Manažment výskumných dát

kurz

Elektronické informačné zdroje pre vedu - Publikačný poradca

Jitka Dobbersteinová

Centrum vedecko-technických informácií SR

NISPEZ IV

Investícia do Vašej budúcnosti

Európska únia
Európsky fond regionálneho rozvoja

OPERAČNÝ PROGRAM VÝSKUM A INOVÁCIE

MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY

CENTRUM VEDECKO-TECHNICKÝCH INFORMÁCIÍ SR

Manažment výskumných dát

Ciele

Odpovedať na otázky:

- Prečo si výskumné dáta zaslúžia starostlivosť?
- Čo sú princípy FAIR?
- Ako sa robí plán manažmentu dát?
- Ako môžu vedcom pomôcť pri manažmente dát knižnice?

Krátky, ale informatívny dánsky videokurz o význame dátového manažmentu: https://vidensportal.deic.dk/RDMelearn
Danish National Forum for Data Management: doi: 10.11581/dtu:00000048, doi: 10.11581/dtu:00000049, doi: 10.11581/dtu:00000050

Dátová explózia

•**90 % svetových dát sa vyprodukovalo za posledné 2 roky** – vďaka novým nástrojom a metódam vo vede, ako: urýchľovače, teleskopy namierené do vesmíru, družice skúmajúce Zem, sekvenovanie génov, magnetická rezonancia a pod. v biomedicíne (ale aj v iných vedách, napr. archeológii), 3D skenovanie a modelovanie, nástup počítačového spracovania vo všetkom, od environmentalistiky po lingvistiku...

•**80 % z dát, ktoré pribúdajú, nie sú štruktúrované** – sú užitočné len pre tých, ktorí ich získali, ale pri dobrom spracovaní a manažovaní by mohli poslúžiť aj mnohým ďalším.

•Dáta však musia byť:

- ❑ **kvalitné**,
- ❑ dobre **opísané**,
- ❑ dobre **štruktúrované**,
- ❑ dobre **uskladnené** a **prístupné**

(v rámci možností – napr. po anonymizácii osobných údajov, vyriešení licencií a pod.)

Typy výskumných dát

- Dáta plynúce z pozorovaní: dáta získané v reálnom čase, zvyčajne nenahraditeľné
- Experimentálne dáta: laboratórne výskumy (génové sekvencie,chromatogramy...)
- Simulačné dáta: testovacie modely (ekonomické, klimatické,,,)
- Skompilované dáta: text mining, data mining...
- Tzv. referenčné dáta: zbierka menších datasetov (napr. databanky génových sekvencií)

Formy výskumných dát:

- Text, tabuľky
- Laboratórne protokoly,
- Archeologické správy,
- Dotazníky, prepisy,
- Audio/video pásky,
- Obsah databáz,
- Modely, algoritmy, programy,
- Metodológie a pracovné postupy,
- Artefakty,

Korešpondencia...

Grantové žiadosti...

Technické správy...

Findable, Accessible, Interoperable, Reusable

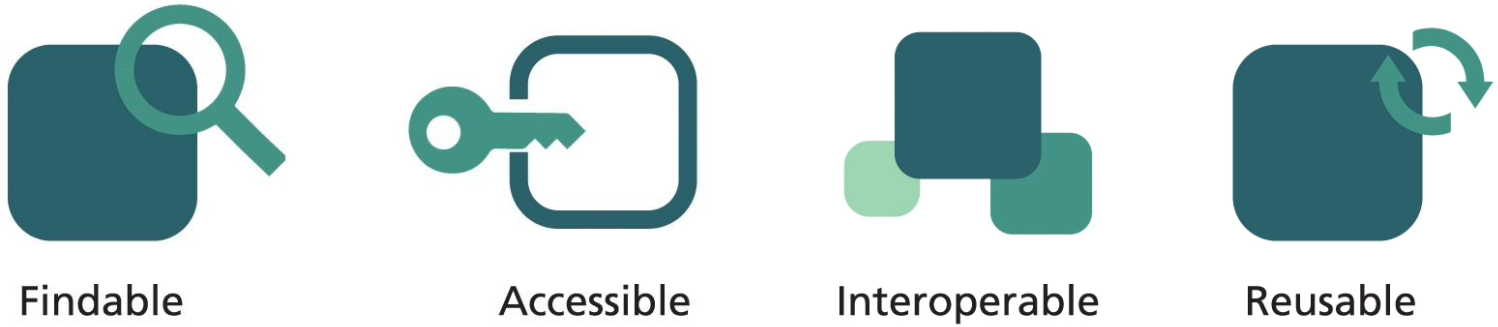
Schéma: Open Access kurz (UNESCO) a CVTI SR (Z. Stožická)

Manažment výskumných dát

Princípy FAIR

Vaše dáta musia byť **FAIR**

FAIR znamená dať Vašim dátam šancu, aby sa v nich mohli vyznať (použiť ich) aj úplne neznámi výskumníci po mnohých rokoch.



Findable — Vyhľadateľné

Accessible — Prístupné

Interoperable — Interoperabilné

Reusable — Opätovne využiteľné

- Princípy FAIR treba dodržiavať **počas celého výskumného cyklu**.
- Keď dáta získavate, myslíte si, že je všetko jasné. Môže sa vám zdať zbytočné zaznamenávať banálne veci o organizácii súborov, významе skratiek a pod. Ale skúste sa na výsledky projektu pozrieť o päť rokov („Čo znamená toto? Ako to, že neviem nájsť x...?“)
- Pozor, **FAIR nerovná sa OPEN data!** Dáta môžu (a mali by) byť FAIR, aj keď nebudú prístupné v plnej podobe celej verejnosti.
- Aj keď dáta z nejakého dôvodu (osobné údaje, práva priemyselného vlastníctva...) nemôžu byť zdieľané otvorene, môžete **vytvoriť opis dát a zdieľať ho** verejne, aby tí, pre ktorých by dáta mohli byť užitočné, mohli podľa pravidiel požiadať o prístup k dátam.

Pôvod FAIR: stretnutie na univerzite v Leidene v 2014
V roku 2016 boli princípy publikované v článku Wilkinson et al. 2016: https://doi.org/10.1038/sdata.2016.18

Schéma: Martínez-Lavanchy et al. 2019, doi: 10.11581/dtu:00000049

Manažment výskumných dát

pixabay.com

Manažment výskumných dát

Aplikácia princípov FAIR berie do úvahy:

- **Dokumentácia** – poskytuje kontext a robí dáta zrozumiteľné aj pre ostatných vedcov.
- **Metadáta** – štruktúrované dáta o dátach, zvyšujú objaviteľnosť dát a uľahčujú ich strojové spracovanie
- **Formáty dát** – dôležité najmä keď treba kombinovať viac súborov dát z rôznych zdrojov
- **Prístupové práva k dátam** – kto a za akých podmienok môže k dátam pristupovať (musí sa určiť vopred, napr. v informovaných súhlasoch pacientov v rámci klinických štúdií musí byť vopred špecifikované, za akým účelom a za akých podmienok sa dáta použijú, čo z nich sa môže zverejniť a čo nie)
- **Trvalé identifikátory** – DOI sa môže použiť aj na vedecké dáta, ktoré budú vďaka nemu ľahko vyhľadateľné. Okrem DOI ale existujú aj identifikátory fyzických vzoriek, organizácií a mnohých iných vecí vo výskume (tie môžeme použiť pri tvorbe dokumentácie k dátam).

Manažment výskumných dát

Dokumentácia

• Na začiatku projektu sa dohodnúť s kolegami na budúcej štruktúre dát (ako sa budú zapisovať, do akých adresárov sa bude čo dávať, ako sa budú nazývať súbory...)

• Dokumentácia by mala obsahovať všetko potrebné o zbere dát...
  - **metódy**
  - **nástroje**
  - **softvér**

• ... a všetky relevantné záznamy počas celého výskumného procesu:
  - **kto s dátami pracoval**
  - **ako sa spracovávali**
  - **ako súvisia** s inými súbormi dát alebo publikáciami

Manažment výskumných dát

Metadáta

Dublin Core Metadata Initiative

• Uznávaným štandardom pre metadáta je Dublin Core

• Rôzne vedné odbory a oblasti ľudského záujmu majú vlastné metadátové štandardy, napr.:

- Darwin Core – pre geografický výskyt druhov
- EML – Ecological Metadata Language pre ekológiu
- VRA Core – Visual Resources Association pre vizuálne umenie
- V behaviorálnych, spoločenských a ekonomických vedách sa často používa Data Documentation Initiative (štandard pre prieskumy, dotazníky, štatistické súbory)...

• Ak štandardy vo Vašej oblasti neexistujú, spíšte do zvláštneho súboru všetky informácie, ktoré považujete za potrebné pre porozumenie daným dátam.

DDI®
DATA DOCUMENTATION INITIATIVE

FINDABLE, INTEROPERABLE, REUSABLE

Manažment výskumných dát

Formáty dát

• Je dôležité používať formáty súborov zaužívané vo Vašom vednom odbore

• Stáva sa, že výrobcovia prístrojov majú vlastné softvéry spracovávajúce výstupy z prístrojov vo vlastných formátoch (ktoré výskumníci s prístrojmi od inej firmy nemusia vedieť otvoriť)

• **Ideálny formát je:**
  - neproprietárny (nie je krytý patentom či copyrightom firmy)
  - nešifrovaný
  - nekomprimovaný
  - bežne používaný výskumnou komunitou
  - v súlade s otvorenými zdokumentovanými štandardmi

Dobrá prax pre formáty súborov podľa Stanfordskej knižnice: https://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-formats

- Metasúbory (containers): TAR, GZIP, ZIP
- Databázy: XML, CSV
- Zemepisné: SHP, DBF, GeoTIFF, NetCDF
- Video: MOV, MPEG, AVI, MXF
- Zvuk: WAVE, AIFF, MP3, MXF
- Štatistika: ASCII, DTA, POR, SAS, SAV
- Obrázky: TIFF, JPEG 2000, PDF, PNG, GIF, BMP
- Tabuľkové údaje: CSV
- Text: XML, PDF/A, HTML, ASCII, UTF-8
- Webové archívy: WARC

INTEROPARABLE, REUSABLE

Manažment výskumných dát

Prístupové práva k dátam

• Niektoré dáta sú citlivé a nie je žiaduce, aby ich mohol vidieť hocikto (či už kvôli právam pacientov, alebo právam priemyselného vlastníctva)

• V takom prípade je dobré **zverejniť metadáta**, aby mohol o prístup k dátam požiadať výskumník, ktorý na to má relevantný dôvod

• Je kľúčové zverejniť aj **podmienky**, za akých môže byť udelený prístup a **kontakt** na oprávnenú osobu, ktorá môže prístup udeliť.

•...dosť často sa však stáva, že sa vedci bránia myšlienke zdieľania dát len z nezvyku, alebo nešpecifických obáv.

Manažment výskumných dát

Licencie pre otvorené dáta

- Najpoužívanejšie licencie sú Creative Commons, napr. CC-BY, CC-BY-SA

- Open Source licencia pre otvorené dáta v IT (softwar, skripty, programy): GPL (General Public Licence)

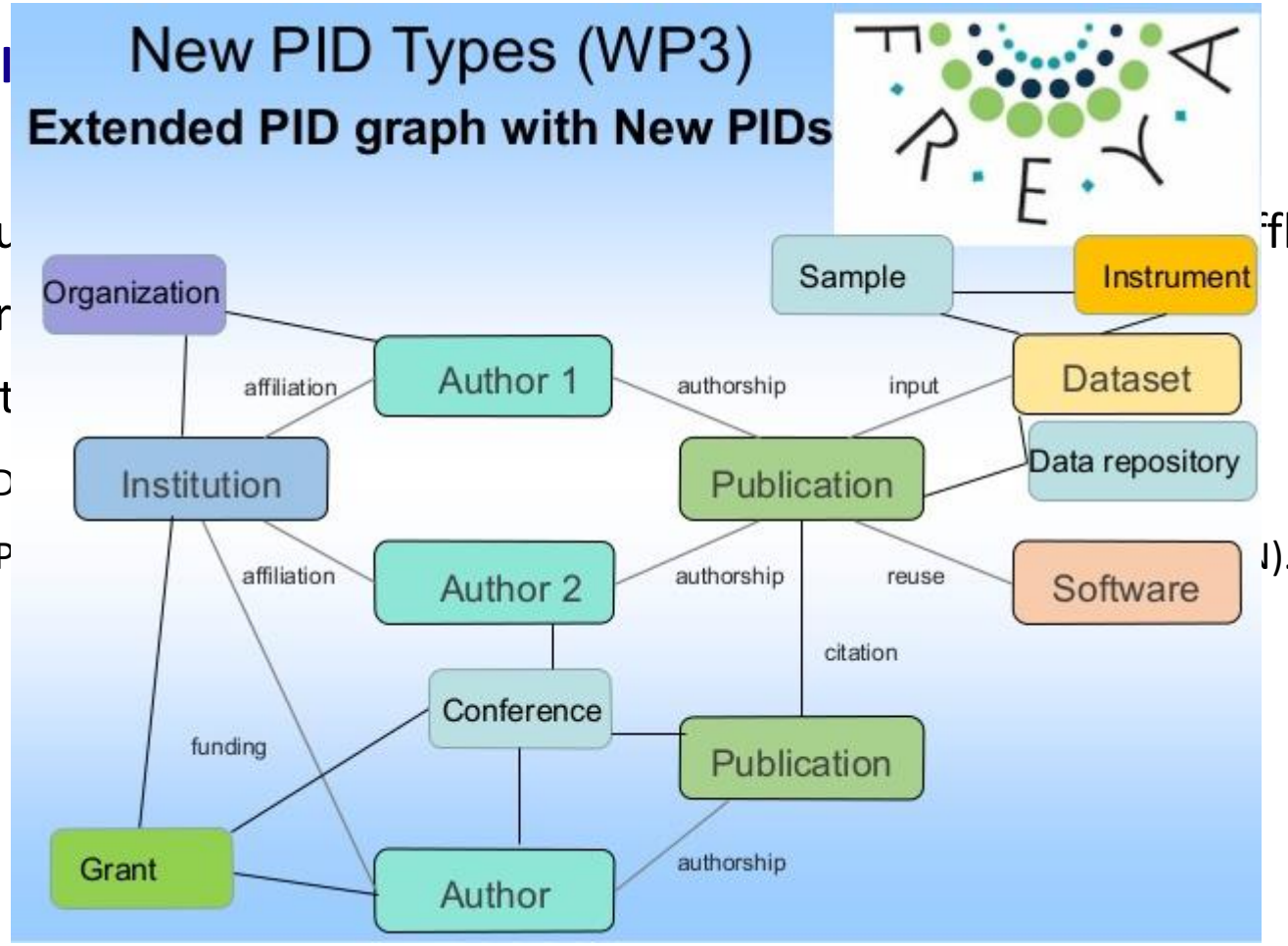
- Aj metadáta majú mať svoju licenciu (dosť často to býva CC0 – public domain)

- https://citeas.org/

Jednoduchým spôsobom ako zdieľať dáta je zverejniť ich v **repozitári** (dátovom, odborovom, inštitucionálnom, alebo všeobecnom ako napr. **Zenodo**, ktoré už pri vkladaní obsahu dáva užívateľom návod ako postupovať v súlade s princípmi FAIR. Poskytuje aj DOI a licencovanie podľa voľby vkladajúceho a zároveň sa stará o dlhodobé uchovávanie).

Manažment výskumných dát

Trval



- pou ffline)
- udr
- Inšt

D

P)...

iniciatíva FREYA

FINDABLE, ACCESSIBLE, INTEROPERABLE, REUSABLE

Manažment výskumných dát

F	A	I	R
Metadáta Trvalý identifikátor	Kto a ako má prístup	metadátové štandardy štadardné formáty dát	Presná dokumentácia

Manažment výskumných dát

DATA MANAGEMENT PLAN – Plán manažovania dát

Manažment výskumných dát

Plán manažmentu dát – Data Management Plan (DMP)

- Viac než formalita – DMP je nástroj, ktorý umožňuje vedcovi
  - systematicky premýšľať o dátovom manažmente
  - robiť strategické rozhodnutia v správnom čase
  - ujasniť si so spolupracovníkmi očakávania ohľadom projektu
  - zjednotiť si so spolupracovníkmi postupy a spôsob dokumentácie.
- O dátach treba rozmýšľať (a rozhodovať sa) PRED, POČAS aj PO uskutočnení výskumu. DMP sa vypracováva na úplnom začiatku, ale v priebehu výskumu sa môže modifikovať podľa aktuálneho vývoja situácie.
- Čoraz viac sa razí politika povinných DMP (ako súčasť žiadosti o grant)
- Niektoré grantové agentúry a donori (napr. ERC) mávajú vzory DMP, do ktorých výskumník vyplní údaje špecifické pre svoj projekt.
- Výskumné inštitúcie vo vyspelých krajinách už majú vo vnútorných predpisoch povinnosť vypracovávať DMP
- Existujú v nich aj pracovné pozície „podpora manažmentu dát“.

Manažment výskumných dát

Na aké otázky odpovedáte v pláne manažmentu dát

- Aký druh dát budete zbierať počas výskumu? (dotazníky, merania, texty, fotografie, skeny...)
- Budú medzi nimi špeciálne, citlivé kategórie dát? (osobné údaje, biologické vzorky, komerčné údaje, ktoré by mohli ovplyvniť obchodnú súťaž...)
- Musíte pred začatím projektu žiadať o súhlas (osoby – napr. pacientov, alebo organizácie, napr. Štátnu veterinárnu a potravinovú správu, Štátnu ochranu prírody, Etickú komisiu...)?
- Potrebujete pred začatím projektu splniť nejaké právne povinnosti?
- Ste si vedomí práv a povinností pri používaní dát iných strán (napr. z komerčných databáz – niektoré môžu mať v zmluvách zákaz zdieľania podkladových dát po ukončení projektu, alebo obmedzenie objemu dát, ktoré môže výskumník extrahovať)?
- V akých formátoch budete dáta ukladať?
- Aký je predpokladaný objem získaných dát?
- Máte stratégiu pomenovávania súborov a adresárov?

Manažment výskumných dát

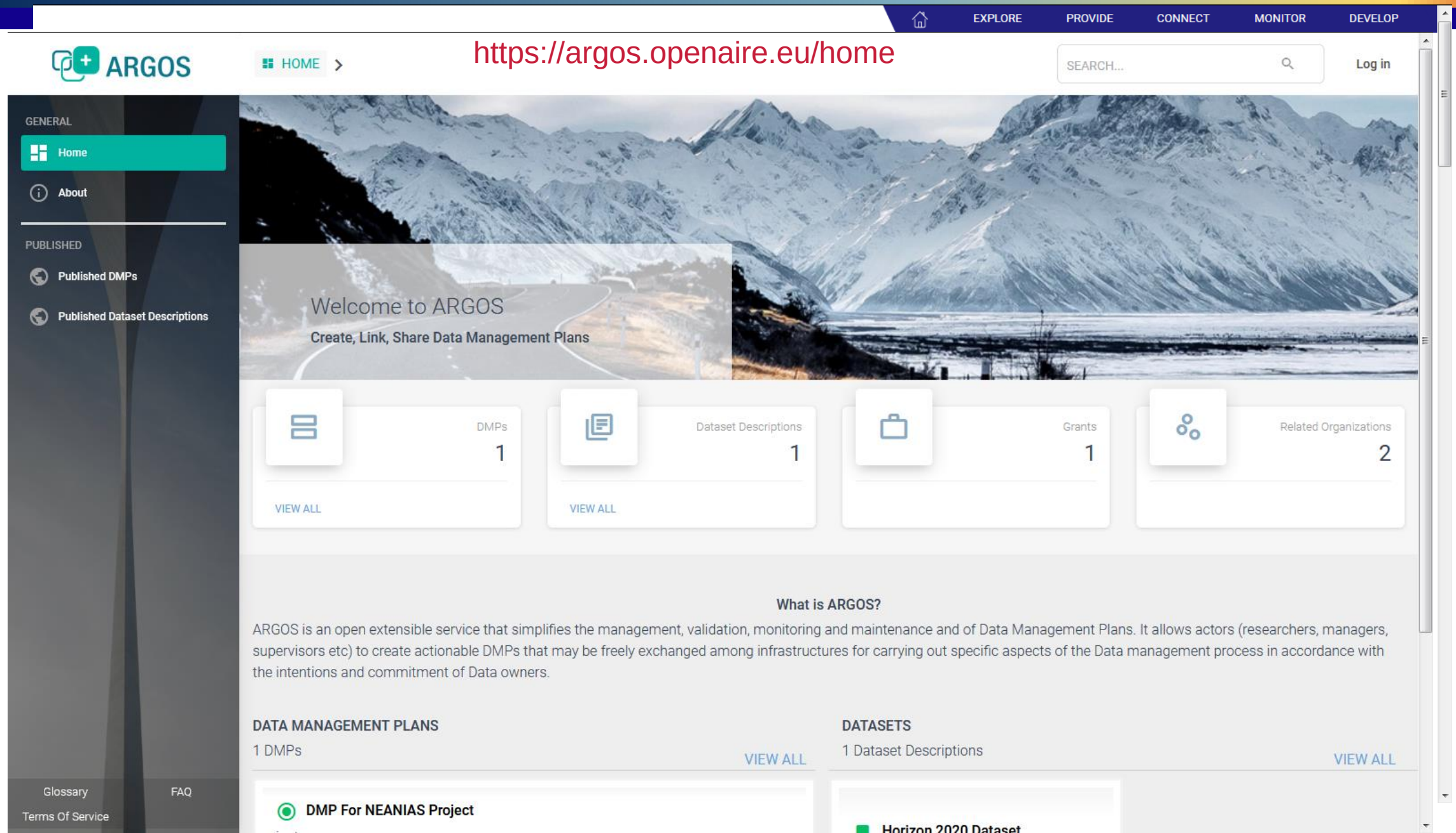
Na aké otázky odpovedáte v pláne manažmentu dát

- Ako budete dáta dokumentovať, opisovať? Metadát je veľa – ale ktoré budú kľúčové pre váš projekt?
- Ako zabezpečíte, že budú vaše dáta zrozumiteľné pre iných?
- Ako zabezpečíte svoje dáta (aby k nim nemali prístup neoprávnené osoby – napr. dáta sú chránené heslom na univerzitnom serveri, citlivé dáta v papierovej podobe zamknuté v skrinke)?
- Ako zaistíte zálohovanie dát?
- Akým spôsobom budete zdieľať dáta, ktoré majú byť zverejnené (napr. v repozitári)?
- Ako zabezpečíte dlhodobé uchovávanie? Ako dlho (napr. minimum 5 rokov pre dáta, ktoré sú podkladom pre publikácie)?
- Ktoré z množstva získaných dát majú byť uchovávané dlhodobo a ktoré majú byť zničené/vymazané?

Pomoc pri vytváraní DMP:
Pokyny pre data management pri projektoch **Horizont 2020**:
https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
Webový nástroj **DMP online**
https://dmponline.dcc.ac.uk/

Manažment výskumných dát

https://argos.openaire.eu/home



Findable, Accessible, Interoperable, Reusable

Schéma: Open Access kurz (UNESCO) a CVTI SR (Z. Stožická)

Manažment výskumných dát

Manažment výskumných dát: ako môžu pomôcť knihovníci?

- Hypotéza - plánovanie (DMP)
- Zber dát (rešeršné služby)
- Spracovanie a analýza – asistenčné služby, školenia
- Zverejňovanie a zdieľanie (repozitáre, licencie)
- Uchovávanie (persistentné identifikátory)
- Opätovné využívanie (správne citovanie dát)

Ako zodpovedný manažment dát prospieva vedcom a vede

- Zdieľanie dát zvyšuje **dosah** výskumu a **prestíž** výskumníka aj výskumnej organizácie, pre ktorú pracuje.
- Zdieľanie dát zvyšuje ekonomickú efektivitu výskumu (pridaná hodnota opätovného zdieľania + možnosť zabrániť zbytočnému duplikovaniu).
- Zdieľanie dát tiež prispieva k **transparentnosti a reprodukovateľnosti** výskumu a pôsobí ako prevencia „sloppy science“ (nedbalej vedy) , pretože si podkladové údaje môže každý skontrolovať (recenzent v rámci recenzného konania, iný vedec z danej oblasti v rámci prieskumu literatúry, aj spätne po rokoch) a vidieť, čo s nimi výskumník robil.

Ďakujem!

Otázky?