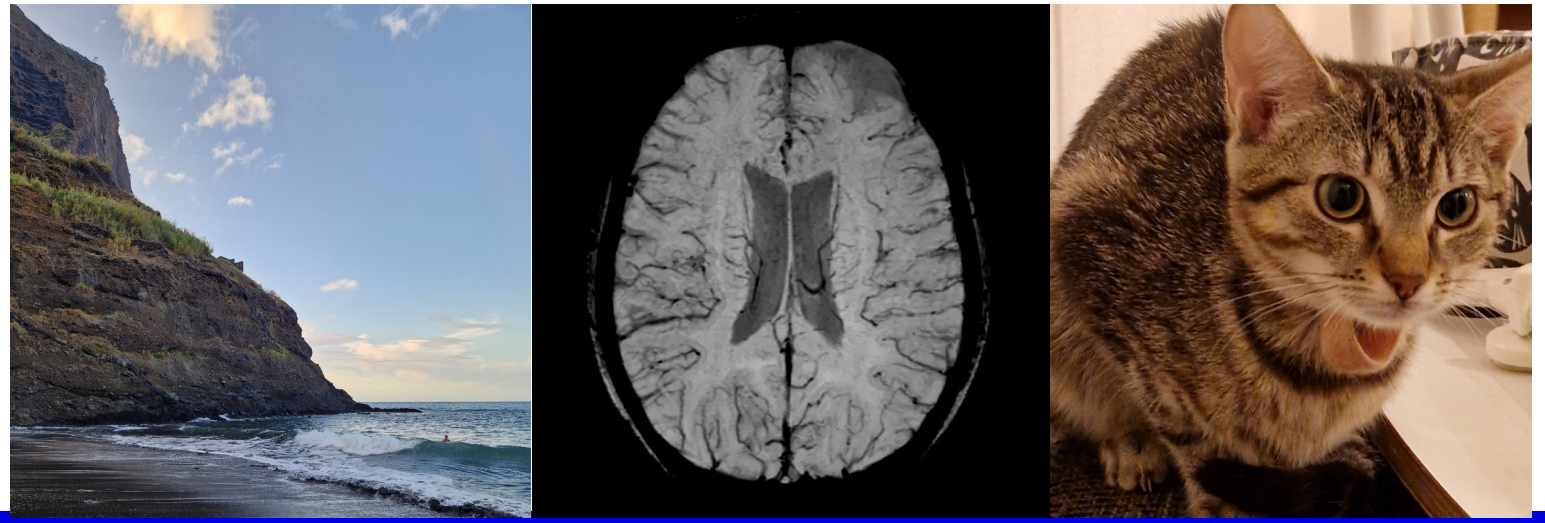
MUNI
ICS

Maj data bez metadat smysl?

Adrián Rošinec

Co jsou data?

- Súbor informácií, faktov
- Reprezentované v štrukturovanej alebo neštrukturovanej podobe
- Napr. obrázok, zvuková stopa, tabuľky, ...



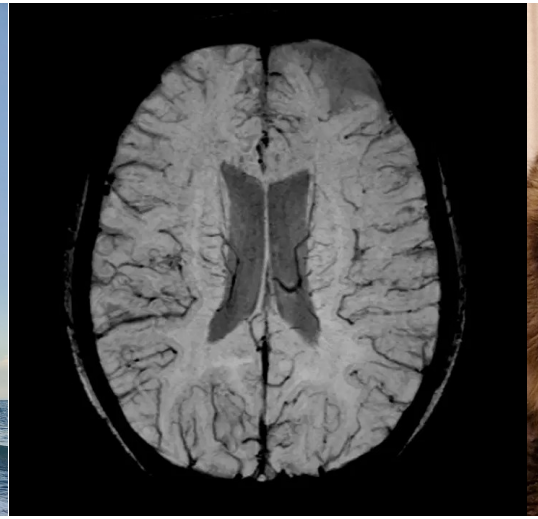
A metadata?

- Doprovodné informácie
- Pomáhajú pochopiť kontext a detaily datasetu
  - Dátum vytvorenia
  - Veľkosť
  - Lokalita
  - Typ/formát

Creation: 07.09.2023

Device: Samsung SM-S906B
f/1.8 1/320 ISO 50 No-Flash

Location: Praia do Faial, Faial, PT



Creation: 17.04.2019

Gender: M

Handedness: Left

SliceLocation: 40.115494018811

SliceThickness: 1.5

EchoTime: 3.13

NumberOfAverages: 4

Creation: 19.02.2023

Device: Samsung SM-S906B
f/1.8 1/25sec ISO 640 No-Flash

Location: Kavárna Pelíšek, tř. Kpt. Jaroše, Brno, CZ

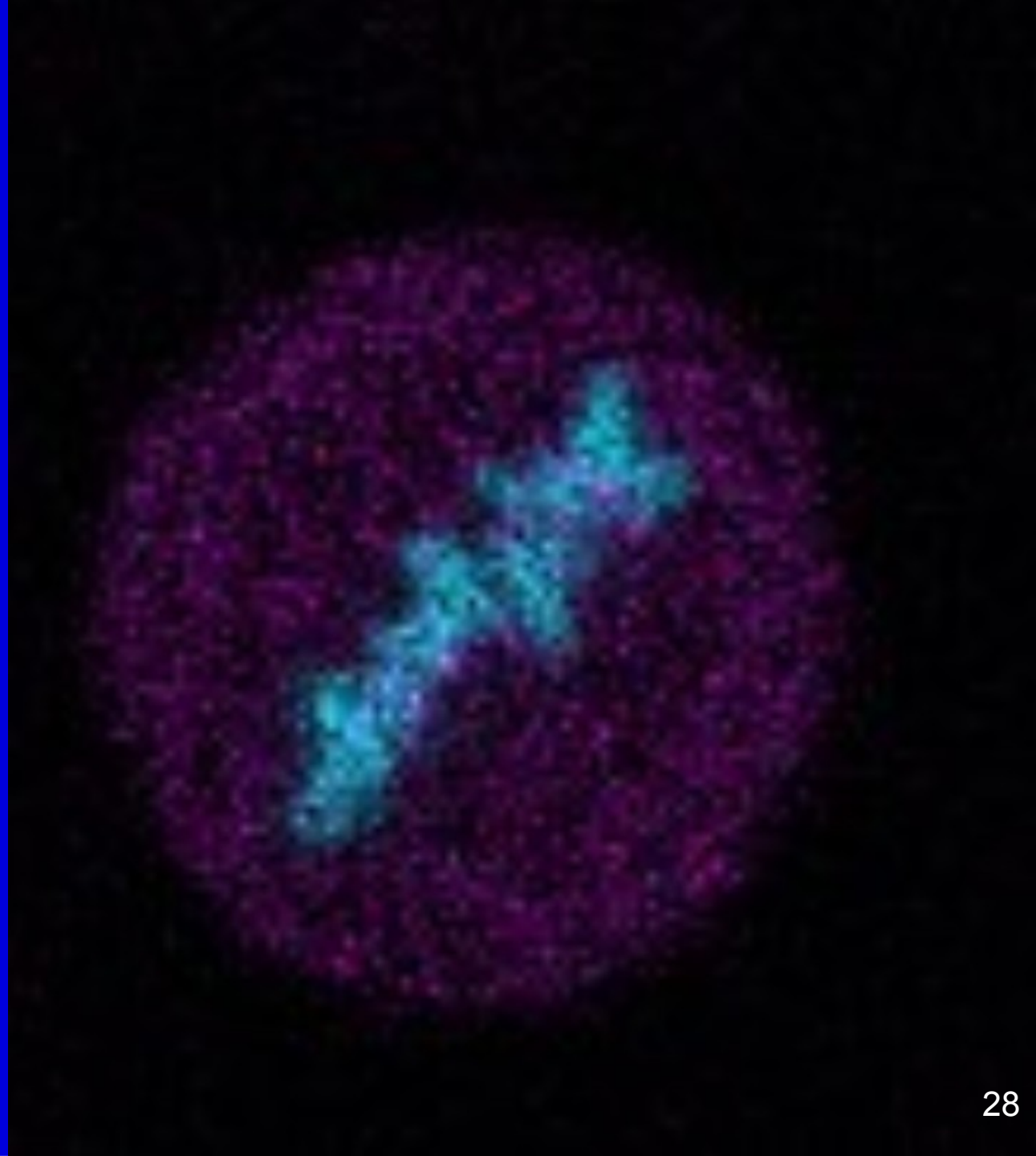
Prečo chceme anotované datasety?

- Zlepšenie prístupu k dátam
  - vyhľadávanie, filtrácia, kategorizácia, identifikácia datasetov
- Pomáha pochopiť kontext vzniku datasetu
  - Ako a prečo dataset vznikol
  - Aká metóda bola využitá pre získanie dat
    - Prístroj (spektrometer/mikroskop/MRI) a jeho parametre
  - Zdroj dát
    - Pacient a jeho diagnóza
- Podporuje interoperabilitu
  - Umožnuje výmena datasetov a využitie inými vyskumníkmi
- Reprodukovatelnosť
- Provenance
- Licensing

Prečo chceme anotované datasety?

- Umožňuje nám kontrolovať kvalitu datasetov, prípadne vedeckých výstupov v čase
  - Vďaka agregáciám a štatistickým prehľadom

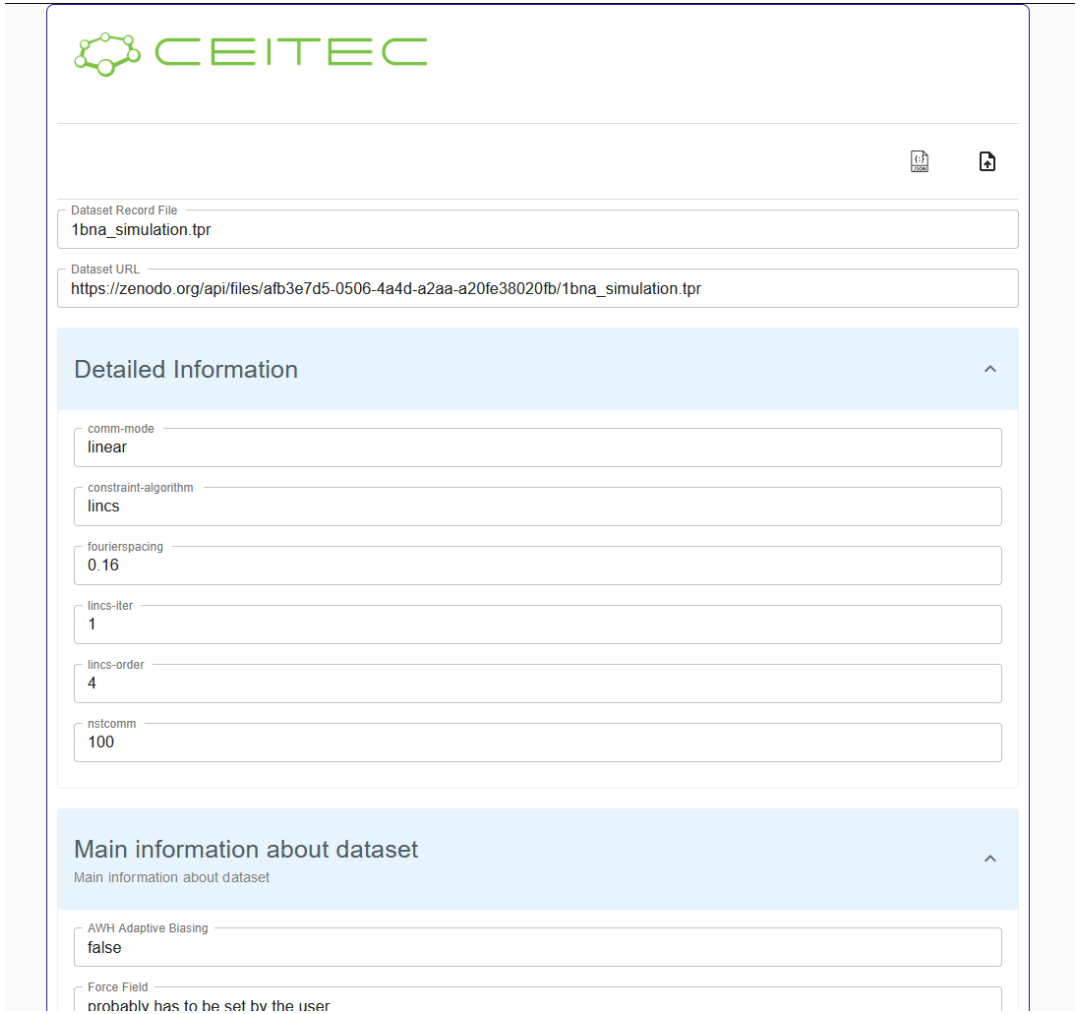
Príklad z praxe



Ako získavame metadáta

- Manuálne anotácie
  - Veľmi pracne s vysokou mierou chybovosti, populárne
  - Cieľ „spríjemňovať“ život anotátorom našeptávačmi, validátormi, rozpoznaním kontextu
- Automatické anotácie
  - Veľmi častý základ sú metadáta z nastavenia prístroja, administratívne metadáta (kto dataset vyrobil), ...
  - Adaptácia analytických nástrojov pre výstup do metadatového súboru
  - Zo štrukturovaných/neštrukturovaných dát potreba vyberať dôležité údaje
    - parametre simulácie, rozpoznávanie v obrázkoch, ...

Editor metadát



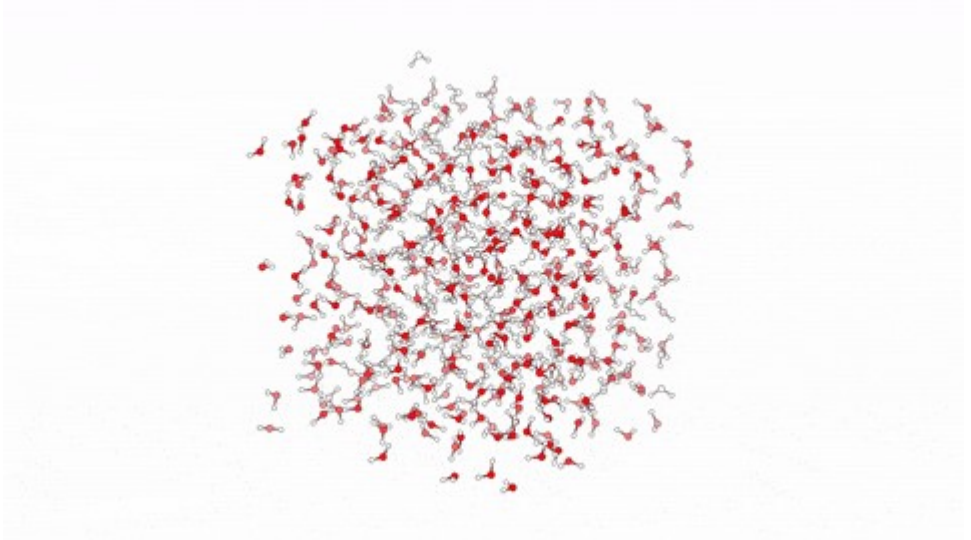
- „Klikátko“
- Skvelé pre manuálne anotovanie datasetu
- Validácie oproti JSON schema
- Potreba rozšírenie o referencie/dependency

Definície metadata schémat

- Snaha definovať/adopovať metadatové schémy pre publikované datasety
  - CEITEC CFs + spolupracujúca komunita
- Pre simulácie molekulovej dynamiky
  - Róbert Vácha a „Molecular Simulations and Design research group“ z Max Planck Insititute
- Formálna schéma ale aj sémantika anotácií

Modifikácia analytických nástrojov

- Gromacs
- Analytické/simulačné nástroje modifikovať aby výstup experimentov obsahoval aj metadatové popisy
  - Parametre spustenia, referencia na vstupné dáta



Metadatový katalog

- Cieľom je budovať metadatové katalogy
  - Pre odborné komunity / vyskumné tímy
- Možnosť jednoducho prehľadávať datasety
  - Zjednodušiť orientáciu v publikovaných ale aj vlastných datasetoch
- Poskytnúť nástroj pre prípravu datasetu pre publikáciu
  - Anotovanie datasetu
  - Validácia schémy
  - FAIR checking

Majú data bez metadat zmysel?

- Jo.
  - V pripade, že ich je veľa, zdieľajú sa, ťažko sa v nich orientuje

MUNI