Host-Enhanced Chemical Indexing in Technical Databases

EDWARD W. BADGER* and CLAUS-DIETER SIEMS

STN International, c/o Chemical Abstracts Service, P.O. Box 3012, Columbus, Ohio 43235, and STN International, c/o FIZ Karlsruhe, Postfach 2465, D-7500 Karlsruhe 1, Federal Republic of Germany

Received June 15, 1989

Many files that index engineering, physics, or other technical literature contain references to chemical compounds. Complex chemical formulas in the title or abstract often contain special characters, for example, (,), [,], +, −, °, and %. Upper and lower case letters often are also included. A search of these formulas and symbols in the basic index is next to impossible because the terms in this field are usually parsed to all alphanumeric characters without sensitivity to case. It is possible for an online host to use a character-recognition algorithm to scan the title and abstract data for special characters or character strings and place them in a separate index field when such files are loaded. Such an algorithm has been designed by Fachinformationszentrun Energie, Physik, Mathematik GmbH in Karlsruhe, West Germany (FIZ Karlsruhe), the European service center of STN International, the scientific and technical information network. This algorithm recognizes and analyzes chemical formulas, material descriptions, alloys, and eutectic systems as well as nuclear reactions and dopings that appear in the title, abstract, or other fields. These character strings are converted into a standardized form and placed in a new field (the element terms field) which is supplied by the online host during the loading process. A checklist of allowed terms (symbols and chemical formulas) is used to prevent irrelevant terms from being mistaken for legitimate chemical symbols. For instance, the algorithm can recognize that CPU (central processing unit) is not a legitimate chemical formula. It is easy to demonstrate the utility this additional index supplies. For example, when this new index field is specified during a search for the character string "bk", it becomes possible to retrieve only studies on the element berkelium (Bk) and not studies on BK-7 borosilicate glass. This specification is possible because the algorithm checklist can detect the difference between Bk and BK. This paper will discuss this algorithm and give search examples to further demonstrate the utility of this innovation.

INTRODUCTION

Complex, nonword combinations of characters often appear in chemical formulas appearing in the title or abstract of a bibliographic citation. These formulas can contain special characters, symbols, and upper and lower case letters. Searching for these symbols or formulas in the basic index of most databases is difficult since the text of this field is parsed at all nonalphanumeric characters; that is, all punctuation and special symbols such as %, +, =, −,), (, [, and] are replaced with blank spaces. Strings containing these characters are divided into component words or word fragments. In addition, most indexes do not differentiate between case: text containing the strings CO, Co, co, or even cO are considered the same in the basic index. Searching on each of these strings in the basic index will result in the same answer set. These limitations present some difficulties to one seeking chemical information in databases that are not specifically designed to index chemical substances. To provide access to the chemical information contained in such files, a special computer program has been devised that recognizes and analyzes symbols and letters that make up the (typically inorganic) chemical formulas, materials descriptions, alloys, and eutectic systems as well as nuclear reactions and dopings that appear in the text of records in engineering, electronics, and other nonchemical technical databases. The algorithm can scan for this information in any existing record fields, usually the title and abstract, and convert the symbols into a standard searchable format. This reformated information is placed in a new indexing field, the element terms (ET) field, which is appended to the record's existing index fields. This algorithm contains character-string recognition checklists, which ensure that only element terms and chemically relevant symbols enter the ET field. The ET field is available exclusively on STN International and appears in these files: COMPENDEX, ENERGIE, ENERGY, INIS, INSPEC, LPHYS, MATBUS, MEET, METADEX, and PHYS.

THE ALGORITHM

This IBM mainframe based proprietary computer program was conceived and developed by programmers and information scientists at Fachinformationszentrun Energie, Physik, Mathematik GmbH in Karlsruhe, West Germany (FIZ Karlsruhe), the European service center for STN International, the scientific and technical information network. The algorithm recognizes chemical symbols and characters by comparing them with a checklist of acceptable material descriptions. Acceptable combinations are then examined to determine whether the material being described is an element, an isotope, an ion, doped material, a nuclear reaction, a chemical compound, or a system (formulas containing two or more metals or semimetals are classified as systems). Appropriate additional index terms are then generated and placed in the ET field. During this process, super- and subscripts are converted to ordinary line numbers, Greek letters are spelled out, and most punctuation symbols are retained intact.

Each formula is then broken up and recombined so that all of the information each formula contains becomes searchable. In the case of a chemical formula for a compound such as uranyl pentachloride (UCl_5), as shown in example 1, the ET field will contain five separate items: the intact formula (UCl5), the member elements given in Hill order separated by asterisks (Cl*U), each of the member elements with the letters cp to indicate that they are part a compound (U cp, Cl cp), and, finally, cp alone to indicate that the record contains a chemical compound.

In the case of more complex formulas, such as for the organic ammonium salt $NH_4(C_6H_2N_3O_7)C_{12}H_8N_2$, all of the symbols and numbers are retained, and the formula is presented just as it appears in the record. All of the element codes and cp terms in the ET field are entered as "bound phrases" and are searched that way. If the intact formula is searched, any grouping symbols such as brackets or parentheses must be masked either by enclosing them in quotes or by substituting

0095-2338/89/1629-0279$01.50/0

EXAMPLE 1

```
FILE 'PHYS' ENTERED AT 15:26:18 ON 11 AUG 88
COPYRIGHT 1988 (c) FACHINFORMATIONSZENTRUM KARLSRUHE

FILE LAST UPDATED: 05 AUG 88      <880805/UP>

=> S UCl5/ET
L1          10 UCL5/ET

=> DIS L1 ALL

L1  ANSWER 1 OF 10
AN  84(8):45741  PHYS
TI  Standard enthalpies of formation of UCl5, UCl6, and UOCl correction
    of errors in previous papers.
AU  Cordfunke, E.H.P.; Ouweltjes, W.; Prins, G.; Vlaanderen, P. van
    (Stichting Energieonderzoek Centrum Nederland, Petten)
SO  J. Chem. Thermodyn. (Nov 1983) v. 15(11) p. 1103-1104
    Published in summary form only
    ISSN 0021-9614; CODEN JCTDA
CY  UNITED KINGDOM
DT  Journal; Short Communication
TC  Info
LA  English
CC  *8260
CT  *URANIUM CHLORIDES; *URANYL CHLORIDES; *FORMATION HEAT; CALORIMETRY
ET  Cl*U; UCl5; U cp; cp; Cl cp; UCl6; Cl*O*U; UOCl; O cp
```

the "!" single-character truncation symbol. For example, the ammonium compound listed above could be searched by using any of these ways:

⇒ S "NH4(C6H2N3O7)C12H8N2"/ET

⇒ S NH4"("C6H2N3O7")"C12H8N2/ET

⇒ S NH4!C6H2N3O7!C12H8N2/ET

The parentheses must be masked because the search software interprets parentheses as proximity operators or nested search term delimiters. However, a search term containing special characters may be selected from an EXPAND command, as shown in example 2.

EXAMPLE 2

```
FILE 'PHYS' ENTERED AT 10:54:36 ON 18 OCT 88
COPYRIGHT 1988 (c) FACHINFORMATIONSZENTRUM KARLSRUHE

FILE LAST UPDATED: 14 OCT 88      <881014/UP>

=> E NH4!C/ET
E1          412 NH4 IP 1/ET
E2            1 NH4 IP 2/ET
E3            0 NH4!C/ET
E4            1 NH4(1-X)D4XH2(1-X)D2XASO4/ET
E5            1 NH4(ALPHA)/ET
E6            1 NH4(AUCL4)/ET
E7            1 NH4(B4O5(OH)4)/ET
E8            1 NH4(B5O6(OH)4)/ET
E9            1 NH4(B5O6(OH)4).2H2O/ET
E10           1 NH4(BETA)/ET
E11           1 NH4(C4H5O5).H2O/ET
E12           1 NH4(C6H2N3O7)C12H8N2/ET

=> S E12
L1            1 "NH4(C6H2N3O7)C12H8N2"/ET

=> DIS L1 TI, IND

L1  ANSWER 1 OF 1
TI  Structure of the complex formed between ammonium picrate and
    1,10-phenanthroline.
AN  82(19):86112  PHYS
CC  *6165
CT  *AMMONIUM COMPOUNDS; PHENANTHROLINES; *DIMERS; *CRYSTAL STRUCTURE;
    LATTICE PARAMETERS; X-RAY DIFFRACTION; SPACE GROUPS
ET  C*H*N*O; NH4(C6H2N3O7)C12H8N2; N cp; cp; H cp; C cp; O cp; P; Mg; Mo;
    H*N; NH4; NH+4; NH ip 1; ip 1
```

The additional information contained in the formula of labeled isotopes or isotopic compounds is added to the element terms field and thus becomes searchable. For instance, labeled potassium hydroxide, ^{39}KOH, will generate many entries in the ET field. First, the formula will appear exactly as in the record. Next, there will be an entry with the isotope superscript removed so a search for the unlabeled compound will retrieve this record. These two entries are followed by two isotope specifiers for each labeled atom. Potassium-39 will be listed as 39K, followed by K is. These entries make it possible to search for all the author-specified isotopes of a particular atom by seeking the atom in a bound phrase with the is term. Finally, the is label appears alone to facilitate the retrieval of papers dealing with general isotopic or radiolabeled compounds. If the citation discusses a range of isotopes, such as $^{24-26}$Mg, then the range will be given, followed by a specific isotope entry for each member in the range. In this example, 24Mg, 25Mg, and 26Mg will all appear in the ET field, which will include the bound phrase Mg is, an element entry for Mg, and the is tag by itself, as in the labeled compound example. Example 3 shows how this information appears in the PHYS database.

EXAMPLE 3

```
FILE 'PHYS' ENTERED AT 11:03:23 ON 18 OCT 88
COPYRIGHT 1988 (c) FACHINFORMATIONSZENTRUM KARLSRUHE

FILE LAST UPDATED: 14 OCT 88      <881014/UP>

=> S 39KOH/ET
            1 39KOH/ET
L4          1 39KOH/ET

=> DIS L4 ALL

L4  ANSWER 1 OF 1
AN  87(21):99572  PHYS
TI  Nuclear quadrupole structure of potassium hydroxide in the 32 GHz
    microwave region.
AU  Raw, T.T.; Yamamura, T.; Gillies, C.W. (Department of Chemistry,
    Rensselaer Polytechnic Institute, Troy, New York 12180-3590 (USA))
SO  J. Chem. Phys. (15 Sep 1987) v. 87(6) p. 3706-3707
    Published in summary form only Current Physics Microform No.:
    8709E0446
    ISSN 0021-9606; CODEN JCPSA
    Contract AC02-81ER10909
CY  UNITED STATES
DT  Journal
TC  Experimental
LA  English
AB  A high temperature Stark and radiofrequency-microwave double
    resonance spectrometer was used to observe the nuclear quadrupole
    structure of the J = 2<-1 transition of 39KOH in a number of
    vibrational states. First order analysis of the hyperfine components
    gave nuclear quadrupole coupling constants of eQq = -7.23 (29) MHz
    for the (0000) state, eQq = -7.16 (33) MHz for the (1000) state, and
    eQq = -7.29 (37) MHz for the (0110) state.
CC  *3320; 3355; 3520
CT  *POTASSIUM HYDROXIDES; *MICROWAVE SPECTRA; *STARK EFFECT; *QUADRUPOLE
    MOMENTS; HYPERFINE STRUCTURE; COUPLING CONSTANTS; DOUBLE RESONANCE
    METHODS
ET  H*K*O; KOH; 39KOH; is; K is; 39K; K cp; cp; O cp; H cp

=> S 24-26MG/ET
L5          2 24-26MG/ET

=> DIS L5 ALL 1

L3  ANSWER 1 OF 2
AN  83(5):19490  PHYS
TI  Excitation of two nucleons into the continuum as a doorway process
    for bound-pion absorption by nuclei.
AU  Birbrair, B.L.; Kalashnikov, Yu.A. (Leningrad Nuclear Physics Inst.,
    Gatchina (USSR))
SO  J. Phys., G (London). Nucl. Phys. (Nov 1982) v. 8(11) p. 1531-1545
    ISSN 0305-4616; CODEN JFHGB
CY  UNITED KINGDOM
DT  Journal
TC  Theoretical
LA  English
AB  The problem of bound-pion absorption by nuclei has been studied in a
    semi-microscopic way. Calculations of the widths of the 1s, 2p and 3d
    levels in 12C, 14N, 16-18O, 19F, 20Ne, 23Na, 24-26Mg, 27Al, 28Si,
    31P, 32S, 35Cl, 39K, 40Ar, 40-48Ca, 46-50Ti, 51V and 52Cr and the 4f
    level in 209Bi pi--atoms based on the nuclear self-consistent theory
    of Birbrair et al and Alkhazov et al are performed assuming that the
    ''two-nucleon'' mechanism is the doorway process for nuclear pion
    absorption (Birbrair and Kalashnikov). Both P and S absorption
    mechanisms together with their interference are included here. This
    enables us to calculate the widths of the 1s levels where S
    absorption dominates. The results are in good agreement with
    experiment. (orig.)
CC  *3610
CT  *PIONIC ATOMS; *ABSORPTION; PION MINUS REACTIONS; BOUND STATE; LEVEL
    WIDTHS; S STATES; P STATES; D STATES; SELF-CONSISTENT FIELD; CARBON
    12; NITROGEN 14; OXYGEN ISOTOPES; FLUORINE 19; NEON 20; SODIUM 23;
    MAGNESIUM ISOTOPES; ALUMINIUM 27; SILICON 28; PHOSPHORUS 31; SULFUR
    32; CHLORINE 35; POTASSIUM 39; ARGON 40; CALCIUM ISOTOPES; TITANIUM
    ISOTOPES; VANADIUM 51; CHROMIUM 52; BISMUTH 209; F STATES; EXCITATION
ET  C; 12C; is; C is; N; 14N; N is; O; 16-18O; O is; 18O; 17O; 16O; F;
    19F; F is; Ne; 20Ne; Ne is; Na; 23Na; Na is; Mg; 24-26Mg; Mg is;
    26Mg; 25Mg; 24Mg; Al; 27Al; Al is; Si; 28Si; Si is; P; 31P; P is; S;
    32S; S is; Cl; 35Cl; Cl is; K; 39K; K is; Ar; 40Ar; Ar is; Ca;
    40-48Ca; Ca is; 48Ca; 47Ca; 46Ca; 45Ca; 44Ca; 43Ca; 42Ca; 41Ca; 40Ca;
    Ti; 46-50Ti; Ti is; 50Ti; 49Ti; 48Ti; 47Ti; 46Ti; V; 51V; V is; Cr;
    52Cr; Cr is; Bi; 209Bi; Bi is; pi; pi-
```

Researchers writing about doped elements use the convention A:B, where A is the major component, the doped element, and B is the minor component, the doping element. Silicon

doped with a trace of silver would be written as Si:Ag. The element term identification algorithm recognizes this and adds information that identifies the doped and doping elements, as well as the fact that the citation deals with doped materials. In the case of Si:Ag, the ET field would contain the following entries: Si:Ag, Ag doping, doped materials, Si, and Ag*Si. Consistent with the indexing added for compounds, Ag doping and doped materials are directly searchable as bound phrases. The algorithm will not generate an entry for the dopant element from the formula convention; however, as the doping elements are generally mentioned by themselves in the title or abstract, they usually appear in the ET field. Example 4 shows how doped materials appear in the ET field.

EXAMPLE 4

```
FILE 'INSPEC' ENTERED AT 11:11:37 ON 18 OCT 88
COPYRIGHT 1988 (c) INSTITUTION OF ELECTRICAL ENGINEERS

FILE LAST UPDATED: 15 OCT 88      <881015/UP>

=> S SI:AG/ET
L1          29 SI:AG/ET

=> DIS L1 TI, IND 3

L1  ANSWER 3 OF 29
TI  Diffusion, solubility and segregation of implanted Cu, Ag and Au in
    amorphous Si.
AN  88:3059564  INSPEC     DN  A88021118
CC  5A6630J; 5A6475; 5A6170T
CT  AMORPHOUS SEMICONDUCTORS; DIFFUSION IN SOLIDS; ELEMENTAL
    SEMICONDUCTORS; ION IMPLANTATION; SEGREGATION; SEMICONDUCTOR DOPING;
    SILICON; SOLUBILITY
ST  150 to 600 degrees C; Si:Cu; diffusion; Si:Ag; semiconductor; Si:Au;
    solubility; segregation; Si; diffusion coefficients; Arrhenius
    relationships; activation energies; solubility; amorphous Si;
    amorphous-crystalline interface; 150 to 600 degC; Si; Si:Cu; Si:Ag;
    Si:Au
ET  Cu; Ag; Au; Si; C; Cu*Si; Cu sy 2; sy 2; Si sy 2; Si:Cu; Cu doping;
    doped materials; Ag*Si; Ag sy 2; Si:Ag; Ag doping; Au*Si; Au sy 2;
    Si:Au; Au doping
CHI Si el; Si:Cu bin, Cu bin, Si bin, Cu el, Si el, Cu dop; Si:Ag bin, Ag
    bin, Si bin, Ag el, Si el, Ag dop; Si:Au bin, Au bin, Si bin, Au el,
    Si el, Au dop
PHP temperature 4.23E+02 to 8.73E+02 K
```

In the case of ionic compounds or elements, the material is indexed as described above with the addition of descriptors indicating the ionic state of the material and a general descriptor indicating the citation contains an ion of a given valence. A negative ion is indexed with in for ion negative; a positive ion is tagged with ip for ion positive. The chloride ion (Cl^-), for example, will appear in the ET field as four separate entries: Cl^-, the ion as it appears in the record; Cl in 1, literally "chlorine, ion negative, valence = −1"; Cl, the element is indexed; and, finally, in 1 is included to indicate the citation contains something with a negative valence of 1. In the case of positive ions, analogous entries are included, substituting ip for in. Example 5 shows how the sulfate anion

EXAMPLE 5

```
FILE 'COMPENDEX' ENTERED AT 11:26:09 ON 18 OCT 88
COPYRIGHT (C) 1988 ENGINEERING INFORMATION, INC.

FILE LAST UPDATED: 03 OCT 88      <881003/UP>
FOR CONFERENCE PAPERS SEE FILE MEET

=> S IP?/ET AND IN?/ET and SO4-2/ET
        16389 IP?/ET
        29831 IN?/ET
           23 SO4-2/ET
L1          5 IP?/ET AND IN?/ET AND SO4-2/ET

=> DIS L1 TI, IND 2

L1  ANSWER 2 OF 5
TI  INFLUENCE OF ALKALINE PARTICULATES ON THE CHEMISTRY OF FOG WATER AT
    DELHI, NORTH INDIA.
    Copyright (c) Engineering Information Inc.
AN  87(11):7593    COMPENDEX    DN  8711113877
CC  443; 451
CT  *METEOROLOGY: *Fog; AIR POLLUTION: India
ST  FOG WATER ANALYSIS; PH VALUES; IONIC CONCENTRATIONS; SOIL
    PARTICULATES
ET  Na; Na+; Na ip 1; ip 1; K; K+; K ip 1; O*S; SO4; SO4-2; S cp; cp; O cp;
    N*O; NO3; NO3-; N cp; NO3 in 1; in 1; SO4 in 2; in 2
```

in the presence of some positive ionic species would be searched and how this information would appear in the ET field.

Nuclear reactions are indexed as completely as possible as demonstrated in Example 6. Target isotopes, particles in-

EXAMPLE 6

```
FILE 'PHYS' ENTERED AT 11:27:34 ON 18 OCT 88
COPYRIGHT 1988 (c) FACHINFORMATIONSZENTRUM  KARLSRUHE

FILE LAST UPDATED: 14 OCT 88      <881014/UP>

=> E 142CE/ET
E1          1 142-XGD F/ET
E2         23 142BA/ET
E3         34 142CE/ET
E4          4 142CE F/ET
E5          6 142CE T/ET
E6          1 142CE(3HE,ALPHA)141CE/ET
E7          1 142CE(D,P)143CE/ET
E8          1 142CE(D,T)/ET
E9          1 142CE(P VECTOR,GAMMA0)/ET
E10         1 142CE(P,GAMMA)/ET
E11         1 142CE(P,GAMMA)143PR/ET
E12         1 142CE(P,GAMMAGAMMA)/ET

=> S E11
L2          1 "142CE(P,GAMMA)143PR"/ET

=> DIS L2 TI, IND

L2  ANSWER 1 OF 1
AN  80(23):94375  PHYS
TI  Proton capture by 176Yb in the giant dipole resonance region.
AU  Palsson, B.; Krumlinde, J.; Bergqvist, I. (Lund Univ. (Sweden));
    Nilsson, L.; Lindholm, A. (Tandemacceleratorlaboratoriet, Uppsala
    (Sweden)); Santry, D.C.; Earle, E.D. (Atomic Energy of Canada Ltd.,
    Chalk River, Ontario. Chalk River Nuclear Labs.)
SO  Nucl. Phys., A. (Aug 1980) v. 345(1) p. 221-231
    ISSN 0375-9474; CODEN NUPAB
CY  NETHERLANDS
DT  Journal
LA  English
AB  The proton capture cross section for the reaction 176Yb(p,gamma)177Lu
    has been measured for incident proton energies between 6 and 24 MeV.
    The excitation function for this deformed nucleus agrees remarkably
    well with the results of previous studies on spherical nuclei, e.g.
    142Ce(p,gamma)143Pr. The results indicate that the giant dipole
    resonance (GDR) is strongly excited as predicted by the
    direct-semidirect (DSD) model. It is found that the model describes
    reasonably well the excitation functions. In the low-energy proton
    range, where the excitation function increases rapidly with proton
    energy, the observed cross section is significantly higher than the
    DSD predictions. The difference can only partly be explained by
    compound nucleus contributions. In the high-energy end, the predicted
    cross section tends to be too high primarily due to an increasing
    contribution of direct capture to orbitals with large angular
    moementa. (orig.)
CC  *2540
CT  *PROTON REACTIONS; *CAPTURE; PHOTONS; *YTTERBIUM 176 TARGET; LUTETIUM
    177; MEV RANGE 01-10; MEV RANGE 10-100; EXCITATION FUNCTIONS;
    DEFORMED NUCLEI; GIANT RESONANCE; E1-TRANSITIONS; COMPOUND-NUCLEUS
    REACTIONS; INTEGRAL CROSS SECTIONS; ENERGY DEPENDENCE
ET  Yb; 176Yb; is; Yb is; Yb(p,gamma)Lu; 176Yb(p,gamma)177Lu; 176Yb t; p
    r; p gamma; 177Lu f; 177Lu; Lu is; Ce(p,gamma)Pr;
    142Ce(p,gamma)143Pr; 142Ce t; 143Pr f; 143Pr; Ce is; 142Ce; Pr is
```

volved (incoming and outgoing), and end nuclei are all identified and indexed. The algorithm recognizes the following conventions for written nuclear reactions: d or D for deuteron reactions, alpha for alpha particles, p for proton, n for neutron, e for electron, t or T for triton, pi or pi+ for pi or pi+ reactions, mu+ or mu− for mu reactions, gamma for photonuclear reactions, and ISOTOPE for isotope reactions. In the case where cerium-142 is converted to cerium-143 by deuteron bombardment with loss of a proton ($^{142}Ce(d,p)^{143}Ce$), the ET field will contain the following entries: 142Ce(d,p)143Ce, which is the way the reaction would appear in a title or abstract online; 142 Ce t, indicating that cerium-142 is the target nucleus; d p, indicating deuteron bombardment with outgoing proton; d r, to show it is a deuteron reaction; 143Ce f, to indicate that cerium-143 is the final nucleus; and Ce(d,p)Ce, indicating this is a deuteron reaction with outgoing proton involving cerium nuclei. The terms 143Ce and 142Ce entries are also included, since these isotopes are mentioned. Finally, Ce is and is show that the citation deals with cerium isotopes or isotopes in general.

Alloys, eutectic systems, and intermetallic compounds are all indexed as systems. Overall, the same information generated for compounds is produced here, except that the sy or system code is inserted rather than the cp compound code. A citation discussing a bimetallic of indium and bismuth, In–Bi, would have the following entries in its ET field: In–Bi, the

term as it appears in the title or abstract; Bi*In, the elements in Hill order; Bi sy 2, coding that indicates bismuth is part of a two-component system; In sy 2, to show that indium is also part of a two-component system; Bi and In, the elements themselves; and, finally, sy 2, to show that the citation contains a two-part system. More complex systems generate analogous indexing with more components. In the case where two or more metals appear in a single compound, the substance is indexed both as a compound and as a system to facilitate retrieval. An example of this "double indexing" appears for the ammonium tantalum tungsten compound NH_4TaWO_6. Each element in this material would appear in the ET field three times: once with the sy tag, once with the cp tag, and once as the element. Alloys in which the proportions of the metals are expressed as a chemical formula will be double-indexed, since the algorithm interprets the formula as a polymetallic compound. Example 7 shows how this double indexing appears.

EXAMPLE 7

```
FILE 'PHYS' ENTERED AT 11:44:38 ON 18 OCT 88
COPYRIGHT 1988 (c) FACHINFORMATIONSZENTRUM  KARLSRUHE

FILE LAST UPDATED: 14 OCT 88      <881014/UP>

=> E NH4TA/ET
E1            2 NH4SNF3/ET
E2            1 NH4SO3CF3/ET
E3            0 NH4TA/ET
E4            2 NH4TAWO6/ET
E5            2 NH4TCO4/ET
E6            1 NH4TCXRE1-XO4/ET
E7            1 NH4TEF5/ET
E8            1 NH4TEO3(OH)/ET
E9            1 NH4TI/ET
E10           1 NH4TI2NBO7/ET
E11           1 NH4U3F13/ET
E12           1 NH4UF6/ET

=> S E4
L10           2 NH4TAWO6/ET

=> DIS L10 TI, IND 2

L10 ANSWER 2 OF 2
TI  Ionic motion in the defect pyrochlore NH4TaWO6.
AN  79(7):24746  PHYS
CC  *6630
CT  SPIN-LATTICE RELAXATION; SPIN-SPIN RELAXATION; PROTONS; *IONIC
    CONDUCTIVITY; MINERALS; *AMMONIUM COMPOUNDS; TANTALUM TUNGSTATES
ET  H*N*O*Ta*W; H sy 5; sy 5; N sy 5; O sy 5; Ta sy 5; W sy 5; NH4TaWO6;
    N cp; cp; H cp; Ta cp; W cp; O cp; H*N; NH4; NH4+; NH4 ip 1; ip 1
```

Finally, this algorithm can recognize romanized Cyrillic abbreviations for steel alloys and translate them into the equivalent Western abbreviations. For instance, KhGNM becomes CrMnNiMo. Both sets of symbols appear in the bibliographic record, but only the Western symbols are placed in the ET field as shown in example 8.

EXAMPLE 8

```
FILE 'INSPEC' ENTERED AT 12:02:57 ON 18 OCT 88
COPYRIGHT 1988 (c) INSTITUTION OF ELECTRICAL ENGINEERS

FILE LAST UPDATED: 15 OCT 88      <881015/UP>

=> S CR*MN*MO*SI*V/ET AND RUSSIAN/LA
            7 CR*MN*MO*SI*V/ET
       154041 RUSSIAN/LA
L2          3 CR*MN*MO*SI*V/ET AND RUSSIAN/LA

=> DIS L2 TI, IND 1

L2  ANSWER 1 OF 3
TI  Formation of austenite during rapid continuous heating of KhG2S2MF
    structural steel.
AN  87:2829601  INSPEC     DN  A87033702
CC  4A8140G; 4A8130H; 4A6470K
CT  ALLOY STEEL; ANNEALING; SOLID-STATE PHASE TRANSFORMATIONS
ST  alpha to gamma transition; high temperature annealing; Cr-Mo-V steel;
    ferrite to austenite transition; carbide dispersion; dilatometric
    analysis; rapid continuous heating; structural steel; X-ray analysis;
    austenite formation
ET  Cr*Mn*Mo*Si*V; Cr sy 5; sy 5; Mn sy 5; Mo sy 5; Si sy 5; V sy 5;
    CrMn2Si2MoV; Cr cp; cp; Mn cp; Si cp; Mo cp; V cp; Cr*Mo*V; Cr sy 3;
    sy 3; Mo sy 3; V sy 3; Cr-Mo-V; Cr; Fe; Mo
CHI Cr ss, Fe ss, Mo ss, C ss, V ss
```

EXAMPLE SEARCHES

The following search examples demonstrate the utility of this additional indexing. In search A, information concerning the energy gap in gallium phosphide (GaP) is desired. Since the algorithm that selects entries for the ET field can tell the difference between upper and lower case letters, the formula for gallium phosphide, GaP, will be indexed in the ET field, while the word gap, as in "energy gap", will appear only in the basic index. In this way, it is possible to differentiate between GaP and gap and search for instances where both terms appear in the same citation.

SEARCH A

```
FILE 'INSPEC' ENTERED AT 12:42:27 ON 18 OCT 88
COPYRIGHT 1988 (c) INSTITUTION OF ELECTRICAL ENGINEERS

FILE LAST UPDATED: 15 OCT 88      <881015/UP>

=> S ENERGY (W) GAP
        392364 ENERGY
         41076 GAP
L1        9164 ENERGY(W)GAP     <----search for energy gap

=> S GAP/ET
L2        4115 GAP/ET           <----search for gallium phosphide

=> S L1 AND L2
L3         122 L1 AND L2        <----search for records containing both terms

=> DIS L3 TI, AB, ET 1

L3  ANSWER 1 OF 122
TI  Study on the properties of GaP:Te crystals.
AB  The electrical and optical properties of the undoped and Te-doped GaP
    crystal grown by the SSD (synthesis, solute diffusion) method have
    been investigated. Zn diffusion into the GaP and the
    electroluminescence properties were investigated. The etch pit
    density of the crystal was 103-104 cm-2. The resistivity, Hall
    mobility and carrier concentration of undoped and Te (0.1
    mole%)-doped GaP at room temperature were 4.94 Omega -cm, 0.07 Omega
    -cm; 187.29 cm2/V.sec, 152.01 cm2/V.sec; 6.75*1015 cm-3, 1.85*1017
    cm-3 respectively. The temperature dependence of the optical energy
    gap was dEg/dT=-5.2*10-4(eV/K), and the main radiative recombination
    observed in photoluminescence was the result of exciton and D-A pair
    emission. The activation energy for the Zn diffusion into the GaP was
    3.23 eV and the temperature dependence of the diffusion coefficient
    was 3.38*10-6 exp (-3.23/KBT). The main emission of
    electroluminescence was due to the edge, D-A pair and Zn-O complex
    recombination. These emission peaks were shifted to longer
    wavelengths as the temperature increased.
ET  Ga*P*Te; Ga sy 3; sy 3; P sy 3; Te sy 3; GaP:Te; Te doping; doped
    materials; Ga cp; cp; P cp; Te; Ga*P; GaP; Zn; D; K*T; KBT; K cp; B
    cp; T cp; O*Zn; Zn-O; Ga; P
```

In search B, any nuclear reactions involving molybdenum-92 as the starting nucleus, cerium-126 as the product nucleus, and some isotope of calcium as the incoming particle are sought.

SEARCH B

```
FILE 'INSPEC' ENTERED AT 14:41:08 ON 12 JAN 89
COPYRIGHT 1989 (c) INSTITUTION OF ELECTRICAL ENGINEERS

FILE LAST UPDATED: 23 DEC 88      <881223/UP>

=> S 92MO/ET
L1          358 92MO/ET  <--search for 92Mo

=> S 126CE/ET
L2           10 126CE/ET <--search for 126Ce

=> S CA IS/ET
L3         3662 CA IS/ET <-- search for calcium isotopes

=> S MO!CA,?/ET
L4            3 MO!CA,?/ET <--search for a fragment of the character string
                            which would comprise the nuclear reaction

=> S L1 AND L2 AND L3 AND L4
L5            2 L1 AND L2 AND L3 AND L4  <--search for all of these ideas
                                             in combination
=> DIS L5 ALL 1-2

L5  ANSWER 1 OF 2
AN  88:3175195  INSPEC     DN  A88090161
TI  Deformation of 126Ce.
AU  Moscrop, R.; Campbell, M.; Gelletly, W.; Geottig, L.; Lister, C.J.;
    Varley, B.J. (Dept. of Phys., Manchester Univ., UK); Price, H.G.
SO  Nucl. Phys. A (Netherlands) (9 May 1988) vol.A481, no.3; p. 559-76;
    37 refs.
    CODEN: NUPABL  ISSN: 0375-9474
    Price: CCCC 0375-9474/88/$03.50
DT  Journal
TC  Experimental
LA  English
AB  The gamma decays of excited states in the nucleus 126Ce have been
    identified in a study of the 92Mo(40Ca, alpha 2p)126Ce reaction at
    150-200 MeV bombarding energy using multiple particle-gamma
    coincidence techniques. Measurements were made of gamma-ray angular
    distributions and production cross sections are a function of energy.
    The lifetimes of the lowest states in the ground state band of 126Ce
    were measured using the recoil distance method. The first excited
    state was found to have a mere life of tau m=949+or-53 ps, which is
```

```
    consistent with a ground state quadrupole deformation of epsilon
    2=0.28+or-0.01 assuming axial symmetry and a uniform charge
    distribution. In the ground state rotational sequence a backbend is
    observed at h(cross)wc approximately=0.32 MeV. This appears to be due
    to the alignment of h11/2 protons. These properties are compared with
    the results of potential energy surface and cranked shell model
    calculations and are assessed in the context of the systematic
    behaviour of even-even nuclei in this mass region.
CC  5A2110F; 5A2760; 5A2320L; 5A2320C; 5A2320E; 5A2110R
CT  GAMMA-RAY ANGULAR DISTRIBUTION; NUCLEAR COLLECTIVE STATES AND GIANT
    RESONANCES; NUCLEAR ENERGY LEVEL LIFETIMES; NUCLEAR ENERGY LEVEL
    TRANSITIONS; NUCLEAR SHAPE; NUCLEI WITH MASS NUMBER 90 TO 149
ST  gamma decays; excited states; 126Ce; 92Mo(40Ca, alpha 2p)126Ce;
    multiple particle-gamma coincidence; angular distributions;
    production cross sections; lifetimes; ground state quadrupole
    deformation; axial symmetry; uniform charge distribution; ground
    state rotational sequence; backbend; potential energy surface;
    cranked shell model; even-even nuclei
ET  Ce; 126Ce; is; Ce is; Mo(Ca,alpha 2p)Ce; 92Mo(40Ca,alpha 2p)126Ce;
    92Mo t; 40Ca r; 40Ca alpha 2p; 126Ce f; Mo is; 92Mo; Ca is; 40Ca

L5  ANSWER 2 OF 2
AN  87:2966449  INSPEC     DN  A87109119
TI  Measurement of lifetimes in the light rare-earth region.
AU  Moscrop, R.; Campbell, M.; Gelletly, W.; Goettig, L.; Lister, C.J.;
    Varley, B.J. (Dept. of Phys., Manchester Univ., England); Price, H.G.
SO  Proceedings of the International Nuclear Physics Conference
    Editor(s): Durrell, J.L.; Irvine, J.M.; Morrison, G.C.
    Bristol, England: IOP 1987. p. 160 vol.1 of 2 vol. (vii+556+xvi+606)
    p.; 1 refs.
    Conference: (IUPAP), Harrogate, Yorks., England, 25-30 Aug. 1986
DT  Conference
TC  Experimental
LA  English
AB  Lifetime measurements for the lowest states of 13262Sm70 and 136Gd72
    have been performed using the recoil distance method. Measurements
    were made using a 180 MeV 40Ca beam on a 0.7+or-0.1 mg/cm2 92Mo
    target set at nineteen positions relative to a fixed gold stopper
    (corresponding to mean flight times of 20 ps to 3.2 ns) and on a
    0.9+or-0.1 mg/cm2 96Ru target set at eleven positions (corresponding
    to mean flight times of 78 ps to 3.2 ns). Lifetime results are
    presented and the deduced B(E2)'s discussed with regard to the
    variation of shape and deformation of the region.
CC  4A2320C; 4A2110F; 4A2760; 4A2320L; 4A2570G
CT  HEAVY ION-NUCLEUS REACTIONS; NUCLEAR ENERGY LEVEL LIFETIMES; NUCLEAR
    ENERGY LEVEL TRANSITIONS; NUCLEAR SHAPE; NUCLEI WITH MASS NUMBER 90
    TO 149
ST  nuclear level lifetimes; 132Sm; 136Gd; B(E2); 132Sm decay into
    126Ce+4He+2p; 132Sm decay into 128Ce+4p; 136Gd decay into 132Nd+4p;
    136Gd decay into 133Pm +3p; 92Mo(40Ca,2p alpha )126Ce;
    92Mo(40Ca,4p)128Ce; 96Ru(40Ca,4p)132Nd; 96Ru(40Ca,3p)133Pm; lowest
    states
ET  Gd72; 136Gd72; is; Gd is; 136Gd; Ca; 40Ca; Ca is; Mo; 92Mo; Mo is;
    Ru; 96Ru; Ru is; Sm; 132Sm; Sm is; Gd; Ce*He; Ce sy 2; sy 2; He sy 2;
    Ce is; 126Ce; He is; 4He; 126Ce+4He; Ce; 128Ce; Nd; 132Nd; Nd is; Pm;
    133Pm; Pm is; Mo(Ca,2p alpha )Ce; 92Mo(40Ca,2p alpha )126Ce; 92Mo t;
    40Ca r; 40Ca 2p alpha ; 126Ce f; Mo(Ca,4p)Ce; 92Mo(40Ca,4p)128Ce;
    40Ca 4p; 128Ce f; Ru(Ca,4p)Nd; 96Ru(40Ca,4p)132Nd; 96Ru t; 132Nd f;
    Ru(Ca,3p)Pm; 96Ru(40Ca,3p)133Pm; 40Ca 3p; 133Pm f
```

This reaction would be written $^{92}Mo(^{x}Ca,...)^{126}Ce$. In this formula x is the isotope number for the calcium nucleus, and the ellipse contains other specifications about the reaction such as leaving particles. To conduct this search on a version of INSPEC that contains the ET field, a successful search strategy would involve searching that field for the specific isotopes ^{92}Mo and ^{126}Ce, any isotopes of calcium, and as much of the reaction formula as possible.

To demonstrate the precision of retrieval the ET field provides, this search was executed in the INSPEC file covering the years 1969 to present as it appears on STN International.

INSPEC now adds a chemical information field (CI or CHI) to their records, which contains the element symbols mentioned in a citation. Each online host has chosen to implement this field differently. For clarity, this field has not been considered for this search example.

CONCLUSION

The element terms field is an access point to chemical information contained in technical databases that do not index their contents from a chemical point of view. The ET field enables users to retrieve more precise answers by enhancing access to chemical information contained in nonchemical databases by making elements, compounds, and reaction notations searchable. This precision is especially useful if the symbol for the element or compound being sought is synonymous with another abbreviation or a real word. Special symbols, especially grouping symbols such as parentheses, are retained in the ET field; in the basic index, these symbols would be replaced with blank characters as if they were a form of punctuation. Reaction or compound expressions containing these symbols can easily be searched by masking the symbols with quotes or single-character truncation. The ET field is available exclusively on STN International and appears in these files: COMPENDEX, ENERGIE, ENERGY, INIS, INSPEC, LPHYS, MATBUS, MEET, METADEX, and PHYS.

Graphics and Natural Language Interface for a Cybernetic Analytical Instrument

HARI GUNASINGHAM* and MUN-LEONG WONG

Department of Chemistry, National University of Singapore, Kent Ridge, Singapore 0511

Received January 12, 1989

The design and implementation of a graphics and natural language interface for a cybernetic analytical instrument are described. A novel feature of the design is the way the symbolic representation of active elements of the experiment is integrated with their functional and procedural characteristics by using a Prolog-based frame formalism. The graphics and natural language interface obviates the need for any programming knowledge on the part of the user. Further, it enables the user to dynamically interact with the experiment in real time.

INTRODUCTION

Recently there has been increasing interest in the concept of a cybernetic analytical instrument that enables the intelligent control of experiments. One of the key characteristics of such instruments that distinguishes them from conventional computer-based instruments is the use of heuristic knowledge to make decisions based on uncertain experimental data.[1]

In an earlier paper[2] we described a design for a cybernetic analytical instrument conceived as a knowledge-based system in closed-loop, real-time interaction with the instrument hardware. The knowledge base is nondeterministic and contains only the logic of the process. In our design, a frame-based formalism was employed. The actual control is directed by the inference engine, which selects and fires the relevant frames at the appropriate time. A key feature of the design is that the inference mechanism, while being deterministic at run time, can be altered by the user to suit the application at hand.

The knowledge-based system was implemented in Prolog. This language has several advantages, arising from its origins in logic, that have no direct analogues in other languages.[3,4] These include declarative programming style, autonomous control of search, independence of logic and control, built-in pattern matching, and modularization.

In our work the version of Prolog used was Turbo Prolog (Borland International). A Turbo Prolog program consists of four essential sections: domains, predicates, database, and

0095-2338/89/1629-0283$01.50/0