

# Digitizing Chemical Structure Images of Japanese Published Patent Applications into a Searchable Format

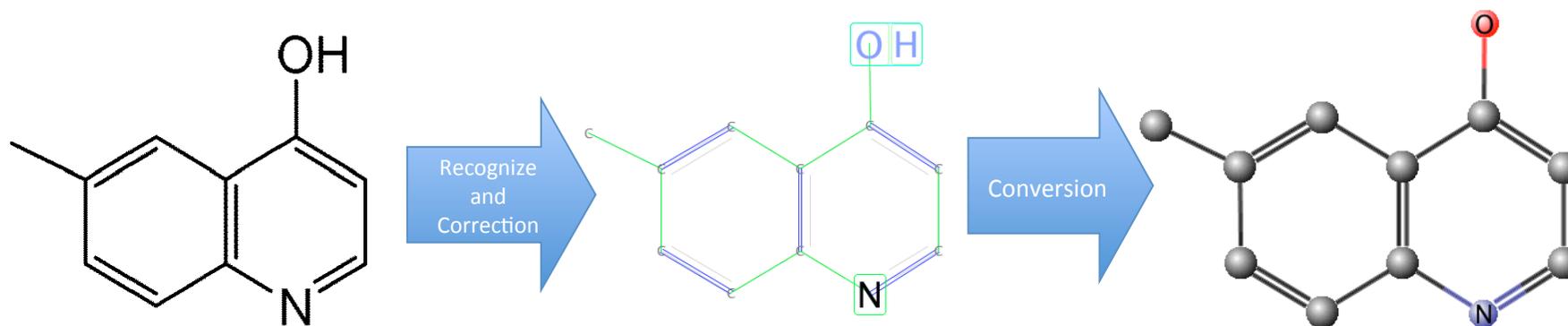
Koji Nakagawa<sup>1</sup>, Daniel Karzel<sup>1</sup>, Akio Fujiyoshi<sup>2</sup>, and  
Masakazu Suzuki<sup>1</sup>.

1 Faculty of Mathematics, Kyushu University, Japan;

2 Faculty of Engineering, Ibaraki University, Japan.

# Digitization into a Searchable Format

An Example taken from  
Japanese Published Patent 001630 in the year 2008



Input Images  
PNG or TIFF etc.

Internal Format  
COF

Searchable Format  
MOL, SDF etc.

Correspondence between  
graphical info.(pos. of chars or lines and its type)  
and  
chemical info. (connection table)

# ChemInfty Project

- Now in the Last Year of 3Years Project
- Universities, Research Institute, and Local Companies are involved in.
- Meant for Business
  - Technology Transfer Program from Univ.
- Based on the technology of Mathematical OCR system: InftyReader

# Business Model

## 1. Digitization Service

- We convert structural images to a searchable format.

## 2. Digitization System Product

- We sell the searchable data creation system.

# ChemInfty System

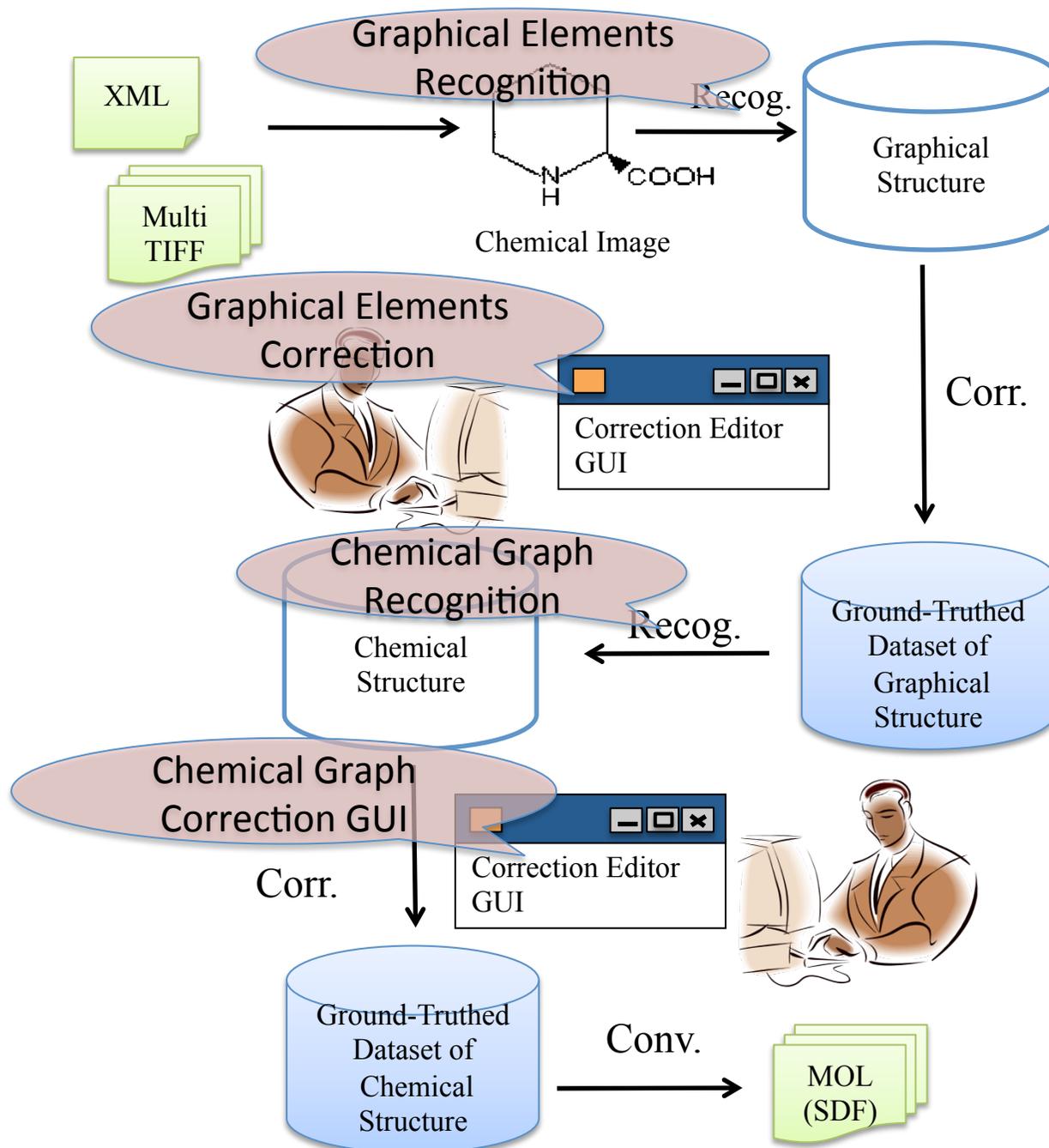
# ChemInfty System

- Recognition result cannot become perfect, so correction by human is needed.
- ChemInfty System consists of Recognition Engine and Correction GUI
- Platform
  - Graphical Recognition (runs on Windows)
  - Chemical Graph Recognition(runs on JavaVM)
  - Correction GUI(runs on JavaVM)

# Supported Recognition Elements

- Bonds
  - Single, Double, Triple
  - Chiral Bonds(Hashed, Wedge)
  - Aromatic Bond (a Circle in a Ring)
- Expansion of Rational formula and Abbreviation formula to Structural formula
  - e.g. -CH<sub>2</sub>COOH, -SO<sub>3</sub>Na, -ONa, -Me, -Et, -tBoc
- Markush structure
  - Position Variation
  - Frequency Variation

# Data Creation Flow



# Alive Systems of Chemical Structure Image Recognition

- ChemInfty
- OSRA
- ChemoCR
- ChemReader
- CLiDE Pro
- Imago
- MolRec

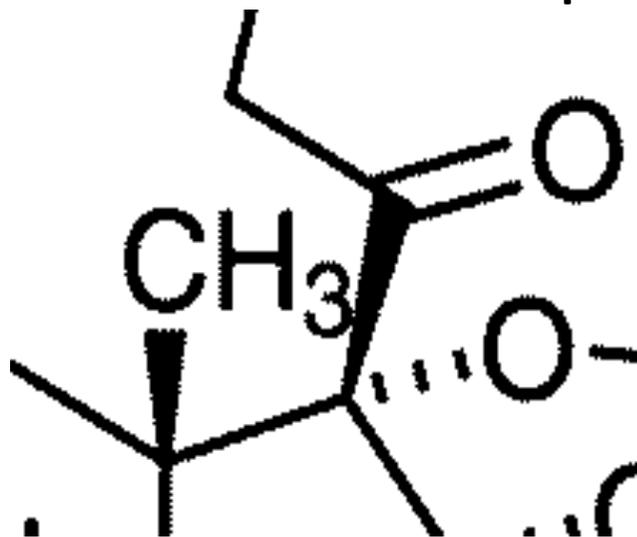
# Alive Systems of Chemical Structure Image Recognition

- ChemInfty
- OSRA
- ChemoCR
- ChemReader
- CLiDE Pro
- Imago
- MolRec

# Characteristics of ChemInfty

# Touching Processing

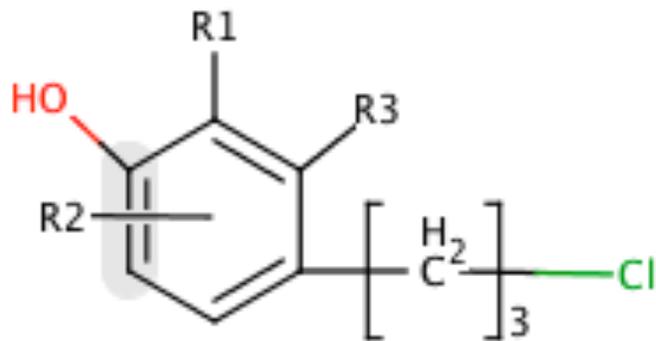
- Touching line-character and character-character causes the bad quality recognition



- ChemInfty treats it well

# Recognition of Markush

- Used frequently in Patents

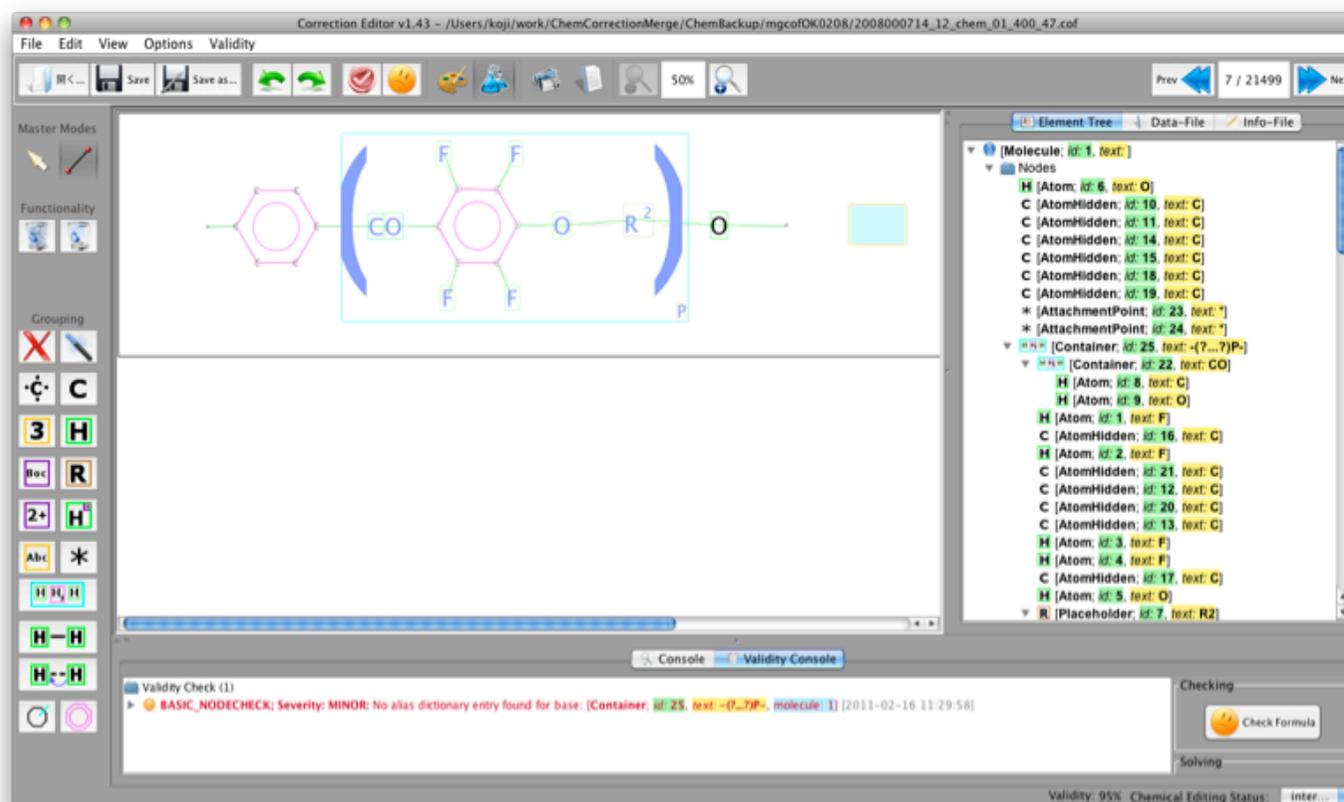


R1 is methyl or ethyl  
R2 is amino  
R3 is alkyl or an  
oxygen-containing heterocycle

- Supported by our system

# Correction GUI

Efficient Correction is possible by using chemical inconsistency



# **Digitization of Japanese Published Patents in 2008**

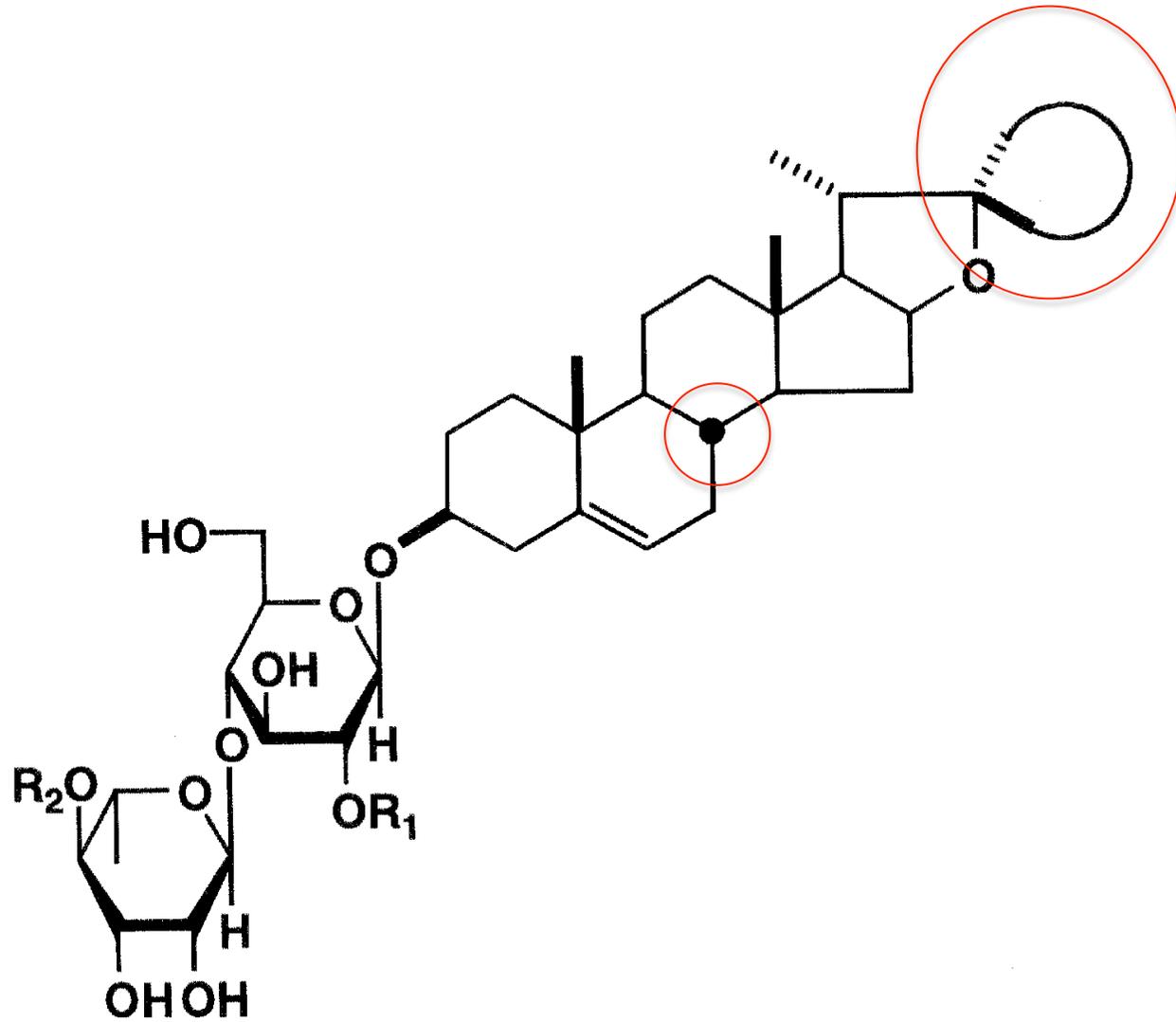
# Target of Digitization

- The first week of Japanese Published Patent issued in the year 2008
  - XML and MultiTIFF
- Tagged by <chemistry>
  - 3813 images
    - Digitized in this turn: 3381 images
    - Not Digitized in this turn: 432 images

# Not Digitized

- Skipped in this turn
  - Chemical Reaction
  - R- substitution of Markush
- The ones that cannot be digitized into common formats (e.g. SDF).
  - Digitization way is not fixed or not commonly accepted.

# Not Digitized Example



# Human Resource(Time and Money) needed to digitize

- Processed Images
  - Graphical Elements: 3608 images
  - Chemical Graph: 3381 images
- 5 persons worked
  - Some are skillful, some are not
  - Some training needed
- During the process the GUI was improved.
- Rough Estimation of Time and Cost
  - 8.47mins./image
  - =3.05mins/chemicalFormula (2.77chemicalFormulae/image)
  - =\$0.662/chemicalFormula(\$13.0/Hour=\$0.217/min)

# Discussion

# Discussion

- Quality Control
  - How much quality do we need?
  - Wrongly Drawn Chemical Formulae
- What to digitize
  - Quality of Images(Clearly printed or Noisy one)
  - Different areas need different ways of digitization
    - Digitization satisfying all users' needs is impossible!

# Future Work

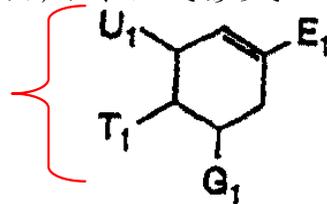
# Text Understanding(Minining)

- Understanding Natural Language text in the description of R-Group

(57) 【特許請求の範囲】

【請求項1】以下の式を有する、化合物ならびにその塩、溶媒和物、分割されたエナンチオマー、および精製されたジアステレオマーであって：

画像



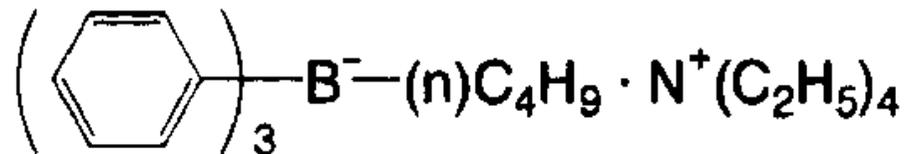
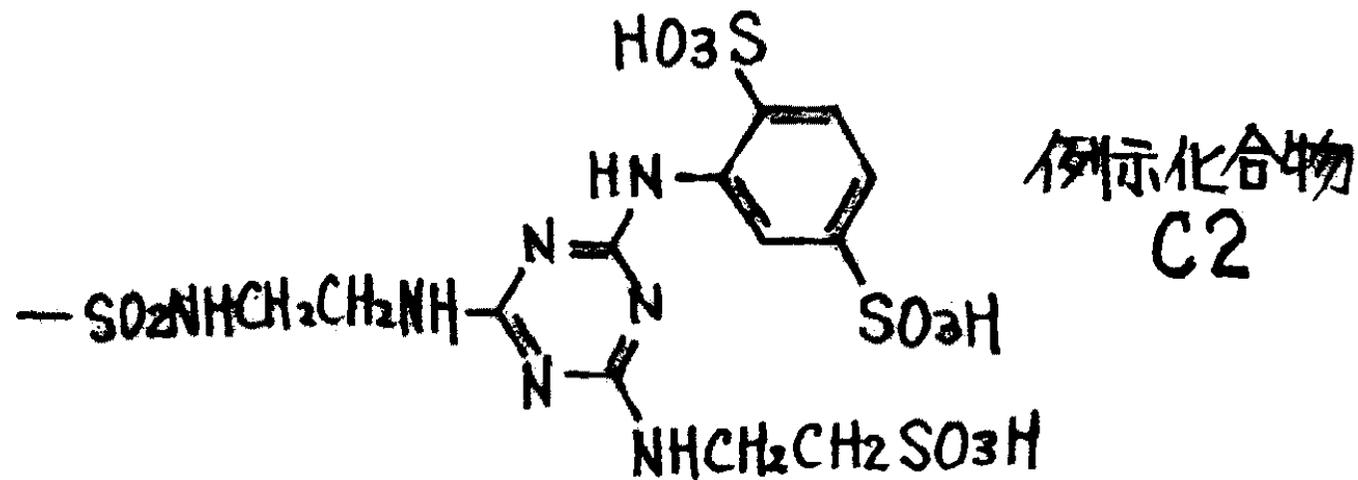
ここで：

E<sub>1</sub>が、-CO<sub>2</sub>H、-CO<sub>2</sub>R<sub>5</sub>、-CO<sub>2</sub>R<sub>5a</sub>W<sub>5</sub>または-CO<sub>2</sub>W<sub>5</sub>であり；

G<sub>1</sub>が、-N(R<sub>11</sub>)<sub>2</sub>、-N(R<sub>11</sub>)C(N(R<sub>11</sub>))(N(R<sub>11</sub>)<sub>2</sub>)、または-C(R<sub>11</sub>)<sub>2</sub>-N(R<sub>11</sub>)<sub>2</sub>であり；

テキスト

# Rational Formulae Expansion



**THANK YOU!**