UNIVERZITA KARLOVA

Studie užití výzkumných dat v rámci vybraných vědních oborů

Univerzita Karlova

Centrum pro podporu open science

Mgr. Dagmar Hanzlíková

Ing. Milan Janíček

Tento výstup vznikl v rámci projektu Rozvoj kapacit VaV UK,

reg.č. CZ.02.2.69/0.0/0.0/18_054/0015222

Praha, 2021

Toto dílo podléhá licenci Creative Commons Uveďte původ 4.0 Mezinárodní Licence.

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

HR EXCELLENCE IN RESEARCH

Obsah

Studie užití výzkumných dat v rámci vybraných vědních oborů 1
Obsah 2
1. Úvod 4
2. Principy open science při správě výzkumných dat 5
2.1 Open science 5
2.2 Replikovatelnost výzkumu a opětovné využití dat 6
2.3 FAIR data 6
2.3.1 Jak zajistit, aby data byla FAIR 8
3. Práce s výzkumnými daty na Univerzitě Karlově 11
3.1 Dotazníkové šetření 11
3.2 Respondenti 12
3.3 Vyhodnocení odpovědí 16
3.3.1 Oblast 1: Práce s výzkumnými daty během výzkumu 16
3.3.2 Oblast 2: Dlouhodobé uchovávání výzkumných dat 24
3.3.3 Oblast 3: Sdílení výzkumných dat 28
3.3.4 Oblast 4: Podpora správy výzkumných dat 35
3.4 Závěrečné shrnutí 37
3.4.1 Respondenti 37
3.4.2 Práce s výzkumnými daty během výzkumu 37
3.4.3 Dlouhodobé uchovávání výzkumných dat 38
3.4.4 Sdílení výzkumných dat 38
3.4.5 Podpora správy výzkumných dat 39
4. Praxe zahraničních výzkumných univerzit 40
4.1 Utrecht University 40
4.1.1 Práce s daty 40
4.1.2 Podpora správy dat 41
4.2 Leiden University 41
4.2.1 Práce s daty 41
4.2.2 Podpora správy dat 42
4.3 KU Leuven 42
4.3.1 Práce s daty 42
4.3.2 Podpora správy dat 43
4.4 Ghent University 43

4.4.1 Práce s daty 43
4.4.2 Podpora správy dat 44
4.5 Shrnutí 44
5. Závěr 45
Poděkování 46
Podkladová data 46
Použité zdroje 47

1. Úvod

V posledních letech vidíme významný tlak ze strany poskytovatelů veřejných financí na změnu ve způsobu zveřejňování výsledků výzkumu. Téma otevřeného přístupu se už více než deset let diskutuje u publikací. Zde došlo k nešťastné situaci, kdy si výzkumné instituce musely draze nakupovat přístup k vlastním výsledkům ve formě předplatného časopisů, do nichž samy přispívaly. S rostoucí cenou předplatných pak vznikla finančně neúnosná situace, jejímž důsledkem je rozšířená podpora pro publikování v režimu open access.

S rozvojem technologických možností ovšem přichází příležitost k rozsáhlejšímu sdílení samotných výzkumných dat. Není nutné se spolehnout jen na jejich interpretaci autorem prostřednictvím článku (byť i nadále zůstávají články v odborných časopisech základním stavebním pilířem vědecké komunikace). Zároveň lze sdílená data využít ke zlepšení možnosti ověření předkládaných tvrzení a v některých případech i přímo využít v dalším výzkumu. Problematika otevřenosti a transparentnosti ve vědě, zahrnutá do termínu „open science“, je samozřejmě mnohem širší. Správa výzkumných dat je ovšem jednou z nejdůležitějších oblastí.

Z těchto důvodů je cílem předkládané studie analyzovat situaci na Univerzitě Karlově a porovnat ji se současnými trendy – především v oblastech správy dat a jejího plánování, vytváření výzkumných dat v souladu s principy FAIR s následnou možností sdílení.

Tento dokument má tři základní části:

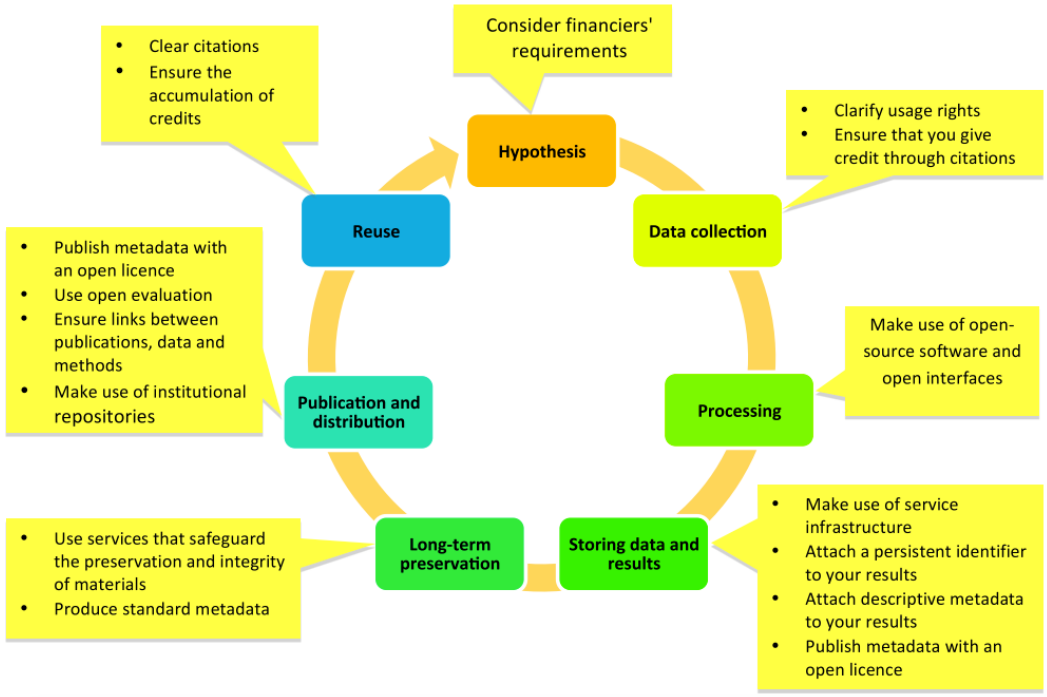
- Obecný popis principů správy vědeckých dat směřujících k vytváření dat splňujících principy FAIR
- Analýza současného stavu na Univerzitě Karlově
- Situace na vybraných zahraničních univerzitách

V závěru shrnujeme aktuální situaci a identifikujeme oblasti, kde vidíme příležitost ke zlepšení a dalšímu rozvoji.

2. Principy open science při správě výzkumných dat

2.1 Open science

Příručka *Open Science Training Handbook* (Bezjak et al. 2018) popisuje open science jako praktikování vědy takovým způsobem, který umožňuje spolupráci a přispívání ostatních k vědeckému poznání. V rámci open science jsou výzkumná data, laboratorní deníky a další výzkumné procesy volně dostupné za podmínek, které umožňují opětovné využití, redistribuci a replikovatelnost výzkumu a jeho podkladových dat a metod. Doporučení UNESCO k otevřené vědě (2021) rovněž popisuje open science jako paradigma, které integruje do výzkumného procesu praktiky vedoucí k reprodukovatelnosti výzkumu, transparentnosti, sdílení a spolupráci díky otevřenosti vědeckého obsahu, nástrojů a procesů. Open science tedy kromě zpřístupňování výsledků výzkumu (publikací, dat aj.) zahrnuje i celou řadu dalších aktivit a principů v různých fázích výzkumného procesu, jako např. používání open source softwaru, otevřené recenzní řízení nebo užívání veřejných licencí pro výsledky (viz Obr. 1). Jejich hlavním cílem je **transparentnost, replikovatelnost a dostupnost vědeckého poznání**.



Obr. 1.: *Podpora otevřenosti v různých fázích výzkumu (Open Science and Research Initiative 2014).*

Význam otevřené vědy si uvědomují také poskytovatelé financí, kteří ve svých podmínkách stále častěji zahrnují principy open science. Evropská Komise dokonce zařadila otevřenou vědu do hodnocení grantových žádostí v programu Horizon Europe: V sekci Excellence musí žadatelé popsat, jak jsou implementovány principy open science jakožto nedílná součást navrhované metodologie (Evropská Komise 2021).

2.2 Replikovatelnost výzkumu a opětovné využití dat

Jak již bylo naznačeno, jedním z důležitých aspektů open science je **replikovatelnost výzkumu**. Tlak na replikovatelnost výzkumu byl způsobený především diskuzemi z posledních let o tzv. replikační krizi. Jedná se o situaci, kdy se ukázalo, že mnoho studií není možné zopakovat a snahy dospět ke stejným výsledkům končí neúspěšně (viz např. Open Science Collaboration 2015; The Academy of Medical Sciences 2015; Errington et al. 2021). S tím dále souvisí otázka důvěryhodnosti publikovaného výzkumu.

Důvodů, proč se výzkum nepodaří replikovat, může být celá řada. Může se jednat například o technickou závadu v průběhu experimentu, či změnu stavu oproti podmínkám původního výzkumu (typické např. v sociologii). Dále může jít také o pochybné vědecké praktiky, jako je úmyslná manipulace s daty, vytváření hypotéz až poté, co jsou známé výsledky (tzv. HARKing) či zatajování proměnných, které mohly mít na výsledek vliv. Ať už je však příčina jakákoli, je zřejmá nutnost zvýšit replikovatelnost výzkumu a existence reálné možnosti postup výzkumu zopakovat a ověřit jeho závěry.

Jedním ze způsobů, jak replikovatelnost výzkumu zvýšit, je zpřístupnění podkladových dat k publikovaným studiím. Podkladovými daty rozumíme nejen tabulky s hodnotami nebo textové soubory, ale také např. laboratorní deníky, skripty, které byly použity k analýze dat, nebo další materiály potřebné pro ověření výsledku prováděného výzkumu (pro další tipy jak zvýšit reprodukovatelnost výzkumu viz Bezjak et al. 2018: Kapitola 4). Za účelem zvýšení důvěryhodnosti publikovaného výzkumu začala řada akademických časopisů a nakladatelství zavádět datové politiky, které do různé míry sdílení podkladových dat k publikovaným článkům vyžadují (např. Springer Nature nebo PLOS ONE[1]). Podle reportu asociace STM (STM 2021) vzrostl v průběhu roku 2020 průměrný počet časopisů s datovými politikami u sledovaných vydavatelů z 29% na 54%. Zpřístupněním podkladových dat umožníme nejen ověřit závěry publikovaného článku, čímž můžeme zlepšit kvalitu vědeckých výstupů a posílit důvěru ve vědu, ale také umožníme tato data opětovně využít.

Kromě replikovatelnosti výzkumu klade otevřená věda důraz také na **opětovnou využitelnost výstupů výzkumu**, ať už se jedná o publikace, výzkumná data nebo jiné typy výstupů. Financování vědeckého výzkumu je velmi nákladné, proto zpřístupněním výsledků výzkumu umožníme efektivně využít zdroje, jelikož neplýtváme prostředky na vytváření stejných dat vícekrát. Zároveň je tím možné přispět k ochraně tzv. over-researched communities (tzn. skupiny lidí, které má zájem zkoumat velké množství vědců) a zvláště vulnerabilních účastníků výzkumu (např. malé děti, osoby vysokého věku, sociálně vyloučené skupiny obyvatel), jejichž data díky zpřístupňování není třeba sbírat opakovaně. Důležité je také zmínit, že přístupná výzkumná data mohou být využita nejen pro další výzkum, ale také pro další účely jako je datová žurnalistika (např. Kočí et al. 2021) nebo užití dat ve výuce (např. Mráz 2021) a jiné.

2.3 FAIR data

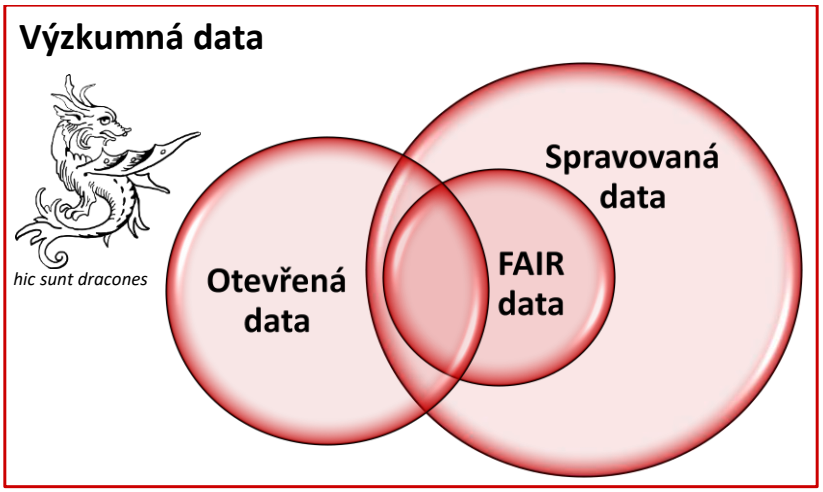
Aby se zvýšila možnost opětovného využití výzkumných dat a tím se přispělo k replikovatelnosti výzkumu, je třeba, aby data měla určité vlastnosti. Tyto vlastnosti jsou shrnuty v principech FAIR

[1] Informace o datových politikách zmíněných nakladatelství naleznete na jejich webových stránkách:
Springer Nature: https://www.springernature.com/gp/authors/research-data-policy,
PLOS ONE: https://journals.plos.org/plosone/s/data-availability

(Wilkinson et al. 2016), které popisují, jak by data měla být zpracována, aby byla lépe nalezitelná (**F**indable), přístupná (**A**ccessible), interoperabilní (**I**nteroperable) a opětovně využitelná (**R**eusable).

Nalezitelnost znamená, že data mohou jednoduše najít jak lidé, tak stroje. Proto je důležité, aby data byla opatřena perzistentním identifikátorem a měla bohatá, ideálně strojově čitelná, metadata. Aby byla data **přístupná**, měl by k nim být zajištěn volný přístup nejlépe prostřednictvím datového repozitáře. Pokud nemohou být data volně přístupná (např. v případě kdy obsahují osobní údaje nebo je výzkumník plánuje využít komerčně), měla by být dostupná alespoň jejich metadata, tedy záznam o tom, že taková data existují. **Interoperabilita**, nebo také vzájemná kompatibilita, zajišťuje, že je možné mezi sebou kombinovat různé datové sady. K tomuto účelu slouží tzv. řízené slovníky (tj. standardizované výrazy pro popis dat v daném oboru, např. MeSH v medicínském výzkumu) a také je vhodné využívat formát dat, který je v daném oboru obvyklý. Pro dosažení **opětovné využitelnosti** je důležité, aby byla data dostatečně popsaná, aby je uživatelé dokázali interpretovat. Data by také měla mít přidělenou licenci, která specifikuje, jakým způsobem a za jakých podmínek mohou být data znovu využita.

Ačkoli principy FAIR zdůrazňují opětovnou využitelnost dat, neznamená to, že všechna data musí být automaticky volně přístupná každému za jakýmkoli účelem. Ne všechna data je totiž možné otevřeně sdílet. Přístup může být omezen obzvláště v případě, že: je otevřené sdílení v rozporu s existujícími pravidly o ochraně osobních údajů, nebo je v rozporu s autorským právem či v případě, že by autor chtěl data využít pro komercializaci. Obecně se zpřístupňování dat řídí pravidlem „**as open as possible, as closed as necessary**“, tedy data by měla být tak otevřená, jak je to možné, ale zároveň uzavřená dle potřeby. Vztah mezi spravovanými daty, FAIR daty a otevřenými daty je znázorněn na Obr. 2.



Obr. 2: *Diagram naznačující vztah mezi spravovanými daty, FAIR daty a otevřenými daty.*

Spravování dat v souladu s principy FAIR a zpřístupňování dle zásady „as open as possible, as closed as necessary“ je obecně považováno za dobrou praxi a ukázku robustnosti a integrity výzkumu. Požadavky na jejich dodržování jsou postupně implementovány jak do podmínek poskytovatelů financí, tak i do legislativy v rámci Evropské Unie. Například příjemci grantové podpory z evropského programu

Horizon Europe[2] nebo z Programu KAPPA Technologické agentury ČR[3] se zavazují, že budou spravovat data v souladu s principy FAIR. Směrnice Evropského parlamentu a Rady (EU) 2019/1024 ze dne 20. června 2019 o otevřených datech a opakovaném použití informací veřejného sektoru[4] uvádí, že „členské státy podporují dostupnost údajů z výzkumu tím, že přijmou vnitrostátní politiky a příslušná opatření s cílem otevřeně zpřístupnit všechny údaje z výzkumu financovaného z veřejných prostředků (‚politiky otevřeného přístupu') podle zásady ‚standardního zpřístupnění dokumentů ve formě otevřených dat' a slučitelných se zásadami FAIR". Dále uvádí, že „by měly být zohledněny obavy týkající se práv duševního vlastnictví, ochrany osobních údajů a důvěrnosti, bezpečnosti a oprávněných obchodních zájmů, v souladu se zásadou ‚co nejotevřenější, uzavřené podle potřeby'". Toto je akcentováno také v návrhu zákona, který slouží k transpozici zmíněné směrnice do českého právního řádu.[5]

V neposlední řadě pak FAIR principy vyzdvihuje rovněž doporučení UNESCO (2021) a také Evropský kodex integrity výzkumu (ALLEA 2017), který k principům FAIR říká, že „výzkumní pracovníci a výzkumné instituce a organizace zajišťují, aby přístup k datům byl při nezbytných omezeních co nejotevřenější a aby v relevantních případech odpovídal zásadám správy dat ‚FAIR', tj. dohledatelnosti, přístupnosti, interoperabilitě a opětovné použitelnosti."

2.3.1 Jak zajistit, aby data byla FAIR

Aby tedy výzkumník zajistil, že data, která vytvořil, jsou FAIR, potřebuje řešit především následující oblasti: dokumentaci, souborové formáty, přístup k datům, perzistentní identifikátory a licenci pro data.

Výzkumná data by měla být doprovázena bohatou **dokumentací**, která jednak zasadí data do kontextu a jednak poskytne doplňující informace o jejich autorovi, podmínkách opětovného využití a další. Informace, které jsou uvedené v související publikaci, často nemusí být pro správnou interpretaci dat dostačující, proto je vhodné do dokumentace doplnit informace, které např. v původní studii nebyly relevantní, ale mohou usnadnit pochopení kontextu, v jakém data vznikala, a tak umožnit jejich využití v dalším výzkumu. Dokumentace by měla dále obsahovat informace o licenci, aby uživatelé dat věděli, co s nimi mohou nebo naopak nemohou dělat. Její součástí může být například také doporučený formát citace dané datové sady pro usnadnění jejího citování.

Aby bylo umožněno co nejširší využití výzkumných dat a také pro zachování dlouhodobé využitelnosti, je vhodné data sdílet v otevřeném **formátu** (aby uživatelé nebyli omezeni tím, že nemají k dispozici placený software, který je k otevření dat potřebný) nebo ve formátu, který se v daném oboru typicky využívá a lze tedy předpokládat, že další výzkumníci v oboru budou moci s takovými daty dále pracovat.

---

[2] Podmínky vztahující se k open science jsou specifikovány v příloze 5 modelové grantové smlouvy, která je k dispozici na stránkách Evropské komise: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/common/agr-contr/general-mga_horizon-euratom_en.pdf. Ve *Studii* vycházíme z verze 1.1 z 15. prosince 2021

[3] Informace o otevřeném přístupu v programu KAPPA jsou k dispozici na stránkách TA ČR: https://www.tacr.cz/otevreny-pristup-v-programu-kappa/

[4] https://eur-lex.europa.eu/legal-content/CS/TXT/?uri=CELEX:32019L1024

[5] Text vychází z materiálu předloženého Vládě dne 31. 3. 2021, dostupný z eKLEPu pro veřejnost pod identifikačním číslem KORNBRYF659P: https://apps.odok.cz/veklep-detail?pid=KORNBRYF659P

Jak bylo zmíněno výše, data by měla být sdílena v souladu s principem „co nejotevřenější, uzavřené podle potřeby“. Zveřejňování dat nesmí být v rozporu s etikou či právem, proto je třeba určit, kdo má mít k datům **přístup** a za jakých podmínek. Některé typy dat není možné sdílet vůbec (například data související s národní bezpečností), u některých může být možnost sdílení omezena. V takovém případě je možné využít řízený přístup k datům, kdy si autor zachová kontrolu nad tím, kdo k datům může přistupovat a co s nimi může dělat. Pokud data mohou být zveřejněna, pak by k nim měl být zajištěn otevřený přístup. K tomuto účelu je v současné době nejlepší využít důvěryhodné **datové repozitáře**. Pro uchování dat jsou nejvhodnější oborové repozitáře, které jsou typicky lépe uzpůsobené potřebám daného oboru, například pokud se jedná o typ dat, která se v daném oboru produkují (např. LINDAT/CLARIAH-CZ[6] provozovaný Ústavem formální a aplikované lingvistiky na MFF UK). Pokud není k dispozici oborový nebo institucionální datový repozitář, je možné využít obecné repozitáře jako je Zenodo[7], Dryad[8] nebo figshare[9]. I v případě, že není možné sdílet data samotná, měla by být přístupná alespoň jejich metadata, tedy záznam o tom, že taková data existují.

Pro zvýšení nalezitelnosti datové sady, především stroji, je třeba opatřit data **perzistentním identifikátorem** (např. DOI nebo handle). Výzkumník ale zpravidla nemůže svým výstupům přidělit perzistentní identifikátor sám, na to je třeba služba, která má oprávnění identifikátory přidělovat. U výzkumných publikací identifikátory zpravidla přiděluje vydavatel, u výzkumných dat mohou tuto službu poskytovat datové repozitáře. Zároveň je také vhodné používat perzistentní identifikátory nejen pro datové sady, ale také pro odkazování na další unikátní objekty v rámci výzkumného procesu (např. identifikátory publikací, datových sad nebo autorů).

Aby případní uživatelé věděli, jak s daty mohou dále nakládat, je vhodné opatřit data **licencí**, která jasně určí, za jakých podmínek je možné data dále využít. Nejznámější a nejčastěji používané licence pro tyto účely jsou Creative Commons[10], prostřednictvím kterých může autor některá svá práva k dílu uživatelům poskytnout a jiná si naopak vyhradit. Existuje také možnost využít dvojího licencování, kdy může autor poskytovat stejná data pod různými licencemi (např. zdarma pro využití ve výzkumu a za úplatu pro komerční využití, viz např. The Cornell Lab of Ornithology: Macaulay Library[11]).

Z pohledu výzkumníka je jednodušší a efektivnější pracovat s principy FAIR v průběhu výzkumného projektu než se snažit je naplnit na jeho konci. Roste proto také význam správy výzkumných dat, a to po celou dobu životního cyklu výzkumných dat: od jejich vytváření přes zpracování a analýzu až po jejich dlouhodobé uchovávání a potenciální sdílení (viz Obr. 3).

S plánováním a správou výzkumných dat může pomoci **data management plan**. Jedná se o dokument, který popisuje, jaká data a jakým způsobem byla během výzkumu vytvářena, a obsahuje informace o jejich dostupnosti a využití.

Následující kapitola popisuje, jak výzkumní pracovníci na UK pracují se svými výzkumnými daty právě ve vztahu k životnímu cyklu výzkumných dat.

[6] https://lindat.mff.cuni.cz/repository/xmlui/
[7] https://zenodo.org/
[8] https://datadryad.org/stash/
[9] https://figshare.com/
[10] http://www.creativecommons.cz/
[11] https://www.macaulaylibrary.org/

Obr. 3: *Životní cyklus výzkumných dat.*

3. Práce s výzkumnými daty na Univerzitě Karlově

Výzkum založený na datech se stává čím dál běžnější napříč různými vědními disciplínami a práce s daty je tedy čím dál častěji nedílnou součástí činnosti každého vědeckého pracovníka. S rozvojem technologických možností a také v souvislosti s probíhajícími diskuzemi o důvěryhodnosti vědy roste tlak na zvýšení kvality, dostupnosti a reprodukovatelnosti výsledků výzkumu a to jak ze strany samotné vědecké komunity (viz např. Open Science Collaboration 2015; Markowetz 2015; Munafò 2017; Errington et al. 2021 aj.), tak také ze strany akademických vydavatelů (viz otevřené datové politiky některých vydavatelů), poskytovatelů financí (viz program Horizon Europe Evropské Komise, program KAPPA TA ČR) nebo evropské či národní legislativy (viz Směrnice Evropského parlamentu a Rady (EU) 2019/1024 ze dne 20. června 2019 o otevřených datech a opakovaném použití informací veřejného sektoru). Objevují se pak zejména požadavky na otevřený přístup k publikacím (open access) a zvýšený důraz na dostupnost a kvalitu produkovaných dat (data spravovaná v souladu s principy FAIR, vytvoření data management plánu).

Odpovědnost za správu vědeckých dat ovšem není možné klást jen a pouze na vědecké pracovníky, kteří je vytvářejí. V mnoha aspektech jsou do značné míry závislí na institucionální podpoře – a to jak technické (např. repozitáře, nástroje a další infrastruktura), tak do určité míry metodické (např. jaké vlastnosti musí data mít, aby plnila požadavky FAIR principů). Neexistence / nedostupnost této institucionální podpory může vést k větší zátěži výzkumníků (mimo oblast samotné podstaty výzkumu), větší náročnosti splnění požadavků poskytovatelů financí a tím i ke konkurenční nevýhodě oproti kolegům z jiných institucí.

S ohledem na tyto nové požadavky v oblasti správy výzkumných dat a jejich zpřístupňování v souladu s principy FAIR bylo potřeba zmapovat aktuální praxi a potřeby vědců v oblasti správy výzkumných dat na Univerzitě Karlově. Za tímto účelem bylo připraveno dotazníkové šetření, jehož výsledky mohou sloužit jako důležitý podklad pro přípravu komplexní infrastruktury a služeb v oblasti správy výzkumných dat na UK.

3.1 Dotazníkové šetření

Centrum pro podporu open science připravilo ve spolupráci s dalšími součástmi rektorátu dotazník *Analýza správy a ukládání výzkumných dat na Univerzitě Karlově*. Struktura otázek vycházela z obdobného šetření prováděného na Masarykově univerzitě a ve stejné době připravovaného na Vysoké škole báňské – Technické univerzitě Ostrava.[12] Dotazníkové šetření bylo provedeno na základě materiálu předloženého Kolegiu rektora 24. 5. 2021 a probíhalo od 8. 6. do 18. 7. 2021. Dotazník byl připraven v nástroji Microsoft Forms, byl distribuován proděkanům pro vědu a oddělením pro vědu jednotlivých fakult Univerzity Karlovy a také sdílen prostřednictvím sociálních sítí Ústřední knihovny UK. Dotazník byl primárně určen akademickým a vědeckým pracovníkům a doktorandům zapojeným do výzkumu. Vyplnění dotazníku probíhalo anonymně a získané údaje nebyly spojovány s konkrétní osobou.

Struktura dotazníku reflektovala životní cyklus výzkumných dat (viz Obr. 3 výše). Otázky se tak týkaly následujících oblastí:

[12] Dotazníky použité na Masarykově univerzitě a VŠB-TUO jsou k dispozici na webových stránkách Pracovní skupiny pro správu vědeckých dat v ČR: https://www.wg-rdm.cz/aktivity/pruzkum-sprava-dat-vedci

1. Práce s výzkumnými daty během výzkumu
2. Dlouhodobé uchovávání výzkumných dat
3. Sdílení výzkumných dat
4. Podpora správy výzkumných dat

Celý dotazník se všemi otázkami a tabulka obsahující odpovědi respondentů jsou k dispozici v repozitáři Zenodo[13]. Z odpovědí byly odstraněny slovní komentáře, e-mailové adresy (pokud je respondenti vyplnili) a informace o zajímavých datových sadách (tzv. datové perly).

Následující sekce studie nejprve stručně charakterizují respondenty a následně představí výsledky dotazníku dle jednotlivých oblastí (viz výše). Každou sekci věnující se práci s daty uzavírá krátké shrnutí slovních komentářů, které respondenti k dané oblasti uvedli. Vyplnění slovních komentářů nebylo povinné, často se jich tedy nesešlo mnoho a nelze z nich tudíž vyvozovat obecné závěry. Respondenti nicméně využili možnost komentářů pro upozornění na oblasti, které jim přišly důležité, případně na oblasti, kde měly pocit, že se jim dotazník dostatečně nevěnoval, proto uvádíme jejich shrnutí.

3.2 Respondenti

Dotazníkového šetření se zúčastnilo celkem **405 respondentů** a zastoupeny byly všechny fakulty Univerzity Karlovy. Vzhledem k tomu, že někteří zaměstnanci působí na více fakultách zároveň, bylo možné v dotazníku zvolit více možností. Z toho důvodu výstupy šetření z hlediska zastoupení fakult početně nesouhlasí s počtem respondentů.

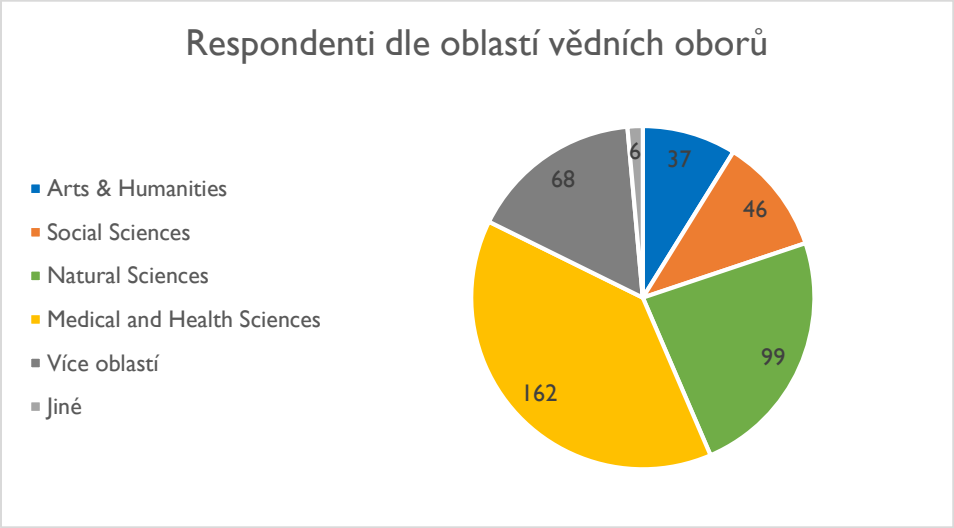
Fakulty nebyly zastoupeny rovnoměrně (viz Tab. 1), více odpovědí obecně dodaly fakulty s medicinským a přírodovědným zaměřením, nejméně pak teologické fakulty, pedagogická a právnická fakulta. Vzhledem k existujícím oborovým rozdílům nejsou výzkumná data a jejich správa pro všechny obory stejně významná, což se mohlo projevit i na skladbě respondentů dotazníku.

[13] https://doi.org/10.5281/zenodo.5841472

Matematicko-fyzikální fakulta	59
Farmaceutická fakulta v Hradci Králové	57
2. lékařská fakulta	47
Přírodovědecká fakulta	47
1. lékařská fakulta	43
Lékařská fakulta v Plzni	43
Lékařská fakulta v Hradci Králové	25
Filozofická fakulta	24
Fakulta humanitních studií	20
Fakulta sociálních věd	16
Fakulta tělesné výchovy a sportu	9
Katolická teologická fakulta	5
Husitská teologická fakulta	5
3. lékařská fakulta	5
Pedagogická fakulta	5
Evangelická teologická fakulta	2
Právnická fakulta	2
Centrum pro ekonomický výzkum a doktorské studium (CERGE-EI)	2
Centrum pro otázky životního prostředí	1
Ústav dějin Univerzity Karlovy a archiv Univerzity Karlovy	1
Ústřední knihovna	1
Jiné	1

Tab. 1: *Počet respondentů dle organizačních jednotek UK.*

Respondenti volili jeden či více **vědních oborů** z oblastí dle vnitřní klasifikace UK. Respondenti mohli vybrat více vědních oborů a poměrně mnoho (více než 17%) jich zvolilo kombinaci z více vědních oblastí (např. 1.5. Arts & Humanities: Linguistics; 3.2. Natural Sciences: Computer Science). Graf 1 znázorňuje počet odpovědí dle oblastí vědních oborů a následně Tab. 2 uvádí podrobné zastoupení jednotlivých vědních oborů.



Graf 1: *Počet odpovědí podle oblastí vědních oborů dle vnitřní klasifikace UK.*

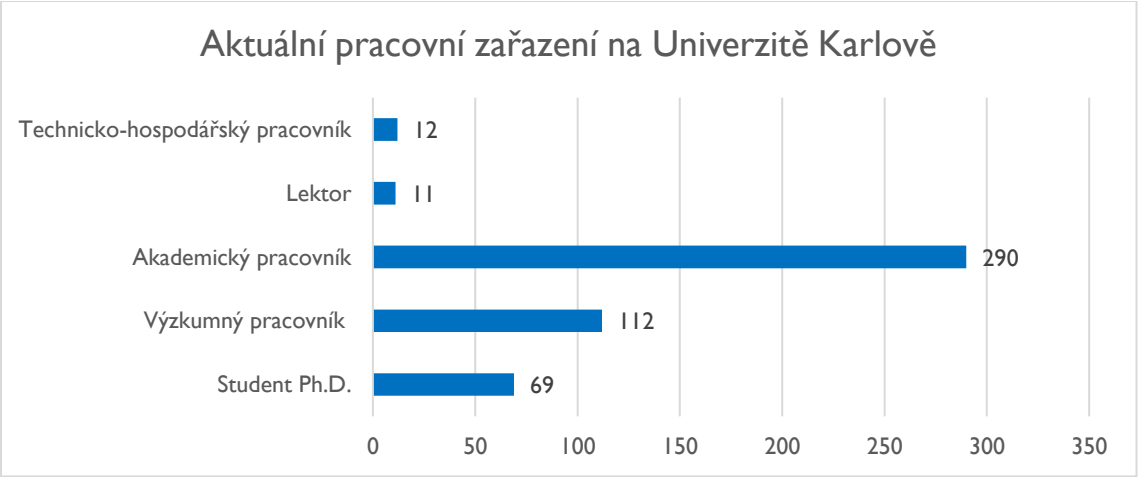
4.13. Medical and Health Sciences: Pharmaceutical Sciences	57
3.1. Natural Sciences: Biology	49
4.11. Medical and Health Sciences: Oncology and Haematology	40
3.6. Natural Sciences: Chemistry	37
3.2. Natural Sciences: Computer Science	31
3.8. Natural Sciences: Physics	30
1.5. Arts & Humanities: Linguistics	29
4.10. Medical and Health Sciences: Neurosciences	26
4.1. Medical and Health Sciences: Cardiovascular Science	23
4.8. Medical and Health Sciences: Medical Diagnostics and Basic Medical Sciences	23
4.3. Medical and Health Sciences: Health Sciences	19
2.9. Social Sciences: Sociology and Applied Social Sciences	15
4.4. Medical and Health Sciences: Immunity and Infection	15
4.15. Medical and Health Sciences: Surgical Disciplines	15
Jiné	14
4.6. Medical and Health Sciences: Internal Disciplines	13
1.4. Arts & Humanities: History	11
3.4. Natural Sciences: Geography	11

1.8. Arts & Humanities: Theology and Religious Studies	10
4.12. Medical and Health Sciences: Pediatrics	9
2.6. Social Sciences: Political Science	8
4.9. Medical and Health Sciences: Metabolic Diseases	8
2.2. Social Sciences: Economics	7
2.10. Social Sciences: Sport Sciences – Social	7
2.11. Social Sciences: Subject Specific Education Research	7
3.7. Natural Sciences: Mathematics	7
4.14. Medical and Health Sciences: Sport Sciences – Biomedical & Rehabilitation Medicine	6
2.1. Social Sciences: Area Studies	5
2.8. Social Sciences: Social and Cultural Anthropology	5
1.1. Arts & Humanities: Archaelogy	4
2.3. Social Sciences: General Education and Pedagogy	4
3.5. Natural Sciences: Geological Sciences	4
4.5. Medical and Health Sciences: Intensive Care Medicine	4
4.7. Medical and Health Sciences: Maternal and Childhood Care	4
5.1. Engineering and Technology	4
1.7. Arts & Humanities: Philosophy	3
2.4. Social Sciences: Law	3
2.5. Social Sciences: Media and Communication Studies	3
2.7. Social Sciences: Psychological Sciences	3
3.3. Natural Sciences: Environmental and Sustainability Research	3
4.2. Medical and Health Sciences: Dental Medicine	3
1.3. Arts & Humanities: Asian, Middle Eastern and East European Studies	2
1.2. Arts & Humanities: Arts and Culture Studies	1
3.9. Natural Sciences: Transdisciplinary Research of the Anthropocene	1
6.1. Agricultural and Veterinary Sciences	1
1.6. Arts & Humanities: Literature	0

Tab. 2: *Počet odpovědí podle jednotlivých vědních oborů dle vnitřní klasifikace UK.*

Nerovnoměrné zastoupení oborů a jejich skupin je nutné brát na zřetel při vyvozování závěrů – u některých oblastí mohou být výsledky považovány za reprezentativnější než u jiných. Jsme si vědomi toho, že typy dat a oborové zvyklosti jsou často velmi odlišné. To ovšem nebrání ve vyhodnocení některých otázek k obecnější problematice, které se objevují napříč všemi skupinami.

Podle **aktuálního pracovního zařazení** se dotazníkového šetření zúčastnilo nejvíce akademických (n=290) a výzkumných (n=112) pracovníků, nejméně pak lektorů (n=11). I v této kategorii mohli respondenti zvolit více možností, protože mohou být na UK zaměstnáni ve více rolích. Za důležité v kontextu dotazníku považujeme zejména skupiny „Výzkumný pracovník" a „Student Ph.D." u nichž lze předpokládat velmi aktivní práci s daty.



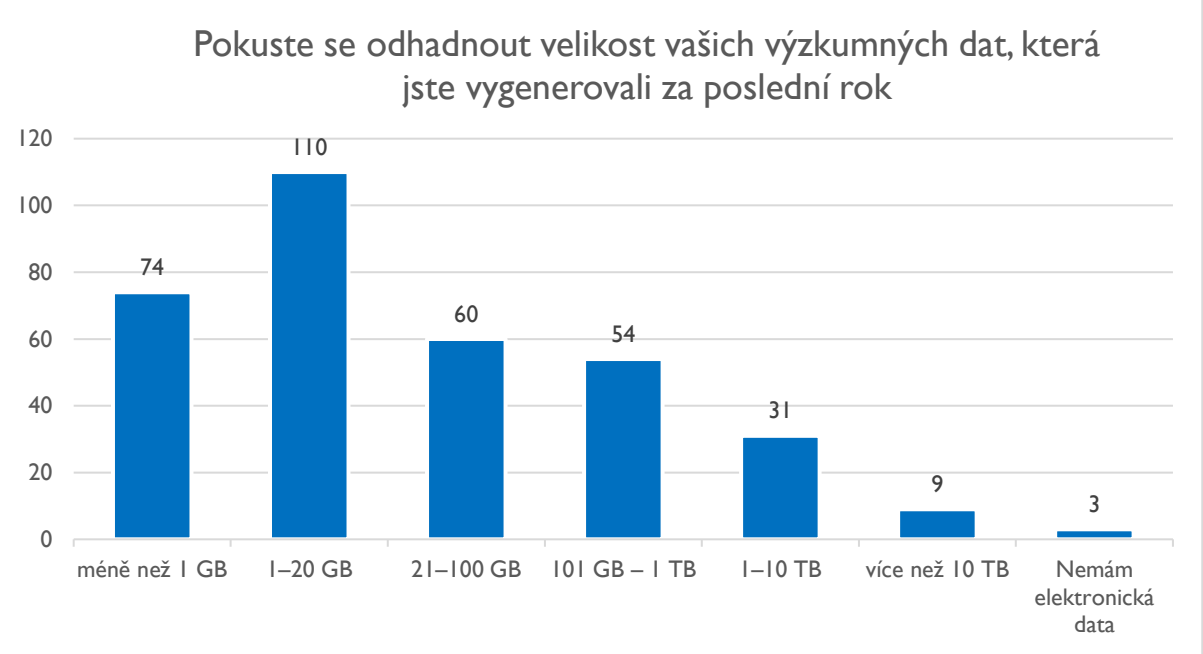
Graf 2: *Počet odpovědí dle pracovního zařazení respondenta.*

3.3 Vyhodnocení odpovědí

3.3.1 Oblast 1: Práce s výzkumnými daty během výzkumu

Objem dat

Množství vyprodukovaných dat se velmi významně liší mezi obory – obrázky z mikroskopů, nebo z některých fyzikálních měření jsou v řádově objemnější než např. textová data. Velké objemy dat kladou významné nároky na infrastrukturu, důležité je u nich zvážit i aspekty případné další dostupnosti a dlouhodobého uchování (v rámci tzv. long-term preservation systemů – LTP). Menší datové sady však nejsou o nic méně důležité a v řadě aspektů (zejména metodických) se od velkých datových sad příliš neliší. Graf 3 zobrazuje odhad respondentů, jaký objem dat vygenerovali za poslední rok (z grafu jsou vynechány odpovědi, kde respondenti uvedli, že nejsou schopni objem dat odhadnout [n=63]).



Graf 3: *Odhad objemu výzkumných dat vyprodukovaných za jeden rok. Zobrazeny jsou pouze odpovědi respondentů, kteří odhad uvedli či odpověděli, že elektronická data nemají.*

Pokud budeme uvažovat jen počet respondentů, kteří dokážou zhodnotit množství dat vygenerovaných za rok (n=338), můžeme říci že:

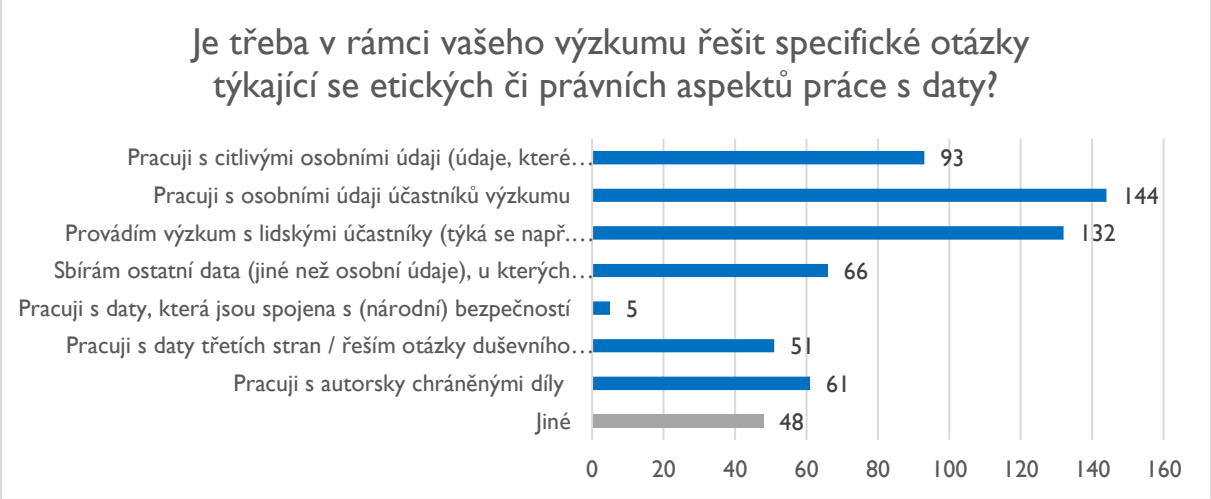
- 54,4% respondentů vygeneruje méně než 20 GB dat za rok (184/338)
- Data o velikosti nad 100 GB ročně vygeneruje 27,8% respondentů (94/338)
- Data o velikosti nad 1TB ročně generuje 11,8% respondentů (40/338)

Chráněná data

Existuje celá řada legitimních důvodů, proč je třeba omezit přístup k datům, např. pokud se jedná o ochranu osobních údajů, ochranu duševního vlastnictví nebo o data, která jsou spojena s národní bezpečností. Principy FAIR jsou v souladu i s těmito požadavky (uplatňuje se princip „as open as possible, as closed as necessary“). V některých případech je však třeba sdílet tento typ dat v rámci výzkumného týmu, proto je důležité poskytnout výzkumníkům bezpečné úložiště s možností řízeného přístupu, což klade další nároky na infrastrukturu.

Je třeba v rámci vašeho výzkumu řešit specifické otázky týkající se etických či právních aspektů práce s daty?	
Pracuji s citlivými osobními údaji (údaje, které vypovídají o rasovém či etnickém původu, politických názorech, náboženském vyznání či filozofickém přesvědčení, členství v odborech, zdravotním stavu či o sexuálním životě nebo sexuální orientaci fyzické osoby)	93
Pracuji s osobními údaji účastníků výzkumu	144
Provádím výzkum s lidskými účastníky (týká se např. respondentů dotazníku i v případě, že nesbíráte osobní údaje)	132
Sbírám ostatní data (jiné než osobní údaje), u kterých je nutná ochrana (např. z důvodu komercializace, patentové ochrany apod.)	66
Pracuji s daty, která jsou spojena s (národní) bezpečností	5
Pracuji s daty třetích stran / řeším otázky duševního vlastnictví	51
Pracuji s autorsky chráněnými díly	61
Jiné	48

Tab. 3: *Výzkumná data z hlediska potřeby jejich ochrany.*



Graf 4: *Výzkumná data z hlediska potřeby jejich ochrany.*

Poměrně velká část respondentů dotazníku pracuje s daty, která nějakým způsobem spadají do kategorie chráněných dat (kategorie se mohou překrývat):

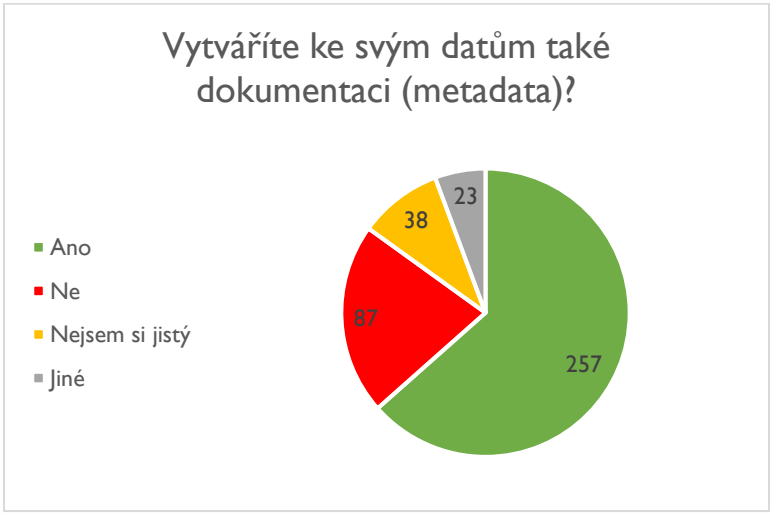
- 35,5% (144/405) pracuje s osobními údaji účastníků výzkumu, 22,6% (94/405) pracuje s citlivými osobními údaji
- 16,2% (66/405) pracuje s jinými než osobními daty, u kterých je potřeba ochrana
- 12,5% (51/405) pracuje s daty třetích stran a řeší otázky duševního vlastnictví

Dokumentace a metadata, oborové standardy

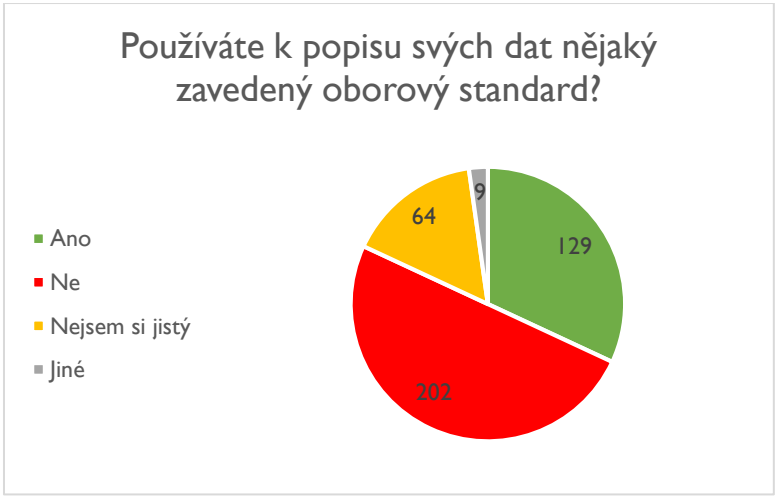
Popis dat (vytvoření metadat) je jedním z klíčových prvků FAIR principů. Pokud data nejsou řádně dokumentovaná a zasazená do kontextu, výrazně se sníží možnost jejich opětovného využití. To platí

jak pro potenciálního uživatele, tak i pro samotného autora dat, pokud se k nim v budoucnu vrátí. Zároveň se také zvýší riziko, že dojde k mylné interpretaci dat. Na druhou stranu tvorba nějaké formy dokumentace či metadat tvoří další zátěž na původce dat a její výhody mohou být v samotný okamžik tvorby dat podceněné. V dotazníkovém šetření více než polovina respondentů (63%) uvedla, že dokumentaci ke svým datům vytváří.

Využití oborových standardů je klíčové pro poskytování dalších služeb k datům a pro interoperabilitu. Standardy do značné míry zajišťují kompatibilitu a porovnatelnost s daty z jiných zdrojů a usnadňují dokumentaci dat. Při interpretaci výsledků šetření a také při budování institucionální podpory je třeba mít na paměti, že oborový standard nemusí existovat pro každý typ dat, nebo může konkrétní situace vyžadovat širší datové schéma.



Graf 5: *Rozdělení respondentů podle toho, zda ke svým datům vytváří dokumentaci či nikoli.*



Graf 6: *Rozdělení respondentů podle toho, zda popisu dat používají oborový standard či nikoli.*

Téměř polovina respondentů (49,9% = 202/405) nepoužívá oborové standardy, nějaký oborový standard používá téměř třetina (31,9% = 129/405) a zbývající respondenti si nejsou jisti, nebo zvolili slovní odpověď.

Z dotazníku není možné přesně zjistit, z jakého důvodu respondenti oborové standardy nepoužívají. Může jít o oborová specifika, existenci/neexistenci takových standardů, nebo i o (ne)vůli přizpůsobovat svá data existujícím strukturám.

Ukládání dat v průběhu výzkumu, zálohy a sdílení

Z hlediska běžné činnosti v rámci správy výzkumných dat jsou klíčovými aspekty jejich ukládání, zálohování a sdílení. FAIR principy zdánlivě nastavují požadavky především na finální produkt – datovou sadu, kterou je možné v ideálním případě dále použít (např. pro replikaci publikovaných výsledků, v dalším výzkumu apod.). Nicméně i procesy probíhající před samotnou publikací výzkumu kladou nároky (v některých případech značné) na infrastrukturu a dodržování pravidel správy výzkumných dat. Jde například o bezpečné uchování a ochranu citlivých a jinak chráněných dat.[14]

Na svém/týmovém lokálním disku ukládá svá data 80% (324/405) respondentů. Sdílené datové úložiště svého pracoviště a externí úložné jednotky používá na uložení dat necelá polovina respondentů (48,9% 198/405 resp. 48,6% 197/405).

Významná část respondentů používá Google Drive (27% 109/405), MS Office 365 (23,5% 95/405) a Dropbox (13,3% 54/405). Využití těchto služeb naznačuje potřebu snadno použitelného prostředí pro sdílení dat. Na druhou stranu ale v případě zejména Google a Dropboxu vyvstává otázka, za jakých licenčních podmínek respondenti tyto služby využívají (ty mohou být problematické z hlediska bezpečnosti dat).

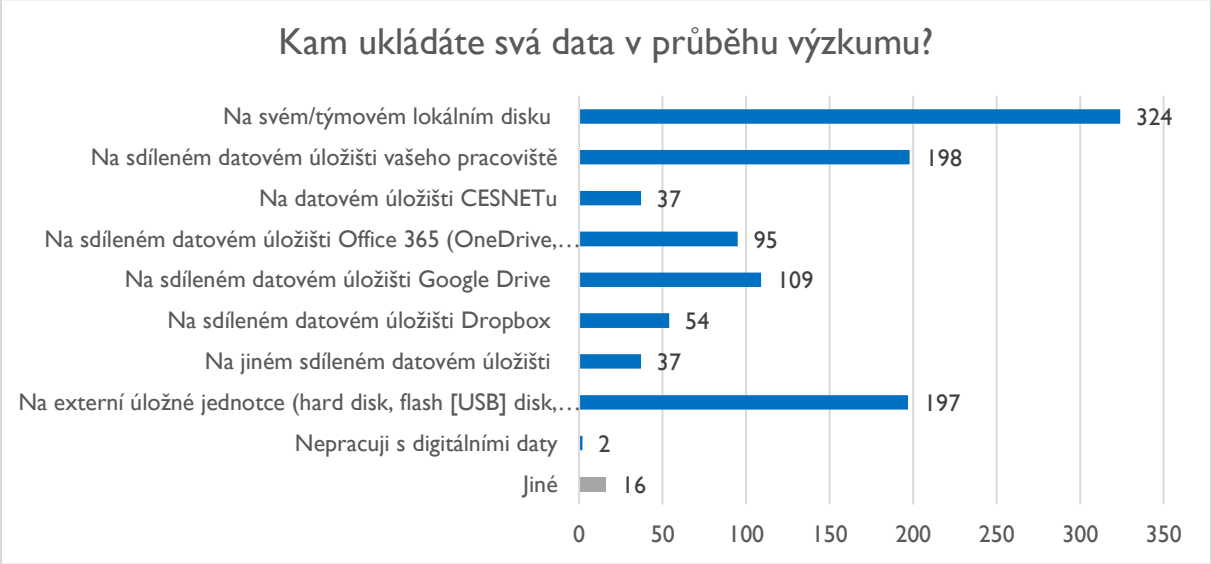
Datové úložiště CESNETu, které je z hlediska bezpečnosti jedním z nejlepších řešení (viz metodické pokyny Ústavu výpočetní techniky a Pověřence pro ochranu osobních údajů[15]), využívá relativně malé množství respondentů (9,1% 37/405).

Ve slovních doplněních k variantě „Jiné“ (celkem 16 odpovědí) pak respondenti zmiňovali využívání GitHubu nebo repozitáře LINDAT/CLARIAH-CZ, které můžeme brát jako příklad služeb pro obory, které někdy otevřeně sdílejí data už i v ranné fázi výzkumného procesu.

[14] Ústav výpočetní techniky UK ve spolupráci s univerzitním Pověřencem pro ochranu osobních údajů proto připravili metodické pokyny pro bezpečné nakládání s daty a výběr datového úložiště, které jsou k dispozici na stránkách Centra pro podporu open science: https://openscience.cuni.cz/OSCI-94.html
[15] https://openscience.cuni.cz/OSCI-94.html

Kam ukládáte svá data v průběhu výzkumu?	
Na svém/týmovém lokálním disku	324
Na sdíleném datovém úložišti vašeho pracoviště	198
Na datovém úložišti CESNETu	37
Na sdíleném datovém úložišti Office 365 (OneDrive, SharePoint, MS Teams)	95
Na sdíleném datovém úložišti Google Drive	109
Na sdíleném datovém úložišti Dropbox	54
Na jiném sdíleném datovém úložišti	37
Na externí úložné jednotce (hard disk, flash [USB] disk, CD/DVD apod.)	197
Nepracuji s digitálními daty	2
Jiné	16

Tab. 4: *Využití typů úložišť pro práci s daty v průběhu výzkumu.*



Graf 7: *Využití typů úložišť pro práci s daty v průběhu výzkumu.*

V otázce **zálohování** téměř polovina (48,9% 198/405) respondentů uvedla, že v průběhu výzkumu používá externí úložné jednotky – ty mají sice výhodu snadné manipulace, zároveň však mají i některá specifická rizika (fyzická ztráta zařízení, fyzické poškození, ztráta podrobností o uložených datech po delší době apod.).

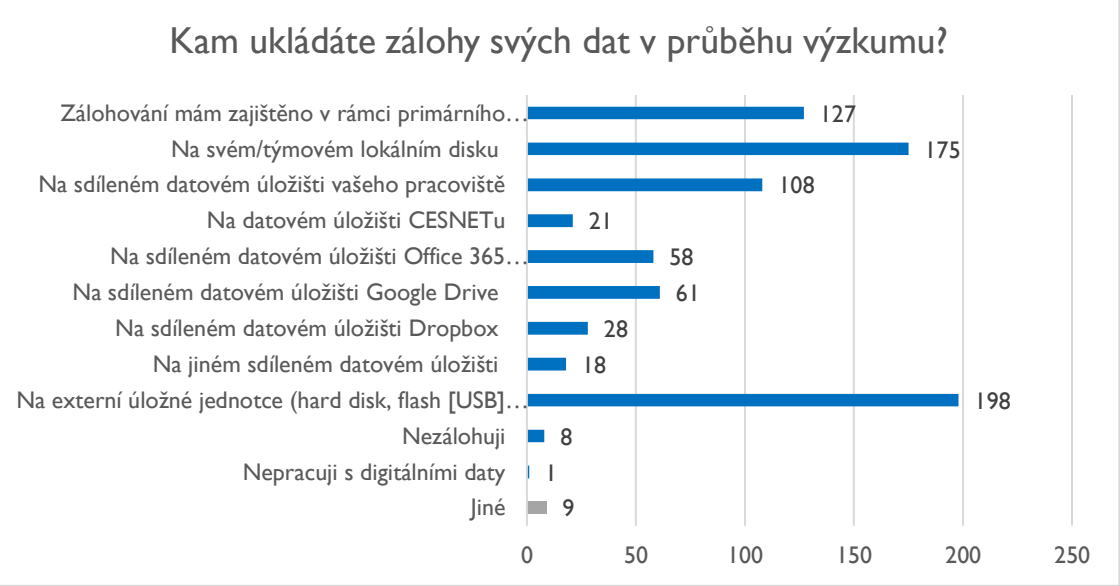
Další poměrně velké skupiny respondentů zálohují na svém či týmovém lokálním disku (43% 175/405), mají zálohování zajištěno v rámci primárního úložiště (31,4% 127/405), nebo v rámci sdíleného úložiště pracoviště (26,7% 108/405).

Online služby Google Drive, Office 365 a Dropbox používá k zálohování poměrně malé množství respondentů. Služby CESNETu využívá 5% (21/405) respondentů.

8 respondentů uvedlo, že nezálohuje, což ovšem může být dáno i povahou některých (zejména velmi objemných) dat.

Kam ukládáte zálohy svých dat v průběhu výzkumu?	
Zálohování mám zajištěno v rámci primárního úložiště	127
Na svém/týmovém lokálním disku	175
Na sdíleném datovém úložišti vašeho pracoviště	108
Na datovém úložišti CESNETu	21
Na sdíleném datovém úložišti Office 365 (OneDrive, SharePoint, MS Teams)	58
Na sdíleném datovém úložišti Google Drive	61
Na sdíleném datovém úložišti Dropbox	28
Na jiném sdíleném datovém úložišti	18
Na externí úložné jednotce (hard disk, flash [USB] disk, CD/DVD apod.)	198
Nezálohuji	8
Nepracuji s digitálními daty	1
Jiné	9

Tab. 5: *Využití typů úložišť pro zálohování dat v průběhu výzkumu.*



Graf 8: *Využití typů úložišť pro zálohování dat v průběhu výzkumu.*

Sdílení prostřednictvím e-mailu je mezi respondenty nejčastěji používaný způsob sdílení dat v průběhu výzkumu (59,3% 240/405). Tato možnost je z uvedených nejjednodušší, i když není vhodná pro všechny typy dat a otázkou může být i dodržení pravidel bezpečnosti práce s daty.

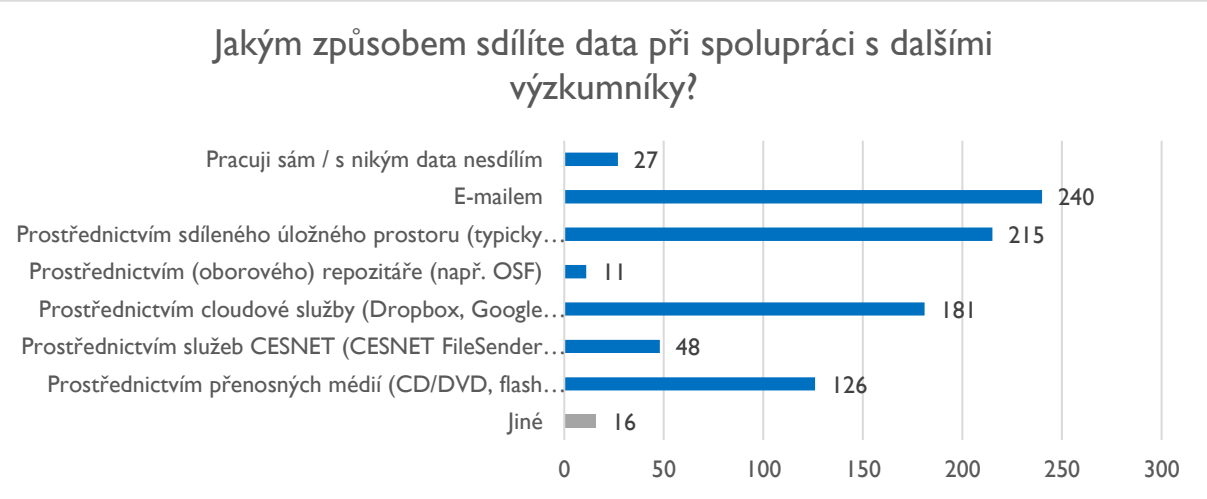
Sdílení prostřednictvím sdíleného úložného prostoru (53,1% 215/405) a cloudových služeb (44,7% 181/405) využívá podobné množství respondentů. Poměrně významné zastoupení (11,9% 48/405) mají i služby CESNET (FileSender apod.)

Nezanedbatelná část respondentů (31,1% 126/405) využívá pro sdílení dat přenosná média.

Jen relativně malá část respondentů uvedla, že data s nikým nesdílí (6,7% 27/405).

Jakým způsobem sdílíte data při spolupráci s dalšími výzkumníky?	
Pracuji sám / s nikým data nesdílím	27
E-mailem	240
Prostřednictvím sdíleného úložného prostoru (typicky sdílené síťové disky)	215
Prostřednictvím (oborového) repozitáře (např. OSF)	11
Prostřednictvím cloudové služby (Dropbox, Google Drive apod.)	181
Prostřednictvím služeb CESNET (CESNET FileSender apod.)	48
Prostřednictvím přenosných médií (CD/DVD, flash [USB] disky apod.)	126
Jiné	16

Tab. 6: *Způsoby sdílení dat v rámci výzkumného týmu.*



Graf 9: *Způsoby sdílení dat v rámci výzkumného týmu.*

Slovní komentáře

V závěru sekce měli respondenti možnost doplnit slovní komentář týkající se práce s daty během výzkumu. V této části respondenti často zmiňovali potřebu zlepšení možností ukládání a (automatického) zálohování dat, zvláště pokud se jedná o objemná data, jako jsou např. snímky z mikroskopů. Dále také upozornili na důležitost prostorových dat, která z jejich pohledu nebyla v dotazníku dostatečně pokryta. Ta jsou hojně využívaná zvláště u některého typu výzkumu na Přírodovědecké fakultě UK a váže se k nim i specifický způsob tvorby, zpracování či vizualizace a často je k tomu nutný specializovaný software. Respondenti také zmínili problematiku práce s chráněnými daty (data obsahující osobní údaje apod.), především pak potřebu získat stanoviska Pověřence pro ochranu osobních údajů nebo Etické komise / Komise pro etiku ve výzkumu (např. někteří poskytovatelé financí či zahraniční partneři mohou tato stanoviska požadovat). V neposlední řadě také zmiňovali problematiku bezpečného přístupu k výzkumným datům pro výzkumné partnery mimo UK, kteří nemají přístup k univerzitní infrastruktuře.

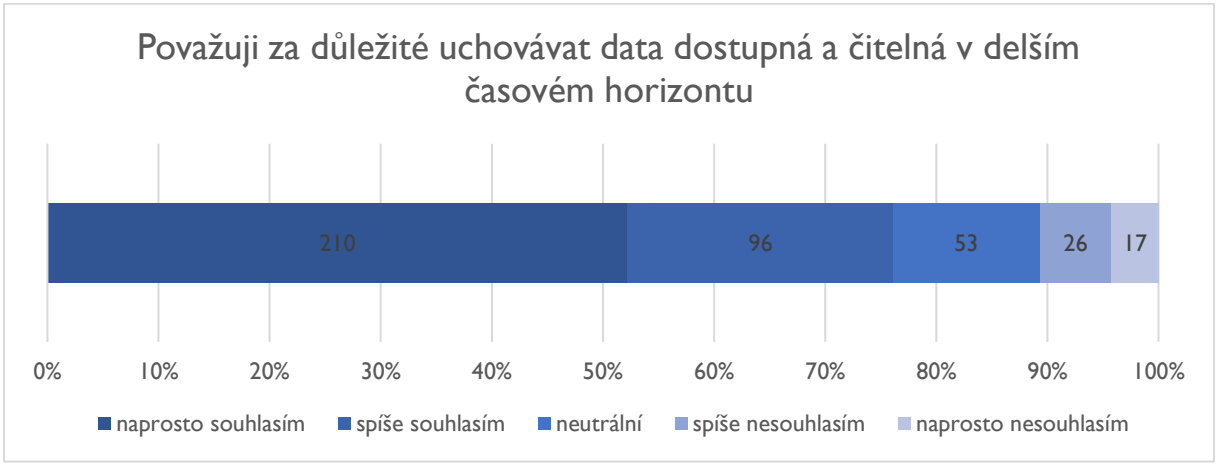
3.3.2 Oblast 2: Dlouhodobé uchovávání výzkumných dat

Dlouhodobé uchovávání výzkumných dat (LTP) je všeobecně jednou z méně podchycených problematik z oblasti správy výzkumných dat. V případě LTP nejde o přímý ekvivalent zálohování (i když k němu má principiálně blízko). V dlouhodobém uchování má důležitou roli nejen zajištění bitové ochrany (uložené soubory se v průběhu času nepoškodí/nezmění) a fyzické zabezpečení dat (například vícenásobné uložení v několika geografických lokacích), ale i volba dat která se mají (či nemají) dlouhodobě uchovávat, způsoby, jak nakládat se zastaralými formáty apod. Nezanedbatelná je vzhledem k dlouhodobosti uchování i cena. I v této oblasti tak jde jak o infrastrukturu a nástroje (úložiště a vhodný LTP systém), tak o metodické postupy (co ukládat, na jak dlouho, jak financovat).

Repozitář pro zveřejňování/publikování dat může v tomto procesu hrát důležitou úlohu. Pokud budou data splňující FAIR principy uložena v repozitáři, může takový repozitář sloužit jako zdroj pro LTP řešení – u zpracovaných a zveřejněných dat je větší pravděpodobnost, že by měla být uchována dlouhodobě.

Více než polovina respondentů (51,9% 210/402) naprosto souhlasí s tvrzením, že je důležité uchovávat data dostupná a čitelná v delším časovém horizontu.

S tvrzením spíše a naprosto souhlasí (76,1% 306/402) respondentů a spíše a naprosto s ním nesouhlasí 10,7% 43/402 respondentů.



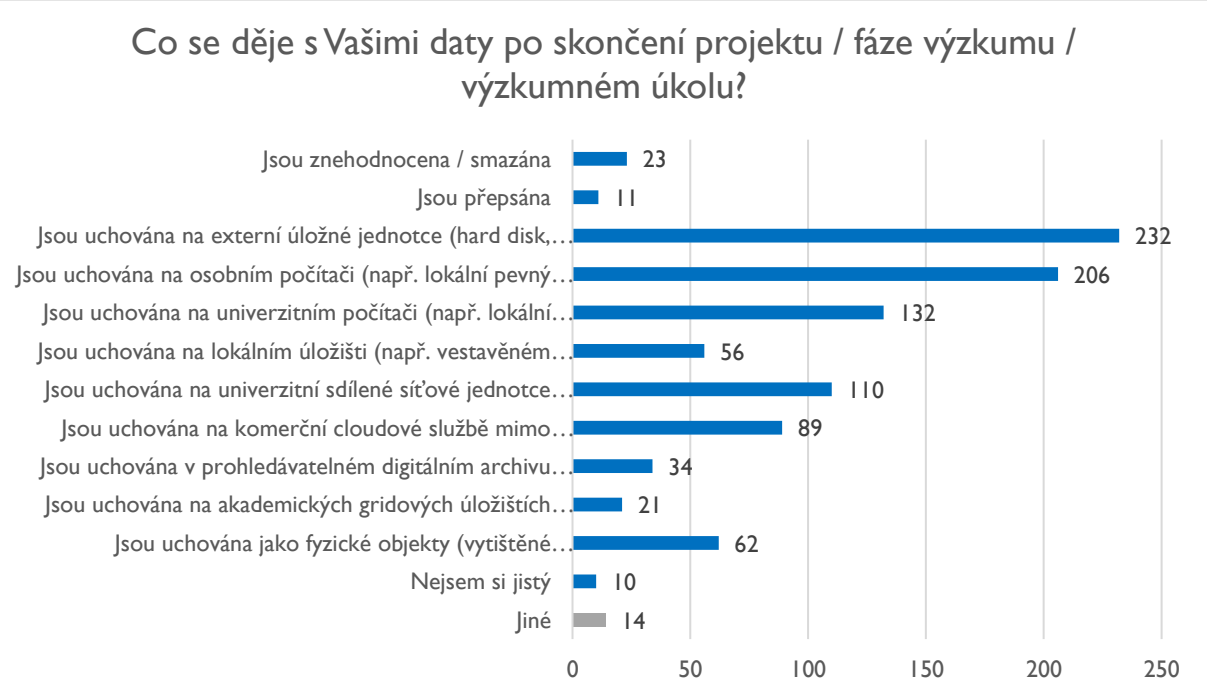
Graf 10: *Postoj respondentů k dlouhodobému uchovávání dat.*

Většina respondentů (57,3% 232/405) uchovává svá data po skončení projektu na externí úložné jednotce a/nebo na svém osobním počítači (50,8% 206/405). Na univerzitním počítači uchovává data 32,6% (132/405) respondentů, na univerzitním serveru (27,2% 110/405). Cloudové služby využívá 22% (89/405) respondentů.

Oproti tomu repozitář nebo jiný digitální archiv používá jen 8,3% (34/405).

Co se děje s Vašimi daty po skončení projektu / fáze výzkumu / výzkumném úkolu?	
Jsou znehodnocena / smazána	23
Jsou přepsána	11
Jsou uchována na externí úložné jednotce (hard disk, flash [USB] disk, CD/DVD apod.)	232
Jsou uchována na osobním počítači (např. lokální pevný disk osobního PC nebo notebooku)	206
Jsou uchována na univerzitním počítači (např. lokální pevný disk pracovního PC nebo notebooku)	132
Jsou uchována na lokálním úložišti (např. vestavěném pevném disku) univerzitního přístroje/sensoru, který generuje data	56
Jsou uchována na univerzitní sdílené síťové jednotce (úložném serveru), jehož obsah je průběžně zálohován	110
Jsou uchována na komerční cloudové službě mimo univerzitu (např. Dropbox, Google Drive / Google Cloud Storage, Amazon Cloud Storage, Microsoft OneDrive / Microsoft Cloud Storage)	89
Jsou uchována v prohledávatelném digitálním archivu (repozitáři)	34
Jsou uchována na akademických gridových úložištích nebo výpočetním clusteru (např. HPC centrum typu CESNET MetaCentrum apod.)	21
Jsou uchována jako fyzické objekty (vytištěné dokumenty v krabicích, chemické vzorky ve skladech apod.)	62
Nejsem si jistý	10
Jiné	14

Tab. 7: *Způsoby dlouhodobého uchovávání výzkumných dat.*

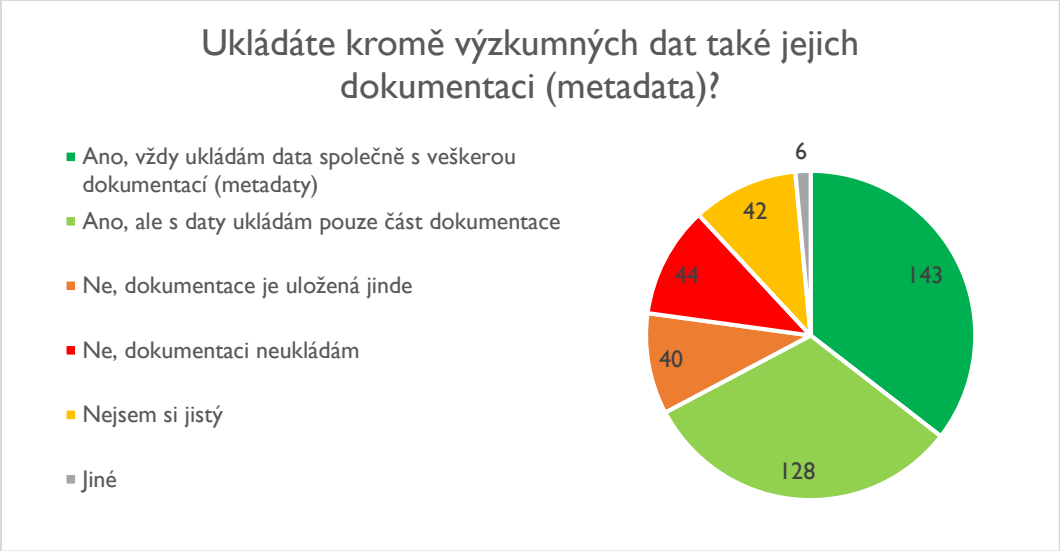


Graf 11: *Způsoby dlouhodobého uchovávání výzkumných dat.*

Dokumentace

Největší skupina respondentů (35,3% 143/405) ukládá data společně dokumentací, o něco méně respondentů (31,6% 128/405) s daty ukládá jenom část dokumentace.

Případná nedostupnost dokumentace k datům nicméně může vést k horší možnosti využít data v budoucnu – a to jak jinými uživateli, tak autorem dat, či jeho kolegy.

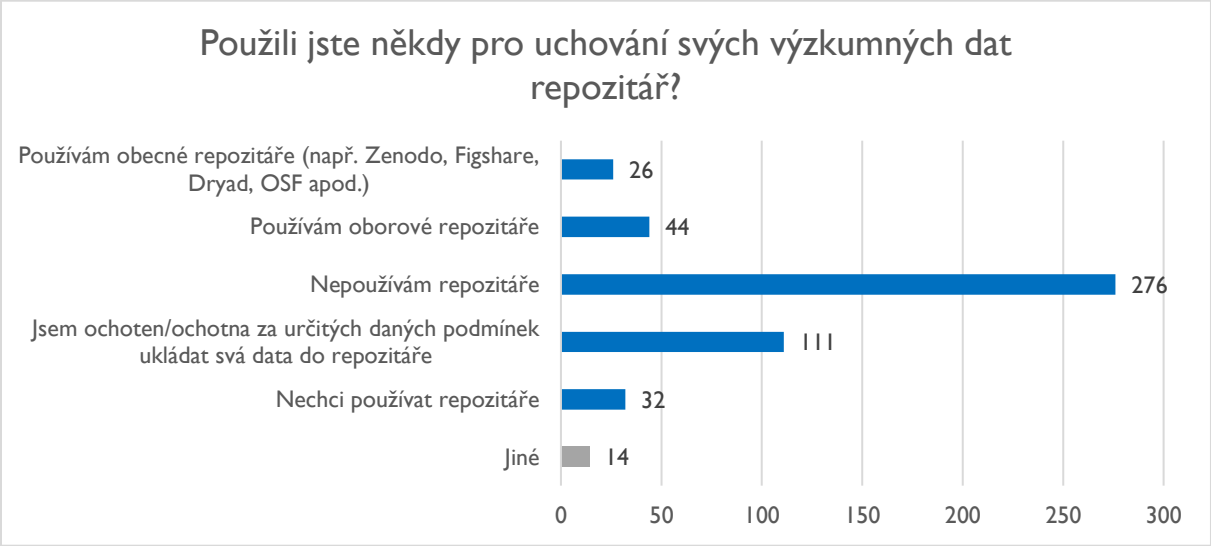


Graf 12: *Rozdělení respondentů podle toho, zdali zároveň s daty ukládají také dokumentaci.*

Repozitáře

Většina respondentů (68,1% 276/405) nepoužívá repozitáře, nicméně pouze relativně malá skupina (8% 32/405) uvedla, že repozitáře používat nechce. Poměrně velká skupina respondentů by byla ochotná za určitých podmínek svá data do repozitáře ukládat (27,4% 111/405).

Skupina respondentů, která používá oborové repozitáře, je větší 10,9% 44/405 než skupina používající obecné repozitáře (6,4% 26/405). Ke kontextu této otázky připomeňme, že oborové repozitáře jsou z hlediska implementace FAIR principů považovány za optimální cestu pro zveřejňovaní dat; obecné repozitáře jsou vhodné v situaci kdy vhodný oborový repozitář není k dispozici.

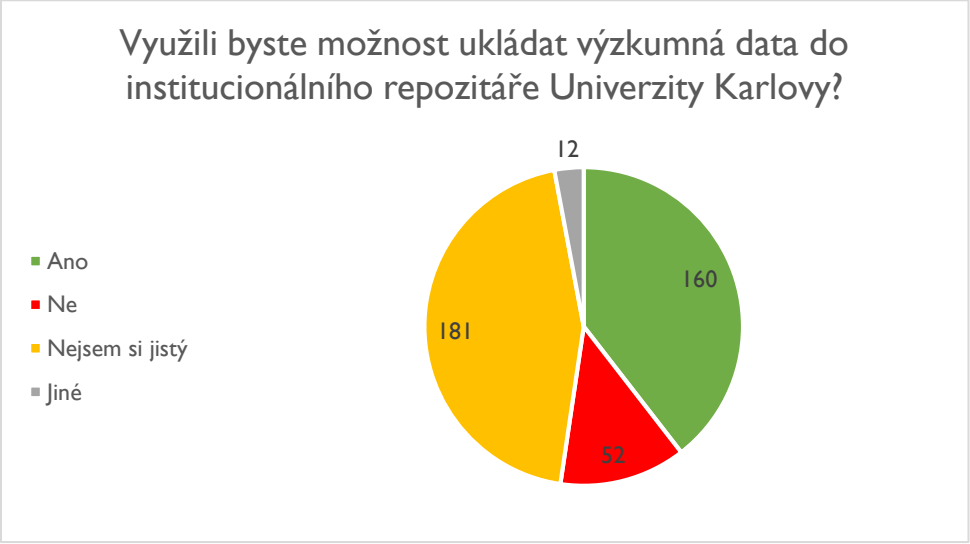


Graf 13: *Postoj respondentů k používání datových repozitářů.*

Na otázku týkající se možnosti využívání institucionálního datového repozitáře UK odpovídali respondenti spíše kladně:

- ano 39,5% (160/405),
- ne 12,8% (52/405)

Největší skupina respondentů (44,7% 181/405) si není jistá.



Graf 14: *Ochota respondentů využívat institucionální repozitář Univerzity Karlovy, pokud by tato možnost existovala.*

Využívání repozitáře se respondenti v zásadě nebrání, roli by zřejmě hrála konkrétní implementace a nabízené služby – repozitář musí vědcům přinášet jasnou přidanou hodnotu a ve finále i usnadnění jejich práce. Neochota používat institucionální repozitář nemusí také nutně znamenat neochotu používat repozitáře obecně, protože část respondentů může používat plně vyhovující oborové repozitáře (viz předchozí otázka „Použili jste někdy pro uchování svých výzkumných dat repozitář?“).

Slovní komentáře

V závěrečných slovních komentářích (celkem 41 komentářů) k této sekci respondenti zmiňovali preference možnosti využití oborových repozitářů (např. ČSDA pro sociálněvědní data nebo LINDAT/CLARIAH-CZ pro jazykovědná data) oproti institucionálnímu repozitáři. Uchovávání a zpřístupňování dat prostřednictvím oborových repozitářů je v současné době obecně považováno za dobrou praxi a je také v souladu s doporučením Centra pro podporu open science pro sdílení dat.[16] Je však třeba mít na paměti, že ne každý obor má k dispozici oborový datový repozitář. Dále respondenti zmiňovali otázku dat vhodných pro dlouhodobé uchovávání: někdy bývá obtížné odhadnout, která data budou v dlouhodobém horizontu užitečná, jiná data naopak rychle zastarávají a není tedy třeba řešit jejich dlouhodobé uchování. V této souvislosti by respondenti ocenili například metodiku, kde a za jakých podmínek data uchovávat nebo příklady dobré praxe, aby výzkumník věděl, jak a která data dlouhodobě uchovávat. Někteří respondenti otázku dlouhodobého uchovávání dat spojovali také s jejich sdílením a vyjadřovali obavy, že ne všechna data je možné sdílet. Jak již bylo zmíněno výše, sdílení dat by se obecně mělo řídit principem „as open as possible, as closed as necessary“, tudíž v odůvodněných případech je možné data nezveřejňovat, avšak i v takových případech může být vhodné jejich dlouhodobé uchování. Někteří respondenti také upozornili na potřebu uchovat s daty zároveň jejich dokumentaci a s tím související zvýšenou zátěž výzkumných pracovníků.

3.3.3 Oblast 3: Sdílení výzkumných dat

Sdílení výzkumných dat může být pro mnohé kontroverzním tématem. Na jedné straně vzniká v posledních letech poptávka po co největší transparentnosti, otevřenosti a opětovné využitelnosti vědeckých výsledků, na straně druhé pak stojí obavy o zajištění bezpečnosti chráněných dat, kontrolu duševního vlastnictví a často i strach o „ukradení“ výzkumu. Jedním z důležitých faktorů je i způsob hodnocení vědy, který zatím tvorbu, sdílení a opětovné využívání dat příliš nepodporuje.

Jak již bylo zmíněno v Kapitole 2.3, v evropském prostoru se v současné době prosazují 2 základní pravidla týkající se přístupnosti výzkumných dat:

- pravidlo „as open as possible, as closed as necessary“
- FAIR principy – nalezitelnost, přístupnost, interoperabilita a opětovná využitelnost

Pravidlo „As open as possible, as closed as necessary“ umožňuje v opodstatněných případech omezit přístup k vytvářeným datům. Podobně funguje i „přístupnost“ z FAIR principů – autor musí určit, za jakých podmínek budou data přístupná a způsob, jak se k nim může potenciální zájemce dostat. Ani v tomto případě to ale neznamená, že každý musí mít nárok na jakákoliv data.

[16] https://openscience.cuni.cz/OSCI-67.html

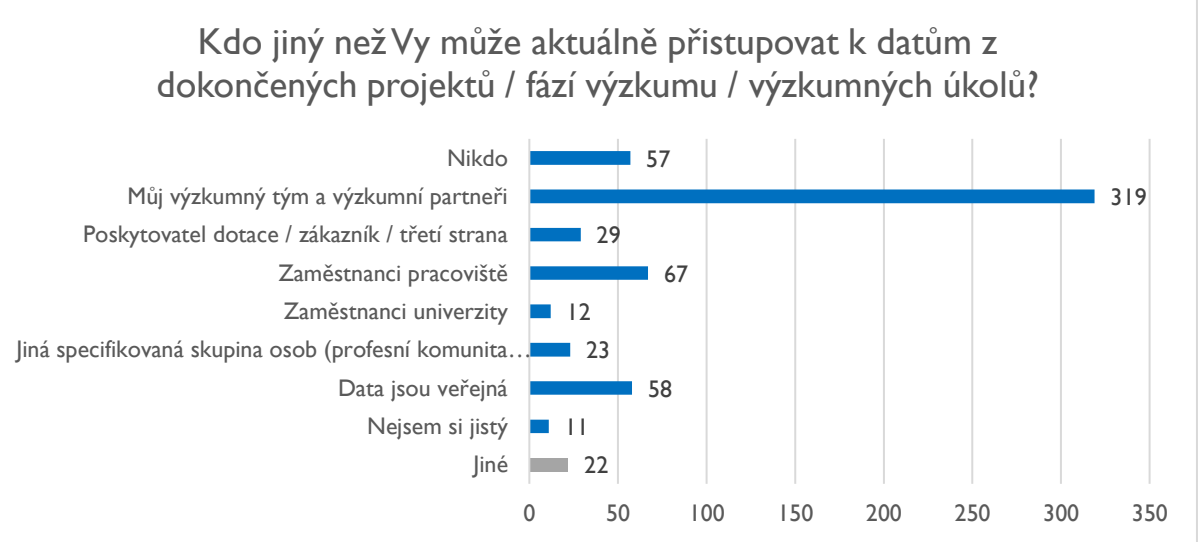
Sdílení dat je tak sice obecným cílem, ke kterému by vědci měli směřovat, nemusí ale jít o sdílení za každou cenu. Instituce by měla podporovat možnost sdílení dat – např. prostřednictvím institucionálního datového repozitáře, nicméně jde jen o jeden z aspektů moderní správy výzkumných dat.

S kým respondenti sdílí data

Většina respondentů (78,8% 319/405) sdílí data z dokončených projektů pouze v rámci týmu a s výzkumnými partnery. V rámci pracoviště sdílí data 16,5% (67/405) a v rámci univerzity 3% respondentů (12/405).

Data zveřejňuje 14,3% (58/405) respondentů, naopak nikdo k nim nemá přístup u 14,1% (57/405) respondentů.

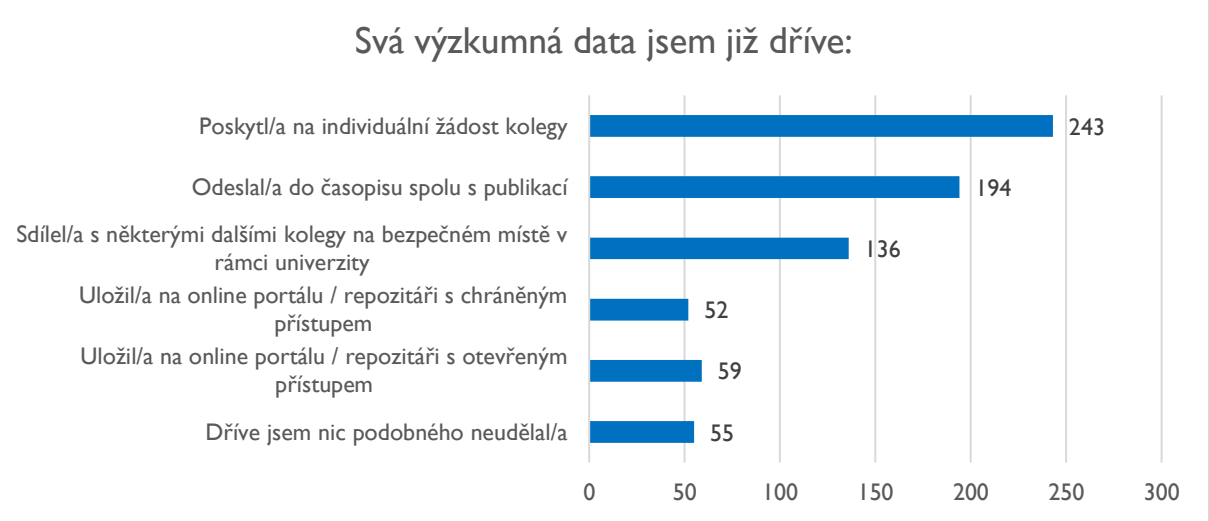
Sdílení dat veřejně není zatím obecně běžné, nejčastěji mají k datům i po skončení výzkumu přístup zejména členové samotných výzkumných týmů.



Graf 15: *Skupiny uživatelů, kteří mohou přistupovat k výzkumným datům z dokončených projektů.*

Ačkoli se na základě předchozích odpovědí může zdát, že výzkumná data často neopustí výzkumnou skupinu, další část dotazníku ukazuje, že tomu tak nutně nemusí být. Nejprve jsme zjišťovali, zdali respondenti již nějakou zkušenost se sdílením dat mají. Více než polovina respondentů (60% 243/405) uvedla, že svá data poskytla na individuální žádost kolegy a téměř polovina (47,9% 194/405) odeslala data do časopisu společně s publikací.

Zkušenosti s ukládáním dat do repozitáře má více než 10% respondentů a to jak s otevřeným (14,6% 59/405), tak omezeným (12,8% 52/405) přístupem.

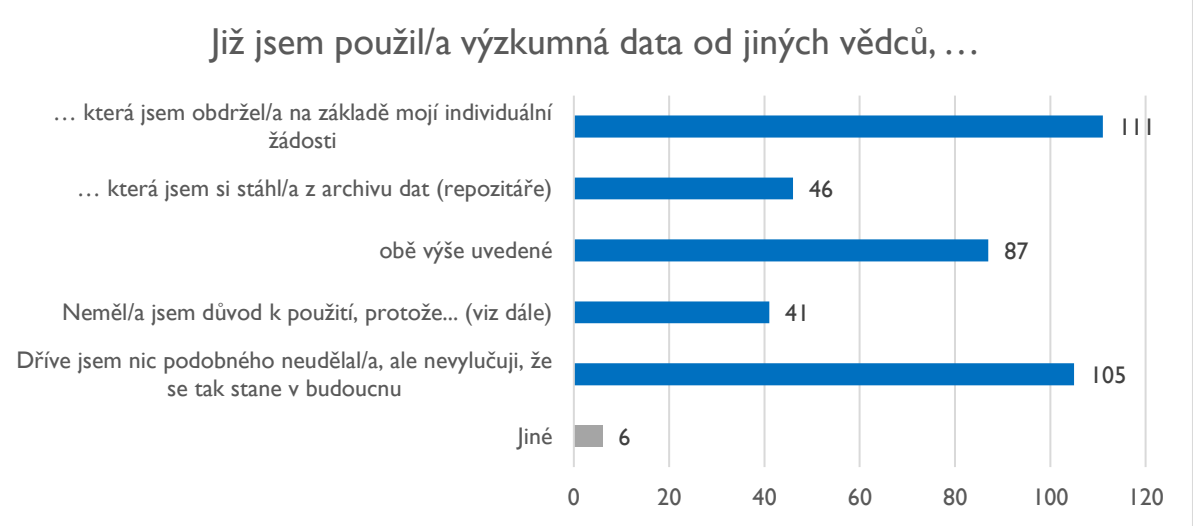


Graf 16: *Zkušenosti výzkumných pracovníků se sdílením dat.*

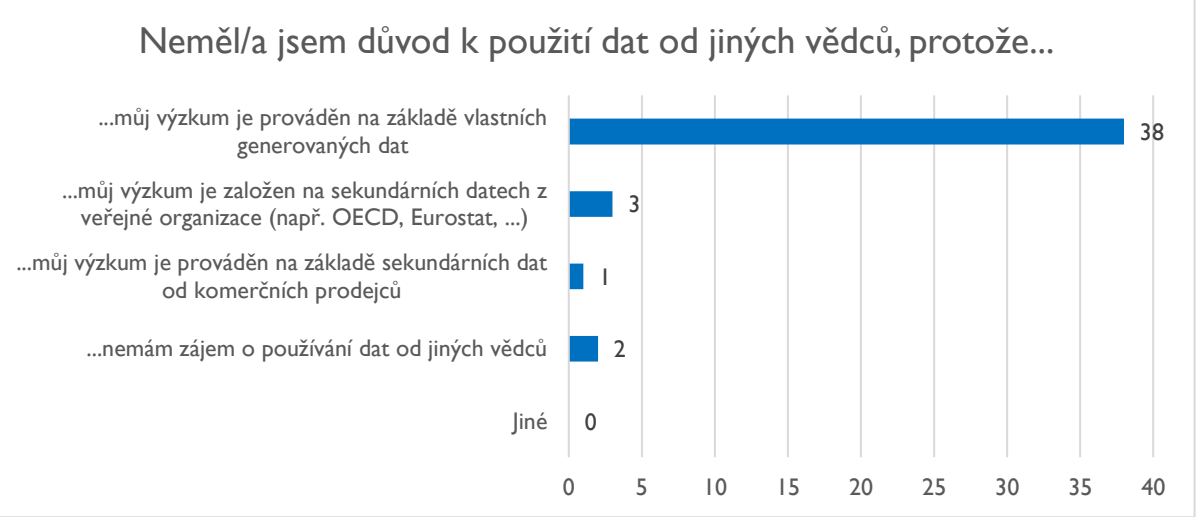
V následující otázce jsme pak zjišťovali, jestli mají respondenti zkušenost se sdílením výzkumných dat také ze strany „příjemce" dat a ukázalo se, že téměř polovina respondentů již použila data, která byla získaná na základě individuální žádosti (48,9% 111+87/405). O něco méně respondentů použilo data stažená z repozitáře (32,8% 46+87/405)

Čtvrtina respondentů (25,9% 105/405) nevylučuje, že data od jiných vědců použije v budoucnu, i když to zatím neudělala.

Nejčastěji uváděným důvodem pro nepoužívání dat jiných vědců je provádění výzkumu na základě vlastních generovaných dat (92,7% 38/41).



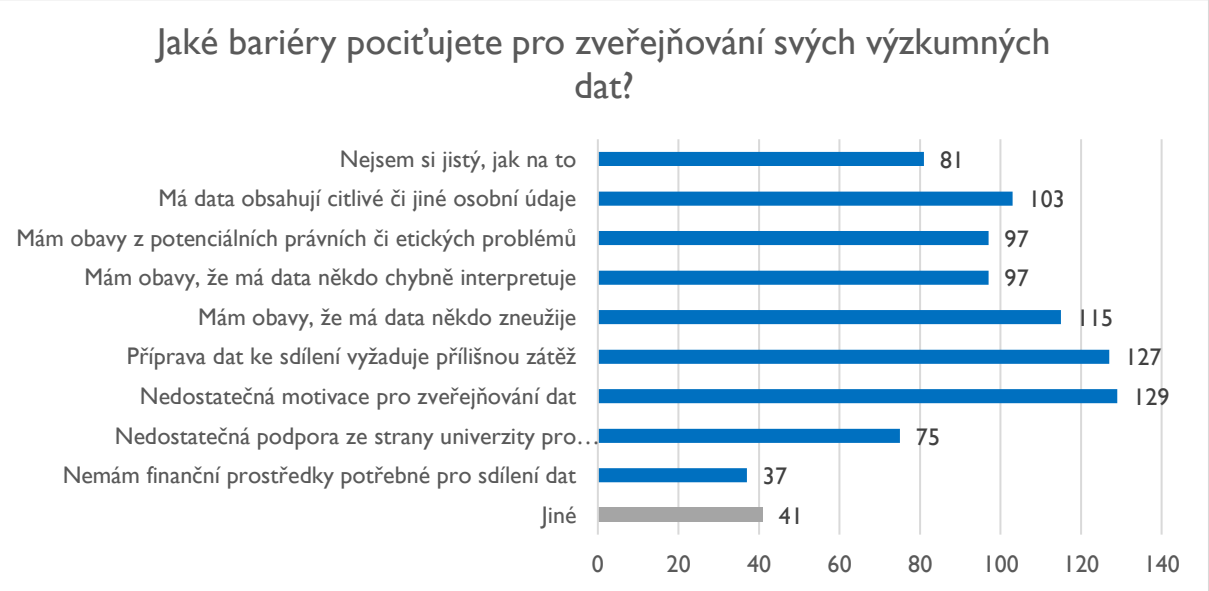
Graf 17: *Zkušenosti výzkumných pracovníků s využíváním sdílených dat.*



Graf 18: *Důvody proč respondenti nevyužívají data od jiných vědců.*

Bariéry a výhody sdílení dat

Jak bylo zmíněno již na začátku podkapitoly o sdílení výzkumných dat, jedná se o oblast, kolem které panují různé obavy. Mezi nejčastěji uváděnými bariérami se objevovala nedostatečná motivace pro zveřejňování dat (31,9% 129/405), obavy ze zvýšení zátěže vědeckých pracovníků (31,4% 127/405) a také obavy, že data mohou být zneužita (28,4% 115/405) či mylně interpretována (24% 97/405). Za zmínku také stojí, že ve slovním doplnění k variantě „jiné" se často objevovaly komentáře, že respondenti žádné bariéry nepociťují. Další slovní doplnění zmiňovala například, že sdílení dat není pro jejich výzkum relevantní či nejistota v oblasti správného nastavení licenčních podmínek.



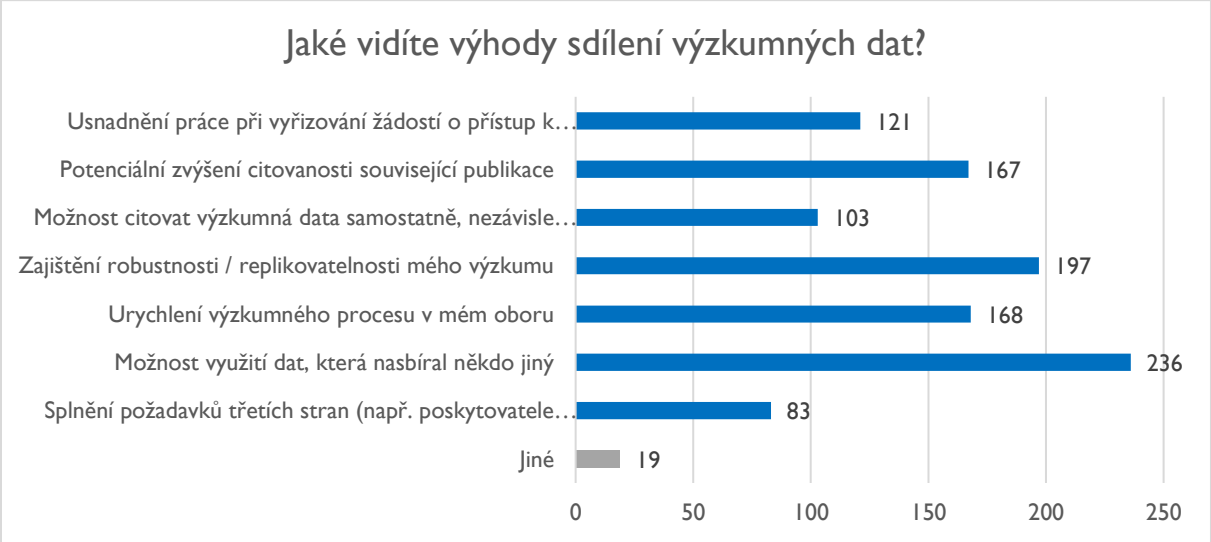
Graf 19: *Bariéry, které respondenti pociťují při sdílení svých výzkumných dat.*

Co se týče výhod sdílení výzkumných dat, nadpoloviční většina (58,2% 236/405) respondentů považuje za největší výhodu sdílení výzkumných dat možnost využít data, která nasbíral někdo jiný (srovnejte s 244 respondenty, kteří již využili data, která obdrželi na základě žádosti a/nebo je získali z otevřeného repozitáře; srovnejte s 243 respondenty, kteří již poskytli svá data na vyžádání).

Z hlediska vlastního výzkumu vidí největší množství respondentů přínosy sdílení k zajištění robustnosti / replikovatelnosti svého výzkumu (48,6% 197/405), potenciální zvýšení citovanosti související publikace (41,2% 167/405) a usnadnění práce při vyřizování žádostí o přístup k datům (29,8% 121/405).

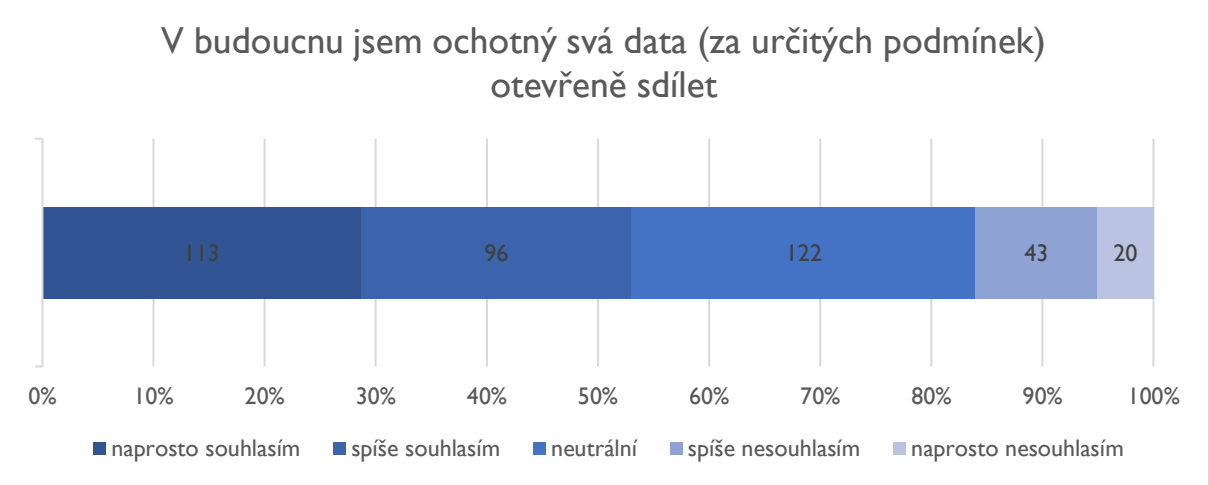
Nezanedbatelné je také množství respondentů, které za výhodu sdílení výzkumných dat považuje urychlení výzkumného procesu v jejich oboru – 41,5% (168/405).

Pouze jednotky respondentů ve slovním doplnění uvedly, že nevidí žádné výhody sdílení, a naopak jiní zmiňovali, že sdílení dat je službou veřejnosti a patří k ideálům vědy.



Graf 20: *Pociťované výhody sdílení výzkumných dat.*

Obecně se ke sdílení svých dat staví pozitivně 53% (113+96/394) a negativně jen 16% (43+20 / 394) respondentů. Neutrálně se vyjádřila necelá třetina respondentů (30,1%).



Graf 21: *Postoj respondentů ke sdílení svých výzkumných dat.*

Požadavky třetích stran

V úvodu studie je uvedeno, že sdílení dat začínají vyžadovat akademičtí vydavatelé a někteří, zatím především zahraniční, poskytovatelé financí. Toto je částečně reflektováno v odpovědích v dotazníku, kde se značná část respondentů již setkala s externími požadavky na sdílení dat, a to buď ze strany vydavatele / časopisu (27,4% 111/405) nebo ze strany poskytovatele financí (8,9% 36/405). Většina respondentů se však s externím požadavkem zatím ještě nesetkala, což může být částečně dáno také tím, že mezi českými poskytovateli financí se zatím nejedná o frekventovaný požadavek.

Dá se očekávat, že tato čísla v budoucnu porostou v souvislosti s rostoucím počtem časopisů, které datové politiky zavádí (viz report STM 2021) a také v důsledku implementace Směrnice Evropského parlamentu a Rady (EU) 2019/1024, která bude pravděpodobně klást na české poskytovatele financí a potažmo na příjemce podpory nové nároky v oblasti správy a sdílení výzkumných dat.

Graf 22: *Zkušenost respondentů s požadavky třetích stran ohledně sdílení výzkumných dat.*

Zkušenosti respondentů se sdílením výzkumných dat

V rámci snahy lépe pochopit obavy, které se ke sdílení výzkumných dat vážou, dostali respondenti prostor připojit slovní komentář, ve kterém mohli sdílet své **pozitivní** či naopak **negativní zkušenosti se sdílením dat**. Respondenti připojili celkem 89 komentářů k negativním zkušenostem (z nichž 52 uvádí, že žádnou negativní zkušenost nemají) a 91 komentářů k pozitivním zkušenostem (z nichž 34 uvedlo, že pozitivní zkušenost nemají).

Ačkoli se mezi pociťovanými bariérami často objevovala obava ze zneužití zveřejněných dat (viz otázka výše), mezi sdílenými negativními zkušenostmi se příklady zneužití dat (např. zcizení dat recenzentem a jejich následné publikování v cizojazyčném časopise nebo „krádež“ nápadu) objevovaly jen zřídka. Častěji se pak objevovaly negativní zkušenosti v rámci výzkumného týmu či katedry. Mezi nejčastěji zmiňované negativní zkušenosti však jednoznačně patřily negativní zkušenosti ze strany „příjemce“ sdílených dat, typicky pak neochota kolegů data poskytnout nebo zveřejněná / poskytnutá data ve špatné kvalitě.

Mezi pozitivními zkušenostmi se objevovaly pozitivní zkušenosti s využitím dat z veřejných databází či poskytnutých jinými výzkumníky, nebo naopak pozitivní zkušenosti sdílením vlastních dat, která mohli následně využít studenti nebo jiní výzkumníci. Další slovní komentáře se často týkaly výhod, které byly zmíněné v předchozí otázce, např. navázání nové spolupráce díky sdílení dat, urychlení práce a komunikace, nebo možnost odhalit chyby v práci s daty (s odkazem na projekt Retraction Watch[17]).

European Open Science Cloud (EOSC)

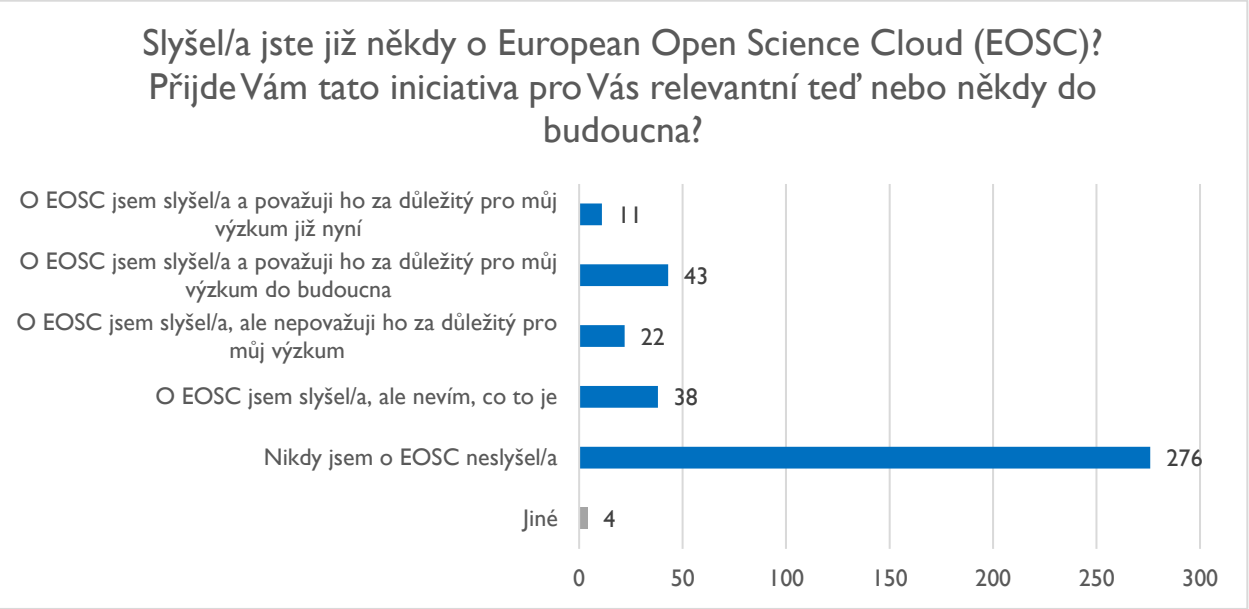
V České republice se v současné době připravuje podpora a implementace EOSC[18], což je evropské prostředí pro sdílení a zpracování výzkumných dat. V ČR se rozvoji Národní datové infrastruktury a implementaci EOSC věnuje infrastruktura e-INFRA CZ ve spolupráci s výzkumnými institucemi a dalšími zúčastněnými stranami[19]. Vzhledem k potenciálním možnostem využití této platformy v rámci evropského vědeckého prostředí a také v souvislosti s chystanými aktivitami v této oblasti na národní úrovni jsme se respondentů ptali, zda o této iniciativě slyšeli a zda jim přijde pro jejich práci relevantní.[20]

Většina respondentů (68,1% 276/405) se dosud s termínem European Open Science Cloud / EOSC nesetkala, menší část (28,1% 11+43+22+38/405) tento termín do nějaké míry zná.

Z respondentů, kteří EOSC znají a mají nějaký názor (11+43+22=76) jej

- 14,4% (11/76) považuje za důležitý už nyní,
- 56,6% (43/76) ho považuje za důležitý do budoucna,
- 28,9% (22/76) ho nepovažuje za důležitý.

[17] https://retractionwatch.com/
[18] https://eosc-portal.eu/about/eosc
[19] https://www.e-infra.cz/eosc
[20] Aktivity národní iniciativy zahrnující širší vědeckou komunity byly zahájeny až na podzim 2021, tedy po ukončení sběru dat pro tento dotazník.



Graf 23: *Povědomí respondentů o European Open Science Cloud (EOSC).*

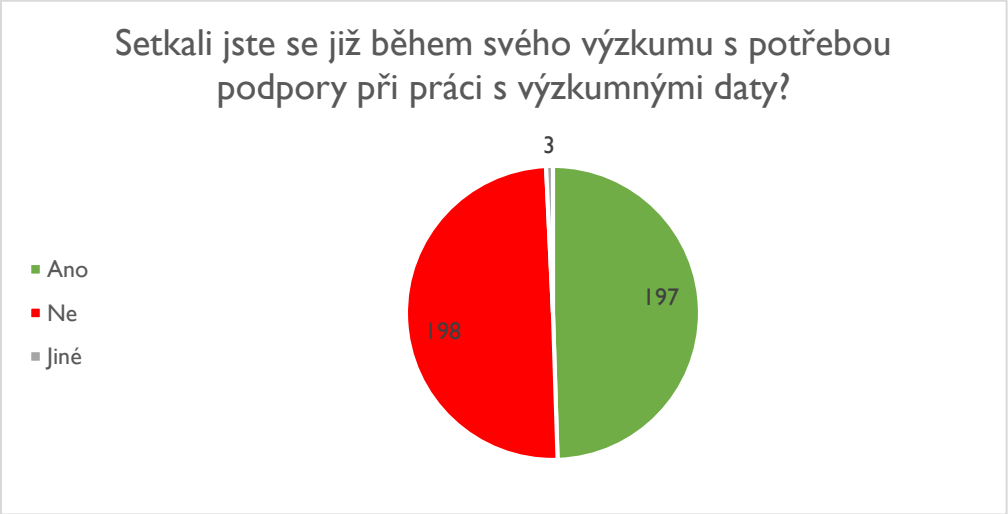
Slovní komentáře

V závěrečných volných komentářích k sekci sdílení (celkem 19 komentářů) zmiňovali respondenti důležitost tématu sdílení dat, ať už z osobního přesvědčení či proto, že se setkávají s požadavky časopisů a univerzita nenabízí dostatečnou podporu pro řešení těchto požadavků. Zároveň však také akcentovali nutnost zohlednění pravidel ochrany osobních údajů, otázku autorství a duševních práv.

3.3.4 Oblast 4: Podpora správy výzkumných dat

Přestože je infrastruktura zřejmým stěžejním bodem práce s výzkumnými daty, v žádném případě nelze opominout ani přímou vědeckou podporu ("research support"). Správná práce s daty vyžaduje nové dovednosti, nové postupy a ani ten nejsofistikovanější nástroj se nebude obsluhovat sám. Úkolem a cílem instituce by mělo být usnadnit vědcům adaptaci na nové postupy a nástroje, tak aby se mohli v co největší možné míře věnovat své hlavní výzkumné práci.

V otázce potřeby podpory při práci s výzkumnými daty se respondenti rozdělili téměř přesně na poloviny: 49,5% (198/400) už během svého výzkumu narazila na potřebu podpory při práci s výzkumnými daty, zatímco podobně velká část respondentů (49,25% 197/400) se s takovou potřebou ještě nesetkala.



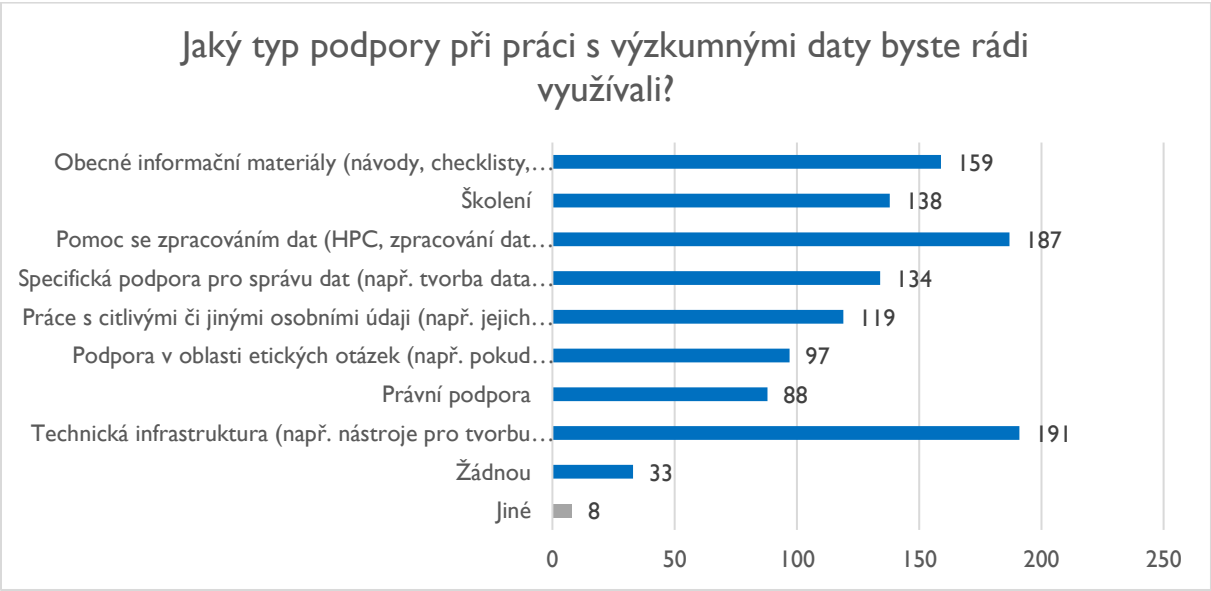
Graf 24: *Dosavadní potřeba podpory při práci s výzkumnými daty.*

Nejvíce respondentů (48,6% 197/405) by rádo využívalo technickou infrastrukturu a pomoc se zpracováním dat (46,2% 187/405).

Z metodické podpory by byl mezi respondenty největší zájem o obecné informační materiály (39,3% 159/405), školení (34,1% 138/405) a specifickou podporu (33,1% 134/405).

Specifické otázky práce s citlivými či jinými osobními údaji zajímají 47,2% (119/405), pomoc s etickými otázkami by rádo využilo 24% (97/405) a s právními otázkami 21,7% (88/405) respondentů.

Žádnou podporu by nevyužilo jen 8,1% (33/405) respondentů.



Graf 25: *Typ podpory, který by respondenti při práci s daty využili.*

Slovní komentáře

Volné komentáře na konci sekce (celkem 18 komentářů) zmínily mimo jiné potřebu podpory v oblasti ochrany osobních údajů, anonymizace dat a zájem o podporu open science jako celku (tedy nejen zpřístupňování dat). Dále pak nutnost vzdělávání technicko-administrativních pracovníků, kteří by pak mohli výzkumníkům při správě dat pomoci.

V závěrečném oddílu dotazníku měli respondenti možnost zanechat obecný komentář k celému dotazníku či problematice práce s daty, která nebyla dostatečně pokryta v předchozích částech (celkem 26 komentářů). Respondenti zde vyjádřili obavy ze zvýšení zátěže, která se s prací s daty pojí, a zmínili, že by bylo užitečné zavést na univerzitě pozici tzv. data stewardů, kteří by výzkumníkům mohli pomoci se správou dat a také s jejich analýzou, vyhodnocováním nebo vizualizací. Dále pak někteří uvedli, že by bylo vhodné zaměřit se také na studenty bakalářského a magisterského studia, kteří mohou v rámci závěrečné práce připravit zajímavou a hodnotnou datovou sadu. V neposlední řadě pak připomínali, že je třeba, aby podpora a metodické materiály byly dostupné také v anglickém jazyce.

3.4 Závěrečné shrnutí

3.4.1 Respondenti

Nejsou rovnoměrně zastoupeni respondenti ze všech fakult, do určité míry to ovšem může být ovlivněno tím, že různé obory pracují s daty do různé míry. Nejvíce zastoupené obory můžeme řadit mezi obory, kde se s daty pracuje poměrně hodně, z čehož může plynout větší zájem o problematiku správy dat, a tudíž i větší zájem o vyplnění dotazníku.

Nejvíce zastoupeny jsou medicinské a přírodovědné obory, znatelně méně respondentů uvedlo obory z humanitní a společenskovědní oblasti. Celkově je ovšem skupina nejvíce zastoupených oborů poměrně pestrá.

3.4.2 Práce s výzkumnými daty během výzkumu

V oblasti práce s výzkumnými daty během výzkumu existuje prostor pro zlepšení. Z hlediska budování infrastruktury je důležitá otázka na množství ročně vyprodukovaných dat. U více než poloviny respondentů množství dat vytvořené za rok nepřesáhne 20 GB, zhruba desetina pak vygeneruje významně velké množství dat – více než 1 TB ročně.

Významné téma tvoří chráněná data. Jde zejména o data obsahující osobní údaje, ale i další typy chráněných dat. Práce s nimi přináší větší rizika vzniku škod a vyžaduje jak dostupnou zabezpečenou infrastrukturu, tak všeobecně známé a dodržované metodické postupy.

Z hlediska naplnění FAIR principů je důležité, že velké množství respondentů nějakým způsobem data dokumentuje. Na druhou stranu je zde velký prostor pro zlepšení s ohledem na využívání oborových standardů.

Při práci s výzkumnými daty většina respondentů spoléhá v první řadě na infrastrukturu pracoviště. Poměrně časté je i využívání služeb dalších stran (např. Google, Dropbox). Obdobné služby jsou nicméně na univerzitě k dispozici i od infrastruktury CESNET, je tedy otázka, zda je spektrum služeb

dostatečné, nebo je třeba je jen více propagovat. Problematické je zejména využívání bezplatných služeb, které nejsou vhodně smluvně ošetřeny. Do stejné kategorie lze zařadit i sdílení dat prostřednictvím e-mailů, což může být v některých případech velmi nevhodné.

K zálohování se často používají externí disky, což je sice relativně jednoduché řešení, ale přináší také značná rizika (fyzická ztráta disku, možnost poškození, technologické zastarávání).

Obecně řečeno jsou lokální řešení v zásadě v pořádku, i u nich by ale měla být nastavena a dodržována metodická úroveň. Nezodpovězenou otázkou je, do jaké míry mají všichni pracovníci přístup k odpovídajícím službám – z odpovědí v dotazníku je vidět, že někteří z nich odpovídající infrastrukturu využívají, odpovědi velkého množství respondentů však zároveň naznačují, že volí jiná řešení.

3.4.3 Dlouhodobé uchovávání výzkumných dat

Respondenti si uvědomují důležitost dlouhodobého uchovávání dat. Technické řešení ovšem není dostatečné – data jsou nejčastěji uchovávána na vlastních médiích a ve vlastních počítačích. To klade na výzkumné pracovníky další nároky a v krajních případech může vést až k neočekávaným problémům – například pokud by byli požádáni o kopie dat ke svým starým článkům za účelem kontroly a nejsou je schopni v takto udržovaném systému dohledat. V této souvislosti je také důležité stanovit, zda a jakým způsobem budou data uchovávána v případě odchodu výzkumníka z Univerzity Karlovy, aby se zamezilo jejich znehodnocení či ztrátě.

V oblasti dlouhodobého uchování dat je důležité rozlišovat mezi uchováváním výsledných zpracovaných a publikovaných dat a uchováním surových, živých dat – vhodná by byla přímá podpora, např. datový kurátor (zejm. pro rozlišení co je a co není nutné dlouhodobě ukládat). Obzvláště u surových dat je také možné že dojde k významným nárokům na infrastrukturu.

Využití repozitářů pro dlouhodobé uchování není v současnosti běžné, nicméně poměrně velká část respondentů se k takové možnosti nestaví negativně. Dalo by se říci, že existuje řada potenciálních zájemců o ukládání dat do repozitáře Univerzity Karlovy.

Bylo by vhodné vytvořit základní univerzitní standard a odpovídající řešení pro dlouhodobé uchovávání dat – jednodušší by bylo navázat takové řešení na repozitáře pro publikovaná data, nicméně je nutné se touto problematikou zabývat i v kontextu surových/živých dat.

3.4.4 Sdílení výzkumných dat

V současnosti není sdílení dat mimo výzkumný tým běžnou praxí. Data bývají poskytována především na přímou žádost kolegy, nikoliv otevřeně; část respondentů nicméně už má zkušenost s posíláním dat do časopisu. Používání dat od jiných vědců je nicméně častější než sdílení a řada respondentů vidí potenciál ve využití cizích dat.

Řada respondentů vidí přínosy sdílení dat – nejčastěji jde o možnost využití dat, která nasbíral někdo jiný – důležité jsou ale i obecnější výhody, jako je například replikovatelnost a robustnost výzkumu a všeobecné urychlení výzkumu v oboru. Bariérami sdílení dat jsou především nedostatečná motivace a náročnost přípravy; obavy týkající se zneužití, ukradení či špatné interpretace dat jsou o něco méně časté. Většina respondentů nezavrhuje možnost otevřeného sdílení svých dat v budoucnu. Zatím se ale většina respondentů ani nesetkala s požadavkem na sdílení dat ze strany poskytovatele financí, nebo časopisu, v němž chtějí publikovat.

Oblast sdílení výzkumných dat je v tuto chvíli na pomezí – na jednu stranu si respondenti uvědomují přínosy tohoto sdílení a jsou ochotní využívat cizí data, na druhou stranu necítí dostatečnou motivaci k tomu, aby sami data sdíleli. Existuje celá řada bariér pro sdílení, s jejichž odstraněním může univerzita pomoci. Nejzákladnější formou této pomoci může být vznik metodické podpory a poskytování technických nástrojů pro tvorbu výzkumných dat splňujících FAIR principy.

Přímý tlak na sdílení dat ze strany poskytovatelů financí a časopisů se zatím příliš neprojevil, lze ale očekávat, že s průběhem času poroste.

3.4.5 Podpora správy výzkumných dat

Potřebu podpory pro práci s výzkumnými daty už pocítila polovina respondentů.

Největší poptávka by mezi respondenty byla po technické infrastruktuře a pomoci se zpracováním dat; poměrně velký zájem by ale byl i o více metodické otázky (obecná školení, nakládání s osobními údaji, příprava DMP). Na druhou stranu jen poměrně malá skupina respondentů uvedla, že by žádnou podporu v této oblasti nevyužila. Respondenti si uvědomují důležitost a komplikovanost celé problematiky správy výzkumných dat. Neměla by se opomíjet ani problematika zpracování dat, která nebyla tímto dotazníkem příliš podchycena.

4. Praxe zahraničních výzkumných univerzit

V Evropské unii se problematika správy výzkumných dat a open science obecně intenzivně řeší už několik let. Vzhledem k tomu, že Česká republika nepatří v tomto ohledu k nejaktivnějším zemím, poprosili jsme kolegy z několika významných evropských výzkumných univerzit z Holandska a Belgie o konzultaci za účelem sdílení zkušeností. Původně plánované zahraniční cesty nebylo možné kvůli pandemii covid-19 realizovat, díky velké vstřícnosti zahraničních kolegů je ale bylo možné nahradit virtuálními setkáními s pracovníky odpovědnými za plánování a realizaci institucionálních strategií správy dat a s členy týmů poskytujících přímou podporu vědcům.

4.1 Utrecht University

Holandská Univerzita v Utrechtu (Universiteit Utrecht, dále UU) je jednou z nejznámějších a nejaktivnějších výzkumných univerzit v oblasti prosazování principů Open Science. UU má 7 fakult a přes 7000 zaměstnanců a více než 35000 studentů.

4.1.1 Práce s daty

Základní pravidla pro práci s daty stanovuje dokument University Policy Framework for Research Data[21] z roku 2016, který definuje pojem "výzkumná data" a nastavuje pravidla a zodpovědnosti pro práci s nimi.

Součástí požadavků na kvalitní správu dat je v Utrechtu i příprava **data management plánů** (DMP). Univerzita má vlastní specifickou šablonu pro DMP a oficiálně podporovanou instalaci nástroje DMPonline.

Pro práci s daty v průběhu výzkumu nabízí UU systém Yoda[22]. Yoda je open source systém který umožňuje uživatelům z UU a spolupracujících institucí **bezpečně ukládat, sdílet, publikovat a uchovávat velké objemy dat během všech fází výzkumného projektu**. Základem systému je open source data management software iRODS[23], další části systému (například uživatelské rozhraní a nastavení specifických pravidel a workflow) byly vytvořeny na UU.

Systém Yoda je provozován na infrastruktuře UU, která se zároveň sama stará o jeho správu. Jedním z původních cílů bylo vytvoření úložiště pro pohodlnou a bezpečnou práci s citlivými daty, která bylo zároveň potřeba sdílet v rámci výzkumných týmů nejen v rámci univerzity, ale i napříč více institucemi. Systém zároveň umožňuje uložení zabezpečené kopie dat, nebo jejich publikování v souladu s principy FAIR.

Pro (finální) publikování výzkumných dat používá UU také systém Dataverse[24], hostovaný národní infrastrukturou DANS. Dataverse slouží jako místo pro ukládání datových sad, pro které neexistuje vhodný oborový repozitář, který by jinak měl být upřednostněn. DANS zajišťuje technickou stránku

[21] https://www.uu.nl/sites/default/files/university_policy_framework_for_research_data_utrecht_university_-_january_2016.pdf
[22] https://www.uu.nl/en/research/yoda/what-is-yoda
[23] https://irods.org/
[24] https://dataverse.nl/dataverse/UU

řešení, obsah samotný je ale spravovaný přímo specialisty z UU. Uživatelé tak mají k dispozici hned několik cest, pomocí kterých mohou svá data spravovat a případně publikovat přímo s podporou univerzity. Různé scénáře mohou být závislé například na velikosti a typu dat, které následně ovlivňují cenu uložení a dalších služeb.

4.1.2 Podpora správy dat

Podpora správy dat je v Utrechtu velmi dobře rozvinutá. UU poskytuje asistenci už v průběhu plánování projektu (tvorba DMP) a připravuje řadu kurzů pro průběžné vzdělávání vědeckých pracovníků. Základem systému podpory jsou RDM consultants, což jsou specialisté z různých oborů, kteří mají široké znalosti o různých aspektech správy dat a fungují jako první linie podpory, odpovídají na e-mailové dotazy, starají se o webové stránky a pořádají školení a workshopy. Dalším důležitým pilířem jsou Reseach engineers, kteří mohou výzkumníkům pomoci s techničtějšími aspekty zpracování a analýzy dat jako je například skriptování, či vývoj SW. Navíc ale také disponuje skupinou odborníků (Data managers), kteří mohou být „pronajati" výzkumným týmem, například v rámci projektu, a podílí se na nastavování workflow, formátování, čištění a verzování dat a zajišťují jejich uchovávání a publikaci.

4.2 Leiden University

Univerzita v Leidenu (Universiteit Leiden, LU) je výzkumnou univerzitou v Holandsku. Má 7 fakult, necelých 7 tisíc zaměstnanců a přes 28000 studentů.

4.2.1 Práce s daty

Univerzita má pravidla pro správu výzkumných dat už od roku 2016 (Research Data Management Regulations[25]). Politika se zaměřuje na různé fáze výzkumného procesu – přípravu práce s daty, práci s daty během projektu a po jeho skončení. Součástí pravidel je i vytváření data management planu a obecným cílem je tvorba dat splňujících principy FAIR. Data musí být uchovávána nejméně 10 let po skončení projektu. Počítá se s tím, že fakulty mohou definovat oborově specifičtější pravidla a postupy.

V přípravné fázi projektu existuje na LU povinnost konzultovat data management plan s odborníkem na instituci, což je podpořeno i ze strany národních poskytovatelů financí. LU nemá vlastní software na tvorbu DMP, nicméně uvažuje o nástroji, kde by byla problematika správy dat spojena ještě s dalšími významnými aspekty – ochranou osobních údajů a citlivých dat a s posouzením etické stránky projektu. LU má vlastní šablonu pro DMP, kterou je možné stáhnout z repozitáře Zenodo[26].

LU pracuje na implementaci vlastní infrastruktury pro práci s daty, která vychází ze systému Yoda na Univerzitě v Utrechtu a rovněž využívá software iRods. Kromě budování vlastní infrastruktury se snaží využít také kooperativních projektů na národní úrovni a některé fakulty používají služeb repozitáře Dataverse[27] na národní platformě.

[25] https://www.library.universiteitleiden.nl/binaries/content/assets/ul2ub/research--publish/research-data-management-regulations-leiden-university_def.pdf
[26] https://doi.org/10.5281/zenodo.4423065
[27] https://dataverse.nl/dataverse/leidenuniversity

4.2.2 Podpora správy dat

LU chce, podobně jako UU, zavádět nové pracovní pozice pro Data Stewards (RDM consultants na UU) a Research Engineers. Oproti jiným univerzitám budou tyto pozice vytvořeny přímo na fakultách, předpokládá se pak snazší navázání kontaktu s vědeckými pracovníky a lepší orientace v podporované problematice. Na druhou stranu se úroveň podpory na fakultách může lišit více než u centralizovaného řešení.

4.3 KU Leuven

Katolická univerzita v Lovani (Katholieke Universiteit Leuven, KUL) je belgická univerzita na které studuje více než 60000 studentů na 15 fakultách. Univerzita má více než 20000 zaměstnanců, z nichž je přes 7000 profesorů a výzkumníků.

4.3.1 Práce s daty

Také KUL má stanovenu **politiku správy výzkumných dat**. Ta definuje základy práce s daty v pěti bodech týkajících se:

1. ukládání a dokumentace dat
2. zajištění zpřístupnění metadat
3. podmínek mazání dat
4. pravidel pro uchování dat po dobu 10 let po skončení výzkumu
5. zohlednění všech dalších relevantních podmínek a legislativních norem

Zároveň také stanovuje základní odpovědnosti jak pro výzkumníky[28], tak pro instituci[29].

KUL má vlastní instanci nástroje DMPonline[30], s jehož pomocí mohou výzkumníci vytvořit vlastní data management plan, ať už podle šablony svého poskytovatele financí, nebo podle šablony KUL.

KUL nemá k dispozici pro systém pro správu dat podobného typu jako je Yoda na univerzitě v Utrechtu. Studentům a uživatelům z univerzity je k dispozici sdílený prostor na univerzitní infrastruktuře, služby Microsoft Suite a komerční služba Box Cloud Storage pro ukládání dat do 100GB velikosti.

Repozitář pro publikování výsledných datových sad nazvaný RDR [radar] se v současné době připravuje a bude spuštěn do plného provozu v roce 2022. Je postaven na systému Dataverse a provozován je přímo KUL. Cílem implementace je umožnit uživatelům z KUL publikovat a sdílet data v souladu s principy FAIR a tím usnadnit plnění požadavků belgických organizací financujících výzkum. Systém Dataverse je napojen i na lokální systém pro správu bibliografických záznamů LIRIAS, díky čemuž je možné spojovat související výzkumná data a publikace v obou systémech.

[28] https://www.kuleuven.be/rdm/en/rdm-policy/responsibilities-of-all-involved-in-research
[29] https://www.kuleuven.be/rdm/en/rdm-policy/responsibilities-of-the-university
[30] https://dmponline.kuleuven.be/

4.3.2 Podpora správy dat

Podpora pro oblast správy vědeckých dat je poskytována prostřednictvím Research Data Management Competence Centre[31], což je virtuální struktura, v níž se setkávají odborníci z IT oddělení, knihovny, oddělení pro výzkum a dalších). Klíčovými aktivitami této skupiny jsou

1. koordinace podpory a postupů týkajících se správy vědeckých dat
2. podpora s plněním podmínek poskytovatelů financí
3. popora dokumentace, archivace a publikovaní datových sad
4. technická podpora a infrastruktura

4.4 Ghent University

Univerzita v Gentu (Universiteit Gent, UGent) je významnou výzkumnou univerzitou v Belgii s více než 49000 studenty a více než 15000 zaměstnanci na 11 fakultách.

4.4.1 Práce s daty

Správa výzkumných dat se na UGent řídí rámcovou politikou správy výzkumných dat vydanou v roce 2016[32]. Tato politika uznává data jako důležitý výstup vědecké činnosti, který by měl být dostupný pro ověření výsledků a další využití, při zachování odůvodněné možnosti omezit přístup k datům v případě kdy je to nutné. Správa vědeckých dat je sdílenou odpovědností instituce (která musí vytvořit odpovídající podmínky a služby) a vědeckých pracovníků (kteří musí s vědeckými daty nakládat odpovídajícím způsobem).

Vědečtí pracovníci musí

1. dodržovat relevantní zákonné a další vnější požadavky na správu dat (např. GDPR, požadavky financujících organizací, etická pravidla a smluvní závazky vůči třetím stranám),
2. plánovat správu dat, v ideálním případě vytvořit DMP,
3. odpovídajícím způsobem dokumentovat výzkumná data,
4. bezpečně ukládat a zpracovávat data během výzkumného procesu,
5. Uchovávat relevantní data a související dokumentaci po dobu nejméně 5 let po skončení výzkumného projektu na publikaci výsledků (cokoliv nastane později),
6. poskytnout přístup k datům v rámci UGent za účelem kontroly integrity a sdílet je šířeji když je to možné, ale v souladu s případnými nutnými omezeními.

Plánování správy dat je povinné pro všechny doktorandy od akademického roku 2020/2021. UGent využívá společné konsorciální instance nástroje DMPonline[33], na jejímž vývoji se i podílí.

[31] https://www.kuleuven.be/rdm/en/RDM-CC
[32] https://www.ugent.be/en/research/datamanagement/policies/rdm-policy.pdf
[33] https://dmponline.be/

Pro práci s živými daty jsou na UGent k dispozici sdílené diskové prostory[34] poskytované a spravované IT oddělením univerzity. Je možné využívat sdílený prostor v základní (2TB na osobu, jednodušší správa přístupu), nebo rozšířené verzi (pokročilejší správa přístupů, pro komplexnější projekty).

UGent nemá vlastní repozitář pro publikování výzkumných dat – využívá existujících repozitářů. Uvažuje se ale o vytvoření systému „Data Vault" pro data, která nemohou být uložena do externího repozitáře – může jít o citlivá, příliš velká, či jiná data pro tyto typy repozitářů nevhodná. Dalším zajímavým prvkem infrastruktury by se měl v budoucnu stát systém, do nějž by se sbíraly informace o vytvořených datových sadách, a to i těch, které se nacházejí v externích repozitářích.

4.4.2 Podpora správy dat

V oblasti podpory lze vyzdvihnout vytvoření skupiny data stewardů na úrovni centrální knihovny. Stewardi jsou rozděleni podle oborů, resp. oborových clusterů (do každého clusteru spadá typicky několik fakult). Úlohou stewardů je pomáhat se správou výzkumných dat v rámci celého výzkumného cyklu. Nejde však jejich o přímé přiřazení ke konkrétním výzkumným týmům, jsou spíše primární kontaktní a informační bod pro problematiku správy dat pro své fakulty.

4.5 Shrnutí

Všechny výše uvedené univerzity mají už několik let stanovenu politiku pro práci s výzkumnými daty. Každá má také vlastní šablonu pro plán správy dat (DMP) a nabízí v nějaké formě i systém (všude DMPonline), kde lze tyto šablony vyplnit.

Liší se do určité míry tím, jakou nabízejí infrastrukturu – část infrastruktury je vytvářena až na základě přijaté politiky správy dat, která typicky určuje odpovědnosti jak na straně vědce, tak na straně instituce. Tam, kde je to možné, se infrastruktura vytváří na národní úrovni.

Pro živá data mají nebo implementují speciální řešení Utrecht a Leiden, Leuven a Gent pracují více v rámci tradiční infrastruktury. Všechny instituce vnímají jako specifickou problematiku ochranu citlivých dat a objemných datových sad.

Většina univerzit (kromě Gentu) má nebo připravuje nějakou formu institucionálního repozitáře pro publikování výsledných datových sad. Ve všech zmiňovaných případech se používá systém Dataverse, který plní základní požadavky z pohledu FAIR principů.

Podpora se zaměřuje jak na vzdělávání a osvětu, tak na konkrétní pomoc se správou dat – mezi nejtypičtější činnosti patří asistence s přípravou plánů správy dat. Významně se rozvíjí role a význam pracovníků specializovaných na práci s daty v konkrétní tematické oblasti. Nejdále je v tomto ohledu UU, kde je možné najmout si odborníka přímo pro svůj projekt.

[34] https://www.ugent.be/en/research/datamanagement/during-research/storage.htm

5. Závěr

Způsob práce s výzkumnými daty se v posledních letech poměrně dynamicky vyvíjí. Zřejmě nejvýznamnějším krokem bylo definování FAIR principů a jejich následné všeobecné přijetí jak vědeckými komunitami, tak poskytovateli financí v čele s Evropskou komisí. Díky tomuto posunu v definování dobré praxe získávají vědci jasně definovaný cíl, k němuž by měla správa výzkumných dat směřovat: data mají být nalezitelná, přístupná, interoperabilní a opětovně využitelná. Data v této formě by pak měla být zveřejňována v maximální možné míře, za předpokladu, že pro jejich nezveřejnění není legitimní důvod.

Proces tvorby dat a jejich správy v průběhu výzkumu je na samotných vědeckých pracovnících – nelze ji přenést mimo jejich zodpovědnost. Na druhé straně je zřejmé, že jde o proces, který vytváří značné nároky zejména v oblasti technické infrastruktury, osvojení si (v některých případech nových) digitálních dovedností a v neposlední řadě vyžadující přímou podporu ze strany vlastní instituce.

Vzhledem k tomu, že tento proces už několik let probíhá, můžeme sledovat, jakým způsobem tuto problematiku řeší zahraniční univerzity. Na příkladech uvedených v této studii vidíme, že pro instituce je důležité určit vlastní směřování v oblasti správy dat. Za tímto účelem bývají přijímány institucionální politiky správy výzkumných dat, které definují jak základní parametry k řešení celé problematiky, tak především jasně stanovují odpovědnosti jak pro vědecké pracovníky, tak pro instituce samotné.

Vzhledem ke specifikům různých oborů se všeobecně považuje za optimální co nejspecifičtější řešení. Obecnější pravidla na vyšší úrovni (například národní či institucionální) však pořád hrají velmi významnou roli: I když nejsou schopny postihnout problematiku správy dat do detailu, nastavují základní rámec, z nějž mohou oborové politiky vycházet, a jehož detaily mohou specifikovat podrobněji. Tato pravidla potom mohou být aplikována v rámci jednoho oboru napříč institucemi a je proto vhodné s jejich potenciální existencí počítat.

Prostředí správy dat na Univerzitě Karlově je oproti osloveným zahraničním univerzitám v jiné fázi vývoje. Univerzita Karlova v tuto chvíli nemá vlastní politiku směřující k tvorbě FAIR dat. To samozřejmě rozhodně neznamená, že by vědečtí pracovníci z UK žádná data splňující FAIR principy neprodukovali a nesdíleli. Motivovat je k tomu můžou například už zavedené postupy v jejich oboru, spolupráce v rámci výzkumných infrastruktur, požadavky zahraničních poskytovatelů financí, nebo i některých časopisů.

Z odpovědí v dotazníku (Kapitola 3.3) nicméně vyplývá, že infrastruktura poskytovaná univerzitou je poměrně roztříštěná a často se používají i volně dostupné služby typu Google či Dropbox, což může vést k potenciálním problémům s ochranou citlivých a chráněných dat. Zároveň je jako důležitý problém vnímáno dlouhodobé uchování dat, pro které není na univerzitě k dispozici vhodné technické a metodické řešení. Na UK zatím není obecně obvyklé využívaní repozitářů k publikování výzkumných dat. Pro mnoho respondentů dotazníků není sdílení v repozitářích dat typickou součástí jejich pracovního postupu a v tuto chvíli nepociťují významný vnější tlak. Přesto, že sdílení není úplně časté, existuje poptávka po možnosti využití cizích dat. Za významnou bariéru sdílení dat považujeme obavu z další pracovní zátěže.

Ve srovnání s vybranými zahraničními univerzitami tak Univerzita Karlova postrádá zejména celouniverzitní politiku určující směřování v oblasti správy dat. Zatímco vědečtí pracovníci mohou být (a zřejmě čím dál častěji i budou) k dodržování pravidel správy vědeckých dat zavázáni vnějšími subjekty (například v rámci grantových smluv), nastavení odpovídající infrastruktury a podpory by mělo být odpovědností univerzity samotné. Významným cílem politiky by mělo být, aby bylo výzkumným

pracovníkům na UK umožněno vytvářet, spravovat a sdílet svá data v prostředí srovnatelném s tím, které je k dispozici na zahraničních výzkumných univerzitách.

Na základě politiky by se pak měla připravovat infrastruktura a podpůrné služby. Je důležité brát ohled na oborové rozdíly a specifika. Na druhou stranu, pokud jsou v nějaké oblasti již k dispozici infrastruktura (např. repozitáře) a používané postupy, nemusí univerzita vytvářet nová specifická řešení, a naopak by měla využít těchto již existujících. Díky tomu lze budovat hlavně obecná řešení a zaměřit se na některá specifika práce s daty, která nejsou z pohledu oborových repozitářů podporována (například sledování a dlouhodobá ochrana zveřejňovaných výzkumných dat z institucionálního pohledu).

To, že UK není v celosvětovém měřítku mezi prvními institucemi vytvářející prostředí pro moderní správu vědeckých dat, přináší i některé výhody – není nutné vytvářet všechna řešení od začátku a prošlapávat nové cestičky. Nejen, že už existuje řada řešení, které mohou být použity jako vzor, ale s tím, jak téma nabývá na důležitosti, roste také tendence ke spolupráci mezi institucemi.

Příklady této spolupráce můžeme vidět nejen na příkladech společných projektů (například Open Science v rámci projektů Aliance 4EU+), ale i v celoevropském měřítku kde se velké množství energie a finančních prostředků využívá k přípravě platformy European Open Science Cloud (EOSC). I v tomto případě nejde o pouhou softwarovou platformu, ale o snahu komplexně řešit problematiku výzkumných dat v kontextu digitální politiky Evropské unie, a to jak z hlediska infrastruktury a služeb, tak i z pohledu lidských zdrojů a metodických postupů.

Podchycení správy vědeckých dat tak před univerzity staví nejen před poměrně náročné problémy, ale zároveň jim přináší i řadu příležitostí k jejich širšímu uplatnění nejen ve světě (otevírající se) vědy, ale i ve společnosti jako takové.

Poděkování

Rádi bychom poděkovali zástupcům zahraničních univerzit v Utrechtu, Leidenu, Lovani a Gentu, kteří s námi během online videohovorů sdíleli cenné zkušenosti se správou dat a nastavením institucionální podpory. Dále bychom rádi poděkovali našim kolegům z Masarykovy univerzity a Vysoké školy báňské – Technické univerzity Ostrava za sdílení podkladových materiálů k dotazníku a podnětné diskuze nad jeho obsahem, a také kolegům z dalších součástí Rektorátu UK, kteří se podíleli na přípravě finální verze dotazníkového šetření. Velmi bychom chtěli poděkovat všem respondentům, kteří si našli čas, dotazník vyplnili a sdíleli s námi cenné postřehy z oblasti práce s výzkumnými daty. Věříme, že se nám podaří nasbírané materiály využít a zlepšit tak prostředí a podporu, kterou Univerzita Karlova nabízí.

Podkladová data

Podkladová data k dotazníkovému šetření *Analýza správy a ukládání výzkumných dat na Univerzitě Karlově* jsou k dispozici v repozitáři Zenodo. DOI: https://doi.org/10.5281/zenodo.5841472

Použité zdroje

Bezjak, Sonja, April Clyburne-Sherin, Philipp Conzett, Pedro Fernandes, Edit Görögh, Kerstin Helbig, Bianca Kramer, Ignasi Labastida, Kyle Niemeyer, Fotis Psomopoulos, Tony Ross-Hellauer, René Schneider, Jon Tennant, Ellen Verbakel, Helene Brinken & Lambert Heller. 2018. *Open Science Training Handbook* [online] Dostupné z: https://open-science-training-handbook.github.io/ [cit. 7. 12. 2021]

Errington, Timothy M., Maya Mathur, Courtney K. Soderberg, Alexandria Denis, Nicole Perfito, Elizabeth Iorns & Brian A Nosek. 2021. Investigating the replicability of preclinical cancer biology. *eLife* 10:e71601. DOI: 10.7554/eLife.71601

Evropská komise. 2021. Horizon Europe (HORIZON) Euratom Research and Training Programme (EURATOM): General Model Grant Agreement EIC Accelerator Contract v 1.0 [online dokument]. Dostupné z: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/common/agr-contr/general-mga_horizon-euratom_en.pdf [cit. 7. 12. 2021]

Evropská Komise. 2021. *Horizon Europe Programme: Standard Application Form v 3.0* [online]. Dostupné z: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/temp-form/af/af_he-ria-ia_en.pdf [cit. 7. 12. 2021]

Kočí, Petr, Michal Zlatkovský & Jan Cibulka. 2021. Rozděleni svobodou [online]. *iRozhlas.cz*. Dostupné z: https://www.irozhlas.cz/zpravy-domov/ceska-spolecnost-vyzkum-tridy-kalkulacka_1909171000_zlo [cit. 7. 12. 2021]

Markowetz, Florian. 2015. Five selfish reasons to work reproducibly. *Genome Biology* 16(274). DOI: https://doi.org/10.1186/s13059-015-0850-7

Mráz, Miroslav. 2021. Využití otevřených dat v mezipředmětové výuce na základní škole [online]. *Otevřená data*. Dostupné z: https://data.gov.cz/články/využití-otevřených-dat-v-mezipředmětové-výuce [cit. 7. 12. 2021].

Munafò, Marcus R., Brian Nosek, Dorothy V. M. Bishop, *et al.* 2017. A manifesto for reproducible science. *Nature Human Behaviour* 1(0021). DOI: https://doi.org/10.1038/s41562-016-0021

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349(6251). DOI: https://doi.org/10.1126/science.aac4716

Open Science and Research Initiative. 2014. *The Open Science and Research Handbook v1* [online] Dostupné z: https://www.fosteropenscience.eu/content/open-science-and-research-handbook [cit. 7. 12. 2021].

STM. 2021. *The Research Data Year 2020 Dashboard* [online]. Dostupné z: https://dashboard.stm-assoc.org/ [cit. 8. 12. 2021].

Technologická agentura ČR. *Otevřený přístup v Programu KAPPA.* [online] Dostupné z: https://www.tacr.cz/otevreny-pristup-v-programu-kappa/ [cit. 7. 12. 2021].

The Academy of Medical Sciences. 2015. Reproducibility and reliability of biomedical research: improving research practice [online]. Symposium report. Dostupné z: https://acmedsci.ac.uk/viewFile/56314e40aac61.pdf [cit. 7. 12. 2021].

The Cornell Lab of Ornithology. *Macaulay Library* [online]. Dostupné z: https://www.macaulaylibrary.org/ [cit. 7. 12. 2021].

UNESCO. 2021. *Recommendation on Open Science* [online]. Dostupné z: https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en [cit. 8. 12. 2021]

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3(160018). DOI: https://doi.org/10.1038/sdata.2016.18