Chem. Listy *108*, 81–106 (2014) Bulletin

POROVNÁNÍ NEJDŮLEŽITĚJŠÍCH CHEMICKÝCH BÁZÍ DAT – PROGRAMU SCIFINDER A DATABÁZOVÉHO SYSTÉMU REAXYS

JAROSLAV ŠILHÁNEK

Ústav organické technologie, Vysoká škola chemicko-technologická v Praze, Technická 5, 166 28 Praha 6
silhanek@vscht.cz

Došlo 12.7.13, přijato 11.10.13.

Klíčová slova: chemické báze dat, SciFinder, Reaxys, Beilstein, Gmelin, Patent Chemistry Database, reakční báze dat

Obsah

1. Úvod
2. Inherentní vlastnosti chemických bází dat
   2.1. Systém bází Chemical Abstracts – program SciFinder
   2.2. Systém Reaxys – báze dat Beilstein, Gmelin a Patent Chemistry Database
3. Experimentální studie rozdílů mezi nejdůležitějšími chemickými bázemi dat
   3.1. Rešerše využívající textové předmětové formulace
   3.2. Porovnání rešerší zaměřených na nalezení chemických sloučenin
   3.3. Vyhledávání chemických reakcí
4. Vyhledávání fyzikálně-chemických dat
5. Patentové rešerše
6. Závěr

1. Úvod

Existence počítačových chemických bází dat je už považována za naprostou samozřejmost s jistým povědomím, že je to drahé, a že občas hrozí ukončení přístupu pro nedostatek finančních prostředků. Stejně tak zůstává v povědomí chemiků, že nejdůležitější chemická báze dat zůstává Chemical Abstracts (CA), i když se skrývá pod ochrannou značkou pro počítačovou aplikaci SciFinder, s jejíž pomocí s informacemi z Chemical Abstracts pracujeme. A obdobně organičtí a anorganičtí chemikové vědí, že klasické „Handbuchy“ Beilstein a Gmelin existují jako báze dat, dnes provozované vydavatelstvím Elsevier pod jinou komerční značkou Reaxys doplněné ještě bází dat Patent Chemistry Database. S ohledem na vysoké náklady pro zpřístupnění právě těchto klíčových chemických bází dat, z celorepublikového hlediska v řádech desítek milionů, vyvstává otázka, do jaké míry se tyto systémy překrývají či doplňují a zda má smysl investovat do jednoho či druhého nebo do obou těchto databázových systémů. Tato práce se především snaží zdůraznit rozdílné inherentní výchozí charakteristiky obou systémů a na reálných příkladech je potvrdit a ilustrovat a alespoň orientačně na omezeném souboru konkrétních příkladů dokumentovat rozdíly v efektivitě, rozsahu i relevanci výsledných rešerší.

2. Inherentní charakteristiky chemických bází dat

2.1. Systém bází Chemical Abstracts – program SciFinder[1]

Název SciFinder je počítačová aplikace, která umožňuje pracovat s veškerým digitalizovaným materiálem, který producent, Chemical Abstracts Service (CAS), součást Americké chemické společnosti, od r. 1907 nashromáždil a dále shromažďuje. Budiž velmi zdůrazněno, že koncepce zpracovávání primárních informačních zdrojů je i v případě elektronické verze stále stejná a je založena na „ručním“, tedy intelektuálním zpracování článku nebo patentu vysoce kvalifikovaným specialistou do podoby záznamu obsahujícím vedle bibliografické identifikace stručný, ale odborně fundovaný věcný popis obsahu daného dokumentu formou předmětových hesel a výčet všech chemických sloučenin v dané práci studovaných. Takže i elektronická báze dat je především bibliografickou bází dat, jejíž funkcí je odkazovat na primární zdroje, které by měly obsahovat hledané informace. Vzhledem k zásadě přijaté na samém počátku této činnosti v r. 1907, že totiž z každé zpracované publikace jsou do vzorcových rejstříků zaevidovány všechny v dané publikaci studované nebo vůbec použité chemické sloučeniny, představují tyto rejstříky dnes v elektronické formě pod označením REGISTRY nejautoritativnější seznam všech známých, tedy popsaných chemických sloučenin. Poměrně nedávno byl tento systém bází doplněn o bázi CASREACT, což je samostatně vytvářená báze chemických reakcí a informace o existenci chemických sloučenin je doplňována o jejich vlastnostech převzatých z původních zdrojů a dále je vytvářena báze dat komerčně dostupných sloučenin a báze právních předpisů regulujících jakékoliv manipulace s chemickými látkami. Proto hovoříme o systému bází CA zpřístupňovaných programem SciFinder.

2.2. Systém Reaxys[2] – báze dat Beilstein, Gmelin a Patent Chemistry Database[3]

Báze dat Beilstein[1] a Gmelin[1], které představují základ dnešního systému Reaxys, jsou naproti tomu přede-

vším digitalizovanou verzí původních tištěných „Handbuchů“, ovšem doplněnou o období v tištěných svazcích nezpracované. Jak v tištěné podobě, tak i jako elektronické báze dat jsou díla Beilstein a Gmelin vytvářena jako soupis všech známých sloučenin dané kategorie, způsobů jejich vzniku nebo přeměn a všech jejich vlastností do konkrétní doby a nikoliv jako průběžně doplňovaný zdroj bibliografických informací. Na obr. 1 je uveden typický záznam sloučeniny v kompendiu Beilstein, v bázi dat je takový záznam doplněn především o grafickou reprezentaci v něm obsažených informací.

Je evidentní, že kromě zmínek o reakční teplotě nebo fyzikálně-chemických veličinách, nejsou uvedeny žádné další údaje a nemůžeme je proto ani vyhledávat. Ale obdobně jako producent bází dat CA, tak i vydavatelství Elsevier se snaží tuto původní koncepci seznamu chemických sloučenin a jejich vlastností rozšířit do oblasti bibliografických bází dat. Děje se to jednak doplňováním názvů a abstraktů článků, tak i propojováním báze Reaxys s dalšími bázemi, které nakladatelství Elsevier provozuje – především bází Scopus.

Významný je dále i rozdíl v počtu zpracovávaných primárních publikací. CA se od samého počátku snažila zachytit a zpracovat chemické i chemicko-inženýrské informace v nejširším možném rozsahu oborovém i regionálním včetně patentů. Naproti tomu materiál do děl Beilstein i Gmelin byl vybírán především z časopisů zaměřených na chemii organickou, resp. anorganickou, tedy na principiálně omezený okruh titulů. A i když jej nakladatelství Elsevier postupně rozšiřuje, okruh titulů zpracovaných do formátu bází Beilstein a Gmelin je stále menší než v případě CA.

3. Experimentální studie rozdílů mezi nejdůležitějšími chemickými bázemi dat

V zásadě pracujeme s oběma nejdůležitějšími chemickými bázemi dat analogicky, jako s většinou ostatních elektronických bází dat kromě jedné specifické výjimky, kterou je možnost formulace strukturních dotazů graficky. Takže buď hledáme relevantní informace zadáváním předmětových hesel, klíčových slov nebo jiných textově vyjádřených problémů – včetně slovní identifikace struktury chemických sloučenin triviálním nebo systematickým názvem, nebo použijeme strukturní editory pro zadání struktury sloučeniny. Tomu odpovídá i nabídka rešeršních strategií, která je v obou případech prakticky totožná a skrývá se pod označením *Explore Topics* v programu SciFinder nebo *Literature* v systému Reaxys, kde pracujeme především s předmětovými hesly, jmény autorů, apod. Hledáme-li konkrétní sloučeninu nebo skupinu sloučenin a jejich přeměn, vycházíme z *Explore Substances*, resp. *Explore Reactions* v programu SciFinder a *Reactions* v systému Reaxys, a to nejčastěji s využitím tzv. strukturních editorů. Hledáme-li rozdíly v rešeršním potenciálu mezi oběma bázemi, je účelné je hledat při jednotlivých rešeršních strategiích.

3.1. Rešerše využívající předmětová hesla – *Search Topics* nebo *Literature*

Mějme za úkol vyhledat relevantní práce o vlivu rutheniových katalyzátorů na stereoselektivitu hydrogenačních reakcí, pro což se nabízejí následující předmětová hesla: *hydrogenation, stereoselective, ruthenium* a *catalyst*.

4-Brom-1-acetoxy-benzol, Essigsäure-[4-brom-phenylester], [4-Brom-phenyl]-acetat
$C_8H_7BrO_2$, Formel XVIII (R = CO-CH_3) (H 200). *B.* Aus 4-Brom-phenol und Acetylchlorid (KLARMANN, Mitarb., *Am. Soc.* **55** [1933] 4657, 4661). Durch Erwärmen von Essigsäure-phenylester mit *N*-Brom-succinimid in Tetrachlormethan (BUU-HOI, *A.* **556** [1944] 1, 7). — Kp_2: 100° (KL., Mitarb.). — Beim Erhitzen mit $AlCl_3$ auf 150–160° entsteht 5-Brom-2-hydroxy-acetophenon (KL., Mitarb.).

Obr. 1. **Ilustrace záznamu sloučeniny v tištěném svazku Beilsteinova kompendia,** Beilstein Handbook, **6** III 747. Přetištěno se svolením Beilsteinova Institutu

Tabulka I
Výsledky rešerše problematiky stereoselektivní hydrogenace na rutheniových katalyzátorech v programu SciFinder

29 references were found containing **"Stereoselective hydrogenation with ruthenium catalyst"** as entered.	29^{10}
1009 references were found containing the two concepts **"Stereoselective hydrogenation"** and **"ruthenium catalyst"** closely associated with one another.	1009
1653 references were found where the two concepts **"Stereoselective hydrogenation"** and **"ruthenium catalyst"** were present anywhere in the reference	1653

SciFinder nám doporučuje formulovat dotaz jako větu: *Stereoselective hydrogenation with ruthenium catalysts*, což poskytne výsledek sumarizovaný v tab. I.

Systém Reaxys na stejný dotaz v maximálním rozsahu, tedy:

stereoselective AND hydrogenation AND ruthenium AND catalysts*

poskytne jen 34 odkazů, což svádí k závěru o shodě s nejvíce relevantním souborem s počtem 29 odkazů z SF, ale diametrálně se lišící s množinou *„closely associated“*, která obsahuje 1019 odkazů a už vůbec nesrovnatelné s maximální odpovědí v počtu 1653 citací. Pokud zcela opustíme dotaz na chemickou entitu, tedy prvek ruthenium, a budeme hledat informace o analýze fenolů v odpadních vodách, dostaneme na dotaz *Analysis of phenols in waste water* 1512 odkazů charakterizovaných jako *closely associated* v programu SciFinder a 934 odkazů na obdobný dotaz (analysis AND phenol* AND waste AND water) v Reaxysu. A ještě daleko výraznější rozdíl je, že soubor citací z SciFinderu vede k více než 1000 sloučeninám a reakcím, zatímco soubor odkazů z Reaxysu jen k 59 sloučeninám ve třech reakcích. Čím jsou způsobeny tyto velké rozdíly ve výsledných souborech odkazů?

Především asi zkušenějšího chemika s konkrétnější představou o charakteru obou bází dat příliš nepřekvapí rozdíly ve výsledcích rešerše analýzy odpadních fenolů, protože takovou rešerši by asi nikdy v Beilsteinově či Gmelinově kompendiu nedělal. Ovšem řada nalezených odkazů obsahuje jen název článku a abstrakt a nebyla tedy zpracována jako chemická publikace. Je to ilustrace výše zmíněné snahy rozšířit záběr do oblasti bibliografických zdrojů propojením báze Reaxys např. s bází Scopus.

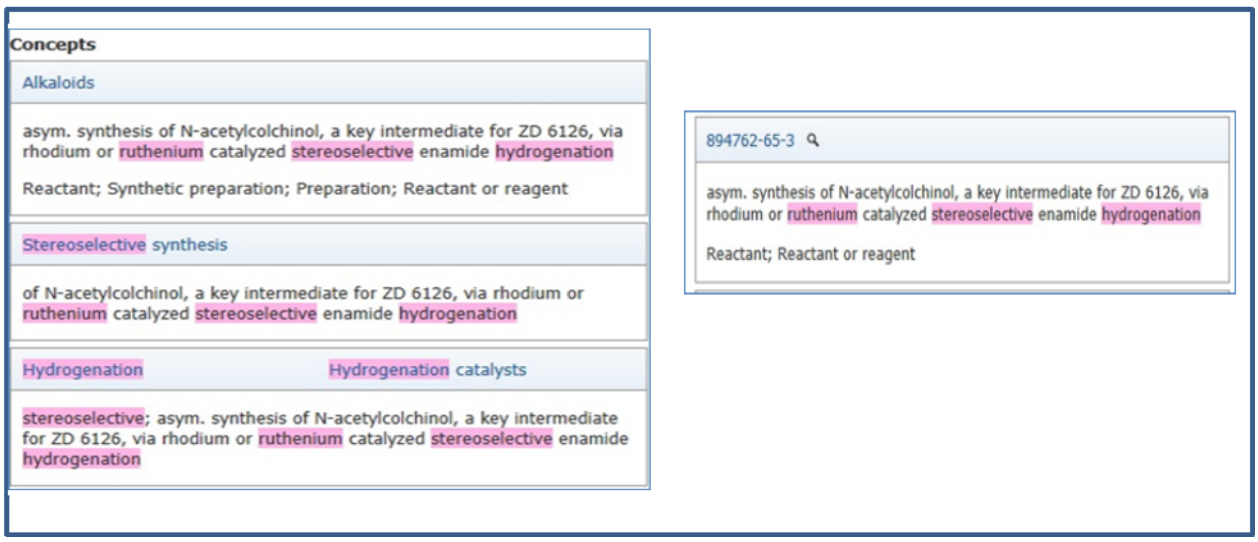
Odpověď na rozdíly ve výsledcích rešerše stereoselektivní hydrogenace poskytne analýza konkrétních dokumentů. Tak např. článek s názvem: *Asymmetric enamide hydrogenation in the synthesis of N-acetylcolchinol: A key intermediate for ZD6126*[4], obsahuje ve svém názvu jen jeden ze zadaných termínů, „*hydrogenation*“, a ani jeho abstrakt (viz obr. 2) ostatní termíny neobsahuje, a přesto byl programem SF nalezen.

Tento článek nebyl nalezen v Reaxysu, přesto, že obsahuje naprosto stejný název i abstrakt. Ostatně v současné době jsou texty abstraktů univerzálně přebírány do bází dat z původních publikací a není proto překvapující, že jsou v obou bázích stejné.

Protože se jedná o nepochybně důležitou chemickou publikaci, její záznam v bázi Reaxys je, a to včetně slouče-

A synthesis of (S)-N-acetylcolchinol I (R = H), a key intermediate in the synthesis of the pharmaceutically useful phosphate prodrug ZD 6126 I (R = PO_3H_2), via catalytic asym. hydrogenation of enamide II was developed. After screening a range of metal and ligand combinations it was found that (S,S)-iPr-FerroTANE Ru(methallyl)$_2$ and [(S,S)-tBuFerroTANE Rh(COD)]BF_4 gave both high enantioselectivity (>90% ee) and high catalyst utility (molar S/C = 1000).

Obr. 2. **Abstrakt článku[4] v originální publikaci a obou bázích dat**



Obr. 3. **Ilustrace rozsáhlého předmětového popisu obsahu článku v CA.** Převzato z programu SciFinder laskavostí Americké chemické společnosti

nin a reakcí, ale nebyl nalezen ve výše zadané rešerši, protože ani jeho název, ani abstrakt neobsahují zadaná hesla. Naproti tomu báze SciFinder je bohatě doplněna předmětovými hesly vytvořenými vysoce kvalifikovanými abstraktory producenta, mezi kterými jsou všechny termíny, které byly použity ve formulaci dotazu (obr. 3).

Hlavní rozdíly ve výsledcích rešerší programem SciFinder a Reaxys spočívají především v rozsáhlém předmětovém popisu dokumentu systémem předmětových hesel a klíčových slov prováděném specialisty producenta, který v systému Reaxys a před tím v bázích Beilstein a Gmelin chybí. I když v bázi Scopus a dalších jsou uváděna klíčová slova nebo další předmětová hesla, jsou výsledkem počítačového zpracování textů abstraktů a maximálně „ručního" zařazení do některé velké oborové skupiny. Ale přesto není rozumné paušálně předmětové rešerše uvedeného typu z Reaxysu odmítat, bližší prozkoumání výše uvedených výsledků rešerší ukáže, že rešerše z Reaxysu není jen podmnožinou souboru ze SciFinderu, ale obsahuje i řadu relevantních odkazů, které program SciFinder nenašel, např. publikace[5,6], což platí i obráceně. Důvody spočívají opět především v rozdílném předmětovém popisu primárního dokumentu, jak je možné se přesvědčit ze záznamů v SciFinder nebo naopak v Reaxysu.

Zajímavé je i porovnání formulace předmětových dotazů v obou bázích dat. Systém Reaxys používá standardní Booleovské operátory včetně operátorů proximitních, jejichž použití je ponecháno na rozhodnutí uživatele, zatímco SciFinder nutí chemika formulovat jeho dotaz větou, tedy včetně předložek, spojek, členů apod., tedy bez operátorů. V žádném případě to ale neznamená, že mají sémantický význam, ale hrají roli při spuštění evidentně sofistikovaného algoritmu, který nikdy nebyl zveřejněn, jen zvnějšku byly ojedinělé snahy mu přijít na kloub[7]. Tato strategie, kterou SciFinder používá, představuje typickou „černou skříňku", která ale překvapivě dobře funguje.

3.2. Rešerše chemických sloučenin – *Search Substances*, nebo *Substances and Properties*

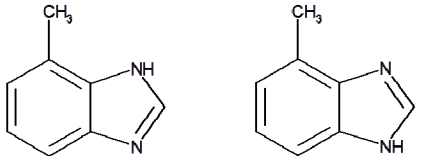
Kromě nakreslení strukturního vzorce umožňují oba systémy i standardní možnost zadání triviálním nebo systematickým názvem a molekulovým (sumárním) vzorcem. Problematičnost názvosloví je všeobecně známá, za zmínku stojí rozsáhlé soubory triviálních a komerčních názvů dříve v CA nepřípustných z důvodů úspory paměti. Sumární vzorce kromě toho, že mohou odpovídat i značně velkému počtu různých sloučenin, vyžadují pozornost v případě sloučenin sestávajících z více složek, typicky anorganických sloučenin. Reaxys akceptuje princip tištěných rejstříků Gmelina, které obsahovaly jak sumární vzorec s abecedně seřazeným výčtem všech prvků bez ohledu na to, zda jsou v kationtu, aniontu či jiné částici, tak i vzorce „rozepsané". Tedy např.

C6FeK3N6 pro hexakyanidoželezitan draselný, ale také formu $[Fe(CN)6]^{(3-)}*3K^{(1+)}=K_3[Fe(CN)6]$, resp. jen $[Fe(CN)6]^{(3-)}*3K^{(1+)}$. Tyto formy jsou pro využití poněkud těžkopádné. SciFinder naproti tomu striktně vyžaduje v molekulových vzorcích samostatné zadávání jednotlivých strukturních jednotek, tedy typicky iontů, které odděluje tečkou jako tzv. „*dot-separated formula*". Tedy ve výše uvedeném příkladu bude sumární vzorec zadávaný pro SciFinder vypadat následovně: C6FeN6.3K. Forma C6FeK3N6 v SciFinderu neposkytne žádnou sloučeninu.

Zadávání struktur organických sloučenin prostřednictvím strukturních editorů je většinou bezproblémové, a to jak v programu SciFinder, tak i v systému Reaxys. Přesto, že Reaxys nabízí možnost nastavení z několika nejčastěji používaných a oblíbených strukturních editorů, je doporučitelné pracovat s implicitně nastaveným editorem, pro který jsou adjustovány různé potenciální strukturní varianty, např. různé strukturní vyjádření nitroskupiny, s náboji či bez nich. Obě báze dat nabídnou po zadání struktury v režimu *Exact* zpravidla řadu příbuzných struktur lišících se přítomností isotopů, nábojů či radikálů apod., kterých bývá u Reaxysu méně a v SciFinderu je jejich nabídka přehlednější. Budeme-li např. hledat informace o 1*H*-benzimidazolu, dostaneme v případě *Exact Search* u SciFinderu pro neutrální uzavřenou strukturu s CAS RN = 51-17-2 9129 odkazů a v Reaxysu 2179 odkazů. Uvážíme-li, že citace v Reaxysu budou více „organické" a v SciFinderu zahrnující širší aplikační oblasti, rozdíly nejsou dramatické a je možné je celkem snadno vysvětlit právě rozsahem zpracovávaných primárních zdrojů. Přesto i zde se můžeme setkat se zajímavými případy. Budeme-li např. hledat 7-methyl-1*H*-benzimidazol, dostaneme v SF informaci pouze o existenci jednoho tautomeru s CAS RN 4887-83-6, a to přesto, že program SciFinder, který tautomery ignoruje, by měl najít současně i druhý předpokládaný tautomer. Báze dat Reaxys naproti tomu umožňuje rozlišit tautomery v zadání nebo je ignorovat, a pokud zvolíme tuto možnost dostaneme jako odpověď obě předpokládané struktury (obr. 4). Sice pro obě uvádí Reaxys totožné CAS RN, tedy 4887-83-6, ale pro každou rozdílné způsoby přípravy a fyzikální data. Ovšem už záznam v CA obsahuje nejistotu ohledně názvu, jako správný je uveden název 7-methyl-1*H*-benzimidazol, ale mezi ostatními názvy jsou uvedeny i alternativně 4-methyl-1*H*-benzimidazol nebo varianta vyjadřující nejistotu – 7(4)-methyl-benzimidazol.

Na úrovni rešerše nejde o to, zda skutečně oba tautomery samostatně existují či nikoliv, ale tento případ ilustruje problematiku převodu dat z primárních zdrojů do databázového záznamu a užitečnost v detailech se lišících zásadách přebírání informací o strukturách.

Výraznější rozdíly nalezneme v případě anorganických sloučenin, což logicky vyplývá z faktu, že součástí Reaxysu je báze dat Gmelin, nejautoritativnější zdroj in-



Obr. 4. **Tautomery 7(4)-methyl-1*H*-benzimidazolu**

formací pro anorganické sloučeniny, i když neúplný. Tak např. při hledání komplexu stříbra s amoniakem akceptují jak Reaxys, tak i SF strukturní formulaci bez označení nábojů, které jsou implicitně doplněny a výsledkem je celá paleta nejrůznější strukturních modifikací, solí i komplexů. V SciFinderu je to 603 různých sloučenin obsahujících komplex Ag(NH3) , v Reaxysu 97. Ovšem hlavní předností Reaxysu jsou bezprostřední odkazy na informace o reakcích a vlastnostech jednotlivých sloučenin. Tak např. pro samotný kation nenabídne SciFinder nic více, než jen odkazy na primární zdroje v počtu 217 citací. Reaxys naproti tomu odkazuje jen na 89 citací, ale na druhou stranu poskytne informace o cca 34 reakcích z toho 20 preparativních. A samozřejmě informace o fyzikálně-chemických datech, např. 18 polí pro faktografická data a 5 různých spekter kromě odkazů na publikace s kvantově-chemickými studiemi. Na druhé straně rešerše z Reaxysu neobsahuje žádný nitrid, který nalezneme až prostřednictvím názvu *silver nitride*.

Není překvapující, že v případě hledání konkrétních chemických sloučenin jsou výsledky v obou zdrojích s jistým nadhledem srovnatelné, ale v každém případě více či méně rozdílné, což je mimo veškerou pochybnost užitečný závěr. SciFinder představuje jistou normu vyplývající z role systému Registračních čísel. Dodejme ale, že CAS RN stále nejsou spolehlivým identifikátorem pro vyhledávání chemické sloučeniny v Reaxysu, protože zatím nedošlo k dohodě mezi CAS a vydavatelstvím Elsevier. V každém případě by měl mít chemik na paměti, že CA registruje chemickou sloučeninu jako jedince, jehož existence je zmíněna v primárním dokumentu a je tudíž spojena s jeho bibliografickým záznamem, zatímco Reaxys, potažmo tedy Beilstein či Gmelin, registrují záznam o sloučenině jako centrum své činnosti včetně informací o jejím vzniku, přeměnách a vlastnostech. O tyto údaje je dnes záznam sloučeniny obohacován v bázi REGISTRY a programem SciFinder propojován s bází CASREACT.

3.3. Rešerše chemických reakcí – *Explore Reaction, Reactions*

Tato možnost se otevřela teprve v elektronických verzích, v tištěných verzích bylo možné vyhledávat informace o reakcích jen jako součást informací o individuálních sloučeninách nebo pomocí jmenných názvů reakcí nebo jiného slovního popisu. Zásadní rozdíl mezi oběma bázemi lze vyjádřit následující charakteristikou:

Chemické reakce jako takové byly v systému bází dat CA zaznamenávány do samostatné báze CASREACT teprve od cca osmdesátých let minulého století, a to ještě z omezeného okruhu disciplín (organická chemie a obory příbuzné).

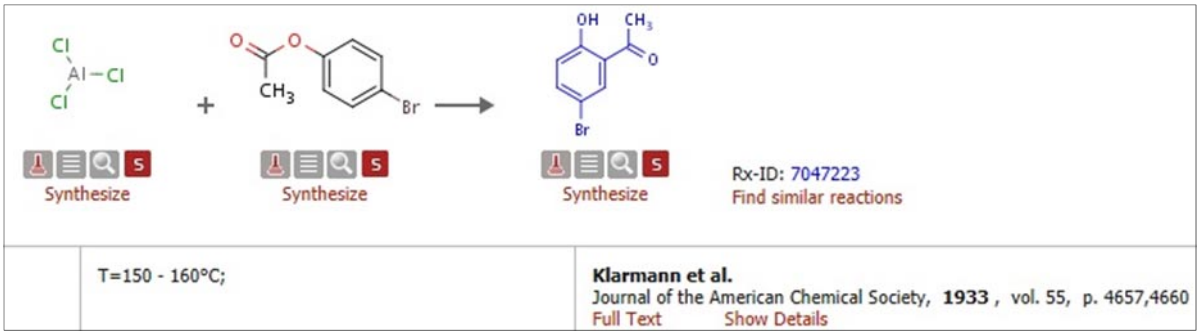
V systému Reaxys jsou přeměny chemických struktur neoddělitelnou součástí bází Beilstein a Gmelin, jinak řečeno, v těchto bázích nemůže být uvedena sloučenina bez jejích reakcí a naopak.

Z toho vyplývá, že programem SciFinder nalezneme využitím *Explore Reactions* jen reakce z posledních cca 30 let a postupně i starších, ale převážně z oblasti organické a příbuzné chemie. Systém Reaxys, jak vyplývá z této charakteristiky, je reakční bází dat v pravém slova smyslu, zatímco SciFinder pracuje se samostatnou reakční bází CASREACT. Odpověď z báze Reaxys bude proto komplexnější, protože zahrnuje reakční informace z daleko širšího časového období, ovšem na druhé straně z daleko menšího počtu primárních zdrojů.

Celkové porovnání obou informačních zdrojů jako reakčních bází dat je proto prakticky nemožné, lze jen porovnávat konkrétní záznamy z překrývajících se primárních zdrojů. Především připomeňme, že reakční informace v tištěných svazcích obou „Handbuchů“ jsou ve formě co nejstručnější věty, jak je ilustrováno v obr. 1. Tak např. věta „*Beim Erhitzen mit $AlCl_3$ auf 150-160^{o}entsteht 5-Brom-2-hydroxyacetophenon (Kl., Mitarb., citace)*“, byla převedena do elektronické báze jako rovnice v záznamu na obr. 5, což je nepochybně slušný programátorský výkon. Kromě názornosti získáváme i propojení na plný text, případně možnost vyhledávat příbuzné reakce.

Na tomto příkladu je možné ilustrovat zásadní problém reakčních bází dat. Uvedená práce ve skutečnosti popisuje přípravu serie 2-alkyl-4-bromfenolů ze 4-bromfenolu acylací, přesmykem na 2-acyl-4-bromfenol a jeho konečnou redukcí. Je to tedy vícestupňová syntéza, což sice z uvedeného záznamu není patrné, ale postupným využíváním nabídek *Synthesis* nebo otevřením plného textu se to dozvíme. Protože ale další zmíněné kroky jsou rovněž součástí dané reakční báze, nabízí se možnost naprogramovat zobrazení celé vícestupňové syntézy jako schématu. Tato cesta je v bázi Reaxys postupně realizována jako nabídka *View Scheme* v případě zatím některých reakcí. SciFinder a báze CASREACT má zobrazování vícestupňových reakcí zabudováno standardně v případě, že v experimentální části publikace jsou uvedeny jen jednotlivé syntetické kroky, které jsou postupně kumulovány do vícestupňových schémat. Takže např. bude-li v publikaci experimentální popis čtyř reakcí vedoucí ke konečnému produktu označenému např. číslem 5, tedy formálně: 1 → 2, 2 → 3, 2 → 4 za alternativních podmínek, 4 → 5, neboli z výchozí sloučeniny 1 jsou postupně připraveny sloučeniny 2,3,4 a konečná sloučenina 5. Tyto reakce jsou samozřejmě převedeny do báze CASREACT individuálně jako jednostupňové, ale navíc jsou vytvořeny i jejich vícestupňové kombinace. Nejdříve tedy tři 2-stupňové reakce: 2 → 4 → 5; 1 → 2 → 4; 1 → 2 → 3; a konečně jedna 3-stupňová reakce: 1 → 2 → 4 → 5, protože produkt 3 není využit k žádnému dalšímu kroku.

Systém Reaxys zatím takto důsledný není, ale postupně je tímto směrem rozšiřován. Tak např. práce Wernera a spol.[8] je zpracována v bázi CASREACT do 119 reakcí, zatímco v Reaxysu jen do 41 reakcí, rozdíl jde na vrub rozsahu kombinací doplněných individuálních kroků vícestupňových reakcí. V každém případě možnost zobrazit vícestupňovou syntézu jako celkové reakční schéma ukáže na první pohled, jaké jsou výchozí sloučeniny a následné



Obr. 5. **Ilustrace převedení popisu reakce z tištěného díla Beilstein**

meziprodukty. Kromě toho jak SF, tak i Reaxys ve svých současných verzích nabízejí nástroje pro retrosyntetické studie (*Synthesis Planner*), Reaxys je začíná používat právě pro znázornění vícestupňových reakcí.

Nejvýraznější rozdíly při hledání chemických reakcí, a to ve prospěch Reaxysu, najdeme podle očekávání u anorganických sloučenin. Hlavním důvodem je skutečnost, že CAS nezpracovává informace o přípravách a reakcích anorganických sloučenin do báze CASREACT. Není pak překvapující, že při pokusu vyhledat programem SciFinder, jak připravit kyselinu tetrathionovou nenalezneme pod nabídkou „*Get Reaction*“ buď vůbec žádný odkaz, nebo jen výjimečné případy zpravidla nesouvisející s problémem, který hledáme. Tak např., budeme-li hledat informace o volné či ionizované kyselině tetrathionové nebo jejích solích, v zásadě nenarazíme na problémy názvoslovné (*tetrathionic acid*) nebo problémy s grafickou reprezentací a jak při práci s programem SciFinder, tak i systémem Reaxys se celkem srozumitelně vyrovnáme se strukturními alternativami, jako je volná kyselina, dianion, diradikál apod. SciFinder nabídne takových variant o dvě nebo tři více. Ovšem u všech těchto alternativních struktur dostaneme sice více odkazů na původní publikace, ale pro jednotlivé varianty nedostaneme v programu SciFinder žádnou preparativní reakci. Informace z Reaxysu jsou naproti tomu bohaté, viz následující tab. II.

Analogické výsledky dostaneme i v případě anorganických sloučenin tradičně strukturně formulovaných kovalentně, např. v případě pentabromfosforu. SciFinder nám poskytne sice 5169 reakcí, ale žádná z nich neobsahuje tuto sloučeninu jako produkt. Naproti tomu z Reaxysu dostaneme jen 432 odkazů na reakce této sloučeniny, ale 36 z nich se týkají její přípravy v pravém slova smyslu. Což ovšem neznamená, že se z bází dat CA nedozvíme, jak se pentabromfosfor dá připravit. Tento úkol ale zůstává na úrovni předmětového popisu, tedy práce v režimu *Explore Topics*, např. požadavkem *Preparation of Phosphorus pentabromide*, nebo *Explore Substance*, a dále omezení souboru odkazů jen na takové, které spadají do kategorie *Preparation*. Jedná se *de facto* o standardní postup v tištěných verzích, kde bylo nutné v seznamu upřesňujících hesel u požadované sloučeniny zachytit termín prep. nebo manuf. případně další, které jsou do elektronické báze převedeny z tištěné verze a nadále standardně vytvářeny.

Phosphorus bromide, PBr5, 7789-69-7P
manuf, from Br and PBr3
Preparation

4. Vyhledávání fyzikálně-chemických dat

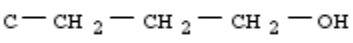
Tato možnost se v užším a realistickém slova smyslu nabízí jen v systému Reaxys, kde ve výchozí volbě je přímo uvedeno *Substances and Properties*. Podobná nabídka, *Property*, se dnes objeví také v programu SciFinder jako další možnost při volbě *Explore Substances*, ale je evidentní, že se jedná jen o první pokus nabídnout vyhledávání numerických dat, a to jak vypočtených, tak i experimentálních, která jsou v poslední době excerpována z primárních zdrojů kromě bibliografického záznamu a chemických sloučenin. V této oblasti je ale stále Reaxys vůči programu

Tabulka II
Výsledky hledání způsobů přípravy kyseliny tetrathionové v bázích SciFinder a Reaxys

	CAS RN	SciFinder	Reaxys
Volná kyselina	13760-29-7	Žádná reakce	64 prep ze 71 reakcí
Její dianion	15536-54-6	1 prep z 8 reakcí	224 prep z 334 reakcí
Draselná sůl	13932-13-3	Žádná prep z 20 reakcí	21 prep ze 75 reakcí

SciFinder bezkonkurenční, díla Beilstein a Gmelin jsou na sběru numerických dat pro každou uvedenou sloučeninu přímo založena. Roztřídění a hierarchický systém faktografických informací v Reaxysu je velmi podobný, ne-li přímo stejný, jako v předchozím nástroji CrossFire a je nepochybně impresivní a užitečný. A to jak pro skutečná numerická data nebo propojení na zdroje např. spekter (v současné době už podobná možnost existuje i z programu SciFinder), tak pro principiální informaci o rozsahu vůbec existujících naměřených dat pro individuální sloučeniny (alternativa dotazu *Exist*). Samozřejmě, problematika fyzikálně-chemických dat je mnohem rozsáhlejší a každý kvalifikovaný chemik se v ní musí orientovat individuálně.

SciFinder sice nenabízí tak názorně strukturovaný systém fyzikálně-chemických dat, ale výsledek takového dotazu je oproti Reaxysu ve výhodě díky zpracovávání daleko širšího spektra primárních zdrojů. Takže např. při hledání odkazů na studie teplotní závislosti hustoty 1-butanolu vede dotaz *Density temperature dependence of 1-butanol* v programu SciFinder k souboru cca 55 odkazů, mezi nimiž je např. odkaz na práci Komarenka a spol.[9], který se v Reaxysu nevyskytuje evidentně proto, že tento časopis nebyl (a asi stále není) do Reaxysu excerpován. SciFinder jej najde na základě detailnějšího věcného popisu pro danou sloučeninu v následující podobě.

- 71-36-3, propertie
  - 

density and viscosity of, temp. dependence of
Properties

Obecně je možné předpokládat, že problematika vyhledávání fyzikálně-chemických dat je mezi profesionálními chemiky natolik známá, že má smysl na tomto místě jen připomenout snad zažité zkušenosti a důvody pro volbu toho či onoho zdroje.

5. Patentové rešerše

Součástí systému Reaxys je báze dat Patent Chemistry Database, relativně nový pokus o vyplnění mezery v nabídce informačních zdrojů[3], protože vzhledem k mimořádně vysokým nákladům na sekundární zpracovávání chemických patentů přestaly být patenty zpracovávány jak do tištěného Beilsteina, tak i do jeho elektronické verze. CA tak zůstal jediným producentem patentových informací jako zdroj pro průmysl i akademickou sféru, mimo čistě komerční a příslušně drahé patentové báze společnosti DERWENT. CAS ve zpracovávání i přes vysoké náklady pokračuje a zpracovává chemické patenty ze 63 zemí, resp. jejich patentových úřadů. Naproti tomu báze Patent Chemistry Database se omezuje jen na evropské a tzv. světové patentové spisy, tedy spisy EU a WO a patentové spisy americké s časovým omezením od r. 1976. Další omezení je na patentové třídy C07 *Organic Chemistry*, A01N *Agrochemicals/Biocides*, A61K *and secondary* IPC C07 *[Medicinal, Dental, Cosmetic Preparations]* a C09B *Dyes*. Oba systémy, program SciFinder i systém Reaxys proto obsahují nástroje pro omezení vybraných souborů jen na patentové spisy a v případě patentů prakticky vždy funguje propojení na plné texty do báze Espacenet nebo báze Amerického patentového úřadu.

Z výše uvedeného je tedy evidentní, že výsledky patentových rešerší z programu SciFinder by měly být úplnější než z báze Reaxys. CAS je si vědoma své vysoké přidané hodnoty v této oblasti, a proto věnuje patentům velkou pozornost. Ale i v záznamech patentů v Reaxysu je patrná snaha o detailnější zpracování a není výjimkou, že záznam patentu je v Reaxysu zpracován z hlediska popsaných reakcí úplněji než v SciFinderu. Na druhé straně předmětový popis patentů v SciFinderu bývá podrobnější. Mimořádně cenné jsou zde informace o patentech z asijských zemí, z japonštiny, čínštiny a korejštiny. Přes své regionální a oborové omezení jsou patentové informace získané z Reaxysu cenné a často s překvapením zjišťujeme, že jeden i druhý systém nezachytily i velmi důležité a problému přesně vyhovující patenty. Další komparativní výhodou Reaxysu je možnost vyhledávat patenty na základě patentového třídění, které SciFinder zatím neumožňuje, ale pravděpodobně brzy nabídne.

6. Závěr

I přes omezený rozsah testovaných rešeršních případů je možné spolehlivě konstatovat, že koncepční zásady, na kterých byly a jsou vytvářeny sekundární informace v bázích CA a Reaxys/Beilstein, Gmelin, jsou stále nejdůležitějším faktorem rešeršních strategií a výrazně ovlivňují výsledky rešerší a to zatím i přes programové doplňování výchozích koncepcí dříve neuváděnými aspekty. Současně ale grafická, v zásadě ovšem počítačová reprezentace struktur a prostředí počítačových bází otevírají dosud nevyčerpané možnosti práce s chemickými informacemi, které se postupně naplňují, a to v konkurenčním prostředí producentů. V každém případě ale platí, že jak rozdílné výchozí koncepce, tak i současná praxe zpracovávání primárních dokumentů u obou nejdůležitějších bází dat vedou k výsledkům, které se s vysokou pravděpodobností nikdy nebudou naprosto překrývat, ale naopak umožní zachytit informace, které by při omezení na jednu nebo druhou bázi dat nebyly nalezeny a využity pro další výzkum.

Tato práce byla realizována jako součást projektu Chemické elektronické informační zdroje pro VaV, č.projektu: CZ.1.05/3.2.00/12.0231, řešeného v rámci programu OP VaVpI prioritní ose 3, podpořeného z prostředků Evropského fondu pro regionální rozvoj a MŠMT ČR.

LITERATURA

1. Šilhánek J.: *Chemická informatika*. Vydavatelství VŠCHT, Praha 2002.
2. http://en.wikipedia.org/wiki/Reaxys, staženo 10.7.2013.
3. http://www.akosgmbh.de/pdf/PCD_Brochure.pdf, staženo 10.7.2013.
4. Lennon I. C., Ramsden J. A., Brear C. J., Broady S. D., Muir J. C.: Tetrahedron Lett. *48*, 4623 (2007).
5. Trindade A., Gois P.M. P., Afonso C. A. M.: Chem. Rev. *109*, 418 (2009).
6. Church T., Andersson P.: Coord. Chem. Rev. *252*, 513 (2008).
7. Wagner B. A.: J. Chem. Inf. Model. *46*, 767 (2006).
8. Werner J., Cerbone L. R., Frank S. A., Ward J. A., Labib P., Tharp-Taylor R. W., Ryan C. W.: J. Org. Chem. *61*, 587 (1996).
9. Komarenko V. G., Manzhelii V. G., Radtsig O. V.: Ukr. Fiz. Zh. (Ukr. Ed.) *12*, 676 (1967).
10. Konkrétní počty nalezených odkazů se vztahují ke konkrétním datu provedení rešerše, což ale v časovém horizontu let v žádném případě neovlivní formulované závěry.

J. Šilhánek *(Department of Organic Technology, Institute of Chemical Technology, Prague):* **Comparisons of the Most Important Chemistry Databases – SciFinder Program and Reaxys Database System**

The posibility of comparing textual description of a problem is a clear advantage of the SciFinder resulting from abundant indexing of the paper content done by specialists in contrast to the description based on the paper title and abstract only. Search for individual compounds affords comparable results for both databases because both use the same principles of excerpting all compounds studied in a given document. In reaction search the advantage of Reaxys is distinct, much like the former databases Beilstein and Gmelin, which are built as reaction databases. The reaction module CASREACT in SciFinder is comparable with the Reaxys in the last 30–40 years only but offers no data for inorganic reactions. The patent search in Chemical Abstracts (CA) with traditional in-depth coverage of patents and in the Reaxys are strong in looking for data of both organic and inorganic compounds. Despite the attempts of both databases at extending coverage and being competitive, the concept based on bibliographic records in CA and compound listing in Beilstein or Gmelin lead to different results, which is advantageous for completeness and relevance of the results.