cesnet

TA DATA NEJSOU FAIR!

David Antoš

CESNET

21. 5. 2019

V pěti kapitolách

- Potíže s vědeckými daty
- Data Management Pretending
- FAIR
- Technická implementace FAIR
- Diskuse

Kapitola I
Potíže s vědeckými daty

Máme tady problém

- s ukládáním dat jsou samé potíže
  - relativně malá životnost technologie
  - nevalná odolnost proti poškození
  - špatná kontrolovatelnost poškození
- ale také
  - jak data najít
  - jak je popsat
  - jak je použít, když už je najdeme

O čem se mluví dlouho I

- zaklínadlo knihovnické komunity:
  Long Term Preservation
  - sada procesů a technických opatření k zachování obsahu informace po dlouhou dobu
    1. střednědobě: uchování binárního obsahu (mnoho kopií, kontrolní součty, pravidelné kontroly, . . .)
    2. dlouhodobě: konverze formátů dat do modernějších (včetně kontrol, že se neztratila informace [což se musí ad hoc])
  - metadata: několik standardů pro popis publikací
    - Dublin Core, MARC

O čem se mluví dlouho II

- problém hledání dat je zde „snadný“
  - hledá se publikace
- pláč knihovnické komunity:
  „nikoho LTP nezajímá a nechce na to dát peníze“
- politicky nekorektní pohled:
  je tohle snad o něčem jiném než penězích?
  - D.A. tvrdí, že téměř ne
  - potřebujete technologii (drahou) a kurátora (drahého)
  - procesy jsou jasné
  - Nějaké otevřené principielní otázky v LTP? Opravdu? A kde?

Doba datová I

- s nárůstem vědeckých dat se problém posouvá
- na škále
  exaktní vědy – živá příroda – humanitní a čirá esoterika
  - „zlatí fyzikové, ti dopředu odhadují, co dokáží zpracovat“
    - i urychlovače si podle toho postavili
  - méně IT-pozitivní obory nemívaly mnoho dat
    - a je to pro ně revoluce
    - a kulturní šok

Doba datová II

- zejména ve vědách o živé přírodě (Life Sciences)
  - publikace jako forma vědecké komunikace ztrácí význam proti datům
  - a dat začalo být zatraceně mnoho
  - konfokální mikroskopy, DNA sekvenátory, . . .
- životní cyklus dat „pořídit – vytěžit – zahodit“ přestal stačit
  - chtěli bychom i „archivovat“
    - nejlépe vše a věčně
- n.b. dnes už je problém i se samotným „pořídit a udržet aspoň do vytěžení“

Doba datová III

- běžná situace
  - doktorský student provozuje databázi
    - na PC v labu
    - pod stolem
    - bez záloh
  - jak chcete její záznam citovat?
  - kdy si někdo vzpomene na PC pod stolem?
    - poté, co student odešel?
    - o dva roky později?
    - až umře disk?
  - „tu aplikaci programoval bývalý manžel paní docentky a běží to tady na tom sedm let starém desktopu na Win XP, máme tam všechno“ #TheBellyOfMUNI #RealStory

cesnet

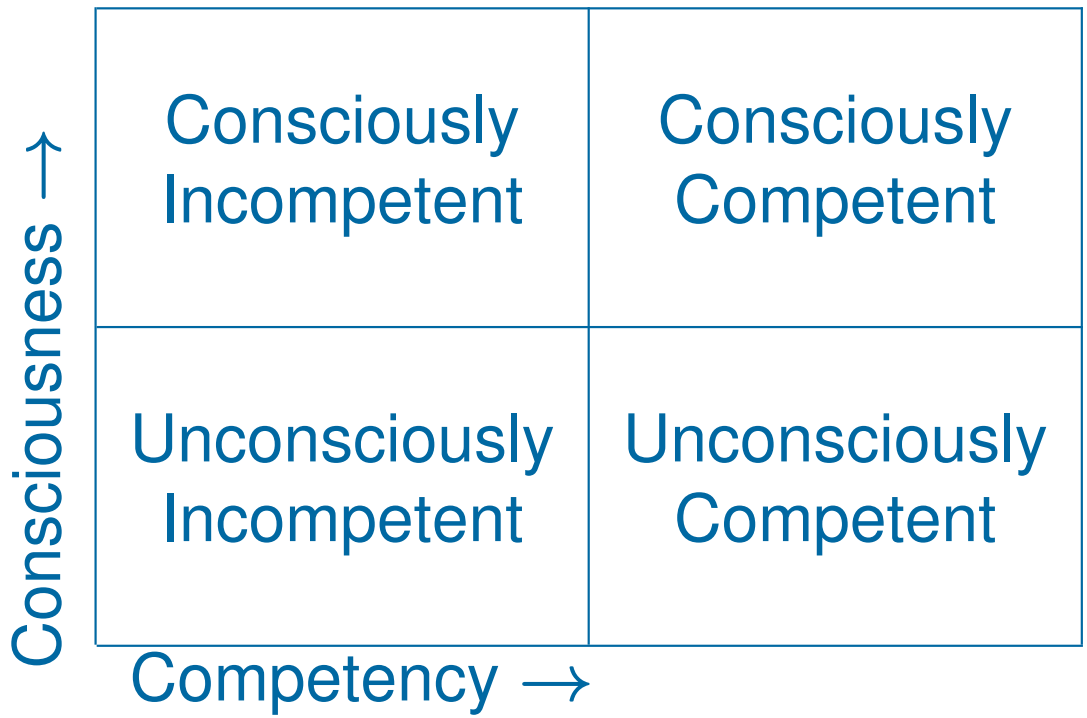
Kapitola II
Data Management Pretending

Pokus grantovek: DMP

- Data Management Plan
  - pokus grantových agentur podchytit práci s daty
  - jako povinná součást některých projetkových žádostí
- DMP je formální dokument popisující zacházení s daty během a po skončení výzkumného projektu. Má za cíl popsat management dat, vytváření metadat, ochrany dat, . . . , a jejich uchování do budoucna.
- na konceptu DMP jsme se mnoho naučili. . .
  - zejména, že vůbec nefunguje

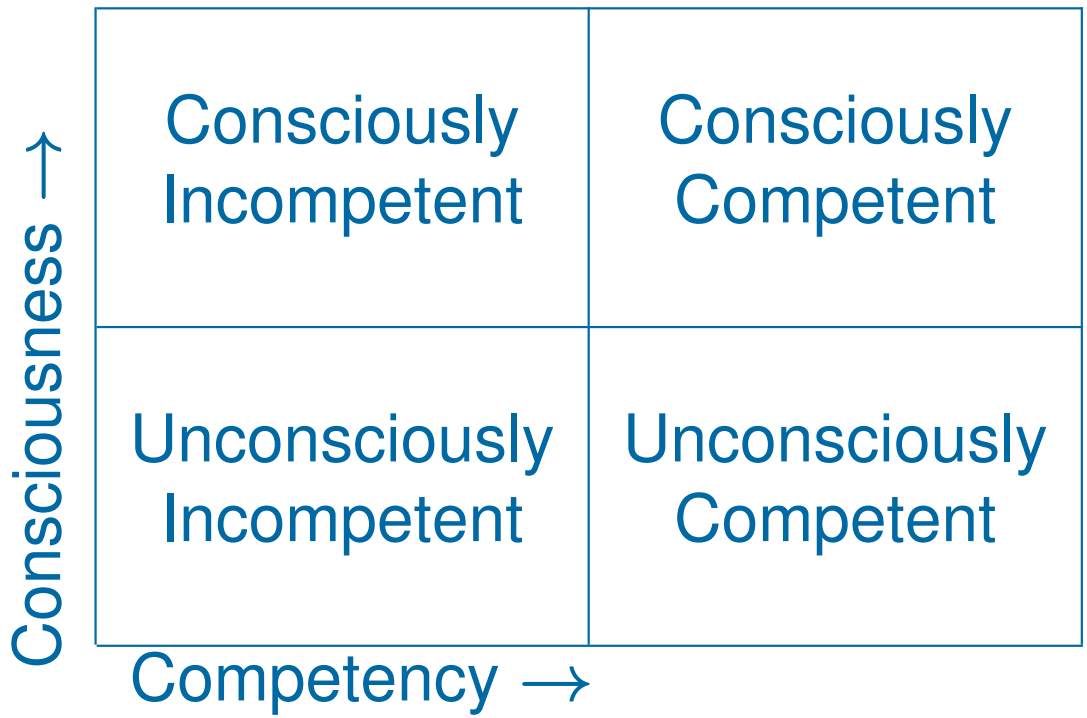
Problém s vědci

- problém s vědci podle Roba Hoofta:



Problém s vědci

- problém s vědci podle Roba Hoofta:



- a co pánové Dunning a Kurger?

Data Management Plan

- většina správy dat v projektových žádostech se redukuje na napsání DMP
  - je to přece povinná kapitola
- DMP vs. “unconsciously incompetent” autor projektu
  - který vůbec nerozumí, co tam má psát
  - odněkud to opíše
  - celé to projde, protože recenzent tomu taky nerozumí
- nebylo popsáno, co je „dobrý management dat“
  - vědče, dělej, šak ty víš co

Kapitola III
FAIR

Stručná historie FAIR I

- pra-FAIR
- březen 2011, Mons et al., The value of data, Nature Genetics
  - vztahy mezi entitami nalezitelné v datech je těžké popsat textem, a popsané textem je nelze snadno používat
  - navrhují (strojově zpracovatelné) „nanopublikace“
- únor 2013, Bechhofer et al., Why linked data is not enough for scientists, FGCS

Stručná historie FAIR II

- leden 2014, Leiden, NL, workshop Designing a Data Fairport
  - formulace FAIR principů
- září 2014, 4th RDA plenary, Amsterdam Barend Mons, “Bringing Data to Broadway”
  - to se ještě sbíraly komentáře k formulacím
- březen 2016, Wilkinson et al., The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data
  - první publikace FAIR principů
  - prakticky v současném znění

Stručná historie FAIR III

- téměř současně se formoval ELIXIR
  - ELIXIR [. . .] brings together life science resources from across Europe. These resources include databases, software tools, training materials, cloud storage and supercomputers. The goal of ELIXIR is to coordinate these resources so that they form a single infrastructure.
- autoři FAIR jsou s ELIXIRem silně spojeni
- od 2014 Barend Mons propagoval a propagoval a propagoval
  - až se věci chytili i EU úředníci
  - a všechna data musí být FAIR

FAIR – co to je?

- zaklínadlo?
  - samozřejmě
    - dnes už evropský projekt bez zmínky o FAIR neprojde
- EU mávatko do průvodu?
  - to rozhodně, ale hlavně docela užitečný koncept
    - reakce na „grantovky chtějí data management, ale není jasné, co by to mělo znamenat“
    - seznam, o čem přemýšlet
- dobrá stránka toho humbuku: začalo se o problému mluvit

- FAIR
  - Findable, Accessible, Interoperable, Reusable
- podrobný popis viz `https://www.force11.org/group/fairgroup/fairprinciples`
  - s pasážemi pro fanoušky ontologií a filosofování o reprezentaci znalostí
- za „daty“ ve FAIR lze vidět i algoritmy, nástroje a workflow
- tip: “(meta)data” značí „data i metadata“

Findable

- F1. (Meta)data are assigned a globally unique and persistent identifier
- F2. Data are described with rich metadata
- F3. Metadata clearly and explicitly include the identifier of the data they describe
- F4. (Meta)data are registered or indexed in a searchable resource

Accessible

- A1. (Meta)data are retrievable by their identifier using a standardised communications protocol
  - A1.1 The protocol is open, free, and universally implementable
  - A1.2 The protocol allows for an authentication and authorisation procedure, where necessary
- A2. Metadata are accessible, even when the data are no longer available
- všimněte si: metadata jsou věčná, data mohou zmizet

Interoperable

- I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
- I2. (Meta)data use vocabularies that follow FAIR principles
- I3. (Meta)data include qualified references to other (meta)data

- stručně: v metadatech používejte řízené slovníky a specifické odkazy
  - „X je řízeno Y“ je lepší než „X souvisí s Y“
  - to vše samozřejmě formálním jazykem

Reusable

- R1. Meta(data) are richly described with a plurality of accurate and relevant attributes
  - R1.1. (Meta)data are released with a clear and accessible data usage license
  - R1.2. (Meta)data are associated with detailed provenance
  - R1.3. (Meta)data meet domain-relevant community standards

Co FAIR je a není

- FAIR jsou stručné doménově i technologicky nezávislé principy
- FAIR není standard
- FAIR $\not\Rightarrow$ veřejně přístupný
- FAIR $\not\Rightarrow$ dostupný zdarma
  - jen má být jasné, jak se věci mají
- FAIR je svatý grál
  - „nad čím bychom se měli zamyslet“
  - ne „vše z toho je třeba splnit dokonale“
- FAIR je proces
  - level of FAIRness, FAIRification (ehm)

Komu to poslouží

- výzkumníkům, kteří chtějí svá data sdílet
  - a být za to řádně citováni
  - a v datech lovit, integrovat je, analyzovat
- grantovým agenturám
  - které chtějí nějaký datový management
- strojovému zpracování – velký důraz na něj
  - znalost vyhledávající výpočetní agent
  - který potřebuje explicitně popsanou sémantiku a kontext
  - tj. nemá intuici o významu digitálního objektu

Strojové zpracování I

- znalost hledající výpočetní agent má být schopen
  1. rozpoznat typ objektu, který nikdy předtím nepotkal
     - jeho strukturu a účel
  2. rozpoznat, zda je užitečný pro řešený problém
     - analýzou metadat nebo dat
  3. rozpoznat, zda je použitelný z hlediska licence, souhlasu, . . .
  4. provést s ním adekvátní akci
- . . . nebo aspoň malý kousek z toho

Strojové zpracování II

- srovnej:
  „Sovětští vědci jsou přesvědčeni, že do konce příští pětiletky vytvoří automatický stroj, který bude schopen vypracovat národohospodářský plán SSSR a provést jeho analýzu.“
  - – zpráva ze začátku 50. let

Kapitola IV
Technická implementace FAIR

Technická implementace

- co potřebujeme pro implementaci?
- minimalisticky vystačíme s persistentními identifikátory
  - to je zvládnutá technologie
  - kromě nápadů jako persistentní identifikátory pro verzovaná data[1]
    - ale RDA na to má pracovní skupinu
  - projekty na PIDy
    - FREYA
- a katalogy metadat
- aplikace řídící workflow se hodí
- . . . ale neexistuje univerzální řešení

[1] tohle mi fakt někdo vysvětlete. . .

DTL Data FAIRport I

- The data FAIRifcation process includes
  1. Original data retrieval
  2. Dataset identification and analysis
  3. Definition of the semantic model
  4. Data transformation
  5. License assignment
  6. Metadata definition
  7. FAIR Data resource deployment (data, metadata, license)

  Currently, this process is done manually, which limits its scalability.

  – `https://www.dtls.nl/fair-data/find-fair-data-tools/`

DTL Data FAIRport II

- sada nástrojů vyvíjených v DTL a provozovaných SURFsara
  1. FAIRifier and Metadata Editor (to create)
  2. FAIR Data Point (to publish)
  3. FAIR Search Engine (to find)
  4. ORKA (to annotate)
- podrobně na `https://www.dtls.nl/fair-data/find-fair-data-tools/`

DTL Data FAIRport III

- The FAIRifier is an online software tool designed to address the commonly encountered problems and data-manipulation tasks in the FAIRification process.
  - založeno na Google OpenRefine
  - parser na obskurní datové formáty a jejich převod do uspořádanější formy
- The FAIRifier [. . .] allows the user to mash together data and metadata, data license, the data model, and the chosen ontologies and identifiers.
  - vysvětlení od Google jsou srozumitelnější ;)

DTL Data FAIRport IV

- samotná FAIRifikace:
  (dle `https://www.go-fair.org/fair-principles/fairification-process/`)
  - analýza původních dat, jaké koncepty reprezentují, jaký mají formát, . . .
  - definice sémantického modelu: popis významu jednotlivých položek přesně, jednoznačně a strojově zpracovatelně
  - často s využitím standardních ontologií a slovníků
    - (ontologie $\approx$ slovník s hierarchií pojmů)
  - aplikace sémantického modelu (“make it linkable”)

DTL Data FAIRport V

- Metadata editor – účel zjevný
- FAIR DataPoint – webový publikační systém
  - webové rozhraní
  - API
- FAIR Data Search Engine sbírá metadata, indexuje, hledá v nich
  - čím se to liší od OpenAIRE metadat?
- anotační nástroj ORKA (Open, Reusable Knowledge graph Annotator)
  - grafický anotátor grafů znalostí
- Data FAIRport je „to všechno dohromady“

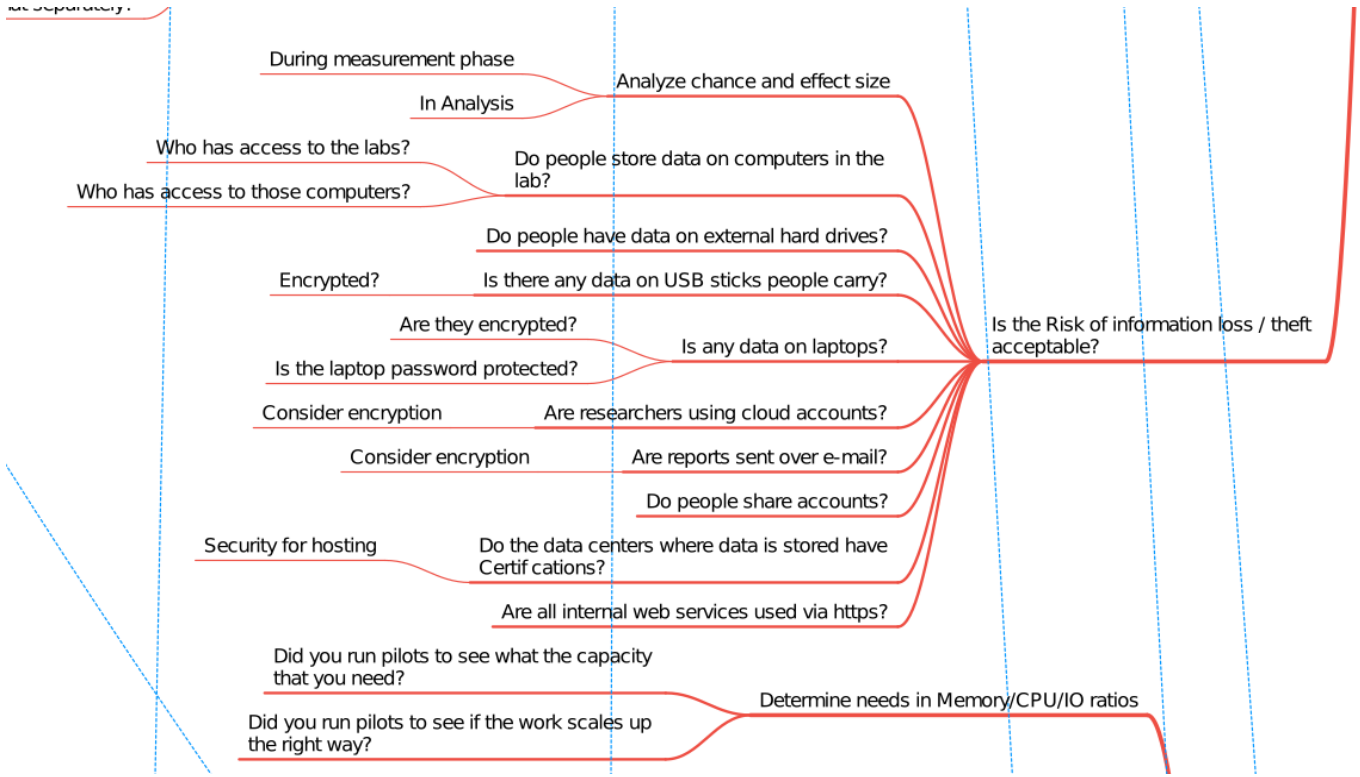
Strojové zpracování

- nástroje na strojové odvozování znalostí z FAIR dat:
  - …
- nevím o žádném

DMP, DS

- Data Management (Planning)
  $\rightsquigarrow$ Data Stewardship
- Data Steward
  - specialista na správu dat
  - včetně definice datových elementů a metadat
  - zajištění, že jsou data stále užívána
  - …
- popisy činnosti DS se značně liší
- některé univerzity zřizují DS pozice
  - i v rozsahu FTE na fakultu
- ať tomu říkáte jakkoli, do projektu stejně musíte napsat DMP

- Rob Hooft – myšlenková mapa souvislostí správy dat



DS Wizard

- Robova myšlenková mapa jako základ
- nad ní expertní systém
  - který provede uživatele tvorbou DMP
  - kladením relevantních otázek
  - formulovaných srozumitelně
  - připraveno na Life Science
  - ale snadno rozšiřitelné na jiné domény
- srovnej s DMPonline
  - užitečný nástroj
  - „obsahuje správné formuláře a mírný návod k jejich vyplnění“
  - ale jinak uživatele nevede

Kapitola V
Diskuse

Rizika FAIR I

- každá dobrá myšlenka se musí zvrhnout na kvantifikaci
- např. `http://fairmetrics.org/`
  - měření úrovně naplnění FAIR (které samy nejsou kvantifikovatelné)
  - autoři metrik si to plně uvědomují:
    - First, there is no such thing as “FAIR”, and neither is there “unFAIR”! [. . . W]e view FAIR as a continuum of ‘behaviors’ exhibited by a data resource that increasingly enable machine discoverability and (re)use. [. . .] “FAIR” will have different requirements for different communities!
  - výstup: FAIR Maturity Indicator

Rizika FAIR II

- existují i další
- lze odhadovat, že metrika v očích úředníka bude
  - snadno hodnotitelný
  - triviálně porovnatelný
  - (zcela nesmyslný, ale to úředníkovi nedojde)

  požadavek grantových agentur
- hovoří se i o certifikacích FAIR úložišť
  - když jsme se nedohodli ani na metrikách? OK…

Rizika FAIR III

- klasický případ nahrazení postupu
  - vzdělávat, pochopit, použít rozumně

  postupem
  - naučit všechny naplnit metriku, ať to má smysl, nebo ne
- snad se to tak moc nezvrhne
- a když, zaměstnanost je třeba udržovat
  - viz též

    `https://strikemag.org/bullshit-jobs/`

Shrnutí

- problém s vědeckými daty
  - a srovnání s problematikou LTP
- Data Management ve vědě
  - proč nefunguje
- FAIR
  - historie, formulace
  - co je a co není
- technické nástroje
  - PID
  - FAIRifikace
  - Data Stewardship (Wizard)
- rizika

cesnet

Jděte a učiňte vaše data FAIRovějšími!