Úvod do problematiky dat

Práce s daty, 16. února 2023

Životní cyklus dat

1. Plánování správy dat
2. Organizování a dokumentace (příprava před zpracováním)
3. Zpracování dat
4. Uložení dat
5. Ochrana dat
6. Archivování a zveřejnění dat
7. Objevování dat

Zdroj: https://dmeg.cessda.eu/

Plánování správy dat

Pokud to s daty myslíme „vážně“, pak je nutno vzít v úvahu aspekty:

- technické, organizační, strukturální, právní, etické a udržitelné

DMP (Data Management Plan):

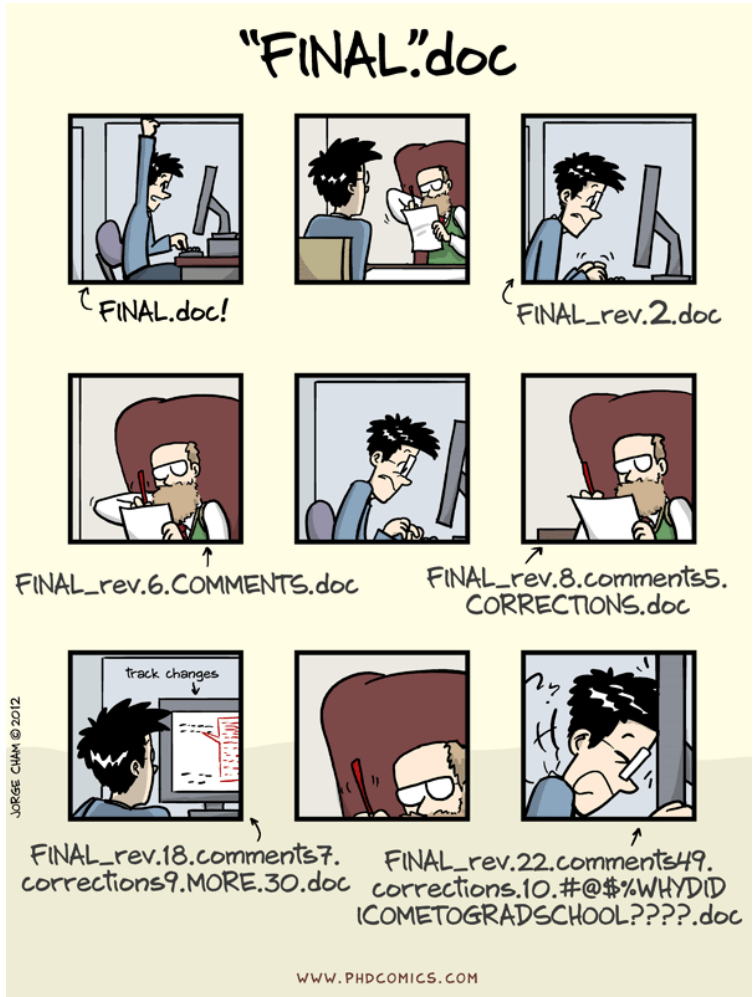
1. umožňuje získat vhled do problematiky
2. zjednodušuje projektový management
3. dodá podklady pro potřebný rozpočet
4. zaštítí zodpovědnost za spravovaná data
5. napomáhá „FAIRifikaci“ dat

dsw.muni.cz

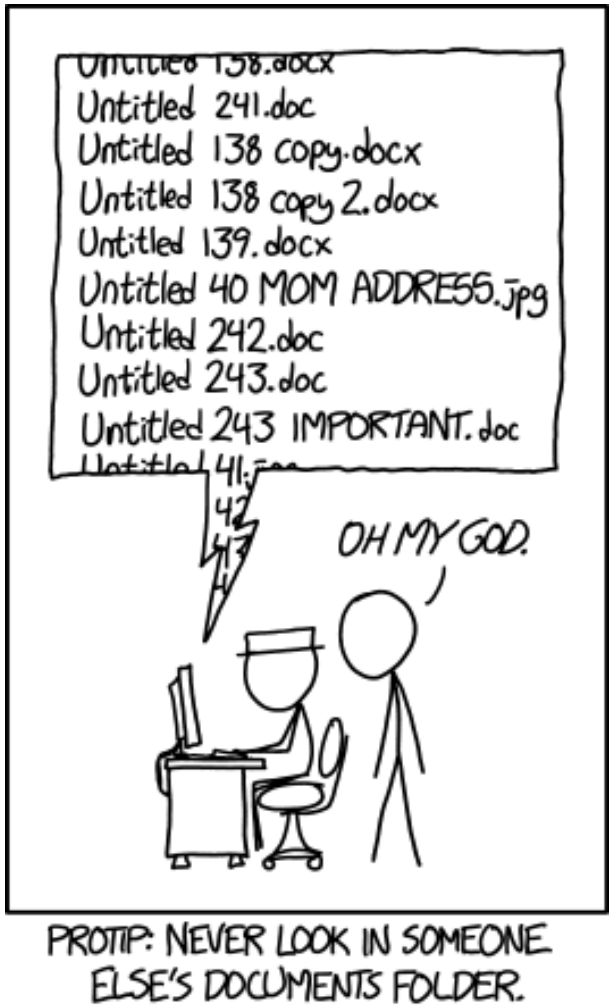
FAIR data

- Findable
  - (meta)data jsou snadno dohledatelná strojově i člověkem
  - (meta)data mají přiřazen jednoznačný perzistentní identifikátor
- Accessible
  - (meta)data jsou přístupná pomocí otevřeného přenosového protokolu
  - protokol v případě potřeby umožňuje autentizační a autorizační proceduru
- Interoperable
  - (meta)data jsou dále strojově zpracovatelná
- Reusable
  - (meta)data jsou srozumitelně popsána a mají přiřazenu licenci k jejich užití

Organizování a dokumentace



Zdroj: https://phdcomics.com/comics.php?f=1531



Zdroj: https://xkcd.com/1459/

Organizování a dokumentace

- Jak jsou data uložena?
  - datová matice, *relační vztahy*, hierarchický soubor
- Jak pojmenovat soubory a identifikátory (*proměnné*)?
  - systematicky, krátce, přenositelně, bez diakritiky
- Jak zajistit integritu datového souboru?
  - vhodný software, *kontrola vstupů* a kontrola úplnosti záznamů
  - „master file“, dokumentovat provedené změny, používat verzování
  - zálohovat a používat vhodné formáty i *kódování znaků*
- Jak dokumentovat data?
  - *metadata* pro projekt, databázi a samotné případy (proměnné)

Více na https://archiv.soc.cas.cz/cz/prace-s-daty/management-dat-v-prubehu-vyzkumneho-projektu

Datové typy proměnných

- Číslo
  - binární, oktanové, decimální, hexadecimální
  - přirozené, celé, reálné … různé rozsahy (2^8, 2^16, 2^32, …)
- Text
  - Znak, krátký řetězec (max. 255 znaků), dlouhý text
- Ano / Ne (True × False; 1 × 0)
- Datum
  - jako číslo nebo řetězec znaků
- Kombinovaný typ
  - pole, množina, index (lineární, strom, …), objekt, …

Kódování znaků (národní abecedy)

- ASCII – 7bitové schéma kódování, nejstarší a stále užívané
- CP852, ISO-8859-2, Windows-1250, … – 8bitová rozšíření ASCII
  - Vždy jen max. 256 různých znaků v rámci dané sady
- Unicode – rodina unifikovaných kódovacích schémat
  - UTF-16, UTF-32 – 16bitové nebo 32bitové schéma
  - UTF-8 – proměnná délka znaku: 1–4 bajty
    - Aktuálně nejrozšířenější a nejoblíbenější způsob kódování znaků
- Převodníky: www.freeformatter.com, string-functions.com

Zpracování dat

- Kódování vstupních dat (kvantitativní, kvalitativní)
- *Kontrola dat*, *čištění*, dopočítávání chybějících hodnot
- *Transformace do požadované podoby*
- *Propojování datasetů*
- *Jednoduché datové analýzy*
  - Četnosti, průměry … popisná statistika – není zapotřebí statistický software
- Pokročilé datové analýzy
  - Korelace, regrese, testování hypotéz – je zapotřebí statistický software
- Vizualizace dat

Uložení dat

Zdroj: https://dilbert.com/strip/2016-04-04

Uložení dat

- Kde jsou data uložena?
  - Soubory, SQL databáze, NoSQL databáze, jiný způsob
  - Lokální disk, síťový disk, cloud
- Jak jsou data zálohována?
  - Manuálně, automaticky
  - Externí lokální médium, síťové úložiště, …
  - Co je co není potřeba zálohovat?
  - Jak dlouho bude zálohy drženy a jak budou zničeny?
- Kolik místa živá data i zálohy potřebují?
- Kdo je za uložená a zálohovaná data zodpovědný?

NoSQL databáze

- Jsou obecně takové, které nejsou založené na relačním modelu
- Kategorie odlišných přístupů k uložení dat
  - Key-value – do DB se ukládá klíč a hodnota (známe-li klíč, získáme obsah)
  - Column-oriented – tabulka bez pevné struktury (různé řádky mohou mít různé sloupce, nelze tak tak spojovat tabulky stejně jak v relačním modelu)
  - Document-oriented – data uložena jako dokumenty ve vhodném strukturovaném formátu (např. XML, JSON)
  - Graph – relace mezi daty lze dobře reprezentovat jako graf (propojené uzly) – časová osa, linky veřejné dopravy, počítačová síť, mapové aplikace
- Visual Guide to NoSQL Systems, CAP teorém

Ochrana dat

Zdroj: https://cloudtweaks.com/wp-content/uploads/2016/03/comic-att-1.jpg

Ochrana dat

- Řízení přístupu k datům
  - Autentizace uživatele – uživatelské účty, hesla, …
  - Autorizace uživatele – práva pro čtení, zápis, změnu oprávnění, …
  - Šifrování dat – v úložišti, při přenosu po síti
  - Bezpečná likvidace dat
- Osobní údaje
  - Zdravotní údaje, vyznání, kontakty, sexuální orientace, biometrická data
  - Právní aspekty – GDPR, informovaný souhlas, další zákonné předpisy
  - Anonymizace dat
- Vlastnictví dat, spoluautorství

Archivace a zveřejnění dat

- Problematika archivace (dlouhodobého uložení) dat
  - Výběr média / služby – spolehlivost
- Výběr dat ke zveřejnění
  - Potenciál dat ke znovupoužitelnosti
  - Zveřejnění v rámci instituce (společnosti) / komunity / komukoliv
  - Popis dat, dokumentace ke zveřejnění
- Licencování dat a související právní otázky
  - Licence Creative Commons; CC0 = je možné s daty dále dělat cokoliv
  - Je součástí dat metodika / skript / aplikace?

Objevování dat

- Propagování dat
  - Skrze publikovanou výzkumnou zprávu (např. v odborném časopise)
  - Vložením do indexovaných veřejných datových repozitářů
    - Univerzální, např. Zenodo, Figshare, …
    - Oborově specifické – registr např. na re3data.org
- Citování dat
  - Hodí se perzistentní identifikátor, např. DOI
- Veřejné rejstříky
  - Národní, evropské (zejména EOSC), celosvětové
- Jsou dostupná data důvěryhodná?

Další čtení a odkazy

- E-knihy
  - Data Management: A Practical Guide for Librarians
  - How to be FAIR with your data
  - Data Stewardship for Open Science: Implementing FAIR Principles
- Zajímavosti
  - Bezpečné užívání technologií
  - Anonymizace a pseudonymizace v ochraně dat a informací
  - Proč klasická anonymizace (a pseudonymizace) nevede k anonymním datům
  - Amnesia – aplikace na anonymizaci dat