JSOU ZAHRANIČNÍ INFORMAČNÍ ZDROJE ZPŘÍSTUPŇOVANÉ V ČR DUPLICITNÍ?

Marie Paráková, Ústřední knihovna UK
Petr Boldiš, ÚISK FF UK

Článek byl zpracován v říjnu 2001 na základě studie – výstupu z grantu LI 002043 Zabezpečení vědy a výzkumu v humanitních oborech základními informačními zdroji.

1 Úvod

Trvale se snižující dostupnost primárních i sekundárních informací roztrpčovala informační pracovníky ve všech knihovnách v České republice. Ve snaze zabránit tomuto propadu vznikla v roce 1999 iniciativa Ústřední knihovnické rady podporovaná Ministerstvem kultury ČR a na svět přišel program, jehož oficiální název byl *Optimalizace dostupnosti informací ze světových periodických zdrojů v českých knihovnách*. Této myšlenky se ujalo Ministerstvo školství, mládeže a tělovýchovy ČR, které je odpovědné za výzkum a vývoj Radě vlády ČR, jež cítila potřebu doplnit vzniklé vakuum v oblasti zajištění a pokrytí těchto aktivit informacemi. Výsledkem bylo otevření programu LI „Informační zdroje pro výzkum a vývoj“, který měl 3 podprogramy:

A. podporu/vytvoření multifunkčních knihovnických center
B. získání konkrétních zdůvodněných titulů informačních zdrojů/dokumentů
C. získání multilicencí/velkoplošných licencí pro přístup k informačním zdrojům.

Program byl otevřen jako víceletý, pokrývající léta 2000–2003. Pro potřeby tohoto článku je zajímavý podprogram o multilicencích pro přístup k informačním zdrojům.

Přestože vznik tohoto programu nebyl procesem náhodným, ale velmi promyšleným, jenž vycházel jak z potřeb široké uživatelské základny, tak z podnětu centrálního řídícího orgánu, nebylo při sepisování grantových přihlášek mnoho prostoru pro obsáhlou diskusi nad nabídkou komerčních firem. Ze široké nabídky dodavatelských firem byly vybrány informační zdroje, které již byly v minulosti ověřeny na malých vzorcích uživatelů jako lokální řešení pro konkrétní instituce, nebo nové celky vytvořené producenty dat, které svým profilem vyhovovaly našim uživatelům. V grantové soutěži pak uspěly především ty projekty, které pokrývaly co nejširší uživatelskou základnu a odpovídaly současným potřebám výzkumu a vývoje.

Tehdy se zrodila myšlenka porovnat některé vybrané multioborové zdroje, které splňovaly podmínku uvádění plných textů časopiseckých článků a u nichž bylo možné očekávat překrytí titulů časopisů. Tato kvantitativní analýza byla provedena u databázových celků eIFL Direct, Periodical Contents Index, ProQuest 5000 a Springer Link. Výsledky studie byly prezentovány na konferenci Inforum 2001 v Praze a vzhledem k zájmu, který vyvolaly, jsme přistoupili k rozšíření studie i na velmi významný bibliografický zdroj, jenž je ve světě vědy a výzkumu chápán jako vrchol a záruka kvality excerpovaných časopiseckých titulů – citační rejstříky Science Citation Index a jejich elektronická podoba nazvaná Web of Science. U tohoto zdroje nás zajímala skutečnost, jakou vazbu mají bibliografické údaje o článcích na výše uvedené plnotextové zdroje.

Studie začíná popisem jednotlivých zdrojů, pokračuje vlastní kvantitativní analýzou, kde je popsána především metodika sběru vstupních údajů, proces porovnávání včetně vytvoření pracovní terminologie a zveřejnění výsledků s uvedením omezujících faktorů, které mohou způsobit případné zkreslení výsledků. Závěr obsahuje obecné shrnutí všech postřehů a návrhů, které by mohly vést k eliminaci omezujících faktorů.

2 Popis jednotlivých databází/databázových celků

Většina údajů o počtech excerpovaných titulů je přibližná a platná k datu vzniku studie – tj. přelomu dubna/května 2001.

eIFL Direct – Electronic Information for Libraries[1)]

Producent: projekt EBSCO Publishing (USA) & OSI
http://www.epnet.com (EBSCO Publishing)
http://www.osi.hu (Open Society Institut)
http://www.nkp.cz/eifl/ (stránky české licence eIFL)

Obsažené druhy dokumentů: články z časopisů, monografické publikace (příručkové dokumenty)
Formáty záznamů: bibliografické citace, plný text
Retrospektiva: od r. 1990 –
Přibližný počet excerpovaných zdrojů: 3300 (periodika), 1300 (příručkové publikace)
Přístup ke službě: http://search.epnet.com

Řešitel projektu pro ČR: PhDr. Hana Nová, Národní knihovna ČR
Trvání licence: 2000–2003 (program MŠMT ČR „Informační zdroje pro výzkum a vývoj“)

eIFL Direct zpřístupňuje databáze společnosti **EBSCO Publishing** obsahující přibližně **3300** titulů časopisů s plnými texty (převážně od r. 1990), novin a zpráv informačních agentur a přes **1300** publikací příručkového charakteru s plnými texty z oblasti humanitních věd a lékařství. eIFL Direct je určen v každé z účastnických zemí především pro akademické, výzkumné a národní knihovny.

Předmět pokrytí: medicína, ekonomie a obchod, vzdělání, zeměpis, historie, humanitní vědy, knihovnictví a informační věda, filozofie, politologie a veřejná správa, psychologie, sociální vědy, teologie a religionistika.

Systém je dále ve studii uváděn jako databáze EBSCO.

Kompletní přehled titulů na: *http://www.epnet.com/maglists/maglist.htm*

Periodical Contents Index Full Text (PCI Full Text)

Producent: ProQuest Information and Learning (USA), divize Chadwyck-Healey (Velká Británie)
http://pcift.chadwyck.co.uk

Obsažené druhy dokumentů: články z časopisů
Formáty záznamů: bibliografické citace, plný text
Retrospektiva: od r. 1770–1990 (1995)
Počet excerpovaných zdrojů: (100 v plném textu k dubnu 2001; 120 k říjnu 2001)
Přístup ke službě: http://www.proquest.cz

Řešitel projektu pro ČR: Mgr. Marie Paráková, Ústřední knihovna UK
Trvání licence: 2000–2003 (program MŠMT ČR „Informační zdroje pro výzkum a vývoj")

Periodical Contents Index je jedinou historickou databází poskytující bibliografické záznamy z oblasti společenských a humanitních věd. Pokrývá více než 3000 časopisů od počátku jejich vydávání až do r. 1990 (1993) a obsahuje cca 11 000 000 záznamů. Databáze pokrývá vůdčí časopisy po celém světě, tj. i v různých jazycích.
Nadstavbou PCI je služba PCI Full Text, která poskytuje i plné texty časopisů (faximilní kopie). V současnosti[2)] obsahuje plné texty 120 časopisů (v době studie 100) a ročně by mělo přibývat dalších 75 titulů. Databáze je z časového hlediska uzavřená – pouze se posouvá, průběžně se inovuje ovládání, indexace a pokrytí plným textem.

Předmět pokrytí: antropologie a etnologie, archeologie a starobylé civilizace, umění a architektura, ekonomie a obchod, vzdělání, zeměpis, historie, americká historie, humanitní vědy, hebraistika, právo, knihovnictví a informační věda, lingvistika a filologie, literatura, hudba a veřejné vystupování, filozofie, politologie a veřejná správa, psychologie, sociální vědy, teologie a religionistika.

Kompletní přehled titulů na: *http://pcift.chadwyck.co.uk/titles/titles.html*

ProQuest 5000

Producent: ProQuest Information and Learning (dříve Bell & Howell Information and Learning), USA
http://www.proquest.com

Obsažené druhy dokumentů: články z novin a časopisů, knihy, tiskové, vládní zprávy, cizí databáze
Formáty záznamů: bibliografické záznamy, plný text
Retrospektiva: od r. 1971 – (bibliografické záznamy), od r. 1987 – (plný text)
Přibližný počet excerpovaných zdrojů: 8378 (cca 4000 v plném textu)
Přístup ke službě: http://www.proquest.cz

Řešitel projektu pro ČR: Mgr. Marie Paráková, Ústřední knihovna UK
Trvání licence: 2000–2003 (program MŠMT ČR „Informační zdroje pro výzkum a vývoj")

ProQuest 5000 je hlavním databázovým produktem společnosti ProQuest Information & Learning. Nabízí přístup k záznamům přibližně 4000 časopisů v plném textu a dalším zdrojům ze širokého spektra oborů. Systém nabízí články v několika dostupných formátech – jako HTML text, HTML text s grafikou, formát PDF a naskenovaný text článku. U každého článku je alespoň anotace.
Do konce roku 2001 mohou uživatelé v České republice vyhledávat i ve specializovaných odborných databázích jiných producentů **Agricola** (zemědělství), **ERIC** (výchova a vzdělání) a **Medline** (oblast medicíny).

Předmět pokrytí: humanitní obory, společenské vědy, filozofie, teologie a religionistika, medicína, mezinárodní problematika, vojenství, vzdělání, ekonomika a obchod, bankovnictví, účetnictví, výpočetní technika, telekomunikace, marketing, management.

Kompletní přehled titulů na: *http://pcift.chadwyck.co.uk/titles/titles.html*

Springer Link

Producent: Springer (Německo)
http:/www.springer.de

Obsažené druhy dokumentů: články z časopisů (elektronické verze), knihy (edice), zprávy pro odbornou veřejnost, software
Formáty záznamů: bibliografické citace, anotace, plný text (elektronická verze časopisu ve formátu PDF)
Přístup ke službě: http://link.springer.de
Počet excerpovaných zdrojů: 480 (časopisy), 17 (knižních edic)

Řešitel projektu pro ČR: PhDr. Anna Patočková, Státní technická knihovna, Praha
Trvání licence: 2000–2003 (program MŠMT ČR „Informační zdroje pro výzkum a vývoj")

LINK je online službou, která nabízí elektronické verze časopisů a knih nakladatelské skupiny Springer.
Úsilí nakladatele směřuje k tomu, aby elektronické verze vybraných časopisů v rámci LINKu byly uveřejněny dříve, než vyjde jejich tištěná verze. Tituly pro zpřístupnění online jsou vybírány odbornými editory časopisů skupiny Springer.

Předmět pokrytí: medicína, farmacie, fyzika, přírodní vědy, výpočetní technika.

Web of Science (WoS)

Producent: Institut for Scientifical Information (ISI), USA
http://www.isinet.com

Obsažené druhy dokumentů: články z časopisů, patenty, „šedá literatura"

Formáty záznamů: bibliografické citace, abstrakt (u cca 60 % záznamů)
Přístup ke službě: http://bimbam.cuni.cz
Retrospektiva: od r. 1974
Přibližný počet excerpovaných zdrojů: cca 8500

Řešitel projektu pro ČR: PhDr. Ivana Kadlecová, Akademie věd ČR
Trvání licence: 2000–2003 (program MŠMT ČR „Informační zdroje pro výzkum a vývoj")

Služba Web of Science umožňuje přístup k unikátním databázím citačních rejstříků ISI, které jsou zcela jedinečné pro sledování citačních vazeb. Jednotlivé báze citačních rejstříků:

1. **Science Citation Index Expanded –** citační rejstřík z oblasti přírodních věd a techniky
2. **Social Science Citation Index –** multidisciplinární báze z oblasti společenských věd
3. **Arts & Humanities Citation Index –** citační rejstřík z oblasti humanitních věd

3 Kvantitativní analýza překrytí excerpčních zdrojů databází

3.1 Metodika porovnávání excerpčních zdrojů

Z důvodu stálého růstu excerpčních zdrojů musel být stanoven den, ke kterému jsou dané údaje platné, a to k 25. 4. 2001. Některé seznamy titulů nezachycují aktuální stav databází k 25. 4. (pomalá aktualizace seznamů titulů), nicméně změny, které by tímto vznikly, nemohou ovlivnit celkové výsledky kvantitativního porovnání jednotlivých databázových celků.

Seznamy databázových sektorů EBSCO v projektu eIFL DIRECT – Comprehensive MEDLINE Full Text a Health Source byly přidány do srovnání později (5. 5. 2001), ale tyto seznamy byly aktualizovány stejně jako ostatní – tj. na přelomu dubna a května.

3.2 Získávání a zpracování podkladových dat

Veškeré podkladové seznamy jsme získali z webovských stránek jednotlivých producentů databází. Tyto seznamy existují v různých formách a také v různé kvalitě.

Nejúplnější data poskytuje společnost EBSCO Information Services, která seznamy nabízí ve třech formátech: HTML, PDF a XLS (tabulka formátu MS EXCEL). Po stránce struktury vypadají záznamy následovně:
název titulu – ISSN – vydavatel – časové pokrytí titulu v databázi: abstrakt, plný text, naskenovaný formát (PDF).

Tyto seznamy existují pouze pro jednotlivé databázové sektory. Proto bylo nutné sloučit všechny seznamy do jednoho a vyřadit duplicity v jednotlivých seznamech tak, aby byl titul v celkovém seznamu databází EBSCO obsažen pouze jednou.

U databáze ProQuest je seznam titulů generován do formátu HTML podle požadavků uživatele na zobrazení. Systém umožňuje zobrazení v těchto výstupních sestavách:

Název titulu
ISSN
Další informace k titulu v databázi:
datum první citace v databázi
datum prvního abstraktu v databázi
datum, od kdy je v databázi plný text
datum, od kdy je v databázi grafika
datum, od kdy je v databázi plný text s grafikou
data, od kdy jsou citace sledovány systémem Wilson nebo MEDLINE
data, od kdy jsou abstrakty sledovány systémem Wilson nebo MEDLINE

Dále bylo možné zvolit si výstupní formát: HTML, text (ASCII) a „Comma-delimited ASCII".

Jak u databáze EBSCO, tak u databáze ProQuest, je možné zobrazit seznam titulů ve stručné podobě přímo v databázi.

Databáze PCI Full Text má svůj seznam plnotextových titulů (obsahoval ve sledovaném období 100 položek) ve formátu HTML. V tomto seznamu jsou pouze následující položky:
název titulu – země a místo publikování – datum, od kdy je titul sledován – předmětové zaměření titulu.
Tento seznam bylo nutné doplnit o další údaje o jednotlivých titulech, které jsou v databázi obsaženy v samostatných souborech.

Springer LINK je spíše sbírkou dokumentů vydávaných skupinou Springer. Seznam časopisů je k dispozici pouze jako jednoduchý soupis ve formátu HTML; stejně jako u databáze PCI Full Text bylo nutné doplnit z dalších stránek potřebné údaje.

Producent citačních rejstříků Web of Science zveřejňuje seznam excerpovaných titulů ve formátu HTML (stránky po 500 titulech). Seznam obsahuje:
název titulu – periodicitu titulu – ISSN – jméno vydavatele společně s jeho adresou.

Jednotlivé seznamy bylo potřeba doplnit do stanoveného formátu potřebného pro zpracování studie: *název titulu, ISSN, data sledování titulů v databázích.*

Srovnání bylo zaměřeno na překrytí titulů časopisů, takže veškeré další typy databázových sektorů (faktové databáze, databáze typu „news" apod.) byly vynechány. Rovněž jsme ze seznamů vyřadili neperiodickou literaturu (encyklopedické a referenční zdroje).

3.3 Metodika porovnávání databází

3.3.1 Srovnání databází programem GNU Diff pro zjištění duplicit

Pro první srovnání duplicit byl použit program GNU Diff pod operačním systémem Linux, který umožňuje porovnávání různých textových souborů. Pro potřeby studie byly připraveny textové soubory s názvy titulů a ISSN. Metoda tohoto zpracování byla zvolena především pro obsáhlost jednotlivých seznamů titulů v databázích, v některých případech i pro jejich obtížnou zpracovatelnost.

Ukázka výstupu programu GNU Diff:

„Qatar - Saudi Arabia Economic Studies“	|	„QST“
„Qatar Country Monitor“	<	
„Quaker Studies“	<	
„Qualitative Health Research“		„Qualitative Health Research“
„Qualitative Inquiry“		„Qualitative Inquiry“
„Quality Assurance“	<	
„Quality in Higher Education“	<	
„Quality“		„Quality“
„Quarterly Journal of Austrian Economics“	|	„Quarterly Journal of Nuclear Medicine“
„Quarterly Journal of Business & Economics“	|	„Quarterly Journal of Speech, The“
„Quarterly Journal of Economics“	|	„Quill, The“

3.3.2 Vizuální kontrola duplicit

Program GNU Diff byl naprogramován jako pomůcka pro porovnávání programového kódu – tj. porovnává v souborech identické znaky v sekvenci za sebou. U porovnávání titulů periodik jsou jeho slabou stránkou podkladová data. Jednotliví producenti databází mají každý svou metodiku zpracování seznamů a i v nich se identické tituly mohou objevit pod různými variantami názvů.

Příklady variant názvů (identické tituly):

Antioch Review	Antioch Review, The
Intervention in School and Clinic	Intervention in School & Clinic
Graphic Arts Monthly; the magazine of the printing industry	Graphic Arts Monthly

Z výše uvedených důvodů bylo nutné vizuálně porovnat programové výstupy se zdrojovými soubory za účelem nalezení duplicit s jinou variantou názvu.

3.3.3 Ověřování titulových duplicit porovnáváním dalších identifikačních znaků

Po převedení duplicitních titulů do samostatných souborů byla provedena další identifikace titulů periodik – porovnáním názvu, ISSN a, nebylo-li dostupné ISSN, místa vydání. Hlavní důraz byl přitom kladen na ISSN, které se vyskytovalo ve všech seznamech titulů od producentů databází (existuje-li). Z těchto seznamů byl připraven seznam *„titulových duplicit“*, tj. seznam shodných titulů periodik, které se vyskytují v obou porovnávaných databázích.

Z důvodu různé metodiky zpracování se některé tituly periodik neshodovaly ve všech identifikačních znacích – tj. stejná varianta názvu, ale odlišné ISSN. Všechny takto nalezené tituly byly převedeny do samostatných oddílů na konci každého seznamu pod označením *„podezřelé duplicity“*. U těchto titulů periodik nebylo možné z důvodu chybějících údajů v seznamech titulů producentů ověřit, že nalezené tituly jsou skutečně identické. V některých případech může jít o chybu zpracování (první číslo v ISSN se u dvou nalezených liší – možný překlep) a u dalších se pravděpodobně jedná o jiný titul se shodným názvem nebo o změnu ve vydávání titulu (zánik a obnovení, převod pod jiného vydavatele apod.).

Termín *„duplicita vydání“* znamená shodu jak v titulu, tak v dalších identifikačních znacích a překrytí v jednotlivých číslech ročníku.

Pokud se některé tituly odlišují částí názvu (podnázev, označení regionálního vydání), ale v ostatních identifikačních znacích se shodují, je na to upozorněno v poznámce.

V databázi ProQuest 5000 se vyskytují tituly, u kterých je plný text dostupný jako volitelný placený doplněk. To znamená, že není k dispozici ve standardní nabídce. Tato skutečnost je také uvedena v poznámce.

3.3.4 Porovnání časového překrytí jednotlivých titulů

Po kontrole duplicit bylo provedeno porovnání časového překrytí jednotlivých titulů srovnáním časového rozsahu excerpce v jednotlivých databázích.

Jak již bylo výše uvedeno, liší se jednotlivé seznamy také údaji o časovém pokrytí jednotlivých titulů v databázi. U některých databází (EBSCO, ProQuest) se uvádí podrobně rozepsané datum – např. 1. 1. 1996, zatímco u dalších (PCI Full Text) se vyskytuje označení „Spring 1991“. Jednotlivé tituly mají také odlišnou periodicitu a z tohoto důvodu bylo zvoleno porovnávání překrytí **na jednotlivé roky** (nejmenší uváděnou jednotkou je 0,5 roku). To se týká titulů, které jsou v jedné nebo druhé databázi pouze v určitém uzavřeném období a pro které používáme označení *„uzavřené duplicity“* (např. od roku 1994 do roku 1997).

Překrytí na roky jsme odvozovali z časových údajů u posledních vydání titulů následovně:

1. u titulů v období do letních měsíců (Summer) jako překrytí 0,5 roku
2. u titulů v období od září/října (Fall/Winter) jako překrytí 1 rok.

Toto odvozování překrytí je pouze pracovní, neboť by bylo potřeba u každého titulu zjistit údaje o číslování jed-

notlivých ročníků a období, ve kterém nový ročník začíná. U některých vydavatelů přechází ročník z roku na rok (např. ročník začíná v roce 2000 a končí v roce 2001), přičemž běžnější je situace, kdy se kalendářní rok překrývá s ročníkem.

V databázích existují také tituly, které se překrývají od určitého roku a v obou databázích překrytí pokračuje. Pokud je titul nadále v databázi sledován, je v poli „do" použit znak „—" jako označení pokračujícího sledování titulu v databázi. Pro označení těchto duplicit byl zvolen termín *„perspektivní překrytí"* a je uváděn od data, od kterého se titul v obou databázích vyskytuje.

Příklad:

Pokrytí		od	do
databáze A	Banking Strategies	1992	—
databáze B	Banking Strategies	1996	—
Překrytí je v tomto případě uvedeno od roku 1996.			

3.3.5 Limitující faktory dostupnosti titulů

U některých titulů se mohou vyskytnout i další omezení. Nejčastější variantou bylo:

a) Full Text Embargo (databáze EBSCO)
b) Full Text Availability Delayed after …Days due to Publisher Restriction (databáze ProQuest)

V prvním případě (*Full Text Embargo*) se jedná o opatření, kterým vydavatel periodika dává povolení k uveřejnění v databázi až po určitém čase. Ten se zpravidla pohybuje v řádech několika měsíců. Z hlediska určování duplicit tak dochází v jedné z databází ke zpoždění, ale toto omezení nemá zásadní vliv na srovnávání.

V případě *„Full Text Availability Delayed after ..."* dochází na základě licenční smlouvy s vydavatelem k vymazání plného textu po stanoveném počtu dní. Tento fakt ovlivňuje srovnání na duplicity tím, že plný text daného titulu se **překrývá vždy pouze po daný počet dní**. Po určené době je v databázi už pouze citace/abstrakt.

Příklad:

Pokrytí	od	do	
databáze A Car and Driver	1/1/98	current	[Full Text Availability Delayed by 14 Days due to Publisher Restriction]
databáze B Car and Driver	1/1/98	current	
v tomto případě není překrytí	„1998 — "	ale **pouze 14 dní**	
Překrytí se neustále opakuje (je u každého nového vydání), ale pořád jde o překrytí 14 dnů. Proto také není tato duplicita zařazena do „*perspektivních duplicit*".			

3.3.6 Metodika porovnávání výsledků

Jak je výše zmíněno, překrytí titulů je vyjádřeno v letech (tzv. uzavřené duplicity) nebo v roce, kdy překrytí začíná (a dále pokračuje). Veškerá srovnávání jsme prováděli především s tzv. *„duplicitami vydání"* – tj. skutečnými titulovými duplicitami. Případy, kdy jsme srovnávali i tzv. *„titulové duplicity",* jsou označeny. Do výsledků srovnání nejsou zahrnuty tzv. *„podezřelé duplicity"*. Tyto údaje jsou shrnuty do tabulek.

Pro porovnání překrytí jsme zvolili následující statistický vzorec pravděpodobnosti (příznivé případy ke všem možným):

$$P(P_1, P_2) = \frac{|\ P_1 \cap P_2\ |}{|\ P_1 \cup P_2\ |}$$

který jsme upravili pro tento účel na:

$$P(DB_1, DB_2) = \frac{|\ DB_1 \cap DB_2\ |}{|\ DB_1 \cup DB_2\ |}$$

Vysvětlení prvků vzorce:

$DB_1 \cap DB_2$ – průnik plnotextových titulů v obou databázích (tj. společné tituly v obou databázích)

$DB_1 \cup DB_2$ – počet sjednocených (tj. sečtených) plnotextových titulů v obou databázích (výchozí stav na začátku studie)

Výsledkem je procentuální vyjádření vzájemného překrytí obou databází.

4 Výsledky srovnání databází s plnými texty

4.1 Springer LINK – Periodicals Content Index FullText

Celkem duplicit vydání: *0*

Při srovnání se potvrdila původní teze, že mezi těmito databázemi není titulové překrytí. Vycházeli jsme z předpokladu naprosto odlišného zaměření obou databází – přírodní vědy (Springer LINK) a společenské vědy (Periodicals Content Index FullText).

4.2 Springer LINK – EBSCO

Celkem duplicit vydání: *59*
Celkem titulových duplicit: *59*

Na rozdíl od předchozího srovnání jsme zde porovnávali úzce profilovanou databázi s databází všeobecného zaměření. Springer LINK není klasickou databází článků z excerpovaných titulů – spíše se jedná o sbírku vybraných titulů vlastní vydavatelské skupiny Springer-Verlag.

Celkově bylo nalezeno 59 duplicit vydání. Z tohoto počtu je 51 titulů producentem databáze označeno jako „nově přidané tituly", které zatím nemusí být dostupné. U databází EBSCO je u 51 titulů z celkového počtu 59 roční embargo na zveřejnění plného textu; toto ochranné opatření tak zajišťuje Springer LINKu zákazníky.

Zajímavostí je, že databáze EBSCO obsahují i tři tituly tohoto vydavatelství, které nejsou excerpovány ve Springer LINKu. To je pravděpodobně z důvodu snahy Springeru zveřejňovat prostřednictvím této služby pouze významné tituly, o které je mezi odbornou veřejností zájem.

4.3 Springer LINK – ProQuest

Celkem duplicit vydání: *0*
Celkem titulových duplicit: *0*

V případě tohoto srovnání nebyly nalezeny ani duplicity vydání, ani titulové duplicity. Pravděpodobným důvodem je jednak odlišné zaměření obou databázových systémů (exaktní vědy a databáze multioborového zaměření), jednak geografické pokrytí.

Skupina Springer publikuje vlastní tituly, které vycházejí v Evropě, zatímco ProQuest excerpuje tituly především ze Spojených států.

4.4 Periodicals Content Index FullText – ProQuest

Celkem duplicit vydání: *0*
Celkem titulových duplicit: *57*

V této části studie došlo k zajímavé situaci – byly srovnávány dva informační zdroje, které v současnosti patří jednomu producentovi – společnosti ProQuest Information & Learning. Databáze Periodicals Content Index FullText je produktem společnosti Chadwyck-Healey z Velké Británie, která se před nedávnem stala divizí výše uvedené společnosti ProQuest. Z hlediska zaměření databází (společenskovědní a multioborové) byly duplicity vydání možné.

Zajímavé je, že zdroje se překrývají pouze titulově – tj. 57 titulů je stejných, ale liší se jejich roky vydání. Důvodem je historické zaměření databáze Periodicals Content Index FullText, která v době realizace studie pokrývala tituly plným textem do roku 1990.[3)] Srovnání také ovlivnil počet titulů s plným textem – v době srovnání 100 titulů Periodicals Content Index FullText. Ve většině případů je časový odstup mezi sledovanými tituly 1–2 roky. Tento rozdíl v nejbližších letech patrně přestane platit z důvodu rozhodnutí společnosti pokrýt plným textem tituly až do roku 1995 v co nejširším rozsahu.[4)]

V budoucnu je tak možné očekávat překrytí několika let v obou databázích (první polovina 90. let).

4.5 Periodicals Content Index FullText – EBSCO

Celkem duplicit vydání: *30*
Podezřelých duplicit: *1*

Překrytí:

počet let	počet titulů
1 rok	23
1,5 roku	5
2 roky	1
31 let	1

Celkem titulových duplicit: *49*

Při srovnávání Periodicals Content Index FullText s databázemi EBSCO již byly nalezeny duplicity vydání. Celkem se jedná o 30 titulů (a jednu podezřelou duplicitu, která se do konečného výsledku nezahrnuje). Překrytí je v případě 23 titulů 1rok (cca 80 % všech nalezených duplicit), 5 titulů se překrývá v 1,5 roce a jeden po 2 roky. Poslední titul – Sloan Management Review (ISSN 0019-848X) se překrývá 31 let a můžeme ho chápat jako výjimku.

Databáze EBSCO začíná tituly excerpovat na počátku 90. let, zatímco PCI FullText v tomto období excerpci končí. I zde je nutné počítat s větším budoucím překrytím z důvodu probíhajícího rozšíření záběru PCI FullText do roku 1995 a pokrytí zbývajících titulů plným textem.

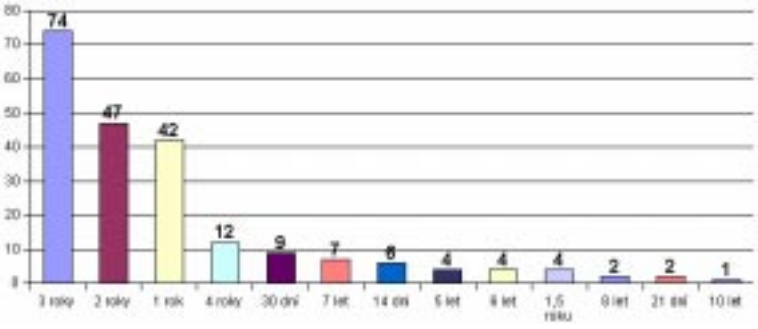
4.6 ProQuest – EBSCO

Celkové překrytí databází – duplicity vydání: *14,05 %*
Celkem duplicit vydání*: 1193*

Jak ProQuest, tak EBSCO jsou svým zaměřením informační zdroje univerzální – multioborové, počet titulů s plným textem je řádově stejný a oba producenti jsou ze Spojených států, což také ovlivňuje excerpční základnu titulů. Tyto údaje naznačovaly už předem, že zde půjde o velké překrytí především v duplicitách vydání.

Výsledky srovnání jsou poměrně překvapivé – překrývá se pouze 14 % celkového obsahu obou databází. Tento výsledek jsme dále rozdělili na dvě části – uzavřené duplicity a duplicity vydání.

Část I. – uzavřené duplicity

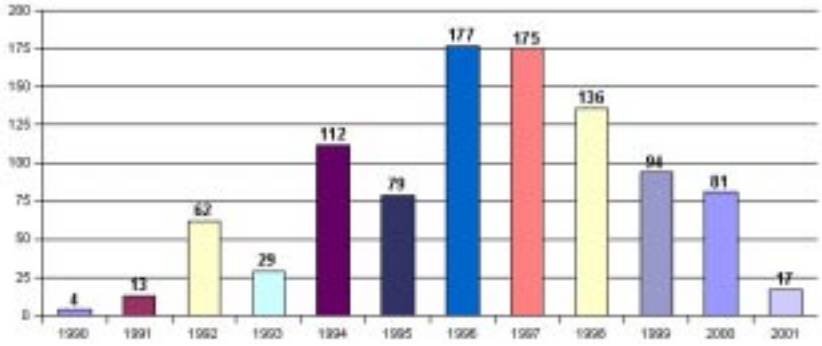


Obr. 1 Délky překrytí u uzavřených duplicit

***Celkem uzavřených duplicit*:** *214*

Jak je z grafu vidět, nejvíce titulů – 74 – se překrývá po dobu 3 let, 47 po dobu 2 let a 42 po dobu 1 roku. Ostatní tituly (cca 23 % všech) mají délku překrytí jinou. Nejdelší nalezenou délkou překrytí bylo 10 let (1 titul).

Část II. – perspektivní duplicity



Obr. 2 Nárůst perspektivních duplicit v letech

Celkem perspektivních duplicit: *979*

Tato část je zajímavá tím, že ukazuje nárůst nově excerpovaných titulů v obou databázích u duplicit vydání. Jak je z grafu patrno, nejvíce společně odebíraných titulů pochází z let 1996–1997 (177, resp. 175 titulů). Od roku 1997 se počet nově sledovaných identických titulů, které jsou v obou databázích, neustále snižuje. Počet titulů uvedený za rok 2001 (17 nově přidaných titulů) nelze brát jako konečný, protože údaj pochází z přelomu dubna/května – tj. první třetiny roku. Při spekulativní tezi, že v dalších dvou třetinách roku bude překrytí novými tituly přibližně stejné, bychom dospěli k číslu 51 titulů v roce 2001. Zde se však jedná o pouhou spekulaci, kterou by bylo vhodné ověřit na konci roku 2001.

5 Srovnání na možnost využití v kombinaci s citačními rejstříky WOS

5.1 Záměr a metodika

Do druhé části srovnání byla zařazena i služba Web of Science, přestože svým charakterem neodpovídá dalším srovnávaným zdrojům a v základní verzi se jedná o citační rejstříky, které neobsahují plné texty časopisů.[5)]

Tato část měla prověřit, zda lze na základě citace nalezené v rejstřících Web of Science (dále WOS) vyhledat plný text v některé z přístupných databází. Zde se tedy nejednalo o zjišťování duplicit, ale o porovnání excerpční základny citačního a plnotextového zdroje za účelem jejich využití v kombinaci.

Výchozí seznam titulů u WOS jsme získali stejným způsobem – tj. z webovských stránek producenta. U jednotlivých titulů jsou uvedeny následující údaje: *název titulu – periodicita – ISSN – vydavatel a jeho adresa.*

Seznam titulů obsahující celkem 8674 položek byl pro zpracování zcela nevhodný. Seznam je ve formátu HTML, údaje jsou uvedeny v odstavcích pod sebou. Seznam je rozdělen na 18 souborů po 500 položkách. Z tohoto důvodu byl programovacím skriptem z těchto seznamů utvořen jediný seznam obsahující ISSN. Tento seznam jsme pak srovnávali s ostatními seznamy titulů (resp. jejich ISSN) již zmíněným programem GNU Diff. U titulů, které čísla ISSN neobsahovaly (cca 30 titulů), bylo provedeno ruční dohledání údajů a jejich porovnání s dalšími seznamy.

5.2 Obecná zjištění

Citační rejstříky WOS pokrývají jak tituly z USA (cca 33,5 % sledovaných titulů), tak i z Evropy (cca 39,3 % sledovaných titulů), Asie a z oblasti Pacifického oceánu (cca 15,7 % sledovaných titulů), a z dalších lokalit (11,5 %).[6)] Tituly, jež sleduje Institut for Scientific Information, jsou vybírány podle různých citačních analýz, kterými se zjišťuje přínos jednotlivých titulů pro daný obor. Z tohoto pohledu měla studie také ukázat, zda lze nalézt v dostupných databázích s plnými texty také prestižní periodika sledovaná citačními rejstříky.

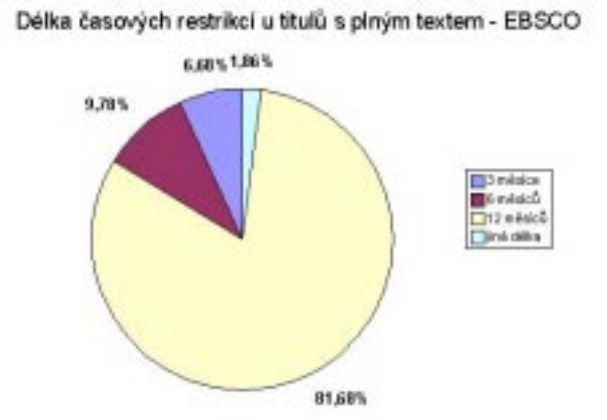
Zjistili jsme, že všechny srovnávané databáze obsahují tituly, které jsou excerpovány i citačními rejstříky WOS. Často je ale přístup k plnému textu v těchto databázích omezen na základě dohody s vydavateli jednotlivých titulů. Toto opatření je nazýváno jako *„moving wall“*. Jeho podstatou je dohodnuté časové období (např. 12 měsíců), které určuje rozestup mezi aktuálně vydaným číslem a číslem uveřejněným v databázi s plnými texty. Restrikce takto zaručuje vydavatelům odběr jejich titulů a databáze jsou pro potenciální uživatele zajímavé spíše jako ucelený retrospektivní zdroj.

5.3 Výsledky srovnání WOS – EBSCO

***Celkem identických titulů*:** *1307*

Databáze EBSCO
- s časovou restrikcí: 808
- volně dostupné: 459

V databázích společností EBSCO jsme nalezli nejvíce titulů excerpovaných citačními rejstříky WOS. Z tohoto počtu (1307) je asi 37 % titulů (459) dostupných bez časových restrikcí. Na většinu titulů (808 – tj. 63 %) je ale uplatněna časová restrikce (*„moving wall“*), která se pohybuje v délce od 3 do 36 měsíců. Rozložení délky časových restrikcí ukazuje následující graf:



Jak vyplývá z grafu, nejvíce se objevuje časový posun titulů v délce 12 měsíců (81,68 %, tj. 660 titulů). S velkým odstupem následuje časová restrikce v délce 6 měsíců (9,78 %, tj. 79 titulů) a 3 měsíců (6,68 %, tj. 54 titulů). Zbylé tituly (1,86 %, tj. 15 titulů) mají časové restrikce jiné – od 4 do 36 měsíců.

Nalezli jsme zde několik titulů, které již nejsou dále v databázích EBSCO excerpovány. Celkem 4 tituly jsou v seznamech uvedeny pod označením „ uzavřené období sledování plných textů".

5.4 Výsledky srovnání WOS – ProQuest

***Celkem identických titulů*:** *500*
***Neověřené tituly („podezřelé duplicity")*:** *2*

Databáze ProQuest
- s časovou restrikcí (vymazání plného textu): 5
- volně dostupné: 495

ProQuest 5000 obsahoval nejvíce titulů bez časových restrikcí – 459. U zbývajících časopisů následuje vymazání plného textu po 30 dnech (1 titul), 180 dnech (1 titul) a 730 dnech (3 tituly). Charakter časových restrikcí je oproti databázím EBSCO úplně jiný. Při využívání obou databází v kombinaci je tedy teoreticky možné získat plný text dostupných periodik vždy.

I u ProQuestu jsme nalezli tituly, které již nejsou dále excerpovány – celkem 59 titulů.

5.5 Výsledky srovnání WOS – PCI FullText

Celkem identických titulů: *70*

V poměru k celkovému počtu titulů s plným textem v databázi PCI FullText (100 titulů), obsahuje WOS nejvíce časopisů, které jsou excerpovány citačními rejstříky – celkem 70 titulů.

V případě PCI FullText je ojedinělé retrospektivní zaměření databáze (1770–1990). Vzhledem k tomu citační databáze WOS pokrývají pouze určitý časový úsek jednotlivých titulů v databázi. Přibližnou délku můžeme odvodit z let zahájení excerpce do citačních rejstříků *„Social Science Citation Index (SSCI)"* (oblast společenských věd) a *„Arts & Humanities Citation Index (AHCI)*" (oblast umění a humanitních oborů), které jsou přes WOS přístupné již od roku 1974. Z tohoto hlediska může být u jednotlivých titulů pokryt přibližně časový úsek 26 let.

Několik titulů v době zpracování studie i v době vzniku tohoto článku[7)] nebylo úplně pokryto plným textem. Chybějící pokrytí je v rozmezí několika málo let (například pokrytí: 1967–1973, 1976–1977) a je výsledkem rychlé přeměny původně citační databáze v databázi s plnými texty, kdy se zpřístupňovaly nejprve nejžádanější ročníky jednotlivých titulů.

5.6 Výsledky srovnání WOS – Springer LINK

Celkem identických titulů: *45*

Springer LINK je, jak již bylo uvedeno výše, službou vydavatelské skupiny Springer a zpřístupňuje pouze vlastní tituly. Tato role tzv. „*agregátora*" měla vliv i při srovnání s citačními rejstříky – všech 45 titulů je vydáváno nakladatelskou skupinou Springer.

6 Limitující faktory spolehlivosti výsledků

6.1 Chyby při zpracování – statistická chyba

Metoda zpracování seznamů – program GNU Diff a následná vizuální kontrola nebyla optimální. Byla ale zvolena z důvodů obtížné převoditelnosti dat, především ze seznamů titulů Web of Science. Je tedy třeba počítat s obvyklou statistickou chybou – 5 %. Pro další srovnání této povahy se nám jeví jako vhodnější použít převod seznamů titulů do vytvořené databáze, případně zpracování seznamů titulů programovým skriptem (převod do vhodného vstupního formátu pro databázi).

6.2 Spolehlivost podkladových dat producentů

Ačkoliv jsme získali jednotlivé seznamy titulů ve stejném období (25. 4. a 5. 5.), seznamy neodrážejí aktuální stav databází k tomuto období, neboť nejsou označeny přesným datem aktualizace. Výjimkou je databáze EBSCO, která měla své seznamy titulů označeny jako aktuální k přelomu dubna a května 2001. U dalších seznamů titulů je již spolehlivost nižší. U Web of Science a Springer LINK je seznam pravděpodobně aktualizovaný v měsíci dubnu, u databáze PCI FullText je aktuálnost seznamů nejasná. Seznam titulů databáze ProQuest podle některých dalších informací pochází z prosince roku 2000.

V případě některých „podezřelých duplicit" se ISSN od sebe lišilo pouze jedním číslem, což je pravděpodobně způsobeno překlepem v jednom ze seznamů. Máme důvod se domnívat, že i v podkladových seznamech se vyskytují chyby.

6.3 Dynamika pohybu titulů v databázích

Jednotlivé databáze se neustále rozšiřují o nové tituly a tento nárůst je velmi rychlý. Příkladem je databáze ProQuest, která k 15. 10. 2001 obsahuje celkem 8800 titulů,[8)] což představuje nárůst o 422 titulů oproti stavu k 25. 4. 2001 (8378 titulů).

Nově excerpované tituly jsou přebírány na základě licenčních smluv s jednotlivými vydavateli periodik. Tyto smlouvy by mohly rovněž ukázat na některá skrytá omezení v případě jednotlivých databází (trvání smlouvy, dodatky k zveřejnění plných textů atd.), ale vzhledem k povaze dokumentů o nich můžeme pouze přemýšlet.

7 Závěry studie

Při realizaci původního záměru, kterým bylo zjistit, zda a k jakému překrývání titulů dochází u vybraných informačních zdrojů, se objevily určité skutečnosti a teze, jejichž vzetí v úvahu by v mnohém usnadnilo provádění obdobných analýz.

1. Ke zjišťování titulových duplicit nelze použít pouze automatický způsob zpracování, ale je třeba tuto metodu doplnit o kontrolu dat a vyhodnocení zjišťovaných skutečností, tzn. ruční dohledání a doplnění chybějících údajů.
2. Pro kvantitativní analýzu překrytí titulů byly vytipovány následující údaje jako rozhodující: název titulu,

ISSN a období, po které je nebo byl titul sledován. Naprosto jednoznačnou rozlišovací schopnost má pouze ISSN. Z tohoto hlediska se nám jeví jako účelné a smysluplné vytvořit standardizovaný strukturovaný záznam a to ve tvaru: název titulu, vydavatel, místo vydání, ISSN, období excerpce v databázi vyjádřené rokem a měsícem. Ze sledovaných producentů databází se tomuto stavu nejvíce přibližuje firma EBSCO, která neuvádí pouze místo vydání. I při dodržení tohoto standardizovaného strukturovaného záznamu zůstane do budoucna problémem nejednotný postup při označování ročníků časopiseckých titulů – některé se kryjí s kalendářním rokem, ovšem existují i takové tituly, u nichž ročník přechází z roku na rok (začíná v určitém období roku 2000 a končí v roce 2001).

3. Při automatickém zpracování je velkým nedostatkem uvádění variantních názvů identických titulů časopisů. Variantnost se projevuje v (ne)uvádění členů, rozepsaném slovním vyjádření zástupných znaků (and, versus, &) a podnázvů. Ideálním stavem by bylo využití stávající databáze ISSN jako východiska pro standardizovaný zápis požadovaných údajů. Další podmínkou by byla jejich pravidelná aktualizace.
4. Určitou hrozbou pro sledované informační zdroje je absence titulů z dalších geografických oblastí. Američtí producenti se soustřeďují především na angloamerickou provenienci a opomíjejí jiná významná teritoria (Evropa, Asie atd.).

Po vyhodnocení všech dosažených výsledků jsme nuceni konstatovat, že oproti původnímu odhadu je titulový překryv u plnotextových databází minimální nebo vůbec žádný. Nejvyšší hodnoty vykazují dvě obdobné databáze ProQuest 5000 a eIFL Direct, což jsme jistě před zahájením studie předpokládali, neboť se jedná o zdroje podobné svým zaměřením. Procentuální vyjádření překryvu titulů v těchto databázích vykazuje hodnotu 14 %, což není z hlediska využitelnosti nijak podstatné překrytí. Výběr informačních zdrojů z hlediska kvantitativní analýzy plně odpovídá cílům a zadání programu LI „ Informační zdroje pro vědu a výzkum“.

Porovnání dostupnosti plného textu na základě bibliografické citace z Web of Science vyznívá nejlépe ve prospěch databáze eIFL Direct, ve které lze nalézt 1305 identických titulů.

Upozorňujeme na skutečnost, že studie obsahuje pouze výsledky kvantitativní analýzy a nelze tudíž od této skutečnosti odvozovat kvalitu databáze. Ke stanovení kvality databáze by bylo třeba zohlednit ještě další hlediska, jako např. uživatelské rozhraní, doplňkové služby, nárůst nových titulů a jejich pokrytí plným textem.

Poznámky:

1) Některé databáze mohou mít i další části, pokud k nim ale uživatelé v ČR nemají přístup, nejsou zmíněny a studie se jimi dále nezabývá.
2) Říjen 2001.
3) Nyní se postupně pokrývají plným textem tituly až do roku 1995.
4) V době vzniku tohoto článku je již 144 titulů dostupných s plným textem.
5) U citačních bází se objevuje celosvětová tendence dodávat v rámci doplňkových služeb i plný text.
6) The Web of Science DEMO [CD-ROM]. Institut for Scientific Information, 1997.
7) Říjen 2001.
8) Jedná se i o tituly, které nejsou pokryty plným textem.

Použité zdroje:

ISI Master Journal List [online]. Institut for Scientific Information, 2001 [cit. 2001-04-25]. Dostupné z: <http://www.isinet.com/cgi-bin/jrnlst/jlresults.cgi?PC=MASTER>.

JSTOR : Moving Wall [online]. JSTOR, 2001, last updated January 07, 2001 [cit. 2001-10-05]. Dostupné z: <http://www.jstor.org/>.

Konsorcium SPRINGER LINK [online]. Praha : Státní technická knihovna 2001, poslední aktualizace 20.6.2001 [cit. 2001-10-18]. Dostupné z: <http://www.stk.cz/zdroje/springer.htm>.

Kubíková, Věra: Program „Informační zdroje pro výzkum a vývoj“. In *Knihovny současnosti 2000.* Brno: Sdružení knihoven ČR, 2000, s.11-15.

PCI Full Text Title Lists [online]. Bell & Howell Information and Learning Company, 2001 [cit. 2001-04-25]. Dostupné z: <http://www.chadwyck.co.uk/> .

ProQuest [online]. Bell & Howell Information and Learning Company, 2001 [cit. 2001-04-25]. Dostupné z: <http://www.proquest.com/>. Název společnosti změněn na ProQuest Information and Learning.

ProQuest® 5000 International Title List. Bell & Howell Information and Learning Company, 2001 [cit. 2001-04-25]. Dostupné z: <http://tls.il.proquest.com/hp/Support/Titles/PQ5000Intl/>. Název společnosti změněn na ProQuest Information and Learning.

Springer LINK [online]. Springer, 2001 [cit. 2001-04-25]. Dostupné z: <http://link.springer.de/>.

The Web of Science DEMO [CD-ROM]. Institut for Scientific Information, 1997.

■ ■ ■ ■ ■