Bibliometrická analýza výzkumníků s použitím MS Access a R

doc. Ing. Pavel Šenovský, Ph.D.
VŠB – Technická univerzita Ostrava, Fakulta bezpečnostního inženýrství, Katedra ochrany obyvatelstva, pavel.senovsky@vsb.cz

Anotace
Tento whitepaper obsahuje popis shromažďování dat potřebných pro bibliometrickou analýzu zvoleného autora od shromáždění dat o jeho publikacích a citacích až po generování reportu a interpretaci vypočtených metrik. Pro usnadnění analýz jsou volně dostupné přílohy obsahující databázi v MS Access určenou pro shromáždění potřebných dat a jejich následný export a analytický skript v R markdown umožňující tato data zpracovat a vygenerovat samotný report.

Klíčová slova
Bibliometrická analýza, MS Access, R, R markdowm, h-index, h-i index, g-index, i10 index, WoK, SCOPUS

Obsah

1 Úvod do bibliometrických analýz

Účelem *bibliometrických analýz* je získat představu o charakteristikách publikačního výkonu autora, instituce nebo např. časopisu. Účelem takových analýz může být získání informací použitelných v „kvantitativní“ části hodnocení autora nebo instituce, identifikace oblastní výzkumu instituce se silnými dopady na vědeckou komunitu apod.

Analýzy tohoto typu jsou požadovány pro kariérní postup v akademické sféře a také jsou využívány v zjednodušené formě pro prokazování kompetencí osob podílejících se na výuce pro účely akreditace studijních programů.

Důvodů pro realizaci bibliometrických analýz tedy může být celá řada a představují jednak základní motivaci pro realizaci analýzy, jednat také s sebou mohou nést jistá omezení např. ve smyslu volby *datasetu* pro realizaci analýzy.

Z hlediska zdroje dat jsou v akademickém prostředí rozlišovány obvykle tři typy:

- Web of Knowledge (WoK) – https://webofknowledge.com/
- SCOPUS – https://www.scopus.com
- ostatní

V případě WoK pak mohou existovat ještě další omezení na použití pouze určité dílčí databáze, nejčastěji pak Web of Science Core Collection.

Ostatními se pak rozumí ostatní zdroje bibliometrických informací jako je Google Scholar, ResearchGate, CiteSeer a řada dalších.

Pro účely analýzy se v tomto whitepaperu zaměříme pouze na analýzu jednotlivých autorů. Pro takový typ analýz má smysl sledovat např. následující metriky:

- počet publikací autora v datasetu
- publikační aktivita autora v čase (např. počty publikací v letech)
- celkový počet citací
- počet citací v letech
- indexy:
    - Hirshův index (h-index)
    - h-i index – normalizovaný Hirshův index
    - g-index
    - i10 index

Některé metriky jsou poměrně přímočaré – např. *počet citací*. Tato metrika říká, kolikrát byl výzkum publikovaný v citovaném článku použit v dalším výzkumu. Sledování vývoje počtu citací v čase lze získat představu o rychlosti přejímání výsledků výzkumu autora a také „délce života“ takového výsledku.

Pro většinu publikovaného výzkumu lze předpokládat nejprve nárust aktivity (počtu citací v čase), následně jejich stagnace, následovaná postupným úpadkem aktivity. Rychlost nástupu obvykle souvisí se způsobem publikace, jako je dostupnost on-line, zaindexování v publikace v databázi, existence paywall apod. Citovanost jako celek pak závisí na celé řadě dalších faktorů od věhlasu autora, aktuálnosti tématu, závažnosti publikovaných poznatků, formy jejich prezentace nebo prostého štěstí.

V případě citací se obvykle předpokládá, že dataset neobsahuje autocitace, nebo že autocitace je možno v datasetu identifikovat a v případě potřeby z analýzy vyřadit.

Z hlediska hodnocení se také používá celá řada *indexů* umožňujících získat celkový přehled o významnosti publikační činnosti analyzovaného autora. Získané indexy je následně možno použít také pro porovnání napříč autory (např. v rámci instituce) nebo nastavení základních kvantitativních požadavků např. pro kariérní postup (takové hodnocení by ale vždy mělo být doplněno také hodnocením dalších aktivit autora).

Při srovnávání je potřeba vždy pamatovat na to, že indexy nejsou univerzálně srovnatelné – porovnatelní jsou pouze autoři, kteří publikují ve stejném oboru nebo v různých oborech, které ale mají stejné publikační zvyklosti.

Nejčastěji používaným citačním indexem je tzv. *Hiershův index*, neboli *h-index*. Tento index udává počet článků autora, jejichž počet citací je roven, nebo vyšší než *h*. Prakticky tedy pro h-index 5 autor musí mít minimálně 5 článků s 5-ti nebo více citacemi, pro h-index 10 pak minimálně 10 článků s 10-ti citacemi.

H-index tedy roste lineárně. Vyšší hodnotu indexu mají autoři s větším množstvím citovaných článků. S h-indexem je ale spojeno také několik problémů.

Prvním z nich je, že význam jednotlivých publikací obvykle není stejný. Jako příklad můžeme uvést Petera W. Higgse. Jeho odhadovaných celkový h-index je okolo 11, pokud bychom ale dataset omezili pouze na WoK, tato úroveň by nebyla dosažena, neboť autor nebyl publikačně dostatečně činný. Znamená to, že se jedná o nevýznamného autora? Očividně nikoliv, ačkoliv i na toto se názory liší. Higgs je laureátem Nobelovy ceny za fyziku (2013) za studii předpovídající existenci částice pro niž se vžilo jméno Higgsův boson, jejíž existence byla experimentálně potvrzena v roce 2012.

H-index tedy neposkytuje celkový obrázek o významu autora. Pro zohlednění celkového dopadu nejcitovanějších publikací autora lze použít alternativní měřítko – *g-index*.

Druhým problémem jsou publikace, na jejichž vzniku se podílely rozsáhlé vědecké týmy. Takové publikace jsou obvykle velmi obsáhlé, s desítkami, v některých případech i stovkami autorů a jsou obvykle silně citované. Takové publikační zvyklosti jsou pro experimentální fyziku, ale také řadu dalších oborů. H-index v takovém případě nepopisuje vědecký výkon autora, protože relativní příspěvek autora ke vzniku takové publikace je logicky velmi malý, ale spíše schopnost autora zapojit se do velkých, významných výzkumných skupin.

To samo o sobě není na škodu, pokud s takovou možností při analýze počítáme. Pokud je ale cílem zjištění skutečného vědeckého výkonu autora, pak je nutno použít jiný typ metriky, např. *h-i index*.

H-i index je h-index normalizovaný na počet autorů. Vypočítat jej lze pomocí vzorce (1).

$$hi = \frac{h^2}{n_{auth}} \tag{1}$$

Kde h-i index *hi* se vypočte jako druhá mocnina h-indexu *h* na počet autorů těchto publikací n_{auth}.

Ze vzorce (1) lze dovodit, že $h = hi$ pouze pokud publikace vstupující do hodnocení autorsky připravila jediná osoba, v ostatních případech bude platit, že $h > hi$.

Požadavky *g-indexu* rostou exponenciálně (proti lineárnímu růstu požadavků h-indexu). Pro hodnotu *g* indexu je potřeba g^2 citací v *g* článcích. Např. pro $g = 10$ je potřeba, aby součet citací z 10-ti nejcitovanějších publikací autora byl minimálně 100.

Tím, že je brán součet citací vstupují do hodnocení všechny citace nejvíce citovaných článků autora. Takže např. pokud by autor publikoval jedinou studii, která by ale měla 1 000 citací, jeho g-index by byl 31 ($31^2 = 961$), ale jeho h-index by byl pouze 1.

Konečně *i10 index* odpovídá počtu publikací autora v datasetu, které přesáhly hranici 10-ti citací.

2 Příprava bibliometrických dat pro analýzu

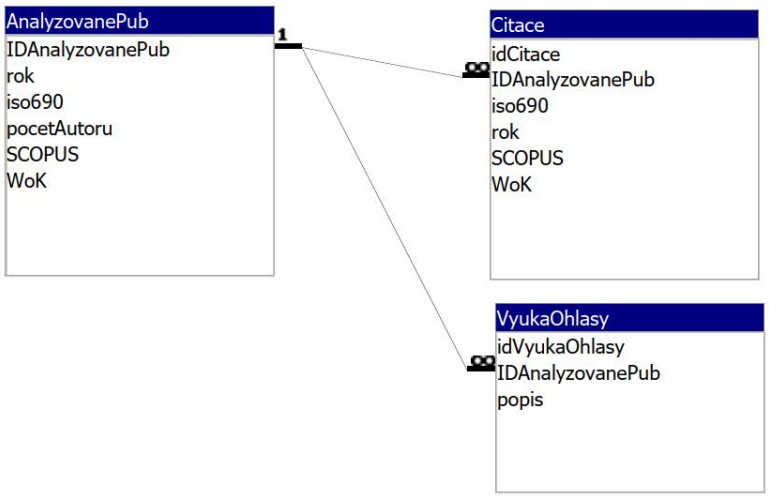
V případě, že cílem je provedení analýzy pouze základních bibliometrických ukazatelů na základě datasetu jedné z populárnějších citačních databází, je obvykle jednodušší použít přímo její vestavěné analytické nástroje.

Pokud je ale potřeba získat přehled na základě dat, která nejsou výhradně z jednoho zdroje je obvykle lepší připravit si externí dataset obsahující zájmová data o publikacích autora a jejich citacích.

Pro účely tohoto whitepaperu byla navržena jednoduchá databáze v systému MS Access. Tato databáze byla zvolena pro svou schopnost manipulovat s daty, stejně jako dostupností elegantních nástrojů pro návrh formulářů, které umožní pohodlný přístup uživatele k bibliometrickým datům.

Technicky ale není problém použít jakoukoliv jinou databázi. Cílem je pouze připravit data k exportu v konzistentní formě.

Struktura báze dat je znázorněna na obr. 1.



Obr. 1: Struktura báze dat pro evidenci citací a jiných ohlasů

Základ databáze tvoří tabulka *AnalyzovanePub*, která je určena pro evidenci publikací určených pro analýzu. Použitá datová struktura je velmi jednoduchá – předpokládá, že analyzovány budou všechna data, která jsou obsažena v databázi.

Tabulka má následující strukturu:

- IDAnalyzovanePub – automatické číslo, identifikátor publikace používaný pro identifikaci odkazované publikace v citacích a ohlasech
- Rok – rok publikování

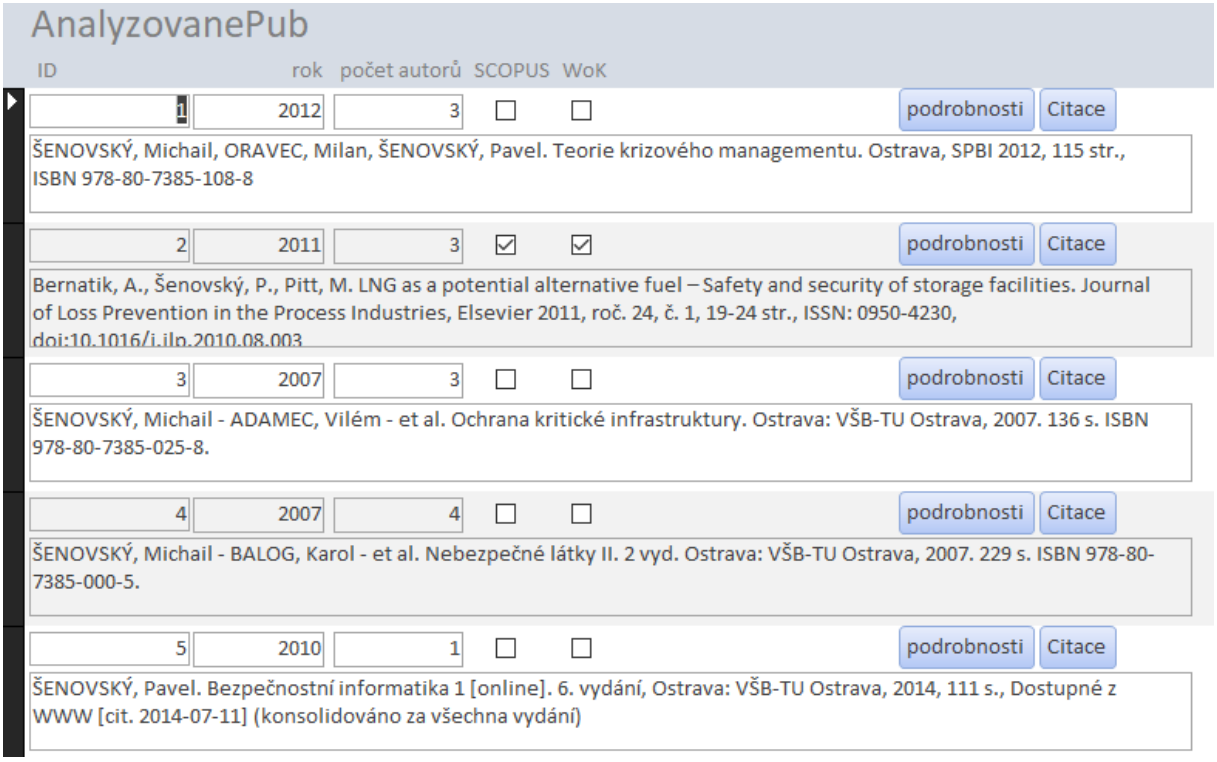
- Iso690 – citace publikace ve formátu ISO 690. Technicky se jedná o volný text, takže případná citace může být v jakémkoliv formátu. Údaj se používá např. pro výpis počtu citací vztažených k publikaci.
- pocetAutoru – počet autorů publikace. Tento údaj se používá pro výpočet h-i indexu.
- SCOPUS – A/N – je publikace evidována ve SCOPUS?
- WoK – A/N – je publikace evidována na WoK?

Tabulka *Citace* má následující strukturu:
- idCitace – identifikátor citace, automatické číslo, nevstupuje do analýzy
- IDAnalyzovanePub – identifikátor publikace, která je citována
- Iso690 – citace citující publikace ve formátu ISO 690.
- Rok – rok publikování
- SCOPUS – A/N – je publikace evidována ve SCOPUS?
- WoK – A/N – je publikace evidována na WoK?

Konečně tabulka *VyukaOhlasy* má následující strukturu:
- idVyukaOhlasy – identifikátor ohlasu
- IDAnalyzovanePub – identifikátor publikace, na kterou je zaznamenán ohlas
- Popis – text samotného ohlasu

Ohlasy se v tomto případě rozumí jakékoliv ohlasy na publikaci, které nemají charakter citace. Z hlediska použití se jedná o doplňkovou informaci, kterou databáze umožňuje evidovat, ale ani samotná databáze, ani analytický skript s touto informací dále nepracuje.

Z hlediska dostupného grafického uživatelského rozhraní jsou v databázi připraveny 2 základní formuláře:
- Analyzované publikace – obsahuje přehled všech publikací analyzovaného autora bez citací a ohlasů, viz obr. 2
- Publikace – obsahuje formulář publikace autora s podformuláři citace (viz obr. 3) a ohlasy (viz obr. 4)

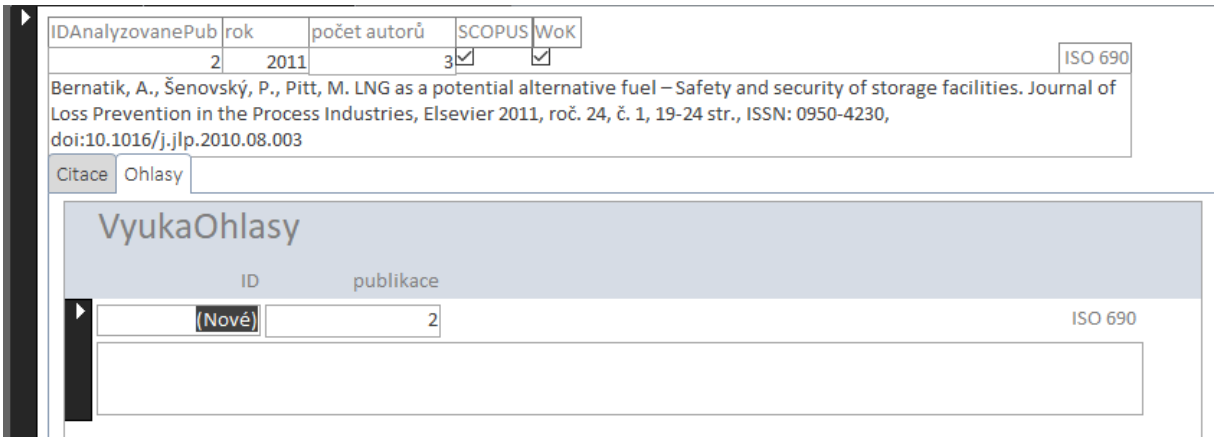


Obr. 2: Seznam publikací analyzovaného autora

Obr. 3: Podrobnosti analyzované publikace s podformulářem citací

Údaje v podformuláři citací jsou řazeny, podle pořadí, v jakém byly do databáze vkládány, nikoliv např. podle roku. Řazení podle odlišných položek je možno kdykoliv zapnout kliknutím do pole podle, kterého se má řadit a kliknutím na ikonu *vzestupně,* popř. *sestupně* na ovládacím panelu, v záložce *domů*, části *seřadit a filtrovat*.

Při editaci je potřeba dát pozor na položku *ID odkazované p.* Jedná se identifikátor publikace, která je citována. Pokud tedy bude toto číslo přepsáno, citace se odebere původní odkazované publikaci a připojí se k nové, podle specifikovaného ID.



Obr. 4: Podrobnosti analyzované publikace s podformulářem ohlasů

Formulář s ohlasy funguje analogicky k formuláři s citacemi.

V databázi jsou dostupné také některé dodatečné nástroje pro přímé získání některých údajů. Např. je připravena sada výběrových dotazů umožňující zjistit počet citací. Tento typ údajů je ale jednoduše zjistitelný pomocí analytického skriptu.

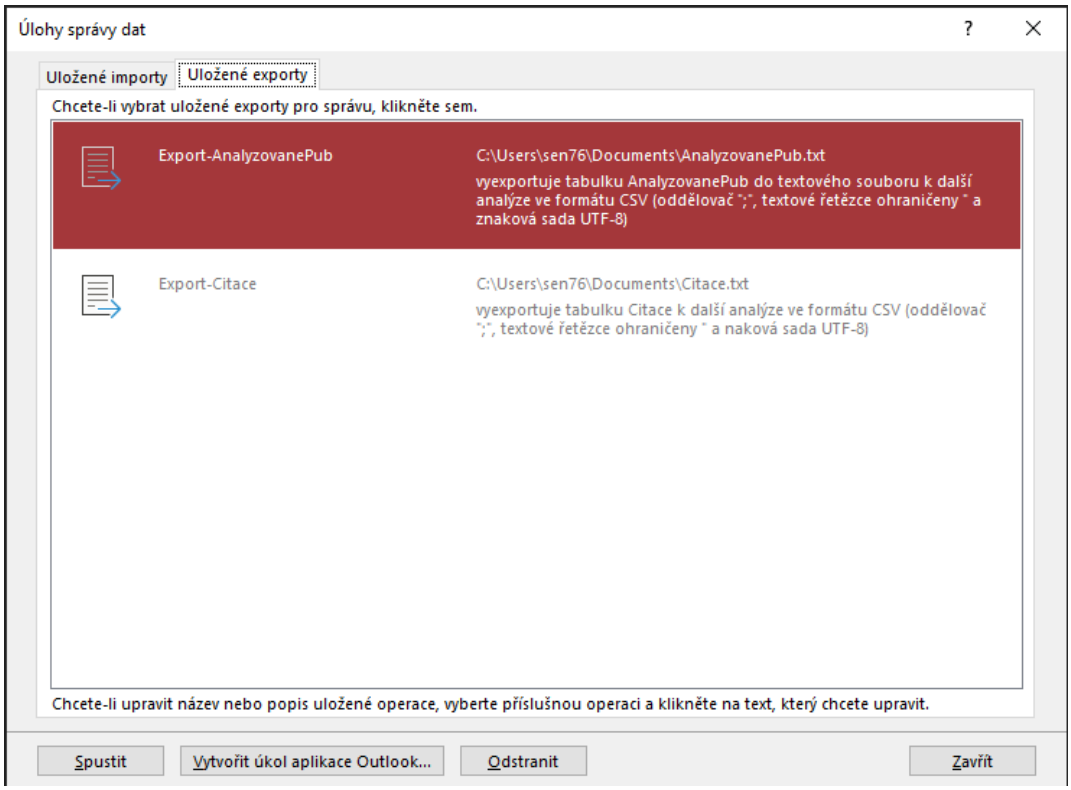
Čistě servisní charakter má dotaz *kontrola citace rok=0*. Tento dotaz zobrazí seznam citací, které nemají vyplněný rok. Pokud jsou údaje správně vyplněny, neměl by dotaz vrátit žádný údaj.

Chybně zavedenou citaci lze najít z položky IDAnalyzovanePub, která identifikuje publikaci autora, ke které je chybně zadaná citace zavedena, a idCitace, které přímo identifikuje chybně zavedenou citaci.

Po provedení opravy spusťte dotaz znovu, abyste ověřili, že problém byl skutečně odstraněn.

Vyplněním údajů do formulářů a jejich kontrolou konzistence jste připravili data pro analýzu. Tato data je pouze potřeba exportovat tak, aby je bylo možno pohodlně analyzovat pomocí R.

K tomuto účelu jsou v databázi připravené uložené exporty, které je možno spustit z lišty nástrojů – záložka *externí data – uložené exporty*. Kliknutím na ikonu se spustí dialogové okno jako na obr. 5.



Obr. 5: Uložené exporty tabulek

Analytický skript v současnosti používá pouze data z tabulek AnalyzovanePub a Citace. Těmto tabulkám odpovídají uložené exporty *Export-AnalyzovanePub* a *Export-Citace*. Je potřeba postupně vybrat a spustit oba exporty.

Výsledkem exportů jsou soubory ve formátu CSV (Comma Seprated Values) *AnalyzovanePub.txt* a *Citace.txt*. Soubory budou exportovány do složky, ve které se nachází databáze, ze které byl export spuštěn. Oba tyto soubory jsou potřeba pro provedení analýzy.

V případě, že v databázi provedete změny je potřeba export opakovat, tak aby analytický skript dostal příležitost pracovat s aktuálními údaji.

3 Bibliometrická analýza pomocí R

Analytický skript byl připraven v prostředí R. Pro jeho spuštění jsou tak potřeba:

- Samotné prostředí R dostupné z https://www.r-project.org/
- RStudio, dostupné z https://rstudio.com/, postačuje open source licence nástroje RStudio Desktop.

Skript je dostupný v souboru *citace.Rmd*. Skript byl připraven metodou literate programing v jazyku R markdown. Metoda je výhodná v tom, že v sobě kombinuje možnost psát volným textem komentáře a doplňovat je fragmenty kódu, které doplňují funkčnost celého výstupu.

Tímto způsobem je tak možno poměrně elegantně generovat reporty s tabulkami a grafy, případně dalšími informacemi prezentovanými volným textem.

Při prvotním otevření skriptu v RStudio, budete vyzváni k doinstalování potřebných komponent pro R markdown. Po povolení instalace budou všechny potřebné komponenty staženy automaticky z repozitářů R a nainstalovány na počítač. Rozhraní RStudio je znázorněno na obr. 6.

K vykonání skriptu je potřeba doinstalovat také některé další knihovny, což lze provést zadáním jednotlivých řádků z výpisu 1 do konzole R a spuštění stisknutím klávesy enter, viz obr. 7. Jedná se o jednorázovou operaci, která si patřičnou knihovnu automaticky stáhne s repozitářů R a naistaluje.

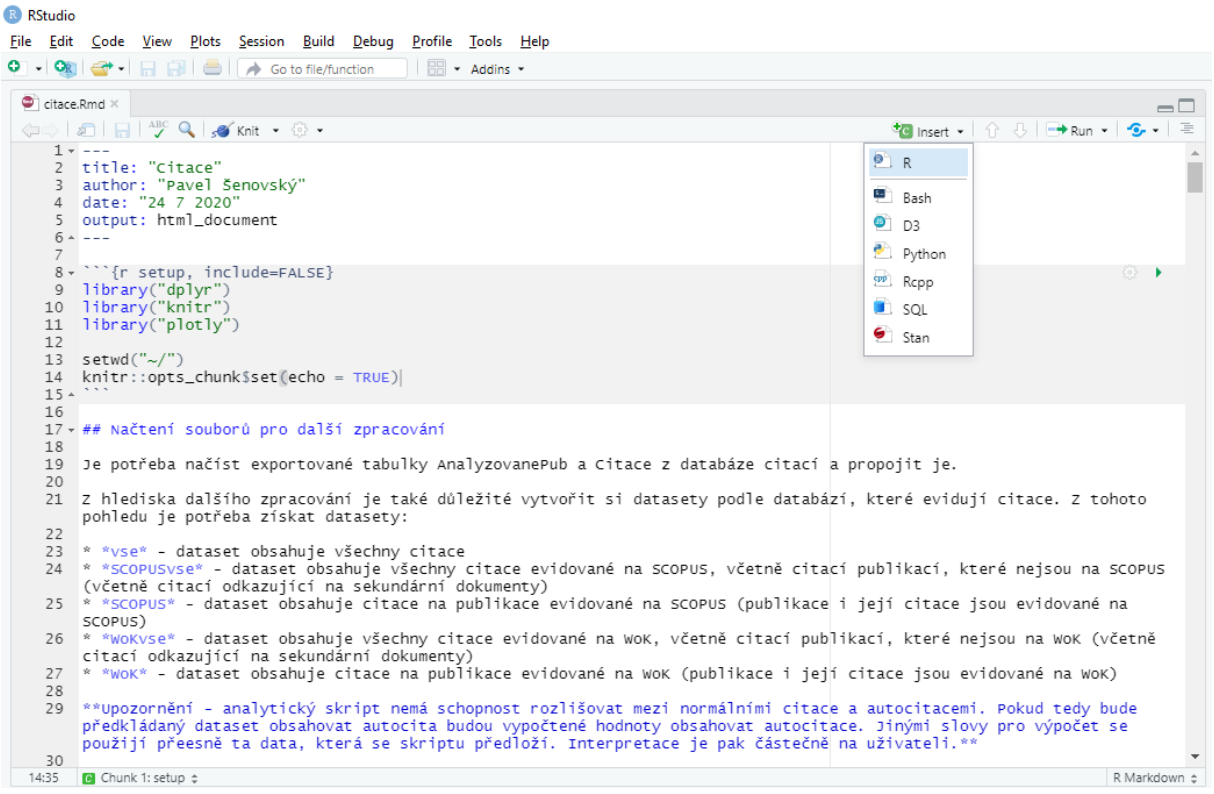
Výpis 1: Instalace knihoven do R

```
install.packages("dplyr")
install.packages("plotly")
```

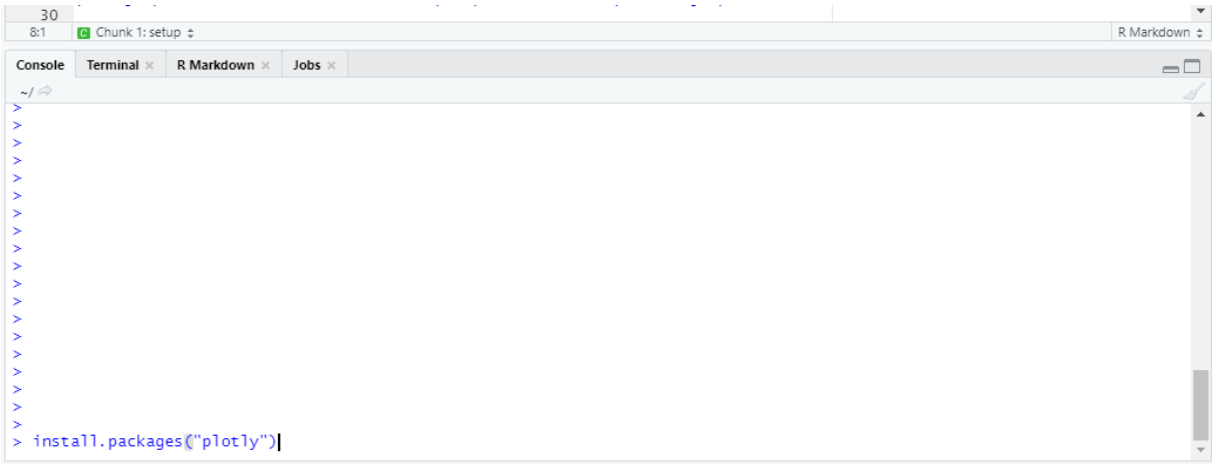
Pokud chcete použít pouze předpřipravený skript ke zpracování dat jediné, co je potřeba udělat, je kliknout na tlačítko *Knit* (viz obr. 6), které vygeneruje report na základě předpřipravených dat CSV souborech (viz předchozí kapitola).

Report se implicitně generuje do podoby webové stránky. „Rozkliknutím“ ikony *Knit* (viz obr. 8) je ale možné toto chování změnit a report si vygenerovat třeba do formátu PDF nebo MS Word.

Zbytek textu v této kapitole je věnován podrobnějšímu popisu R markdown a ve skriptu použitých postupů výpočtu pro případ, že by skript bylo potřeba upravit.



Obr. 6: RStudio – rozhraní pro manipulaci s R markdown soubory



Obr. 7: Konzole R – instalace knihoven

Obr. 8: Možnosti generování reportů pomocí nástroje Knitr

Každý skript v R markdown začíná hlavičkou, která vypadá podobně jako ve výpisu 2.

Výpis 2: Hlavička R markdown

```
---
title: "Citace"
author: "Pavel Šenovský"
date: "24 7 2020"
output: html_document
---
```

Hlavička dokumentu začíná a končí třemi pomlčkami, mezi kterými jsou klíčová slova. Title obsahuje název skriptu, author obsahuje jméno autora, date obsahuje datum poslední aktualizace skriptu a konečně output obsahuje informaci, do jakého formátu se má výsledný report generovat. Podporovány jsou:

- html_document – generování do WWW stránky, implicitní výstup
- pdf_document – výstup do PDF formátu
- word_document – výstup do formátu MS word (.docx)
- beamer_presentation – prezentace ve formátu PDF
- ioslides_presentation – prezentace ve formátu HTML (webové stránky)

K jednotlivým cílovým formátům je potřeba dodat, že ně všechny jsou si úplně rovny. Např. výstup do webové stránky zachovává schopnost manipulovat s grafy ve smyslu možnosti

průzkumu jednotlivých datových bodů, výběru určité oblasti grafu a možnosti uložení grafu do podoby bitmapy ve formátu PNG. Tyto možnosti se při výstupu do formátu PDF, nebo MS Word ztrácejí.

Zbytek dokumentu je tvořen segmenty kódu a volným textem. Segment kódu ve v editoru vždy podbarven šedivou barvou (viz řádky 8 – 15 na obr. 6). Sémantická struktura takového segmentu je zachycena na výpis 3.

Výpis 3: Struktura segmentu kódu v R

```
```{r název, include=FALSE}

Kód v R

```
```

Segment kódu je vždy ohraničen trojicí znaků „backtick“ (`). První řádek segmentu má ale ve složených závorkách specifikovány další parametry. Prvním z nich je jazyk, ve kterém je segment napsán, v tomto případě r. Následuje mezera a název segmentu. Název může být víceslovný, může také obsahovat interpunkci, ale nesmí obsahovat čárku. Ta je používána pro oddělování parametrů.

Poslední parametr je include. Tento parametr je implicitně nastaven na TRUE, což znamená, že segment kódu bude vykonán a zároveň se přepíše v textové formě do výsledného reportu. Pokud takové chování vyhovuje, lze parametr include vynechat.

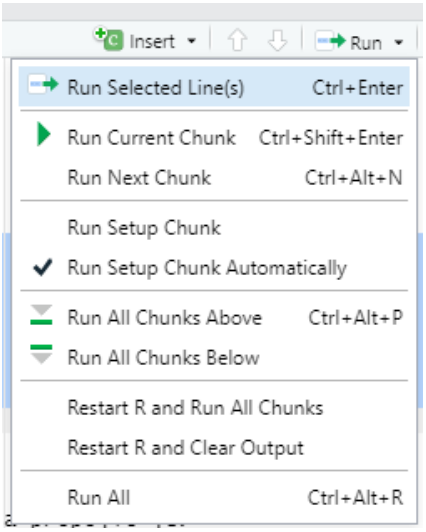
Pokud segment kódu nechceme mít ve výstupu, je možno nastavit include=FALSE a daný segment se pouze vykoná (do reportu se ale nepřidá). Tento postup se používá často pro servisní části kódu jako je připojování knihoven apod.

Volný text je formátován pomocí jednoduchého značkovacího jazyka:

- # nadpis – za znakem # musí vždy následovat mezera, počet # pak odpovídá úrovni nadpisu (# nadpis první úrovně, ### nadpis třetí úrovně)
- [link]{www.vsb.cz} – vložení URL odkazu do textu
- *text kurzívou* nebo _text kurzívou_
- **text tučně** nebo ___text tučně___
- Vložení rovnice: $A = \pi*r^2$
- Vložení statického obrázku:

Podporovány jsou také tabulky, odrážky, číselné seznamy apod. Krátký návod pro R markdown je možno nalézt na adrese: https://rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf.

Jednotlivé segmenty kódu je možno spouštět buďto samostatně, kliknutím na zelenou šipku na pravém okraji prvního řádku segmentu nebo kliknutím na tlačítko *Run* a volbou toho, co přesně se má přepočítat, viz obr. 9.



Obr. 9: Spouštění kódu

Při spouštění kódu je potřeba pamatovat na to, že jednotlivé segmenty nemusí nutně být na sobě nezávislé. Změna kódu v segmentu si tak může vynutit přepočítání segmentů, které po něm následují (k tomuto účelu slouží volba *Run All Chunks Below* na obr. 9), nebo je dokonce „rozbít" zavedením nekompatibilních změn – např. odstraněním sloupce v tabulce, se kterým další segmenty kódu pracují.

Skript obsahuje řadu segmentů, které zpracovávají postupně data a generují výsledky. Základním je segment ***import CSV***, který provádí import dat z CSV souborů vyexportovaných z databáze (viz kapitola 2) a generování základních datových sad:

- *vse* - dataset obsahuje všechny citace
- *SCOPUSvse* - dataset obsahuje všechny citace evidované na SCOPUS, včetně citací publikací, které nejsou na SCOPUS (včetně citací odkazující na sekundární dokumenty)
- *SCOPUS* - dataset obsahuje citace na publikace evidované na SCOPUS (publikace i její citace jsou evidované na SCOPUS)
- *WoKvse* - dataset obsahuje všechny citace evidované na WoK, včetně citací publikací, které nejsou na WoK (včetně citací odkazující na sekundární dokumenty)
- *WoK* - dataset obsahuje citace na publikace evidované na WoK (publikace i její citace jsou evidované na WoK)

Výše uvedené datasety jsou využívány pro provádění dalších kroků analýzy. Z hlediska fungování segmentu se nejprve pomocí funkce read.table načtou data z CSV souborů. Předpokládá se přitom, že první řádek souboru obsahuje hlavičku, jako oddělovače jsou použity středníky, textová pole jsou ohraničena uvozovkami a je použito kódování znaků UTF-8. Výše uvedené nastavení odpovídá nastavení uložených exportů v databázi MS Access.

Základní dataset *vse* je sestaven vnitřním spojením tabulek publikace a citace přes ID analyzované publikace. Vzhledem k tomu, že obě původní tabulky obsahují stejně se

jmenující položky (ale s odlišným významem), je použita funkce *rename* pro přejmenování jednotlivých sloupců.

Zbývající datasety jsou odvozovány z datasetu *vse* pomocí aplikací filtru na sloupců signalizujících přítomnost publikace, popř. její citace v některé ze sledovaných databází.

Generování seznamu publikací se spočteným počtem citací pro jednotlivé publikace se generuje pomocí kódu znázorněného ve výpisu 4.

Výpis 4: Generování seznamu publikací s citacemi – příklad pro dataset WoK

```
tabWoKVse = WoKvse %>%
  select(rokPublikovani, iso690, pocetAutoru) %>%
  group_by(iso690, rokPublikovani, pocetAutoru) %>%
  summarise(pocetCit = n()) %>%
  arrange(desc(pocetCit), desc(rokPublikovani))
tabWoKVse %>%
  select(-pocetAutoru) %>%
  kable(caption="Citace evidované ve WoK (včetně sekundárních publikací)")
```

Tabulky obsahující rok vydání publikace, citaci ve formátu ISO 690 a počtu citací se generují pro všechny datasety obdobně. Výpis 4 tak obsahuje příklad pro dataset WoKvse. Pro další zpracování je odvozen také další dataset tabWoKVse, který obsahuje navíc také počet autorů publikace. Tato informace je využívána v dalších segmentech pro výpočet h-i indexu.

Dataset je seřazen podle počtu citací a podle roku publikování, obojí sestupně.

Do výpisu tabulky generovaného funkcí kable se už nepoužívá sloupec počet autorů.

Příklad výstupu je znázorněn v tab. 1.

Tab. 1: Citace evidované ve WoK (včetně sekundárních publikací)

iso690	**rok-Publiko-vani**	**po-cet-Cit**
Bernatik, A., Šenovský, P., Pitt, M. LNG as a potential alternative fuel – Safety and security of storage facilities. Journal of Loss Prevention in the Process Industries, Elsevier 2011, roč. 24, č. 1, 19-24 str., ISSN: 0950-4230, doi:10.1016/j.jlp.2010.08.003	2011	19
ŘEHÁK, D., ŠENOVSKÝ, P., HROMADA, M., LOVĚČEK, T. Complex Approach to Assessing Resilience of Critical Infrastructure Elements. International Journal of Critical Infrastructure Protection. Vol. 25, June 2019, pp. 125-138, DOI: 10.1016/j.ijcip.2019.03.003	2019	9
…	…	…

Počty citací v letech jsou počítány opět pro všechny datasety pomocí kódu analogickému výpisu 5.

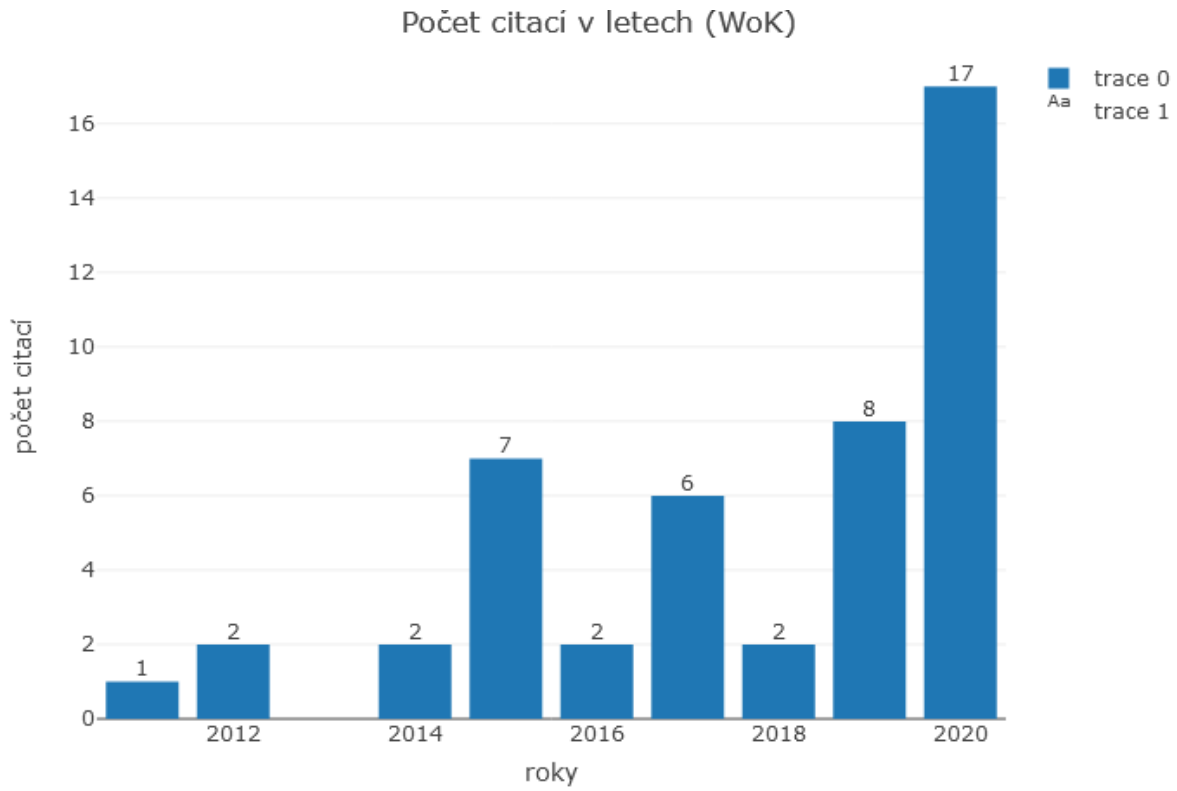
Výpis 5: Počet citací v letech

```
t = WoK %>%
  select(rokCitace) %>%
  group_by(rokCitace) %>%
  summarise(pocetCit = n()) %>%
  arrange(rokCitace)
fig <- plot_ly(data = t,
```

```
            x = ~t$rokCitace,
            y = ~t$pocetCit,
            type = "bar") %>%
        layout(
          title = "Počet citací v letech (WoK)",
          xaxis = list(title = "roky"),
          yaxis = list(title = "počet citací")) %>%
        add_text(
          text = ~t$pocetCit,
          textposition = "top middle",
          cliponaxis = FALSE
        )
fig
```

Jako základ výpočtu se používá vždy patřičný dataset, ve výpisu 5 je použit dataset obsahující primární publikace evidované ve WoK. Z tohoto datasetu se používá pouze sloupec rokCitace, na který je aplikovaná agregační funkce *n* umožňující zjistit, kolik citací se objevilo v daném roce.

Výsledek výpočtu je následně vykreslen pomocí knihovny *plotly*, do podoby sloupcového grafu. Příklad výsledku operace je dostupný na obr. 10.



Obr. 10: Počet citací v letech - příklad

Výpočet h-indexů je realizován pomocí funkce, viz výpis 6.

Výpis 6: Výpočet h-indexu

```
h_index = function(tab){
  hIndex = 0
  nAutori = 0
```

```
  i = 1
  for(cit in tab$pocetCit){
    if(cit > hIndex) {
      hIndex = hIndex + 1
      nAutori = nAutori + tab$pocetAutoru[i]
      i = i + 1
    } else { break }
  }
  result = c(hIndex, nAutori)
}

hVse = h_index(tabVse)
hSCOPUS = h_index(tabSCOPUS)
hWoK = h_index(tabWoK)
```

Funkce v sobě obsahuje jednak výpočet samotného h-indexu, jednak počet autorů, kteří se podíleli na přípravě publikací použitých pro výpočet h-indexu.

Výpočetní funkce využívá datasety odvozené pro zobrazení výpisu publikací s počtem citací. Využívá přitom, že data v takových datasetech jsou už agregovaná a seřazená podle počtu citací.

Výpočet samotný probíhá pomocí cyklu. Procházeny jsou postupně publikace od nejcitovanější směrem dolů a je porovnáváno, zda počet citací právě vyhodnocované publikace je roven nebo vyšší než pořadí dané publikace. H-index odpovídá pořadí poslední publikace, která takovou podmínku splňuje.

V cyklu se zároveň postupně načítají počty autorů, kteří se podíleli na přípravě daných publikací. Tento údaj je pak využíván pro výpočet h-i indexu.

Výpočet h-i indexu je pak jednodušší, viz výpis 7.

Výpis 7: Výpočet h-i indexu

```
hi_index = function(hIndex) {
  result = round((hIndex[1]^2)/hIndex[2], digits = 2)
}
hiVse = hi_index(hVse)
hiSCOPUS = hi_index(hSCOPUS)
hiWoK = hi_index(hWoK)
```

Jelikož h-i index je pouze h-indexem normalizovaným na počet autorů, všechny podklady potřebné pro výpočet byly připraveny při přípravě h-indexu. Funkce v tomto segmentu pak pouze tyto údaje vezme a použije, viz vzorec (1).

G-index je z hlediska výpočtu trochu složitější nežli h-index, viz výpis 8. Složitost vyplývá z toho, že do výpočtu vstupují všechny citace nejcitovanějších publikací autora. G-index tak může být vyšší, než je počet publikací, které analyzovaný autor napsal.

Výpis 8: Výpočet g-indexu

```
g_index = function(tab){
  citace = 0
  i = 1
  for(cit in tab$pocetCit){
    if((cit + citace) >= (i*i) ){
      citace = citace + cit
```

```
      i = i + 1
    } else {
      return(i)
    }
  }
  return(floor(sqrt(citace)))
}
gVse = g_index(tabVse)
gSCOPUS = g_index(tabSCOPUS)
gWoK = g_index(tabWoK)
```

Výpočet tak probíhá tak, že publikace jsou procházeny v cyklu postupně od těch nejvíce citovaných směrem dolů, podobně jako v případě h-indexu. Vyhodnocováno je, zda-li kumulativní počet citací je větší nebo roven druhé mocnině pořadí. Výpočet končí, pokud tato podmínka není splněna nebo se došlu na konec seznamu publikací.

G-index je vypočítán jako druhá odmocnina kumulovaného počtu citací, zaokrouhlená dolů.

Hrubý počet citací představuje pouze agregovanou hodnotu počtu citací napříč datasety, viz výpis 9.

Výpis 9: Hrubý počet citací – příklad pro WoK

```
nWoK = tabWoK %>%
  group_by() %>%
  summarize(n = sum(pocetCit))
```

Koneční *i10* index je vypočítáván aplikací filtru na datasety publikací počet$Citaci \geq 10$ a použití agregační funkce *n* pro zjištění toho, kolik publikací tuto podmínku splňuje, viz výpis 10.

Výpis 10: Výpočet i10 indexu na datasetu WoK

```
i10WoK = tabWoK %>%
  filter(pocetCit >= 10) %>%
  group_by() %>%
  summarize(n = n())
```

Závěr

V předchozím textu byly popsány některé metriky a postupy pro vyhodnocování bibliometrických údajů o jednotlivcích. Jedná se o poměrně mocný nástroj, o kterém ale platí ono okřídlené *„dobrý sluha, ale zlý pán“*.

Existuje sada doporučení, které mají za cíl tyto negativní stránky hodnocení minimalizovat – v roce 2015 byly shrnuty do tzv. *Leidenského manifestu*, který specifikuje 10 základních pravidel pro provedení kvalitního hodnocení:

1. Kvantitativní hodnocení by mělo sloužit jako podpora kvalitativního, odborného posouzení.
2. Měřte výkonnost ve vztahu k výzkumným cílům instituce, skupiny nebo výzkumníka.

3. Je třeba chránit vynikající výzkum regionálního významu.
4. Sběr a analýza dat by měly být otevřené, transparentní a jednoduché.
5. Ti, kteří jsou hodnoceni, by měli mít možnost ověřit data a analýzy.
6. Je nutné zohledňovat rozdíly mezi obory v publikační a citační praxi.
7. Hodnocení jednotlivých výzkumníků by mělo být založeno na kvalitativním posouzení jejich portfolií.
8. Vyhněte se nemístné konkrétnosti a falešné přesnosti.
9. Věnujte pozornost vlivu hodnocení a indikátorů na systém.
10. Indikátory by měly být pravidelně přezkoumávány a aktualizovány.

Celý text manifestu v češtině: http://www.leidenmanifesto.org/uploads/4/1/6/0/41603901/leiden_manifesto_cz.pdf.