



# Databases

**Helen Cooke**  
**GlaxoSmithKline**

A Celebration of the History of Chemical Information  
29 Nov 2010

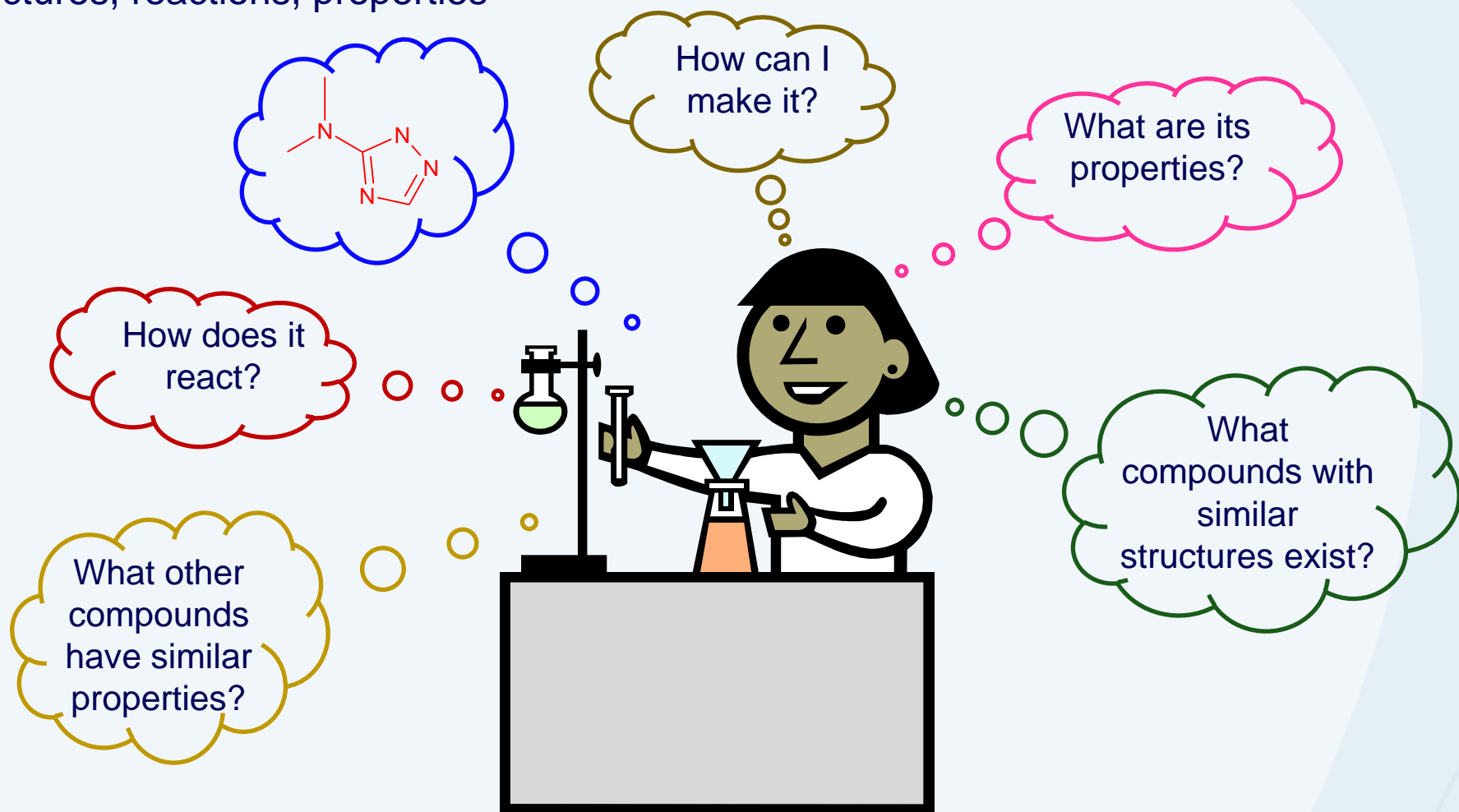
# Agenda

- Information seeking habits of chemists
- Precursors
- Drivers and enablers
- Chronology
- Experiences of a chemical information specialist

Scope: the focus of this presentation is structure-based databases concerned with published literature, not organisations' internally generated or proprietary information

# Needs / information-seeking behaviours of chemists

The majority of questions chemists ask are focused on compounds - their structures, reactions, properties

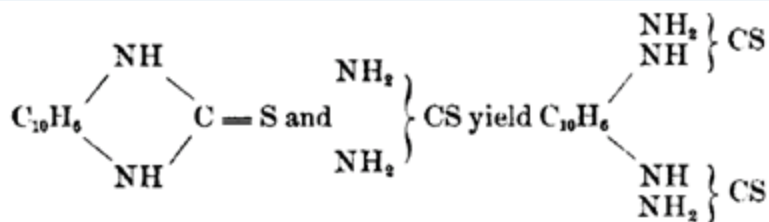


# Communicating chemical compound information

The structure diagram has been the internationally understood face-to-face and print communication language for chemists since the early 1900s

- Although structure diagrams had been around since the mid-1800s, due to printing technology limitations, they only started to appear regularly in publications in the early 1900s

Example of a reaction from 1879:



# Solutions devised by producers of secondary information sources

Publishers recognised chemists' unique language of structures and reaction schemes, and sought ways to provide structure-focused searching, through:

- Innovative classification schemes, as deployed through, e.g. *Beilstein*, *Gmelin*, *Chemical Abstracts*, *British Chemical Abstracts*
  - Some were also used to help organise and retrieve information in electronic versions of the content, e.g. *Chemical Abstracts* section headings
- Indexing in accordance with nomenclature standards, e.g. IUPAC, *Chemical Abstracts*
- Standards for ordering chemical formulae, e.g. Hill system

The challenge for chemical database producers was to replicate and improve upon printed secondary information sources

# Predecessors and enablers

Key Word In Context (KWIC) systems (from early 1960s), e.g.

- CAS's *Chemical Titles*, from 1961
- Specialist KWIC *Index to Neurochemistry*, produced by *Excerpta Medica* and IBM in 1961

Chemical registration systems, e.g.

- Chemical Abstracts Registry System, initiated in 1964

Punched and edge-notched cards



# Edge-notched card from SCIFAX DTA data index

*A Powerful Aid to Scientific Research*

## SCIFAX D.T.A. DATA INDEX

*A Punched Card Index of Data and References for*

### DIFFERENTIAL THERMAL ANALYSIS

*with Mineral, Inorganic and Organic Sections*

Please Return to:-  
INFORMATION SECTION,  
Unilever Research Laboratory,  
PORT SUNLIGHT.

PRINCIPAL PEAK CODING: 925-1050, 775-825, 14,23,7

SECOND PEAK CODING: ACTINOLITE

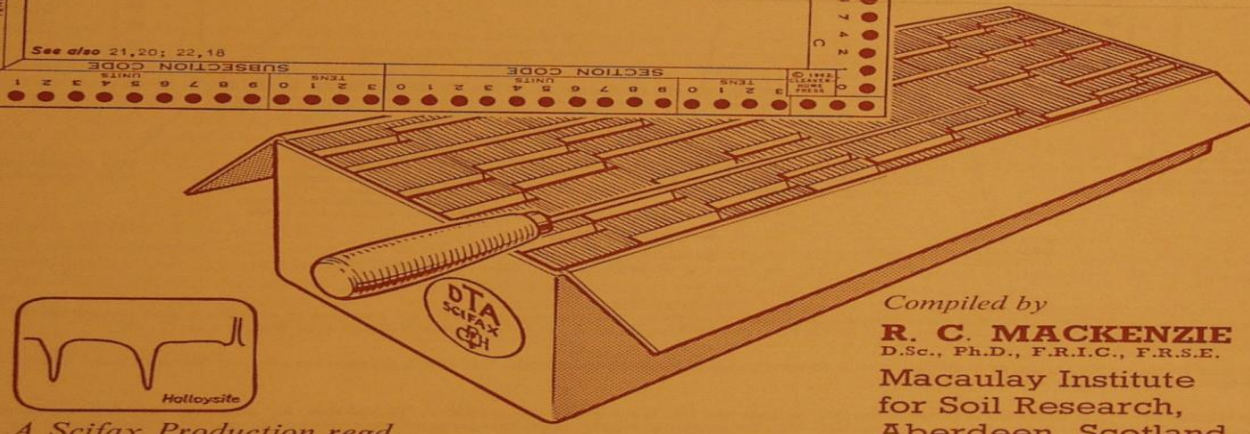
ENTRY CODE: 21,18

Chemical formula:  $Ca_2(Mg,Fe)_5Si_8O_{22}(OH)_2$

T(-)	S	T(+)	S	T(-)	S	T(+)	S	T(+)	S	Ref
780	vs	1020 <sup>L</sup>	ms	1190	vs			1275 <sup>L</sup>	vs	1
789	vs	1002 <sup>L</sup>	ms							2
		1049 <sup>p</sup>	ms							

1. KORZHINSKII, Doklady Akad. Nauk SSSR, 1956, 111, 445.  
2. KORZHINSKII, Trudy i Soveshch. Termosk., 1955, 2, 266.

See also 21,20; 22,18



*A Scifax Production regd.*

Compiled by  
**R. C. MACKENZIE**  
D.Sc., Ph.D., F.R.I.C., F.R.S.E.  
Macaulay Institute  
for Soil Research,  
Aberdeen, Scotland

**CLEAVER-HUME PRESS LTD. LONDON**

1955

1956

# Findex

Regd.

## PUNCHED CARD System

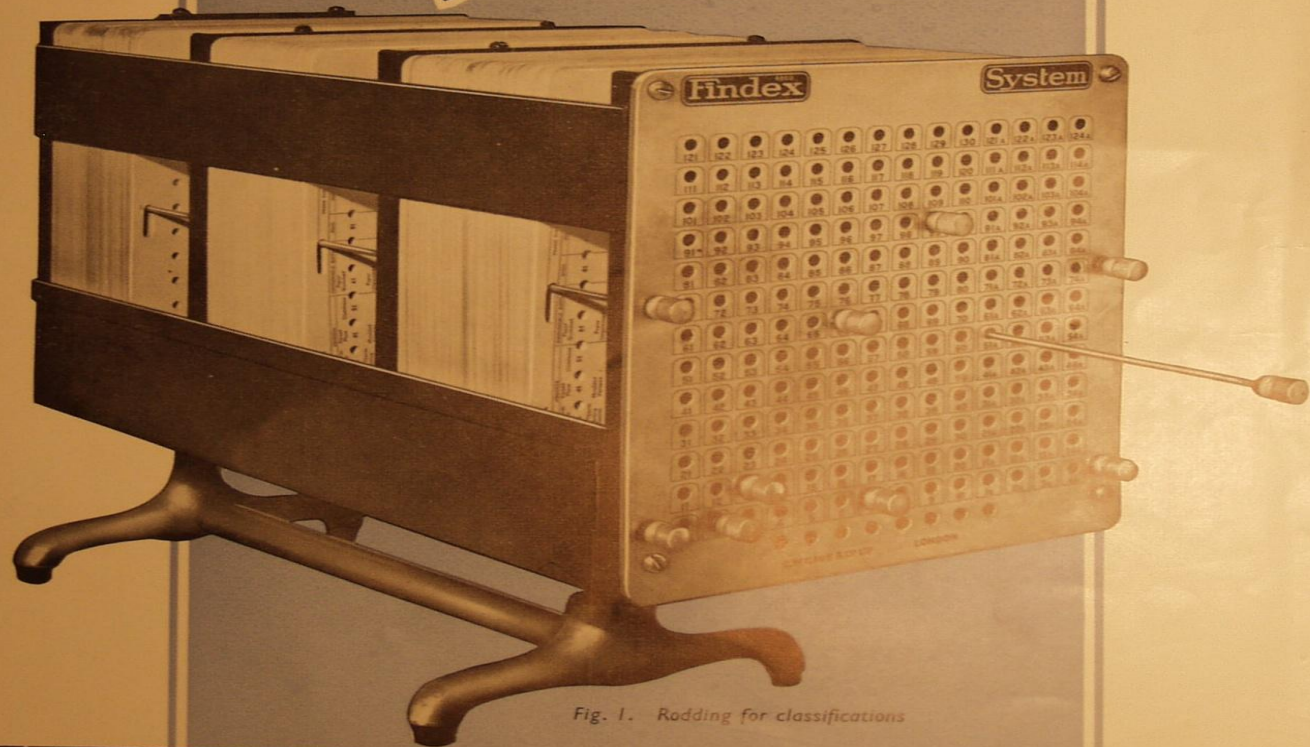


Fig. 1. Rodding for classifications

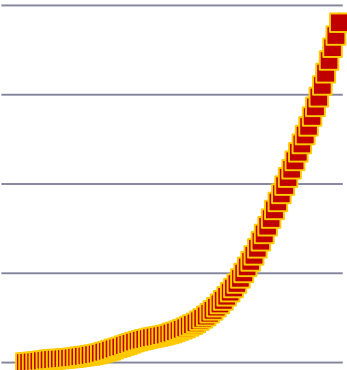
*1,000 Indexes in One*

FOR PERMANENT OR  
SEMI-PERMANENT RECORDS

# Chemical databases: drivers and enablers



Advances in  
IT, telecoms,  
networks and  
research



Literature  
explosion



Commercial  
impetus from  
pharma and  
chemical  
industries

A grid of chemical codes and notations. The grid is organized into rows and columns, with each cell containing a small icon or symbol representing a specific chemical structure or notation. The symbols include numbers, letters, and geometric shapes like circles and hexagons.

Development  
of chemical  
codes,  
languages  
and notations

# Advances in telecoms, IT and research

## Computers used for the production of printed information sources

- Realisation that the repositories created could be made searchable

## Computer power increased

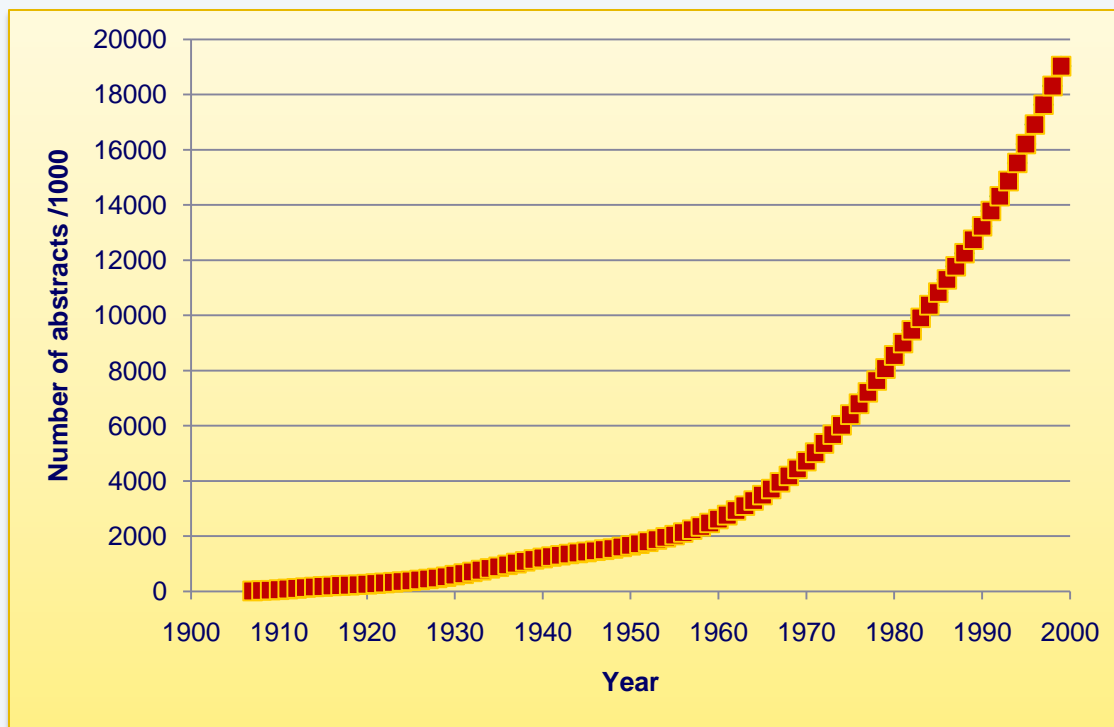
## Telecommunications improved

## Research at universities and publishers

- e.g. Sheffield University Department of Information Studies, partnered with industry and publishers to develop solutions to chemical information needs

# Literature explosion

Growth in literature made it increasingly difficult for chemists to find the information they were looking for, and better methods were needed



Number of abstracts in *Chemical Abstracts*

## The pharmaceutical and chemical industries:

- Needed efficient methods of finding chemical information
- Were able to provide financial backing for research
- Could afford the necessary hardware, professional staff, and subscriptions to chemical information systems

# Development of languages / notations

Derwent  
Ringdoc  
chemical  
code

1	2	3	4	5	6	7	8	9	10
					<b>G</b>				

CLEARTEXT:

11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
12	SPECIAL CODE	1	ISOLATED	ISOLATED	LINEAR	ISOLATED	POLY	POLY	POLY	POLY	POLY	POLY	POLY	POLY	POLY	POLY
11	STERIOD	2	CONDENSED AROMATIC	CONDENSED AROMATIC	ANGULAR	CONDENSED HETERO-CYCLIC	SEVERAL LIKE X IN RING	X TO SEVERAL RINGS	ANG. SUBSTN.	HET. CYC.						
0	CARBO-HYDRATE	3	CONDENSED ALICYCLIC	CONDENSED ALICYCLIC	ANG-ANG	CONDENSED HETERO-CYCLIC	SEVERAL DIFF. X IN RING	SEVERAL DIFF. X IN RING	GEM. SUBSTN.	HET. AT SUB.						
1	PROTEIN	4	CONDENSED HETERO-CYCLIC	CONDENSED HETERO-CYCLIC	PERI	CONDENSED HETERO-CYCLIC	SEVERAL DIFF. X IN RING	SEVERAL DIFF. X IN RING	ALICYCLIC	ALICYCLIC						
2	NUCLEIC ACID	5	1	2	BRIDGE	2	N	1,2	2	2	2	2	2	2	2	2
3	OTHER NATURAL SUBSTANCE	≥6	2	3	SPIRO	3	N	1,3	β	β	β	β	β	β	β	β
4	POLYMER	ISOLATION	3	7	WITH 3 DB	≥4N	1,2,4	1,2	1,2,3	≥7	CH,X	5	4	4	4	4
5	5	ADJAC. DETECTION	4	≥5	UNSATURATED	1,0	≥3,1M	1,3	1,3,5	4	1	5	5	5	5	5
6	INORGANIC	PROPERTIES	≥5	≥7	SATURATED	≥2,0	≥3,2M	≥1,4	≥1,4,5	2	2	2	2	2	2	2
7	CLEARTEXT	SYNTHESIS	APPL.	Δ	Δ	S	≥3	1,2,3	1,2,3	6	3	3	3	3	3	3
8	PREP.	STRUCTURE UNDETERMINED	ANALYSIS	□	□	≥7	3,5,8	1,2,4	1,2,3,7	4,6	4,6	4,6	4,6	4,6	4,6	4,6
9	FROM THE LITERATURE	ORGANIC FORMULA	9	9	ALDEHYDE	Δ	OTHER HETERO-CYCLIC	HETERO-CYCLIC	2,4	2,4	2,4	2,4	2,4	2,4	2,4	2,4

CLEARTEXT:

46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62
12	SPECIAL CODE	1	ISOLATED	ISOLATED	LINEAR	ISOLATED	POLY	POLY	POLY	POLY	POLY	POLY	POLY	POLY	POLY	POLY
11	STERIOD	2	CONDENSED AROMATIC	CONDENSED AROMATIC	ANGULAR	CONDENSED HETERO-CYCLIC	SEVERAL LIKE X IN RING	X TO SEVERAL RINGS	ANG. SUBSTN.	HET. CYC.						
0	CARBO-HYDRATE	3	CONDENSED ALICYCLIC	CONDENSED ALICYCLIC	ANG-ANG	CONDENSED HETERO-CYCLIC	SEVERAL DIFF. X IN RING	SEVERAL DIFF. X IN RING	GEM. SUBSTN.	HET. AT SUB.						
1	PROTEIN	4	CONDENSED HETERO-CYCLIC	CONDENSED HETERO-CYCLIC	PERI	CONDENSED HETERO-CYCLIC	SEVERAL DIFF. X IN RING	SEVERAL DIFF. X IN RING	ALICYCLIC	ALICYCLIC						
2	NUCLEIC ACID	5	1	2	BRIDGE	2	N	1,2	2	2	2	2	2	2	2	2
3	OTHER NATURAL SUBSTANCE	≥6	2	3	SPIRO	3	N	1,3	β	β	β	β	β	β	β	β
4	POLYMER	ISOLATION	3	7	WITH 3 DB	≥4N	1,2,4	1,2	1,2,3	≥7	CH,X	5	4	4	4	4
5	5	ADJAC. DETECTION	4	≥5	UNSATURATED	1,0	≥3,1M	1,3	1,3,5	4	1	5	5	5	5	5
6	INORGANIC	PROPERTIES	≥5	≥7	SATURATED	≥2,0	≥3,2M	≥1,4	≥1,4,5	2	2	2	2	2	2	2
7	CLEARTEXT	SYNTHESIS	APPL.	Δ	Δ	S	≥3	1,2,3	1,2,3	6	3	3	3	3	3	3
8	PREP.	STRUCTURE UNDETERMINED	ANALYSIS	□	□	≥7	3,5,8	1,2,4	1,2,3,7	4,6	4,6	4,6	4,6	4,6	4,6	4,6
9	FROM THE LITERATURE	ORGANIC FORMULA	9	9	ALDEHYDE	Δ	OTHER HETERO-CYCLIC	HETERO-CYCLIC	2,4	2,4	2,4	2,4	2,4	2,4	2,4	2,4

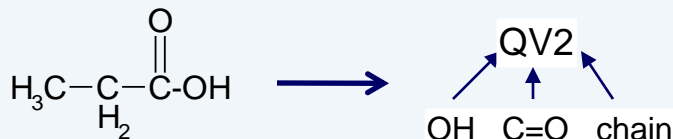
R2

28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
EXTRA PUNCHES H																	

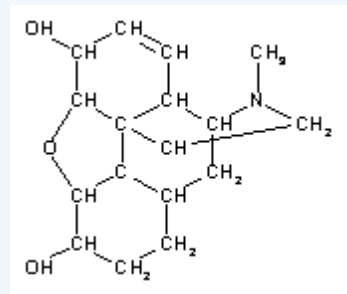
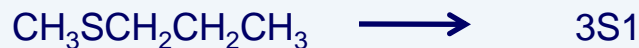
11	12	13	14	15	16	17	18	19	20	21
12	SPECIAL CODE	1	ISOLATED	ISOLATED	LINEAR	ISOLATED	POLY	POLY	POLY	POLY
11	STERIOD	2	CONDENSED AROMATIC	CONDENSED AROMATIC	ANGULAR	CONDENSED HETERO-CYCLIC	SEVERAL LIKE X IN RING	X TO SEVERAL RINGS	ANG. SUBSTN.	HET. CYC.
0	CARBO-HYDRATE	3	CONDENSED ALICYCLIC	CONDENSED ALICYCLIC	ANG-ANG	CONDENSED HETERO-CYCLIC	SEVERAL DIFF. X IN RING	SEVERAL DIFF. X IN RING	GEM. SUBSTN.	HET. AT SUB.
1	PROTEIN	4	CONDENSED HETERO-CYCLIC	CONDENSED HETERO-CYCLIC	PERI	CONDENSED HETERO-CYCLIC	SEVERAL DIFF. X IN RING	SEVERAL DIFF. X IN RING	ALICYCLIC	ALICYCLIC
2	NUCLEIC ACID	5	1	2	BRIDGE	2	N	1,2	2	2
3	OTHER NATURAL SUBSTANCE	≥6	2	3	SPIRO	3	N	1,3	β	β
4	POLYMER	ISOLATION	3	7	WITH 3 DB	≥4N	1,2,4	1,2	1,2,4	1,2,4

# Development of languages / notations – linear notations

- Wiswesser



- SMILES

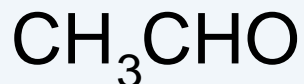


- Dyson

- Greimas

- 
- 
-

# Development of languages / notations – connection tables

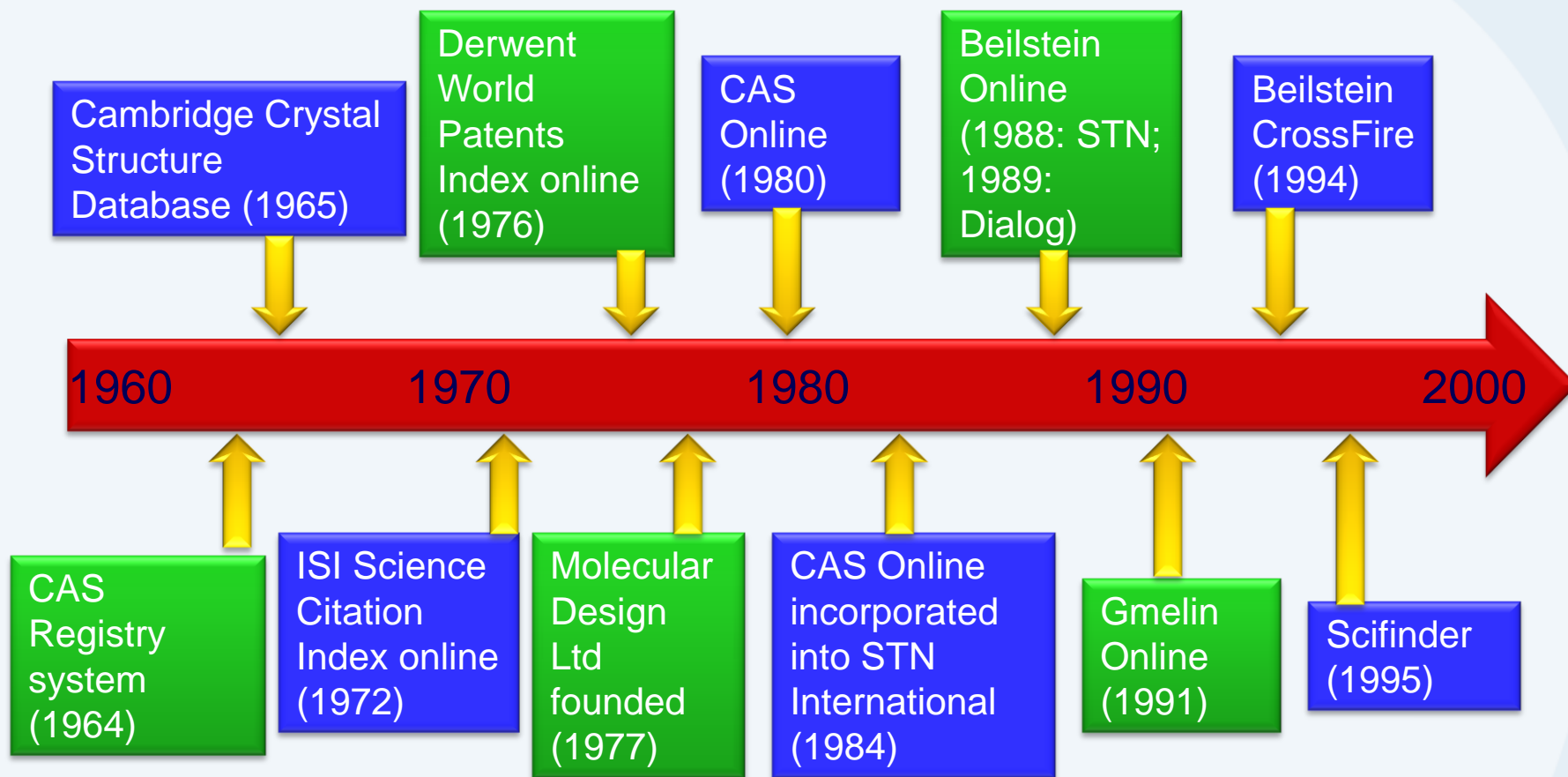


	C	C	H	H	H	H	H	O
C	-	1	1	1	1	0	0	0
C		-	0	0	0	1	2	
H			-	0	0	0	0	
H				-	0	0	0	
H					-	0	0	
H						-	0	
O							-	

0 = no bond, 1 = single bond, 2 = double bond,  
3 = triple bond, - = no bond possible

- Chemical Abstracts Service uses connection tables to enable structure searching capabilities for CAS's products.
  - First devised by Harry L Morgan, starting in 1962
- MDL's MOL files are connection tables used in ISIS databases, CrossFire, and many proprietary database systems

# Some landmarks in chemical database history



# The search experience

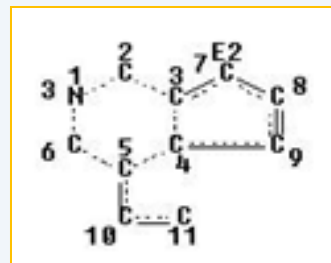
- Printed indexes: nomenclature, classification schemes
- Early databases: chemical codes (e.g. Derwent), command languages (e.g. CAS Online)

=> GRA R65, 5 C2

=> BON ALL S, 8-9 D, R 1 2 N

=> NOD 1 N, VAL 1 E3

=> HCO 7 E2



- Recent systems: structure drawings and reaction schemes

# End-user experience

-1980

## Printed secondary sources

- Chemical Abstracts
- Beilstein and Gmelin Handbooks
- British Chemical Abstracts

Empowered users aided by chemistry librarians

~1980 -1995

## Online pay as you go systems, e.g.,

- CAS Online (CA, Registry, etc.)
  - Derwent World Patent Index, World Drug Index
  - Beilstein Online
- Delivered by online service providers, e.g. Dialog, Questel, STN, Datastar

Disempowered end users, enhanced role for info scientists, librarians

~1995 -

## End-user tools

- CD-ROM
- CA on CD
- Medline
- SCI
- Client-Server, e.g.
- SciFinder
- CrossFire

Re-empowered end-users, librarians/info scientists become trainers, and provide tech and application support

# Searching for all in UK academia: key innovations leading to empowered end-users

Daresbury Chemical Database Service

STN academic discount programme made chemical structure searching affordable

Government-funded services, e.g. MIMAS

Science Citation Index (Web of Knowledge)

CrossFire Beilstein and Gmelin

SciFinder

# Spectrum of end-user structure-searchable chemistry databases available today

SciFinder

Beilstein /  
Reaxys

Gmelin

Cambridge  
Structural  
Database

Available  
Chemicals  
Directory

Merck Index

Materials  
Safety Data  
Sheets

e-EROS

ChemSpider