K některým nesrovnalostem v H15 a okolí

Pavel Šmerk

FI MU

4. 5. 2017

Jak to vzniklo: H14

- do H13 vyhledávání obdobné „hlavnímu“ isvav.cz
  - vizuálně, funkčností, prezentací výsledků, . . .
  - jen přibylo pár políček formuláře relevantních daným datům
- vyhledávání v H14 funkčně nesrovnatelně horší
  - nepodařilo se mi zjistit, co bylo důvodem omezení možností
  - nerozumím, proč nikomu nevadilo, že to odporuje smlouvě (s. 71)
  - přitom stačilo do existujícího SW přidat pár políček pro II. pilíř
    - tvůrci tedy nelitovali práce navíc, aby výsledek odporoval smlouvě i praktickým potřebám — pro to se těžko hledá rozumné vysvětlení
- vytvořil jsem si tedy obdobu H13 nad daty H14
  - když bylo zjevné, že nejde o provizorium, ale o finální stav
  - mohl jsem aspoň přidat funkce, které mi chyběly v H13 a starších
    - hledání podle více hodnot, podle bodů(?), vlastní řazení a výpis
- bohužel mi nedošlo, že nefunkčnost H12 a starších není náhoda
  - dřív bylo k dispozici mnoho let zpět; aspoň bych stáhl H13 a RIV

Jen pro vážné zájemce: oficiální H14

- prvních pár měsíců šlo kombinovat rok, druh a oborovou skupinu
  - + nezdokumentované fulltextové hledání v názvech výsledků a zdrojů, anotacích a klíčových slovech + OR, AND, NOT a "..."
  - z principu to tedy neumožňovalo ani hledání dle autorů
- cca v dubnu fulltext zahrnul i autory + přibyla nápověda
  - bylo nutno hledat "příjmení, jméno"či podle vedidk
  - hledání dle VO, oborů, ISSN/ISBN, bodů, ... stále nemožné
  - stejně tak kód hodnocení, tedy důvod hodnocení/vyřazení
    - v datech přitom poměrně podrobně specifikován
- čísla vlevo ukazují, že čtvrtina výsledků není v Pilíři I + III
  - k jejich nalezení buď nutno vědět vhodný dotaz pro fulltext
  - nebo sekvenčně procházet celou oborovou skupinu
  - dost možná významný překryv s výsledky mimo hodnocení
    - stránka, co „prezentuje Hodnocení výsledků výzkumných organizací“
    - běžně vracela výsledky s textem „nebyl zahrnut do Hodnocení 2014“
    - ⇒ při neomezení druhem zřejmě v podstatě šum

O rok později: H15, III. etapa

- H15e3 zveřejněno pouze v .xlsx ⇒ skripty jako když najde
  - v .xlsx se chyby mezi výskyty výsledku u různých VO hledají těžko
    - ČVUT jistě dělá totéž, ale nic o nich nevím, jinak těžko říct
  - postupně jsem k existujícímu přidával, co jsem zrovna potřeboval
- ⇒ *reklamní vložka, místy v kontrastu s rvvi.cz/riv*
  - informace o sjednoceném výsledku (jako v H14), srovnání s H14
  - vlastní výpis obohacen o ořezání dlouhých hodnot či bodové součty
  - skládání dotazů včetně expanzí výskytů na celé výsledky
  - seznamy oddělené jen mezerami (Ctrl+C/V z nějaké tabulky ap.)
  - vše v URL (HTTP GET), bez cookies/sessions, „Zpět“ dle referer
    - lze mailem poslat URL výsledku i vyplněného formuláře
    - pro hledání není k rvvi.cz sessions ap. důvod, je to dotaz–odpověď
      - navíc při hledání různých věcí z více karet se ty dotazy nějak ovlivňují, ale nepodařilo se mi to deterministicky nasimulovat
  - rvvi.cz je na mnoho typů dotazů velice pomalé
    - ve skutečnosti ale toto jediný praktický rozdíl: dle logů lidé pokládají téměř výhradně dotazy, které by zvládlo i rvvi.cz (autor/VO/druh)

H15, III. etapa

- na první verzi dat jsme (na MU, ale asi i jinde) strávili neuvěřitelné množství času, na fakultách ručně, já se snahou o automatizaci
  - zpětně se zdá, že KHV/RVVI si nepoužitelnost dat uvědomovali, jen se nějak zapomnělo pošeptat VO, že to nemusejí příliš hrotit
- jedním z výsledků byl dlouhý soupis vesměs systematických chyb
  - detaily už neaktuální, zvlášť, když ta data nebyla myšlena vážně
  - ale dokládá to přístup Sekce, jaké šílenosti se neleká zveřejnit
    - chyby jsou vždy, zvlášť v časovém presu, ale mnohé je jen šlendrián
    - a třeba ty hodnoty 42 ap. těžko mohly vzniknout náhodou
  - přijetí výpočtu by odpovídalo ruletě o cca 3 miliardy Kč
    - NIC se nestalo, případné jednotlivé chyby hlaste přes poskytovatele
    - — v jiné aféře stačila poloviční suma, aby po měsících padla vláda
- další verze a deadliny pro jejich reklamace šly rychle po sobě
  - většina energie šla na dohledávání chyb ve výsledcích MU
  - (na okraj: zdroje H14 shodné s rvvi.cz, starší nenabízejí ani oni)

H15, finální data — úvodní příklad: ukázka „digrese“

- disclaimer: jsem jen „knihovník“, co umí programovat :-)
  - nemám schopnosti ani ambice komentovat, jak by to *mělo* fungovat, z mého pohledu byla dána pravidla a podle nich se mělo něco stát
  - problém je, jsou-li pravidla formulována či chápána nejednoznačně
- např. hned výběr dat: s. 5 bere *záznamy*, s. 7 až pak konsoliduje
  - ve skutečnosti (i v H14 a dříve) se konsolidace dělá už v IS VaVaI
  - takže např. výsledek s rokem uplatnění 2009 dostane body
    - takových 109 (proč jen starší?), s body 47, viz ale rozdíl prvních dvou
    - často chybná sloučení — a absence kontrol extrémů (výskytů, let)
    - hned by viděli, že slučují různé ročníky/autory/UT WoS/...
      - v posledním navíc jiný počet výskytů — dvojí sjednocování?
      - často naopak (asi budiž, byť proti Metodice), obecně.zmatek
  - ale pozor, na rvvi.cz se zobrazí rok uplatnění 2010 — je změněn!
    - knihu šlo koupit 2009, kdy zřejmě i vyšla $\Rightarrow$ ČVUT to mělo dobře
    - dle RIVu jakoby špatně, takových 618, takto i počty.autorů, co ještě?
      - pak se říká, že to VO vykazují blbě — ano, ale tímto se to stírá
    - $\leq$H14 byly kopie dat, tam ať si, toto jsou změny v „hlavních“ datech...

H15, finální data

- probíhá kontrola záznamů dle s. 7 a Přílohy č. 9? I těchto?
- nedochází ke konsolidaci oborových skupin dle s. 8, viz zde
  - toto je ale úmyslné rozhodnutí — a je otázka, není-li to i rozumnější
    - navíc Metodika neřeší situaci, že toho dojde stejně ve stejný den
- jak jsou prováděny kontroly, když tyto články ve Scopusu nejsou?
  - ISBN tam není, ISSN naposledy 2010, jak to může mít body?
  - a třeba toto ve WoS je, ale s rokem 2015
    - to se přece taky mělo kontrolovat
  - toto se pochopitelně obtížně hledá obecně
    - zejména u těch, co na sebe nenapráskali UT WoS či Scopus EID
- zůstalo 414 výsledků, u nichž součet podílů přesahuje 1
  - kromě 14 jde vždy o kombinaci starých a nových záznamů
  - rekordman ovšem není tento případ, nevím, jak to vzniká
- tatáž VO má vícekrát body za tentýž výsledek (191 ks), i 3×
  - 81× nové věci, kdy se podíl VO dělí na dvě (ne nutně stejné) části

H15, finální data

- krásný příklad kombinace dvou systematických chyb
  - VÚVL to v H14 vykázal přes MŠMT a MZe a dostal body
  - kvůli špatnému roku to vykázal znovu v H15, MZe zrušil
  - pro názornost, snadný součet ap. jsou při více výskytech v datech body jen u náhodného z nich, zde MZe, ale „platí“ pro celý výsledek
  - nové výskyty (včetně spolupracujících VO) byly sjednoceny, k původnímu byla převzata nula, nikoli patřičné body
    - (jiná věc je, že předtím neměl ty body dostat)
  - co mělo v H14 0 a pak se objevilo ve WoS/Scopus, zůstalo s 0
    - je pravda, že na s. 6 je na konci 2. odstavce *mohou*
    - dosud ale byly nulové vždy znovu hledány a případně hodnoceny
      - jinak by je všichni vždy mazali a znovuvkládali, aby vyvolali případné (pře)hodnocení, zcela zbytečná a absurdní činnost
  - toto se nejen objevilo ve WoS, dokonce to tam našli: kód 1JI
    - proč je u starého výskytu 1JS, netuším, taky běžné
  - reklamace nemožná, protože prostě nereagovali

H15, finální data

- body za 1JI/J_{imp} zřejmě nejsou počítány správně
  - pokud jsem
    - neočišťoval o nepřiměřený podíl autocitací (s. 31 Metodiky)
    - při více ISSN se stejným IF nebral počítal průměr pořadí, jak odpověděli na dotaz při veřejné zakázce, ale to nejlepší
    - a počítal s přesností pouze na 5 desetinných míst
  - na 95 % ISSN z MU jsem došel ke stejným výsledkům
  - někdy se to nevysvětlitelně liší
    - v JCR 2014 je 3. z 92, čemuž by odpovídalo cca 218 bodů
- mnoho a mnoho dalšího, ale dělat to pořádně je hrozně pracné
  - přechod od příkladu ke všem instancím je těžký, protože se ty chyby různě mixují
  - no a co potom s tím, když H15 už je uzavřené?
- je to dilema i vzhledem k H16
  - člověk bude nepřipraven a času na reklamace bude málo
  - ztratí se tím teď hromada času a pak to bude úplně jinak
    - (prý se to učí ap. a příště to už budou umět)