Analytica Chimica Acta, 265 (1992) 305–312
Elsevier Science Publishers B V , Amsterdam

The Integrated Gmelin Information System

New developments in information processing

A. Nebel, U Tolle, R Maass, G Olbrich and R Deplanque
*Gmelin Institut für Anorganische Chemie der Max-Planck-Gesellschaft, Varrentrappstrasse 40–42,
W-6000 Frankfurt am Main 90 (Germany)*

P Lister
Juniper Systems, London (UK)

(Received 22nd November 1991)

Abstract

The Gmelin Information System consists of the Gmelin Handbook of Inorganic and Organometallic Chemistry and the Gmelin Factual Database This information system is constructed such that the factual database is used as the primary source for the handbook production process The development of some novel software tools is described These software tools enable the handbook authors to examine the factual data from the database and to add text, in a further step these parts are processed to form a draft manuscript The completed handbook manuscripts are transferred to the printer using a newly implemented SGML procedure These developments enable the direct creation of a printed handbook volume from the Gmelin Factual Database

Keywords Gmelin Factual Database, Information systems

DESCRIPTION OF STARTING POINT

Gmelin handbook production

The Gmelin Handbook of Inorganic and Organometallic Chemistry is the only comprehensive collection of published factual information in the areas mentioned The handbook authors, scientists who are experts in their respective fields, provide an evaluated overview of the published research results The Gmelin Handbook is now in its 8th edition, it covers the literature from 1924, and comprises more than 600 volumes

During the last 20 years the number of publications in inorganic, organometallic and physical chemistry has increased exponentially To keep abreast with this increase, new methods of handbook production and new media for distributing the collected and the evaluated factual information have to be introduced The first step to build up a system for computer aided handbook production was a detailed system analysis to find out the essential steps which the authors are following from the original literature as primary information to the complete handbook The following list gives an overview

(1) The information in the Gmelin Handbook is compound-oriented, the definition of a handbook project consists of a specific group of inorganic or organometallic compounds These can be thousand different compounds in the field of organometallic chemistry or just one compound like Si_3N_4 in inorganic chemistry The number of

Correspondence to A. Nebel, Gmelin Inst für Anorganische Chemie, Varrentrappstrasse 40/42, D-6000 Frankfurt am Main 90 (Germany)

compounds in one defined project depends in most cases on the number of published facts The different elements in the periodic system are described one after the other following the Gmelin system After collecting the relevant literature the handbook author has to read the publications very carefully During this study he extracts information about the compounds and their factual data and stores these in an electronic card index

(2) After completion of this work an author has to sort the extracted data. Essentially this is a step of inverting the data from a citation–compound–facts hierarchy to a compound–facts–citation orientated one In the handbook the compound is the key entry to the factual information

(3) After the sorting step, an author selects parts of the excerpted data to create the handbook manuscripts Typically he describes at first the methods of preparation for a given compound The order of factual information in the Gmelin Handbook follows a convention

(4) During the creation of the handbook manuscript the most important step is the comparison of different publications about the same fact and the same compound In consequence, the handbook presents factual information which is completely reviewed and evaluated, based on all publications dealing with a specific compound

(5) After different steps of reviewing, the editor collects all the partial manuscripts to the complete manuscript of the volume which is then transferred to the printer

Before 1989, all the steps described were done by conventional methods without electronic data processing with the exception of one author group working in the field of organometallic chemistry who utilized text processing on PCs

The Gmelin Factual Database

In December 1991, after five years of development the Gmelin Institute presented the Gmelin Factual Data Base to the scientific public In this project an effort was made to store factual data from research in the fields of inorganic, organometallic and physical chemistry Thus the user now has online access to factual and numerical data in a systematic and compound-related way The Gmelin Database represents the largest electronic storage of information in the areas mentioned The following list gives an overview for the different facts covered by the data acquisition and stored in the database formation and synthesis, chemical reaction, electric properties, electrochemical reaction, condensed phase, citation, magnetic properties, mechanical properties, molecular properties, multi component systems, optical properties, phase transition, spectroscopic properties, thermal properties, thermodynamic data, transport phenomena, structure, substance identification

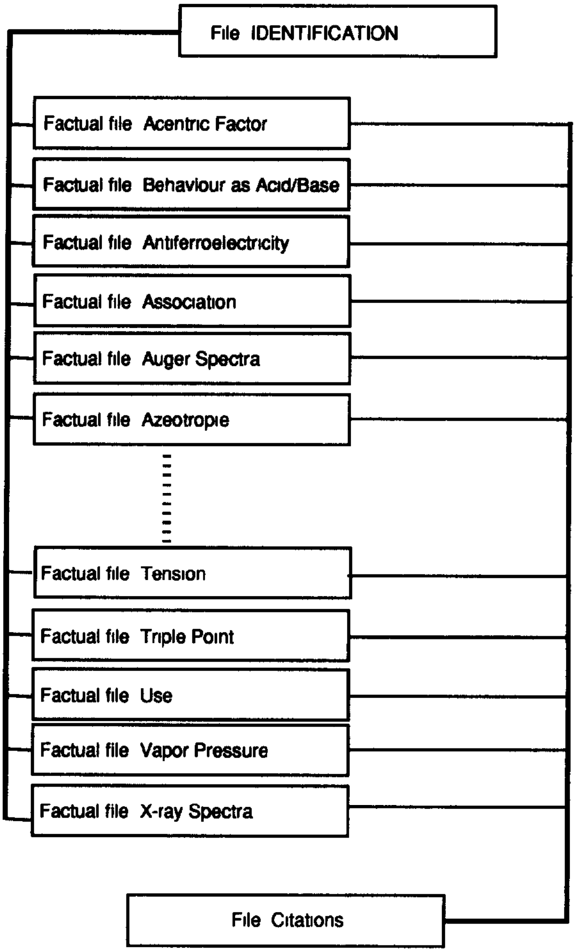

File IDENTIFICATION
Factual file Acentric Factor
Factual file Behaviour as Acid/Base
Factual file Antiferroelectricity
Factual file Association
Factual file Auger Spectra
Factual file Azeotropie
Factual file Tension
Factual file Triple Point
Factual file Use
Factual file Vapor Pressure
Factual file X-ray Spectra
File Citations

Fig 1 Hierarchical structure of the Gmelin Factual Database

Figure 1 shows the hierarchical scheme of the database and the connections between the different levels of information This scheme is important for the next sections for describing the integration of both media, the handbook and the database

THE CONCEPT OF THE INTEGRATED GMELIN INFORMATION SYSTEM

The Gmelin Institute offers two different media for the distribution of factual information to the public Each of these is characterized by its own specific attributes The handbook presents complete information, reviewed by scientists In many fields of inorganic chemistry descriptive text is necessary for this task On the other hand, the database enables the online retrieval of factual and numerical data for inorganic and organometallic compounds

The following points are important for the integration of handbook and database into one information system

(1) The Gmelin Factual Database is the primary source of data in the handbook production
The Gmelin Factual Database will be used as the main source of primary data for the handbook production In those cases where the content of the database is not complete, additional literature will be added, when available

(2) Additional evaluation of database contents
The Gmelin Factual Database as a source of primary data enables evaluation of the data in addition to that done by the quality control before data processing A handbook author at the beginning of a project will be provided with a complete information package (compounds, facts, and citations) instead of the isolated citation information in the conventional procedure During the handbook production an author reads the original citations very carefully and then compares the contents in the database with the published data As a consequence, the online version of the Gmelin Factual Database presents factual and numeric data on a high quality level

(3) Additional acquisition of factual data during handbook production
The study of the original literature by the authors is the source of additional data which will be stored in the database For example, line positions of IR spectra which are not collected during the primary data acquisition for the Gmelin Factual Database can be stored during the creation of a related handbook article

(4) Additional acquisition of bibliographic data during handbook production
An author who is excerpting original literature will find additional citations not present in the database With the aid of a new component for ordering the literature in the Integrated Information System the data acquisition from these articles will be initiated

(5) Handbook manuscripts stored in the database
Handbook manuscripts created in the system can be stored together with the database for future online retrieval

THE INTEGRATED GMELIN INFORMATION SYSTEM

The Integrated Gmelin Information System is a repository for scientific information in the fields of inorganic, organometallic and physical chemistry It enables the presentation of the stored data in different forms and media

- The Gmelin Factual Database as an online version,
- The Gmelin Handbook Volumes of the 8th edition,
- The Gmelin Factual Database as an inhouse version,
- Specific selected parts of the Gmelin Factual Database as inhouse versions

The different components and their role in the Integrated Gmelin Information System are shown in Fig 2

The Production Database GFDB (Gmelin Factual Database)

This reference database was implemented on an IBM mainframe of the Gmelin-Beilstein-Computer Center The data processing and the registration of the different chemical compounds is also done on the mainframe Every data manipulation for corrections etc is stored in the refer-

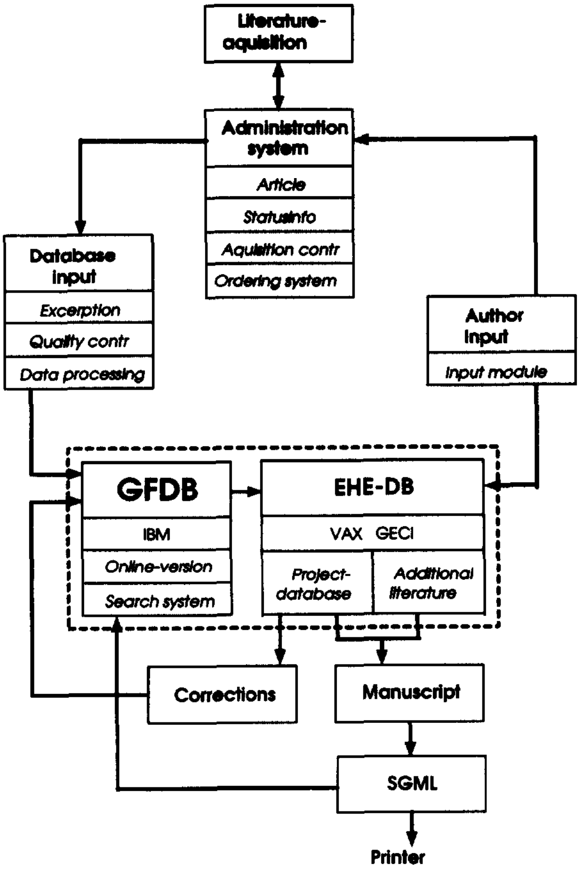


Fig 2 The Integrated Gmelin Information System

ence database A special download procedure transfers the data to the handbook system

The EHE-Database GHDB (Gmelin Handbook Database)

The Gmelin Handbook Database was implemented on the central database server in the handbook department (VAX6000-320) The file structure consists of 141 files which conforms to the number of facts in the database This file structure was chosen for flexibility in adding new facts (as a group of fields) or elementary fields to the database The database management system for the GFDB and for the GHDB is ADABAS (Software AG) The data transfer was realized via network connection A copy of those parts of the GFDB which are relevant to a handbook project are stored in the Gmelin Handbook Database

The administration system

The administration system consists of a database where the citation data of publications on a single article level will be stored Additionally, information on the actual status of an article will be incorporated This database enables controlling all of the data acquisition for the Gmelin Factual Database Especially the acquisition from additional literature ordered by handbook authors will be realized by the administration system This additional acquisition will cover noncontiguous time intervals The following list shows the main tasks of the administration system

- Storage for every citation in the whole integrated system,
- Storage of status information for every single article,
- Control of the whole data acquisition process for the GFDB,
- Ordering service for the library,
- Integration with the reference list controlling database

Database input

The database input follows three different steps

(1) Excerption of articles from a selected catalogue of 120 Gmelin relevant periodicals The excerption is done with a data acquisition program running on PCs

(2) The data are then transferred on floppy disks to the Gmelin Control Department Before data processing and registration, the excerpted data are checked partly or in some cases completely to ensure high scientific quality

(3) During data processing the registration of the different compounds based on their 3-D structure is the most important step of the database input The registration process is described in an accompanying paper

The EHE system for computer-aided handbook production

The authors will use two main software applications The Gmelin Electronic Card Index forms the user interface to the Gmelin Handbook Database, as the text processing system DECwrite

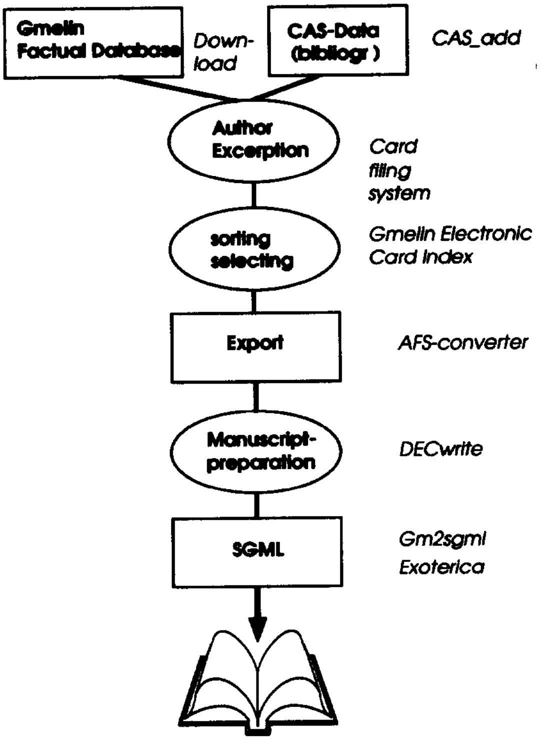


Fig 3 The dataflow from the database to the handbook volume

is used The Gmelin Electronic Card Index is described in detail in the last section of this paper The data flow from the production database to the complete handbook volume is described in the next section

THE DATA FLOW FROM THE PROJECT DATA BASE TO THE HANDBOOK VOLUME

The details of the data flow in the handbook production process are shown in Fig 3 Within the card index, the author will get the citation information from CAS or from the Gmelin Handbook Database The authors will use the citation data to order the original literature During the study, excerpted data can be stored in special notes fields which are logically connected to one specific citation and a compound After selecting and reordering, which can easily be done in the computer aided system with the card filing system, an export function is started Within this process a completely formatted draft document is created It is possible to create automatically documents consisting of paragraphs, tables and reference lists The export procedure uses the AFS-format (DEC) which is an ASCII format with tags for formatting information A subsequent program indentifies linearized formulas of compounds and adds the markups for sub- and superscript which cannot be stored in database The export files in AFS format are then transformed into a draft manuscript in DDIF format from which the authors prepare final handbook manuscripts using the document system DECwrite

DECwrite is a WYSIWYG system which uses the style file constructed for storing layout and formatting information for document types After a detailed analysis of the typesetting characteristics in the Gmelin Handbook a Gmelin style file was created, use of this style file leads to uniformly formatted manuscripts For the data transfer to the printer, the ISO standard SGML was chosen This necessitated the development of a converter from DDIF to SGML and, furthermore, the construction of a Document Type Definition for the Gmelin handbook The Document Type Definition according to the SGML standard contains the description of the structure of a given document type For further processing of SGML documents a commercial software package from Exoterica Co was used, especially for parsing a converted document The complete flow of data from the database to the printed volume as described is now in production and a first handbook volume was created as a result of the new computer aided handbook production technique

During handbook production an author can concentrate on the intellectual parts of his work These are study and excerption of data relevant for a Gmelin handbook project, reordering, selection and sorting of the contents of the card index, comparison and evaluation of published data and creation of a handbook manuscript

The next section describes the development of the Gmelin Electronic Card Index This software system is the card filing system which presents specific parts of the Gmelin Factual Database as a primary source of information to the handbook authors It is the key software for connecting database and handbook

THE GMELIN ELECTRONIC CARD INDEX

Based on the hypertext system CAMS4 under MS-DOS, the first step of the development was a portation to VAX/VMS Results from a detailed system analysis [1–3] in the Gmelin handbook department showed the necessity of presenting the contents of the Gmelin Database to the handbook department in an individual format with different functions for sorting and selection An author needs only a small portion of the complete catalogue of data fields for writing a particular article or section in the Gmelin Handbook During conventional handbook production described in the first section, there is a complete inversion of the hierarchy of data structure from the primary information (conventional) as citation–compound–facts, to the handbook as compound–fact–citation The organization of the data structure in the Gmelin Database is similar to the handbook By using the Gmelin Database as a source of primary information, the author does not need to invert the data before writing the handbook manuscripts

For reasons described in the previous paragraphs a flexible template generator was developed With the aid of a template, an author is able to create his own individual "view" of the related database The main feature of this template is its function as a "filter" for an ADABAS database

The program modules of the Gmelin Electronic Card Index

To be compatible with the other application programs in the handbook department the implementation of the Gmelin Electronic Card Index was realized on the graphical user interface DECwindows (Digital Equipment Corporation) The software package for direct database access includes three modules The communication to the database management system ADABAS (software AG) was obtained by means of a special data dictionary The program "FILEPREP" allows creation of this data dictionary, which can be accessed by the other modules It is based on the "field definition table" of ADABAS The main function of this module is to add descriptive long names to the structure of the underlying database structure

The module "PAINT" is the template generator The designer of the template is able to place descriptive text and field cells on a window which can be compared with a working area Within a selection box, the fields which are available in the connected database file can be selected After a selection process involving moving through the list with up and down arrows, the field is placed at the cursor position The location of the field cell on the template can be changed by mouse pointer dragging The display format of a field

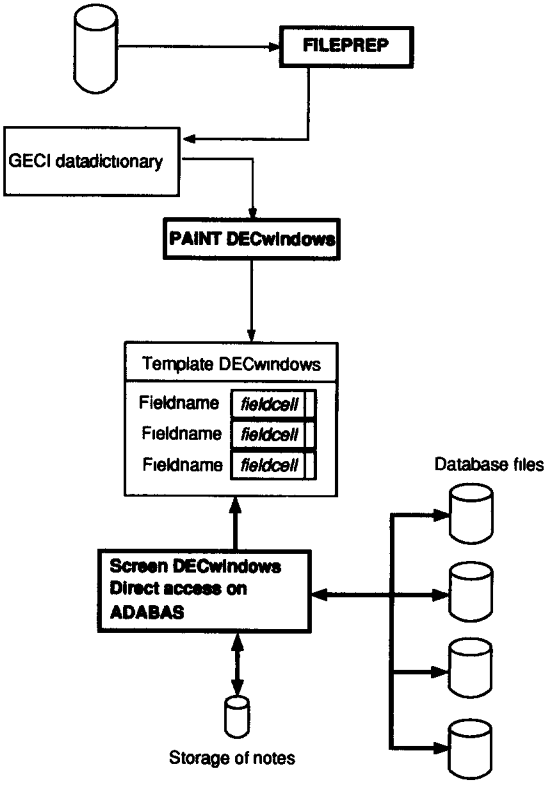
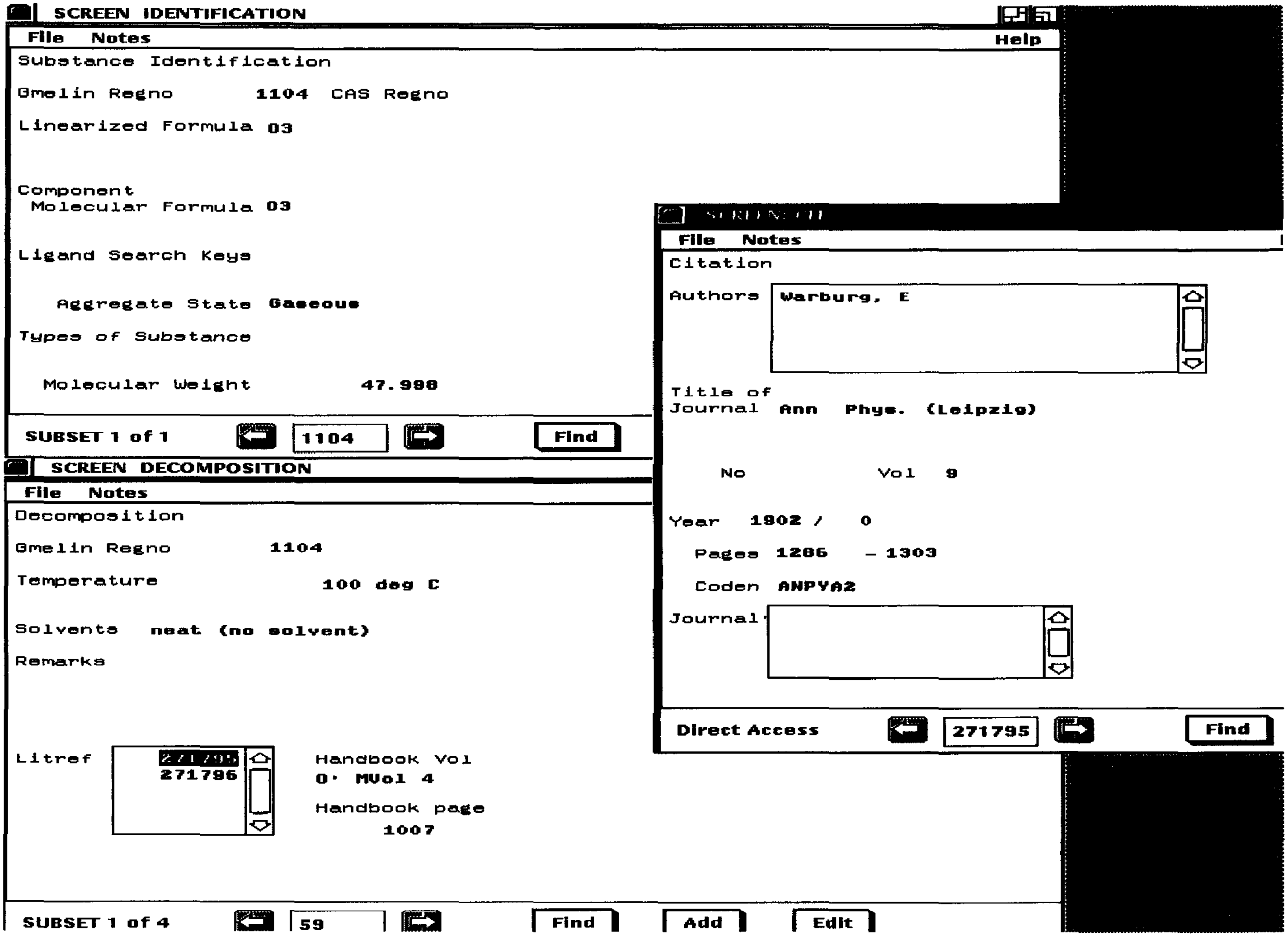


Fig 4 The programs of the Gmelin Electronic Card Index


SCREEN IDENTIFICATION

Substance Identification

Gmelin Regno 1104 CAS Regno

Linearized Formula O3

Component
Molecular Formula O3

Ligand Search Keys

Aggregate State Gaseous

Types of Substance

Molecular Weight 47.998

SUBSET 1 of 1 1104

SCREEN DECOMPOSITION

Decomposition

Gmelin Regno 1104

Temperature 100 deg C

Solvents neat (no solvent)

Remarks

Litref 271795
271796

Handbook Vol
O: MVol 4

Handbook page
1007

SUBSET 1 of 4 59

[illegible]

Citation

Authors Warburg, E

Title of
Journal Ann Phys. (Leipzig)

No Vol 9

Year 1902 / 0

Pages 1286 – 1303

Coden ANPYA2

Journal

Direct Access 271795

Fig 5 Templates for direct access to three different ADABAS files

may be edited All the different ADABAS field types can be displayed, including multiple fields and periodic groups A template can be stored under a specific name after creation The third module, "SCREEN", enables the user to directly access the database using the template Figure 4 shows the program system

Module "SCREEN" has several different functions for working in the database In addition to the functions for scrolling through the database sequentially, the user is able to build complex search criteria for working on a subset of records These functions are the result of a detailed system analysis [1–3] in the handbook department There is a text window within the screen module, which can be opened by a pull-down menu Text that is entered here can be stored in a separate text file The connection between the actual record and the text block is accomplished by means of a combination of the citation number and the Gmelin registry number The textblocks stored here form the basis of a draft manuscript which is created by the previously described export function Figure 5 shows templates for direct access to the database files "Substance Identification", "Decomposition" and "Citation"

Fields from different database files can be placed on the templates created by "PAINT" In consequence, an author will get a collection of facts (collection of fields from different database files) to work on In most of the cases a template will contain data for identifying the compounds (linearized formula, aggregate state etc.), factual data and fields for the complete citation information Together with these templates, containing fields from more than one database file, a correlation file is stored holding the information of which records must be shown together when the template is used for access to the Gmelin Factual Database

The Gmelin Institut acknowledges financial support given by the German Ministry of Research and Development (BMFT) for this work

REFERENCES

1 EDV-unterstutzte Erstellung des Gmelin-Handbuchs, Pflichtenheft, Batelle-Institut, Frankfurt am Main, 1986

2 Gmelin-Handbucherstellung (EHE Gmelin), Planungsunterlagen, Version 2 0, Softron GmbH, Munchen, 1987

3 EHE-Gmelin, Elektronischer Karteikasten, Systemanalyse, Gmelin-Institut für anorganische Chemie, Frankfurt, 1988