

MUNI

Úvodní přednáška

CORE042: Data – odpověď na základní otázku života, vesmíru a vůbec...

1. přednáška

Michal Růžička <ruzicka@ics.muni.cz> a kol.
Správa dat – Ústav výpočetní techniky MU

Přednášející a spoluautoři



- **Michal Růžička**
- ÚVT MU, Správa dat
- FAIR data, Open Science, citlivá data, bezpečnost dat
- Digitální knihovny



- **Miroslav Bartošek**
- ÚVT MU, Open Science
- Elektronické informační zdroje
- Automatizace knihoven
- Digitální knihovny



- **Jiří Marek**
- ÚVT MU, Sekretariát EOSC-CZ
- Manažer Open Science
- Citizen Science



- **David Antoš**
- ÚVT MU, CESNET, z.s.p.o.
- Oddělení datových úložišť
- Digital Preservation

MUNI

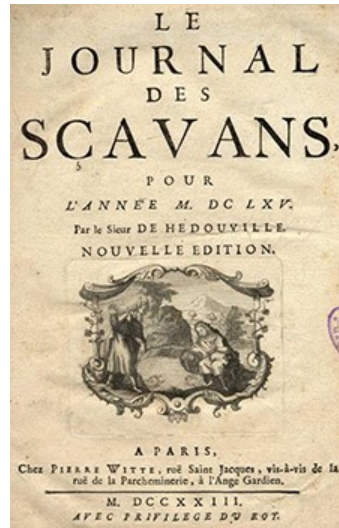
Základní informace o předmětu

aneb Troška administrativy na začátek

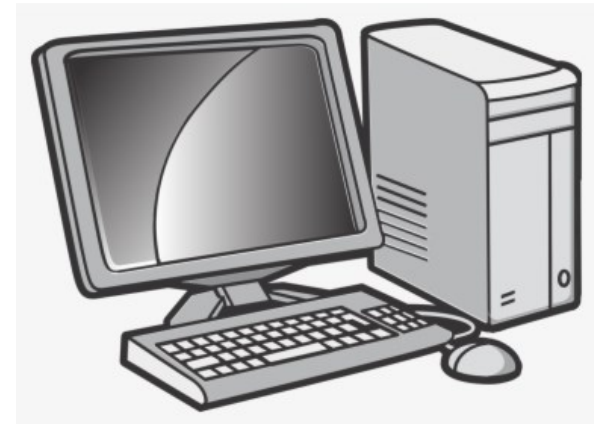
Vědecká komunikace – historie a současnost



osobní komunikace



tištěné časopisy



**globální
digitální
komunikace**

O čem náš kurz bude

- **O datech a metodách** ve vědě a výzkumu. 😊
 - A dalších náležitostech výzkumného procesu v 21. století.
- **Data jsou základem** vědeckého bádání snad ve všech současných oborech.
- **Data se stávají základem vědecké komunikace a spolupráce.**
 - **Už nestačí** vědecký článek, kde je vytištěná tabulka s výsledky.
 - **Publikace** je až **výsledek**, **základem** jsou **data**, na kterých je výsledek postaven.
- **Chceme primární data**, ze kterých byly výsledky odvozeny, která můžeme **využít pro opakování experimentu.**
 - Verifikace výsledků.**
Základ dalšího (návazného) výzkumu. Znovuvyužití nákladně/unikátně získaných dat.
- **Chceme konkrétní software**, který byl **použitý pro zpracování**, ne jen popis metody.
 - Reimplementace „na zelené louce“ není vždy žádoucí.
 - Odchytky skutečné implementace** od popisované metody mohou vést ke **zkreslení závěrů** (verifikace...).
- Ovšem **rozdíly** ve zvyklostech, přístupech a potřebách **mezi** různými vědními **obory** jsou **značné**.

Na co se můžete těšit

- **Přednášky** úspěšných výzkumníků MU **napříč fakultami a ústavy MU.**
- Ukáží vám **praxi** práce s výzkumnými daty **dle zvyklostí svého oboru**, prakticky, **na konkrétních příkladech.**
 - Bude tak mít možnost nahlédnout „**pod pokličku**“ **výzkumu na jiných fakultách**, než kterou sami důvěrně znáte.
- Doufáme, že když nahlédne do způsobu uvažování, potřeb a praxe v jiných oborech, **usnadní vám to v budoucnu spolupráci** s kolegy z jiných **oborů.**
 - I v rámci přednášek našeho kurzu se vám budeme snažit ukazovat praktické **možnosti mezioborové spolupráce.**

Co doufáme, že si z našeho kurzu odnesete

– Přehled o

- **životním cyklu** výzkumných **dat**,

- postupech pro **designování experimentů a výzkumu**,

data nestačí jen sbírat a intuitivně vyhodnocovat. Je potřeba se nad nimi zamyslet a správně je pochopit,

- **vlastnostech** tzv. **FAIR dat** a způsobech jejich **naplnění v praxi**,

- konkrétních praktických **příkladech využití dat** ve výzkumu a **přenosu výsledků do praxe/komerce**,

- **podobnostech a odlišnostech** přístupů k práci s výzkumnými daty a

- využívání dat v praxi/komeraci **napříč odlišnými** výzkumnými **obory**.

– Díky tomu budete

- lépe **připraveni** pro práci s daty při svém **studiu** nebo **výzkumu**, ale

- budete také lépe **schopni porozumění a spolupráci** s kolegy z **jiných oborů**.

Co budeme chtít po vás

- **Abychom se fyzicky vídali na přednáškách.** 😊
- **Abyste se nebáli na cokoliv zeptat.**
 - Klidně v průběhu přednášky otevřete diskusi s přednášejícím a ostatními posluchači.
Může podléhat změnám dle potřeb konkrétního přednášejícího. 😊
- **Abyste nám dali zpětnou vazbu na kurz.**
 - Co se vám **líbilo**.
 - Co se vám **nelíbilo**.
 - Co byste doporučili **dělat jinak**.
 - Buďte **upřímní**, otevření, **konstruktivní**.
 - CORE042 je **mladý kurz**, zpětná vazba od vás posluchačů je pro nás **nesmírně důležitá**.
Kdykoliv během semestru, určitě na konci semestru.
Mailem, přes Teamsy, osobně, v předmětové anketě v ISu, ...
- **Abyste odevzdali závěrečnou esej.**

Závěrečná esej

- **3 kredity** za absolvování CORE042 **nedostanete** úplně zadarmo. 😊
 - **Nebudeme** zkoušet naučené **konkrétní informace**, technické **znalosti** apod.
 - Chceme, **abyste se** nad tématy **zamysleli**.
- Na konci semestru **odevzdáte krátkou esej** na některé ze zadaných témat.
 - Krátká = **1–2 strany A4** (≈ 6–9 tisíc znaků včetně mezer)
 - Někdo to po vás všech musí také přečíst, takže za větší rozsah nejsou kladné body navíc. 😊
 - Velikost písma, řádkování a rozložení stránky musí být „rozumné“.
 - **Odevzdat do 14. 1. 2024** (tj. do měsíce od zadání témat na poslední přednášce), abychom eseje do konce zkouškového období zvládli ohodnotit.
 - Nyní 42 témat esejí k výběru, navazují na přednášky, které v průběhu kurzu uslyšíte.
 - Akceptujeme i jiné závěry a **argumentací podložená** zpochybnění našich tvrzení z přednášek.
- Závěrečné **hodnocení** jako kolokvium, tj. **binární: prospěl/neprospěl**

- Jak témata typově vypadají?
 - Patentová ochrana – motor nebo brzda pokroku?
 - Proč je důležité rozumět důvodům rozhodování umělé inteligence?
 - Je replikační krize hrozbou pro důvěru veřejnosti ve vědu?
 - Ticho jako prostředek komunikace? Nicota a prázdno jako důležitá data?
 - Je anglická Wikipedia důvěryhodný zdroj informací?
 - Digitální laboratorní deníky – výhoda nebo komplikace proti sešitu s propiskou na laboratorním stole?

Témata přednášek podrobněji (1/2)

1. (ÚVT) **Úvodní přednáška** – o čem CORE042 bude, zasazení do **širšího kontextu**.
2. (ÚVT) **Zpracování a analýza (velkých) dat** – datové modelování a analýza dat, jak se na data dívát, kde s daty **pracovat**, **sběr dat** pro analýzu **stavu krajiny**, zpracování hromady **nestrukturovaných citlivých dat** pro Policii ČR.
3. (ESF) **Rozhodnutí založená na datech: o myších a lidech** – jak **designovat experimenty** a jak si poradit, když experimenty provést nelze (etika apod.).
4. (FSpS) **Základy umělé inteligence** – využití AI metod ke zpracování dat **i bez znalosti programování**, typy strojového učení, využití AI pro **podporu sportovců při tréninku**, jak výzkum naplánovat, **sesbírat data v terénu**, zpracovat pro AI, ...
5. (FSS) **Poučení z „krize důvěryhodnosti“ v psychologické vědě** – ilustrace **nezbytnosti sdílení dat a procedur** na **zkušenosti** posledních cca 12 let s **replikační krizí** v psychologii.
6. (PřF) **Data v mikrobiologii – krok za krokem laboratoří** (i tou **antarktickou**) životním cyklem výzkumných dat s bakteriemi, mikrobiálním genomem, klinickými daty, **laboratorními deníky**, ...
7. (FF) **Data a vědecká komunikace – filozofie dat, jak vědecká data komunikovat, trendy vědecké komunikace**, způsoby **měření vědecké komunikace**.

Témata přednášek podrobněji (2/2)

8. *(PedF)* **Data pro vzdělávání – jak data správně sesbírat a interpretovat, datová věda v pedagogice, přehled pastí a propastí sběru a analýzy dat.**
9. *(PrF)* **Etické a právní aspekty práce s datami – právní aspekty vědecké práce, etika a integrita ve vědě, příklady eticky/právně pochybné vědecké praxe a publikování, právní otázky práce s daty ve výzkumu.**
10. *(CTT)* **Od akademického výzkumu k praxi – komercializace vědeckých výsledků, i věda může být business a je to dobře, sdílení vědeckých výsledků v praxi, komercializace výzkumu, ochrana duševního vlastnictví, příklady úspěšné komercializace.**
11. *(CEITEC)* **Vědecký proces – jak dělat výzkum správně, jak nám pomáhají data a kde naopak jen samotná data nestačí.**
12. *(FI)* **Jazykové korpusy – jak je vytvořit (a zbavit se nežádoucího obsahu a doplnit obsah užitečný), k čemu jsou pak dobré, jak data analyzovat a následně zpřístupnit dalším.**
+ *(ÚVT)* **zadání témat kolokviálních esejí, závěrná diskuse a zhodnocení tohoto běhu předmětu CORE042.**

MUNI

Informace a data

aneb Vezměme to od Adama

Informace a data

– Informace.

- **Údaj** o prostředí a změnách jeho **stavu**.
- **Znalost**, kterou lze předávat.
- Způsob snížení entropie systému, který informaci přijímá.
- Opak šumu.

– Informace je zachycena jazykem.

- **Přirozený** (např. čeština).
- **Formální** (**zápis** matematické formule, **formát** datového souboru v počítači).

– Data.

- Údaje **popisující stav** a vlastnosti nějakých objektů.
- V informatice: záznam informací v podobě zpracovatelné počítačem.
- **Posloupnost symbolů**, která má **přiřazenu interpretaci**.

– Data jsou uložena na **médium**.

- **Uložení dat do reálného objektu**.
- **Fyzikální reprezentace dat**.

Příklad pro historiky

Kosmova kronika

– Informace.

- „Historie“ původu Čechu od stavby babylónské věže po smrt knížete Vladislava I.

– Jazyk.

- Přirozený – latina.

– Data.

- Text (posloupnost písmen) popisu jednotlivých událostí.

– Médium.

- Rukopisy, později tištěná knižní vydání.

Příklad pro malíře

Johannes Vermeer: Stojící dáma u spinetu (1673–75, t. č. National Gallery London)

– Informace.

- Dvourozměrná vizualizace pohledu na dámu stojící u spinetu.

– Jazyk.

- Olejomalba.

– Data.

- Vyjádřená fyzickým objektem. (Nebo jednotlivé tahy štětce?)

– Médium.

- Olej na plátně.

Příklad pro metalisty

Rammstein: Amerika (singl z alba Reise, Reise, vydáno 2004)

– Informace.

- Záznam zvuku skladby. (Nebo zvuk skladby?)

– Jazyk.

- Formální – formát uložených dat (např. MP3).

– Data.

- Posloupnost bitů reprezentující záznam zvuku skladby ve formátu MP3.

– Médium.

- Pevný disk, CD, flash paměť přehrávače, ...

Médium vs. uložená informace



– Médium vs. informace.

- Rembrandtova Noční hlídka je **jen jedna**.

Je to **ten fyzický objekt** vytvořený rukou výjimečného malíře.

- Věstonická venuše je také **jen jedna**.

Ale **vystavena** je její **kvalitní kopie**.

- Shakespearovy Sonety jsou posloupnost slov.

V **řadě různých fyzických výtisků**.

- **Digitální objekty** většinou existují v **mnoha shodných kopiích**.

Médium se stává **nezajímavým**.

– Čemu **přisuzujeme hodnotu**?

- Médiu? Informaci?

- Fenomén non-fungible token: „koupím“ si virální video na YouTube.

I když je snadno kopírovatelné a veřejně dostupné.
(ehm)

- Důležité pro **dlouhodobé uchování**.

Zachovat médium?

Zachovat informaci?

- **Reprezentaci čeho (jakou informaci)** potřebujeme **zachovat**?

Příklad: Kniha – uchovat **jen text**,

nebo i jeho **rozložení** na stránce, **font** písma, **barvu**, **tloušťku**, **kvalitu** a **strukturu** papíru, **tíhu** v ruce, **zvuk** při obracení stránek, **vůni** papíru, ...?

Kopírování informací



– Bez ztráty kvality:

- Opis knihy (při pečlivé korektuře).
- Digitální soubor.
- Není to nový koncept – texty se opisují tisíce let.

Otázka je, kolik to dá práce.

Dubenkový inkoust na pergamenu je zase trvanlivý (papír ovšem leptá).

– Se ztrátou kvality:

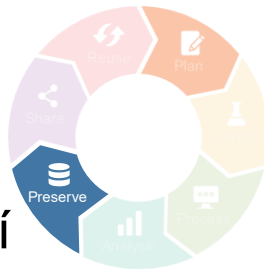
- Typicky libovolné analogové médium.
- Kopie obrazu (nebo falzifikát).
- Fotografie libovolného artefaktu.
- Video na VHS.
- Kopírování magnetofonového pásu.

Trvanlivost informací



- Pozor: I **digitální data jsou na médiu nakonec zachycena v analogové podobě.**
 - Po čase nemusí jít přečíst.
- Knihovny a další paměťové instituce.
 - Byly **vytvořeny** (a následně knihovníci po tisíciletí vychovávání) k **uchovávání médií.**
 - Před 20 lety seriózně řešily „jak zařídit, aby CD dlouho vydrželo“.
 - Teprve **nedávno** začaly **chápat specifika digitálních dat.**
 - **Dnes** masivně **digitalizují.**
 - **Chystá se fáze masivních ztrát dat.** 😊

Jak (ne)přijít o informace – příklady I



- Babylónské hliněné destičky, Svitky od Mrtvého moře, Rosettská deska a hieroglyfy.
 - **Nesmíte je fyzicky poškodit.**
- **Filmy** cca před rokem 1950 na **nitratové podložce**.
 - **Prudce hořlavé**, chemicky degraduje, zápalná teplota klesá až ke 30 °C.
 - Takto vyhořela např. budova International Museum of Photography, Rochester, NY v roce 1978.
 - Archivy pro nitratový film se staví jako bunkry.
- „Safety film“ **černobílý** (chemicky **stabilní**, založený na stříbru): Životnost i přes 100 let.
 - **Reálné zkušenosti**, ne jen **simulace** a předpoklady stárnutí – černobílý film založený na stříbru tu s námi těch 100 let už opravdu je.
- **Barevný film**: chemicky **degraduje** (ztrácí barvy).
 - Dělají se **barevné separace**.
 - Např. Disney každých zhruba 5–7 let vydává Sněhurku a 7 trpaslíků z roku 1937.
 - Pořídili červený, modrý a zelený separát.
 - Na barevných **negativech z Měsíce** dnes nejspíš **nebude vidět téměř nic**.
 - NASA pořídila několik generací kontaktních kopií.
 - Originály pouze skladuje v chlazeném trezoru.

Jak (ne)přijít o informace – příklady II



- **Analogový magnetický záznam (zvuku).**
 - Pás se **rozpadá**, dá se někdy zachránit zahřátím („zapečením“).
- **Analogové video na magnetickém pásu.**
 - Např. BBC měla politiku, že pás je médium k vysílání, nikoli k archivaci.
 - **Některé epizody Dr. Who se považují za ztracené.**
 - Monty Python si **koupili** Flying Circus od BBC, tím ho **zachránili**.
- **NASA nemá digitální video z přistání Apollo 11.**
 - Existuje TV záznam z kamery snímající obraz z projekce v řídicím středisku.
- **Papír: Noviny z 20. století jsou na kyselém papíru.**
- Občasný mail typu: „Mám tu půlpalcové magnetické **pásky** z IBM z **80. let, dokážete to přečíst?**“
 - (Nepomohla ani IBM.)

Hrozby pro (digitální) informace



– Malá životnost nosiče.

- Pevné **disky roky**.
- Zapisovatelná **CD roky**.
- Magnetické **pásy „desítky let“**.
- Čím **vyšší hustota** dat, tím **hůř**.

– Rychlé **technologické** („morální“) **zastarávání**.

- Často **nepomůže** ani **výrobce**.
- Najdete zařízení **v muzeu**, ale to není **funkční**.
- **Udržovat** zastaralé **systemy** je velmi **drahé**.

Určitě to nikdo nechce dělat jen „co kdyby se to někdy hodilo“.

– **Formát dat**.

- I když **přečtete**, často to **nepomůže**.
- **Dokumentace** k mnoha formátům se **ztratila**.

– **Software** a jeho **provozoschopnost** po delší dobu.

- Software je obvykle **vytvořený** pro nějakou **konkrétní platformu** (typ počítače, verze operačního systému).
- Nové **verze** často **nekompatibilní** se staršími.
- **Cíleně „uměle“** zaváděné **restrikce** (ochrany proti kopírování, aktivace licencí apod.).

– Pro **srovnání**: **Klasická knihovna** nebo **nápis v kameni**.

Volba datového nosiče



- **Obecné řešení neexistuje.**
- Vždy je to **kompromis** založený na
 - **dostupné technologii,**
 - **ekonomické** zvladatelnosti,
 - úvaze, **proti čemu se chráníme,**
 - **praktičnosti** řešení,
 - komerční **dostupnosti** / masovosti výroby.
- Pro interní úložiště v počítačích **disky (magnetické a SSD).**
- Pro archivy a externí úložiště většinou **disková pole** nebo **páskové knihovny.**

- **Jiná média postupně mizí.**
 - optická: CD, DVD, Blu-ray
 - magnetooptická: minidisc
 - magnetická („pružné disky“): diskety, ZIP drive
- **Narůstá ukládání dat jako služba.**
 - „Do cloudu“, což řeší **část** problémů.
- **Příklad speciálního řešení:**
 - „Černá skříňka“ v letadle.
 - **Ukládání dat do DNA.**

Kapacita na **1 gram DNA** teoreticky 455 EB, v praxi toho času (září 2022) **30 PB**; každé písmeno DNA kóduje jednu dvojici 00/01/10/11.

Dobře se kopíruje, není třeba **žádné napájení** (uložené *de facto* v krystalu cukru), odhaduje se, že **po 100 letech** zůstane **zachováno 93 % dat**. Metoda drahá, relativně – uvažte celkové náklady na uložení takového množství dat po 100 let na jiném médiu. 😊

(Zdroj: Ústní report kolegy o prezentaci vyslechnuté na konferenci ILIDE v září 2022: [OAIS-compliant digital archiving of research and patrimonial data in DNA](https://olos.swiss/); <https://olos.swiss/>)

Kategorizace úložišť na MU a doporučení pro jejich využívání



- Datová úložiště na MU: <https://it.muni.cz/kategorie/datova-uloziste>
- Doporučení pro užívání úložišť: <https://it.muni.cz/prehledy/doporuzeni-pro-uzivani-ulozist>
 - Kategorizace dat.
 - Kategorizace úložišť.
 - **(Ne)vhodnost** různých **úložišť** pro různé **typy dat**.

Ochrana dat vs. ochrana informací



– Ochrana dat.

- Tj. zajištění existence **nezměněné** posloupnosti **bitů**.
- **Obvyklá rizika**: **Selhání** médií, **chyby** administrace/software, **přepětí** v elektrické síti, **požáry/potopy**, **krádeže**, **války**, ...

– Obecné řešení: zálohování, více kopií na více místech.

- Umíme tím zvýšit **pravděpodobnost**, že uložený proud bitů (data)
 - půjde přečíst a bude **stejný**, jako na začátku,
 - nebo **když** bude **poškozený**, umíme to **poznat** a máme **další kopii**,
 - nebo** máme **smůlu**.

– To ale **nezaručí**, že přečtená data **po čase** také **k něčemu** opravdu **budou**.

– **Potřebujeme** zařídit, abychom

- data mohli **interpretovat**,
- měli **software**, který je dokáže **zpracovávat/zobrazovat** apod.

***Digital information lasts forever – or five years,
whichever comes first.***

— *Jeff Rothenberg, RAND, 1995*

MUNI

Životní cyklus výzkumných dat

Náš průvodce přednáškami CORE042

Životní cyklus výzkumných dat



Zdroj: ELIXIR RDMkit, <https://rdmkit.elixir-europe.org/>

- Jaká **data (znovu) používáte**
 - včetně licencí, které vám to umožňují,
- jaká **data generujete** a jakým způsobem,
- **kde je ukládáte, zálohujete, dlouhodobě uchováváte,**
- jak je **trvale a jedinečně identifikujete,**
- jak je **zpracováváte,**
- jak je **analyzujete,**
- kde je **zveřejňujete a sdílíte,**
- kdo tohle všechno **zaplatí;**
- **o čem data skutečně jsou,**
- **k čemu** jsou data **(ne)vhodná,**
- **kdo může data znovu použít,**
- jaká konkrétní data podporují vaše výsledky,
- jak je použít pro **opakování** vašich **experimentů** atd.

Práce s daty v projektu vyžaduje péči



- Data jsou **základ**, vyžadují patřičnou **péči**.
- Je to **proces**.
 - **Nestane** se to **jen jednou** během celého projektu (např. na začátku).
 - Je **třeba** plán **průběžně aktualizovat**.
- Práce s daty v projektu tak důležitá, že se stává **standardním požadavkem grantových agentur** apod.
 - Požadavek odevzdání tzv. Data Management Plan (DMP) **několikrát v průběhu projektu**.
- Různé technické **nástroje pro pomoc s DMP jako dokumentem i procesem**.
 - Na MU např. institucionální instance Data Stewardship Wizard: <https://dsw.muni.cz/>



Proč sdílet výzkumná data

– Rozvoj vědecké metody.

*„Metoda, která charakterizovala přírodní vědy od 17. století, zakládající se na systematickém pozorování, měření a experimentování za účelem formulování, testování a upravování vědeckých hypotéz. **Kritika je základem vědecké metody.**“*

Zdroj: Scientific method. Oxford English Living Dictionaries [online]. Oxford University Press. Dostupné z: https://en.oxforddictionaries.com/definition/scientific_method

Jak najít/vybrat datový repozitář



– OpenAIRE: [Jak najít důvěryhodný repozitář pro vaše data](#)

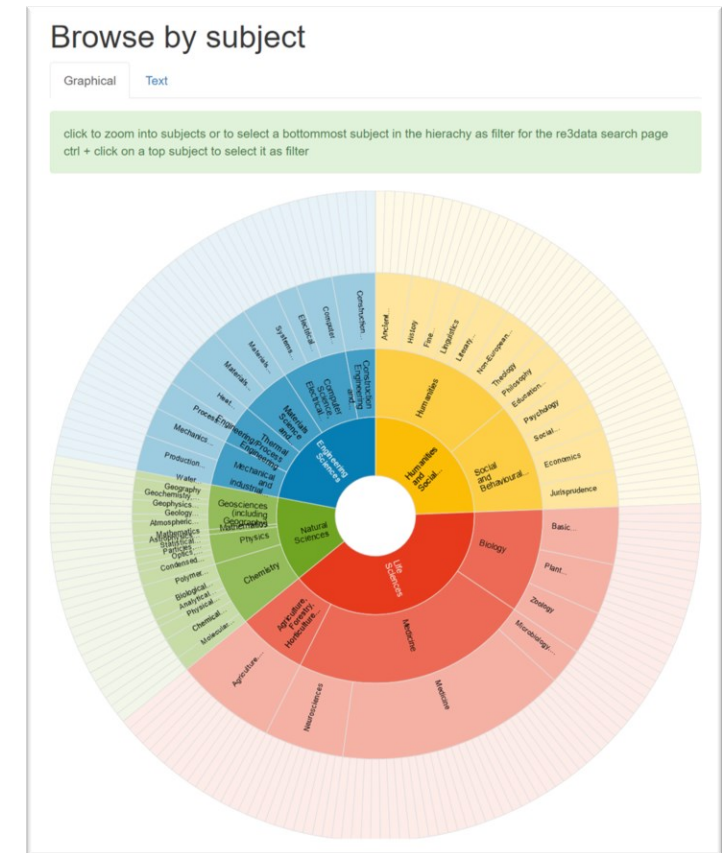
– Preferovány jsou **důvěryhodné** certifikované **repozitáře**.

- [CoreTrustSeal](#) ([seznam certifikovaná repozitářů](#)).
- [Nestor Seal](#) (verifikace dle DIN 31644).
- [ISO 16363](#).

– Např. ale známé [Zenodo](#) žádnou certifikaci nemá...

– Mezi **nejpoužívanější** obecné repozitáře patří

- [Zenodo](#),
- [Figshare](#) nebo
- [Dryad](#).



Zdroj: <https://www.re3data.org/browse/by-subject/>

Jak najít/vybrat datový repozitář (2)

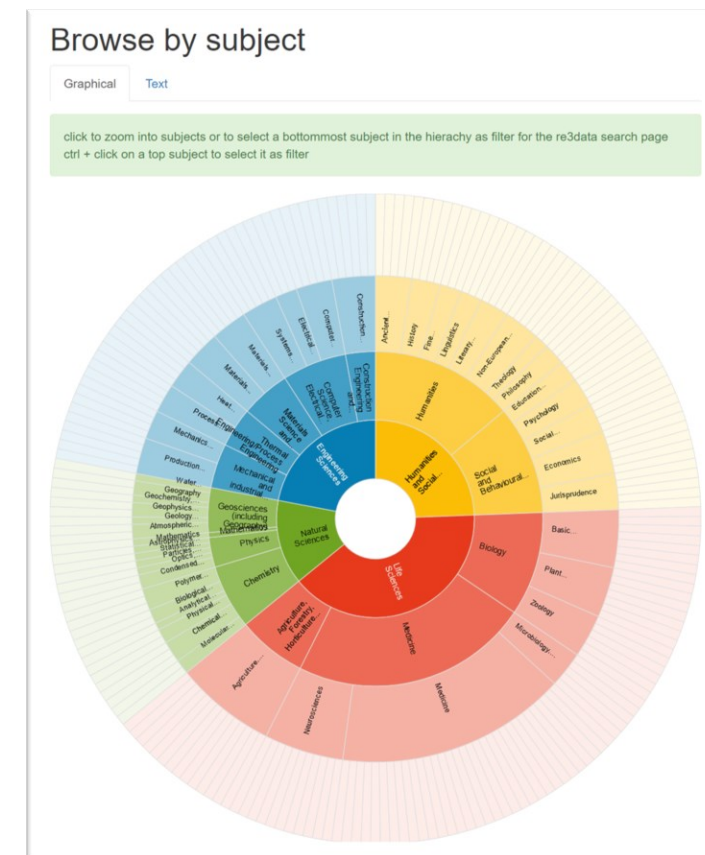


– Adresáře repozitářů:

- Open Access repozitáře: [OpenDOAR](#)
- Datové repozitáře: [re3data.org](#)

– Novinka: Datový repozitář CESNET

- Pilotní provoz: <https://data.narodni-repozitar.cz/>
- Návody: https://du.cesnet.cz/cs/navody/narodni_repozitar/start



Zdroj: <https://www.re3data.org/browse/by-subject/>

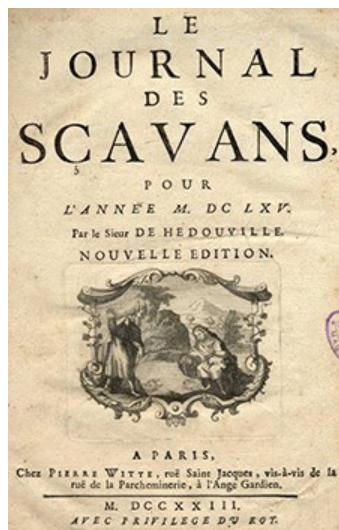
MUNI

Otevřená věda, otevřená/FAIR data

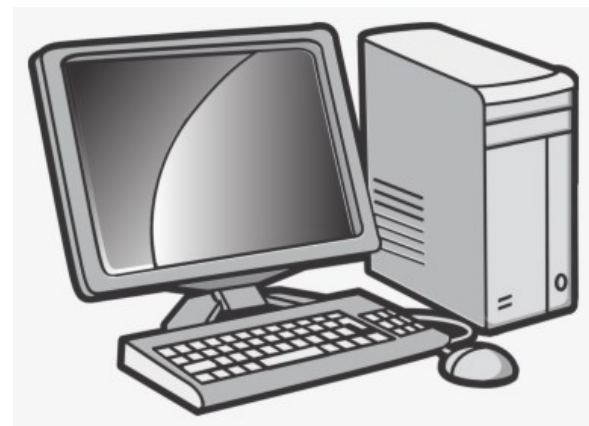
Vědecká komunikace – historie a současnost



osobní komunikace



tištěné časopisy



**globální
digitální
komunikace**

Otevřená věda a data



- **Open Access** – hnutí za otevřenou dostupnost kvalitních vědeckých publikací.

- Proč je důležité otvírat nejen publikace ale i výzkumná data?

- **Ověření správnosti výsledků.**

- Kontrola (nesprávné postupy, pomínutí nevhodných dat, falšování).

- **Reproducibilita vědy.**

- Možnost opakovat experiment a porovnat míru shody výsledků.

- **Znovuvyužití dat.**

- Úspora (**neopakovat stejné drahé experimenty**).

- Jedinečnost (**data, které již nelze nikdy získat**).

- Využití **dosud nepoužitých dat** (snímek širšího okolí sledované hvězdy).

- Využití existujících dat v novém kontextu a pro nové účely.**

- Urychlení inovačního cyklu, přístup veřejnosti,

- ...

- **Open Science** – „**všechny** kostičky dohromady“.

- **Publikace, data, citizen science, open peer review, ...**

Výzkumná data

– Doprovodná data k publikaci

– *„Data potřebná k validaci výsledků ve vědecké publikaci a s nimi související metadata.“*

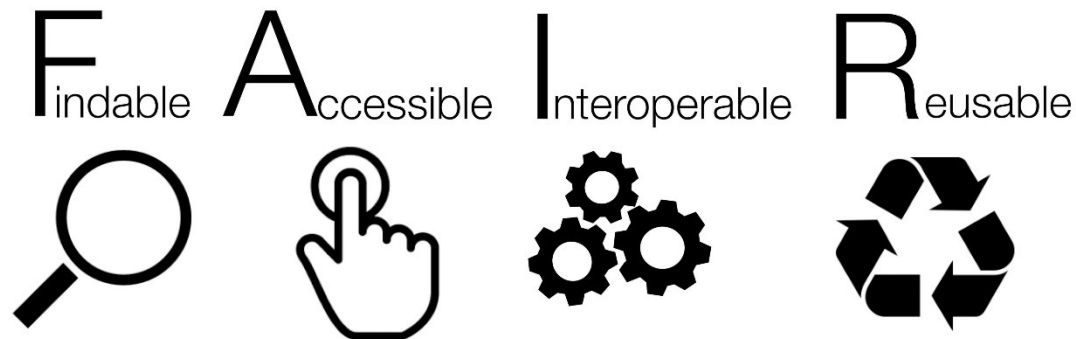
– Samostatné výzkumné sady

– *„Další data a související metadata, která se pojí s daným výzkumným projektem“
(a jsou uvedena v tzv. Data Management Planu)*

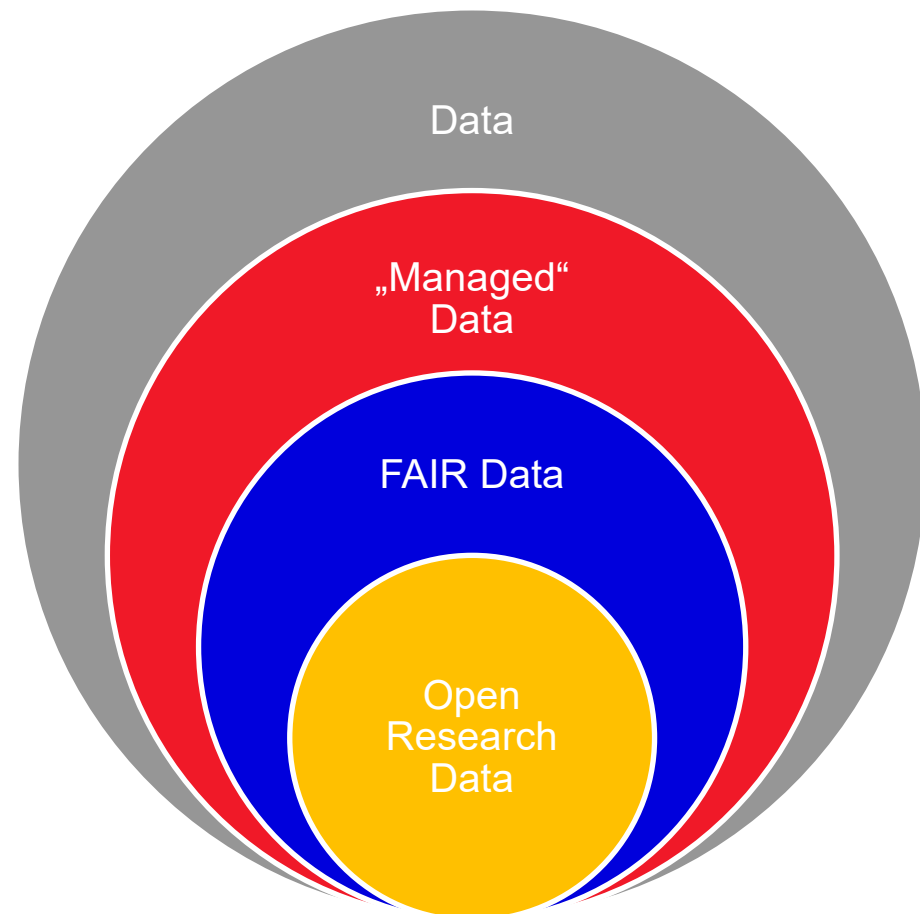
Zdroj: H2020 Programme AGA – Annotated Model Grant Agreement Version 5.2 ze dne 26 června 2019. [online] s. 248. Dostupné z: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/amga/h2020-amga_en.pdf

Jak otevírat data? – Bud'me FAIR!

- Základní moto: ‘**As Open as Possible, As Closed as Necessary**’
- FAIR data:
 - **Findable** – Dostatečně podrobná **metadata**, globální **persistentní identifikátory (PID)**.
 - **Accessible** – Metadata i data **srozumitelná lidem i strojům**, důvěryhodné repozitáře.
 - **Interoperable** – **Strojově zpracovatelná** data i metadata v zavedených **standardech**.
 - **Reusable** – Jasná **licence**, přesná **data o původu (data provenance)**.



Úrovně dat



Perzistentní identifikátory (PID)



- Mají zajistit **oddělení identifikace objektu** jako takového,
 - osoba,
 - instituce,
 - publikace,
 - dataset,
- **od jeho momentálního fyzického umístění.**
- **Příklad** – datová sada *https-set*
 - **Identifikátor** datové sady: <https://doi.org/10.48791/4mxxp-r725>
 - **Současné fyzické umístění:** <https://ucnmuni.sharepoint.com/teams/mu-UVT-https-set/Shared%20Documents/Forms/AllItems.aspx?id=%2Fteams%2Fmu%2DUVT%2Dhttps%2Dset%2FShared%20Documents%2Fhttps%2Dset%2Dv1%2E0%2E0&p=true&ga=1>
 - Fyzické umístění se **bude pravděpodobně** v budoucnu **měnit** – zvažován přesun do [pilotně provozovaného datového repozitáře CESNET](#).
 - **Změny nevadí** – **uživatelům** je jako odkaz na data **vždy** prezentováno [DOI 10.48791/4mxxp-r725](https://doi.org/10.48791/4mxxp-r725), které je vždy **zavede** na **aktuální umístění**.

Perzistentní identifikátory (PID) (2)



- Mají zajistit **jednoznačnost**.
- **Příklad** – jména fyzických osob.
 - **Více forem zápisu** jména **jedné fyzické osoby**.
 - **Více různých fyzických osob** se **stejným jménem**.
- Mají zajistit **trvalost** (perzistenci).
 - **Metadata** fyzicky umístěna **u třetí strany** nezávisle na fyzickém umístění odkazované entity.
 - Vlastník identifikátoru se stará o **aktualizaci metadat** a aktualizace **směřování** na aktuální umístění.
 - Třetí strana **pečuje** o **zachování** poslední verze a historie **metadat** a **existenci identifikátoru** jako takového, i pokud se vlastník identifikátoru o něj starat přestane. A dokonce i tehdy, pokud identifikovaná entita jako taková nebude zachována.

<input type="checkbox"/>	NovakD (1)	Novák, David (1)	Join	Delete
<input type="checkbox"/>	NovakE (5)	Nováková, Eva (5) Nováková, E. (0)	Join	Delete
<input type="checkbox"/>	NovakJ (52)	Novák, Josef (38) Novák, Jos. (1) <u>Novák, J. (13)</u>	Join	Delete
<input type="checkbox"/>	NovakJ2 (7)	Novák, Jiří (7) Novak, Jiri (0)	Join	Delete
<input type="checkbox"/>	NovakJ7 (19)	Novák, Josef (16) <u>Novák, J. (3)</u>	Join	Delete
<input type="checkbox"/>	NovakK (2)	Novák, Karel (2)	Join	Delete
<input type="checkbox"/>	NovakM (2)	Novák, Mirko (2) Novak, Miroslav M. (0) Novak, M. M. (0) Novák, M. (0)	Join	Delete
<input type="checkbox"/>	NovakM2 (2)	Nováková, Markéta (2)	Join	Delete
<input type="checkbox"/>	NovakM3 (1)	Novák, Miroslav (1)	Join	Delete
<input type="checkbox"/>	NovakO (2)	Novák, Ondřej (2)	Join	Delete
<input type="checkbox"/>	NovakP (1)	Novák, Petr (1)	Join	Delete
<input type="checkbox"/>	NovakS (1)	Novák, Stanislav (1)	Join	Delete
<input type="checkbox"/>	NovakV (57)	Novák, Vítězslav (55) Novák, V. (2)	Join	Delete
<input type="checkbox"/>	NovakV2 (7)	Novák, Vilém (7) Novák, V. (0)	Join	Delete
<input type="checkbox"/>	NovakV3 (53)	Novák, Vladimír (53)	Join	Delete
<input type="checkbox"/>	NovakZ (1)	Novák, Zdeněk (1)	Join	Delete

Zdroj: Autoritní databáze projektu [DML-CZ](#)

Řada typů PID



CODEVALUE	PREFLABEL_EN	HREF	Maturity	Globally resolvable	For which object type	Comments	
ADSbibcode	Astrophysics Data System - Bibliographic Reference Code	https://ui.adsabs.harvard.edu/	High	Needs token	Publication		
ARK	Archival Resource Key	https://arks.org/	High	Local	Other (in comments)	Everything	
arXiv	arXiv identifier scheme	https://arxiv.org/	High	Yes	Publication		
ASIN	Amazon Standard Identification Number	https://sellercentral.amazon.ca/gp/help/external/200317470?language=en-CA&ref=mpbc_200576730_cont_200317470	High		Other (in comments)	Things sell by Amazon	
ConfID	Conference identifier	https://indico.cern.ch/event/780651/attachments/1776614/2888642/Conference_PIDs_and_Crossmark.pdf		?	Event	Not clear if it is a Crossref service	
Crossref DOI			High	Yes	Publication		
Crossref_funders	Crossref Funder Registry	https://www.crossref.org/services/content-registration/grants/	?	Yes	Organisation		
Crossref_grants	Registering research grants	https://www.crossref.org/community/grants/			Other (in comments)	Grants	
DataCite DOI							
DOI	Digital Object Identifier	https://www.doi.org/	High	Yes	Other (in comments)	28 different resources and outputs	
			High		Publication	Services supporting PIDs and metadata for 40+ resource and output types	
EAN13	The 13-digit International Article Number	https://www.gs1.org/standards/barcodes/ean-upc	High	?	Other (in comments)	Physical product identifier. A Whole famili of id: UPC-A, UPC-E, EAN13, EAN8	
eISBN	electronic International Standard Book Number	https://www.isbn-international.org/			Publication		
eISSN	Electronic International Standard Serial Number	http://portal.issn.org/	High	https:// portal .issn .org/ resource/ ISSN/ 0376 -4583<	Publication	Identifies various types of serial publications (eg. journals, websites, blogs)	
GRID	Global Research Identifier Database	https://www.grid.ac/	Closed	?	Organisation	Transitioned to ROR	
Handle	Handle	http://www.handle.net/	High	Yes	Dataset	It is the base of DOI also	
ISSN	International Geo Sample Number	https://www.igsn.org/	High	Yes	Other (in comments)	Physical Samples and Sampling Features	
ISAN	International Standard Audiovisual Number	https://www.isan.org/		Yes?	Publication		
ISBN	International Standard Book Number	https://www.isbn-international.org/	High		Publication		
ISLI	Identifies the links between different entities	https://www.isbn-international.org/content/isli-introduction	?	Yes	Other (in comments)	Link bw. entities	
ISMN	International Standard Music Number	https://www.ismn-international.org/		No	Publication		
ISNI	International Standard Name Identifier	https://isni.org/page/search-database/			Person	Contributors to creative works and their distribution and organizations	
ISSN	International Standard Serial Number	http://portal.issn.org/	High	https:// portal .issn .org/ resource/ ISSN/ 0376 -4583<	Publication	Identifies various types of serial publications (eg. journals, websites, blogs)	
ISTC	The International Standard Text Code	http://www.istc-international.org/	Defunct	No	Publication	Ceased in 2017	
LSID	Life Sciences Identifier	http://www.lsid.info/	?	Yes	Other (in comments)	Metadata for life science items	
ORCID	Open Researcher and Contributor ID	https://orcid.org/	High	Yes	Person		
PIC	(EC) partner identity code	https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/support/faq/1055					
PMID	PubMed ID	https://www.ncbi.nlm.nih.gov/pmc/pmcotpmid/			Publication		
PURL	persistent uniform resource locator	https://archive.org/services/purl/	High	Yes	Other (in comments)	Resources on the Web	
QID	Wikidata identifier	https://www.wikidata.org/wiki/Wikidata:Identifiers			Other (in comments)	Knowledge item	
RAID	Persistent Identifier for research projects	https://www.raid.org.au/			Needs token	Other (in comments)	Research projects
Ringgold	Unique numerical identifier applied to organizations in the scholarly supply chain	https://www.ringgold.com/			Organisation		
ROR	Research Organization Registry	https://ror.org/	High	Yes	Organisation		
RRID	Research Resource Identifier	https://scicrunch.org/resources					
ScopusAuthorID	Scopus Author ID	https://service.elsevier.com/app/answers/detail/a_id/11212/supporthub/scopus/	?	No?	Publication		
SWHID	Software Heritage persistent Identifiers	https://docs.softwareheritage.org/develop/sw-h-model/persistent-identifiers.html		Local	Source Code		
UPC	Universal Product Code	https://www.gs1.org/standards/barcodes/ean-upc			Other (in comments)	Synonym of EAN13? Product identifier	
URI	Uniform Resource Identifier						
URL	Uniform Resource Locator						
URN	Uniform Resource Name						
VAT-number	VAT number	http://ec.europa.eu/taxation_customs/vies/vatRequest.html			Organisation		

Zdroj: Interní pracovní materiály [EOSC Task Force PID Policy and Implementation](#)

Proč jsou výzkumná data „těžká“?

- **Nelze vždy požadovat okamžitý přístup.**
 - Právo prvního využití.
- **Nelze vždy otevřít.**
 - Citlivé osobní nebo komerční údaje.
- **Velmi velký rozsah.**
 - TiB a více, miliony souborů, rychlý růst v čase.
- **Velká variabilita formátů a forem.**
 - Často netextové.
- **Rozdílné oborové standardy.**
 - Pokud vůbec existují.
- **Různé třídy dat.**
 - Raw data – zpracovaná data – analyzovaná data.
- **Velká pracnost se zpřístupněním dat někomu jinému.**
 - Uspořádání, popis, přenos, řízení přístupu.
- **Málo prozkoumaná oblast.**
 - Důvěryhodnost, úplnost, kvalita, vlastnictví, dlouhodobé uchování, kurátorství, ...
 - Ocenění akademickou komunitou?
- **Soudobý trend: pojd'me data otevřít!**
 - I přes tu velkou složitost.

MUNI

Data ve výzkumu a vývoji

aneb Širší pohled na začátek

Big Data

– Buzzword posledních let.

- Když **mám** opravdu **hodně dat**, **umím** z nich **lecos vyčíst**.

Statistika už **opravdu funguje**, AI (artificial intelligence, umělá inteligence; další buzzword).

- **Základ businessu největších technologických gigantů dneška**: Google, Facebook, ...

– Big Data = objem dat na hranici zpracovatelnosti soudobými technologiemi.

- Výzkum: Aktuální hranice kolem exabytu dat (10^{18} bytů), cca objem denní světové produkce.
- V praxi: Jakýkoliv *hóóódně* velký soubor dat.

– Nové výzvy.

- Přenos, uchování, zpracování, vyhledávání, získávání výsledků v reálném čase, ...

– Nejrůznější zdroje dat.

- Vědecký výzkum, provozní data (platby kartami, mobility), sociální sítě, státní správa.
- (Částečně) strukturovaná nebo i nestrukturovaná data.

– Obrovské možnosti/potenciál využití v nejrůznějších oblastech.

- Marketing (i politický), výzkum, zdravotnictví, vojenství, business aplikace.

– Údaje vytěžené z velmi velkých souborů dat již mají charakter zákonitostí.

- Chování lidí, společenské a přírodní jevy (COVID-19), podpora strategického rozhodování.

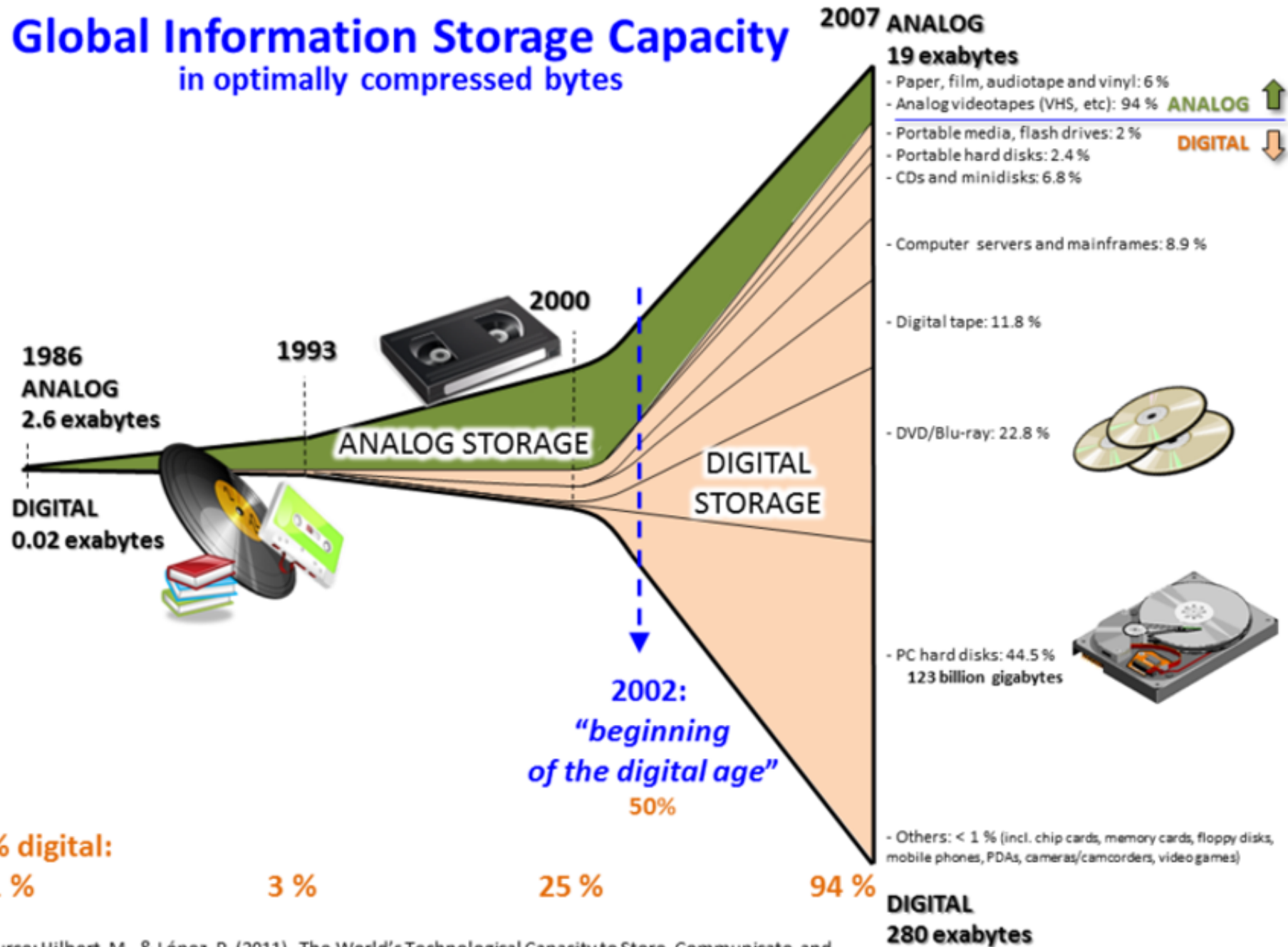
Vsuvka: Kilo, Mega, Giga – jak dál?

– Kilobyte	KB	2^{10}	10^3	tisíc	
– Megabyte	MB	2^{20}	10^6	milion	počet živočišných druhů
– Gigabyte	GB	2^{30}	10^9	miliarda	počet obyvatel Indie
– Terabyte	TB	2^{40}	10^{12}	bilion	počet všech ryb v oceánech
– Petabyte	PB	2^{50}	10^{15}	biliarda	počet mravenců na Zemi
– Exabyte	EB	2^{60}	10^{18}	trilion	inflace v Zimbabwe 2009
– Zettabyte	ZB	2^{70}	10^{21}	triliarda	počet zrněk písku na Zemi
– Yottabyte	YB	2^{80}	10^{24}	kvadrilion	počet hvězd ve Vesmíru
– Počet atomů na Zemi			10^{50}	(10^{78} – 10^{82} ve Vesmíru)	

Role IT v rozvoji vědy

- Od poloviny 20. století **uplatnění IT ve výzkumu** (projekt Manhattan).
- **Počítače vzácné a drahé, přístup jen pro znalé a „vyvolené“.**
- **Boom IT počátkem 90. let (3C):**
 - **Computing** – růst výpočetního výkonu, rozšíření osobních počítačů (osobní IT).
 - **Communications** – sítě, zrychlení přenosu dat (text, audio, video), penetrace internetu.
 - **Content** – výrazný nárůst paměťových kapacit, růst obsahu dostupného v digitální formě.
- **Digitální éra.**
 - **Široké pronikání IT do výzkumu** – výpočet, data.
 - **Zrychlení komunikace** – internet, web.
 - **Rozvoj SW nástrojů a aplikací.**
 - **Automatizace** postupů a procesů.
 - **Online dostupnost výsledků.**

Global Information Storage Capacity in optimally compressed bytes



Příklad z praxe: *Human Genome Project* (HGP)

- Projekt **mapování lidského genomu**.
- **1990–2003**, **ambiciózní**, srovnáván s projekty Manhattan, Apollo.
- **3 miliardy USD**, **mezinárodní** (20 laboratoří z USA, UK, JP, FR, DE, CH).
- **Cíl: Přechíst** kompletní genetickou informaci (**DNA**) **člověka** (sekvence **3,1 miliard** nukleotidů A, G, C, T).
- **Obrovský objem prací**, **nezvládnutelný bez nových postupů a technologií**.
- Získaná data a technologie **otevřít komukoliv**.
- **Očekáván obrovský přínos** pro medicínu, genetiku, molekulární biologii, další.
- **Etické**, společenské a právní **otázky**.
- Základní metoda: Sekvence DNA.
 - **Vynalezena** teprve pár let před projektem.

Sekvenace DNA



- **Zjišťování pořadí nukleových bází** („písmen“ A, C, G, T) v krátkých sekvencích DNA pomocí **biochemických metod a počítačového zpracování**.
- Sangerova metoda sekvenování (1977).
 - Část **DNA rozsekáme na malé úseky** (tisíce písmen), ty **přečteme a seskládáme**.
 - **Zjednodušený příklad** (dle Storchová Z: *Homo sapiens sapiens: přečteno!*, Vesmír 97, 2000/8, 427–429).

Chcete přečíst větu: „*Tak dlouho se chodí se džbánem pro vodu, až se ucho utrhne.*“

Neznáte ale jazyk, takže se nemůžete domýšlet, a umíte přečíst vždy jen pár znaků.

Celý text rozdělíte na malé úseky, náhodně. Získáte např. něco podobné tomuto:

Na počítači vyhledáte překrývající se úseky (např. „*trhn*“ a „*hne.*“).

Seřadíte z toho kratší části a nakonec celou větu.

e chod	louho	ž se u
vod	o se	hodí
e džb	trhn	í se d
trhn	hne.	ak dl
u, a	až	s dlou
vod	ro vo	u, a

- U lidského genomu má ale ta věta celkem **3.1 miliardy znaků!**

Pokud bychom ji přepsali do běžného textu **knihy A4**, dostaneme **sloupec knih 30 metrů vysoký!**

Good guys vs. Bad boys



- **Postup** prací HGP byl **velmi zdlouhavý, pomalý**.
- 1998: Craig Venter **odešel** z projektu a **založil** komerční **firmu** Celera Genomics.
 - **Cíl: Předběhnout HGP**, získat **patenty** na geny a **prodávat** je zájemcům!
Financování od farmaceutických firem, soukromých investorů.
- **Nové zjednodušené postupy** sekvenace – **ne tak přesné, ale rychlejší**.
- Obrovská **rivalita** (nepřátelství) a soutěžení **mezi** oběma **týmy**.
- **Remíza**: 2000 **zveřejnili společně** pracovní verzi genomu (finální 2003) (de-facto porážka firmy Celera).
- 2013: Nejvyšší soud USA: DNA je **produktem přírody** a **nelze ji patentovat!**

Výsledky a dopady HGP

- **Velký úspěch:** S vysokou přesností **zmapován** kompletní **genom člověka!**
 - **Historie genetiky** rozdělena na „před“ a „po“.
- **Vznik celých nových vědních oborů.**
 - Bioinformatika, computational genomics, ...
- **Rozsáhlé veřejně přístupné databáze** a genové banky (GenBank, ...).
- **Obrovské zrychlení a zlevnění** sekvenování.
 - **Dnes** celý člověk za pár hodin a pár set USD.
- **Pokroky ve zpracování velkých objemů dat.**
 - Dnešní sekvenátory TB dat/den, viz CEITEC-MU.
- **Sekvenovány genomy** velkého množství **různých organismů.**
 - I vyhynulých – neandrtálec, mamut.
- **Rozvoj poznání** v mnoha oblastech.
 - Původ a vývoj druhů, migrace, ...
- **Ale:**
 - **Genomu** dosud **až tak nerozumíme**, je mnohem komplikovanější, než jsme si mysleli.

Pačes: „Jsme na tom stejně, jako bychom přečetli celou knihu v portugalštině a neuměli portugalsky.“
 - „**Odpadní**“ část **DNA** (nekóduje geny, 98 % DNA) hraje **mnohem větší roli**, než jsme si mysleli.
 - **Některá očekávání** se zatím **nenaplnila**, nebo jen z části.

Personalizovaná medicína.
- **COVID-19:**
 - **Vakcína od Moderna:** Na počátku pandemie COVID-19 – **Fyzicky virus nikdy neměli** k dispozici, stačila jim **pouze genetická informace** jako **digitální data**. Virus vnímali jako kus softwaru.
 - O dva dny později byla **vakcína hotová**. **Zbytek** roku zabraly **klinické testy**, první země vakcínu schválily kolem Vánoc.

Etické a společenské otázky

- **Ochrana (vysoce citlivých) osobních údajů.**
- **Patentování (uzavírání) informací.**
- **Psychologické aspekty.**
- **Genetické inženýrství (dítě na přání).**
- **Eugenika.**
- **Dostupnost benefitů jen pro někoho (bohaté).**

MUNI

Shrnutí

1. přednáška CORE042

Shrnutí

– **Data ≠ informace!**

- Data musí být správně pochopena, interpretována.

– **Data a nástroje a metody jejich zpracování jsou základem i výsledkem soudobého výzkumu a vědecké komunikace.**

– **Vyžadují náležitou péči a porozumnění.**

Abychom o ně nepřišli, a to i v dlouhodobém horizontu.

Abychom správně designovali výzkum a interpretovali výsledky.

- **Životní cyklus dat** platí **obecně napříč obory**,
- ale **implementace** jeho fází je **oborově závislá**.

– **O čem náš kurz bude.**

- **O datech** ve vědě a výzkumu.
- **O metodách** na datech založeného výzkumu **21. století**.

– **Na co se můžete těšit.**

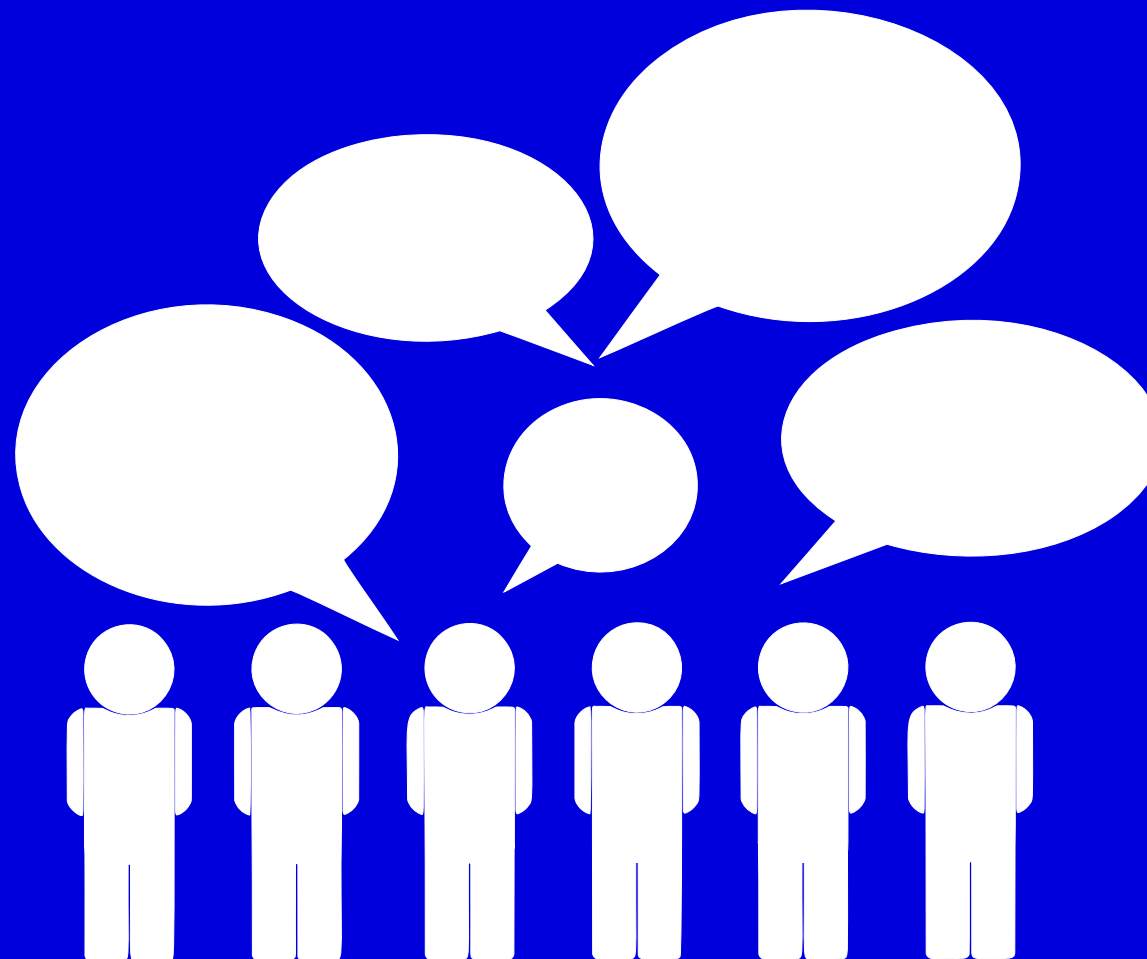
- **Přednášky** úspěšných výzkumníků MU **napříč fakultami a ústavu MU**.
- **Příklady z praxe** z mnoha **různých úhlů pohledu**.
- **Přehled o životním cyklu dat** ve výzkumu a **postupech designování výzkumu**.

– **Nebudeme toho po vás chtít mnoho.**

- **Fyzická přítomnost** na přednáškách.
- **Kratičká eseje** na konci semestru.
- **Konstruktivní zpětná vazba** k mladému předmětu.

MUNI

Diskuse



Zdroj: [Communicate_communication_conference_2028004](#) od [OpenClipart-Vectors](#) z [Pixabay](#)