



## **MMS Training Workshop**

January 2011

## MMS

Combining the patent office expertise of



with the indexing expertise of



THOMSON REUTERS

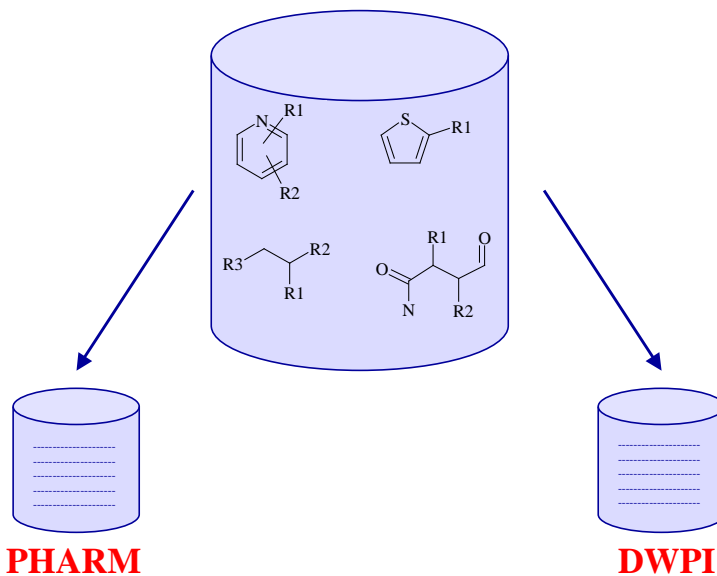
### • Introduction

The Merged Markush Service (MMS) is an extensive structure searchable patent information service for the Pharmaceutical and Chemical communities. MMS is the result of the cooperative agreement between INPI and Thomson Reuters.

• **INPI**, *Institut National de la Propriété Industrielle*, is the French Patent & Trademark Office and the French Register of Commerce & Trade. INPI registers patents, trademarks, industrial designs and companies. One of its missions is to make information concerning patents, trademarks, designs and companies available. To achieve this mission, INPI relies on its public libraries, on its Competitive Intelligence Search Service and on its information products (15 databases and 4 CD-ROM titles on international intellectual property). INPI provides worldwide access to over 35 million records of technical and business information.

• **Thomson Reuters** is a leader supplier of business-critical information to companies and research institutes across the world. It is a leader in technological intelligence and has earned an international reputation for its analysis of patent and scientific data. The company has been supplying the needs of Fortune 50 companies for over 50 years. Thomson Reuters headquarters are in London, it also has branch offices around the world, including the United States and Japan.

## MERGED MARKUSH SERVICE

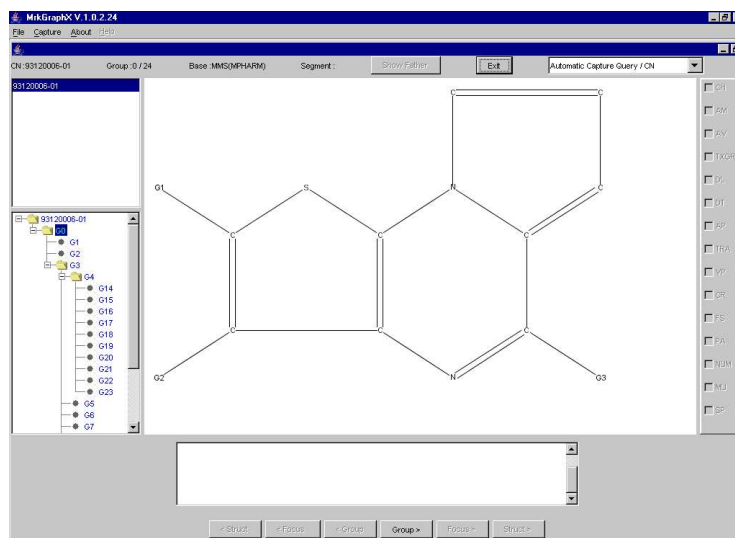


MMS 01-2011

1 - 2

- **Introduction**

**Approx. 3,000,000 structure records**

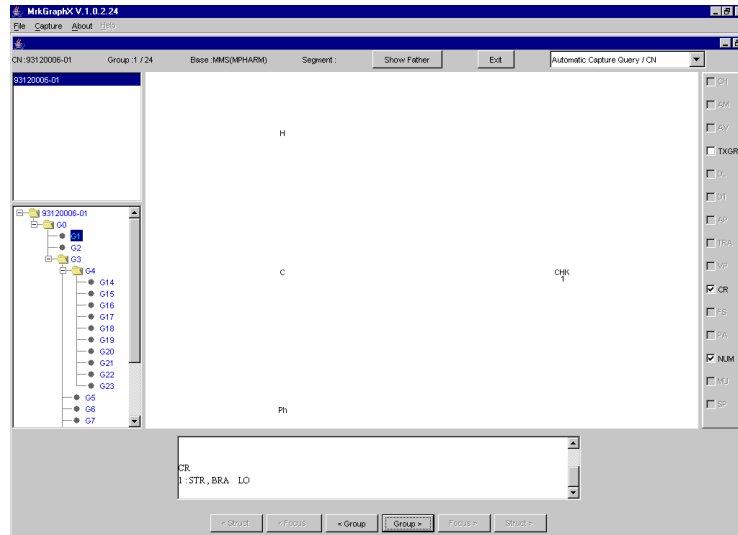


MMS 01-2011

I - 3

- **Introduction**

## Approx. 2,800,000 structure records

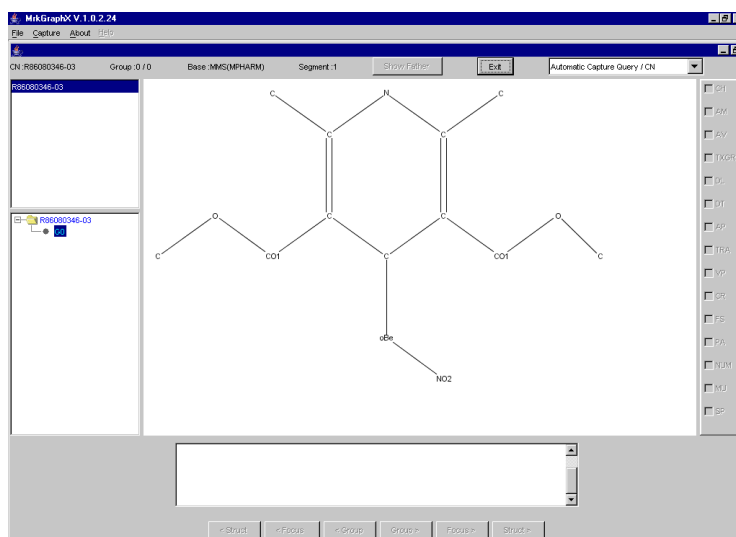


MMS 01-2011

I - 4

- **Introduction**

## Approx. 3,000,000 structure records



MMS 01-2011

I - 5

- Introduction

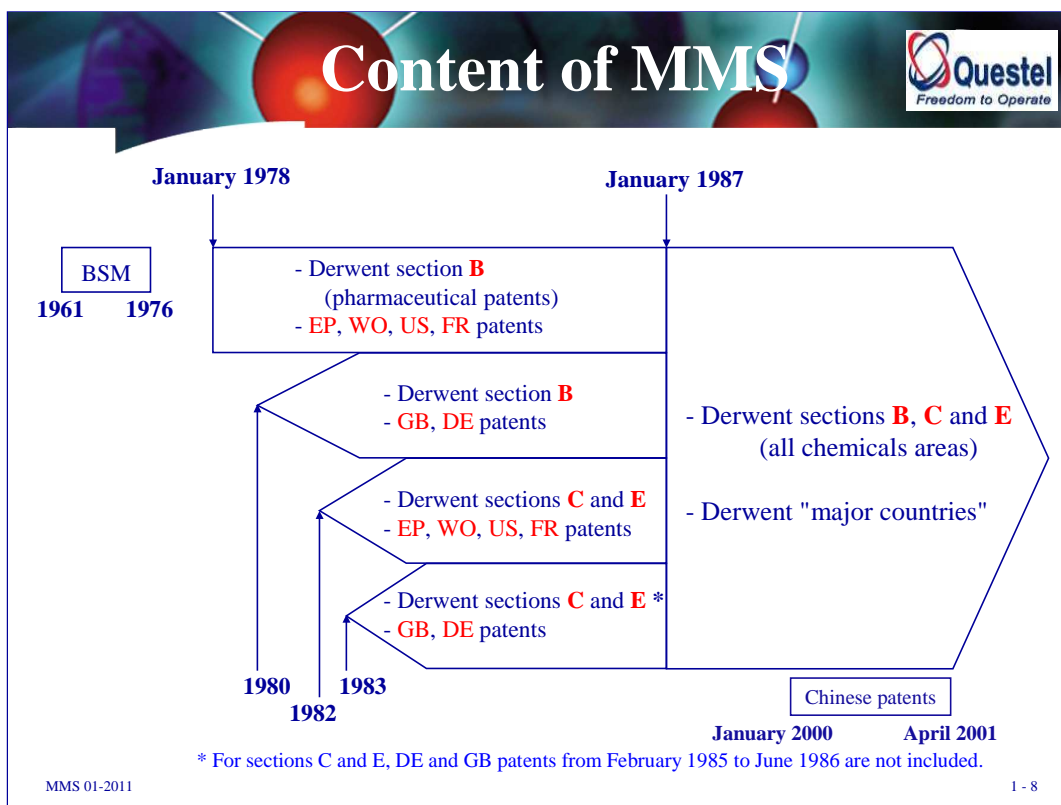
## **MMS contains:**

- $\approx$  50% Markush structures
- $\approx$  50% single compounds

## **• Introduction**

- **Content of MMS**
  - Subject, period and country coverage
  - Structure coverage
  - Depth of indexing
- **Indexing policy**
  - Markush structures
  - Indexing tools
  - Structure representation conventions
- **Structure of the MMS file**

- **MMS overview**



## • Content of MMS

### - Post-87 patents:

Structures from patents in Derwent Sections B (Pharmaceutical and veterinary patents), C (Patents relevant to agriculture and veterinary medicine) and E (General chemistry including dyes and pigments) are indexed in MMS.

Only structures from patents published in those countries considered by Thomson Reuters to be "major countries" are indexed. Structures from "Equivalent" patents are not indexed if the parent "Basic" patent has already been indexed in MMS.

### - Pre-87 patents:

As far as the pharmaceutical area is concerned, EP, US and FR patents and PCT applications are indexed back to January 1978. DE and GB patents in the same area have been added back to 1980. Chemical patents from EP, US, FR, and PCT are indexed back to 1982 and the DE and GB patents in the chemical areas go back to 1983.

In addition, MMS also offers the complete BSM collection (Specific French Drug Patents).

## Structure coverage

- Compounds described as new
- Products of new processes, including materials purified in new ways
- Compounds which are removed\*
- Compounds used to effect removal\*
- Compounds that are analyzed or detected\*
- Compounds used in analysis or detection\*
- Catalysts that are new
- Important ingredients of new compositions
- Known compounds with new activities

*\* when this is important to the novelty of the invention*

## • Content of MMS

Patents for all areas of chemistry (Derwent's Sections B, C, and E) issued from January 1987 onward by one of the patent authorities considered as a major country by Derwent are covered in MMS. The Derwent Major Countries are:

- Austria - AT
- Australia - AU
- Belgium - BE
- Canada - CA
- European Patent Office - EP
- France - FR
- Germany - DE
- Japan - JP
- Netherlands - NL
- New Zealand - NZ
- Russia - RU
- South Africa - ZA
- Sweden - SE
- Soviet Union - SU
- Switzerland - CH
- United Kingdom - GB
- United States - US
- World Intellectual Property Organization - WO
- Research Disclosures - RD

# Content of MMS



## **Depth of indexing:**

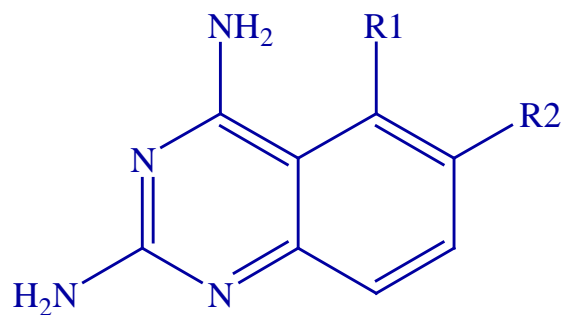
- Claims (generic and specific terms)
- Examples
- Disclosure (listed compounds)

- **Content of MMS**

- **Markush structures**
  - Markush structure description
  - Markush structures in MMS
- **Indexing tools**
- **Structure representation conventions**

- **Indexing policy**

## Markush structure description



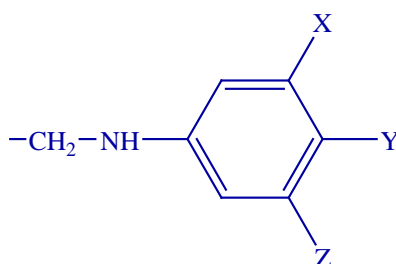
- **Indexing policy**

- **Markush structure description**

Consider the Markush formula which is shown on this slide.

## Markush structure description

- **R1** is methyl, ethyl or a lower alkyl group
- **R2** is cyano, Br, Cl, halogen or a group of formula:



**X, Y and Z:** H, Br, Cl, halogen, -O-CH<sub>3</sub>, alkoxy or -CF<sub>3</sub>

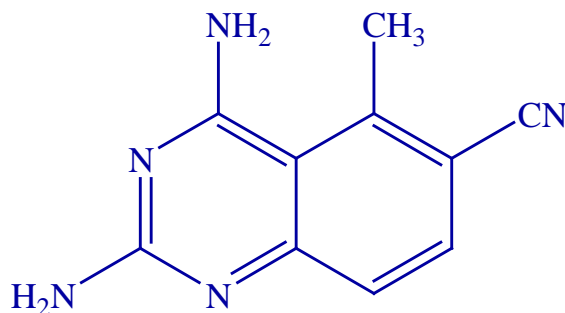
- **Indexing policy**
  - **Markush structure description**

In this formula there are two variable parts - R1 and R2 - which can be:

R1: methyl, ethyl or a lower alkyl group

R2: cyano, Br, Cl, halogen or a phenylaminomethyl group where X, Y and Z, which may be the same or different, are hydrogen, Br, Cl, halogen, methoxy, lower alkoxy, -O-CH<sub>3</sub> or trifluoromethyl.

## Markush structure description



**R1: methyl / R2: cyano**

- **Indexing policy**

- **Markush structure description**

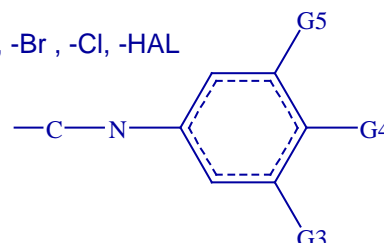
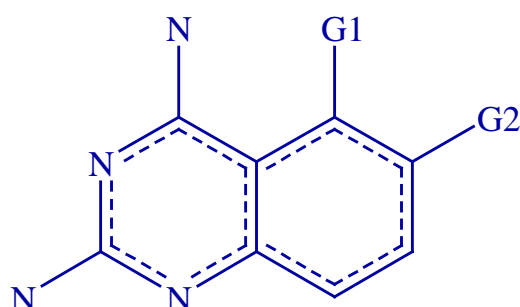
One of the specific compound encompassed by the Markush formula is:  
2,4-amino-5-methyl-6-quinazolinonitrile.

This compound is formed by the combination of the fixed part (2,4-quinazolinediamine) with R1 as methyl and R2 as the cyano group.

## Markush structures in MMS

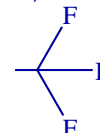
$G_1 = -C, -ET, -CHK^{LO}$

$G_2 = -CN, -Br, -Cl, -HAL$



$G_3 = G_4 = G_5 = -H, -Br, -Cl, HAL,$

$-O-C, -O-CHK^{LO}$



MMS 01-2011

I - 15

Compounds which are described in a patent document as a possibility of a Markush structure are recorded in the MMS structure database as one or more Markush structures covering the specific examples as well as the generic information which is extracted from the claims.

Structures are represented in the MMS file as graphs, i.e. nodes (or positions) and the connections between the nodes.

Markush structures stored in the structural MMS file have the following features:

- A fixed part and variable groups (symbol G followed by a numerical suffix in the range 1 to 50)
- The fixed parts and the variable parts attached thereto form the main group G0 (group zero)
- A variable group may contain additional variable groups, i.e. a variable value may be an isolated variable group or a variable group attached to a group of atoms; there may be up to 4 levels of nesting
- Generic expressions such as "alkyl" are expressed as Superatoms
- Specifications relating to atoms (e.g. charges, abnormal mass, etc...) or Superatoms (e.g. "lower for alkyl") are expressed as attributes on these atoms or Superatoms
- Information which cannot be conveyed graphically (e.g. carbon chain length) is expressed as strings of Text Notes which are displayed under the graphical information.

### Indexing tools

- Atoms
- Shortcuts
- Superatoms
  - Superatom attributes
  - Text notes
- G groups

- **Indexing policy**
  - **Indexing tools**

## Indexing tools

- **Atoms:**

All atoms from the Mendeleiev periodic table

- **Shortcuts:**

- Carbon chains: ET, NPR, IPR, NBU, SBU, TBU, IBU, Cn (C1, C2 ... )
- Benzene ring: PH, OBE, MBE, PBE
- Functions: CO1, CO2, SO2, SO3, PO3, PO4, CN, NO2, ACE

- **Indexing policy**

- **Indexing tools**

<b>Shortcuts</b>	<b>Definition</b>	<b>Shortcut</b>	<b>Definition</b>
<b>CO2</b>	-COOH	<b>ET</b>	ethyl
<b>CO1</b>	-C(O)- (divalent)	<b>NPR</b>	n-propyl
<b>SO2</b>	-SO2- (divalent)	<b>IPR</b>	isopropyl
<b>SO3</b>	-S(O)2(OH)	<b>NBR</b>	n-butyl
<b>PO3</b>	-P(O)2(OH)	<b>SBU</b>	s-butyl
<b>PO4</b>	-P(O)(OH)2	<b>TBU</b>	tert-butyl
<b>CN</b>	-CN	<b>IBU</b>	isobutyl
<b>NO2</b>	-NO2	<b>PH</b>	phenyl
<b>ACE</b>	-C(O)CH3		

## Indexing tools: Superatoms

- **Acyclic hydrocarbons:**

**CHK** Alkyl or alkylene

**CHE** Alkenyl or alkenylene

**CHY** Alkynyl or alkynylene

- **Carbocyclic hydrocarbons:**

**ARY** Carbocyclic system, optionally fused, containing at least one benzene ring (aryl)

**CYC** Cycloaliphatic, optionally fused, non-aromatic carbocycle

- **Indexing policy**

- **Indexing tools**

- **Superatoms**

Superatoms are used in the MMS structural database to express chemical families which are described in patent documents by generic expressions, such as alkyl, halogen, protecting group, etc...

Each Superatom is represented by a two or three letter symbol, similar to the symbols for atoms in the periodic table.

Superatoms are used to represent five classes of structural features.

## Indexing tools: Superatoms

- **Heterocyclic systems:**

- HEA** Monocyclic aromatic heterocycle (heteroaryl)

- HET** Non-aromatic monocyclic heterocycle

- HEF** Fused heterocycle

- **Indexing policy**

- **Indexing tools**

- **Superatoms**

## Indexing tools: Superatoms

- **Metals:**

- MX** Any metal

- AMX** Alkali or alkaline earth metals

- A35** Group IIIA - VA metals

- TRM** Transition metals (excluding lanthanum)

- LAN** Lanthanides (including lanthanum)

- ACT** Actinides

- **Indexing policy**

- **Indexing tools**

- **Superatoms**

## Indexing tools: Superatoms

- **Other:**

**HAL** Halogens

**ACY** Acyls

**PRT** Protecting group

**POL** Polymer or polypeptide residue

**DYE** Dye group residue

**XX** Any atom or group excluding hydrogen (display only)

- **Indexing policy**

- **Indexing tools**

- **Superatoms**

## Indexing tools: Superatom attributes

- **Acyclic hydrocarbon attributes:**

- Chain length

<b>LO</b>	Low (1 to 6 carbon atoms)
<b>MID</b>	Middle (7 to 10 carbon atoms)
<b>HI</b>	High (11 or more carbon atoms)

- Chain type

<b>STR</b>	Straight
<b>BRA</b>	Branched

Some precise pieces of information relating to Superatoms are expressed as attributes associated with the corresponding Superatom. Attributes may be associated with the acyclic hydrocarbon, cyclic system and amino acid Superatoms.

– **Acyclic hydrocarbon attributes**

Acyclic hydrocarbon Superatoms may have two classes of attributes associated with them. These attributes describe:

- Chain length

<b>LO</b>	for low (1 to 6 carbon atoms)
<b>MID</b>	for middle (7 to 10 carbon atoms)
<b>HI</b>	for high (11 or more carbon atoms)

The default value for chain length is low, middle, high (LO, MID, HI); thus, if no specification relating to chain length is given in the patent document for alkyl, alkenyl or alkynyl, these chains are assumed to be low, middle, high (LO, MID, HI).

- Chain type

<b>STR</b>	for straight
<b>BRA</b>	for branched

The default value for chain type is both straight and branched; thus, if no specification is given in the patent relating to chain type, both the STR and BRA attributes are assigned.

## Indexing tools: Superatom attributes

- **Cyclic system attributes:**

- Ring type

**MON** Monocyclic

**FU** Fused

- Degree of saturation

**SAT** Fully saturated

**UNS** Unsaturated

### Cyclic system attributes

Cyclic system Superatoms may have two classes of attributes associated with them. These attributes may not apply to all cyclic Superatoms. The cyclic system attributes describe:

- Ring type

**MON** for monocyclic

**FU** for fused

These attributes may be applied to the ARY and CYC Superatoms. The default value for the ring type attributes is both monocyclic and fused; thus, if no specification is given in the patent document relating to ring type, both the MON and FU attributes are assigned.

- Degree of saturation

**SAT** for fully saturated

**UNS** for unsaturated

These attributes may be applied to the CYC, HEF and HET Superatoms. The default value for the degree of saturation attributes is both saturated and unsaturated; thus, if no specification is given in the patent document relating to the degree of saturation, both the SAT and UNS attributes are assigned.

## Indexing tools: Text notes

- **Number of carbon atoms:** CHK, CHE, CHY, CYC
- **Number of double bonds:** CHE, CHY
- **Number of triple bonds:** CHY
- **Number of rings:** ARY, CYC, HEF
- **Number of ring members:** ARY, CYC, HEA, HET, HEF
- **Number of specific heteroatoms:** HEA, HET, HEF
- **Atom of attachment:** HEA, HET, HEF, CYC
- **Number of repetitions for repeating units**

- **Indexing policy**

- **Indexing tools: Text Notes**

There are some pieces of information given in a patent document which cannot be conveyed graphically in the structure file. This kind of information is conveyed by Text Notes which are displayed with the graphical representation of the structure. There are three types of information which are given in this fashion:

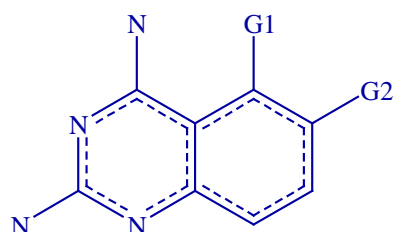
**Information applying to the entire Markush structure**, such as stereochemical information or provisos mentioned in the patent document. This information, which may be in a free text form, is provided at the main group (G0) level. Free text is always preceded and followed by a slash (/).

**Information relating to a particular Superatom**, such as "1 to 3 carbon alkyl". This information is provided in a controlled text form at the level of the group which contains the Superatom to which the information applies.

**Information on the number of repetitions in a repeating unit:** this information is indicated by controlled text at the level of the group which contains the repeating unit.

## G groups

Fixed part



Variable part

G1 = -C ; -ET ; -CHK<sup>LO</sup>

*50 G groups, 4 levels of nesting, 1023 atoms*

- Indexing policy
  - Indexing tools

## **Structure representation conventions: bond normalization**

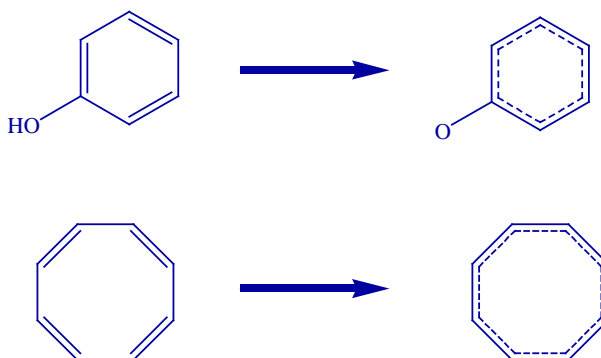
- Aromaticity
- Tautomerism
- Enol-Keto tautomerism

- **Indexing policy**
  - **Structure representation conventions**

## Structure representation conventions

- **Aromaticity**

The bonds in rings containing  $2n$  atoms with  $n$  alternating single and double bonds are normalized. This normalization takes priority over other rules such as tautomerisation.



MMS 01-2011

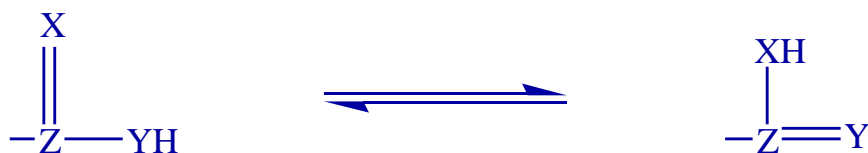
1 - 27

- **Indexing policy**

- **Structure representation conventions**

## Structure representation conventions

- **Tautomerism**



*Where:*

*X, Y are O, S, Se, Te, N*

*Z is B, C, Si, N, P, As, S, Se, Te, F, Cl, Br or I*

- **Indexing policy**

- **Structure representation conventions**

## Structure representation conventions

- **Tautomerism**

If X and Y are different, the double bond is placed preferentially on the bond to the atom which is first in the sequence:

**O, S, Se, Te, N**



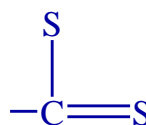
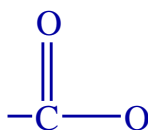
- **Indexing policy**

- **Structure representation conventions**

## Structure representation conventions

- **Tautomerism**

If X and Y are the same but not N, one double bond and one single bond are used.



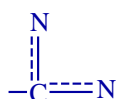
- **Indexing policy**

- **Structure representation conventions**

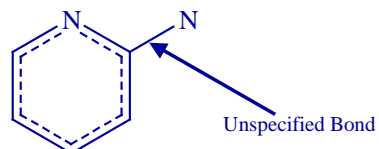
## Structure representation conventions

- **Tautomerism**

If both X and Y are N, normalized bonds are used.



but



**Note:** This rule IS NOT consistently applied to amino groups on normalized ring systems. Therefore, we recommend that you use an unspecified bond on the ortho bond to the N.

- **Indexing policy**

- **Structure representation conventions**

## Structure representation conventions

- **Enol-Keto tautomerism**

If the heteroatom is O, S, Se, or Te, the keto form is preferentially indexed.



X = O, S, Se, Te

- **Indexing policy**

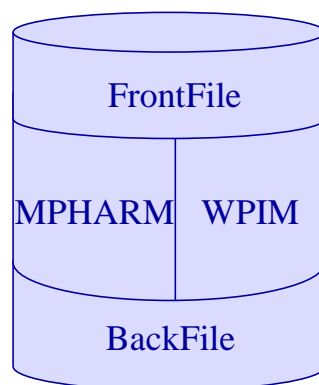
- **Structure representation conventions**

- Introduction
- Content of MMS
  - Subject, period and country coverage
  - Structure coverage
  - Depth of indexing
- Indexing policy
  - Markush structures
  - Indexing tools
  - Structure representation conventions
- Structure of the MMS file

- **MMS overview**

## Four segments:

- FrontFile - from Derwent week 9816
- MPHARM – 1987-1999
- WPIM – 1987-1998 (week 15)
- BackFile – Before 1987

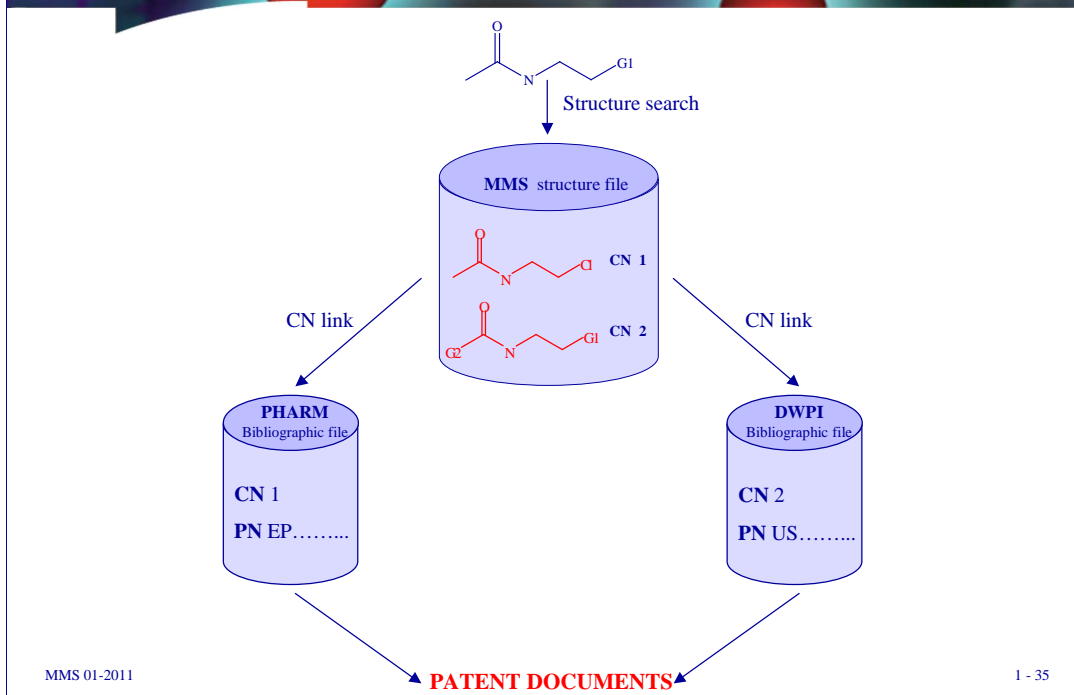


## • Structure of the MMS file

The MMS file comprises four segments:

- **BackFile:** Retrospective coverage of the chemical, agrochemical and pharmaceutical patents.
- **MPHARM,** corresponding to the MPHARM file up to the end of 1999: this is a non-growing segment
- **WPIM,** corresponding to the WPIM file up to June 1998: this is a non-growing segment
- **FrontFile:** it is growing with the coverage of the newly published patents.

Structure searches may be performed in all, one or more segments.



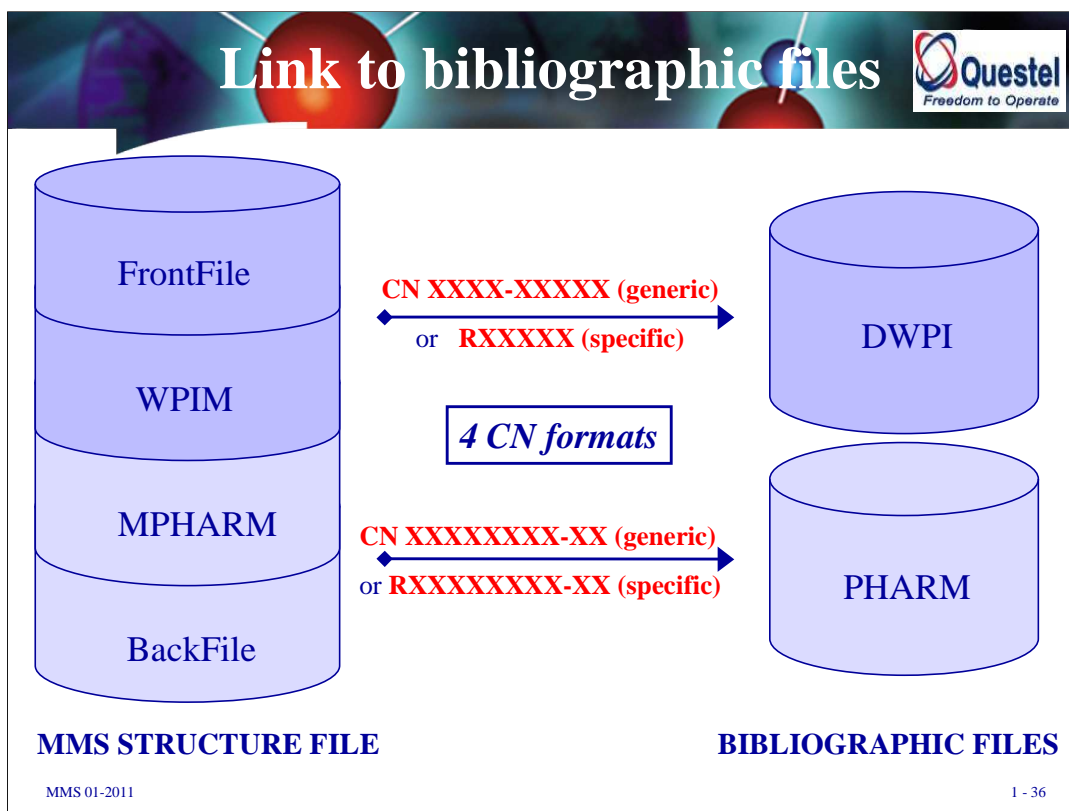
## • Link to bibliographic files

Each structure record of the MMS structure file is identified by a unique Compound Number (CN) and is associated with a reference in at least one of the bibliographic files PHARM (produced by INPI) and DWPI (produced by Thomson Reuters).

The Compound Number of an MMS structure is contained in the CN field of the associated bibliographic reference(s) of the bibliographic files (PHARM and/or DWPI).

**The Compound Number is thus the link between the MMS structure file and the bibliographic files.**

A structure search performed in the MMS structure file using a Markush query (or a specific query) provides a list of Compound Numbers, which are transferred to the bibliographic files for patent retrieval.



### Link to bibliographic files:

#### Two CN formats for Markush structures:

The **XXXXXXXX-XX** or **RXXXXXXXX-XX** Compound Number format identifies structures of the MPHARM and BackFile segments of the MMS file. Structures of the MPHARM and BackFile segments only have a reference in the PHARM bibliographic file.

The **XXXX-XXXXX** or **RXXXX-XXXXX** or **RXXXXXX** Compound Number format identifies structures of the WPIM and FrontFile segments of the MMS file. Structures of the WPIM and FrontFile segments only have a reference in the DWPI bibliographic file.

All CNs starting with the letter R represent specific structures.