



Foto: Carlos Muza, unsplash <https://unsplash.com/photos/hpJSKU2UYSU>

# Otvorená veda v praxi: Otvorené vedecké dáta

prezenčný kurz

Centrum vedecko technických informácií SR

17. 10. 2023

9.30 - 12.00

Zuzana Stožická

Matej Harvát

**NISP EZ<sup>IV</sup>**  
Investícia do Vašej budúcnosti



EURÓPSKA ÚNIA  
Európske štrukturálne a investičné fondy  
OP Integrovaná infraštruktúra 2014 – 2020

 **MINISTERSTVO  
DOPRAVY A VÝSTAVBY  
SLOVENSKEJ REPUBLIKY**

 **MINISTERSTVO  
ŠKOLSTVA, VEDY,  
VÝSKUMU A ŠPORTU  
SLOVENSKEJ REPUBLIKY**



# Otvorená veda v praxi: Otvorené vedecké dáta

- Všeobecné princípy otvorenosti
- Reprodukovateľnosť výskumu
- Typy a rozdelenie otvorených dát
- Benefity a využitie otvorených dát
- Publikovanie vedeckých dát
- Citovanie vedeckých dát

Prečo by nás mali zaujímať výskumné dáta?

## Prečo žiadame otvorenosť?

Čo sa môže pokaziť, keď sú dáta „uzavreté“?

Čo všetko vlastne môžu byť výskumné dáta?

V čom sa môže zlepšiť fungovanie vedy, keď budú otvorené?

Ako publikovať otvorené dáta?

Ako citovať datasety druhých vedcov?



# Všeobecné princípy otvorenosti



Znalosti (nástroje...) považujeme za **otvorené**, ak umožňujú používateľom slobodne:

- Prístupovať k ich obsahu (čítať, stiahnuť do počítača...)
- Používať ho
- Pozmeniť (preložiť, vylepšiť...)
- Zdieľať, šíriť.

## Aké možnosti nám dáva otvorenosť?

### 1. Vidieť do vecí

Transparentnosť, dôveryhodnosť, reprodukovateľnosť, obhájiteľnosť / accountability (autor signalizuje, že má čistý štít, použil korektné metódy),

Otvorenosť nám môže pomôcť vidieť na sebe (svojich dátach) veci, na ktoré by sme sami neprišli

Príležitosť opraviť chyby, alebo nájsť nečakané benefity

### 2. Možnosť veci **ovplyvniť**, slobodné prostredie

Z otvorených údajov druhých sa môžeme učiť, používať ich, bez zbytočných bariér

### 3. Pripravenosť **odpovedať** na podnety zvonka

Možnosť rýchlejšie zareagovať na zmeny, prispôbiť svoje stratégie aktuálnemu vývoju situácie

### 4. **Zdieľanie** s druhými, budovanie vzťahov a spoluprác, tvorenie nových hodnôt

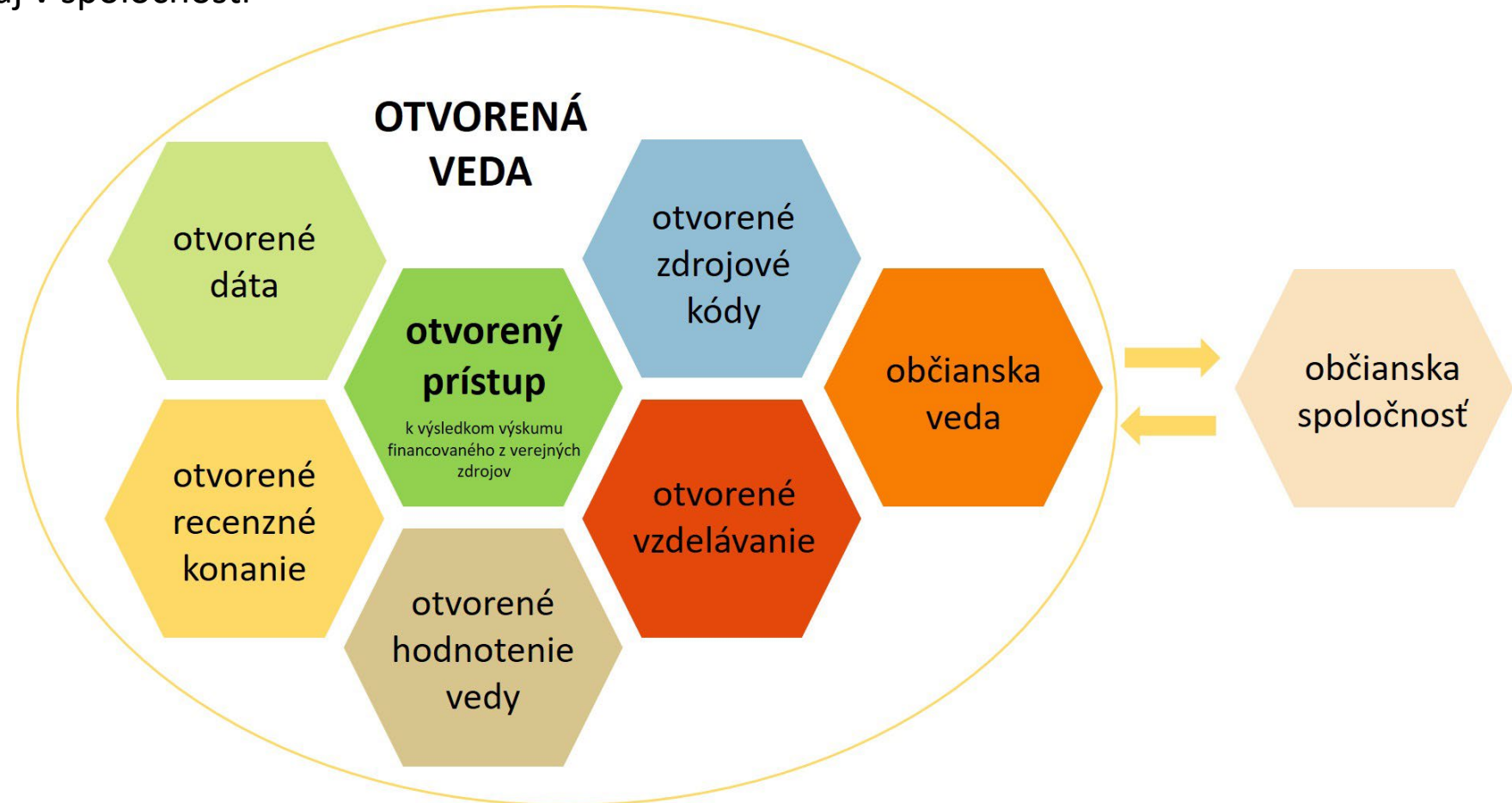
Umožniť druhým prístup k svojim výsledkom, zhodnocovať výsledky novými spôsobmi (čím viac ľudí vaše výsledky využije, tým viac sa financovateľovi vyplatí do vášho výskumu investovať – pričom nemusí ísť o peniaze, ale napr. o verejný prospech), posilňovať spoluprácu, povzbudzovať v zdieľaní ostatných

# Všeobecné princípy otvorenosti

**Otvorená veda** – prístup k výskumnému procesu založený na spolupráci, zdieľaní a otvorenom šírení poznatkov pomocou optimálneho využitia digitálnych technológií a aplikovania princípov otvorenosti v každom aspekte výskumného cyklu.

**Význam otvorenosti** – vo vede a aj v spoločnosti

- efektivita (finančná, časová)
- viditeľnosť
- širší dosah (vo vedeckej komunite aj pre verejnosť)
- atmosféra spolupráce
- Integrita, spravodlivosť, etika



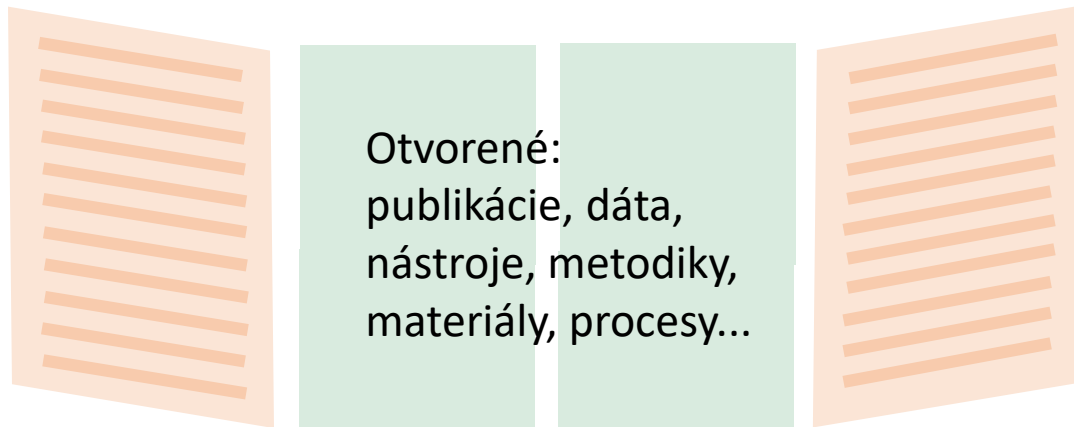
**OPEN SCIENCE:  
JUST  
SCIENCE  
DONE RIGHT**

Schéma: Z. Stožická podľa Foster Open Science

<https://www.fosteropenscience.eu/content/what-open-science-introduction>

# Všeobecné princípy otvorenosti

Otvorené môže byť všetko, čo vedcovi napadne, aj to, čo mu ešte nenapadlo:



- Zdieľanie detailných metodík
- laboratórných kníh
- dát
- voľného času na výskumných prístrojoch medzi laboratóriami/inštitúciami
- zdieľanie reagensí (príklad)



*Home / Archive / July 2022, Issue 2 / Careers*

## What If Scientists Shared Their Reagents for Free?

Some researchers have decided to provide their products without financial compensation or expectations of authorship on resulting papers, prompting a flurry of new work.

# Všeobecné princípy otvorenosti

**Prekážky**, ktoré vedcom bránia v používaní užitočných zdrojov vytvorených druhými, môžu byť:

- Informačné (nedozvieme sa, že taký zdroj existuje, nenájdeme ho)
- Technické (nemáme prístup)
- Právne (autorské práva, licencie)
- Psychologické (obavy, predsudky)



Foto: Tim Collins, unsplash,  
<https://unsplash.com/photos/B5ox94ZFGgl>

## Riešenia:

- Informačné (propagácia otvorených zdrojov)
- Technické (zvyšovanie viditeľnosti zdrojov, indexovanie, použitie vhodných formátov, identifikátorov, otvorený prístup)
- Právne (vytvorenie právneho rámca pre zdieľanie: verejné licencie, formulovanie politik/mandátov obsahujúcich povinnosť otvorene zdieľať výsledky výskumu financovaného z verejných zdrojov)
- Psychologické (šírenie povedomia, vyvracanie mýtov, povzbudzovanie/motivácia k otvorenosti, propagácia postupov dobrej praxe)

# Cesta k otvoreným dátam – prvé kroky

- Zhromažďovanie a zdieľanie dát odlišuje človeka od zvierat. Zvieratá si dokážu odovzdať informácie, ale len v obmedzenej miere. Dôkazy o zhromažďovaní a zaznamenávaní dát ľuďmi existujú už od vynálezu písma.
- Sumeri pred 5000 rokmi zapisovali na hlinené tabuľky dane a úrodu (boli to v podstate dáta verejnej správy, ale vzhľadom na povahu záznamov a vtedajšej spoločnosti boli prístupné len úzkemu okruhu používateľov. V súčasnosti sú tieto dáta základom pre výskum, ktorý hovorí o vtedajšej spoločnosti – a aj o vývoji klímy), ale aj astronomické mapy (okrúhla tabuľka).
- Nástroje potrebné pre správu dát sme zbierali postupne počas histórie: Prvé knižnice v Egypte (2000 pnl) rozvíjali techniky kurácie a uchovávanía písomných záznamov.
- To, čo dnes nazývame výskumnými dátami, bolo v rôznych záznamoch meraní, výskumných denníkoch a vedeckých spisoch, ktoré mohli byť prístupné v knižniciach, neprístupné, alebo prístupné na požiadanie len pre iných vedcov.



# Cesta k otvoreným dátam – štatistika a počítaacie stroje

- Nestačí údaje mať a uchovávať, treba s nimi vedieť účinne a presne pracovať. **Štatistické metódy** na spracovanie údajov sa začala rozvíjať v 17. storočí (obchodník [John Graunt](#), porovnával záznamy o úmrtnosti v jednotlivých častiach Londýna).
- Strojové spracovanie veľkých objemov dát: **Hollerithov tabelátor** (spracovanie dát zo sčítania ľudu v USA trvalo v roku 1880 až 8 rokov, [Herman Hollerith](#) ho v roku 1890 vďaka svoju elektromechanickému stroju na triedenie diernych štítkov skrátil na 2,5 roka) – položil základy pre rozvoj počítačov a založil firmu IBM.
- Systém spracovania diernych štítkov sa rozšíril v USA aj v Európe, v štátnom sektore aj komerčných firmách.
- 1965: objavuje sa termín **Data Science** (spočiatku v súvislosti s dátami spoločenských vied, Raban a Gordon 2020, <https://doi.org/10.1007/s11192-020-03371-2>)
- systém diernych štítkov naráža na fyzické limity – napr. už v 1966 zástupcovia National Weather Records Center (Národného centra pre záznamy počasia) v USA vyjadrili obavy, že ich budova skolabuje pod váhou diernych štítkov.
- 1969: pojem **Big Data** (volume, velocity, variety) – dáta príliš veľké, aby si s nimi poradili tradičné systémy a preto vyžadujú nové technológie spracovania;





## 20. storočie – rodí sa pojem „výskumné dáta“

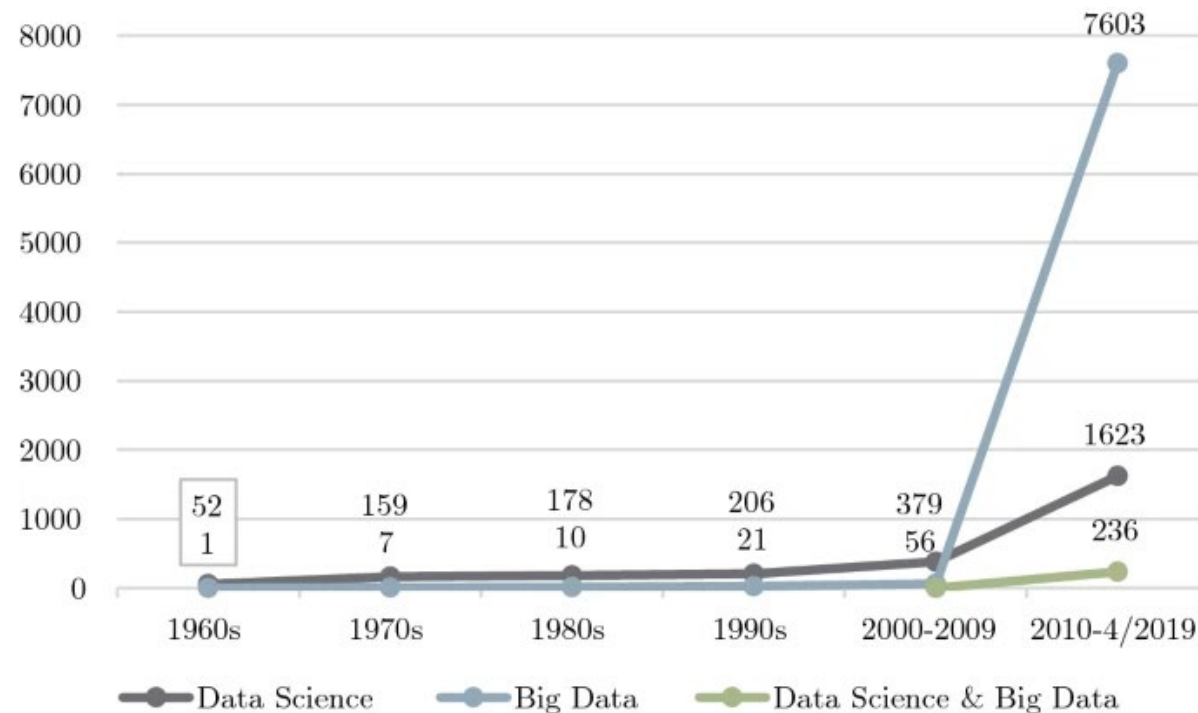
- Rozvoj technológií pre ukladanie veľkého množstva dát: Fritz Pfleumer **magnetické pásky** (patent: 1928, použitie v počítačoch: 50-80-te roky 20. storočia)
- Základy pre počítačové **cloudy** položil už v r. 1960 Joseph Carl Robnett Licklider
- Edgar F. Codd (1970) – základ pre relačné **databázy** (ukladanie dát v efektívnejšie štruktúrovanej podobe)

- 1990: Tim Bernes-Lee, CERN: **internet**

### Zdieľanie vedeckých dát na internete

S rozvojom internetu, výskumných infraštruktúr a nových nástrojov (teleskopy, mikroskopy, sekvenovanie DNA, proteomika...) naberá pojem „Big Data“ nový rozmer.

- Vznikajú prvé **repozitáre** (napr. 1991: arXiv).
- Niektoré časopisy, ktoré mali elektronickú verziu, začali umožňovať publikovanie podkladových dát v doplnkových materiáloch k článku.



## 20. storočie – rodí sa pojem „výskumné dáta“

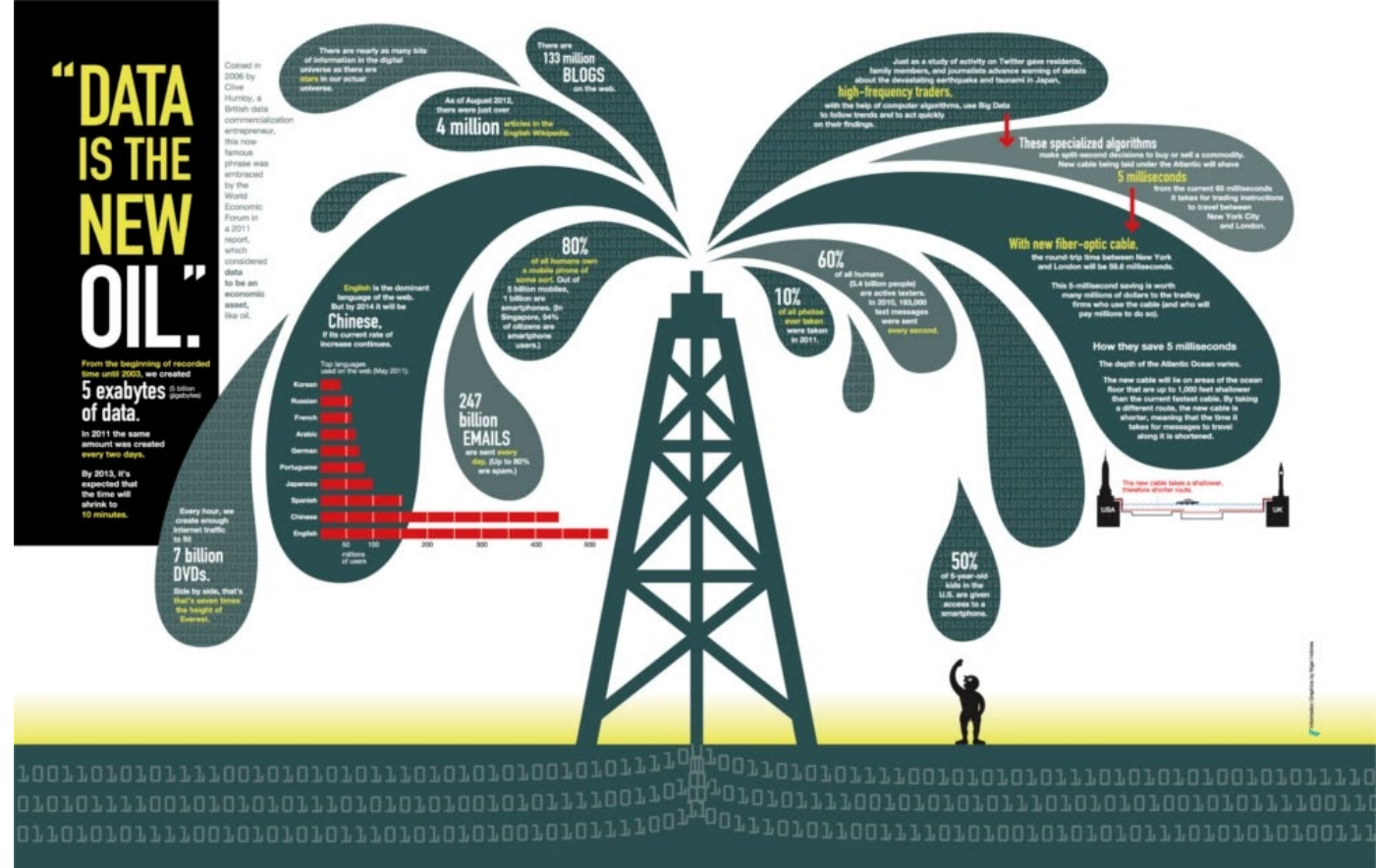
- **Otvorené dáta ako nevyhnutnosť pri širokej medzinárodnej spolupráci:** od výskumu *Coenorhabditis elegans* až k Human genome project (1990-2003) – zverejnenie sekvencií na internete do 24 h po sekvenovaní
- **Vyhľadávanie** (1997: Google, vyhľadávacie algoritmy, indexovacie mechanizmy)
- 1999, National Academies of Sciences, USA: Prvá vplyvná **politika definujúca výskumné dáta**. V politikách, napr. v pilotnej politike Horizontu 2020 sa definícia obmedzuje na digitálne, výpočtové výskumné dáta.
- Po roku 2000: **OECD** (Organisation for Economic Co-operation and Development), definícia Open Data, 2007: ***Principles and Guidelines for Access to Research Data from Public Funding***

**Výskumné dáta:** faktografické záznamy (číselné výsledky, textové záznamy, obrázky a zvuky), ktoré sa používajú ako primárne zdroje pre vedecký výskum a ktoré sú vo vedeckej komunite všeobecne akceptované ako potrebné na overenie výsledkov výskumu.

**OECD:** „Prístup k výskumným údajom zvyšuje návratnosť verejných investícií do tejto oblasti; posilňuje otvorené vedecké skúmanie; podporuje rozmanitosť štúdií a názorov; podporuje nové oblasti práce a umožňuje skúmať témy, ktoré pôvodní výskumníci nepredpokladali.“

# Dáta ako komodita

- V komerčnej sfére – **dáta ako komodita**: „**Dáta sú ropou súčasnosti.**“ (matematik [Clive Humby](#), 2006 – v roku 1995 stál pri vzniku prvého vernostného programu Clubcard pre spoločnosť Tesco)
- „**Dáta sú ropou súčasnosti. Sú cenné, ale skutočne použiť sa dajú len v spracovanej podobe.**“
- „**Údaj nemožno považovať za fakt a fakt za vhl'ad.**“ ([Michael Palmer](#), 2006)



Obrázok: ONS digital, <https://digitalblog.ons.gov.uk/2016/01/25/open-data-new-oil-fuels-society/>

**V skutočnosti boli dáta vždy cenné – len s digitálnymi nástrojmi narástla schopnosť ľudí dáta získavať, skladovať, spracovávať a analyzovať.**

- Odvtedy: Rýchly **nárast kvantity** dát (zdvojnásobenie objemu každé dva roky), potreba organizácie, manažmentu, čistenia (špinavé dáta – dáta v pôvodnom stave s chybami alebo chýbajúcimi údajmi), ochrany, korektnej analýzy a interpretácie

# Dáta a „trenie“

- [Paul N. Edwards](#) (2011) konštatuje, že dáta netečú tak hladko ako ropa: **Data friction** (trenie – čo sa deje medzi dvomi povrchmi, napr. keď majú spolupracovať dvaja vedci z rôznych disciplín – konflikty, ťažkosti pri rozhodovaní, skreslenia, vznik nepresností pri každej manipulácii / prechode od jedného aktéra k druhému cez „rozhranie“, problémy súvisiace s pohybom dát, ktoré spomaľujú ich tvorbu, prenos a využívanie) – aj vo vzťahu ku **konfliktu medzi pohľadom na dáta ako komoditu a dáta ako prostriedok pre verejný prospech** („knowledge commons“). Edwards to nevidí fatálne – trenie podľa neho môže viesť ku vzniku nového, efektívnejšieho a spravodlivejšieho usporiadania.

Príklad: rozdielne perspektívy **meteorológov** zaoberajúcich sa **predpoveďami počasia** (okamžitá komunikácia výsledkov, malý význam archivácie starých výsledkov) a **klimatológou** (dôraz na prístupnosť archívnych údajov o počasí). Počas histórie meteorológie dáta o počasí čelili treniu napr. aj na hraniciach jednotlivých štátov. Klimatológovia nepovažujú dáta meteorológov za stabilné.

- „Na rozdiel od ropy, dáta sa dajú **využiť znova**.“ (dátový vedec [Francis Adrian Viernes](#), 2021)
- **Environmentálna stopa** dát (ak by boli dátové centrá krajinou, boli by piatym najväčším konzumentom energie na svete) – potreba efektivity (Hoosain et al. 2022, DOI <https://doi.org/10.1007/s43615-022-00191-9>)

# Replikovateľnosť, reprodukovateľnosť, integrita...

... aby sme sa na výsledky výskumu mohli spoľahnúť

- Vychádza zo snahy potvrdiť, že to, čo sa zistilo v pôvodnom výskume, naozaj platí a výsledok nebol len dielom náhody alebo dôsledkom chýb.

- **Prečo sa výsledky vedeckého výskumu často nedarí zopakovať?**

Dôvody môžu byť legitímne (napr. jedinečný materiál, špecifiká miestnej kultúry...), ale aj nelegitímne (nedodržanie štandardov, slabá dokumentácia / nedostatočný opis metód, zamlčanie dôležitých informácií, sloppy science...)

- **Kríza reprodukovateľnosti** (psychológia, medicína, onkológia, [ekológia](#), [strojové učenie](#)...)
- John Ioannidis (2005): Why most published research findings are false? (o medicíne). Brian Nosek (2008), Reproducibility Project: Psychology (iba 36 % štúdií sa podarilo zreplikovať).
- Predpoklad: vedci vždy dajú k dispozícii podkladové dáta k článku (realita: len asi 38 % autorov)
- Systém hodnotenia zameraný na novosť v súčasnosti nedáva vedcom veľa motivácií overovať tvrdenia (replikovať štúdie) iných
- Dôsledok, napr.: zlyhávajúce klinické testy pre liečebné stratégie „úspešné“ na animálnych modeloch (strata financií, životov...)

Video o kríze reprodukovateľnosti: <https://www.youtube.com/watch?v=FpCrY7x5nEE>

Softvérová reprodukovateľnosť: <https://www.activestate.com/blog/trust-security-and-the-reproducibility-crisis-in-software/>

Goodman et al., 2016, [DOI: 10.1126/scitranslmed.aaf5027](https://doi.org/10.1126/scitranslmed.aaf5027)

# Replikovateľnosť, reprodukovateľnosť, integrita...

... aby sme sa na výsledky výskumu mohli spoľahnúť

- **Reprodukovateľnosť** – získať tie isté výsledky z rovnakých dát (podmnožina: výpočtová reprodukovateľnosť – získať tie isté výsledky z rovnakých dát pomocou rovnakého kódu/softvéru),
- **Replikovateľnosť:**
  - **priama** – získať také isté výsledky metódami, ktoré sú blízke pôvodnému výskumu, ako je to len možné,
  - **systematická** – experimentálna procedúra sa zopakuje, s niektorými zámernými zmenami,
  - **konceptuálna** – získať také isté výsledky inými metódami (zistíme, do akej miery môžeme zovšeobecniť výsledky pôvodného výskumu).
- kontext **integrity výskumu**
- Čo pomáha? Predregistrácia (Aspredicted.org, Open Science Framework), validácia metód, otvorené dáta.
- Častý zdroj problémov – nedostatky vo zverejňovaní metodík, nekorektné štatistické vyhodnocovanie, p-hacking, HARKing
- Nová vedná disciplína – **metaveda** (metascience)

Ako zlepšiť reprodukovateľnosť? Reakcia učenej spoločnosti American Society for Cell Biology na krízu reprodukovateľnosti:  
<http://www.ascb.org/wp-content/uploads/2015/11/How-can-scientist-enhance-rigor.pdf>

„Keby som písal jazykmi ľudskými  
aj anjelskými, ale integritu by som  
nemal, bol by som ako cvendžiaci  
kov a zuniaci cimbal.“



Miriam Bauer, Humboldt Kosmos: <https://www.humboldt-foundation.de/web/kosmos-cover-story-104-3.html>

**Integrita výskumu** (aby „držal pokope“ – ktorýkoľvek z  
elementov integrity vypadne, práca stráca hodnotu a uplatniteľnosť)

- zmysluplné hypotézy a design výskumu (plánovanie, pre-registrácia – napr. na stránkach <https://aspredicted.org/> alebo <https://osf.io/prereg/>),
- validné, dostatočne citlivé a spoľahlivé metódy,
- poctivá realizácia, reprodukovateľnosť,
- korektná analýza výsledkov,
- transparentná komunikácia o výskume,
- vyhýbanie sa predpojatosti (bias),
- vedecká a publikačná etika.

Platforma **FORRT** (Framework for Open and Reproducible Research Training))

<https://forrt.org/>

Učebnica **The practice of reproducible research** (Kitzes et al. 2018): <https://www.practicereproducibleresearch.org/>

**ENAI** – stráky európskej siete pre akademickú integritu: <https://www.academicintegrity.eu/wp/>

**ERA portál** <https://eraportal.sk/eraportal/etika-vo-vyskume/>

**Deklarácia o posilnení kultúry vedeckej integrity na Slovensku:**

[https://eraportal.sk/wp-content/uploads/2021/08/DEKLARACIA\\_3\\_8\\_2021\\_everzia2.pdf](https://eraportal.sk/wp-content/uploads/2021/08/DEKLARACIA_3_8_2021_everzia2.pdf)

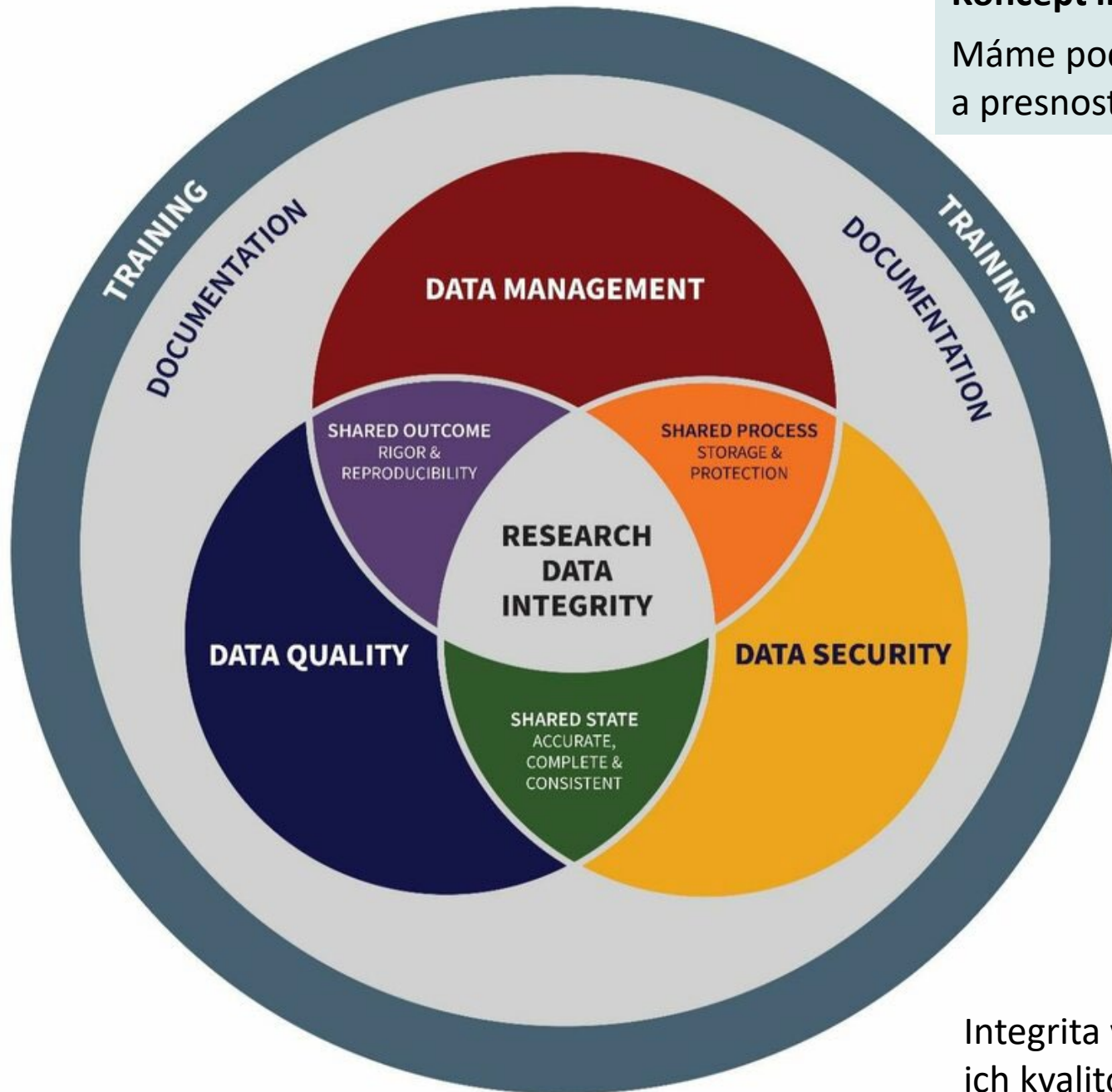
Národná komisia pre vedeckú integritu, ombudsman, odnímanie titulov získaných nepoctivo, postihy za „contract cheating“ ...

- **Integrita výskumu** sa týka: plánovania, navrhovania, realizácie, komunikácie (šírenia výsledkov) a recenzovania/hodnotenia – teda všetkých fáz životného cyklu výskumu.
- Mnoho stiahnutí článkov sa týka problémov s integritou dát ako sú chyby, nespoľahlivé dáta, falzifikácia / fabrikácia dát, plagiátorstvo dát, neposkytnutie pôvodných dát (Retraction Watch, kategória [Data issues](#)).
- Na dodržanie integrity je nevyhnutné dodržiavať princípy objektivity, čestnosti, otvorenosti, obhájiteľnosti / sledovateľnosti (accountability), férovosti a zodpovednej správy (stewardship) (National Academies of Sciences (2017) [Fostering integrity in research](#)).
- Podľa [Singapúrskej deklarácie o integrite výskumu](#) (2010) má vedec:
  - viesť / udržiavať jasné, presné záznamy všetkého výskumu,
  - zdieľať dáta a výsledky otvorene a rýchlo,
  - vytvárať a udržiavať prostredie priaznivé pre integritu výskumu,
  - hlásiť nezodpovedné výskumné praktiky (falzifikácia, fabrikácia dát, nezodpovedné nakladanie s dátami a pod).



**Koncept integrity výskumných dát: sú validné (platné) pre vedu?**

Máme podmienky/procesy na **zabezpečenie a udržanie** ich validity a presnosti?



**Princípy 4R pre integritu výskumných dát:**

**Research rigor** – rigoróznosť/kvalita výskumu (validné a exaktne použité metódy pri zbere a spracovaní dát, interpretácii a komunikácii výsledkov)

**Reproducibility** – reprodukovateľnosť (získanie konzistentných výsledkov z tých istých vstupných dát s použitím tých istých metód a kódu)

**Replication** – replikovanie výskumu, teda získanie konzistentných výsledkov (s prihliadnutím na stupeň neistoty vlastný skúmanému systému) z rôznych štúdií odpovedajúcich na tú istú výskumnú otázku, pričom každá štúdia získala vlastné dáta

**Reuse of data** – vytvoriť podmienky pre opätovné využitie dát.

Integrita výskumných dát je úzko prepojená s manažmentom dát, ich kvalitou a bezpečnosťou. Súčasťou tréningu výskumníka by mal byť aj tréning dokumentácie.

## Scientists/Researchers

Protecting against research bias  
Mentorship/supervision  
Standardised experimental methods

Support  
collaboration/  
team science

Support/  
fund validation  
works

Training  
Documentation

Quality assurance  
Tools/resources  
Accessibility/transparency

Incentives/awards  
Open science

Independent oversight  
Long-term funding

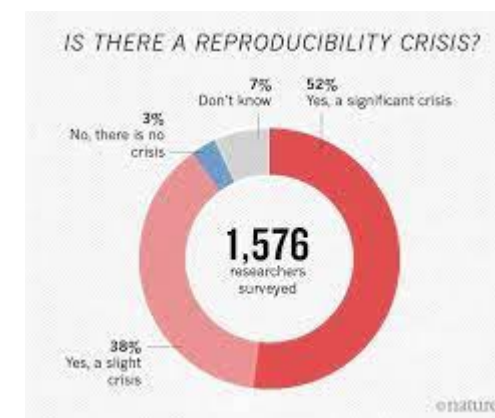
Research Institutions

Funding Institutions

## Úlohy vedcov ako jednotlivcov, výskumných inštitúcií a financovateľov na zabezpečenie integrity výskumu a kvality dát

Nedostatok reprodukovateľnosti neznamená hneď neplatné výsledky alebo pokus o podvod, ale reprodukovateľnosť zvyšuje dôveryhodnosť a spoľahlivosť výsledkov výskumu.

Podľa výsledkov prieskumu 70 % vedcov nebolo schopných zreplikovať prácu iných, 50 % nedokázalo niekedy zreplikovať ani vlastné výsledky (Baker, 2016, <https://doi.org/10.1038/533452a>)





## Transparency and Openness Promotion (TOP) guidelines

- Šírenie **zásad transparentnosti a otvorenosti v politikách a redakčných procesoch časopisov**
- Na rozdiel od princípov COPE/DOAJ/OASPA/WAME (zameraným skôr na to, ako časopis komunikuje o sebe) sa zásady TOP venujú konkrétnym praktikám s presahom na reprodukovateľnosť výskumu:
  - **Citovanie** (nielen článkov, z ktorých autor vychádzal, ale aj dát, kódu, materiálov...) – **dôsledné citovanie rôznych foriem prínosu druhých vedcov zvyšuje ich motiváciu k otvorenosti**,
  - Transparentnosť ohľadom **dát**,
  - Transparentnosť ohľadom **analytických metód** (kódu),
  - Transparentnosť ohľadom **materiálov**,
  - Transparentnosť ohľadom **dizajnu a analýz**,
  - **Preregistrovanie štúdií**,
  - Preregistrovanie **analýz**,
  - **Replikovanie** výskumu.
- V samotnom materiáli je niekoľko verzií navrhovaného znenia na rôznych úrovniach (Level 1, 2, 3), podľa toho, do akej hĺbky sa časopis vo svojich politikách chce/potrebuje venovať danej téme.

- **Úroveň 1:** povedať, či budú dáta/materiály otvorené (alebo aspoň za akých podmienok budú prístupné iným vedcom),
- **Úroveň 2:** časopis publikuje len články, kde sú dáta/kód/materiály zdokumentované a budú prístupné ďalším vedcom pre potreby replikovania experimentu,
- **Úroveň 3:** časopis publikuje články, kde sú dáta/kód/materiály maximálne prístupné časopisu (recenzentom) pred publikovaním na overenie reprodukovateľnosti.

„Článok v časopise je ústredným prvkom procesu vedeckej komunikácie. Pokyny pre autorov definujú, ktoré aspekty výskumného procesu by mali byť prístupné komunitě na hodnotenie, kritiku, opätovné použitie a rozšírenie. Vedci uznávajú hodnotu transparentnosti, otvorenosti a reprodukovateľnosti. **Zlepšenie zásad časopisov môže pomôcť, aby sa tieto hodnoty stali zjavnejšími v každodennej praxi a v konečnom dôsledku zlepšili dôveru verejnosti vo vedu samotnú.**“



**Full text**

Figures and data

Side by side

p=0.05); above this dose participants tend to guess better than random. For this calculation, psilocybin containing mushroom doses were converted to LSD equivalents as described in the *Additional information on statistical models* section.


## Data availability


All data is shared in CSV format at [https://github.com/balazs1987/mcrds\\_public/tree/master/data](https://github.com/balazs1987/mcrds_public/tree/master/data) (copy archived at <https://archive.softwareheritage.org/swh:1:rev:43bb44cc8b6b79536e4b2afd0a7c724e90f137f7/>) together with appropriate documentation.


## References


### Version of Record

Accepted for publication after peer review and revision.

 Download

 Cite

 Share

 Comment (3)

46,493 views, 68 citations

### Version of Record published

March 2, 2021 (This version)

Accepted

January 20, 2021

Matej Harvát

Typy a rozdelenie výskumných dát

# Benefity a využitie otvorených dát



## Prečo je publikovanie dát dôležité?

- Transparentnosť (kontrolovateľnosť, viac očí skôr nájde chyby / odhalí prípadný podvod),
- Príležitosť pre tvorivú interpretáciu (viac mozgov – viac nápadov, možnosť odhaliť nečakané súvislosti),
- Integrita (reprodukovateľnosť, dôveryhodnosť, šanca lepšie si porovnávať validitu metód),
- Etický aspekt dostupnosti pre všetkých (inkluzívnosť, rovnejšie príležitosti napr. pre vedcov z chudobnejších krajín / inštitúcií),
- Príležitosť získať kredit pre vedcov, ktorí dáta zbierali,
- Spolupráca (Otvorené dáta – nevyhnutnosť pri veľkej medzinárodnej spolupráci: začínajúc projektom *Coenorhabditis* a sekvenovaním ľudského genómu, cez dáta zo satelitov, končiac občianskou vedou),
- Rozvoj rôznorodého využitia / aplikácií (porovnanie užitočnosti otvorených a uzatvorených/proprietárnych dát),
- Stimulácia podnikania (otvorené dáta ako východiskový materiál pre podnikateľov – napr. aplikácie s využitím voľne dostupných dát, otvorené dáta vo verejnej správe, využitie dát novým spôsobom, ktorý by ich tvorcovia nečakali),
- Motivácia k riešeniu problémov etiky a bezpečnosti, namiesto ich obchádzania.

# Benefity a využitie otvorených dát – príklady

- Otvorené dáta z [OpenStreetMap](#) umožňujú v kombinácii s družicovými zábermi a miestnymi dobrovoľníkmi v rámci projektu občianskej vedy [Missing maps](#) mapovanie terénu pre výjazdy dobročinných / humanitárnych organizácií,
- Geografické a družicové dáta – mnohostranné využitie (mapovanie stavu lesov, život v mestách, prepojenie s pocitovými mapami...),
- **Mejora tu escuela** – program v Mexiku (dostupné dáta o školách, motivujú k zlepšeniu ich výkonu a umožňujú občanom lepšie informované rozhodnutia, <https://datos.gob.mx/blog/mejoratuescuelaorg>),
- [COVID19](#) data repository – dostupnosť dát – vývoj diagnostiky, protilátok, testovanie perspektívnych liečiv.



Odpor voči otváraníu dát, potreba komunikácie, šírenia povedomia, riešenia prípadných konfliktov a vyvracania mýtov o otvorených dátach (mentalita zhromažďovania dát „data hoarding“ vs zdieľania dát „sharing“).



P. N. Edwards, foto: [Stanford University](#)

„The opening of scientific data is both a data friction in itself and a way to collectively manage data frictions by weakening complex issues of data ownership.“ (Paul Edwards)

Otvorenie dát vytvára „trenie“ (zádrhely na ceste dát, konflikty, možnosť chýb a dezinterpretácií), ale zároveň pomáha riešiť väčšie problémy.



# Bingo „Výhoviek proti otvoreným dátam“

Budeme dostávať spam.

Ja sa o to nestarám, hoci niekto iný by mohol.

Nemáme API.

Sú príliš veľké.

Bude nám prichádzať priveľa otázok (požiadaviek).

Dáta nebudú chránené.

Dáta nemajú dostatočnú kvalitu.

Čo ak by ich využili teroristi?

Je to príliš zložité.

Možno ich budeme chcieť použiť v článku.

Ukradnú nám ich zloději.

Už existuje taký projekt...

Právnicki budú vyžadovať zvláštne licencie.

Čo ak tie dáta neskôr budeme chcieť predať?

Nie sú veľmi zaujímavé.

Ľudia ich možno nesprávne pochopia.

# Bingo „Výhoviek proti otvoreným dátam“

Budeme dostávať spam.  
Spam dostávame aj tak.

Ja sa o to nestarám, hoci  
niekto iný by mohol.  
Ak by si to hovoril každý,  
nikto by nikdy nezdieľal.

Nemáme API.  
Dáta sa dajú zdieľať aj bez  
API.

Sú príliš veľké.  
Skúste sa pozrieť na  
dátové repozitáre vo  
Vašom odbore.

Bude nám prichádzať  
priveľa otázok  
(požiadaviek).  
Pravdepodobne nie.

Dáta nebudú chránené.  
Zvážte, ktorá časť dát  
potrebuje aký stupeň  
ochrany.

Dáta nemajú dostatočnú  
kvalitu.  
Na tom sa dá pracovať.

Čo ak by ich využili  
teroristi?  
Nezverejňujte citlivé dáta,  
anonymizujte.

Je to príliš zložité.  
Nie je to až také zložité.  
Existujú zrozumiteľné  
návody.

Možno ich budeme chcieť  
použiť v článku.  
V článku len odkážete na  
zverejnené dáta.

Ukradnú nám ich zloději.  
Predstavte si, čo by s nimi  
mohli robiť, rozoberte si  
jednotlivé možnosti.

Už existuje taký projekt...  
Aspoň si môžete porovnať  
výsledky, napomáha to  
dôveryhodnosti.

Právnicki budú vyžadovať  
zvláštne licencie.  
Verejné licencie sú  
štandard.

Čo ak tie dáta neskôr  
budeme chcieť predať?  
Zvážte, či je to naozaj  
pravdepodobné.

Nie sú veľmi zaujímavé.  
Človek nikdy nevie, na aké veci  
môžu budúcim používateľom  
poslúžiť dnešné dáta.

Ľudia ich možno  
nesprávne pochopia.  
Priložte k nim vysvetlenie.

Príklad: [M. F. Maury](#) (1807-1073), lodné denníky, vetry a morské prúdy a klimatická zmena



# Mýty o otvorených dátach

- **Otvorené dáta sú zlé pre podnikanie**

Otvorené dáta môžu byť cenným zdrojom pre súkromný sektor rovnako ako pre verejný (na otvorených dátach napr. o doprave vznikajú úspešné aplikácie ako Citymapper)

- **Chcete, aby boli otvorené všetky dáta?!**

Nie. Existujú oprávnené dôvody, ktoré bránia otvorenosti dát – čo však nič nemení na skutočnosti, že sa netýkajú väčšiny výsledkov výskumu financovaných z verejných zdrojov. Princíp „*As open as possible, as closed as necessary.*“

- **Pri otvorených dátach sa nedá zachovať súkromie**

Dá. Aj v rámci otvorenej vedy existujú jasne definované podmienky, za ktorých je lepšie nechať dáta uzavreté, nezverejniť všetky, sprístupniť ich iba vopred definovaným osobám / skupinám, alebo ich pred zverejnením upraviť / anonymizovať (napr. dáta o prepojení jednotlivých variantov génov a ochorení sú veľmi žiadané pre vedu, aby sa zistili príčiny chorôb a našla sa liečba. Poistovne by tiež veľmi radi poznali genetické riziká poistencov, aby tomu prispôbili svoje obchodné stratégie, ale také informácie sú chránené)

- **Ľudia budú otvorené dáta zneužívať**

Dáta sa zneužívajú, aj keď sú uzavreté, stále je potrebné s tým bojovať. No riziko je vyvážené ostatnými benefitmi, napr. keď sú dáta otvorené, viac očí má príležitosť nájsť a opraviť chyby.

- **Stačí ak dáta „dáte von“, používatelia prídu**

Komunikácia o dátach a spätná väzba je dôležitá, aby dáta slúžili svojmu účelu.



Každé „otvorenie“ / sloboda prináša aj riziko. Musíme sa však pohnúť od pasívneho postoja ‘Nerobme radšej nič’ k riešeniu otázky ‘Ako to môžeme zvládnuť bezpečne a udržateľne?’

Otvorené dáta podporujú praktiky lepšieho manažmentu dát.

Možno jeden Váš dataset nezachráni svet, ale jeho zdieľanie pomáha budovať kultúru otvorenosti, ktorá svet mení k lepšiemu.

Matej Harvát

Publikovanie a citovanie výskumných dát



Ďakujem za pozornosť