

Organic Reaction Database Translation from REACCS¹ to ORAC²

Todd M. Miller, Jan-Willem Boiten, Martin A. Ott, and Jan H. Noordik*

CAOS/CAMM Center, University of Nijmegen, Toernooiveld, 6525 ED Nijmegen, The Netherlands

Received August 25, 1993*

An account of the process of transferring a large number of organic reactions from one reaction retrieval system (REACCS) to another (ORAC) is given. These systems use different external file formats. The translation between those formats is treated in detail for both structural and textual data. The basic attributes of an organic reaction database program are described as they are encountered within the translation process.

INTRODUCTION

The ability to exchange data between different organic reaction databases,³⁻⁵ such as REACCS,¹ ORAC,² and SYNLIB,⁶ allows one flexibility in dealing with structurally oriented chemistry programs. Differences among these database programs, which might not be evident during normal use,^{7,8} are highlighted by translations. The translation process has to overcome these differences to provide a standardization across the databases of organic reactions, as well as a platform for further development of techniques⁹ which may be applied to these reaction databases. From a user's point of view, these translations allow all reactions to be placed in a single reaction database system and thus make them all accessible with minimal user actions. Furthermore, the knowledge of a translation process provides us with the basic information to construct database entries for either system involved in the translation.

The content and quality of the organic reactions databases for both systems involved in the translation described in this paper (REACCS¹⁰ and ORAC) is substantial. A Theilheimer¹¹ database of about 42 000 reactions, covering mainly the older (up to 1980) literature, exists for both REACCS and ORAC. ORAC has 12 additional boxes of 5000 reactions each (60 000 reactions) covering mainly the literature of the period 1980-1991, and REACCS covers a similar period (1981-1991) in its 36 601 reactions CLF (Current Literature File). The most recent literature is accessible through both systems by the CSM (Current Synthetic Methodology) or ChemInform RX-databases,¹² both based on the (printed) abstracting service of ChemInform.¹³ Apart from these general databases, both systems offer additional databases dedicated to special branches of chemistry (e.g., heterocyclic and organometallic) or based on well-known printed issues (e.g., *Organic Syntheses*).

The translation described in this paper concerns the translation of the REACCS CLF into ORAC format; of course, a program that is able to translate CLF should be able to translate the other databases as well. The actual translation takes place between two ASCII reaction file formats: RDfile (Reaction-Data file)¹⁴ and SMD¹⁵ (Standard Molecular Data) v4.3.¹⁶ These files constitute external file formats for REACCS and ORAC, respectively; neither program writes the other program's format. The process of transferring organic reactions from one database to another is a three-step process. The first step is the write out of the entire CLF database in RDfile format, then this file is translated into SMD format by a newly developed program (*Reactrans*)

described in this paper, and finally the SMD file is imported into ORAC.

OVERVIEW OF THE DATABASE STRUCTURES

REACCS. The unique property that distinguishes one REACCS database entry from another is the structural information of the reaction. If the products and the reactants are identical, but data such as reaction conditions, reagents, or literature references are different, REACCS stores each procedure as a separate VARIATION of the same database entry. These variations are the first level of a hierarchical structure of the textual data, which is shown for the REACCS CLF database in Figure 1.

Below the variation level, a top level of datafields has been defined, several of which are parent datafields for datafields of the second level. An example is LITREF, which is the parent datafield of AUTHOR and JOURNAL, the latter being the parent for a number of datafields of the third level. Many of the datafields at the top level may occur more than once. Lower level datafields cannot have multiple occurrences within a single top-level datafield, but only implicitly by multiple occurrences of the parent datafield. As an example, the full datafield name of YEAR is RXN(k):VARIATION-(l):LITREF(m):JOURNAL:YEAR for the *m*th reference of the *l*th variation of the *k*th reaction in the database.

REACCS offers many ways to retrieve and display data¹⁷ requiring powerful graphics drivers. Textual data are displayed within boxes using resizable fonts. The location and size of these boxes is user-definable inside an interactive menu (forms definition mode). It is possible to produce a scrolling list of textual data while inside REACCS.

ORAC. The ORAC databases have a sequential structure unlike the hierarchical structure of the REACCS databases. The textual data are stored in datafields which can contain only one data item and datatables which consist of arrays of datafields of the same type. For example, the author information is stored in a datatable AUTHORS which consists of six text datafields AUTHOR which can hold one author name each. The size of the datatable, i.e. the number of datafields that it may hold, is predefined upon database definition. For some datafields (e.g., JOURNALS, SOLVENT-KEYS), a thesaurus has been defined, which is a collection of all permitted values for the datafield. For instance, the SOLVENT-KEYS datafield can contain the value "Benzene", because this is a keyword defined in the SOLVENT thesaurus. Searches for "C6H6" and for "Benzene" give the same results, since these two values are defined as synonyms in the SOLVENT thesaurus. The thesauri are user-accessible through menu screens, displaying alphabetically all existing keywords with their synonyms.

* Abstract published in *Advance ACS Abstracts*, March 1, 1994.

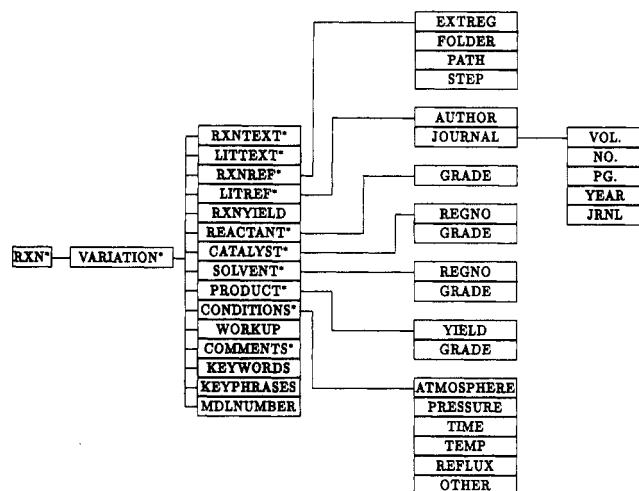


Figure 1. REACCS CLF database hierarchy. The asterisks mark items that may occur more than once.

ORAC displays its data only through menu screens. There is no method in ORAC to display textual data from the database without previously defining the appropriate menu from outside ORAC.

STRUCTURAL DATA TRANSLATION

Structural Data Formats. A reaction entry in a REACCS RDfile starts with an internal registry number and the number of reactants and products, followed by a number of MOLfile¹⁴ blocks, one for each reactant or product molecule. Each MOLfile contains the following information:

atom information	bond information
no. of atoms	no. of bonds
atom type (atomic symbol)	bond type (bond order)
formal charge	attached atoms
radical status	bond stereo
atom stereo parity	bond change
isotope	
2-D atom coordinates	
atom mapping index no.	

A reaction entry in an SMD file is logically organized, with separate blocks for each item of information. The reactant and product molecules are put together into one reactant and one product connectivity table, each constituting a so-called CT block. The atom-to-atom mapping is implicit in the sequence numbers of the atoms in the CT blocks. Atoms that are lost or added in the reaction, and thus have no counterpart on the opposite side, are mapped onto a dummy atom on the opposite side. Separate blocks are provided for atom coordinates and reaction centers. Together, the CT, atom coordinate and reaction center blocks contain the following information:

atom information	bond information
no. of atoms	no. of bonds
atom type (atomic symbol)	bond type (bond order)
formal charge	attached atoms
radical status	bond stereo
atom stereo parity	
2-D atom coordinates	
reaction center status	

Molecular Structures. The reactant and product connectivity tables (CT blocks) to be supplied to ORAC may contain

more than one structure. However, ORAC does not accept more than three reactant molecules or more than two product molecules. Our translation program (Reactrans) discards reaction entries violating this limit. Some structures (e.g., alkali metal salts) that are treated by REACCS as one structure are treated by ORAC as consisting of two molecules and cause additional entries to be refused by ORAC during import. Reactrans also discards entries in which the total number of atoms or bonds in a reactant or product connectivity table exceeds the ORAC limit of 128.

In a number of instances, REACCS uses the atomic symbol "X" to denote attachment to a polymer support, something ORAC has no representation for. Reactrans converts these atoms to carbon atoms for simplicity. Although the RDfile format allows special bond types such as aromatic and generic, these do not occur in the CLF database, so no provisions had to be made for them. Stereochemical features are also easily translated. REACCS and ORAC use the same representation of stereochemistry: both the parity number for stereocenters and the bond code for stereo bonds adhere to the same standard, so these values are simply copied. Absolute stereochemistry is supported by both REACCS and ORAC, but not by the SMD format used by ORAC. Absolute stereochemistry in the RDfile, indicated by the chiral flag, was converted to relative by ignoring this flag. In both REACCS and ORAC, the stereochemistry of the double bonds is usually derived from the 2-D coordinates. Therefore most double bond stereochemistry is easily translated. Alternatively, double bonds can be specifically marked as cis or trans, but the number of occurrences of this representation in the CLF database is so small (eight) that we are not sure this is intentional. Because there is no explicit representation for cis or trans double bonds in the SMD file, these bonds were marked as "either". Since ORAC does not use isotope information, Reactrans ignores this information in the RDfile.

We found that REACCS misuses the notion of radical to represent π -coordination (e.g. cyclopentadienyl ligands were given radical marks on all five carbon atoms). No attempt was made to identify or correct such structures. This was not useful because the SMD format used by ORAC does not have a representation for π -coordination either.

Atom Mapping and Reaction Centers. The atom-to-atom mapping can be performed automatically by ORAC during import of a reaction or can be translated from the information in the RDfile. Because ORAC's automatic mapping procedure does not always give correct results and is rather CPU time-consuming, we chose to translate the mapping information. This translation proved to be computationally straightforward. In the RDfile it is possible to map the reactant(s) on more than one product (e.g., the product of a side reaction). This is not possible in SMD version 4.3. To avoid the reactant(s) being mapped onto a minor product, the product with the highest yield is chosen (in cases where there are no product yields, the first product is taken). The mapping is translated by changing the atom sequence numbers, using REACCS' mapping index numbers,¹⁸ for the new atom sequence numbers in the reactant and product connectivity tables. For each unmapped atom, a dummy atom is added on the opposite side. This addition may cause the total number of atoms in a reactant or product connectivity table to exceed 128, in which case the reaction entry has to be discarded.

Reaction center atoms are transmitted to ORAC by way of a reaction center block in the SMD file. The RDfile, however, contains only bond changes.¹⁹ Reaction center atoms are determined in the following way: an atom is a reaction

Table 1. Frequencies of Multiple Occurrences for Some Important Datafields in REACCS CLF

datafield	no. of occurrences within a single variation												
	1	2	3	4	5	6	7	8	9	10	11	12	>12
COMMENTS	13 249	3184	1014	321	167	75	43	19	12	8	7	4	3
RXNTEXT	6 404	7742	6595	5877	3114	2504	283	224	78	79	31	14	28
SOLVENTS	19 327	4353	745	143	40	4	1	1					
CATALYST	16 481	8087	2662	872	299	133	58	30	15	11	4	3	6
JOURNAL	31 952	4030	956	203	79	22	9	12	19	0	6	3	1

center atom when it is on a bond that has changed or when there is a change in charge, radical status, or stereochemistry.

Reactions without reaction center atoms cannot be imported into ORAC and are therefore discarded. This may happen when the major "product" is unchanged starting material.

TEXTUAL DATA

General Approach. The transfer of textual data from REACCS to ORAC is hampered by the difference in data structure between the two databases. The hierarchical structure of REACCS would explode in numerous datafields if an attempt were made to translate it directly into the sequential data structure of ORAC. Difficulties arise at several levels, one of the more serious ones being illustrated by Table 1, which shows the significant frequency of multiple occurrences of datafields within one CLF database entry. This effect will be further compounded by the variable length text strings used by REACCS for some datafields, particularly COMMENTS and KEYPHRASES. As a consequence, each occurrence of these datafields may consist of lines up to several hundred characters, which must be split up into several lines (i.e., several ORAC datafields in a datatable) to conform to the maximum string length in ORAC (70 characters).

The hard truth is that a one to one functional translation of all textual data from a complete database entry of CLF to ORAC is neither possible nor practical. It is not the bulk of the information which causes the problem but its organization. The data are organized according to rules determined by one system, and these rules do not comply with those of the other system. Therefore, several adaptations had to be made and priorities set to enable the conversion.

The explosion of the number of datafields required for a complete translation of a CLF entry to SMD format originates from a single variation. The presence of multiple variations would augment this problem, but fortunately the number of entries in the CLF database having more than one variation is rather limited. As a consequence, we decided to split every CLF entry with multiple variations into separate ORAC database entries (called datacards).

Another consideration is the display of the data in ORAC. To show data from a datatable, we must know how many lines to display. If we show too few lines, the data will be incompletely displayed, but if we reserve too many lines for a datatable, this space would be lost for other datatables, as it is impossible in ORAC to scale the display space to meet the display needs. The space on the first two menu pages in ORAC is too limited to show all data, so choices had to be made.

It is important to note that the datafields to be used in the SMD file should preferably correspond to existing datafields in the ORAC system, as far as searchable data are concerned. The reason behind this is the correlation between searchable datafields and query options in the ORAC search menus. At present these menus are defined, more or less uniformly, across all databases. The corresponding uniformity of datafields

throughout all databases within ORAC is desirable from a user's point of view, as a single query is then sufficient to search all databases. It is not possible to use a single query to search a field in database X and the same field in database Y if they are of a different type or are attached to different thesauri. Therefore ORAC's existing data definitions of searchable datafields determine those of the searchable datafields in ORAC CLF. As a consequence, as many of CLF's searchable data as possible must be fit into the existing (searchable) datafields of the ORAC database. CLF's nonsearchable data can easily be transferred to nonsearchable datafields in ORAC that may be newly created.

Not all searchable CLF fields could be transferred to searchable ORAC datafields. In order to display all of the reaction data for these cases, a large nonsearchable ORAC datatable (called CUME-INFO) was defined to hold most REACCS textual datafields to avoid potential loss of data. Headers are added to the data in this datatable to indicate the original REACCS datafield. The CUME-INFO datatable simply contains all the data for display only, which can be viewed by the user via newly defined menu pages.

Reactrans contains several routines to translate textual data. Which routine is used depends on the datatype of the datafield and the destination in ORAC (Table 2 shows all CLF datafields with their datatype and ORAC destination). The handling of each CLF datafield during the translation is governed by a special input file, which mentions for each datafield the output datafield name and the type of conversion needed, e.g., "Integer", "Keyword", or "Var-text" (variable length text string). This approach guarantees a high flexibility in the composition of the output database without having to change the translation program.

The remainder of this section will discuss the manipulations each individual datafield is subjected to during the translation.

Referencing Data. The first occurrence of the CLF VOL., NO., PG., and YEAR fields can be read directly into corresponding datafields that exist in ORAC. The AUTHOR field is processed into individual lines per author and those lines are parsed into *last, initials* which is the format used in other ORAC databases. This processing of author names is complicated by the variety of formats used in the CLF author fields. These variations originate from the way punctuation is used, from the usage of full first names instead of initials, and from the occurrence of suffixes such as Jr. or III, etc. The ORAC AUTHOR field is used not only for queries on an author name, but also to fill a special datafield, called AUTHOR-TITLE. This field is automatically formatted by the ORAC program. The output is a string of maximum length 50 characters, having commas between the authors, and which is closed with "et al." if the length is about to exceed the 50 characters. AUTHOR-TITLE is used to display the authors on the first page of the ORAC data display menu (see also Figure 2). A similar special datafield exists for the bibliographic part of the reference. This datafield (CITATION) compiles a nice looking reference from the JOURNAL, VOLUME, PAGE, and YEAR datafields.

Table 2. Description and Destination of All CLF Datafields^a

REACCS name	description	datatype	ORAC destination
RXNTEXT	reaction conditions above arrow	VT	CONDITIONS; CUME_INFO
LITTEXT	literature reference	VT	CUME_INFO
EXTREG	external registry number	A12	discarded
FOLDER	used internally in MDL databases	A12	discarded
PATH	path of branching reaction series	A3	discarded
STEP	step in reaction series	VT	discarded
AUTHOR	author	VT	AUTHORS; AUTHOR_TITLE
VOL.	volume of journal	A16	VOLUME; CITATION
NO.	number of journal	A16	JOURNAL_PART_NUMBER
PG.	page of journal	A16	PAGE; CITATION
YEAR	year of journal	I4	YEAR; CITATION
JRNL	journal name	VT	JOURNALS; CITATION
RXNYIELD	reaction yield	RR	YIELD
REACTANT.GRADE(n)	reactant specifier (e.g., 1 mmol)	VT	REACTANT_GRADE_n (n = 1, 2, 3)
CATALYST.REGNO	name of reagent	RN	ACTUAL_REAGENT_KEY
CATALYST.GRADE	reagent specifier	VT	discarded
SOLVENT.REGNO	solvent name	RN	SOLVENT_KEY
SOLVENT.GRADE	solvent specifier	VT	discarded
PRODUCT.YIELD(n)	per product yield	RR	PRODUCT_YIELD_n (n = 1, 2)
PRODUCT.GRADE(n)	product specifier	VT	PRODUCT_GRADE_n (n = 1, 2)
ATMOSPHERE	gas atmosphere	A5	ATMOSPHERE; CUME_INFO
PRESSURE	pressure (atm)	RR	PRESSURE; CUME_INFO
TIME	reaction time (h)	RR	TIME; CUME_INFO
TEMP	reaction temperature (°C)	RR	TEMP; CUME_INFO
REFLUX	reflux or not	A1	CUME_INFO
OTHER	other conditions	VT	CUME_INFO
WORKUP	reaction workup	VT	CUME_INFO
COMMENTS	general comments	VT	COMMENTS; CUME_INFO
KEYWORDS	keywords	VT	REACTION_KEYS; CUME_INFO
MDLNUMBER	MDL number (for each variation)	A12	MDLNUMBER
KEYPHRASES	keyphrases	VT	REACTION_NAME; REACTION_KEYS; CUME_INFO

^a In = integer of size n; An = character string of size n; VT = variable length character string; RR = real range; RN = registry number.

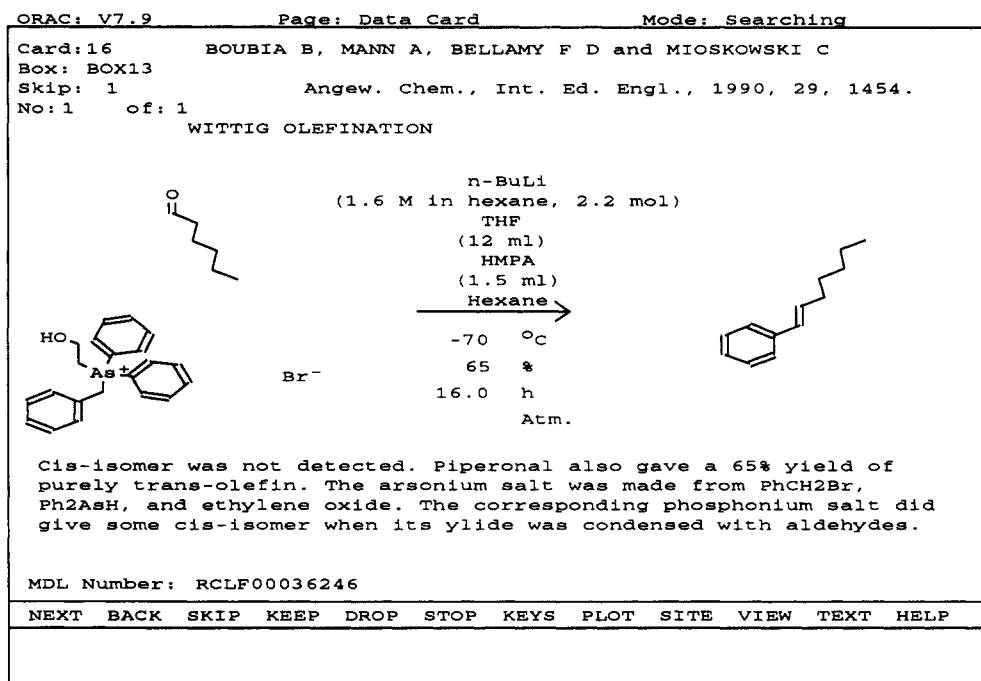


Figure 2. First page (data page) of the ORAC data display defined for CLF featuring (among others) CONDITIONS (translated from RXNTEXT) and COMMENTS.

Unfortunately, the JOURNAL datafield is thesaurus-driven, meaning that only journals defined in the thesaurus can be read into this field. As a consequence, a failure to read the journal results in a datacard without a reference, which is in fact a useless datacard. For this reason, all journals appearing in CLF and not known in the ORAC thesaurus have been identified. The majority of them have been defined as new journals in the thesaurus, or as synonyms of existing ones. The remainder of the unknown journals were misspellings of

known journals. It was not desirable to introduce spelling errors in the thesaurus, as this thesaurus can be browsed by the users. Therefore, we created a list containing the misspelled journals followed by their correct spelling, and this list was used by the translation program to correct the errors.

The special datafields CITATION and AUTHOR_TITLE take over the functionality of the CLF datafield LITTEXT. However, LITTEXT cannot be ignored, because these ORAC datafields can display only one reference. Therefore, LIT-

TEXT is read into CUME-INFO to allow the user access to the subsequent literature references.

Numerical Data. The numerical data cannot be passed directly, as REACCS exports either real numbers or ranges of real numbers for TIME (e.g. 0.1–0.2 h), PRESSURE, TEMP, and RXNYIELD, which are followed by a unit designator. ORAC has no such range specification of unit handling. A real range is replaced with an average value for the TIME, YIELD, and PRESSURE fields in order to maintain its numerical attributes. TEMPERATURE ranges are replaced by the first value of the range, as this value usually seems to indicate the reaction temperature, while the second value is the temperature to which the reaction mixture is warmed up or cooled down; taking the average would lead to ridiculous results, e.g., –78 to +20 °C (reacting at –78 °C, warming up to room temperature) would be translated to –29 °C. YIELDS and TEMPERATURE have to be converted to integer values to conform to the existing ORAC datafields. TIME and PRESSURE are rounded to two significant digits to obtain a more esthetic display (REACCS regularly shows times like 50 min as 0.833333 h).

Solvents and Reagents. The solvents and reagents data are translated in unprocessed form by Reactrans, but during import into ORAC they are checked against the corresponding thesaurus. This results in frequent refusals due to solvents and reagents unknown to the thesaurus. Although the vast majority of the reagents could be read, it was still considered too laborious to add all the unknown reagents to the thesaurus. Some reagents, however, were abbreviated differently in ORAC than in REACCS and were therefore frequently refused. For these last cases, new synonyms have been defined in the ORAC ACTUAL-REAGENT-KEY thesaurus. Typical examples are all solid metal reagents, which are referred to by their atomic symbol in ORAC, while REACCS uses representations of the form "Na (m)". The omission of some solvents and reagents data is not really a loss of information, as most of this data are also present in the RXNTEXT (its processing is treated under "other datafields"); however, any data refused by the thesaurus are not searchable. There were not many undue refusals of solvents. Most refusals were caused by reagents erroneously assigned as solvents in the CLF database or by spelling errors.

Reaction Keywords and Keyphrases. In both REACCS and ORAC, reaction keywords are words characterizing reactions or reaction types, such as "halogenation", "isomerization", "stereoselective", or "cycloaddition". Both systems have a fixed list of keywords; REACCS has 28 keywords and ORAC 486. In ORAC, the REACTION-KEYS datafield is thesaurus-driven. In our situation it is undesirable to add new reaction keywords to the thesaurus; these cannot be used to search reactions in existing databases. Fortunately, practically all REACCS reaction keywords could be used directly as ORAC reaction keywords, so there was no need to expand the thesaurus.

The keyphrase datafield in REACCS contains free-text information such as name reactions, compound classes, or reaction types. Many of the keyphrases used, such as "epoxidation", "desulfurization", "rearrangement", or "intramolecular", correspond to the 486 ORAC reaction keywords but not to the 28 REACCS reaction keywords. As a consequence, keyphrases can be used to generate additional reaction keywords for ORAC. This is very useful, as REACCS keywords can generate only a small fraction of ORAC's reaction keywords. Similarly, these keyphrases can be used to generate the field values for ORAC's REACTION-NAME

datafield, such as "Diels-Alder", "Grignard", "Shapiro", or "Pictet-Spengler".

To speed up the processing of keyphrases, an index file was built to be used by the translation program. In this file, all keyphrases are listed together with the reaction keywords and the reaction name that can be derived from each of them. The index file was built automatically by a separate program using a list of (sub)strings that generate reaction keywords and/or a reaction name. For instance, the strings "(4+2)", "(2+4)", and "Diels-Alder" all generate the ORAC reaction keyword "[4+2] cycloaddition" and the reaction name "Diels-Alder". Strings that could not be handled reliably in this way (e.g., "Wittig" or "Claisen") were processed manually afterwards. This manual processing also allowed us to derive additional ORAC keywords and/or reaction names from misspelled keyphrases, such as "Arndt-Eisert" from both "Arndt-Eisert" and "Arndt-Erstart". The CLF database contains 41 549 keyphrases; after conversion to upper case the number of different keyphrases turned out to be 10 514. From 3555 keyphrases, reaction keywords and/or a reaction name could be derived. This number is much lower because about half of the keyphrases are compound names or classes.

The total number of reaction keywords generated for a single reaction (variation) may easily exceed ORAC's data-table limit of six, especially because duplicate keywords may be generated from different keyphrases or REACCS keywords. Therefore, a priority scheme for ORAC reaction keywords was used, in which a high priority was given to specific keywords (such as "1,3-dipolar cycloaddition") and a low priority to rather general keywords (such as "addition"). The list of reaction keywords generated for a single reaction was sorted according to decreasing priority, at the same time eliminating duplicates. Only the first six keywords were written to the SMD file.

All reaction names generated were written to the SMD file, although ORAC accepts only one (the first); no attempt was made to select one of the reaction names as the most appropriate. Figure 3 shows the second display page in ORAC featuring REACTION-KEYS and REACTION-NAME.

Other Datafields. The remaining REACCS datafields could not be converted to searchable ORAC datafields. The data in these datafields can be transferred to nonsearchable ORAC datafields which do not necessarily have to be kept in precise correspondence with those found in the existing ORAC databases. Nevertheless, existing ORAC fields have been used as much as possible in order to minimize the complexity of the ORAC database definitions. This was possible for RXNTEXT and COMMENTS which would be translated to existing datafields in ORAC, CONDITIONS and COMMENTS, respectively. Both datafields are also read into the CUME-INFO datatable, since the ORAC data display has no space for more than six lines for either datafield, this number being frequently exceeded by one of these two datafields. For RXNTEXT this is caused solely by multiple occurrences of the datafield, as can be seen from Table 1. COMMENTS usually consists of a few lines, but these lines tend to be (very) long. Reactrans cuts these lines into multiple lines at the last space before the 70th character. No reformatting is performed, since the COMMENTS regularly contain formatted texts, such as tables. The (rare) long RXNTEXT lines are truncated at 30 characters automatically by the ORAC display menu. Figure 2 shows an example of the first display page in ORAC featuring both RXNTEXT and COMMENTS.

Apart from CUME-INFO which is a large nonsearchable datatable (140 elements), seven other new nonsearchable

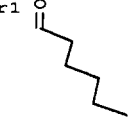
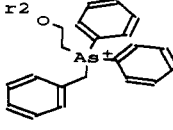
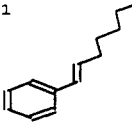
ORAC: V7.9		Page: Keys Card		Mode: Searching	
r1  0.9 mmol	r2  1 mmol	r3 Br ⁻	p1  65 % (E)	p2 *	
Card: 16 Box: BOX13 Journal: Angew. Chem., Int. Ed. Engl. Year: 1990 Vol: 29 Page: 1454 Name of Reaction: WITTIG OLEFINATION					
Authors: BOUBIA B MANN A BELLAMY F D MIOSKOWSKI C		Reagent Keys: n-BuLi			
		Reaction Keys: REGIOSELECTIVE GEOSLECTIVE		WITTIG OLEFINATION OLEFINATION	
		Solvent Keys: HMPA		THF Hexane	
NEXT BACK SKIP KEEP DROP STOP DATA PLOT TEXT VIEW HELP					

Figure 3. Second page (keys page) of the ORAC data display showing (among others) the keyword fields. Below the structures the associated grades and yields (for products) are displayed.

datafields have been derived from seven CLF fields: three for reactant grades, two for product grades, and two for product yields. These fields are displayed underneath the molecule they belong to, on the second menu page defined for ORAC CLF, shown in Figure 3. The solvent and catalyst grades are not translated, as it is impossible to associate them with the solvent and reagent keys in the ORAC database. This was not considered as a major loss, as most information in these datafields is redundant with the RXNTEXT.

CLF contains three datafields (EXTREG, STEP, and PATH) to show the relationship between the steps of a multiple stepsynthesis. Using the information in these datafields, other steps from the same synthesis can be searched in REACCS. This feature does not exist in ORAC, so these datafields have lost their use in the ORAC version of CLF and are therefore not translated. Similarly, the FOLDER datafield is not translated, because it is only used by Molecular Design internally.

Three CLF datafields (REFLUX, WORKUP, and OTHER) do not have an equivalent in the existing ORAC databases. The limited space on the ORAC display menus led us to the decision not to include the information of these fields on either of the two display pages. Instead, the information collected in these fields is read into CUME-INFO, which is user-accessible through the text menu pages. The REFLUX field demanded special treatment; the only existing value in CLF was "d" (meaning refluxed), which was translated to "refluxed" in CUME-INFO.

RESULTS

All 36 601 reactions from CLF were subjected to the translation process. This number is increased to 37 304 by the expansion of 604 multiple-variation reactions into 1307 separate reactions. The final result of the translation process is an ORAC database containing 35 059 reactions. Table 3 shows the resources needed for each translation step. Four individual steps are reviewed in the table: the export of the data from REACCS to an RDfile, the actual translation step from RDfile format to SMD format using the Reactrans

Table 3. Resources Needed for Each Translation Step

	RDfile export	Reactrans	SMD file import	screen
CPU time/reaction ^a (s)	5.9	1.20	10.8	7.0
elapse time/reaction ^a (s)	12.0	1.52	21.4	8.7
total CPU time ^a (h)	60	12.2	109	68
total elapse time ^a (h)	122	15.5	215	85
av CPU load ^a (%)	49	79	51	80
disk space input file (MB)	70	230	110	130
disk space output file (MB)	230	110	130	170

^a CPU time on a μ -VAX 3300.

Table 4. Summary of the Reasons Datacards Were Refused by Reactrans

	no. of cards	total cards (%)
too many atoms or bonds	27	0.07
too many reactants or products	855	2.34
atom mapping data absent	11	0.03
atom map number not on both sides	7	0.02
mapped atoms differ in atom type	54	0.15
no reaction center atoms	93	0.25
total of untranslatable reactions	1047	2.86

program, the import of the SMD file into ORAC, and finally the generation of structural screens using the ORAC screen utility.

The total of 2245 reactions (including variations) lost in the translation process are divided between the second and the third step of the translation process. Reactrans and ORAC import refused respectively 1076 (1047 excluding variations) and 1169 reactions. The frequencies of the several types of Reactrans refusals can be found in Table 4. Most of these failures were due to the ORAC upper limit to the number of reactants (3) and the number of products (2). The datacards that exceed one of these numbers are refused by the Reactrans program, but the differences between REACCS and ORAC in the approach to salts cause problems. REACCS treats a salt as an entity consisting of a cation and an anion, while ORAC considers a salt as consisting of two fragments. The

Table 5. Import Statistics of Textual Data

datafield	total no. of data lines	refused data lines	no. of cards having this field	cards with refusals of this field
ATMOSPHERE	5 722	0	5 722	0
AUTHOR	103 255	660 (0.6%)	35 872	400 (1.1%)
COMMENTS	47 191	5 (0.01%)	17 458	2 (0.01%)
CUME-INFO	635 862	0	36 228	0
JOURNAL	36 216	0	36 216	0
REACTION-KEYS	86 242	0	30 408	0
MDLNUMBER	36 228	0	36 228	0
JOURNAL-PART-NUMBER	29 852	0	29 852	0
PAGE	36 216	0	36 216	0
PRODUCT-GRADE-1	7 126	2 (0.03%)	7 124	2 (0.03%)
PRODUCT-GRADE-2	2 207	0	2 207	0
PRESSURE	942	0	942	0
ACTUAL-REAGENT-KEY	46 235	4102 (8.9%)	28 054	1556 (5.5%)
REACTANT-GRADE-1	12 792	0	12 792	0
REACTANT-GRADE-2	6 215	0	6 215	0
REACTANT-GRADE-3	267	0	267	0
REACTION-NAME	5 478	374 (6.8%)	5 115	333 (6.5%)
CONDITIONS	99 389	974 (0.9%)	32 051	439 (1.4%)
YIELD	25 342	0	25 342	0
SOLVENT-KEY	30 237	461 (1.5%)	24 009	144 (0.5%)
TEMPERATURE	20 068	0	20 068	0
TIME	19 388	0	19 388	0
VOLUME	30 731	0	30 731	0
PRODUCT-YIELD-1	17 030	0	17 030	0
PRODUCT-YIELD-2	3 343	0	3 343	0
YEAR	36 216	0	36 216	0

splitting up of salts in ORAC resulted in additional datacards having too many products or reactants. These cards passed Reactrans unnoticed, but were duly refused by ORAC import. This refusal of datacards by ORAC import for too many fragments occurred 1029 times out of a total of 1185 error messages. Two other problems were detected by ORAC import, accounting for the other rejected cards. A set of 28 cards was rejected for "incorrect valency" and 128 for "illegal stereocentre". Apparently, some cards have more than one problem as we found 1185 error messages for 1169 refusals.

Reactrans does not check whether textual data will exceed ORAC limits, as ORAC import will check the data automatically. Nevertheless, some textual data are already rejected by Reactrans in order to cut down the resources needed by the program. Therefore, Reactrans has an upper limit to the number of occurrences of datafields, which led to the refusal of 15 author data lines (more than 12 authors) and 16 refusals of a text line (more than 24 text lines of the same datafield). Obviously, these data would be refused by ORAC import otherwise. The only datafield that is really preselected by Reactrans is the REACTION-KEYS field, as Reactrans assigns priorities to keywords and selects the most important keywords if the limit (6) is exceeded. With 3142 reactions (out of the 36 228 variations/reactions that were successfully translated) one or more keywords were discarded.

Table 5 gives an overview of the import of the CLF textual data into ORAC. Textual data can be refused by ORAC import for two reasons: an attempt is made to put data into a thesaurus-driven datafield but the data are not known to the thesaurus, or the upper limit to the number of datafields in a datatable is exceeded. The former reason accounts for all refused solvent data and the vast majority of the refused reagent data. The other refusals were all due to overflow of the maximum number of data for a datatable. The PRODUCT-GRADE-1 field (which is a datafield, not a datatable) also has two cases of overflow despite the fact that the field only consists of the first occurrence of the PRODUCT.GRADE field in REACCS. This is caused by two very long data lines which were exceeding the 70 characters and therefore split up in two lines by Reactrans. Obviously,

the attempt to read a second line in a datafield was unsuccessful.

CONCLUSIONS

We consider ORAC to be a browsing system for lead references and not a laboratory journal or a data compilation system. Inclusion of at least the structural aspects and the lead references of the database already given us as a database that can meet this goal. Therefore, we consider the database conversion successful, as we not only managed to translate most structural data and references but also succeeded in the transfer of most of the textual data. The conversion of other REACCS databases can be accomplished easily by applying the translation procedure described in this paper to other REACCS databases. Only the keyphrases datafield will require some manual processing in order to obtain a similar proportion of translated keyphrases as in the transfer of CLF.

The translation procedure forced us to deal with both ORAC and REACCS in great detail, which provided us with a better insight in the structure of both databases. Earlier comparative evaluations^{7,8} of ORAC and REACCS focused merely on the content of the database and the user-friendliness. We are now able to comment on the database structure of both databases and the quality of the data in REACCS. Generally, the textual data structure in REACCS looks more sophisticated compared to that of ORAC. The most pronounced difference is the use of a hierarchical structure versus a sequential one, but remarkably, this distinction was not a major obstacle to the translation. REACCS' hierarchical structure could be flattened easily at the variation level. The real shortcomings of the ORAC database structure were the limitation of the number of reactants and products, the need to predefine datatable sizes, and the upper limit of 70 characters to the length of a data line. These restrictions, leading to frequent refusals of data, emerge from ORAC's use of fixed display menus which the program cannot automatically adapt to the data. REACCS can allow longer lines because it is able to use different fonts to resize the text on the screen. Naturally, this requires a more sophisticated terminal or emulator.

Another significant difference is the use of free text searches versus thesauri-driven searches. A thesaurus, as used by ORAC, allows the user to restrict himself to a single keyword search without bothering about synonyms or alternative spellings (a search for "Diels-Alder" also finds "Diels Alder" and "[4+2]-cycloaddition"). A disadvantage of thesaurus-driven keywords is the extra effort demanded from database compilers to check their data against the thesaurus. On the other hand, input of free text is prone to spelling errors, which would immediately be found using a thesaurus. Spelling errors in searchable data (journals, authors, keyphrases) were the most frequent among the data errors we found in CLF. Some additional problems were detected by Reactrans and ORAC import (such as mapping oxygens on carbons, incorrect stereocenters, etc.), but the total number of errors found by us was certainly not outrageous. Therefore, we consider REACCS' CLF database to be of a sufficient high quality. Obviously, the described data-transfer process did not give us insight in the quality of the data in the existing ORAC databases, so no comparison could be made.

Organic reaction database translation programs, of which the Reactrans program is an example, allow one to exchange data between reaction databases to arrive at a single, structurally based, organic reaction cataloging system. The benefits of such a system for both the user and the database maintainer are obvious.

ACKNOWLEDGMENT

We wish to thank Molecular Design Limited for access to the REACCS and ORAC export and import utilities and for permitting for the data transfer. We also thank D. W. Featherston (CAOS/CAMM Center) for his valuable advice during the preparation of this manuscript. The support of the Dutch National Science Foundation and the use of the services and facilities of the Dutch CAOS/CAMM Center, under Grant Nos. SON 326-052 and STW NCH99.1751, are gratefully acknowledged.

REFERENCES AND NOTES

- (1) REACCS (REaction ACCess System), Version 8.1, Molecular Design Ltd., San Leandro, CA.
- (2) ORAC (Organic Reactions Accessed by Computer), Version 7.9, Molecular Design Ltd., San Leandro, CA.
- (3) Wipke, W. T.; Dill, J.; Hounshell, D.; Moock, T.; Grier, D. Exploring Reactions with REACCS. In *Modern Approaches to Chemical Reaction Searching*; Willett, P., Ed.; Aldershot: Gower, U.K., 1986; pp. 92-117.
- (4) Johnson, A. P. Computer aids to synthesis planning. *Chem. Br.* **1985**, *21*, 59-67.
- (5) Chodosh, D. F.; Hill, J.; Shpilsky, L.; Mendelson, W. L. SYNthesis LIBrary, an expert system for chemical-reaction knowledge-base management. *Recl. Trav. Chim. Pays-Bas* **1992**, *111*, 247-254.
- (6) SYNLIB (SYNthesis LIBrary), Version 3.21, Distributed Chemical Graphics, Inc., Meadowbrook, PA.
- (7) Borkent, J. H.; Oukes, F.; Noordik, J. H. Chemical Reaction Searching Compared in REACCS, SYNLIB, and ORAC. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 148-150.
- (8) Zass, E. A. User's View of Chemical Reaction Information Sources. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 360-372.
- (9) Hendrickson, J. B.; Miller, T. M. Reaction Classification and Retrieval. A Linkage between Synthesis Generation and Reaction Databases. *J. Am. Chem. Soc.* **1991**, *113*, 902-910.
- (10) Hendrickson, J. B.; Miller, T. B. Reactions Indexing for Reaction Databases. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 403-408.
- (11) The Theilheimer databases have been derived from: Theilheimer, W. *Synthetic Methods of Organic Chemistry*; S. Karger: Basel, 1946-1980; Vols. 1-35.
- (12) The Cheminform RX database is supplied by MDL in ORAC and REACCS format since 1992. Since this database consists of several tens of thousands reactions, a subset called the CSM database is also available.
- (13) ChemInform, Fachinformationszentrum Chemie GmbH, Berlin, Germany.
- (14) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244-255.
- (15) Bebak, H.; Buse, C.; Donner, W. T.; Hoever, P.; Jacob, H.; Klaus, H.; Pesch, J.; Römet, J.; Schilling, P.; Woost, B.; Zirz, C. The Standard Molecular Data Format (SMD Format) as an Integration Tool in Computer Chemistry. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 1-5.
- (16) A draft proposal has been published for version 5.0: Barnard, J. M. Draft Specification for Revised Version of the Standard Molecular Data (SMD) Format. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 81-96.
- (17) Kasperek, S. V. *Computer Graphics and Chemical Structures*, Part 4; John Wiley & Sons: New York, 1990; pp 625-636.
- (18) Moock, T. E.; Nourse, J. G.; Grier, D.; Hounshell, W. D. The Implementation of Atom-Atom Mapping and Related Features in the Reaction Access System (REACCS), In *Chemical Structures: The International Language of Chemistry*; Warr, W. A., Ed.; Berlin: Springer-Verlag, 1988; pp 303-313.
- (19) MDL calls this a reacting center status, but it is more a bond attribute.