



DERWENT  
WORLD PATENTS INDEX

**CPI Chemical Indexing Guidelines**  
Indexing of Chemical  
and Pharmaceutical Patents

© 2001 Thomson. All rights reserved

Edition 2  
ISBN: 1 903836 01 X

Copyright 2001 Thomson  
Published by Thomson Derwent  
14 Great Queen Street, London WC2B 5DF,  
United Kingdom

Visit the Thomson Scientific web site at <http://www.thomsonscientific.com/>

First edition published September 1995  
Second edition published April 2001

ISBN: 1 903836 (Edition 2)

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means – electronic, mechanical, recording, photocopying or otherwise – without express written permission from the copyright owner.

Derwent gratefully acknowledges the work of Dr. Paul Bryan, who prepared the bulk of the text for the first edition of the Guide, published in 1995. This second edition has been updated to reflect the changes in Derwent products introduced since the first edition was published.



# Contents

<b>1</b>	<b>About this manual .....</b>	<b>1</b>
<b>2</b>	<b>About Derwent .....</b>	<b>2</b>
<b>3</b>	<b>Help Tools and Customer Assistance .....</b>	<b>3</b>
	Printed Help Materials .....	3
	Classes .....	4
	Derwent Help Desk .....	4
	Markush TOPFRAG .....	5
<b>4</b>	<b>CPI Sections B, C, and E .....</b>	<b>6</b>
	Overview of Derwent Patent Classification .....	6
	Coverage of Sections B, C, and E .....	7
<b>5</b>	<b>Types of BCE indexing .....</b>	<b>12</b>
	Introduction .....	12
	Chemical (Fragmentation) Codes .....	13
	Markush DARC graphical indexing .....	13
	Derwent Chemistry Resource indexing .....	15
	Manual Codes .....	16
	Ring Index Numbers .....	16
	Registry Numbers .....	17
<b>6</b>	<b>Types of compounds indexed .....</b>	<b>22</b>
	Introduction .....	22
	Types of compounds selected to be fully indexed .....	22
	Corresponding roles and essential chemical codes .....	23
<b>7</b>	<b>Levels of indexing detail .....</b>	<b>26</b>
	Introduction .....	26
	Patents of Average Complexity .....	27
	Level 3 : Very involved patents .....	30

<b>8 AutoCPI .....</b>	<b>36</b>
Introduction .....	36
Compounds that require the manual addition of codes .....	37
Compounds that cannot be indexed by AutoCPI .....	38
AutoCPI indexing of superatoms .....	40
Changes to indexing policy due to AutoCPI .....	41
<b>Appendix 1 Text Notes .....</b>	<b>43</b>
Controlled Text .....	43

# 1 About this manual

The material in the first edition of this User Guide was originally written for use as an internal policy document at Derwent, listing guidelines for indexers of chemical and pharmaceutical patents. That document was revised and presented to Derwent's patent information customers in 1995 as the "Chemical Indexing Guidelines" manual, in order to give advanced users insight into how Derwent indexes patents. This 2<sup>nd</sup> Edition is the first published revision and is intended to bring the Guidelines up to date. It also includes information about changes to the original guidelines since the introduction of Derwent Chemistry Resource Indexing as well as details of the Derwent Chemistry Resource Indexing Guidelines.

Section 4 – *CPI Sections B,C and E* introduces the reader to Derwent Sections and Classes, which are the categories used to classify patents at Derwent. Sections B, C, and E are explained in detail, as these are the Sections whose indexing is described in this manual.

Section 5 – *Types of BCE Indexing* describes the different types of indexing tools, languages, codes, and terms used by Derwent to index chemical and pharmaceutical patents. Most are described briefly because they receive extensive treatment in other manuals, but Text Notes and Registry Numbers are discussed in some detail.

Section 6 – *Types of compounds indexed* lists the criteria used by analysts to select compounds from a patent document for full indexing, and identifies the index terms that are automatically assigned to compounds based on their role in the patent.

Section 7 – *Levels of indexing detail* discusses the levels of indexing detail applied to different types of patents.

Section 8 – *AutoCPI* describes AutoCPI, the computer software used internally at Derwent to facilitate the rapid creation of detailed indexing records for patents.

*Appendix 1* give further details about Text Notes.

## 2 About Derwent

Derwent, the leading specialist in scientific and patent information, has for over 50 years provided vital information to companies and research institutes across the world.

Derwent World Patents Index (Derwent WPI) is unrivalled in its comprehensive, enhanced patent information covering more than 8 million separate inventions from 40 patent-issuing authorities including the USPTO, WIPO, EPO, Japanese and German patent offices.

Derwent's products are designed to meet the needs of not only the major multi-nationals, but equally importantly, to fulfil the information demands of smaller, more specialised organisations.

Used by a global audience, Derwent's information products give a comprehensive picture of technological innovations world-wide – providing critical advantage by highlighting new opportunities, identifying competitors and assisting R&D.

As part of The Thomson Corporation, Derwent works closely with global leaders in the information industry to guarantee customers access to unequalled business and technological intelligence.

# 3 Help Tools and Customer Assistance

The resources listed below are provided by Derwent to help customers learn how to search effectively for chemical patent information.

## Printed help materials

- The **Chemical Code Dictionary** is an alphabetical subject listing that refers the user from concepts and nomenclature to the corresponding Chemical Codes.
- The **Chemical Coding Sheet** is a concise summary of the BCE Chemical Codes. It gives an abbreviated description of each code, provides an overview of the logical sequence of the codes, and indicates the year that each code became available for searching.
- The **Chemical Retrieval Manual** is a reference work that describes indexing rules, Chemical Code definitions, and the construction of code search strategies.
- **CPI Manual Codes** describes the use of Manual Codes and lists all Manual Codes, and the time periods for which they are relevant.
- **Global Patent Sources** discusses the structure of patents, describes the sources from which Derwent receives patent documents, and gives an overview of Derwent's patent products and services.
- **Derwent Registry Compounds** listed in alphabetical order contain Registry Numbers, Specific Compound Numbers and Derwent Chemistry Resource numbers are provided in this reference guide.
- The Markush DARC User Manual discusses Markush structure concepts and teaches users how to access and search the Derwent World Patent Index Markush database using Markush DARC search software.
- The host specific Online User Guides discusses the Derwent World Patents Index database and the many methods that can be used to access its contents.

## Classes

The following classes are available for customers who wish to learn more about the various Derwent online products.

- Searching Derwent World Patents Index Online
- Introduction to Chemical Code Searching
- Advanced Chemical Code Searching
- Markush DARC and Derwent World Patents Index Markush. This manual is only available through INPI/Millennium Info. For a copy of the MMS user guide please go to <http://www.millenniuminfo.com>

## Derwent Help Desk

Expert advice and support is available via our Customer Technical staff, to provide a fast and efficient response to all your enquiries. The experienced Technical Support staff have an in-depth knowledge of all Derwent's products and services and are familiar with the command languages of the various online hosts.

From general customer queries through to complex technical questions, the Technical Support department is there to help you.

You can contact your local Technical Support desk by phone, fax or email, or visit the Customer area on the Derwent web site.

**Europe/International**

Tel: +44 (0)20 7344 2999  
Fax: +44 (0)20 7344 2900  
Email: [ts.support.emea@thomson.com](mailto:ts.support.emea@thomson.com)

**North America**

Tel: +1 703 706 4220  
Toll Free: +1 800 451 3551  
Fax: +1 703 838 0450  
Email: [custserv@derwentus.com](mailto:custserv@derwentus.com)

**Japan**

Tel: +81 3 5218 6500  
Fax: +81 3 5218 7840  
Email: [ts.support.jp@thomson.com](mailto:ts.support.jp@thomson.com)

**Derwent Web Site**

<http://www.derwent.com>  
<http://www.derwent.co.jp>

**Markush TOPFRAG**

Markush TOPFRAG is a microcomputer program that converts specific or variable chemical structures drawn by the user into Chemical Code and Markush DARC search strategies. See <http://thomsonderwent.com/products/patentresearch/markushtopfrag/> for details of the latest version and for ordering information.

## 4 CPI Sections B, C, and E

### 4.1 Overview of Derwent patent classification

To facilitate the indexing and retrieval of patent documents, Derwent classifies patents in various subject categories. At the broadest level, patents fall into one of the following three categories:

- Chemical
- Engineering
- Electronic and electrical

Each of these broad subject categories is further divided into **Derwent Sections** as follows:

#### **Chemical Patents Index (CPI)**

A	Polymers, plastics (Plasdoc)
B	Pharmaceuticals (Farmdoc)
C	Agricultural chemicals (Agdoc)
D	Foods, detergents, water treatment, and biotechnology
E	General chemicals (Chemdoc)
F	Textiles, paper
G	Printing, coating, photographic
H	Petroleum
J	Chemical engineering
K	Nucleonics, explosives, protection
L	Refractories, ceramics, cement, electroorganics, electroinorganics
M	Metallurgy

#### **Engineering Patents Index (EngPI)**

P	General
Q	Mechanical

#### **Electronic/Electrical Patents Index (EPI)**

S	Instrumentation
T	Computing and control
U	Semiconductors and electronic circuitry
V	Electronic components
W	Communications
X	Electrical Power Engineering

Section designations serve both as patent subject categories and as the basis of customer subscriptions to materials and services produced by Derwent. A single patent may be classified in more than one Section when warranted by its subject matter. For example, a patent describing a new dyestuff for polyamide textiles would be classified in Sections A, E, and F.

Patent Sections are subdivided into **Derwent Classes**, which are designated by a Section letter followed by two digits. Derwent Classes can be used in searches on specific topics. For example, the otherwise ambiguous word stem WARN can be combined with the Class designation X22 (“Automotive electronics”) to retrieve only those references pertaining to automotive warning devices.

The Derwent Classification System is described in more detail in the Derwent publication entitled **The Derwent Classification For All Technologies**, from which the material in the following section was derived.

The rest of this manual deals only with Sections B, C, and E, i.e. the Sections in which chemical and pharmaceutical patents are classified. The following section lists the criteria for determining which patents should be classified in Sections B, C, and/or E.

## 4.2 Coverage of sections B, C, and E

Derwent’s chemical patent coverage began in 1963 with **Farmdoc**, an indexing and abstracting service for pharmaceutical and veterinary patents. In 1965, the service was extended to agricultural patents via the **Agdoc** service. In 1970, general chemistry coverage was initiated and all of the different chemical services were grouped together into the Central Patents Index, now called the Chemical Patents Index (CPI). The individual services correspond to the CPI Sections as follows:

Farmdoc ..... CPI Section B

Agdoc ..... CPI Section C

Chemdoc ..... CPI Section E

In April 1993 the services were renamed, as shown below, but they continued to provide the same patent coverage.

**Pharmaceutical** indexing ..... **CPI Sections B and C**

**Chemical** indexing ..... **CPI Section E**

### 4.2.1 Coverage of Section B

The following types of patents are classified in Section B:

- Patents stated to be of pharmaceutical or veterinary interest, as well as those that refer to compounds used as intermediates in the manufacture of pharmaceutical or veterinary products.
- Patents on compositions used for diagnosis or analysis in the pharmaceutical and/or veterinary fields (e.g. stains for bacterial pathogens).
- Patents on artificial sweeteners, chemical warfare agents, and plaque-disclosing compositions.
- Patents dealing with the production of formulations, e.g. tablets, pills, capsules, suppositories, aerosols, etc. Also patents on devices specifically designed for dispensing pharmaceuticals, e.g. syringes, child-proof closures, calendar pill boxes, aerosol devices, etc.

Section B includes the following Derwent Classes:

- B01 Steroids.** Including systems with carbocyclic and/or heterocyclic rings fused to the basic steroidal ring structure.
- B02 Fused ring heterocyclics**
- B03 Other heterocyclics**
- B04 Natural products and polymers.** Includes testing of body fluids (other than blood typing or cell counting); pharmaceuticals and veterinary compounds of unknown structure; vaccines; testing of microorganisms for pathogenicity; testing of chemicals for mutagenicity or human toxicity; biotechnology and fermentative production of DNA or RNA. Also **general compositions**.
- B05 Other organics.** Includes aromatics, aliphatics, organometallics, and compounds whose substituents vary such that they would be classified in several of the classes B01 through B05.
- B06 Inorganics.** Including fluorides for toothpastes and other similar compounds.
- B07 General.** Includes tablets, dispensers, catheters (except catheters used for drainage or angioplasty), encapsulation, etc. Excludes systems for administration of blood, saline, intravenous feeding, etc.

### 4.2.2 Coverage of Section C

Patents covering compounds of agricultural and veterinary interest are classified in Section C, including:

- **Pest control agents** such as insecticides, miticides, rodenticides, molluscicides, slugicides, vermicides (nematocides, anthelmintics, etc.), pest repellents and attractants, and soil fumigants. Also biological control using microorganisms, predators, or natural products.
- **Chemical warfare agents**
- **Plant growth control agents** such as herbicides, weedicides, defoliant, desiccants, fruit drop and set controllers, rooting compounds, sprouting inhibitors, growth stimulants and retardants, moss and lichen controllers. Also **plant genetics**.
- **Plant disease control agents** such as fungicides, viricides, timber preservatives, and bactericides.
- **Soil improvement agents** such as fertilisers, trace metal additives, bacterial action control stimulants, and soil consolidation agents (if used for agricultural purposes).
- **Veterinary products** such as disease control agents, nutritional agents, and veterinary vaccines.

Section C includes the following Derwent Classes:

- C01 Organophosphorus, organometallics** - i.e. compounds that include elements other than H, C, N, O, S and halogen.
- C02 Heterocyclics**
- C03 Other organic compounds, inorganic compounds, and multi-component mixtures. Polymers and proteins.**
- C04 Fertilisers** – including urea and phosphoric acid production. Soil modifiers and plant growth media. Chemical aspects of **compost production**.
- C05 Biological control** – including use of microorganisms, predators, and natural products, but excluding veterinary medicine.
- C06 Biotechnology** – including plant genetics and veterinary vaccines.
- C07 Apparatus, formulation, general.** Includes veterinary syringes, general formulations in which the active compound is not central to the invention (e.g. wettable powders), and analysis.

### 4.2.3 Coverage of Section E

Most patents concerning the production, purification, use, detection, removal, or phase changes of **non-polymeric chemical compounds**, as well as the **apparatus** and **novel catalysts** for producing them, are classified in Section E.

**Exceptions** to this rule are:

- Compounds stated to be solely for use as a **pharmaceutical, veterinary medicament, fertiliser, herbicide, or pesticide** are classified only in Section B and/or Section C. However, if an **additional use** is stated, e.g. the compound is also a dyestuff intermediate, the patent is classified in Section B and/or Section C and additionally in Section E.
- **Monomers** taking part in a polymerisation reaction and **starting materials** for a chemical reaction are not classified in Section E unless the patent is concerned with the production or purification of the monomer or starting material.
- **Polymerisation catalysts** are not normally classified in Section E unless the novelty of the invention is the catalyst.
- **Mixtures** of compounds described as a cut (i.e. hydrocarbon feedstock) in a **petrochemical process** are, as a rule, only classified in Section H.
- **Highly complex, non-stoichiometric compounds**, e.g. those used as fluorescent materials, are classified in Section L only, but simpler compounds are normally classified in Section E. Patents on the growth of **single crystals of pure elements or compounds**, e.g. Si, GaAs, or BN, are classified in Sections E and L.
- **Free metallic elements** are not classified in Section E

A patent may be classified in Section E because of the type of compound(s) disclosed, and additionally in other Sections because of the uses or other attributes disclosed. Typically, perfumes, flavourings, and additives to foods and tobacco are classified in Sections D and E. Solvents and very common reagents, e.g. water, are not normally classified in Section E.

#### *E1 General organic*

**E11** Containing **P** and/or **Si**.

**E12 Organometallics** - i.e. compounds that include elements other than H, C, N, O, S, halogens, Si, and P.

**E13 Heterocyclics.**

**E14 Aromatics** – i.e. compounds with at least one benzene ring.

**E15 Alicyclics.**

**E16 Aliphatics that include N and/or halogen.**

**E17 Other aliphatics.**

- E18** General **hydrocarbon mixtures**.
- E19** **Other organic compounds** general – including organic compounds of unknown or indefinite structure and general mixtures of many types. Also organic reactions (e.g. nitration, resolution) when applied generally.
- E2 Dyestuffs*
- E21** **Azo** – including diazonium compounds.
- E22** **Anthracene** – including structures with more than 3 rings.
- E23** **Heterocyclics**.
- E24** **Other dyes, general dyes, all precursors**.
- E3 General inorganic*
- E31** Compounds of V, Nb, Ta, Cr, Mo, W, Mn, Tc, Re, Fe, Ru, Os, Co, Rh, Ir, Ni, Pd, Pt, Pa, and subsequent actinides.
- E32** Compounds of Ti, Zr, Hf, Cu, Ag, Au, Zn, Cd, Hg, Ga, In, Te, Ge, Sn, Pb, As, Sb, Bi.
- E33** Compounds of Be, Mg, Ca, Sr, Ba, Ra, Sc, Y, La, Ac, Al, lanthanides (rare earths), Th.
- E34** Compounds of Li, Na, K, Rb, Cs, Fr.
- E35** **Ammonia, cyanogen, and their compounds** – including HCN and cyanamide. Excludes hydrazine, which is classified in E36.
- E36** **Non-metallic elements, semi-metals (Se, Te, B, Si) and their compounds** (except those covered in Class E35).
- E37** **Other inorganic compounds** general – including mixtures of many components. Also **inorganic reactions** and **inorganic processes** of general applicability.

# 5 Types of BCE indexing

## 5.1 Introduction

Each patent classified in Sections B, C, and E (abbreviated BCE) is analysed using a variety of indexing languages and tools. Derwent's professional analysts attempt to provide as many access points to chemical patent information as possible, within a reasonable period of time after the patent has been received.

Each indexed compound is described according to its **structure**, its **role** in the patent that discloses it, and the **chemical properties and activities** it exhibits within the invention, using a combination of the following indexing tools, or indexing perspectives:

- Chemical (Fragmentation) Codes
- Markush DARC Graphical Indexing
- Derwent Chemistry Resource Graphical Indexing
- Manual Codes
- Ring Index Numbers
- Registry Numbers

Most of these indexing tools are described in detail in other manuals (a list of which appears on page iv of this manual), and therefore receive only a brief treatment in this chapter.

### Note

When this manual states that a particular type of compound is *fully indexed*, it means that the compound is indexed using both Chemical Codes and chemical structures (Markush DARC and from 1999, Derwent Chemistry Resource indexing), in addition to any other applicable index terms discussed in this chapter. The types of compounds that are fully indexed are listed in Section 6.2.

## 5.2 Chemical (Fragmentation) Codes

The first *deep indexing* method introduced by Derwent to comprehensively index the broad range of structures disclosed in chemical patents was the Chemical Fragmentation Coding System (alternatively referred to as the *Chemical Codes*, *Fragmentation Codes*, or simply as *the codes*). The first Chemical Codes, introduced in 1963 with the Farmdoc service, were called punch codes because they represented hole punch positions on a Hollerith computer card, which in turn represented chemical and biological features of patented pharmaceutical compounds. The punch codes were replaced in 1981 by the 4-character alphanumeric codes that are currently in use, and the backfile has been converted to the new format. Therefore patents indexed with punch codes can be searched using the Chemical Codes. However, codes introduced in 1981 representing new features, with no corresponding punch codes, could not be generated for the backfile.

Chemical Codes are applied to the novel features of an invention using both the Documentation Abstract and the patent specification as the source material. The Codes describe:

- chemical structure
- role of the indexed compound
- properties and activities exhibited by the indexed compound

Chemical Codes are discussed in detail in the Chemical Retrieval Manual.

## 5.3 Markush DARC graphical indexing

Beginning in Derwent Week 198701, both Markush structures and single compounds in Sections B, C, and E have been graphically indexed using the Markush DARC indexing language.

The corresponding Markush DARC search language is used to search for patents in the **Derwent World Patents Index Markush** (DWPIM) database, and the answers are presented in a graphical format familiar to chemists. Markush DARC indexing allows search results of higher relevance than Chemical Code indexing, because Markush DARC searches are conducted using complete chemical structures, that can be either specific or highly variable, unlike the relatively disjointed set of chemical fragments used in Chemical Code searching. Moreover, Markush DARC does not sacrifice comprehensiveness to achieve the increased relevance.

Markush DARC indexing and Chemical Code indexing are performed concurrently and, where possible, Chemical Codes are generated from Markush DARC graphics files by computer software called **AutoCPI**, discussed in Section 8. Chemical structures that have been indexed graphically in DWPIM are identified in the Derwent WPI bibliographic database by the Chemical Code M904. (The use of *M904* in search strategies is discussed in the *Markush DARC User Manual*.)

It should be noted that although Markush DARC indexing began in 1987, complete Markush DARC indexing coverage was not achieved until the dates shown below.

<b>Derwent Section</b>	<b>Complete coverage from Derwent Week:</b>
B and C, general	198920
B and C, polypeptides	199101
B and C, oligo- and polysaccharides	199125
E	199116

Prior to these dates, some complex patents may not have been indexed using Markush DARC structural indexing.

In 1998 the DWPIM database was combined with the MPharm database produced by INPI (the French Patent and Trademarks Office) to produce the Merged Markush Service (MMS). From this time Derwent has continued to index patents as they issue, whilst INPI have been indexing backwards in time, to increase the period covered by the database.

At the beginning of 2001, the MMS file contained over 850,000 Markush and specific structures corresponding to:

<b>Range</b>	<b>Coverage</b>
1982 – 2001	EP, FR, US, WO Pharma patents
1987 – 2001	Pharmaceutical patents from other countries covered by Derwent
1987 – 2001	Chemical patents from all countries covered by Derwent

The *Markush DARC User Manual* describes the search methods available for accessing the contents of Derwent World Patents Index Markush and its successor, MMS.

### 5.3.1 Text Notes

Text Notes are used by analysts to supplement the information conveyed by Markush DARC superatoms and their attributes, restricting the definition of superatoms further than superatom attributes. Text Notes are displayed below structure drawings of answers to Markush DARC search queries; they can be seen whenever a G-group is displayed that includes a superatom that has been assigned Text Notes.

There are two types of Text Notes: **Controlled Text** and **Free Text**. Controlled Text Notes have a fixed set of parameters and are limited to a standard format. Free Text Notes, as their name implies, are not limited to a standard format; they are added to indexing records when attributes and Controlled Text Notes are insufficient for conveying relevant information about the coverage of a superatom. They are also added to indexing records to convey general information about the indexed structure. Neither Controlled Text Notes nor Free Text Notes are directly searchable by the user. Controlled Text Notes are, however, consulted by the Markush DARC search software during specific kinds of searches, e.g. when Broad Translation has been applied to a specific structural entity in a search strategy. Instead of allowing a hit based only on superatoms and attributes, the CTN is consulted by the search software to see if it contains information that would indicate that the record is not actually a match for the search query.

For more details on Controlled and Free Text, see Appendix 1.

## 5.4 Derwent Chemistry Resource indexing

The Derwent Chemistry Resource is a chemical structure database for searching specific compounds indexed in Derwent WPI bibliographic records. The database is searchable both by chemical structure and by various text fields, allowing simple access to the Derwent WPI database by specialist and non-specialist chemical searchers alike. The Derwent Chemistry Resource runs in parallel to, and to a certain extent replicates, Chemical Indexing (Fragmentation Codes) for patents classified in Chemical Patents Index (CPI) Sections B (Pharmaceuticals), C (Agrochemicals) and/or E (General Chemicals). Derwent Chemistry Resource indexing commenced in Derwent Update 199916, for ongoing basic patent references classified in CPI Sections B (Pharmaceuticals), C (Agrochemicals) and/or E (General Chemicals).

The Derwent Chemistry Resource Numbers, which are unique identifiers for specific chemical compounds, form the link between the Derwent Chemistry Resource chemical structure database and corresponding bibliographic indexing in Derwent WPI. The online functionality that provides this connection reflects the capabilities of the Host systems involved.

Currently the Derwent Chemistry Resource is only available on STN, where it is integrated within the Derwent WPI files. The Derwent Chemistry Resource is not available on Questel.Orbit as a separate database. However, all specific compounds which are included within Derwent Chemistry Resource on STN, are also indexed as a part of the Merged Markush Service (MMS), and linked to bibliographic records in Derwent WPI on Questel.Orbit via their corresponding Markush DARC Specific Compound.

## 5.5 Manual Codes

Manual Codes are alphanumeric codes that represent terms in a hierarchical controlled-term vocabulary. Manual Codes have developed over a period of more than 30 years, having been introduced in 1963 with Derwent's Farmdoc service (CPI Section B). Manual Codes for Agdoc (CPI Section C) were introduced in 1965, and for Chemdoc (Section E) in 1970.

Manual Codes are applied to novel features of inventions using Derwent's **Documentation Abstracts** as the information source. Manual Codes enable **inexpensive, rapid** indexing and retrieval of patents, but they lack the *deep indexing* coverage of both Chemical Code and Markush DARC indexing. Whereas Chemical Code and Markush DARC indexing are designed to determine if certain chemical structures are present among the thousands or millions of compounds claimed or disclosed in a chemical patent, Manual Codes are intended for general classification of patented compounds, reactions, and uses. They are similar to International Patent Classification (IPC) codes, but are applied more consistently than the corresponding IPC codes.

Manual Codes are discussed in detail in the Derwent publication *CPI Manual Codes*.

## 5.6 Ring Index Numbers

The Chemical Codes described in Section 5.2 includes numerous codes for ring systems. However, some rings are not uniquely described by a single code. To enable more specific searches on ring systems, Derwent began assigning ring numbers from *The Ring Index* (by Patterson, Capell, and Walker, 2nd Edition, American Chemical Society) to patent indexing records in 1972. Ring Index Numbers are five digit numbers listed in the RR field (RIN field on STN) of the Derwent WPI database.

Ring Index Numbers are assigned in the following cases:

- ring systems whose chemical codes additionally cover other ring systems
- rings containing elements other than C, N, O, or S
- spiro rings

Ring Index Numbers only indicate the skeleton of ring atoms present, and therefore do not take into account substituents, tautomeric forms, or degree of unsaturation. Thus, Ring Index Numbers are NOT assigned:

- to ring systems uniquely identified by a single code
- to distinguish between tautomers or different states of hydrogenation
- to metals in rings, unless the metal is bonded to two C atoms

Furthermore, Ring Index Numbers are applied to rings that are **specifically named or drawn** in a patent. Additional rings that could be **deduced or inferred** unambiguously from the patent's disclosure or Markush structures are normally assigned Ring Index Numbers, although it is possible that some of these latter rings have been omitted due to system limits. When indexers encountered patented ring systems that were not included in *The Ring Index*, Derwent created its own system of Ring Index Numbers. The rings in *The Ring Index* are numbered from 00001 to 25000; Derwent's ring numbers start at 40000, with in excess of 35,000 new rings created. In 1999, Derwent ceased creating new Ring Index Numbers, although Ring Index Numbers created prior to 1999 are still applied. **Derwent-generated Ring Index Numbers** were formerly available on cards, microfilm, CD-ROM, and via the microcomputer program Markush TOPFRAG. Currently, they are only available as data files usable with the Markush TOPFRAG or STN Express programs.

Ring Index Numbers are discussed in more detail in the *Chemical Indexing User Guide*.

## 5.7 Registry Numbers

### 5.7.1 Application of Registry Numbers

In 1981, a Registry of about 2100 compounds and elements commonly mentioned in claims and examples was compiled, and each compound was assigned a Registry Number. The purpose of creating the Registry was to improve retrieval of specific compounds that frequently occur in patents.

All Registry compounds in a patent that meet the criteria listed in Section 6.2 for *fully indexed* compounds are indexed with Registry Numbers. In addition, **significant compounds** or **significant non-metallic elements** mentioned in the claims or examples that are Registry compounds, but that do not meet the criteria listed in Section 6.2, are also indexed with their Registry Numbers. (An example of an **insignificant** compound is a solvent mentioned in a process in which any solvent may be used.) This is a departure from most other forms of BCE indexing, which are only used to describe the novel aspects of inventions.

Indexing a compound with its Registry Number automatically generates relevant structural Chemical Codes. The Registry Number is also assigned one of the following Roles:

- P** Compounds **produced**. This Role is assigned to products of reactions and products isolated from purification and extraction reactions.
- S** **Starting materials, catalysts, reagents** (including reaction auxiliaries classified in BCE). Starting materials are defined as compounds which donate one or more carbon atoms to an organic product, or any atom(s) to an inorganic product.
- U** **Use** of a compound. This Role is assigned to ingredients in compositions, and compounds detected, analysed, tested, or removed. It is also assigned to reaction auxiliaries classified in CPI Sections other than A, B, C, and E.

Registry compounds and the guidelines for their use are discussed in the *Chemical Retrieval Manual*, which contains three different lists of the Registry compounds. Each list contains the Registry Number, full name, molecular formula, and standardized name. For example, the following entry is found in the first list, which is in alphabetical order by compound name:



The second list is in chemical structure order, with structural formulae and the list of Chemical Codes used for searching the structure. The second list is a good source of chemical indexing examples. The third list is in activity order.

### 5.7.2 Special Cases

The following notes describe the assignment of Registry Numbers in special cases.

- Compounds that could be derived from Markush structures, but which are not exemplified, are not indexed with Registry Numbers. **Only compounds that are specifically mentioned are indexed with Registry Numbers.**

**Example:** Suppose a patent mentions “alcohols of carbon numbers 1-6”, exemplifying methanol and ethanol. Only the latter two alcohols are indexed with Registry Numbers.

**Example:** Suppose a patent mentions “the oxides and sulphides of copper, cobalt, and zinc.” Both the cation and anion vary; thus, the different possible compounds that could be inferred from this phrase are not indexed with Registry Numbers, unless specifically mentioned elsewhere in the patent.

- If only **one ion in a compound varies**, the listed compounds are considered to be specifically mentioned, and are indexed with Registry Numbers.  
**Example:** A patent mentions “the oxides of copper, iron and tin” and “the oxides, sulphide, and sulphate of copper”. In this case, all possible permutations are indexed with Registry Numbers.
- The **isomers** of a compound referred to collectively by a generic term are considered to be specifically mentioned. Therefore, all isomers indicated by the generic term are indexed with Registry Numbers.  
**Example:** Suppose a patent mentions the generic chemical term cresol in a manner that warrants indexing with Registry Numbers. All isomers of cresol would be indexed with Registry Numbers, i.e. 2-, 3-, and 4-cresol.
- Registry Numbers are assigned for **compounds that are one step removed** from the inventive feature.  
**Example:** Suppose a patent states that “ammonium chloride is recovered from waste liquor obtained from ammonia production.” Ammonium chloride would be fully indexed with Markush DARC graphics and Chemical Codes. Both ammonium chloride and ammonia would be indexed with their Registry Numbers and the Role P.  
**Example:** Suppose a patent states that “claimed compound (I) may be used as an intermediate in the preparation of oxirane.” Compound (I) would be fully indexed and oxirane would be indexed with its Registry Number and the Role P.
- **Metal phosphate salts** are, in the absence of information to the contrary, assumed to be **tribasic**. For example, calcium phosphate is indexed with the Registry Number for calcium triphosphate.
- **Metal salts** of organic acids, alcohols, etc., are indexed with the Registry Number of the parent acid, alcohol, etc., unless a Registry Number exists for the particular metal salt mentioned. For example, sodium acetate is listed as a Registry compound, so it is indexed with its own Registry Number. However, there is no Registry Number for iron acetate, so the Registry Number for acetic acid is assigned when iron acetate is mentioned.
- **Metal phthalocyanines** are indexed with the Registry Number for phthalocyanine. Copper phthalocyanine is an exception to this rule, as it is the only metal phthalocyanine that is a Registry compound (RN=1160).

- **Polymers** are not indexed with Registry Numbers if they are classified in Section E only.
- **By-products** (i.e. small quantities of unintended products) are indexed with Registry Numbers only if the patent states that they are **recoverable and useful**.
- When used as **ingredients in a composition** or **auxiliaries in reaction processes**, the following list of compounds are not indexed with Registry Numbers unless they are the inventive feature of the patent:
  - solvents
  - electrolytes
  - cultivating media
  - drying agents
  - buffers
  - atmospheres
  - water
  - ceramics/refractories
  - glasses
  - silver halides used in photographic compositions
- **Reaction auxiliaries** such as acids, bases, oxidising agents, reducing agents, etc. are only indexed with Registry Numbers if the patent claims a production process of a new or known compound.
- **Optional components of compositions** are not indexed with Registry Numbers.
- **Complex/mixed oxides** defined as ratios or percentages of the constituent individual oxides are not indexed with their Registry Numbers.

**Example:** Suppose a patent mentions a composition that includes the following mixed oxides:

$\text{Fe}_2\text{O}_3$  (30% wt),  $\text{Al}_2\text{O}_3$  (30% wt),  $\text{SiO}_2$  (40% wt).

In this case, the compounds  $\text{Fe}_2\text{O}_3$ ,  $\text{Al}_2\text{O}_3$ , and  $\text{SiO}_2$  are not indexed with Registry Numbers.
- **Monomers**, used in polymerisable compositions or as starting materials for polymers, are not generally indexed with Registry numbers, unless they are the novelty of the invention.
- Analysts do not usually refer to the original patent specification when indexing **Russian or Japanese** patents with Registry Numbers. Instead, the selection of compounds is made from the title and abstract only. However, if structures are drawn out and/or tables are given in the patent specification, these structures are also indexed with Registry Numbers.

### 5.7.3 Limitations on Registry Number Indexing

Selection of Registry compounds from the title, abstract, claims, and examples using the above rules may on occasion give rise to long lists of selected compounds. For example, a patent may include a long list of compounds in the main claim referring to one of the components of a composition. In such cases, analysts may choose to limit the selection of compounds for Registry Number indexing to the **most preferred or most exemplified** compounds, as taken from real examples and/or specific claims. In other cases, the analyst may choose to select only representative examples, e.g. when no preferred examples are given.

### 5.7.4 Changes in application since the introduction of Derwent Chemistry Resource

There has been no change in the application rules since the introduction of Derwent Chemistry Resource. All Derwent Registry Numbers were given new Derwent Chemistry Resource numbers and are being included in the online Derwent Chemistry Resource structure file as soon as they are cited in a new patent. Whenever a Derwent Chemistry Resource compound is applied to a patent reference, the corresponding Derwent Registry Number (and any associated Fragmentation coding) is automatically applied to the reference.

# 6 Types of compounds indexed

## 6.1 Introduction

Most indexing terms and codes applied to patents classified in Sections B, C, and E pertain to the inventive features of patent documents. This is true most notably of Chemical Code, Derwent Chemistry Resource and Markush DARC indexing. The types of compounds selected for Chemical indexing are listed in the following section. Once compounds have been selected for indexing, certain Roles and Chemical Codes are automatically applied to them, based on the reason(s) for their selection. These Roles and Codes are listed in Section 6.3.

## 6.2 Types of compounds selected to be fully indexed

The following types of compounds and compound attributes are *fully indexed* using the BCE Chemical Codes and Markush DARC indexing:

- all compounds and reaction intermediates stated to be novel
- products of new processes
- new uses of known materials
- materials detected and detecting agents
- detection media
- materials recovered or purified in new ways
- materials removed and removing agents (only since 1977, unless they were the only chemicals in the invention that could be indexed)
- components of compositions that are essential to the invention
- novel catalysts since 1970
- activities, properties, and uses
- chemical formulations and apparatus

### 6.3 Corresponding roles and essential chemical codes

Once a compound has been selected for indexing, it is automatically assigned a Role qualifier and the Chemical Codes that are deemed essential, according to its role in the patent. (See the Markush DARC User Manual and Derwent Chemistry Resource User Guide for information about Roles.) The table below lists the types of compounds selected for indexing with their corresponding Roles and essential Chemical Codes. The superscripts refer to the corresponding notes.

Material selected	Essential Chemical Codes	Role qualifier
New compound	M710 <sup>1</sup>	N
New intermediate	M710	N
Compound produced	M720, N process codes	P
Compound stored	M720, N104	P
Compound purified or extracted	M720, applicable N16: code	P
Material detected or analysed	M750, N102	A
Detecting or analytical reagent <sup>2</sup>	M781, N102, P831 or P832, Q505	D
Medium in which a compound is detected <sup>3</sup>	M760, N102	U
Compound removed	M750, applicable N16: code	X
Removal/purifying agent <sup>2, 4</sup>	M781, Q508, applicable N16: code	R
Indicator <sup>2, 5</sup>	M781, N102, Q505, P831 or P832	D
New compound used as a catalyst <sup>6</sup>	M710, Q421	N, C
Known compound whose use as a catalyst is new <sup>6</sup>	M730, Q421	C
New use of a compound	M781	U
Component(s) of a composition <sup>7</sup>	M781 or M782 <sup>8</sup>	U and/or M
Starting material/ Product defined from starting material	M730	S/Q
Reagent	M730	V
Known compound	Any M7: code except M710	K
Therapeutically active agent	<sup>3</sup> P: codes	T
Excipient	in addition to other Roles/M7 codes	E

#### Notes

- <sup>1</sup> The code *M710* (new compound) has not been assigned broadly and regularly enough to be relied upon in **exhaustive** Chemical Code searches for compounds claimed to be novel.
- <sup>2</sup> If **more than one agent** is indexed, the Role qualifier *M* (component of a mixture) is also assigned, and the code *M781* (use of one compound) is replaced with *M782* (use of 2 or more compounds) In such cases, see also Note 8.

- <sup>3</sup> Media in this type of patent are usually body fluids that cannot be indexed with Markush DARC graphics records. However, if the **medium can be indexed** with Markush DARC graphics, the Role U is assigned.
- <sup>4</sup> If a **removal process** is stated to be an important feature of the invention, the code *Q431* (separating solids etc.) is additionally assigned. Patents on removal processes are generally classified in Section J01.
- <sup>5</sup> Refers to **detection by colour change. Indicators** are indexed under the subheading M2 if classified in Section B, and under the subheading M3 if classified in Section E. (*Subheadings* are explained in the *Chemical Retrieval Manual*.)
- <sup>6</sup> From 1970 forward, **novel catalysts** are **fully indexed**, both with Chemical Codes and (from 1987) Markush DARC graphics. From 1977 forward until the introduction of the Derwent Chemistry Resource in 1999, **all other catalysts** were **partially indexed**, but with Chemical Codes only. The Chemical Codes assigned in the partial indexing are:

- *M730* (known compound used in synthesis), *Q421* (catalyst)
- applicable codes from code Parts A, B, and C (except Subset C80:) for elements and/or compounds which are significant components of the catalyst
- the applicable code from Subset M41:

However, in some cases Manual Codes for catalysts (Section N) provide more information than partial indexing with Chemical Codes. For example, when the following compounds used as catalysts, they are not partially indexed:

- **zeolites**, when no metals other than alkali or alkaline earth are present
- **aluminosilicates**, when no metals other than alkali or alkaline earth are present
- **alumina**

From 1999 onwards, when Derwent Chemistry Resource Indexing was introduced, catalysts are fully indexed with Derwent Chemistry Resource and/or Markush structures, and full chemical indexing if they are important to the invention. Therefore, partial coding of catalysts ceased with the advent of the Derwent Chemistry Resource.

- <sup>7</sup> If **more than one component** of a mixture is indexed, the Role qualifier *M* and the Chemical Code *M782* (see also Note 8 below) are assigned; otherwise the Role qualifier *U* and the Chemical Code *M781* are assigned.

*Known compounds that serve a standard function in a composition are not usually indexed* with Chemical Codes and Markush DARC graphics, although they may be indexed with Registry Numbers. (Registry Numbers are discussed in Section 5.6.) However, if the novelty of a composition is the relative proportion of its constituents, all the constituents are fully indexed, even if they are known compounds that serve standard functions.

- <sup>8</sup> The assignment of code *M782* to compounds classified in Section B and/or Section C prompts the additional assignment of either *M430* or *M431* (selected according to the number of active ingredients).

# 7 Levels of indexing detail

## 7.1 Introduction

It is Derwent's aim to index the wide chemical disclosure taught by patent documents, using information from the disclosure, tables, lists, examples, and claims. However, in practice analysts are constrained by both system limits and time limits. The system limits of the Markush DARC graphics indexing software allow an analyst to describe a single Markush structure with up to 1023 atoms, 50 variable *G-groups*, and 5 levels of *G-group nesting* (*G-groups* and *G-group nesting* are discussed in the *Markush DARC User Manual*). In some cases, in an effort to gain the broadest possible protection of inventive features, a patent applicant may incorporate a limitless number of derivatives into a generic structure. In such cases, indexing every possible structure would exceed the limits of the indexing software. Furthermore, because of the volume of patents indexed by Derwent, practical limits must be set on the amount of time that can be devoted to the indexing of a single document.

In order to ensure that the important aspects of patents can be indexed in a commercially acceptable time without breaching system limits, a series of three **indexing levels** were developed. These have not been formally used since the introduction of the Derwent Chemistry Resource in 1999, but they are still useful for describing the complexity of the patent and the type of indexing guidelines that have been applied over time by Derwent as well as those currently in force

- Level 1** indicates the most in-depth indexing possible, i.e. the most complete indexing of the range of compounds claimed, exemplified, or disclosed;
- Level 2** describes an intermediate level of indexing;
- Level 3** is the least comprehensive indexing, although it covers as much as possible of the new chemistry taught by the patent. The level of indexing is chosen at the discretion of the analyst. Most patents are indexed at Level 1 or Level 2, with Level 3 guidelines being used when time and/or system limits present a problem.

Levels 1 and 2 are discussed in the following section on "Patents of Average Complexity", and Level 3 is discussed in the subsequent section on "Very Involved Patents".

## 7.2 Patents of average complexity

### 7.2.1 Level 1 Indexing

Level 1 indexing is thorough and exhaustive – all relevant compounds are indexed from the following segments of the patent document using a combination of Derwent Chemistry Resource and/or Markush structures and chemical fragmentation codes:

- disclosure
- tables or lists
- examples
- claims

First the specific compounds are selected for the Derwent Chemistry Resource from the claims and examples. Any Markush structures or further examples are then indexed using Markush indexing.

The following guidelines are applicable to Level 1 indexing since the Derwent Chemistry Resource commenced.

#### *A Derwent Chemistry Resource Indexing*

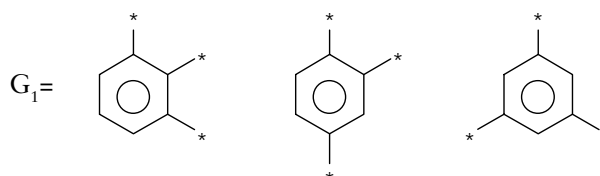
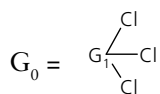
- All specific compounds pertinent to the invention are selected from the Claims
- The main (best) example is then selected
- Any further examples that expand on the structural diversity of previously selected compounds and have accompanying data to indicate they are “real” examples, are then selected at the discretion of the Analyst
- Prior to 2001, a limit of 25 new compounds or 50 known compounds was applied to Derwent Chemistry Resource selection.

#### *B Markush Indexing*

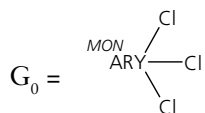
- A Markush structure is drawn to represent the claimed/disclosed Markush
- Specific values of generic terms given in Examples selected for the Derwent Chemistry Resource are not added to the Markush
- Values of generic terms from Examples not selected for the Derwent Chemistry Resource are added to the Markush.

- Specific values are indexed as real atoms, not as superatoms.

**Example:** Trichlorobenzene is indexed as follows:



This is preferable to using the superatom  $\text{ARY}^{\text{MON}}$ , as shown below:



- All specific values of chains such as  $(\text{CH}_2)_n$  are indexed, within system limits.
- Specific/generic combinations are always included.

**Example:** Suppose a patent claims the fragment  $\text{NR}_1\text{R}_2$ , where  $\text{R}_1$  and  $\text{R}_2 = \text{H}$  or alkyl, and the specific example  $\text{N}(\text{CH}_3)_2$  is given. Level 1 indexing of this structure, which allows retrieval of both the specific and generic values, would include:



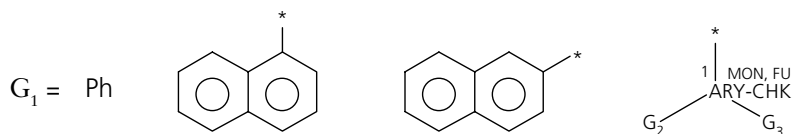
where  $G_1$  and  $G_2 = \text{H}, \text{CHK}, \text{C}$

## 7.2.2 Level 2 Indexing

Level 2 indexing covers the same information as Level 1 and the Derwent Chemistry Resource guidelines are exactly the same. However, Level 2 allows analysts to use indexing shortcuts when atom limits or time limits become a problem. Permitted shortcuts for the Markush indexing are described in the following notes.

- Superatoms may be used to cover **large numbers of permutations of specific values**, and are supplemented by Text Notes when additional information is necessary. (Text Notes are discussed in Section 2.3.1.)

**Example:** Phenyl or naphthyl optionally substituted by 1, 2 or 3 alkyl chains may be indexed as:



Text note: ARY1=C6, NR1; ARY1=C10, NR2.

**Example:** Phrases such as “by alkyl we mean methyl, ethyl, n-propyl...” may be indexed as *CHK* in Level 2 indexing. If a range is given in the claims, the indexing is made more specific using Text Notes.

- **Specific/generic combinations** may be omitted.
- Superatoms may be used to cover **large numbers of specific heterocyclic or carbocyclic rings**. Text Notes are added to allow more precise retrieval.
- The superatoms *HEF*, *CYC<sup>FU</sup>* and *ARY<sup>FU</sup>* may be used to index **spiro rings**, with appropriate supplemental Text Notes. (Several Text Notes examples involving spiro ring systems are presented in Appendix 1.)

## 7.3 Level 3 : Very involved patents

Level 3 indexing is reserved for very complex patents, the kind of patents for which exhaustive indexing is impossible due to time and/or system limits. As a minimum, Level 3 indexing includes all real examples and their corresponding generic terms in the wider disclosure or claims. However, as will be seen, some patents even require a reduction of the number of real examples indexed.

In general, patents receiving Level 3 indexing fall into one or more of the following three categories:

- Patents with **very large numbers of specific examples**
- Patents with **very complex Markush structures**
- Patents claiming **long lists of possible components of compositions**

### 7.3.1 Patents with very large numbers of examples

Some patents list so many examples that indexing only the claimed and real examples in the Derwent Chemistry Resource and Markush DARC would breach time and/or system limits. In such cases, the following structures are indexed (with fragmentation codes produced from Derwent Chemistry Resource and Markush structures):

*Pre-Derwent Chemistry Resource:*

- a representative selection of examples
- the generic structure(s) in the claims

*Derwent Chemistry Resource – 2001:*

- up to 25 new or 50 known specific structures from claims and examples (included in the Derwent Chemistry Resource)
- the generic structure(s) in the claims (as a Markush)

*2001 ongoing:*

- claimed examples (included in the Derwent Chemistry Resource)
- at least one real example (included in the Derwent Chemistry Resource)
- the generic structure(s) in the claims (as a Markush)
- a representative selection of remaining examples (included in Markush)

The phrase *representative selection of examples*, means that analysts do not choose a pre-set number of examples at random, but rather use the expertise they have gained from

reading patents to select those examples that are most indicative of the invention, and that allow the best chances of retrieval of the patent reference.

The following simplified example illustrates a typical generic structure taken from a photographic patent.

**Example:** Suppose a patent discloses the use of a dye of formula (I).



where DYE = a dye type selected from: azo, anthraquinone, phthalocyanine quinacridone, indigo and triphenodioxane

X = halotriazine or halopyrimidine reactive group

Suppose further that a long list of examples is given, too great to index at levels 1 or 2, the majority of which are azo type dyes. If the indexing of all real examples breaches time and/or system limits, then representative examples are chosen by applying the following guidelines, which are given in order of priority:

- examples of each of the different dye types (DYE) are selected
- examples are selected with different reactive groups (X)
- different types of azo dye types are selected, ie monoazo, disazo, polyazo, metal-azo complexes, formazan, metal-formazan complexes etc.
- azo dyes are selected with different types of diazo components and coupling components

Although this example describes a dye patent, the principles of selecting representative examples shown here, may be applied to patents of any type.

### 7.3.2 Patents with very complex Markush structures

Some Markush structures in patents are so broad that indexing them completely would breach system limits. This is not because the system limits are too stringent, but rather because the Markush structure has been defined much more broadly than the norm. When very complex Markush structures are encountered in patents, the following structures are usually selected for indexing:

- all example compounds (following appropriate Derwent Chemistry Resource guidelines)
- the generic Markush structure

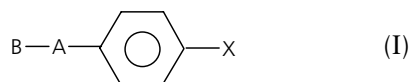
If there are too many examples to index within time constraints, representative examples are indexed (as discussed in the previous section).

The central core of a Markush structure is considered most important. As the essential part of the structure, the central core forms the basis of the Markush group  $G_0$ . The further removed a substituent is from the central core, the less important it is assumed to be. Therefore, complicated generic Markush structures are normally indexed as follows:

- The **central core** is indexed as fully as possible.
- **Substituents bonded to the central core** are indexed as fully as possible, using superatoms if warranted by time and/or system constraints.
- **Remote substituents on substituents** are replaced by the superatom XX, unless very specific substituents are mentioned, in which case they are indexed as fully as time and system limits allow.
- **Optionally substituted rings or chains** which form an important part of the structure are indexed both as the unsubstituted ring or chain (as fully as possible within time and system limits) and as the appropriate superatom (with its substituents either attached or replaced by the superatom XX).

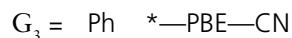
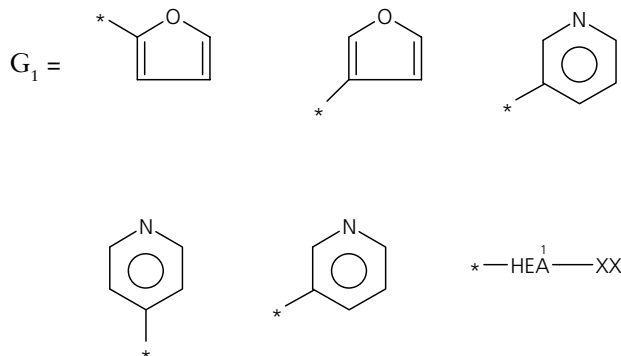
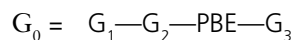
The following example, although *much simpler than the Markush structures indexed at this level*, illustrates the application of the above rules.

**Example:** Suppose a patent claims structure (I) below.



- where A = phenylene ring
- B = furan or pyridine ring optionally substituted by: alkyl chain, aryl ring, fluorine, chlorine, bromine, nitro, alkoxy, azido, guanidino, cyano and thiocyno
- X = phenyl ring or 4-cyanophenyl

The moiety X is very specific. However, the substituents on the heteroaromatic rings (B) are optional and are not at all specific. Level 3 indexing would result in the following graphical indexing record:



Text note: G1\*HEA1=O1, C4, RA5, X0; HEA1=N1, C5, RA6, X0.

*Claimed generic structures that include real atoms* are, as a rule, always indexed; in most cases, they are indexed separately from the selected examples. Generic structures that describe chemical properties and functions rather than chemical structures are not necessarily indexed.

**Example:** Suppose a patent claims the following generic structure:



where TIME = a timing group which cleaves after a specified amount of time

PUG = a photographically useful group

RED = a reducing group

No real atoms are shown in this Markush structure, so no informative indexing can be produced from it. In such cases, only the examples corresponding to the Markush structure are indexed.

### 7.3.3 Patents claiming long lists of possible components of compositions

Patents claiming numerous components of a composition often involve compositions that consist of several types of agents; each agent fulfils a particular function, but, in the most general definition of the invention, the agent is not limited to a specific set of chemical structures. A typical example is a detergent composition that claims long lists of surfactants, sequestering agents, detergent builders, etc.

If a patent claims a list of compounds for which it is not feasible to provide full structural indexing, one of the three indexing guidelines listed below is followed. Preference is given to the guideline that allows the maximum amount of indexing coverage within available time and system limits.

- A All compounds are indexed, and all specific chain/ring systems mentioned in the claims, but not exemplified, are replaced by superatoms with relevant Text Notes.
- B Only compounds exemplified or preferred are indexed, and all specific chain/ring systems mentioned in the claims, but not exemplified, are replaced by superatoms with relevant Text Notes.
- C Only compounds exemplified in addition to being claimed are indexed.

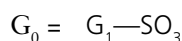
**Note:**

Since the introduction of the Derwent Chemistry Resource, specific compounds would first be selected for the Derwent Chemistry Resource, using the guidelines for the Derwent Chemistry Resource described previously, and then Markush Indexing would be done as above using Guidelines A or B. (For Guideline C, there would be no Markush structure.)

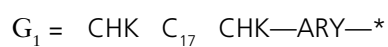
The simplified example below illustrates how these guidelines are applied.

**Example:** Suppose a patent claims an anionic surfactant as part of a composition, and then proceeds to give the following list of anionic surfactants as suitable candidates: dodecyl benzene sulphonic acid (exemplified), nonyl benzene sulphonic acid, naphthalene sulphonic acid, lauryl sulphonic acid, oleyl sulphonic acid, lauryl sulphate, stearic acid (exemplified), myristic acid, and p-stearyl benzoic acid. Dodecyl benzene sulphonic acid, lauryl sulphonic acid, stearic acid and p-stearyl benzoic acid are stated to be preferred.

Indexing all compounds, using superatoms to represent specific chain/rings that are not exemplified (i.e. using guideline A above), would result in two Markush DARC graphics records:



and



In these indexing records the example compounds received full structural indexing, while the compounds mentioned in the claims (but not exemplified) have been indexed with superatoms. Indexing the same patent at the minimum coverage level (i.e. using guideline C above) would include only the real examples, i.e. stearic acid and dodecyl benzene sulphonic acid.

**Note:**

The examples used in this section are presented to illustrate the indexing principles discussed. The patents that actually receive Level 3 indexing are much more complex than these examples.

# 8 AutoCPI

## 8.1 Introduction

Beginning in 1987, chemical structures covered within Sections B, C and E have been indexed using **both Chemical Codes and Markush DARC graphical indexing**. The redundancy involved in double-indexing prompted the development of software to automatically generate structural Chemical Codes from Markush DARC graphics files. The resulting software, **AutoCPI** was first implemented in Derwent Week 199222. AutoCPI works in a fashion similar to Markush TOPFRAG, a microcomputer program that converts standard (topological) chemical structures drawn by the user into a corresponding Chemical Code search strategy. With the introduction of the Derwent Chemistry Resource indexing in 1999, the software was modified to enable it to read the Derwent Chemistry Resource structure files and generate the appropriate Chemical fragmentation codes.

Most organic and inorganic compounds can be converted from structures to Chemical Codes by AutoCPI. Because the codes are automatically generated from Markush DARC or Derwent Chemistry Resource structures, there are no inconsistencies between Chemical Code indexing and the structural indexing. Furthermore, Chemical Code indexing produced by AutoCPI is consistent with the Chemical Code indexing that has been produced manually over the years, although very minor policy changes regarding Chemical Code indexing of superatoms were introduced to compensate for the way that AutoCPI converts such groups. (These policy changes are discussed in Section 8.5.)

Initially, AutoCPI could not generate codes for inorganic compounds, ions or ligands and some organometallic complexes. The software has subsequently been modified to enable it to automatically generate Chemical Codes for most structures indexed by Derwent. Some codes still need to be added manually, however.

Thus, the following Chemical Codes are not always present in AutoCPI output, and must be added manually by analysts on at least some occasions:

- **L710** for some structures
- **Set L8: codes** (sugars and derivatives)
- **M417** through **M431**
- **M630** through **M911** (except **M904**)
- Codes in **Parts N:** through **W:** (i.e. non-structural codes)

## 8.2 Compounds that require the manual addition of codes

### ■ Salts of organic compounds

Salts of organic acids or bases can now be converted to Chemical Codes by AutoCPI. However, Analysts must then add the appropriate salt or complex codes M630, M640, M650, M770 to the AutoCPI output and check that the correct “metal to...” codes have been applied in Sets A: and C: (A9: and C7: codes).

### ■ Organometallic compounds

Some organometallic compounds cannot be fully converted by AutoCPI, nor can compounds with a metal atom drawn with bonds attached. As above, the A9:, C7:, M6:, and/or M7: codes may need adding.

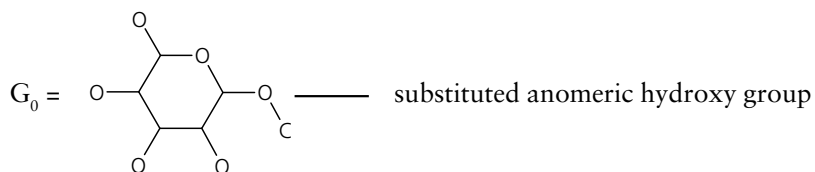
### ■ Complexes and salts

i.e. molecular compounds containing two or more organic fragments may need the M6: and/or M7: codes adding.

### ■ Sugars

By convention, sugars are only indexed in the ring form for Chemical Code indexing when the anomeric hydroxy group is substituted; otherwise, the chain form is indexed. This is not consistent with Markush DARC indexing, which requires that sugars be indexed in the ring form whenever possible, regardless of substitution. Thus, AutoCPI can only convert sugar structures from Markush DARC graphics to Chemical Codes if the anomeric hydroxy group is substituted. Analysts must then add the relevant codes from Set L8:, including K0 and L8 if they are applicable to the structure.

**Example:** Methyl glucoside is indexed in the ring form for both Chemical Code indexing and Markush DARC graphics.



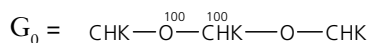
The above Markush DARC graphics record can be used by AutoCPI to generate corresponding Chemical Codes. The following codes must be added by analysts to the AutoCPI output:

K0, L8, L814, L821, L831

### ■ Repeating units

The numbering attributes applied to structures with repeating units are not recognised by AutoCPI, so each repeated unit is converted as though it were only one unit. Therefore, those codes that describe the repeated chemical group must be added to the AutoCPI output.

**Example:** Suppose the following structure had been indexed:



Text note: M100=1-5.

So that the polyether repeat unit is completely indexed with Chemical Codes, the following codes (describing polyethers) must be added to the AutoCPI output:

H583, H584, M322, M323, M392, M393

### ■ Natural products

Non-polymeric natural product structures can be converted by AutoCPI to Chemical Codes. Prior to the introduction of the Derwent Chemistry Resource, indexers added relevant Part V: codes to AutoCPI output for natural products classified in Sections B and C. With the introduction of the Derwent Chemistry Resource in 1999, the V: codes were discontinued. These records can be identified by using the M905 time code in Subs M2 instead of M903, which was used prior to the introduction of the new systems within Derwent.

### ■ Dyes

Analysts must add all relevant Part W: codes to AutoCPI output for dyes.

## 8.3 Compounds that cannot be indexed by AutoCPI

The following types of structures cannot be converted by AutoCPI to Chemical Codes:

### ■ Steroids

### ■ Fullerenes

### ■ Compounds indexed under the subheading M1

Typical examples are polysaccharides, polypeptides, polymers (including addition and condensation polymers).

**■ Sugars with an unsubstituted anomeric hydroxy group**

AutoCPI will not convert this type of sugar from Markush DARC graphics to Chemical Codes, because they are indexed in the ring form for Markush DARC graphics and in the corresponding chain form for Chemical Code indexing.

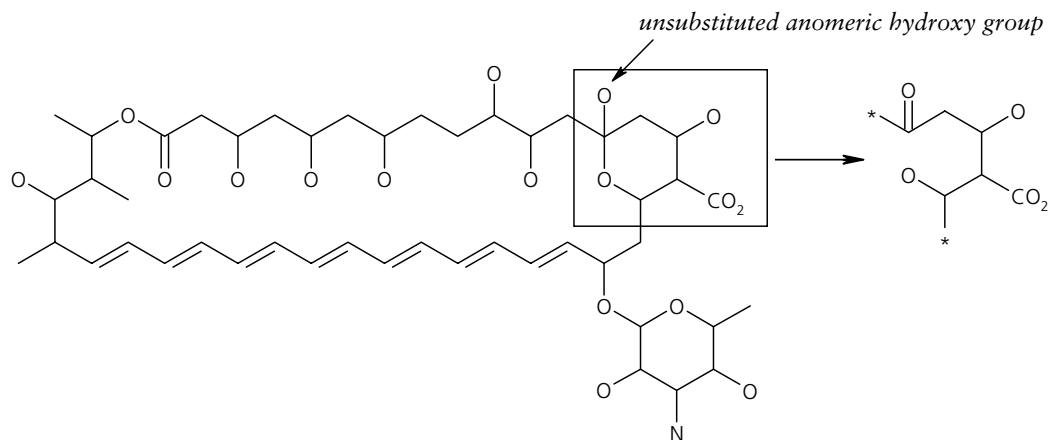
**■ Compounds containing d+ (delta) charges**

The only delta-charged compounds that can be successfully converted by AutoCPI are those in which the delta charges are in an azulinium ring.

**■ Amphotericin B, Nystatin, Rapamycin etc.**

Because these structures contain unsubstituted anomeric hydroxy groups, AutoCPI will not convert their Markush DARC indexing into Chemical Codes. Those rings containing unsubstituted anomeric hydroxy groups are indexed in the corresponding chain form for Chemical Code indexing and indexed in the ring form for Markush DARC graphics.

**Example:** The diagram below illustrates the difference between Markush DARC indexing and Chemical Code indexing of Amphotericin B.



The highlighted ring is opened for indexing with Chemical Codes and remains closed for Markush DARC indexing.

## 8.4 AutoCPI indexing of superatoms

Superatoms and their attributes are, in general, successfully converted by AutoCPI into Chemical Codes. A summary of how AutoCPI translates organic superatoms and their attributes into Chemical Codes is given in the following table.

**Note:**

- The ring superatoms (ARY, HEA, HET, CYC) are assumed to be monoattached, so the Chemical Codes given in the following table correspond to monoattached rings.
- Superatoms CHE (alkenyl) and CHY (alkynyl) are not shown in the table, but the generation of Chemical Codes for these superatoms results in Chemical Code translations similar to those for CHK, with added codes to describe their respective unsaturation.

Superatom	Description	Codes generated by AutoCPI
CHK	Alkyl	Codes for C1 through C $\geq$ 19 alkyl, straight and branched
CHK <sup>LO</sup>	Lower alkyl	Codes for C1 through C6 alkyl, straight and branched
CHK <sup>MID</sup>	Mid alkyl	Codes for C7 through C10 alkyl, straight and branched
CHK <sup>HI</sup>	Higher alkyl	Codes for C11 through C $\geq$ 19 alkyl, straight and branched
CHK <sup>LO, MID, HI</sup>	Lower, mid or higher alkyl	Codes for C1 through C $\geq$ 19 alkyl, straight and branched
CHK <sup>BRA</sup>	Branched alkyl	Codes for C3 through C $\geq$ 19 alkyl, branched only
ARY	Aryl	Codes for general aromatic, mono-substituted benzene and naphthyl (G010, G020, G021, G040, Set G1: code if applicable, G221)
ARY <sup>MON</sup>	Monoaryl	Codes for mono-substituted benzene (G010, Set G1: code if applicable)
ARY <sup>FU</sup>	Fused aryl	Codes for general aromatic and mono-substituted naphthyl (G020, G021, G040, Set G1: code if applicable, G221)
ARY <sup>MON, FU</sup>	Mono or fused aryl	Codes for general aromatic, mono-substituted benzene, and naphthyl (G010, G020, G021, G040, Set G1: code if applicable, G221)
HEA	Monoheteroaryl	Codes for general aromatic heterocycle (F010, F020)
HET	Monoheterocyclic (non aromatic)	Codes for general non-aromatic heterocycle (F010, F021)
HEF	Fused heterocyclic	Codes for general fused heterocycle (D010, D040, D020)
CYC	Cycloaliphatic	Codes for general cycloaliphatic ring (G030, G050, G553, G563)

*cont'd*

Superatom	Description	Codes generated by AutoCPI
CYCMON	Monocycloaliphatic	Codes for general cycloaliphatic ring (G030, G050, G553, G563)
CYCFU	Fused cycloaliphatic	Codes for general polycyclic aliphatic ring system (G031, G032, G051)
CYCUNS	Unsaturated cycloaliphatic	Codes for general cycloaliphatic ring (G030, G050, G551, G552, G561, G562)
CYCSAT	Saturated cycloaliphatic	Codes for general cycloaliphatic ring (G030, G050, G553, G563)
CYCFU, UNS	Fused and unsaturated cycloaliphatic	Codes for general polycyclic aliphatic ring system (G031, G032, G051)
CYCMON, UNS	Unsaturated monocycloaliphatic	Codes for general cycloaliphatic ring (G030, G050, G551, G552, G561, G562)

## 8.5 Changes to indexing policy due to AutoCPI

Since the introduction of AutoCPI, changes have been made to Chemical Code and Markush DARC indexing policies. These changes were made so that indexing is consistently created, whether automatically by AutoCPI or manually by analysts.

### 8.5.1 Chemical code indexing policy changes

Section 8.4 lists the Chemical Codes used by AutoCPI to translate Markush DARC superatoms and their attributes. A number of these translations represent a departure from previous Chemical Code indexing policy. When the translation from Markush DARC superatom is performed manually, analysts follow the rules used by AutoCPI for the process so that consistency of Chemical Code indexing is ensured.

The table below lists superatoms for which pre-AutoCPI Chemical Code indexing differs from the post-AutoCPI Chemical Code indexing presented in Section 8.4.

Superatom	Chemical Codes Pre-AutoCPI	Chemical Codes Post-AutoCPI
ARY	G010, G040, Set G1: code if applicable	G010, G020, G021, G040, G221, Set G1: code if applicable
CYCUNS	G030, G050, G552, G562	G030, G050, G551, G552, G561, G562
HEA	F012, F013, F014, F020, F111, F431	F010, F020
HET	F011, F021, F423, F433	F010, F021
HEF	D010, D020, D040, D601, D621	D010, D020, D040

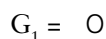
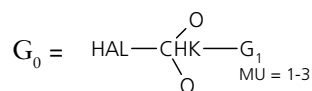
## 8.5.2 Markush DARC indexing policy changes

Since the implementation of AutoCPI, slight modifications in Markush DARC graphics indexing have been introduced to ensure consistent conversions.

### Multiplier attributes

*Multiplier attributes* can only be interpreted by AutoCPI when they are applied to G-groups, not when they are applied to specific atoms. (Multiplier attributes and G-groups are discussed in the *Markush DARC User Manual*.) To produce all Chemical Codes for structures with multiplier attributes, the atoms with multipliers are replaced by G-groups.

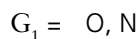
**Example:** For AutoCPI to successfully convert a haloalkyl chain substituted with 3, 4, or 5 hydroxy groups to Chemical Codes, the structure is indexed as follows:



### Essential/Negation Codes

*Essential codes*, also called negation codes, are used by analysts to indicate chemical fragments that are present in all variations of a particular Markush structure. (Essential/negation codes are discussed in the manuals *Introduction to the BCE Chemical Codes* and the *Chemical Retrieval Manual*.) Because AutoCPI produces Chemical Codes from graphics, the original Markush DARC indexing of structures for AutoCPI must be performed in such a way that the proper essential/negation codes are generated.

**Example:** Suppose a patent claims alcohols and primary amines. The following graphics structure would not indicate to AutoCPI that alcohol and amine essential codes are needed, and would therefore not be suitable input to AutoCPI:



Instead it would be necessary for the analyst to divide the graphics diagram shown above in such a way that the amine and alcohol essential codes are generated by AutoCPI. The correct structure diagram is shown below.



and



# Appendix 1 Text Notes

## Controlled Text

*Controlled Text Notes* (CTN's) are applied to superatoms in order to define the entity represented by the superatom more precisely than is possible using superatom attributes. CTN's are always added by analysts when a superatom has been used in lieu of specific atoms or chemical fragments mentioned in a patent.

They are also used to indicate more precisely the range of possible values of a chemical entity described by a superatom, when that range is different than the default range of the superatom. For example, the sentence "C1-18 alkyl, preferably C1-10 alkyl, more preferably C1-6 alkyl" would be indexed with the superatom *CHK*, having attributes *LO*, *MID*, and *HI*, and supplemented by the text note *C1-18*.

On the other hand, CTN's are **not** added when the information they contribute is **included by default** within the definition of the superatom. For example, in the table overleaf, the groups in the left column are fully described by the superatoms in the right column, so no CTN would be necessary to index them. (The default values of all superatoms are listed in the *Markush DARC User Manual*.)

Chemical group	Fully represented by:
Alkyl, 1-6 carbons	CHK <sup>LO</sup>
Alkyl, 1-10 carbons	CHK <sup>LO, MID</sup>
Alkyl, 7-10 carbons	CHK <sup>MID</sup>
Alkenyl, 2-6 carbons	CHE <sup>LO</sup>
Alkynyl, 2-10 carbons	CHY <sup>LO, MID</sup>

CTN's use a **closed set of parameters**, which are listed below. When used in a CTN to describe a particular superatom, each parameter is immediately followed by a numerical value or range of values, indicating the number of times the atom or group can appear in the chemical entity represented by the superatom. (The superscripts used in the list of parameters shown below refer to the following notes.)

CTN Parameter	Definition
E <sup>1</sup>	Number of double bonds
Y	Number of triple bonds
C	Number of carbon atoms
Heteroatom symbol	Number of occurrences of a particular heteroatom
X	Number of heteroatoms not otherwise identified in the Text Note
NR	Number of rings in a ring system
RA <sup>2</sup>	Number of atoms in a heterocyclic ring system
>Atomic symbol	Presence of one attachment to the specified atom
>>Atomic symbol <sup>3</sup>	Presence of more than one attachment to the specified atom

#### Notes:

- <sup>1</sup> *Normalised bonds* in aromatic rings are not counted as double bonds. For example, indene represented by a superatom would receive the annotation *E1*, not *E4*. In ambiguous cases, for instance where a compound may have tautomeric structures with differing numbers of double bonds, the number of double bonds would not be specified.
- <sup>2</sup> The parameter *RA* is used to indicate the number of atoms in a *heterocyclic* ring system, and the parameter *C* is used to indicate the number of atoms in a *carbocyclic* ring system. When used to annotate fused ring systems, *RA* refers to the total number of atoms in the ring system. For example, a superatom used to index quinoline may be described by the CTN parameter *RA10*.
- <sup>3</sup> The parameter *>>atomic symbol*, when used with superatoms *CHK*, *CHE*, *CHY*, and *CYC*, indicates that more than one atom (other than H) is bonded to the same atom.

The following punctuation marks are used in CTN's:

Punctuation	Use
Comma (,)	separates parameters referring to the same superatom
Hyphen (-)	indicates a numerical range (i.e. minimum-maximum)
Period (.)	indicates the end of a CTN
Semicolon (;)	separates portions of the CTN that refer to different superatoms, or portions that refer to alternative definitions of the same superatom

**Example:** Suppose the following CTN appears at the bottom of the graphics screen when viewing the G-group  $G_{10}$  of a patent reference retrieved by a Markush DARC search:

HET1=E1 , C2-4 , S1 , N1-3 , X0 , RA6 ; HEF2=E1-2 , O1 , N1-3 , NR2 , RA10 .

This CTN refers to two superatoms: *HET1* and *HEF2*. The digits appended to *HET* and *HEF* specify precisely which superatoms in  $G_{10}$  are being described. (The superatom numbering doesn't appear automatically on the graphical answer screen; if it is not obvious which superatom is being referred to in a CTN, the attribute *NU* can be displayed so that the superatom numbering will be shown, as explained in the Markush DARC User Manual).

The *HET* superatom described by this CTN represents all unfused heterocycles with one double bond, two to four C atoms, one S atom, one to three N atoms, no heteroatoms other than S and N, and 6 ring members. The *HEF* superatom described by this CTN represents all fused heterocyclic ring systems with one or two double bonds, one O atom, one to three N atoms, two rings, and 10 total ring members.

Note that far fewer rings match the criteria in this CTN than match the superatoms *HET* and *HEF*, even with superatom attributes applied to restrict their values.

Analysts can creatively utilise punctuation in CTN's to focus the coverage of a superatom more narrowly. This is demonstrated in the following example.

**Example:** Suppose a patent mentions an optionally substituted heterocyclic ring that can be any of the following: pyrrole, furan, thiophene, pyridine, or pyrimidine. Suppose further that due to time and/or system constraints (discussed in Section 7.1), the analyst must index these rings with a superatom rather than with specific structures. One possible solution is to use the superatom HEA, supplemented by the following CTN:

$$\text{HEA1}=\text{C4}-5, \text{N0}-2, \text{S0}-1, \text{O0}-1, \text{X0}.$$

This Text Note unfortunately implies many rings not covered by the patent.

The following CTN also includes rings not covered by the patent, i.e. pyridazine and pyrazine, but far fewer than the previous CTN:

$$\text{HEA1}=\text{C4}, \text{N0}-1, \text{O0}-1, \text{S0}-1, \text{X0}, \text{RA5}; \text{HEA1}=\text{C4}-5, \text{N1}-2, \text{X0}, \text{RA6}.$$

Note that the parameter RA5-6 was not needed in the first CTN in this example because, by definition, HEA contains 5 or 6 ring atoms.

**Text for repeating units** is a special form of CTN. The format of the CTN for repeating units is:

$$M100=\text{min}_1-\text{max}_1; M200=\text{min}_2-\text{max}_2.$$

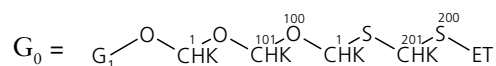
where  $M$  is the symbol that indicates a repeating unit;  $100$  and  $200$  are labels for specific repeating units within a particular G-group; and  $\text{min-max}$  is the number of times the units may repeat. The digits 100 and 200 are displayed as numbering attributes, similar to the 1 in HEA1 of ordinary CTN's, i.e. they identify the repeating group in the display.

**Example:** Suppose a repeating group within  $G_5$ , that is labelled 100 (i.e. NU=100), repeats 1 to 19 times, and another group within  $G_5$ , that is labelled 200, repeats 1 to 10 times. This information could be conveyed in a Text Note at the bottom of an answer display of  $G_5$  as:

$$M100=1-19; M200=1-10.$$

A superatom numbering attribute, when displayed within a repeating unit, will appear as the sum of the repeating unit *NU* and the superatom *NU*. Therefore, superatoms with numbering attributes which form part of a repeat unit, can be identified in the answer display.

**Example:** Suppose the following structure, and its numbering attributes, is displayed in  $G_0$ .



A CHK superatom labelled 1 and included within the repeating unit M100 is displayed as  $\text{CHK}^{101}$  and within repeating unit M200 as  $\text{CHK}^{201}$ . That is to say, all alkyl superatoms, labelled  $\text{CHK}^1$ ,  $\text{CHK}^{101}$  and  $\text{CHK}^{201}$  are described by the same CTN. Suppose the display showed:

$\text{CHK}1 = \text{C}1 - 4.$

This carbon count of 1 to 4 would apply to all the CHK superatoms labelled  $\text{CHK}^1$ ,  $\text{CHK}^{101}$  and  $\text{CHK}^{201}$ . Both parts of the repeating unit,  $\text{CHK}^{101}$  and  $\text{O}^{100}$  are described by the following CTN (to show that the unit can repeat between 1 and 5 times):

$\text{M}100 = 1 - 5.$

Both parts of the repeating unit,  $\text{CHK}^{201}$  and  $\text{S}^{200}$  are described by the CTN (to show that the unit can repeat between 1 and 6 times):

$\text{M}200 = 1 - 6.$

### Free Text Notes

Free Text Notes (FTN's) are added to indexing records when attributes and CTN's are insufficient for conveying relevant information about the coverage of a superatom. FTN's always appear between forward slashes, i.e. /FREE TEXT/, but they are not otherwise limited to a particular format or set of parameters. Unlike CTN's, FTN's may be used to describe either G-groups or superatoms. Also unlike CTN's, FTN's are not consulted by the Markush DARC software; they are provided solely to help the user assess the relevance of an answer to the search question.

The following spiro ring system examples illustrate the combined use of Controlled Text and Free Text. (These structures would usually be drawn out fully; and only indexed with superatoms when time and/or system limits are likely to be breached. Time and system limits are discussed in Section 7.)

**Example:**

Superatom:  $CYC^{FU}$

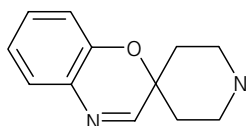
Text Note:  $CYC1=C11, NR2/SPIROBICYCLOHEXANE/$ .

The attribute FU is included to indicate a fused ring system.

**Example:**

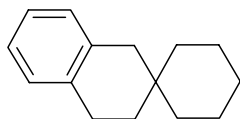
Superatom: HEF

Text Note:  $HEF1=C9, N2, X0, NR2, RA11/4, 4'-SPIROBIPERIDINE/$

**Example:**

Superatom: HEF

Text Note:  $HEF1=C12, N2, O1, X0, E1, NR3, RA15/2H-1, 4-BENZOXAZINE-2-SPIRO-4'-PIPERIDINE/$ .

**Example:**

Superatom:  $ARY^{FU}$

Text Note:  $ARY1=C15, NR3/CYCLOHEXANESPIRO-2'-TETRALIN/$ .

**Note:**

A spiro ring system can only be indexed with the superatoms  $HEF$ ,  $ARY^{FU}$  or  $CYC^{FU}$ . If at least one ring in the spiro system is heterocyclic,  $HEF$  is indexed; if the spiro system has no heterocyclic rings but at least one ring is aromatic,  $ARY^{FU}$  is indexed; otherwise  $CYC^{FU}$  is indexed.