Jak vytvořit FAIR data

Milan Janíček
milan.janicek (at) ruk.cuni.cz
Centrum pro podporu open science
Univerzita Karlova

Open Science Week 2023
17.10.2023, Ostrava

Univerzita Karlova

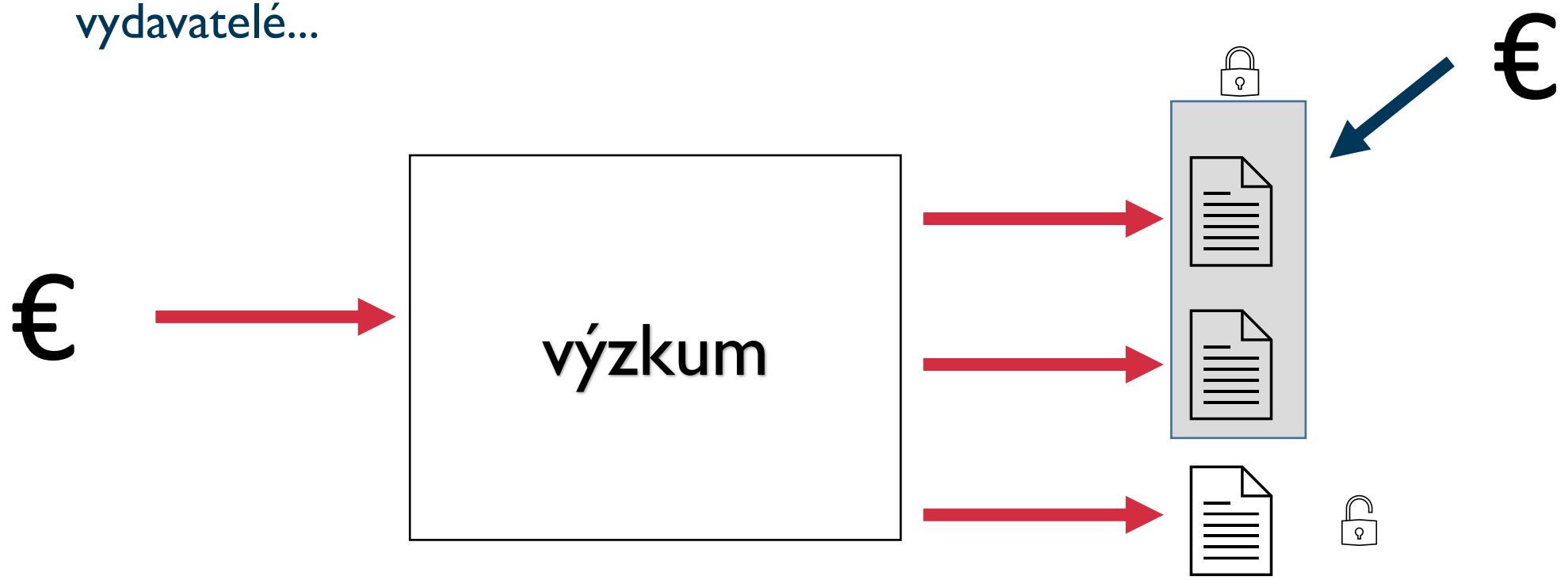
Open Science

Open Science – proč?

- cílem je:
  - větší transparentnost vědecké práce
  - lepší dostupnost výsledků

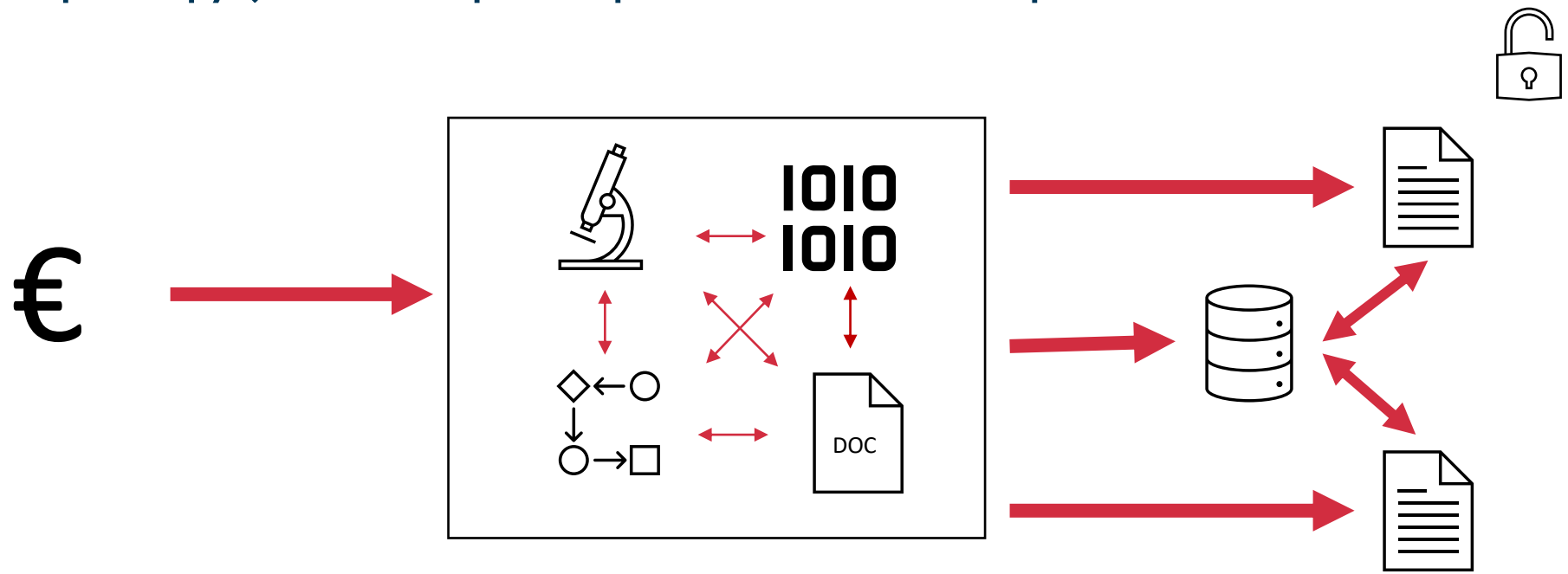
Open Science – proč?

- ve skutečnosti je situace často ještě o něco horší..
  - za výsledky se platí ve formě předplatného časopisů, práva vlastní vydavatelé...



Open Science – proč?

- prostředkem je:
  - článek není jediný výstup
  - postupy je možné pochopit / ověřit / znovupoužít



Open Science – proč?

- reprodukovatelnost výzkumu
- znovuvyužitelnost výstupů
- už jsou k dispozici technické prostředky
  - věda nemusí mít nutně formu článků

- z hlediska zdrojů financování
  - efektivnější financování
  - ideálně se jedna věc platí jednou, pak jde využít znovu
  - větší důvěra k výsledkům
  - transparentnější proces, možnost zkontrolovat nebo navázat

Jak dosáhnout změn?

- Open Science je určitá změna způsobu, jakým se pracuje
  - v některých oborech se ale blíží už zavedené „best practice“
- motivace je možná přes nastavení projektů
  - jeden z hlavních nástrojů je program Horizon Europe
- motivace v (o)hodnocení vědy?
  - zatím ne...
- Open Science = věda dělaná dobře?
  - vnitřní motivace?

Správa výzkumných dat
Research Data Management

Co jsou výzkumná data?

= informace, které byly vytvořeny za účelem zjištění či reprodukování vašich výsledků výzkumu

- Výzkumná data mohou mít různou podobu (digitální i nedigitální)
  - Tabulky, dokumenty
  - Audio a video nahrávky, obrázky, fotografie
  - Dotazníky, odpovědi na otázky, přepisy rozhovorů
  - Laboratorní deníky, terénní poznámky
  - Software, skript
  - Vzorky, exempláře

Správa dat - pojmy

→ Jaká data mají vzniknout?
→ Jak toho dosáhnout?
→ Kdo / co nám s tím pomůže?

- FAIR Data Principles = jaké mají mít data vlastnosti
- Data Management Plan / plán správy dat = jak se s daty pracuje
- Data Steward = pracovní pozice, specialista na práci s daty

FAIR principy

FAIR Principles

FAIR data

= *data, která jsou snadno nalezitelná, dostupná, interoperabilní a opětovně využitelná*

Findable

Accessible

Interoperable

Reusable

FAIR data

- jsou základní principy, jak by měla data vypadat, aby odpovídala současným požadavkům
  - co je potřeba pro budování služeb, která budou propojovat data z různých zdrojů
  - co je potřeba pro sdílení a znovuvyužití dat
  - co je potřeba pro strojové zpracování a data mining (!)
  - co je potřeba k ověření výsledků výzkumu (reproducibilita)

- nestačí vzít svoje soubory a někam je nahrát, kontext je stejně důležitý jako data samotná!

- na druhou stranu – „FAIR“ principy jsou především popisem dobré praxe

FAIR data vs. open data

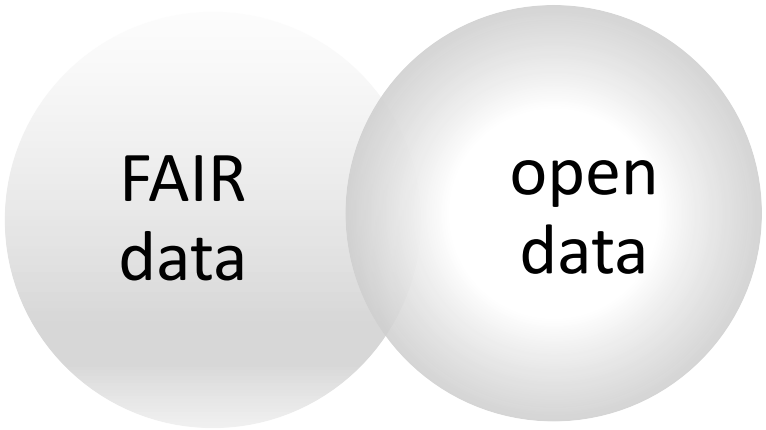
- FAIR ≠ Open (!!!)

Findable

Accessible

Interoperable

Reusable



Open data is data that can be

freely used,

re-used,

redistributed

by anyone - subject only, at most, to the requirement to attribute and sharealike.

FAIR data vs. open data

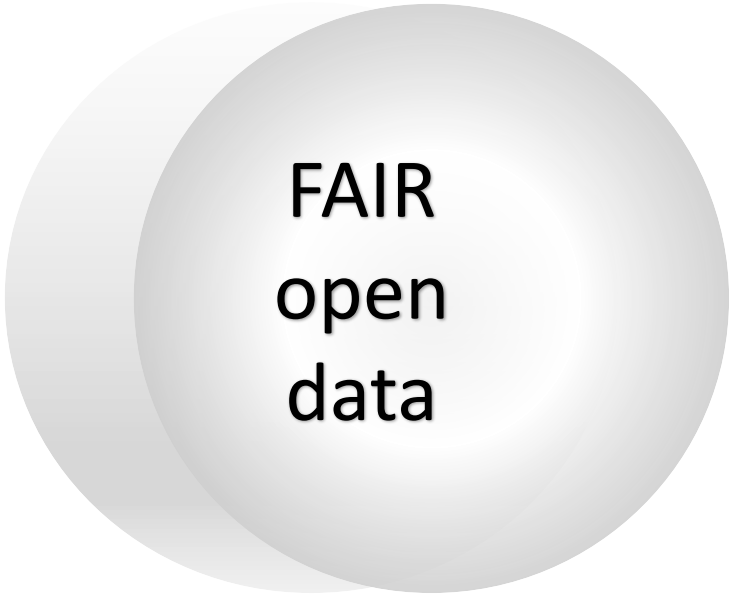
- FAIR ≠ Open (!!!)

Findable

Accessible

Interoperable

Reusable



Open data is data that can be

freely used,

re-used,

redistributed

by anyone - subject only, at most, to the requirement to attribute and sharealike.

Jak moc mají být data otevřená / uzavřená?

- obecně uznávaný princip
  - „**as open as possible, as closed as necessary**“
- existují důvody, proč data nelze poskytnout jako open data
- osobní údaje
- duševní vlastnictví
- ...

- do značné míry záleží na podmínkách poskytovatelů financí, jakékoli neotevření dat (či odklad otevření) by ale mělo být podle této zásady zdůvodněno (v rámci data management planu)

FAIR data

To be Findable:

F1. (meta)data are assigned a globally unique and eternally persistent identifier.
F2. data are described with rich metadata.
F3. (meta)data are registered or indexed in a searchable resource.
F4. metadata specify the data identifier.

TO BE ACCESSIBLE:

A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
A1.1 the protocol is open, free, and universally implementable.
A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
A2 metadata are accessible, even when the data are no longer available.

TO BE 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.

TO BE RE-USABLE:

R1. meta(data) have 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 their provenance.
R1.3. (meta)data meet domain-relevant community standards.

GO FAIR

https://www.go-fair.org/fair-principles/

FAIR Data - Findable

- nalezitelná data
- cílem je zabezpečit možnost **najít** data
- data by měli být schopni najít jak lidé tak počítače

- významnou roli hrají perzistentní identifikátory
  - např. DOI, nicméně mohou existovat i specifické oborové identifikátory
- metadata datové sady jsou základem pro nalezení v rámci různých vyhledávacích služeb
- („F“ principy jsou velmi blízké knihovnám)

Perzistentní identifikátory (PID)

- "kód" jednoznačně identifikující nějaký objekt, osobu, …
  - DOI (= digital object identifier) je perzistentní identifikátor (článku, datové sady..) 10.5281/zenodo.4263217
  - ORCID je perzistentní identifikátor (osoby) 0000-0002-8271-3674
  - URL není perzistentní identifikátor

- PID jsou ale resolvovatelné:
  - https://doi.org/10.5281/zenodo.4263217
  - https://orcid.org/0000-0002-8271-3674

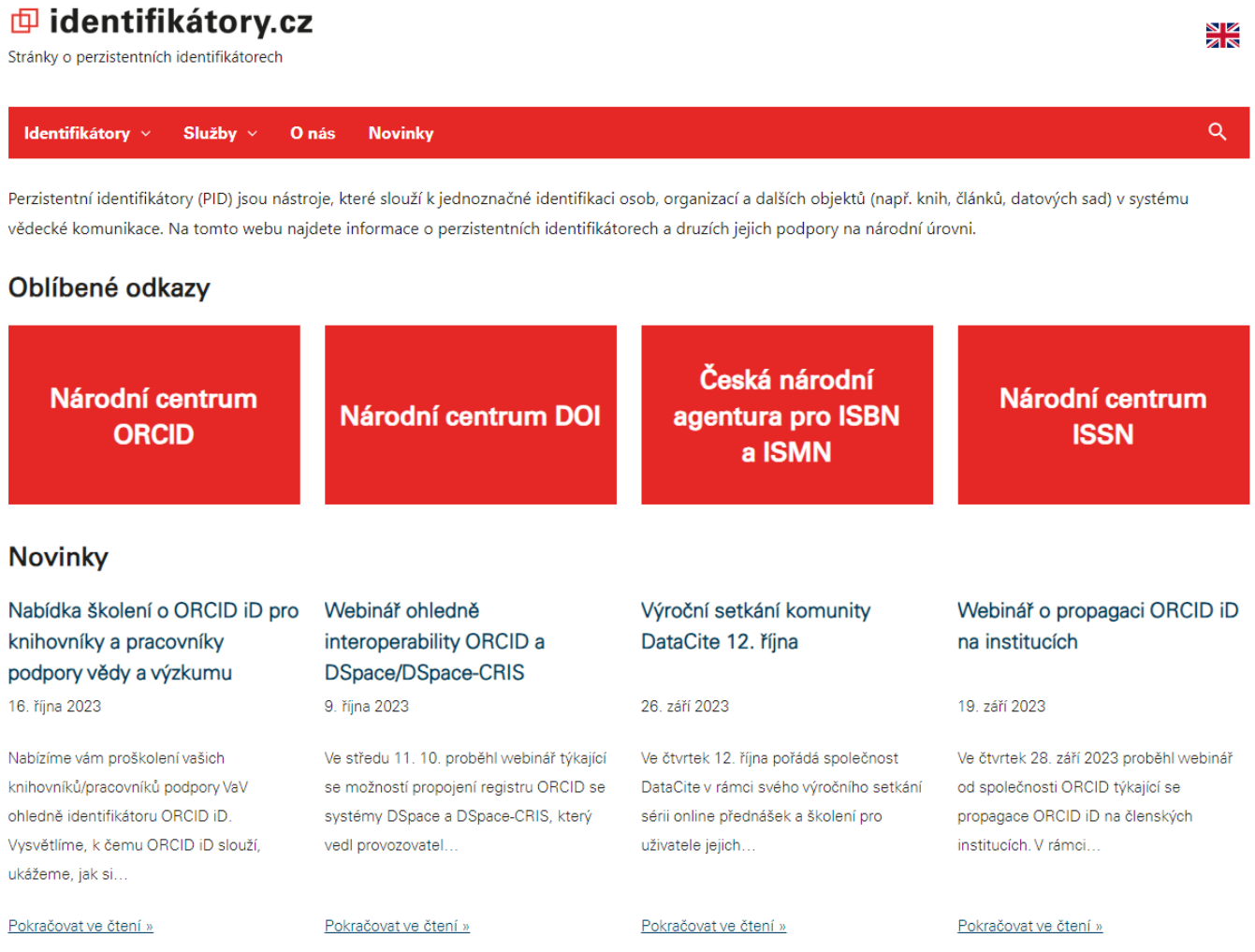
Vlastnosti PID

- globálně unikátní
  - nikdo další už ho nikdy nemůže přiřadit ničemu jinému

- resolvovatelný (překlad, někdo? ☺)
  - existuje způsob, jak získat pomocí identifikátoru samotný identifikovaný digitální objekt, jeho reprezentaci nebo alespoň odpovídající stránku

- perzistentní (trvalý)
  - zůstává trvale unikátní a resolvovatelný
  - objekt, který je identifikátorem popsán, se také v čase nemění
  - to může být složité / drahé

Centrum PID



NTK
50°6'14.083"N, 14°23'26.365"E
Národní technická knihovna
National Library of Technology

https://identifikatory.cz/

Co/kdo všechno může mít PID?

- publikace
  - DOI
- (digitální) objekty
  - handle (např. dspace.cuni.cz)
  - DOI
- lidé
  - ORCID
  - ResearcherID
- organizace
  - ROR.org, GRID, ISNI...
- ... a další

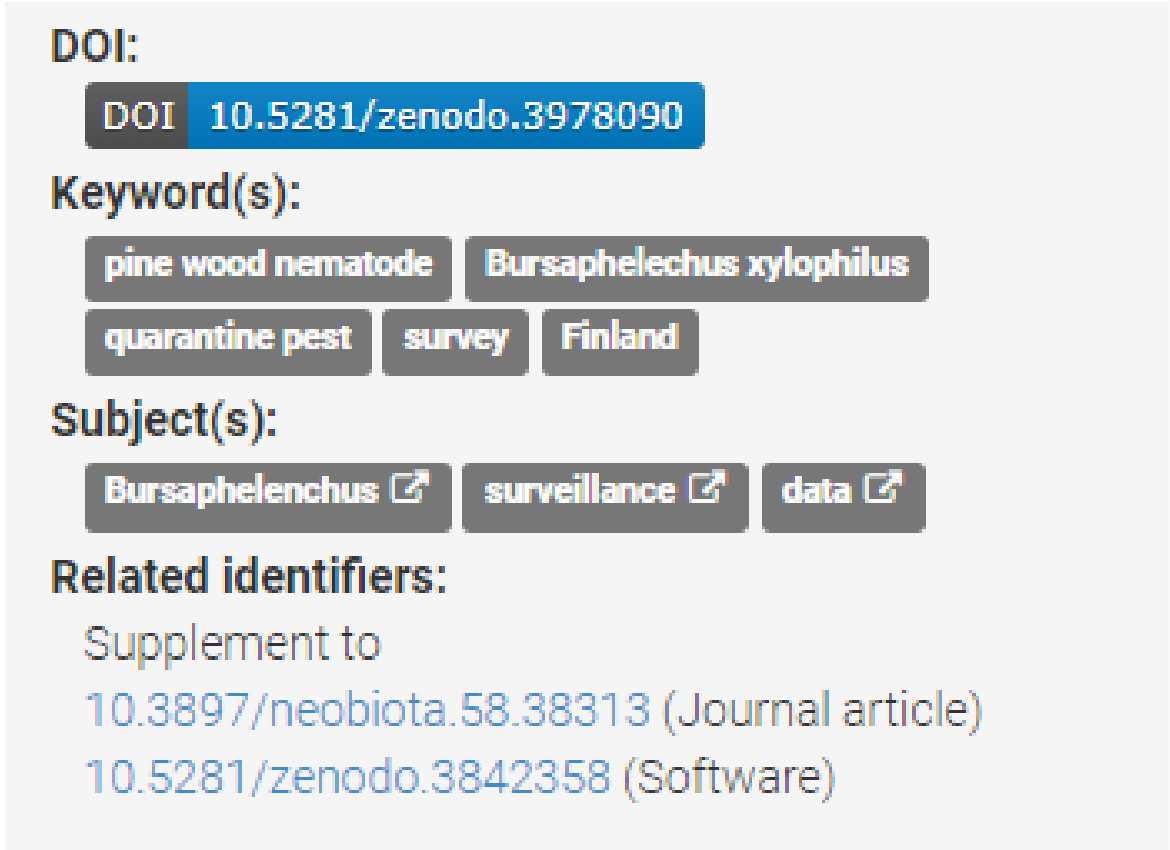
Metadata (nalezitelnost)

- data popisující jiná data
  - například název datové sady, autor, klíčová slova, popis vlastností..
- důležité nepodceňovat - podle metadat záznamu se vyhledává: špatná metadata => nenalezitelná data!
- při popisu je vhodné používat řízené slovníky (a ontologie)
  - záleží na oborových zvyklostech – např. MeSH v medicíně
- je dobré popisovat data takovými termíny, které byste zadali do vyhledávače, kdybyste je sami chtěli najít

FAIR Data - Findable

- identifikátor
- metadata
- odkazy na další zdroje

  → via identifikátory



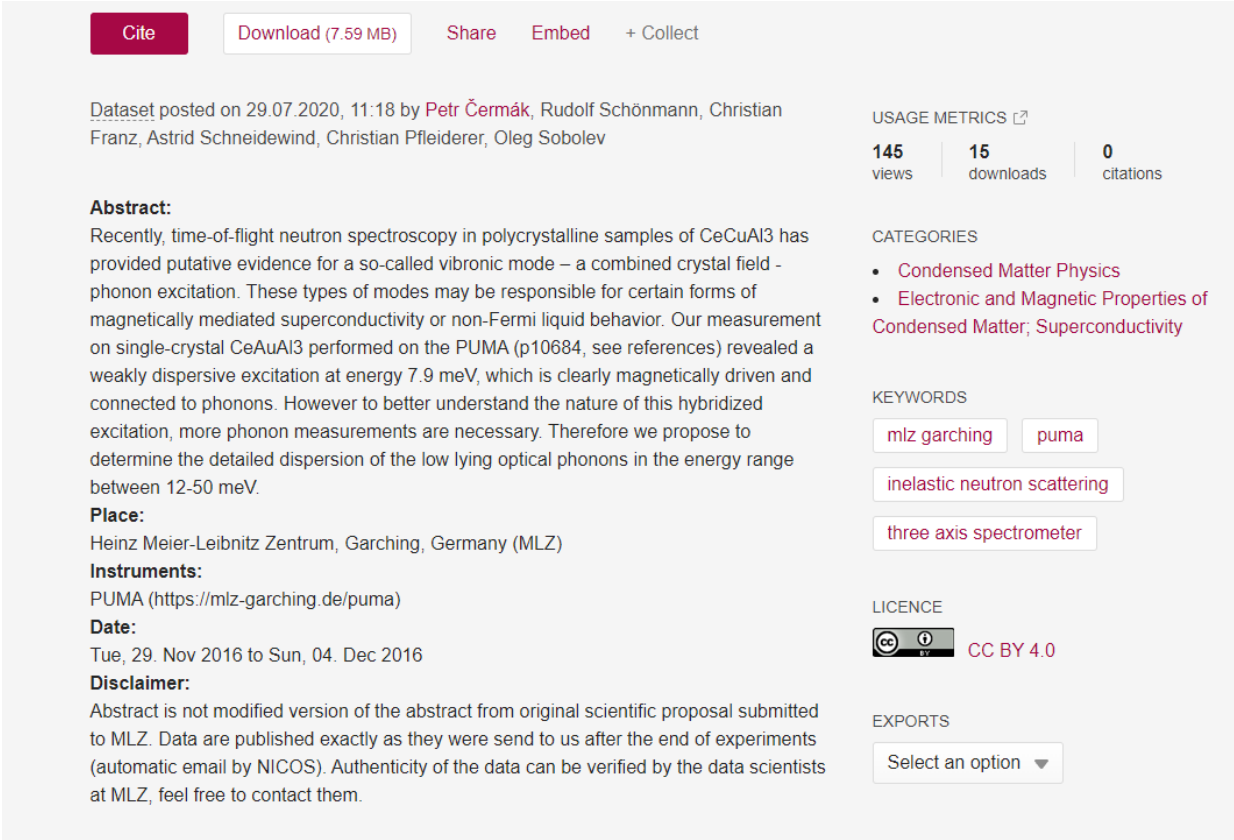
FAIR Data - Accessible

- přístupná data
  - meta(data) by měla být získatelná pomocí svého identifikátoru
  - za použití standardního komunikačního protokolu, který může použít kdokoliv
    - autentizace a autorizace jsou možné
      - nemusí být nutně „online“ protokol
      - ne *všechno* musí být volně dostupné (!)

- metadata by měla být dostupná, i kdyby už samotná data nebyla
  - pokud by už data nebyla dostupná, identifikátor by měl vést na „tombstone“ - náhrobek

FAIR Data - Accessible

- možnost stažení
- publikovaná alespoň metadata



FAIR Data - Interoperable

- interoperabilní data
  - mělo by být možné kombinovat data s dalšími daty a využívat další nástroje
- formát by měl být otevřený a interoperabilní pro různé nástroje a služby
- otevřený formát nemusí existovat, ale pokud existuje, měl by být preferován
  - i metadata by měla být interoperabilní
    - je vhodné používat oborové standardy

Metadata (interoperabilita)

- otevřený formát sám o sobě nestačí k zajištění interoperability
- teplota [°C]:
  - 36,6
  - 36,1
  - 37,8
  - 41,5
- .... bylo teplé léto, nebo měřím teplotu pacientům?

Metadata (interoperabilita)

- data samotná by měla být „vevnitř“ dostatečně popsaná
  - přinejmenším by je měl pochopit člověk
  - optimálně by bylo popsat je tak, aby je mohl zpracovat i stroj
  - tzn. opět použít předem definované řízené slovníky / taxonomie / ontologie
- ideální je využít oborový standard (pokud existuje)
  - → roste tlak na vytváření oborových standardů – data z různých zdrojů by měla vznikat v takové formě, aby byla rovnou interoperabilní
  - užitečný zdroj existujících standardů je https://fairsharing.org/
  - německá iniciativa - 26 oborových konzorcií NFDI https://www.nfdi.de/consortia/?lang=en

NFDI – inspirace?

Humanities and Social Sciences

- BERD@NFDI: NFDI for Business, Economic and Related Data
- KonsortSWD: Consortium for the Social, Educational, Behavioural and Economic Sciences
- NFDI4Culture: Consortium for Research Data on Material and Immaterial Cultural Heritage
- NFDI4Memory: The Consortium for the Historically Oriented Humanities
- NFDI4Objects – Research Data Infrastructure for the Material Remains of Human History
- Text+: Language and text-based research data infrastructure

Engineering Sciences

- NFDI4DataScience: NFDI for Data Science and Artificial Intelligence
- NFDI4Energy: National Research Data Infrastructure for Interdisciplinary Energy System Research
- NFDI4Ing: NFDI for Engineering Sciences
- NFDI-MatWerk: National Research Data Infrastructure for Materials Science and Materials Engineering
- NFDIxCS – National Research Data Infrastructure for and with Computer Science

https://www.nfdi.de/consortia/?lang=en

Life Sciences

- DataPLANT: Plant research data
- FAIRagro: FAIR Data Infrastructure for Agrosystems
- NFDI4Immuno – National Research Data Infrastructure for Immunology
- GHGA: National Research Data Infrastructure for Immunologyv
- NFDI4Biodiversity: Biodiversity, Ecology and Environmental Data
- NFDI4BIOIMAGE: National research data infrastructure for microscopy and bioimage analysis
- NFDI4Health: NFDI personal health data
- NFDI4Microbiota: NFDI for Microbiota Research

Natural Sciences

- DAPHNE4NFDI: Data from PHoton and Neutron Experiments for NFDI
- FAIRmat: FAIR Data Infrastructure for Condensed-Matter Physics and the Chemical Physics of Solids
- NFDI4Cat: NFDI for sciences related to catalysis
- MaRDI: Mathematical Research Data Initiative
- NFDI4Chem: Chemistry consortium for the NFDI
- NFDI4Earth: NFDI Consortium Earth System Sciences
- PUNCH4NFDI: Particles, Universe, NuClei and Hadrons for the NFDI

FAIR Data - Reusable

- znovuvyužitelná data
  - optimalizujte data pro znovuvyužití
    - důkladný a přesný popis dat umožňuje jak replikaci výsledků tak nové využití dat v novém kontextu
    - provenance – (zjednodušeně) informace o původu a úpravách dat
- důležitá je i licence – specifikuje jakým způsobem a za jakých podmínek mohou být data znovu využita
  - není vždy nutné všechno publikovat úplně volně
  - {na druhou stranu – volnější licence usnadňuje další využití dat, o což by v principu mělo jít, obzvlášť pokud byla vytvořena s veřejnou finanční podporou}

FAIR Data - Reusable

- licence
- provenance

Publication date:
April 4, 2021

DOI:
DOI 10.5281/zenodo.4661737

Keyword(s):
COVID-19 | Air Traffic | flight tracks | fuel consumption | Black Carbon emission | contrail | CoCiP

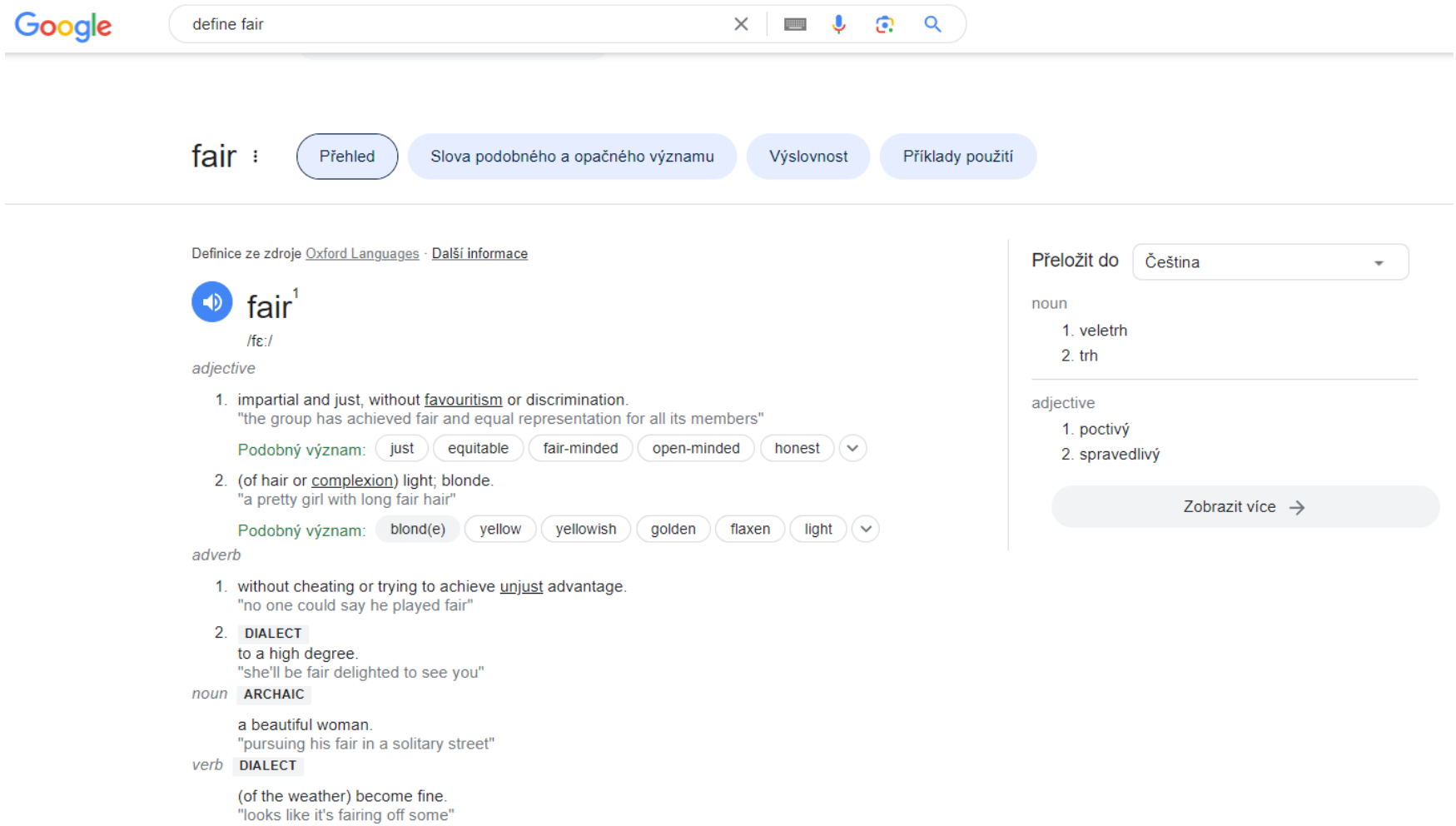
License (for files):
Creative Commons Attribution 4.0 International

Producer	Hubert Mara (IWR, Heidelberg University) (HMara) https://orcid.org/0000-0002-2004-4153 Bartosz Bogacz (IWR, Heidelberg University) (BBogacz) https://orcid.org/0000-0002-2004-4153
Production Date	2019-03-11
Production Place	Heidelberg, Germany
Contributor	Project Member : Bayer, Paul Victor
Deposit Date	2019-02-25
Date of Collection	Start: 2018-07-24 ; End: 2018-08-22 Start: 2019-03-01 ; End: 2019-03-11
Kind of Data	Cuneiform tablets; 3D Measurement data
Software	GigaMesh Software Framework, Version: 181100 to 190300
Related Datasets	Heidelberg Cuneiform 3D Database (HeiCu3Da) for the Hilprecht Collection: https://doi.org/10.11588/heidicon.hilprecht
Origin of Sources	Hilprecht Sammlung, Jena, Germany, https://hilprecht.mpiwg-berlin.mpg.de/ Cuneiform Digital Library Initiative (CDLI) https://cdli.ucla.edu/

Jak vyrobit FAIR data?

- nejefektivnější je počítat s požadavky FAIR už na začátku výzkumu a vše si naplánovat (Data Management Plan; plán správy dat) a pak plán dodržet (Research Data Management)
- začít řešit FAIR na konci procesu je pozdě a bude to stát úsilí navíc!

FAIR == fair?


Google
define fair
fair
Přehled
Slova podobného a opačného významu
Výslovnost
Příklady použití
Definice ze zdroje Oxford Languages · Další informace
fair[1]
/fɛː/
adjective
1. impartial and just, without favouritism or discrimination.
"the group has achieved fair and equal representation for all its members"
Podobný význam: just equitable fair-minded open-minded honest
2. (of hair or complexion) light; blonde.
"a pretty girl with long fair hair"
Podobný význam: blond(e) yellow yellowish golden flaxen light
adverb
1. without cheating or trying to achieve unjust advantage.
"no one could say he played fair"
2. DIALECT
to a high degree.
"she'll be fair delighted to see you"
noun ARCHAIC
a beautiful woman.
"pursuing his fair in a solitary street"
verb DIALECT
(of the weather) become fine.
"looks like it's fairing off some"
Přeložit do Čeština
noun
1. veletrh
2. trh
adjective
1. poctivý
2. spravedlivý
Zobrazit více

Jak vyrobit FAIR data? (výběr otázek)

- Jaká data budete vytvářet?
  - typy, formáty souborů
- Která z nich budete chtít sdílet? Jaké budou sloužit jako podklady k publikacím?
  - struktura dat, ontologie
- Jaké jsou oborové standardy a best practices? Budou vaše data potenciálně někomu užitečná?
- Jak budete data v průběhu práce popisovat? Jakým způsobem budete vést dokumentaci?
- Kde budete data uchovávat? Budete je sdílet? S kým a za jakých podmínek?
- Budete ukládat data do repozitáře? Oborový / institucionální / obecný? Jak budou popsaná? Jaké využijete identifikátory?

Jak vyrobit FAIR data? (výběr otázek)

- Jaká data budete vytvářet?
  - typy (I), formáty souborů (I)
- Která z nich budete chtít sdílet? Jaké budou sloužit jako podklady k publikacím?
  - struktura dat (I,R), ontologie (I,R)
- Jaké jsou oborové standardy a best practices? (I,R) Budou vaše data potenciálně někomu užitečná?
- Jak budete data v průběhu práce popisovat? (I,R) Jakým způsobem budete vést dokumentaci? (I,R)
- Kde budete data uchovávat? (A) Budete je sdílet? (A) S kým a za jakých podmínek? (A,I,R)
- Budete ukládat data do repozitáře? (F,A,R) Oborový / institucionální / obecný? Jak budou popsaná? (F) Jaké využijete identifikátory? (F,A,I,R)

• Kdo/co pomůže se správně zeptat?

- vlastní znalosti
  - řada materiálů na webu, například How to FAIR
- Data Steward
  - pracovní pozice - přímá podpora pro práci s daty na instituci
  - otázky + pomoc s odpovědí (!)
  - interaktivita, přirozená inteligence, pochopení kontextu, odborná znalost práce s daty
- Data Management Plan
  - dokument jehož rolí je mimo jiné i nasměrovat výzkumníka ke správným otázkám
  - někdy jen formulářová šablona...
  - ... jindy „virtuální“ Data Steward – Data Stewardship Wizard

Pár poznámek na závěr / k diskuzi?

- FAIR principy nejsou pevná pravidla, popisuje se ideální stav.
- „Národní“ není tak důležité jako „oborové“.
- Důležité je plánovat práci s daty v průběhu projektu.
- Ne všechno závisí na lokální infrastruktuře. Ne všechno vyřeší její existence.
- Vzdělávání a budování infrastruktury bude postupný a dlouhotrvající proces.
- Každý výzkumník
  - by měl vědět jak má pracovat s daty, včetně toho jak mají (v principu) vypadat požadované výstupy
  - by měl mít k dispozici dostatečnou technickou infrastrukturu
  - by měl mít k dispozici dostatečnou podporu

  ➔ nemělo by být cílem udělat z každého *navíc* experta na data

Děkuji za pozornost :-)

Milan Janíček
milan.janicek (at) ruk.cuni.cz
Centrum pro podporu open science
Univerzita Karlova

Open Science Week 2023
17.10.2023, Ostrava

Univerzita Karlova