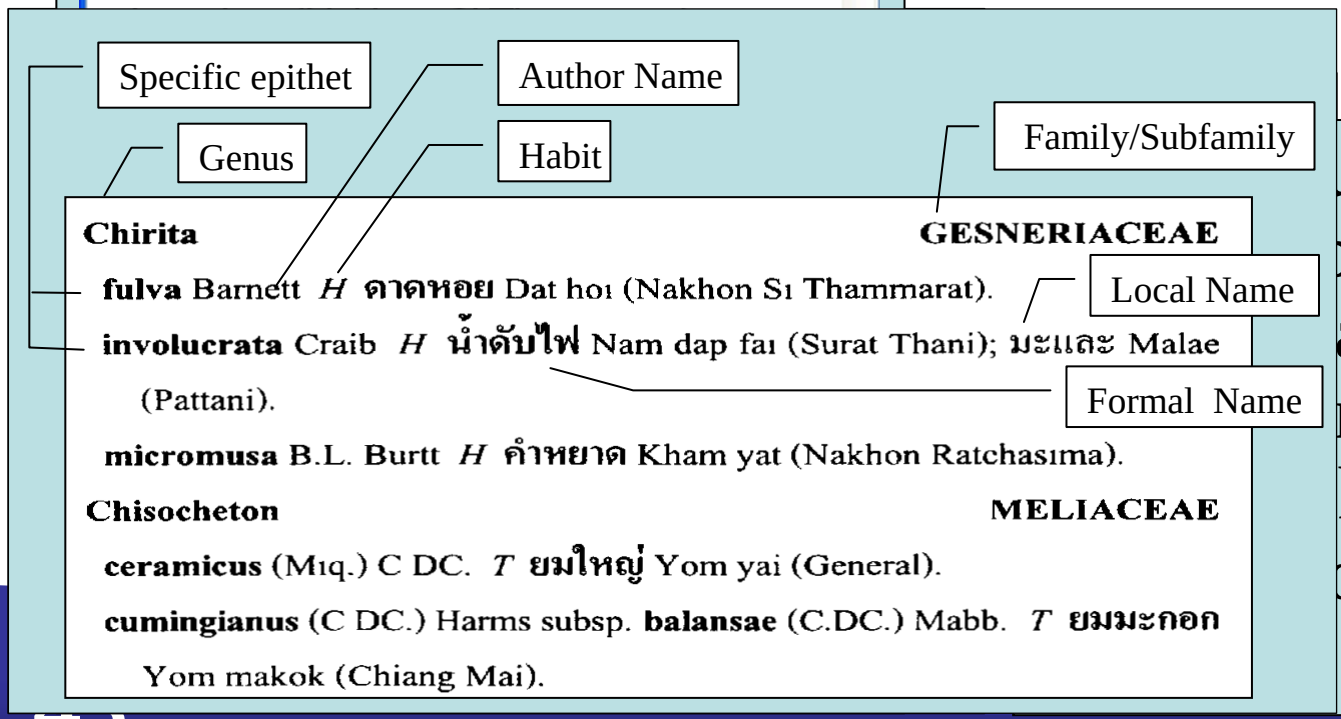
MI(herb, garlic) > MI(medicine, garlic), MI(industry, garlic)

Results : HYPO(garlic, herb)

Dictionary based Ontology Construction







ue:

olied task
arser to extract
rms by
haracteristic
on of terms

Dictionary based Ontology Construction

- Alphabet Characteristic of Dictionary.

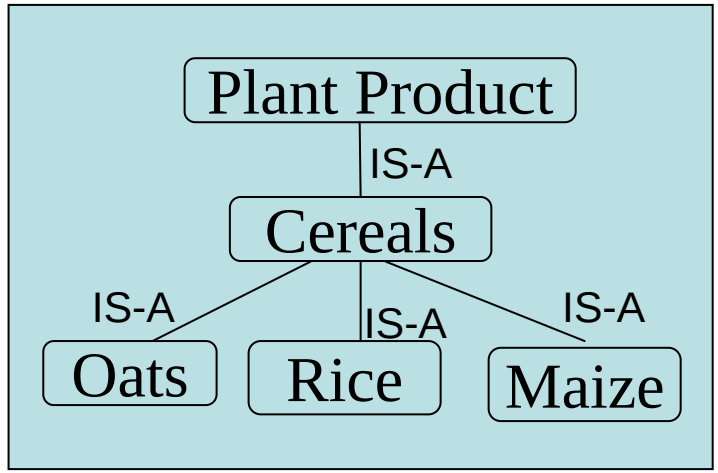
Feature	Database field	Example
All upper case	Family/Sub-Family	EUPHORBIACEAE
Start with upper case	Genus	Acalypha
All lower case	Specific epithet	brachystachya
Thai alphabet with bold font	Formal Name	**ตำแยดอยใบบาง**
Thai alphabet	Local Name	เกี้ยวเกล้า

- **Limitation:**
  - Dictionary has only plant names

AGROVOC Thesaurus based Ontology Construction

- Technique:
  - Convert BT/NT to IS-A Relation

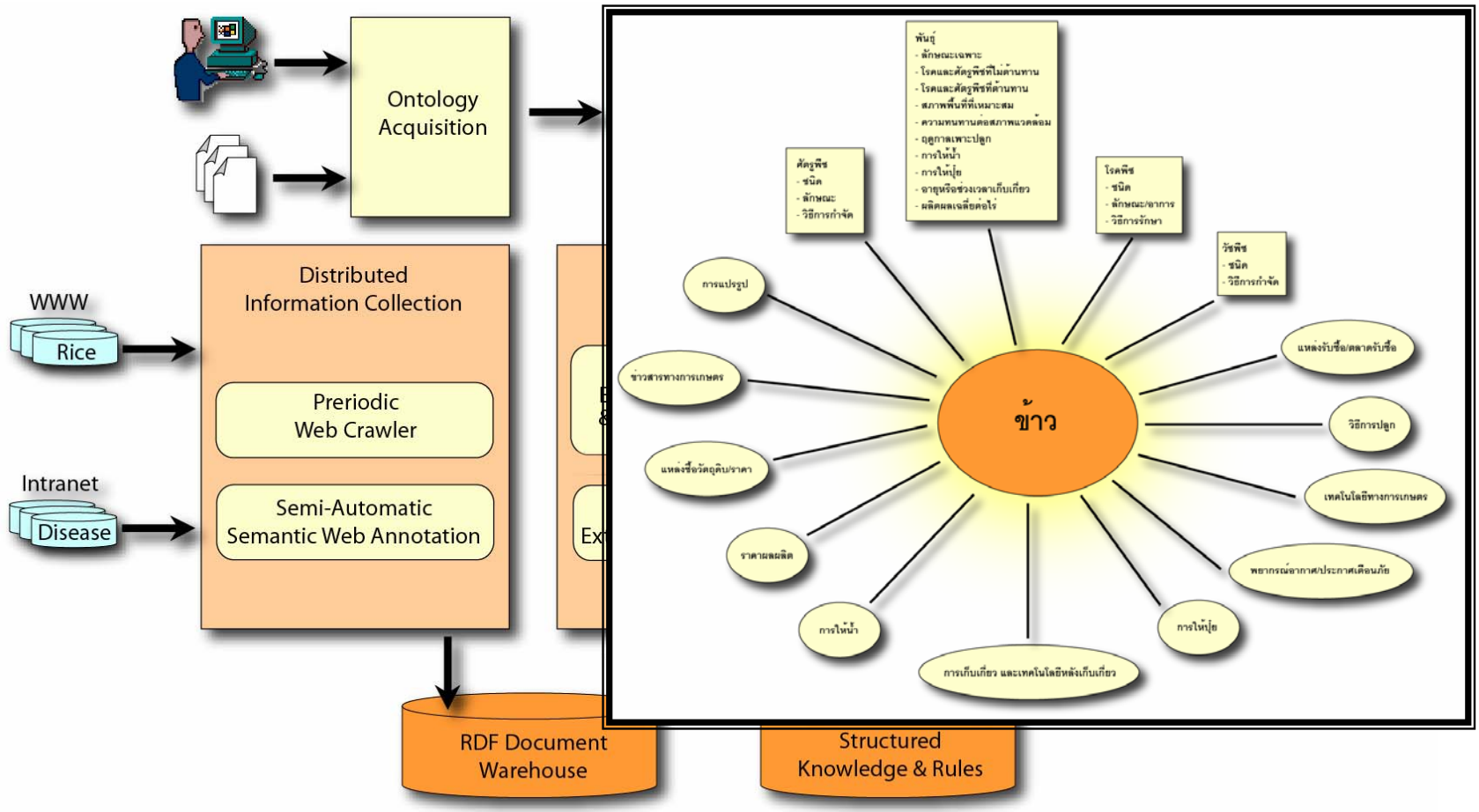
Cereals	BT	Plant Product
	NT	Oats
		Rice
		Maize



Experimental Results

Source	Number of Terms	Number of Relations	Accuracy
Raw Text (150 doc.)	3,720	3,312	73 %.
Dictionary	37,110	21,620	100%.
Thesaurus	27,540	15,628	91%.
3 Sources	43,073	31,387	87 %.

- By random checking with 1,000 united terms, the accuracy of the system is 87 %.


Ontology Acquisition
WWW
Rice
Intranet
Disease
Distributed Information Collection
Preriodic Web Crawler
Semi-Automatic Semantic Web Annotation
RDF Document Warehouse
Structured Knowledge & Rules
ข้าว
พันธุ์
- ลักษณะเฉพาะ
- โรคและศัตรูพืชที่ไม่ต้านทาน
- โรคและศัตรูพืชที่ต้านทาน
- สภาพพื้นที่ที่เหมาะสม
- ความทนทานต่อสภาพแวดล้อม
- ฤดูกาลเพาะปลูก
- การให้น้ำ
- การให้ปุ๋ย
- อายุหรือช่วงเวลาเก็บเกี่ยว
- ผลิตผลเฉลี่ยต่อไร่
ศัตรูพืช
- ชนิด
- ลักษณะ
- วิธีการกำจัด
โรคพืช
- ชนิด
- ลักษณะ/อาการ
- วิธีการรักษา
วัชพืช
- ชนิด
- วิธีการกำจัด
การแปรรูป
ข่าวสารทางการเกษตร
แหล่งซื้อวัสดุดิบ/ราคา
ราคาผลผลิต
การให้น้ำ
การเก็บเกี่ยว และเทคโนโลยีหลังเก็บเกี่ยว
การให้ปุ๋ย
พยากรณ์อากาศ/ประกาศเตือนภัย
เทคโนโลยีทางการเกษตร
วิธีการปลูก
แหล่งรับซื้อ/ตลาดรับซื้อ

Deployment

Knowledge portal as

One Stop Service

Better living condition of Agriculture

Finally,

- We have just initiated an open source Digital Library since it will be the back bone of e-learning for both formal and informal education.

Knowledge based Society and Economy,
Acadamic Knowledge Factory and Knowledge Park

so on.

Acknowledgement

- KURDI: Kasetsart University Research and Development Institute
- Graduate School of Kasetsart University
- IADLC2005 Chairs and Organizer