Chem. Listy *111*, 731−737 (2017) Referát

DATABÁZOVÉ ZDROJE V CHEMII

JINDŘICH JINDŘICH[a,b]

[a] *CZ-OPENSCREEN: Národní infrastruktura pro chemickou biologii, Ústav molekulární genetiky AV ČR v.v.i. Vídeňská 1083, 142 20 Praha 4,* [b] *Katedra organické chemie, Přírodovědecké fakulta Univerzity Karlovy v Praze, Hlavova 2030/8 128 43 Praha 2 jindrich.jindrich@img.cas.cz*

Došlo 24.7.17, přijato 26.9.17.

Klíčová slova: chemické databáze, vyhledávání sloučenin, SciFinder, Reaxys, Web of Science, PubChem, Chem-Spider, Google Scholar

Obsah

1. Úvod
2. Databáze v chemii
3. Komerční chemické databáze
   3.1. Reaxys
   3.2. SciFinder
   3.3. Web of Science
   3.4. Srovnání komerčních databází
4. Volně přístupné chemické databáze
   4.1. Google Scholar
   4.2. ChemSpider
   4.3. PubChem
   4.4. PubMed
   4.5. Ostatní chemické databáze
5. Závěr

1. Úvod

Při dnešním explozivním růstu objemu dostupných informací je schopnost hledat potřebné informace jednou ze základních znalostí, kterou musí každý vědec ovládat. Kromě základních vyhledávacích technik, které již dnes téměř každý používá i v běžném životě, jako je hledání podle klíčových slov v internetovém vyhledávači, existuje v oblasti chemie řada technik, které už tak jednoduché či zřejmé nejsou. Hledání podle chemických identifikátorů[1] je asi nejblíže prostému hledání podle klíčových slov s tím rozdílem, že chemické identifikátory je o něco obtížnější definovat. Ale řada i volně dostupných programů již umí tyto identifikátory vygenerovat, pokud jim umíme zadat strukturu sloučeniny (např. ChemSketch[2] nebo MarvinSketch[3]). To nejobtížnější na hledání informací o chemických sloučeninách nebo jejich chemických přeměnách je pak právě hledání podle struktur či substruktur. Méně běžné je také vyhledávání ve strukturovaných databázích podle chemických, fyzikálních, nebo biologických vlastností chemických sloučenin.

Z komerčních chemických databázových zdrojů jsou nejznámější a nejpoužívanější webové aplikace Reaxys[4], SciFinder[5] a Web of Science[6] (WoS), ve kterých jsou výše uvedené vyhledávací techniky možné a kterým se tento článek bude věnovat podrobněji. Srovnání vyhledávácích možností těchto tří zdrojů bylo v loňském roce publikováno v Journal of Chemical Education[7] a srovnání dvou výrazněji chemicky orientovaných zdrojů – Reaxys a SciFinder – vyšlo před několika lety i v Chemických listech[8]. V Chemických listech byl také popsán důležitý, ale specializovanější, zdroj 3D strukturních informací – Cambridge Structural Database[9].

V tomto článku budou popsány typy databází používaných v chemii a vyhledávací možnosti nejdůležitějších komerčních nástrojů. Bude také podán přehled alternativních volně přístupných webových chemických zdrojů, které mohou v některých případech komerční zdroje nahradit. Na závěr budou shrnuty přednosti i nedostatky popsaných databází a doporučeny metody pro jejich optimální využití.

2. Databáze v chemii

Pro účinné vyhledávání v jakýchkoli databázích je dobré mít povědomí o tom, jak jsou v nich data organizována a vzájemně propojena. Podle typu dat, která obsahují, lze databáze dělit na dvě velké skupiny, které se ale mohou vzájemně prolínat – plnotextové (fulltextové) a strukturované.

Plnotextová databáze je obvykle sada dokumentů, ke kterým jsou vytvářeny indexy umožňující jejich rychlé prohledávání. Tento typ databází obvykle provozují vydavatelé (časopisy, knihy), patentové úřady (patenty) nebo vysoké školy (závěrečné práce studentů). Asi největší databázi na světě tohoto typu ale provozuje Google[10], kde dokumenty jsou webové stránky a ostatní veřejně přístupné soubory na internetu.

Strukturované databáze se naproti tomu obvykle skládají ze sady *tabulek*, které obsahují *záznamy* (nazývaných také *řádky*, anglicky *records* nebo *rows*), které všechny mají stejnou strukturu definovanou sadou *polí* (nazývaných také *sloupce*, anglicky *fields* nebo *columns*). Každému záznamu je vždy přirazen unikátní identifikátor (tzv. *primární klíč*, obvykle v názvu obsahuje „ID“ nebo „Number“), pomocí kterého lze na záznam odkazovat.

Příkladem takového primárního klíče může být hojně používané CAS RN[11] (Chemical Abstracts Service[12] Registry Number), které je primárním klíčem pro strukturu v databázi REGISTRY (viz dále). Podle obsahu lze strukturované databáze dělit na dvě velké skupiny – bibliografické a faktografické.

Bibliografické databáze, jak název napovídá, obsahují informace o publikacích, patentech a podobných dokumentech, čili každý záznam ponese informace o právě jedné publikaci (ale neobsahují zpravidla plný text dokumentu). Typickými poli v bibliografických záznamech jsou – autor, název časopisu, název článku, ročník, rok vydání, stránky apod. Relativně novým parametrem, popisujícím jednoznačně umístění dokumentu na webu, je Digital Object Identifier[13] (DOI). Záznam v bibliografické databázi podává sekundární informaci, tj. odkazuje na primární dokument, který teprve může obsahovat konkrétní fakta.

Faktografické databáze naproti tomu obsahují v záznamech informace o jiných entitách, než jsou dokumenty, tj. obsahují ona konkrétní fakta extrahovaná z primárních dokumentů. V oblasti chemie tedy hlavně fakta o chemických sloučeninách nebo o chemických reakcích, mezi které patří identifikátory sloučenin, fyzikální, reakční, spektrální, toxikologické či bioaktivitní informace.

Ideální databázový systém samozřejmě umožňuje vyhledávat záznamy podle všech hodnot polí a umožňuje vytvářet vyhledávací dotazy za využití logických operátorů. Tento způsob vyhledávání ale vyžaduje podrobné znalosti o struktuře databáze (hlavně názvy polí), syntaxi vyhledávacího jazyka a jisté informatické schopnosti. Proto se snaží producenti databází vytvářet uživatelská rozhraní, která zadávání takových dotazů uživatelům výrazně usnadňuje. Jedna metoda, která je často používána, jsou formuláře již obsahující názvy obvyklých polí daného typu (např. formulář pro „bibliografické údaje“ nebo „fyzikální parametry“). Modifikací předchozí „statické“ formulářové metody jsou formuláře, do kterých lze nová pole dynamicky přidávat podle potřeby (za pomoci rozbalovacích nabídek), případně i s požadovaným logickým operátorem. U formulářů je ale obvykle obtížné zapojit všechny typy logických operátorů nebo je vhodně seskupovat. Další metodou, která už umožňuje konstrukci dotazu jakékoli složitosti, je přímé vytváření vyhledávacího dotazu v textové podobě za pomoci funkcí pro vyhledání názvů polí případně i jejich možných hodnot. Nejpokročilejší uživatelsky přívětivou metodou v tomto směru (kterou zatím nabízí jen nejnovější verze Reaxys) je využití drag&drop[14] pro seskupování libovolných polí do skupin a použití jakékoli kombinace logických operátorů uvnitř skupin a mezi skupinami.

U chemických databází je další metodou, jak tvořit vyhledávací dotaz, zadání chemické struktury nebo substruktury hledané sloučeniny pomocí specializovaného strukturního editoru. U webových aplikací byly řadu let využívány editory instalované do webových prohlížečů jako zásuvný modul – Java applet. Velmi rozšířený strukturní editor tohoto typu byl např. JME editor[15]. Od roku 2015 však začali producenti webových prohlížečů postupně opouštět podporu této technologie (v posledních verzích prohlížečů Firefox a Chrome už Java applety nefungují) a tak museli i producenti chemických databází přejít na editory založené na JavaScriptu, který je zjevně technologií budoucnosti. Mezi nejlepší strukturní editory tohoto typu dnes patří MarvinJS (používaný i v aplikaci Reaxys) od firmy ChemAxon, která se v posledních letech stále více prosazuje v oblasti chemoinformatiky. Z webových stránek této firmy lze získat zdarma řadu programů, z nichž již v úvodu zmíněný editor chemických struktur MarvinSketch[3] je asi nejvíce používaný.

Další, patrně nejpoužívanější program pro kreslení chemických vzorců, je komerční ChemDraw[16] prodávaný původně firmou Cambridgesoft, kterou ale nedávno koupila firma PerkinElmer. Poslední verze tohoto editoru umožňuje nakreslené struktury přímo vyhledat v rozhraní SciFinder.

3. Komerční chemické databáze

V chemických oborech jsou nejčastěji využívány tři komerční webové databázové aplikace – Reaxys, SciFinder a WoS. Z těchto tří jsou si nejvíce podobné co do chemického obsahu a způsobu vyhledávání první dvě (WoS je spíše multioborovou bibliografickou databází). Díky značnému přesahu si proto Reaxys a SciFinder konkurují, což je ovšem pro uživatele výhodné. První zásadní výhodou pro koncové uživatele, se kterou přišel předchůdce Reaxys – CrossFire – v polovině 90. let minulého století, byl jednorázový roční poplatek za přístup k databázím pro instituci. Do té doby se při přístupu k chemické databázi podobného rozsahu, kterou poskytoval CAS, platilo za čas, počet záznamů apod., což neumožňovalo přístup studentům nebo nezkušeným uživatelům. Reakcí CAS na CrossFire byl vznik aplikace SciFinder a obdobného platebního modelu. Od té doby se výrazně rozšířily řady uživatelů a tvůrci těchto systémů se snaží nabízet stále lepší uživatelské prostředí a funkcionalitu. V roce 2008 se z desktopové aplikace SciFinder stala webová aplikace a z desktopové aplikace MDL CrossFire se v roce 2009 stala webová aplikace Reaxys. Obě služby pokračují ve zlepšování nabízených funkcí a interface, obě také zavedly pokročilou aplikaci plánování syntetických strategií. Stále jsou však mezi nimi rozdíly, které stojí za podrobnější rozbor.

3.1. Reaxys

Webová služba Reaxys[4] spojila při svém vzniku[17] roku 2009 dohromady data z databází **Beilstein**, **Gmelin** a **Patent Chemistry Database**. Databáze Beilstein byla vytvořena z údajů v tištěném zdroji Beistein Handbook of Organic Chemistry[18], který byl ve své době nejrozsáhlejším tištěným zdrojem („příručka“ zabírající celou stěnu v knihovně) obsahujícím vlastnosti organických sloučenin a který obsahoval údaje i z těch nejstarších chemických publikací (nejstarší z roku 1771). Databáze Gmelin vychá-

zela z údajů v tištěné Gmelin Handbook[19], obsahující informace o anorganických a organometalických sloučeninách. Patent Chemistry Database byla vytvořena z údajů o patentech z oblastí blízkých chemii v roce 2005 a obsahovala data abstrahovaná z chemických patentů publikovaných v anglickém jazyce od roku 1976.

Z historie jejího vzniku plyne, že databáze Reaxys je výrazně faktograficky zaměřená a obsahuje rozsáhlé informace o přípravě sloučenin a o jejich experimentálně změřených vlastnostech a odkazy na primární literaturu, kde byly dané informace zveřejněny. Databáze je neustále aktualizována údaji o nových sloučeninách. Data pocházela primárně z asi 400 nejdůležitějších chemických časopisů a z chemických patentů. V roce 2013 se počet časopisů, ze kterých jsou data získávána, zvýšil na 16 000. V současnosti obsahuje databáze asi 500 milionů experimentálních vlastností, které lze vyhledávat v 500 různých polích a zahrnují údaje o reakcích, fyzikálně-chemické a spektrální vlastnosti i údaje o biologických aktivitách. Data jsou aktualizována každý týden, ale časová prodleva, která uplyne od publikace a zanesení odpovídajících údajů do databáze, může být v rozmezí dvou týdnů až dvou měsíců.

Data jsou vedena ve třech databázích, vzájemně propojených – Structures (více než 28 milionů záznamů), Reactions (44 mil.), Citations (54 mil.). Vyhledávání v Reaxys je pak možné právě v jednom z těchto tří kontextů a výsledky hledání ukazují záznamy z jedné z těchto databází s odkazy do zbylých dvou. Reaxys prohledává i externí volně dostupné databáze PubChem (pro biologické aktivity sloučenin, bude popsána dále), eMolecules (databáze dodavatelů chemikálií) a LabNetwork (databáze prodávajících a nakupujících – produktů z oblasti vědy).

Největší předností ve srovnání s ostatními databázemi, kterou Reaxys disponuje, je možnost vyhledávání podle hodnot všech (tj. 500) experimentálních vlastností, které jsou v databázi zaneseny. V jiných databázích je obvykle také možné vyhledávat podle vlastností sloučenin, ale vždy jen podle několika málo předem vybraných nejběžnějších vlastností.

I vyhledávání podle struktury nebo reakční přeměny je velmi dobře propracováno. Oproti ostatním databázím nabízí Reaxys širší nabídku strukturních proměnných (Reaxys Generic Groups), které usnadňují substrukturní vyhledávání, i možnost definovat řadu parametrů pro atomy ve struktuře (náboj, isotop, radikál, …).

Video tutoriály k používání Reaxys jsou nejrychlejším způsobem, jak se s ním naučit pracovat[20].

3.2. SciFinder

SciFinder byl firmou CAS uveden na trh roku 1995 jako nástroj (desktopová aplikace) pro hledání chemické literatury. Dnešní webová aplikace poskytuje přístup k většině databází produkovaným firmou CAS a také k volně dostupné bibliografické databázi MEDLINE. Webové rozhraní nabízí přímé prohledávání databází:

CAplus – bibliografická databáze (prohledávána současně s **MEDLINE**[21]), která obsahuje údaje z nejdůležitějších 1500 časopisů z oblastí blízkých chemii[22]. Pro bližší informace o zdrojích, ze kterých CAS čerpá, je k dispozici volně dostupný nástroj CAS Source Index (CASSI)[23], kde lze vyhledávat podle CODEN, ISBN, ISSN, názvu nebo zkratky názvu všechny zdroje použité CAS od roku 1907. Do CAplus jsou také abstrahovány patentové dokumenty z oblastí blízkých chemii z nejdůležitějších devíti patentových úřadů[24]. Pokud v rozhraní SciFinder použijeme některého formuláře ze sekce *REFERENCES*, budou zobrazeny bibliografické záznamy z této databáze. Kromě bibliografických údajů jsou také pro zobrazené záznamy dostupné relevantní odkazy do ostatních CAS databází a ve většině případů i odkazy na plné texty nalezených dokumentů.

CAS REGISTRY – databáze obsahující informace o všech sloučeninách, které byly kdy CAS abstrahovány z literatury. Každé dosud nepopsané sloučenině, která je přidána do databáze, je přirazeno nové CAS Registry Number (CAS RN), které je velmi rozšířeným identifikátorem chemických sloučenin používaným často i v katalozích prodejců chemikálií. Pro nejběžnější chemikálie (kolem 7900) je možné CAS RN dohledat podle jména sloučeniny (nebo naopak podle CAS RN dohledat jméno) ve volně přístupné webové aplikaci provozované CAS – Common Chemistry[25]. Pokud v rozhraní SciFinder použijeme některého formuláře či strukturního editoru ze sekce *SUBSTANCES*, budou zobrazeny výsledky právě z této databáze. V každém nalezeném záznamu jsou pak kromě informací o sloučenině uvedeny i relevantní odkazy do ostatních CAS databází. Momentálně obsahuje CAS REGISTRY více než 130 milionů záznamů o organických a anorganických sloučeninách a více než 67 milionů DNA a proteinových sekvencí. Každý den je do databáze přidáno kolem 15 tisíc nových sloučenin.

CASREACT – databáze s informacemi o chemických reakcích, které byly publikovány v některém z více než stovky klíčových časopisů. Většina reakcí je z publikací od roku 1985, ale lze najít i záznamy publikované už v roce 1840. Pokud je v rozhraní SciFinder použita sekce *REACTIONS*, která umožňuje hledat jen podle údajů zadaných ve strukturním editoru, budou zobrazený záznamy z této databáze. Momentálně obsahuje databáze CASREACT více než 83 milionů záznamů a každý den je přidáno asi 30 tisíc nových reakcí.

Další databáze, které jsou dostupné z odkazů výsledků vyhledávání výše uvedených databází, jsou:

CHEMCATS – databáze dodavatelů chemikálií. U každé nalezené komerčně dostupné sloučeniny je uveden odkaz na její dodavatele, který vede do této databáze (momentálně (rok 2017) má podobu obrázku Erlenmeyerovy baňky s cenovkou). Lze v ní ale nalézt jen ty dodavatele, kteří se zapojí do CHEMCATS programu firmy CAS. I tak je často možné nalézt důvěryhodného dodavatele s výrazně výhodnější cenou, než nabízí obvyklí dodavatelé. Navíc je aktuální cena chemikálie poměrně často odlišná od té, která je uvedena v SciFinderu.

CHEMLIST – databáze regulovaných chemikálií. Zahrnuje chemikálie, které se vyskytují na nějakém sezna-

mu regulovaných chemikálií (toxických, nebezpečných apod.). Pokud je sloučenina nalezená v databázi CAS REGISTRY obsažena i v databázi CHEMLIST, pak lze patřičné odkazy nalézt v záznamu sloučeniny v sekci *REGULATORY INFORMATION*. Momentálně obsahuje tato databáze více než 348 000 záznamů a každý týden je přidáno asi 50 nových látek.

Aktuální způsob základního ovládání rozhraní SciFinder lze nejrychleji nastudovat za pomoci online tutoriálu[26] na výukových stránkách CAS. K dispozici je i tutoriál pro kreslení chemických struktur[27].

3.3. Web of Science

Web of Science (WoS), dříve nazývaný ISI Web of Knowledge, navazuje na služby poskytované firmou Institute of Scientific Information (ISI), která byla založena v roce 1960. Její zakladatel, Eugene Garfield (1926–2017), je považován za otce citačních databází a bibliometrické analýzy. **Science Citation Index**[28] (SCI) byla první databáze tohoto typu, která vyšla z tištěného zdroje Current Contents. WoS vznikl v roce 1997 poté, co firma Thomson Reuters koupila firmu ISI. Od roku 2017 je provozovatelem WoS firma Clarivate Analytics, která byla dříve součástí firmy Thomson Reuters.

Dnes WoS umožňuje přístup k databázím **Science Citation Index Expanded** (SCIE), **Art and Humanitites Index** (AHCI) a **Social Sciences Citation Index** (SSCI), případně podle předplatného i k indexům konferenčních příspěvků a knih. Je to tedy skutečně multidisciplinární citační databáze. WoS aktualizuje obsah svých citačních databází indexací obsahů nově vyšlých článků z vybraných vědeckých časopisů. Tyto časopisy patří do tzv. Web of Science Core Collection (WoS CC), která zahrnuje kolem 12 tisíc časopisů[29]. Tento seznam je ale průběžně aktualizován podle daných kritérií na kvalitu[30].

Další databází, která je dostupná v rozhraní WoS, je **Journal Citation Report** (JCR). Ta je důležitým zdrojem pro hodnocení vědecké kvality časopisů – zde najdeme dnes tak intenzivně diskutované impakt faktory. Obecně lze říci, že časopisy, které najdeme v JCR, patří mezi ty lepší (mají impakt faktor) a je rozumné publikovat právě v nich. Dnes, v době predátorských časopisů, které vydají cokoli bez ohledu na vědeckou kvalitu, jen když zaplatíte, je to vhodný zdroj pro rozhodnutí, zda v příslušném časopise publikovat. Lze říci, že výběr do JCR je ještě o něco přísnější než do WoS CC a aktuálně JCR pracuje s 11 tis. časopisy.

Instituce si mohou připlatit přístup i ke specializovaným chemickým databázím **Current Chemical Reactions** (CCR) a **Index Chemicus** (IC), které také patří do WoS CC. CCR obsahuje přes 1 milion syntetických metod abstrahovaných z více než 100 chemických časopisů od roku 1986. IC poskytuje přístup k více než 2,5 milionu sloučenin abstrahovaných také z více než 100 chemických časopisů od roku 1993. Obě tyto databáze jsou ale řádově menšího rozsahu, než odpovídající databáze v aplikacích SciFinder nebo Reaxys. Navíc má nyní WoS technické problémy v moderních webových prohlížečích s vyhledáváním sloučenin podle struktury, protože používá Java appletový strukturní editor.

WoS vyniká hlavně pokročilými funkcemi pro citační analýzu. Umožňuje počítat H-index – dnes hojně používaný k hodnocení kvality vědců a také umí hledat citace podle více polí, než ostatní databáze. Novější záznamy ve WoS lze tak vyhledávat i podle grantové agentury nebo podle čísla grantu. Je možné také nalezené záznamy exportovat v různých formátech do souboru nebo do webové verze osobní bibliografické databáze EndNote. Pro následný import do jiné bibliografické databáze je vhodné použít pro export strukturovaný formát RIS, který je uznávaným standardem pro tyto účely.

Ovládání WoS je nejsnadnější se naučit za pomoci video tutoriálů[31], které jsou dostupné na stránkách provozovatele.

3.4. Srovnání komerčních databází

Srovnání komerčních databází přináší tabulka I.

4. Volně přístupné chemické databáze

4.1. Google Scholar

Google Scholar[32] je variantou webového vyhledávače Google, dostupnou od roku 2004, ve které se do výsledků začleňují jen dokumenty nebo metadata dokumentů souvisejících s vědeckou činností. Mezi indexované zdroje patří[33] recenzované online dostupné vědecké časopisy, knihy, konferenční příspěvky, diplomové a disertační práce, patenty a další odborná literatura. Indexovány jsou jak zdroje volně přístupné v plném textu, tak i ty nepřístupné, u kterých jsou indexována jen volně přístupná metadata. Podle odhadů je takto dostupných více než 150 milionů dokumentů. Nové dokumenty jsou přidávány několikrát týdně, ale může trvat i několik měsíců až rok či více, než se nový dokument na Google Scholar objeví.

Této volně dostupné službě je řadou vědců dávána přednost i před placenými, jako je např. WoS, díky propracovanému zpracování přirozeného jazyka, ve kterém Google exceluje. K placeným nástrojům je ale vhodné se uchýlit, pokud je potřeba použít strukturovaný dotaz (podle hodnot polí v bibliografickém záznamu) nebo je třeba najít nedávno vydané dokumenty.

Kromě vyhledávání nabízí Google Scholar uživatelům i možnost tvořit si osobní profil se seznamem vlastních publikací a vytváří citační statistiky a počítá H-index podobně jako WoS.

4.2. ChemSpider

ChemSpider[34] je volně dostupná databáze chemických sloučenin nabízející rychlé vyhledávání podle textových nebo strukturních dotazů. Aktuálně popisuje přes 58 milionů sloučenin, o nichž získává informace asi

Tabulka I
Srovnání komerčních databází

Databáze	**Reaxys**	**SciFinder**	**Web of Science**
Adresa aplikace	www.reaxys.org	scifinder.cas.org	webofknowledge.com
Poskytovatel	Elsevier www.elsevier.com	Chemical Abstract Services (CAS) www.cas.org	Clarivate Analytics clarivate.com
Zprostředkovává přístup k interním databázím	Reaxys Reactions (44 mil) Substances (29 mil) Citations (54 mil) experimentální fakta (500 mil)	CAS REGISTRY: sloučeniny (130 mil sloucenin, 67 mil sekvencí) CAplus: bibliografie (436 mil) CASREACT: reakce (83 mil) CHEMCATS: dodavatelé chemikálií CHEMLIST: regulované chemikálie (384 tis) MARPAT: struktury v patentech	Science Cit.Index Expanded (SCIE) Social Sciences Cit.Index (SSCI) Arts & Humanities Cit.Index(AHCI) Current Chemical Reactions (CCR-EXPANDED) (1 mil) Index Chemicus (IC) (2.5 mil) celkem více než 100 mil. Záznamů
Využívá externí databáze	PubChem (91 mil) eMolecules (10 mil) LabNetwork	Medline (23 mil)	Google Scholar
Vychází z historických tištěných zdrojů	Beilstein Hanbook Gmelin Handbook	Chemical Abstracts (1907–2009)	Current Contents
Dnešní primární zdroje dat	• odborné časopisy – chemie (4500) • s chemií spojená periodika, abstrakty konferencí (>16 000) • knihy • patenty	• odborné časopisy – chemie • knihy • patenty • konferenční sborníky, abstrakty • chemické katalogy • disertace	• odborné časopisy • knihy • konferenční sborníky, abstrakty
Typ dat	bibliografické i faktografické DB • bibliografická data • chemická data (reakce, přípravy) • životní prostředí • eperimentální fyzikální data • experimentální spektrální data	bibliografické i faktografické DB • bibliografická data • chemická data (reakce, přípravy) • experimentální fyzikální data • experimentální spektrální data • vypočítaná fyzikální data • vypočítaná spektrální data	biliografické DB, omezená chemická faktografická DB • pokročilá bibliografická data • bibliometrická data
Data z oblastí vědy	chemie, medicinální chemie	chemie, přírodní vědy	multidisciplinární – přírodní vědy, sociální vědy, humanitní vědy
Možno vyhledávat podle	• přirozený jazyk • klíčová slova • struktura (sloučeniny, reakce) • bibliografické údaje • **všechny** experimentální fyzikální, spektrální, biologické vlastností sloučenin	• přirozený jazyk • klíčová slova • struktura (sloučeniny, reakce) • bibliografické údaje • 13 experimentálních, 21 vypočítaných fyzikálních a biologických vlastností sloučenin	• klíčová slova • bibliografické údaje • struktura/vlastnosti sloučeniny (omezené množství parametrů, technické problémy se strukturním editorem)
Max. počet záznamů pro export	5000	100	500
První data z roku	1771	1800	1900
Unikátní vlastnosti a vyhledávací možnosti	• Do detailu propracované vyhledávání podle hodnot všech 500 různých polí za využití logických operátorů. Velice vhodné pro vyhledávání sloučenin podle jejich parametrů (fyzikálních, biologických vlastností). • Obsahuje pouze experimentálně stanovené vlastnosti sloučenin (ne vypočítané). • Možnosti exportu nalezených dat jsou co do množství záznamů a možných formátů nejpokročilejší.	• Nejlépe propracovaný algoritmus pro vyhledávání přirozeným jazykem – lexikální analýza, automatický převod na termíny databázově počítačové. • Kromě experimentálních vlastností sloučenin obsahují jejich záznamy i vypočítané (predikované) vlastnosti. • Vyhledávání sloučenin podle vlastností je omezeno jen na 13 experimentálních a 21 vypočítaných vlastností.	• Rozsáhlé nástroje pro citační analýzu (H-index), snadný export citací, pokročilé vyhledávání podle bibliografických údajů (pracoviště, grantová agentura, číslo grantu). • Omezené možnosti pro hledání prodle struktury (technické problémy se strukturním editorem, malý počet záznamů)

z 500 zdrojů, na které se zpětně odkazuje. Poskytovatelem této služby je od roku 2009 Royal Society of Chemistry (RSC), což dává naději na další pozitivní vývoj této velmi užitečné služby.

Pro nalezenou sloučeninu nabízí informace značného rozsahu – všechny její myslitelné názvy a identifikátory (standardní i nestandardní), experimentální i vypočítané fyzikálně-chemické vlastnosti, údaje o toxicitě a biologické aktivitě, spektra (NMR, IČ, MS, UV-Vis), seznam prodejců, odkazy na publikace, patenty apod. Které informace jsou k dispozici, samozřejmě záleží na tom, co se podařilo získat z původních zdrojů, odkazy na něž jsou rovněž k dispozici.

ChemSpider si klade nelehké úkoly pro svou další činnost:

- Získat na jedno centrální místo informace o všech sloučeninách dostupných na webu, umožnit jejich snadné vyhledávání a standardizovat jejich struktury a názvy.
- Zlepšit kvalitu veřejných chemických zdrojů za využití automatizované kontroly struktury a také manuálních úprav spolupracujících expertů.
- Poskytnout platformu pro vkládání a uchovávání dat. Registrovaní uživatelé mohou uploadovat vlastní struktury a informace o nich a mohou publikovat reakce na ChemSpider SyntheticPages[35], což je volně dostupná databáze chemických reakcí s plnými texty experimentálních procedur tvořená přímo uživateli.
- Snažit se o usnadnění přístupu ke všem datům za využití webového rozhraní optimalizovaného pro mobilní zařízení, mobilních aplikací, webových služeb pro zpřístupnění dat.
- Integrovat data do RSC publikací za využití přímých odkazů a využít validovaných chemických názvů pro vyhledávání v Google Scholar, PubMed a RSC knihách, časopisech a databázích.

To jsou jistě předsevzetí, se kterými nelze než souhlasit a v rámci svých možností je podpořit. Každopádně z toho plyne, že ChemSpider by mělo být to nejvhodnější místo pro nalezení té „nejsprávnější“ struktury i názvu pro hledanou sloučeninu.

4.3. PubChem

PubChem[36] je volně přístupná webová aplikace a databáze založená roku 2004, která poskytuje informace o biologických aktivitách malých molekul. Za malé molekuly se považují látky, které mají méně než tisíc atomů a vazeb. PubChem provozuje National Center for Biotechnology Information (NCBI), které je součástí knihovny National Library of Medicine (NLM). Ta patří do National Institutes of Health (NIH) – největší agentury v USA, která se věnuje výzkumu v oblasti medicíny.

PubChem se skládá ze tří hlavních databází – Substance, Compound a BioAssay, které jsou navzájem propojeny. Compound obsahuje unikátní struktury (každé je přiřazen číselný identifikátor SID) a Substance reálné vzorky, pro které byla zjišťována biologická aktivita (s případnými odkazy do Compound). BioAssay je potom databáze biologických aktivit, které byly zjištěny pro vzorky z databáze Samples.

V databázi lze vyhledávat chemické sloučeniny podle mnoha parametrů včetně jejich (sub)struktury, identifikátorů, názvů či sumárního vzorce. Každý nalezený záznam obsahuje veškeré dostupné informace o sloučenině, včetně její biologické aktivity a odkazů do všech ostatních NCBI databází (např. bibliografické databáze PubMed), ve kterých se daná sloučenina vyskytuje. Od roku 2012 jsou využívána data z PubChem i v aplikaci Reaxys. Dnes je PubChem s více než 91 miliony unikátními strukturami největší volně přístupnou databází chemických sloučenin.

4.4. PubMed

PubMed[37] je volně přístupné webové rozhraní (od roku 1997) určené k vyhledávání záznamů nacházejících se primárně v databázi MEDLINE (23 milionů záznamů). Do této bibliografické databáze se indexují záznamy z časopisů a dalších primárních zdrojů, které souvisí s medicínou. Lze tam tedy najít i informace o publikacích z oblasti medicinální chemie nebo biochemie. PubMed identifier (PMID) je celé číslo (primární klíč) užívané v PubMed pro identifikaci záznamů v této databázi.

4.5. Ostatní chemické databáze

Chemických databází jsou jistě desítky, možná stovky. Liší se obvykle rozsahem a specializací. Relativně aktuální seznam těch nejdůležitějších je možné najít na serveru ChemistryViews[38].

5. Závěr

Z vlastností popisovaných databází plyne, že výběr té, kterou je vhodné použít, záleží vždy na tom, čeho chceme docílit:

- Pokud chceme zjistit, jestli již sloučenina byla popsána v literatuře, nebo o ní zjistit veškeré dostupné informace, je vhodné použít jak SciFinder, tak Reaxys. Tyto dva systémy poskytnou téměř vždy odlišné výsledky, které plynou z použití různých primárních zdrojů a také období, ve kterém byl ten který zdroj používán. Reaxys jde obvykle více do minulosti.
- Pokud chceme najit všechny publikace, které popisují reakci, která nás zajímá, je opět vhodné použít jak Reaxys tak SciFinder. Opět pravděpodobně dostaneme výsledky, které se budou lišit.
- Pokud chceme provést úplnou literární rešerši na téma popsané pomocí klíčových slov, je samozřejmě vhodné použít všechny tři komerční zdroje, ale u Web of Science lze očekávat největší počet záznamů.
- Pokud nemáme zájem o nalezení pokud možno všech informací na dané téma, ale chceme najít alespoň něco, vystačíme obvykle s nějakou volně dostupnou databází. Pro sloučeniny je pak nejvhodnější použít

ChemSpider a pro publikace Google Scholar. U Google Scholar máme navíc šanci (nikoli však jistotu), že nejvíce relevantní výsledek bude v seznamu výsledků na prvním místě.

– Pokud chceme najít sloučeninu, na kterou máme řadu požadavků ohledně jejích vlastností, pak nejlepší volbou bude Reaxys.
– Máme-li nějaký komplexní dotaz formulovaný přirozeným jazykem, tak nejlepším systémem, který poskytne pravděpodobně nejlepší odpověď bude SciFinder.
– Pokud chceme stáhnout z databáze co nejvíce experimentálních dat, nebo bibliografických záznamů v co nejkratším čase a v co nejstrukturovanější podobě, pak jasnou volbou bude Reaxys.

Tento článek vznikl za podpory MŠMT v rámci Národního programu udržitelnosti I projekt LO1220 (CZ-OPENSCREEN).

LITERATURA

1. Jirát J.: Chem. Listy *111*, 710 (2017).
2. http://www.acdlabs.com/resources/freeware/chemsketch, staženo 25. 5. 2017.
3. https://www.chemaxon.com/products/marvin/marvinsketch, staženo 25. 5. 2017.
4. https://www.reaxys.com, staženo 28. 5. 2017.
5. https://scifinder.cas.org, staženo 28. 5. 2017.
6. https://webofknowledge.com, staženo 28. 5. 2017.
7. Bharti N., Leonard M., Singh S.: J. Chem. Educ. *93*, 852 (2016).
8. Šilhánek J.: Chem. Listy *108*, 81 (2014).
9. Hašek J.: Chem. Listy *105*, 467 (2011).
10. https://www.google.cz, staženo 28. 5. 2017.
11. http://www.cas.org/content/chemical-substances/faqs, staženo 27. 5. 2017.
12. http://www.cas.org/, staženo 29. 5. 2017.
13. http://dx.doi.org, staženo 27. 5. 2017.
14. https://www.w3.org/TR/2012/WD-html5-20120329/dnd.html#dnd, staženo 28. 5. 2017.
15. http://www.molinspiration.com/jme/, staženo 27. 5. 2017.
16. http://www.cambridgesoft.com/software/overview.aspx, staženo 28. 5. 2017.
17. Lawson A. J., Swienty-Busch J., Géoui T., Evans D., v knize: *The Future of the History of Chemical Information*. kap. 8, American Chemical Society, Washington DC, 2014.
18. Luckenbach R.: J. Chem. Inf. Comput. Sci. *21*, 82 (1981).
19. Lippert W.: J. Chem. Inf. Comput. Sci. *19*, 201 (1979).
20. https://service.elsevier.com/app/answers/detail/a_id/14526/c/10547/supporthub/reaxys/related/1/, staženo 29. 5. 2017.
21. https://www.nlm.nih.gov/pubs/factsheets/medline.html, staženo 28. 5. 2017.
22. http://www.cas.org/content/references/corejournals, staženo 28. 5. 2017.
23. http://cassi.cas.org, staženo 28. 5. 2017.
24. http://www.cas.org/content/references/patentcoverage, staženo 28. 5. 2017.
25. http://www.commonchemistry.org, staženo 28. 5. 2017.
26. http://www.cas.org/etrain/scifinder/overview/story_html5.html, staženo 28. 5. 2017.
27. https://scifinder.cas.org/help/scifinder/R42/tutorial_draw_structures.htm, staženo 28. 5. 2017.
28. Garfield E: Int. Microbiol. *10*, 65 (2007); http://garfield.library.upenn.edu/papers/barcelona2007.pdf, staženo 28. 5. 2017.
29. http://scientific.thomsonreuters.com/imgblast/JCRFullCovlist-2016.pdf, staženo 28. 5. 2017.
30. http://wokinfo.com/essays/journal-selection-process/, staženo 27. 5. 2017.
31. http://wokinfo.com/training_support/training/web-of-science/, staženo 29. 5. 2017.
32. http://scholar.google.com, staženo 29. 5. 2017.
33. https://scholar.google.com/intl/us/scholar/help.html#coverage, staženo 27. 5. 2017.
34. http://www.chemspider.com, staženo 29. 5. 2017.
35. http://cssp.chemspider.com/, staženo 29. 5. 2017.
36. https://pubchem.ncbi.nlm.nih.gov/search/, staženo 27. 5. 2017.
37. https://www.ncbi.nlm.nih.gov/pubmed/, staženo 27. 5. 2017.
38. http://www.chemistryviews.org/details/education/10015921/Chemistry_Databases.html, staženo 30. 5. 2017.

J. Jindřich[a,b] *([a] CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Institute of Molecular Genetics of the Academy of the Sciences of the Czech Republic, Prague, [b] Department of Organic Chemistry, Faculty of Science, Charles University, Prague):* **Database Resources in Chemistry**

The most widely used specialized chemistry databases and web services allowing access to them will be described. Besides the largest commercial three (Reaxys, SciFinder, Web of Science), some freely available services (Google Scholar, ChemSpider , PubChem, PubMed) will also be discussed. Strong and weak points of the presented databases will be emphasised and recommendations for optimal database selection and searching will be given.