THE SPORE DATABASE: CONTENTS AND DATAFIELDS

1. Database Contents

1.1 Selection Criteria

"SPORE" is the acronym for "**S**olid-**P**hase **O**rganic **RE**actions", and the database therefore contains reactions where reactants and/or products of a chemical transformation are linked to some sort of solid (or potentially solid) support. More specifically, the synthetic target should be a "small molecule" and not a macromolecule, and many syntheses of biologically relevant macromolecules such

as long peptide chains, oligonucleotides or oligosaccharides are thus not included in the database.

There is no rule without exception, however. Publications dealing with these macromolecular natural products may contain synthetic aspects which could conceivably be relevant for the synthesis of smaller molecules (e.g., by using novel solid supports or special protecting groups), and studies of this kind may be included in SPORE if a reaction based on suffiently small molecules can be found in the article.

Furthermore, not all reactants or products in SPORE are necessarily linked to insoluble "solid" polymers during a reaction. First of all, there are soluble polymers such as non-crosslinked polystyrene or polyethylene glycol derivatives, which are nevertheless regarded as "solid supports" because a solid form of these polymers (usually obtained upon precipitation) facilitates the isolation and purification process - which is one of the typical advantages of solid-phase methodologies. Secondly, some reactions of entirely non-supported molecules may be crucial for a solid-phase strategy (e.g., when a ligand is attached to a linker prior to a coupling reaction of the resulting linker-ligand assembly with a solid support), and a few reactions of this type can also be found in the database.

Solution-phase reactions employing solid-supported reagents are typically assigned to the ChemInform RX (CIRX) and not to the SPORE database - but again, there is one exception. An unambiguous distinction between "reactant" and "reagent" is ensured in both databases by the strict definition that a "reactant" must introduce at least one carbon atom into the product structure. As an extension of this definition, solid-supported C-introducing agents are exclusively covered by SPORE and are excluded from CIRX, and this simple rule allows a facile and undisputed assignment of a given reaction to either of the two databases. Reactions featuring carbon-introducing compounds such as polymer-bound acylating agents or Wittig-type ylides (often regarded as "polymer-supported reagents" in the literature) are a topic of the SPORE database, whereas applications of, e.g., polymer-bound oxidants or reductants are included in the CIRX database (the solid-phase syntheses of these polymer-bound reagents, however, may well be found in the SPORE database).

1.2 Sources: Literature and Patents

The reactions in the SPORE database were gathered from articles published in relevant chemical journals and from publications of patents or patent applications. A group of reactions derived from one of these articles or patents represents an "abstract" (similarly to an abstract in the ChemInform journal and reaction database), and these reactions are linked by a special group of data (see below).

While articles from a wide range of different journals have found their way into SPORE (there were 118 different CODEN's in a current* release of the database), the majority of publications were taken from some of the major journals covering studies in organic and/or combinatorial chemistry (e.g., Tetrahedron Lett.: 23.4% of all citations, J.Comb. Chem.: 9.7%, J. Org. Chem.: 7.4%).

About 10% of the overall number of citations refer to patent publications. In a typical case, this would include references to both a patent publication [usually a "WO"-type application under the Patent Cooperation Treaty (PCT) of the World Intellectual Property Organization (WIPO)], and to an issued patent (generally a US patent), if possible. In the case of patent families and reissued patents, just one or two of all possible citations are typically included in the database.

1.3 "Secondary Citations"

In contrast to other databases, the group of reactions belonging to an abstract of the SPORE database may contain not just a "primary citation" [i.e., a set of data collected under the *LITREF(1)*† header of the data tree], but also one or more additional citations as so-called "secondary citations" [i.e., data collected under *LITREF(2)*, *LITREF(3)*, etc.]. One example of secondary citations has just been given: patent-derived reactions may contain citations relating to a patent application and to an issued patent.

---

* All figures refer to the SPORE database version completed in August 2008

† Italicized terms in capital letters always represent family or field names of the SPORE data tree

Generally, two (or more) different publications can be combined to one single abstract with primary and secondary citations if their contents are similar or closely connected. In addition to the (practically identical) contents of patent application and patent, this also applies, e.g., to a patent and the subsequent publication of the reactions in a chemical journal, or to a preliminary account and a following detailed report of a particular study, or the synthesis of a reagent and its applications. The first abstract in SPORE, e.g., is Merrifield's original introduction of solid-phase peptide synthesis, which was published in a J. Am. Chem. Soc. article from 1963 (primary citation in abstract 6963001) and was again described in a more detailed (and extended) form in a Science article from 1965 (secondary citation).

One specific use of secondary citations can be found in the so-called "collective abstracts", where the database user can find references to a number of publications related to a particular topic, which cannot be covered by conventional SPORE abstracts (e.g., due to non-reaction related topics such as compound properties or spectral studies, or because of the size of the synthesized compounds). Some illustrative examples can be found, e.g., by combining the Keyword (i.e., *RXN-TYPE*) "collective abstract" with terms in the *TOPICS* datafield such as "* FT-IR*", " *scavenger*", or "*pyrrole polyamide*".

2. Datafields

2.1 Publication-Related Data

RXN:VARIATION:LITREF:AUTHOR
RXN:VARIATION:LITREF:TITLE
RXN:VARIATION:LITREF:JOURNAL:VOL.
RXN:VARIATION:LITREF:JOURNAL:NO.
RXN:VARIATION:LITREF:JOURNAL:PG.
RXN:VARIATION:LITREF:JOURNAL:YEAR
RXN:VARIATION:LITREF:JOURNAL:CODEN
RXN:VARIATION:LITREF:JOURNAL:JRNL
RXN:VARIATION:LITREF:LOCATION
RXN:VARIATION:LITREF:LANGUAGE
RXN:VARIATION:LITREF:EXPDET
RXN:VARIATION:LITREF:FULL_CITATION

Similarly to various other databases, bibliographic data are situated in the "*LITREF:JOURNAL*" datafields *JRNL* (Journal Name), *CODEN*, *YEAR* (Publication Year), *VOL.* (Volume), *NO.* (Number) and *PG.* (Page Numbers).

The other datafields under "*LITREF*" provide a *FULL_CITATION* as well as the *TITLE* and the *LANGUAGE* of the publication, the names of the *AUTHOR*(s) and the *LOCATION* (i.e., the address of the institute or company responsible for the study). The *EXPDET* datafield indicates whether full experimental details are included in the publication (entry "yes") or in a supplement (entry "supplementary").

2.2 Abstract-Based Data

RXN:VARIATION:SYSNO
RXN:VARIATION:SYSTEXT
RXN:VARIATION:CROSSREF:REFNO
RXN:VARIATION:CROSSREF:ID
RXN:VARIATION:MDLNUMBER

The abstract-based datafields present in the ChemInform RX (CIRX) database have also been incorporated into the SPORE database. All reactions of a given abstract are linked by the presence of the same, unique *REFNO* (7-digit reference number; invariably starting with 69 or 72 in the case of SPORE abstracts, followed by two digits for the publication year - e.g., 6999189 for the 189th abstract from 1999, or 7204106 for the 106th abstract from 2004).

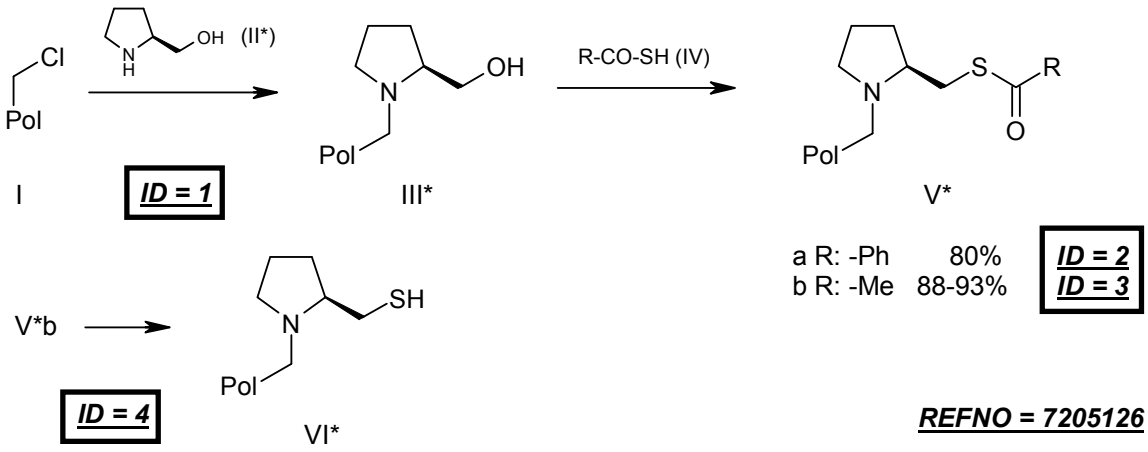
The general contents of a particular publication are characterized by the *SYSNO* code and the *SYSTEXT* term(s) of a ChemInform-specific classification system. Database users should always remember that these terms (e.g. "cycloaddition", "pyrazole derivatives", "steroids") refer to the major topics of the entire publication, but not necessarily to each individual reaction of this abstract (keyphrases relating to individual reactions are the *RXN_TYPE* terms described in chapter 2.4).

Reactions within an abstract are assigned *ID* values in ascending order. Any reaction of the SPORE database is therefore unambiguously characterized both by the *MDLNUMBER* (a datafield especially created for this purpose) and by the particular combination of *REFNO* and *ID* entries. These datafields are also used for

cross-references between reactions or abstracts in the texts of various *TOPICS* entries (cf. chapter 2.4).

```
RXN:VARIATION:PRIMARY
RXN:VARIATION:RXNREF:EXTREG
RXN:VARIATION:RXNREF:PATH
RXN:VARIATION:RXNREF:STEP
RXN:VARIATION:CROSSREF:REFNO
RXN:VARIATION:CROSSREF:BASERXN
RXN:VARIATION:CROSSREF:PRERXN
RXN:VARIATION:CROSSREF:SUCCRXN
```

Some of the datafields of this list also refer to the *ID* values of reactions. Groups of very similar (Markush-type) reactions have entries in *BASERXN* which specify the *ID* of a representative (e.g., high-yielding) member of this group; a further datafield *PRIMARY* allows searches providing just one member (*viz*, the first) of the group. Correlations between reactions of a synthetic sequence are provided by two alternative indexing systems: the *RXNREF* datafields *EXTREG*/*PATH*/*STEP* (which are also present in various other databases), and the datafields *PRERXN* and *SUCCRXN* (both of which contain data referring to an *ID*). The contents of these datafields are illustrated by the following example:



ID	RXNREF	PATH	STEP	PRIMARY	BASERXN	PRERXN	SUCCRXN
1	720512601	A	1 OF 2	Primary	-	-	2
	720512601	B	1 OF 3				3
2	720512601	A	2 OF 2	Primary	3	1	-
3	720512601	B	2 OF 3	-	3	1	4
4	720512601	B	3 OF 3	Primary	-	3	-

2.3 Standard Experimental Reaction Data

A reaction is characterized by the structures of the reactant(s) and product(s) which are connected to provide a Rxn-RegNo for this particular transformation, and by a set of additional data in one or more *VARIATION(*s) which describe the synthetic methodology (or methodologies) to achieve this transformation. Standard experimental data which are present in both SPORE and CIRX are assigned to the datafields described in this chapter. Notably, some of these fields allow the convenient handling of each single step of a multistep procedure.

RXN:VARIATION:COMMENTS
RXN:VARIATION:CONVERSION
RXN:VARIATION:NSTEPS

General information to the reaction or method is given in the datafields *COMMENTS* for data which cannot be assigned to single reactants or products (such as product ratios), in *NSTEPS* which specifies the number of steps in multistep procedures (the value is thus invariably >1), and in *CONVERSION* for reactions where the original publication only gave data for the conversion of the reactants and not for the yields

of the products (the "conversion" data of some publications, however, have been interpreted as being equivalent to "yield" data and were put into the latter datafield in order to improve data consistency).

```
RXN:VARIATION:PRODUCT:YIELD
RXN:VARIATION:PRODUCT:CS
RXN:VARIATION:PRODUCT:DS
RXN:VARIATION:PRODUCT:DE
RXN:VARIATION:PRODUCT:EE
RXN:VARIATION:STEPNO:ST.NO
RXN:VARIATION:STEPNO:ST.YLD
```

The efficiency of the synthetic procedure becomes apparent from the values present in the product-related datafields *YIELD*, *CS* (chemoselectivity), *DS* (diastereoselectivity), *DE* (diastereomeric excess) and *EE* (enantioselectivity). The value in the *ST.YLD* (step yield) field of the *STEPNO* branch is usually identical to the *YIELD* value (especially in single-step reactions), but in the case of multistep procedures it can contain the yield for the step with the number *ST.NO*.

```
RXN:VARIATION:REACTANT:GRADE
RXN:VARIATION:REACTANT:ADD.STEP
RXN:VARIATION:PRODUCT:GRADE
```

Additional data to a *REACTANT* or *PRODUCT* may be included in the corresponding *GRADE* datafields, and the datafield *REACTANT:ADD.STEP* specifies the number of the step in which this reactant was added in a multistep procedure. In the SPORE database, the *PRODUCT:GRADE* field often contains comments to the value given in the *YIELD* field or to the purities of the products (entries such as "overall yield", "HPLC", or "95% purity"), whereas *REACTANT:GRADE* typically specifies the amount or concentration of the reactant (e.g., "5 equiv.", "excess", "25%") or provides the name and/or source of the solid-supported starting material of a reaction sequence as it is given in the publication (e.g., "Wang resin (1.0 mmol/g); Source: NovaBiochem", or "Preparation: cf. 6994040" with a cross-reference to the *REFNO* described in chapter 2.2).

```
RXN:VARIATION:STEPNO:CATALYST:REGNO
RXN:VARIATION:STEPNO:CATALYST:CAT
RXN:VARIATION:STEPNO:SOLVENT:REGNO
```

Similarly to other databases, compounds besides the reactants required for the synthetic process are specified in the *CATALYSt* and *SOLVENT* fields via the *REGNO* of the corresponding molecules in the database. The field *CATALYST* comprises the two terms "catalyst" and "reagent" commonly used in the literature, and to distinguish between the two, the field *CAT* had already been introduced in the CIRX database to identify catalysts in the narrower sense (i.e., compounds having a catalytic effect or being used in catalytic amounts) by the presence of an entry in the field *CAT*.

RXN:VARIATION:STEPNO:CONDITIONS:PRESSURE
RXN:VARIATION:STEPNO:CONDITIONS:TIME
RXN:VARIATION:STEPNO:CONDITIONS:TEMP
RXN:VARIATION:STEPNO:CONDITIONS:WARMUP
RXN:VARIATION:STEPNO:CONDITIONS:COOLDOWN
RXN:VARIATION:STEPNO:CONDITIONS:REFLUX
RXN:VARIATION:STEPNO:CONDITIONS:PH
RXN:VARIATION:STEPNO:CONDITIONS:OTHER
RXN:VARIATION:STEPNO:OPERATION

Further reaction parameters are contained in the self-explaining fields *PRESSURE*, *TIME*, *TEMP* and *PH*. A reaction performed under reflux of the solvent has an entry in the field *REFLUX*, and heating or cooling during the reaction are indicated by the presence of entries in *WARMUP* or *COOLDOWN*, respectively. The field *OTHER* is used for special procedures such as "irradiation", "microwaves", "ultrasound" or "electrolysis", while any additional relevant reaction parameters are kept in the datafield *OPERATION* (e.g., "argon", "3 cycles", "workup with aq. HCl").

RXN:VARIATION:STEPNO:RXNTEXT
RXN:VARIATION:STEPNO:CONDTEXT

The two datafields *RXNTEXT* and *CONDTEXT* are primarily used for display purposes since they contain all the above-mentioned data as formatted text strings. It should be noted, however, that *RXNTEXT* also contains additional information not present in any other datafield, since the display of the reagents or solvents includes not only names for these compounds (which can also be found in the molecule field *SYMBOL*;

cf. chapter 2.5) but also specifications of, e.g., reagent concentrations or solvent ratios [such as "20% piperidine", "3 equiv. 1N aq. NaOH", "toluene/EtOH (9:1)"].

2.4 SPORE-Specific Data

In addition to datafields already known from the CIRX database, a series of additional datafields have been incorporated into SPORE to meet the requirements of a database dealing with reactions on a variety of different supports. Nearly all of these datafields are unique to SPORE and are based on specially developed data structures and classification strategies, which provide a database user with a range of opportunities not present in any other reaction database.

RXN:VARIATION:TOPICS

The *TOPICS* datafield may be regarded as an exception, since datafields of this kind are also present in other databases. It includes various sorts of textual information on the particular reaction, such as the general synthetic strategy, the influence of different reaction parameters, the source of the reactants or catalysts, spectral properties of the solid-supported compounds, applications of the products, results of pharmaceutical investigations, etc. In abstracts dealing with the "historical" period of solid-phase synthesis (let's say, from 1963 to the late 1990's), the entries in the *TOPICS* field are often rather brief and summarize findings in an abbreviated form, but the entries covering more recent studies usually try to include any interesting piece of information provided by the original publication.

Various cross-references to other studies or reactions are included in the *TOPICS* entries. While the reference to other studies is accomplished by simply citing the *REFNO* of the abstract describing the study (e.g., "cf. 7204105"), cross-references to other reactions of the current abstract cite both the *ID* and the *MDLNUMBER* (cf. chapter 2.2) of this particular reaction (e.g., "cf. ID 10 [RSPO69025407]").

RXN:VARIATION:RXN_TYPE

The datafield *RXN_TYPE* contains keyphrases to the particular reaction which - in contrast to the field *KEYPHRASES* of other databases - were not generated by computer programs but were assigned by chemists (although computers help in validating these assigments). Since human beings tend to disagree about the relevance and meanings of appropriate terms, the assignment of the *RXN_TYPE* entries is based on a list of clearly defined specific terms, which has been growing steadily and today covers chemical transformations (such as "C-acylation", "lactam formation", "S-oxidation (thiol)", or "metathesis") as well as name reactions ("Wittig olefin synthesis", "Paal-Knorr pyrrole synthesis", etc., etc.), experimental conditions ("photoreaction", "microwaves", "biocatalyzed reaction"), general terms ("nucleotide synthesis", "multicomponent reaction", "click chemistry"), and specific polymer-related terms (such as "polymer coupling", "traceless cleavage (H-aryl)" or "soluble-polymer reaction").

Currently, >360 different terms have been introduced as *RXN_TYPE* entries in SPORE. The definitions for all of these terms can be found in the accompanying paper, "Definitions of SPORE Keyphrases (RXN_TYPE's)".

RXN:VARIATION:STEPNO:RXN_CONDITIONS

The field *RXN_CONDITIONS* was created to allow searches for general types of reaction conditions without having to include specific molecules or data values in the query. It is based on a fixed list of terms including entries such as "Strong Organic Acids", "Small Nucleophiles", "Protic, Non-Aqueous", or "Thermal".

RXN:VARIATION:STEPNO:LIGAND_LINKAGE:STATUS
RXN:VARIATION:STEPNO:LIGAND_LINKAGE:GROUP
RXN:VARIATION:STEPNO:POLYMER_LINKAGE:STATUS
RXN:VARIATION:STEPNO:POLYMER_LINKAGE:GROUP
RXN:VARIATION:STEPNO:SOLID_SUPPORT:MOLEXTREG
RXN:VARIATION:STEPNO:SOLID_SUPPORT:LINKER
RXN:VARIATION:STEPNO:SOLID_SUPPORT_NAME
RXN:VARIATION:STEPNO:POLYMER:MOLEXTREG
RXN:VARIATION:STEPNO:POLYMER:LOADING
RXN:VARIATION:STEPNO:POLYMER_NAME
RXN:VARIATION:STEPNO:POLYMER_DATA

The above list of datafields describe various aspects of the solid support used in a particular reaction. The two *MOLEXTREG* fields refer to a molecule datafield with the same name (cf. chapter 1.5) and links the reaction to a "generic molecule", i.e. a general structure of a polymer or solid support containing a pseudoatom "R" instead of a specific ligand.

Datafields with identical names are also used in the case of the *LIGAND_LINKAGE* (i.e., the functional group connecting the linker and the ligand) and the *POLYMER_LINKAGE* (i.e., the functional group connecting the linker to the polymer). The functional group is defined by the *GROUP* entries (e.g., "carboxylic ester" or "ether, benzylic"), and the *STATUS* term indicates the stability of the linkage ("linkage", "stable" or "cleavage") under the reaction conditions.

The field *LINKER* shows the presence ("yes") or absence ("no") of a linker in a solid support. If it is present, the data in *LIGAND_LINKAGE*/*POLYMER_LINKAGE* and in *POLYMER:MOLEXTREG*/*SOLID_SUPPORT:MOLEXTREG* will be different - otherwise, they are identical.

All of these datafields are based on a systematic classification of polymers, solid supports, code assignments and linkages specifically devised for the SPORE database. The basic concepts and ideas behind the classification strategy are described in the accompanying paper, "Classification System for Solid Supports Used in Solid-Phase Synthesis", which also includes various specific examples for the entries in these datafields.

The remaining datafields of the list given above contain text entries. They describe the polymer and solid support by appropriate names in *POLYMER_NAME* and *SOLID_SUPPORT_NAME*, respectively, and they specify the *LOADING* of the solid support and any other *POLYMER_DATA* which may be relevant. The entries for the names of the polymers and solid supports are not arbitrary but are assigned according to a special format, which is described in the accompanying paper "Polymer- and Solid-Support-Names; MKeywords".

RXN:VARIATION:STEPNO:PROT_GP:MOLEXTREG
RXN:VARIATION:STEPNO:PROT_GP:STATUS
RXN:VARIATION:STEPNO:PROT_GP:FUNC_GROUP

These datafields refer to protecting groups present in the reactant or product of the reaction. Again, the *MOLEXTREG* entry includes a link to a generic molecule which gives a general representation of the particual protecting group and contains the pseudoatom "R" instead of a specific protected molecule.

The functional group protected by this protecting group is specified by the field *FUNC_GROUP* (e.g., "amine, secondary" or "1,3-diol"), and the field *STATUS* describes the stability of the protection in this reaction step ("protection", "stable" or "deprotection").

RXN:VARIATION:STEPNO:ENCODING:TAG
RXN:VARIATION:STEPNO:ENCODING:COMMENTS

In some combinatorial strategies, special encoding/decoding methods are used to identify individual library members attached to the polymer beads after a mix-and-split synthesis sequence. The application of such a strategy in a particular reaction is indicated by an entry "yes" in the *TAG* field, and the encoding method used is described by the text in the *COMMENTS* field.

2.5 Molecule Datafields

MOL:SYMBOL
MOL:MKEYWORDS
MOL:MOLECULAR.WEIGHT
MOL:CAS
MOL:MOLEXTREG
MOL:REFERENCE:NOTES
MOL:REFERENCE:MREFNO
MOL:REFERENCE:MLITTEXT
MOL:REFERENCE:MAUTHOR
MOL:REFERENCE:MJOURNAL:MVOL.
MOL:REFERENCE:MJOURNAL:MNO.
MOL:REFERENCE:MJOURNAL:MPG.
MOL:REFERENCE:MJOURNAL:MYEAR
MOL:REFERENCE:MJOURNAL:MCODEN
MOL:REFERENCE:MJOURNAL:MJRNL

The datafields used for the molecules in the SPORE database are partly identical to those present in other databases, which also specify the *MOLECULAR.WEIGHT* of a compound, possibly include a *CAS* Registry Number, and give one or more names for this molecule in the *SYMBOL* field. The molecule datafields, however, also include a number of features which are unique to the SPORE database.

The preceding chapter already mentioned the so-called "generic molecules", which are included in the database as generalized structural representations of polymer supports, solid supports with linkers, and protecting groups, and which incorporate the pseudoatom "R" instead of specific structures of ligands or protected compounds. Each of these generic molecules is equipped with a unique *MOLEXTREG,* i.e. a specific code in the case of polymeric supports (cf. the accompanying paper "Classification System for Solid Supports Used in Solid-Phase Synthesis"), and usually a common abbreviation in the case of protecting groups (such as "Fmoc"or "4-Cl-Bn"). Furthermore, all names assigned to the *SYMBOL* field of generic polymers and supports start with the text string "Generic: " in order to avoid the presence of molecules with duplicate names in the database - thus, only the "real" support has the entry "Wang resin" in the *SYMBOL* field, while the corresponding generic structure has the entry "Generic: Wang resin".

Special entries for generic molecules can also be found in the datafield *MKEYWORDS*, which contains (computer-generated) structural keywords for all molecules of the database (e.g., "alcohol", "pyrrolidine", "transition metal"). In the case of polymer-supported compounds, keywords describing the polymeric framework (e.g., "PS", "MPEG", "glass", "polyethylene imine") are present as well, and *MKEYWORDS* such as "generic structure", "linker" or "protecting group" are assigned to generic molecules. A list of the SPORE-specific *MKEYWORDS* is included in the accompanying paper "Polymer- and Solid-Support-Names; MKeywords".

The *REFERENCE* branch of datafields contains *NOTES* as the central piece of information, whereas the contents of all other datafields are taken from the corresponding reaction data of that particular abstract (i.e., from the datafields not containing the leading 'M', with *MREFNO*, e.g., containing the *REFNO* of the abstract; cf. chapter 2.1).

The *NOTES* field gives various textual comments to many molecules of the database and mentions, e.g., pharmacological properties of a lead structure or possible applications of a polymer-supported reagent. Most importantly, however, all generic molecules contain suitable commentary in this field, which specifies the uses and characteristics of the support or protecting group. The number of entries can range from one single comment to a collection of diverse remarks gathered from various abstracts (Wang resin and the oxygen-linked Merrifield-type support, e.g., contain far more than 50 *REFERENCE* entries).