Typy a rozdelenie otvorených (vedeckých) dát

Matej Harvát

Odbor podpory otvorenej vedy

CVTI SR

NISPEZ IV
Investícia do Vašej budúcnosti

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

MINISTERSTVO DOPRAVY A VÝSTAVBY SLOVENSKEJ REPUBLIKY

MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY

CENTRUM VEDECKO-TECHNICKÝCH INFORMÁCIÍ SR

Viacero definícií

- Všeobecná (OpenAIRE, Open Data Handbook)

Otvorené výskumné údaje *sú údaje, ktoré môže ktokoľvek voľne používať, opätovne využiť a ďalej šíriť - pod podmienkou uvedenia autora a zdieľania pod rovnakou licenciou*.

- Právna (Autorský zákon SR, Smernica 2019/1024 – nová PSI smernica)

otvorené dáta *= údaje v otvorenom formáte, ktoré môže ktokoľvek voľne používať, opakovane používať a zdieľať na súkromné i obchodné účely s minimálnymi alebo žiadnymi právnymi technickými alebo finančnými obmedzeniami.*

- Otvorené dáta podľa CVTI SR

údaje, ktoré môže ktokoľvek slobodne používať a následne aj redistribuovať, no s výhradnou podmienkou priznania autorstva a citácie digitálne informácie, ktoré sú k dispozícii kedykoľvek a akémukoľvek používateľovi; ide o dáta alebo obsah, ktoré sú voľne prístupné bez akýchkoľvek

O vysvetlenie pojmu "**výskumné údaje**" sa pokúša viacero definícií. Je to tým ťažšie, že je potrebné nájsť spoločného menovateľa pre mnohé typy údajov (sociologické, ekonomické, medicínske, biologické, vytvorené zo zvukových, obrazových alebo dokonca pachových údajov atď.)

Definícia „**Research data**“ OECD’s (2007):
“Vecné údaje (číselné hodnoty, textové záznamy, obrazové a zvukové materiály), ktoré sa používajú ako primárne zdroje pre vedecký výskum a ktoré sú vo vedeckej komunite všeobecne uznávané ako potrebné na overenie výsledkov výskumu. Súbor výskumných údajov predstavuje systematické, čiastočné zobrazenie skúmaného predmetu. Môžu mať rôzne formy (údaje z experimentov, údaje z pozorovaní, prevádzkové údaje, údaje tretích strán, údaje z verejného sektora atď.)”

Slovník

Printer-friendly version

Dáta – zaznamenané výstupy merania alebo zisťovania, zhromaždené v rámci vedeckého výskumu (často vo forme čísel, znakov, symbolov). Odrážajú stav reality v konkrétnom čase.

Surové dáta (raw data) – nespracované dáta (tak, ako vyšli z meracieho prístroja/experimentu).

Spracovanie dát – zhromažďovanie, triedenie, čistenie, štatistické vyhodnocovanie a ďalšie činnosti, ktorými sa zvyšuje kvalita dát, alebo vďaka ktorým možno z dát získať zmysluplné informácie

Databáza – množina dát organizovaná pomocou logickej schémy, uložená v počítačovom systéme tak, aby z dát bolo možné získavať informácie (napríklad prostredníctvom dopytovacieho jazyka)

Otvorené dáta – dáta voľne, bezplatne, online dostupné na použitie pre ľubovoľných záujemcov, spravidla pod verejnou licenciou (napríklad Creative Commons). Viac o otvorených dátach napr. Open Data handbook (http://opendatahandbook.org/guide/en/what-is-open-data/).

Citlivé dáta – dáta, ktorých strata, zneužitie, pozmenenie alebo neautorizovaný prístup k nim môže viesť k narušeniu súkromia osôb, bezpečnosti, obchodného tajomstva alebo ku škodám na životnom prostredí a biodiverzite.

Veľké dáta (big data) – dáta veľkého objemu, rôznorodé a rýchlo pribúdajúce (viac o nich napr. na stránke: https://www.ecommercebridge.sk/big-data-umela-inteligencia/). Často sa využívajú pri strojovom učení, napr. aplikácia Pl@ntNet má dataset pre strojové učenie určovania rastlinných druhov (viac tu: https://plantnet.org/en/2021/03/30/a-plntnet-dataset-for-machine-learning-researchers/), alebo projekt Monitoring sucha porovnáva reálne dáta od reportérov sucha s vlastným počítačovým modelom.

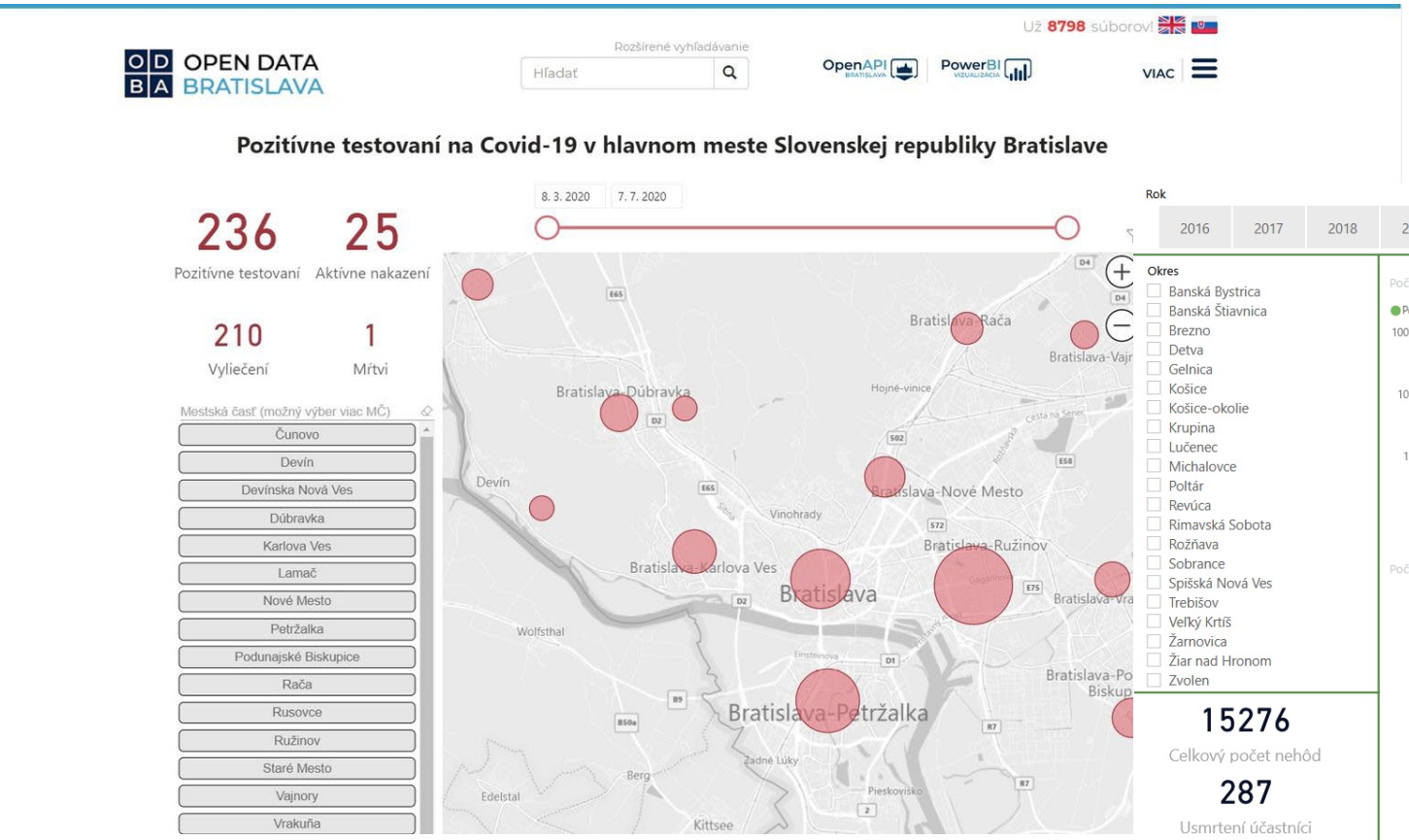
Metadáta – dáta, ktoré vypovedajú o dátach, okolnostiach ich vzniku a využívania (napr. kto ich vytvoril, v akej inštitúcii, kde a kedy boli zozbierané, akou metódou, aký materiál použil, kto má povolenie s dátami pracovať a pod.)

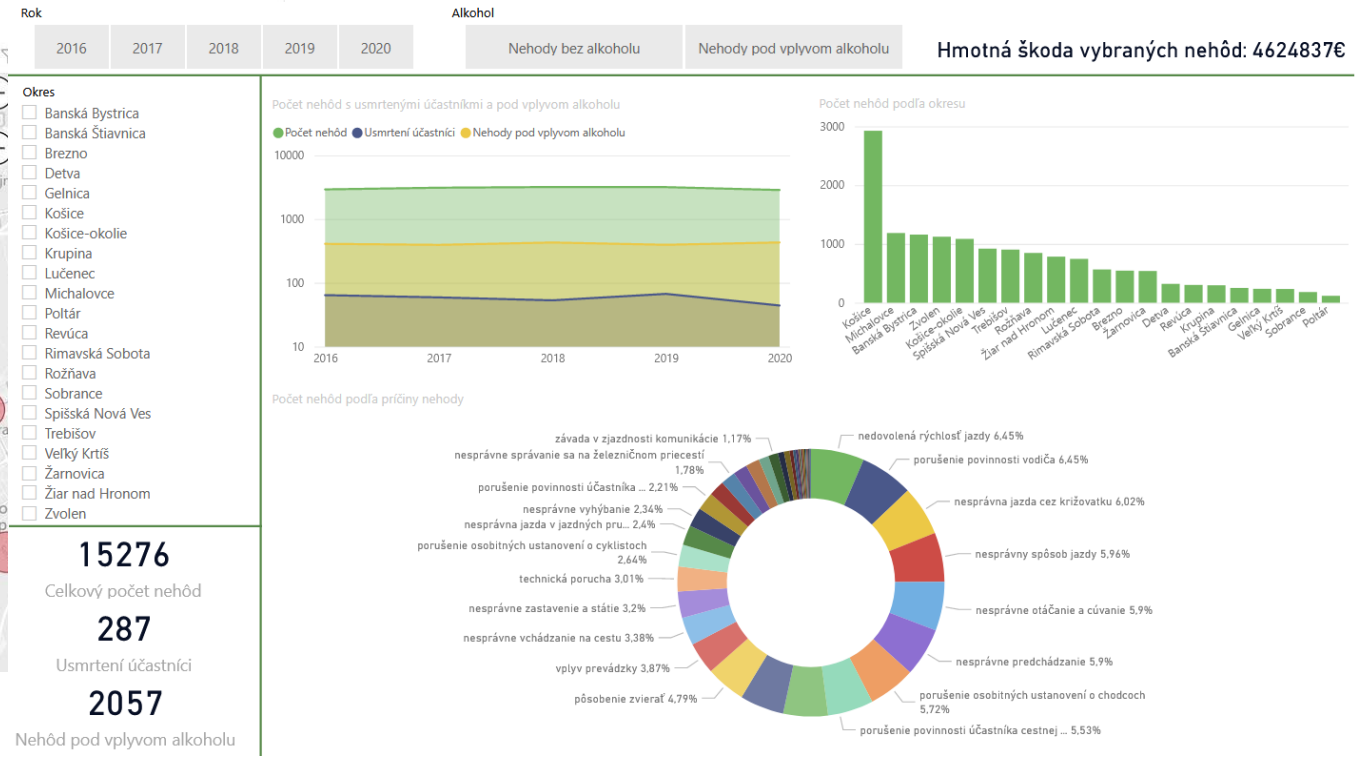
Integrácia dát – kombinovanie údajov z rôznych zdrojov do jedného celku, aby bolo možné získať zmysluplné informácie, alebo jednotný pohľad

Interoperabilita – schopnosť systémov (napr. dátových repozitárov) vzájomne si rozumieť a spolupracovať

Informácie – dáta vložené do kontextu (súvislostí), nadobúdajúce zmysel, takže je možné ich využiť, napríklad pomocou nich odpovedať na výskumné otázky.

Otvorené dáta – príklady





Výskyt koronavírusu v hlavnom meste
Zdroj: reprofoto Opendata.bratislava.sk

Dáta o dopravných nehodách
Zdroj: https://www.alvaria.sk/wp-content/uploads/2022/03/obrazok-23.png

Otvorené dáta/údaje

Otvorené verejné dáta

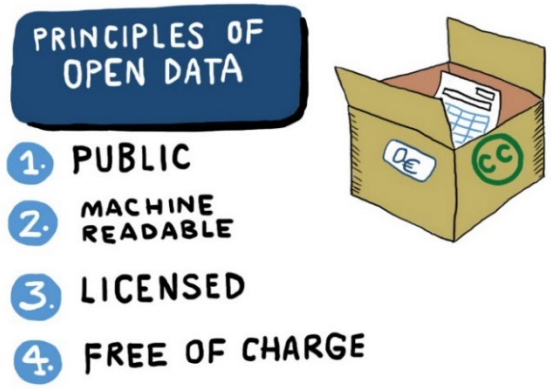
Otvorené administratívne dáta

Otvorené výskumné dáta

Otvorené dáta verejnej správy

OD v neziskovkách a súkrom. sfére

Otvorené dáta vo vede



Čo sú to otvorené dáta (Open Data)?
Otvorené dáta sú informácie alebo údaje voľne a bezplatne dostupné pre každého za rovnakých podmienok, ktoré je možné použiť na akýkoľvek účel komerčného či nekomerčného charakteru. Sú sprístupnené na internete v štruktúrovanej forme, ktorá umožňuje ich hromadné strojové spracovanie.

Čo sú datasety?
Primárnym cieľom Open Data je publikovanie štruktúrovaných datasetov (nielen verejnej správy). Dataset je ucelená a samostatne použiteľná skupina súvisiacich údajov vytvorených a udržiavaných na určitý účel a uložených spoločne podľa rovnakej schémy.

- **Údaje verejnej správy**:

„Open Data“ prístup sa vzťahuje na všetky údaje vytvárané alebo spravované príslušnými organizáciami verejnej správy s výnimkou údajov, ktoré nie je možné publikovať na základe niektorých právnych predpisov (napríklad osobné údaje, utajované skutočnosti a pod.).

- Dátová kancelária MIRRI (**Katalóg otvorených dát**)
- Štatistický úrad (DataCube)

data.gov.sk
ústredný portál verejných služieb ľudom

Vedecké/Výskumné dáta

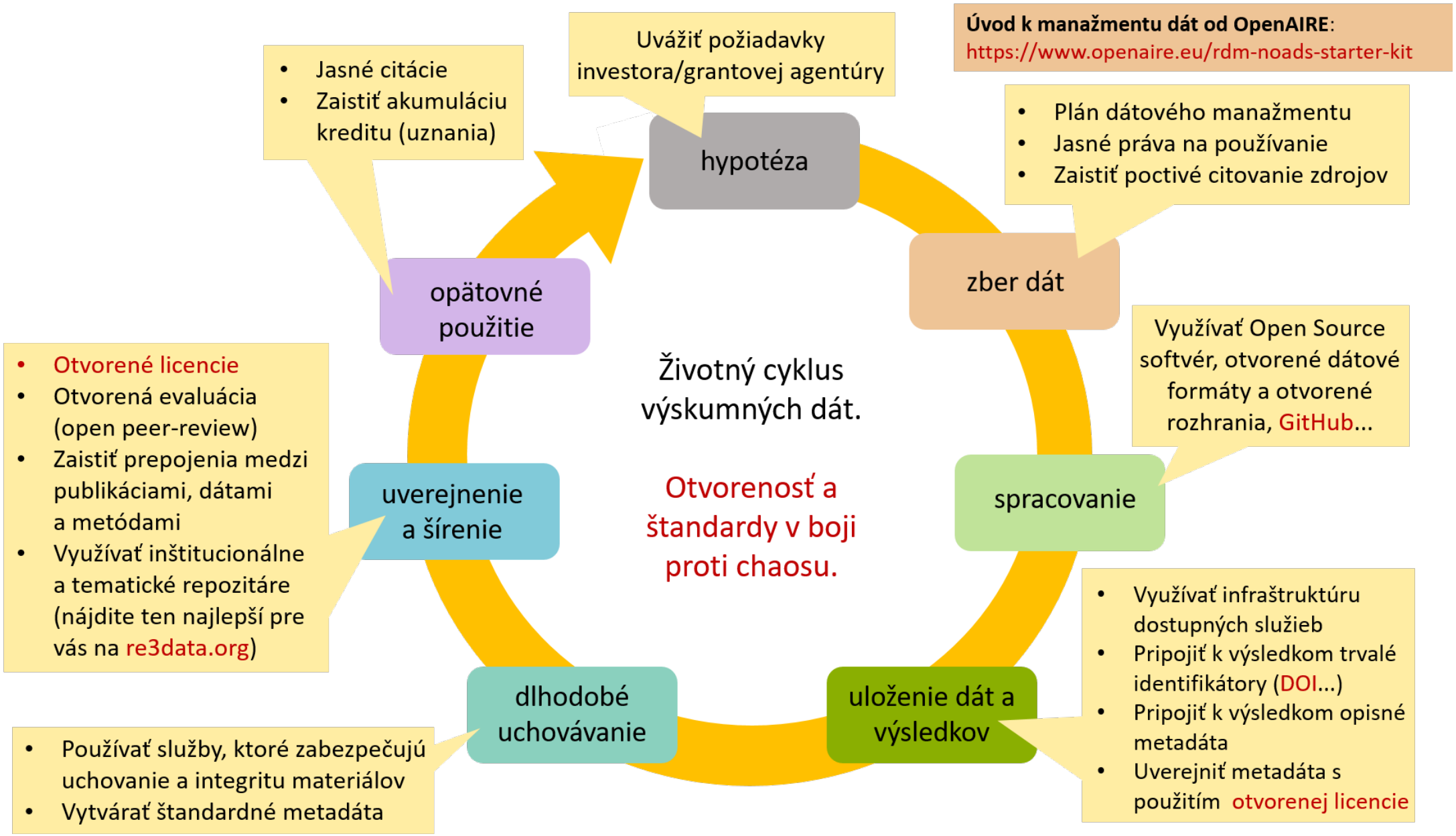
Moderná koncepcia vedeckých dát vznikla v druhej polovici 20. storočia s rozvojom rozsiahlej informačnej infraštruktúry zameranej na spracovanie vedeckých informácií a pozorovaní.

Výskumné dáta sú údaje, ktoré sa používajú ako primárne zdroje pre vedecký výskum.

- Zahŕňajú najmä, ale nie výlučne :
- **primárne dáta**: pôvodné údaje (meranie, text, obraz, zvuk, video, dotazníky atď.) zozbierané alebo vytvorené na účely realizácie výskumného projektu;
- **existujúce dáta** zozbierané alebo skopírované na okamžité alebo budúce použitie v projektoch (najmä administratívne alebo štatistické údaje, digitalizovaný obsah zo zbierok, údaje dostupné v databázach výslovne sprístupnených výskumnej komunite);
- **nové (sekundárne) dáta**, ktoré sú výsledkom spracovania (analýzy, agregácie, transformácie atď.) primárnych dát.

Otvorené výskumné dáta sú dáta, ku ktorým je možné voľne pristupovať, opätovne ich používať, modifikovať a ďalej šíriť na účely akademického výskumu a výučby, ale aj mimo nich. **V ideálnom prípade otvorené dáta nemajú žiadne obmedzenia týkajúce sa opätovného použitia alebo šírenia a sú ako také primerane licencované. Vo výnimočných prípadoch, napr. na ochranu identity ľudských subjektov, sa stanovujú osobitné alebo obmedzené obmedzenia prístupu.**

Otvorené zdieľanie údajov ich vystavuje kontrole, čím sa vytvára základ pre overovanie a reprodukovateľnosť výskumu a otvára cestu k širšej spolupráci. **Otvorené údaje môžu nanajvýš podliehať požiadavke na atribúciu a zdieľanie obdobných údajov (attribute and sharealike).**

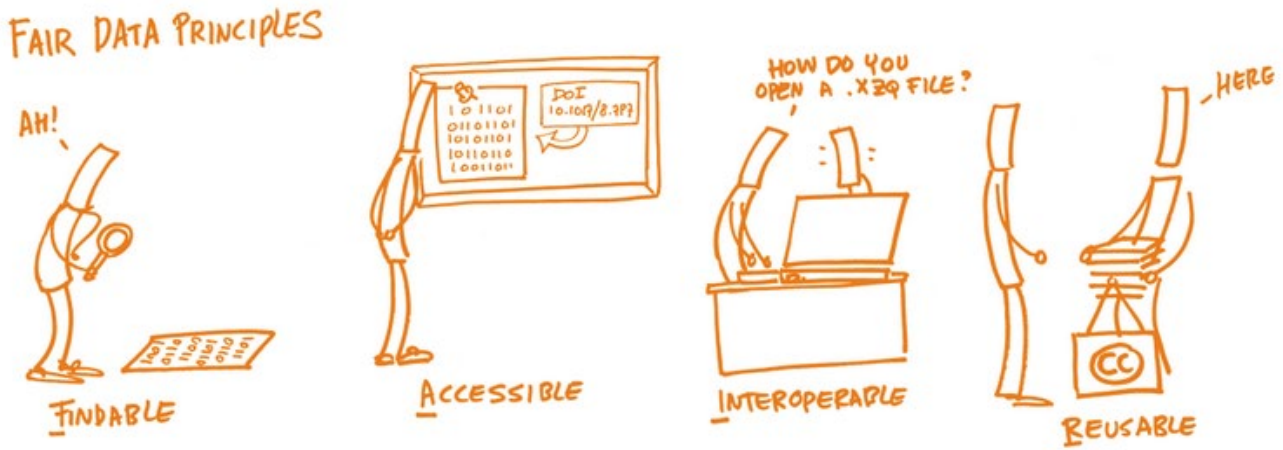

Úvod k manažmentu dát od OpenAIRE:
https://www.openaire.eu/rdm-noads-starter-kit
Uvážiť požiadavky investora/grantovej agentúry
hypotéza
Plán dátového manažmentu
Jasné práva na používanie
Zaistiť poctivé citovanie zdrojov
zber dát
Využívať Open Source softvér, otvorené dátové formáty a otvorené rozhrania, GitHub...
spracovanie
Využívať infraštruktúru dostupných služieb
Pripojiť k výsledkom trvalé identifikátory (DOI...)
Pripojiť k výsledkom opisné metadáta
Uverejniť metadáta s použitím otvorenej licencie
uloženie dát a výsledkov
Používať služby, ktoré zabezpečujú uchovanie a integritu materiálov
Vytvárať štandardné metadáta
dlhodobé uchovávanie
Otvorené licencie
Otvorená evaluácia (open peer-review)
Zaistiť prepojenia medzi publikáciami, dátami a metódami
Využívať inštitucionálne a tematické repozitáre (nájdite ten najlepší pre vás na re3data.org)
uverejnenie a šírenie
Jasné citácie
Zaistiť akumuláciu kreditu (uznania)
opätovné použitie
Životný cyklus výskumných dát.
Otvorenosť a štandardy v boji proti chaosu.

Otvorené vedecké dáta

Za výskumné dáta sa nepovažujú tieto údaje :

- súbory obsahujúce základné prvky publikácie výskumu (texty tvoriace jadro publikácie, ako aj prílohy - tabuľky, grafy, obrázky atď.);
- súbory vytvorené administráciou projektu (vedecké, finančné správy) alebo mediálna komunikácia súvisiaca s projektom.

S cieľom uľahčiť ich publikovanie a/alebo zdieľanie v rámci možného otvoreného prístupu sa výskumné údaje organizujú a spravujú podľa medzinárodných noriem špecifických pre každú oblasť, aby sa dodržiavali **zásady FAIR** (Findable, Accessible, Interoperable, Re-usable).



Zdroj: https://open-science-training-handbook.github.io/Open-Science-Training-Handbook_EN/02OpenScienceBasics/02OpenResearchDataAndMaterials.html

Zdroj https://www.univ-lorraine.fr/en/research-innovation/research-data /

Citlivé dáta

Citlivé údaje – musia byť chránené pred neželaným zverejnením. Prístup k citlivým údajom by mal byť zabezpečený. Ochrana citlivých údajov sa môže vyžadovať z právnych alebo etických dôvodov, z dôvodov týkajúcich sa osobného súkromia alebo z dôvodov vlastníctva.

Tieto osobné údaje sa považujú za „citlivé“ a podliehajú osobitným podmienkam spracovania:

- osobné údaje, ktoré odhaľujú rasový alebo etnický pôvod, politické názory, náboženské alebo filozofické presvedčenie;
- členstvo v odborových zväzoch;
- genetické údaje, biometrické údaje spracúvané len na účely identifikácie ľudskej bytosti;
- zdravotné údaje;
- údaje týkajúce sa sexuálneho života alebo sexuálnej orientácie danej osoby

- EÚ má prísne predpisy týkajúce sa ochrany osobných a citlivých neosobných údajov (napr. všeobecné nariadenie o ochrane údajov - GDPR).
- V programe Horizont 2020 sa pri podávaní žiadosti o grant vyžaduje etické posúdenie
- Anonymizácia a pseudonymizácia citlivých (výskumných) dát

Zdroj obrázku: https://www.fosteropenscience.eu/content/cartoonopen-data

- Čo robiť v prípade, ak nie ste vlastníkom dát?
- Nemôžete publikovať dáta, ktoré ste nezbierali, pokým nemajú licenciu, ktorá umožňuje redistribúciu
- Znovu-využitie dát musí byť zodpovedné a právo na súkromie účastníkov výskumu musí byť dodržané
- Použitie takýchto dát treba správne odcitovať
  - Podmienky podľa použitej licencie
  - Ak dáta nie sú licencované, treba kontaktovať vlastníka dát

Otvorené vedecké dáta

Pochopenie rôznych typov údajov umožňuje používateľom vybrať si ten, ktorý zodpovedá ich potrebám a cieľom.
Pre správnu analýzu dát je rozhodujúce:

- Kvantitatívne – merateľné hodnoty, štruktúrované, štatistické údaje, (vek, výška, počet...)
- Kvalitatívne/ kategorické – triedia sa podľa kategórií (pozorovanie, rozhovor, dotazník, farba vlasov, meno, profesia ....)

Premenná - charakteristika, ktorú možno merať a môže nadobúdať rôzne hodnoty (výška, vek, príjem a národnosť)
Číselné premenné - kvantitatívne
Kategorické premenné – kvalitatívne

Nominálne údaje sa týkajú premenných, ktoré pomenúvajú alebo označujú kategóriu, pozorujú sa, ale nemerajú, pomenúvajú premennú bez použitia akéhokoľvek konkrétneho poradia (počasie, hudobné žánre, farba). Vedci používajú nominálne údaje – na výpočet frekvencií, proporcií a percent.

Ordinálne údaje – určujú poradie, alebo stupnicu, môžu mať číselné hodnoty, časopis: IF pre konkrétny vedný odbor – súbor časopisov

Kvantitatívne údaje sa týkajú premenných s kvantifikovateľnými a číselnými hodnotami. Možno ich objektívne merať.

Intervalové údaje sa týkajú informácií meraných na stupnici s rovnakými vzdialenosťami. časopis: Q1 – Q4, kvartil ako kvalita časopisu

Pomerové údaje sú kvantitatívne údaje, ktoré majú rovnaký a definitívny pomer medzi každou hodnotou. Na rozdiel od intervalových údajov majú
pomerové údaje absolútnu nulu, tj. premenné nemôžu mať záporné hodnoty a nula znamená, že žiadna z týchto premenných nie je
prítomná (výška nemôže mať zápornú hodnotu).

Otvorené vedecké dáta

FAIR princípy

Findable – ľahká dostupnosť strojovo čitateľných dát a ich metadát umožňuje automatické objavovanie súborov dát a služieb.
Accessible - dáta obsahujú jasné informácie o tom, ako k nim pristupovať, prípadne sa vyžaduje overenie alebo autorizácia. Zodpovedajúce metadáta by
mali byť dostupné aj vtedy, ak už dáta nie sú k dispozícii.
Interoperable - dáta sú interoperabilné s aplikáciami (ako je API) alebo s pracovnými postupmi na analýzu, ukladanie a spracovanie.
Reusable - hlavný cieľ FAIR - opätovné použitie dát (dáta musia obsahovať informácie o licencii na používanie, ako aj o pôvode).

Štruktúrované popisné dáta

Povinné polia, dátové schémy a štandardy závisia od vednej disciplíny.
Dôležité doplnky: manuály na používanie softvéru, vysvetlivky skratiek a kódov a pod.
Príklad schém: Data Documentation Initiative
Digital Curation Center, alebo RDA poskytuje prehľad o schémach a štandardoch používaných v mnohých disciplínach

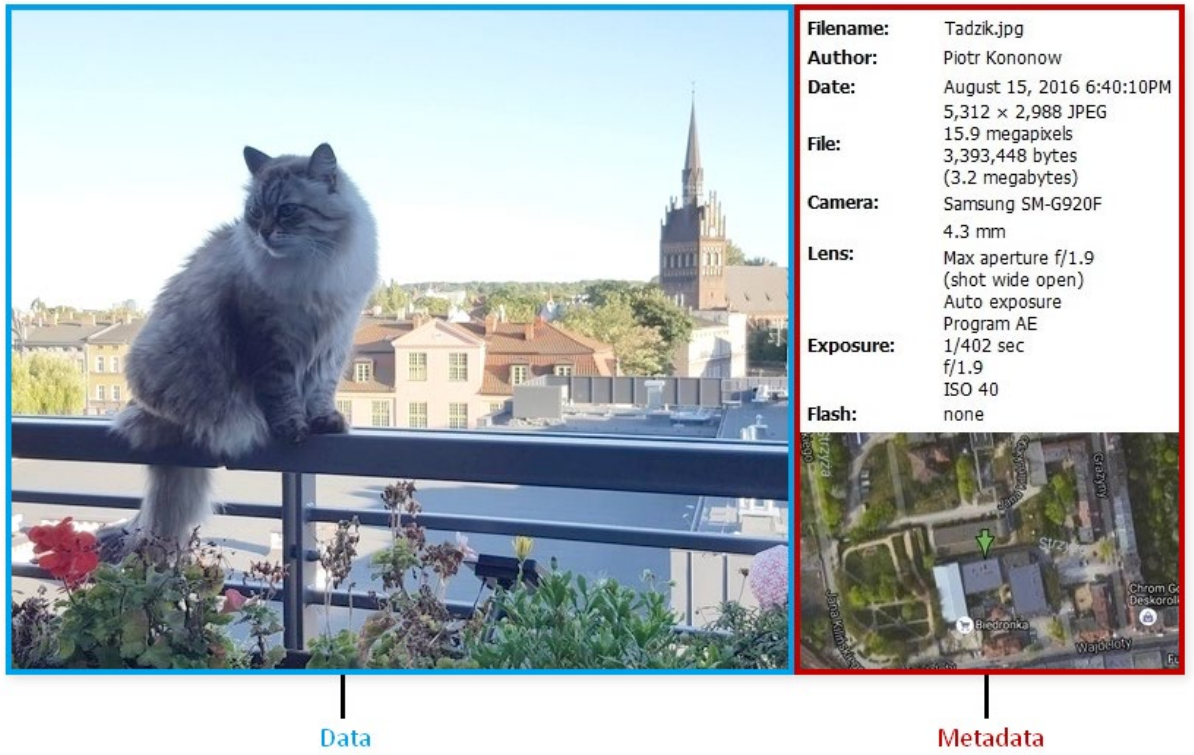
TYP POPISNÝCH DÁT	CIEĽ	PRÍKLAD
Deskriptívne	minimálne potrebné na nájdenie digitálneho objektu	autor, názov, abstrakt, dátum
Štrukturálne	vzťah medzi jednotlivými predmetmi konkrétneho celku	odkazy na súvisiace digitálne objekty (článok – výskumné dáta)
Technické	informácie o technických aspektoch dát	formát dát, použitý hardvér/softvér, verzia, autentifikácia, šifrovanie, štandard
Administratívne	údaje, ktoré sa zameriavajú na používanie (práva) a správu digitálnych objektov.	napr. licencia, možné dôvody embarga, výnimky, sledovanie vyhľadávania a používateľov.

Typy metadát

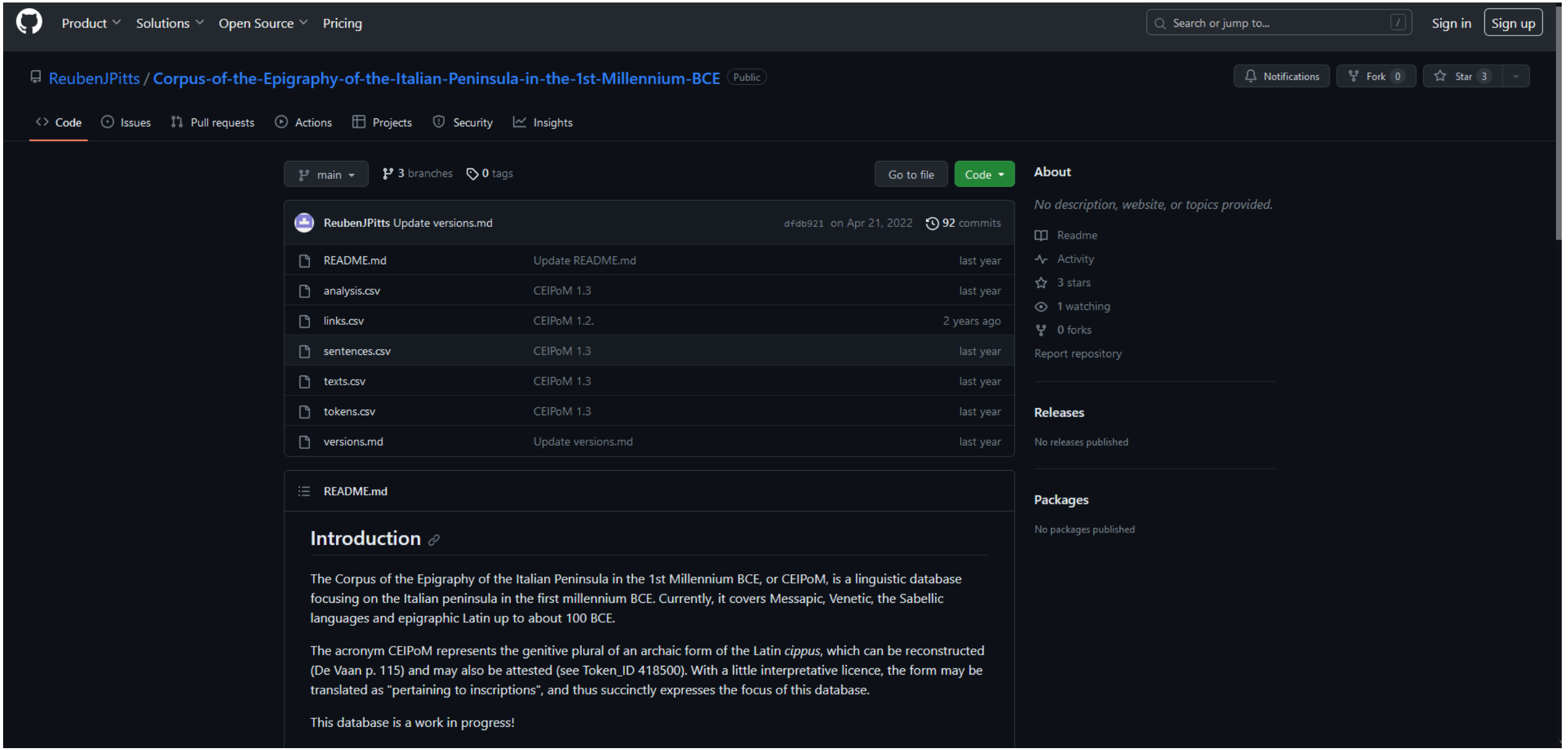
Metadáta môžu byť:

- **Opisné metadáta** – informácie o obsahu a kontexte dát
- *Napríklad:* názov, tvorca, kľúčové slová predmetu a opis (abstrakt)
- **Štrukturálne metadáta** – opisujú fyzickú štruktúru zložených dát
- *Napríklad:* použitý fotoaparát, clona, expozícia, formát súboru a vzťah k iným údajom alebo súborom
- **Administratívne metadáta** – informácie používané na manažment dát
- *Napríklad:* kedy a ako boli vytvorené, kto k nim môže pristupovať, softvér potrebný na ich použitie a autorské práva

- Zdroj: https://guides.lib.unc.edu/metadata/definition



Otvorené vedecké dáta (príklad)



Prepojené dáta / otvorené dáta

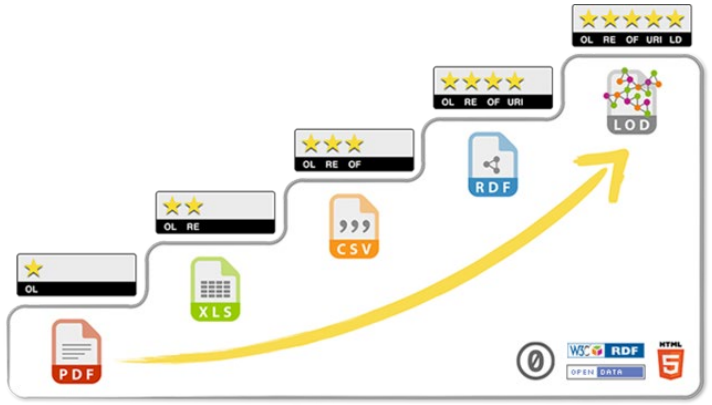
Otvorené dáta - môže voľne používať a distribuovať ktokoľvek (s výhradou požiadavky share-alike)

Otvorené dáta sa však nerovnajú prepojeným dátam. Otvorené údaje môžu byť sprístupnené každému bez prepojenia na iné dáta. Zároveň je možné dáta prepojiť bez toho, aby boli voľne dostupné na opätovné použitie a distribúciu.

komunita W3C vynakladá veľké úsilie na obohatenie cloudu Linked Open Data (LOD), podrobne https://lod-cloud.net/

Linked Open Data/ Linked Data je súbor navrhovaných princípov na zdieľanie strojovo čitateľných prepojených dát na webe. V kombinácii s otvorenými dátami (údajmi, ktoré možno voľne používať a distribuovať) sa nazývajú Linked Open Data (LOD). Prepojenie dát a otvorených dát je možné s použitím tzv. open sources, napríklad: súbor LOD Dbpedia, získavanie štruktúrovaných informácií z Wikipédie
5 stupňov (hviezdičiek) otvorených dát

Prepojené dáta sú jedným zo základných pilierov sémantického webu, známeho aj ako web dát. Sémantický web je o vytváraní prepojení medzi množinami údajov, ktoré sú zrozumiteľné nielen pre ľudí, ale aj pre stroje, a prepojené dáta poskytujú najlepšie postupy na umožnenie týchto prepojení. Umožňujú integrovanie štruktúrovaných dát z rôznych zdrojov.



DBpedia

Prepojené otvorené dáta/ linked open data

Data.europa.eu - pokyny na prepojené otvorené dáta na portáli
podrobne text:
https://data.europa.eu/sites/default/files/d2.1.2_training_module_1.2_introduction_to_linked_data_en_edp.pdf

Share Family – ekosystém prepojených dát. Technológia Share Family využíva LOD Platform, čo je inovatívny rámec vyvinutý špeciálne na konverziu, štruktúrovanie a opakované použitie bibliografických údajov v prepojených otvorených údajoch podľa dátového modelu BIBFRAME .

VÝHODY

- flexibilná integrácia a prepojenie predtým nesúrodej množiny dát
- možnosť zhodnotenia a zlepšenia získaných dát,
- ponuka nových služieb, zníženie nákladov

PODMIENKY ZVEREJNENIA

- štruktúrované dáta (namiesto tabuliek, obrázkov a pod.), jednotné formáty, napr. csv, namiesto Excel,
- používanie jednotných identifikátorov URI – pre rozpoznanie, prelinkovanie súvisiacich dát pre kontext a pochopenie významu dát

Podrobne: https://www.ontotext.com/knowledgehub/fundamentals/linked-data-linked-open-data/

Ako nájsť otvorené výskumné dáta

(Niektoré) užitočné nástroje na vyhľadávanie OVD:

- Google Dataset Search – jednoduchá integrovaná platforma umožňujúca základné vyhľadávanie datasetov podľa názvov, kategórií, kľúčových slov
- FAIRsharing.org – zoznamy dostupných dátových úložísk v oblasti prírodných vied
- Dataverse – open source webová aplikácia na zdieľanie, uchovávanie, citovanie, skúmanie a analýzu výskumných dát. Uľahčuje sprístupňovanie údajov ostatným a umožňuje ľahšie replikovať prácu iných.
- OpenAIRE explore („datasets“): https://explore.openaire.eu/search/find/research-outcomes?type=%22datasets%22

Zdroj: https://authorservices.taylorandfrancis.com/data-sharing/citing-data/