FAIR data: Principy pro správu výzkumných dat

Vít Novotný, 6. ročník
Fakulta informatiky, Masarykova univerzita

Brno, 27. listopadu 2018

Internet umožňuje s malými náklady celosvětově publikovat vědecká data a učinit je tak užitečná pro budoucí výzkum. Toto však vyžaduje přítomnost infrastruktury a obecných postupů pro vyhledávání vědeckých dat na základě jejich parametrů.

Díky rozvoji výpočetní techniky jsme dnes schopni strojově zpracovávat velké objemy vědeckých dat. Toto zpracování však vyžaduje mnohdy týdny ruční práce, která sestává z hledání a stažení požadovaných dat, zjištění jejich původu a podmínek užití a jejich převodu do jednotného formátu pro strojové zpracování.

Findability, Accessibility, Interoperability, Reusability (FAIR) je iniciativa, která se zaobírá formulací a uplatňováním principů přípravy a publikování dat. Cílem těchto principů je zajistit vyhledatelnost, přístupnost, interoperabilitu a znovupoužitelnost vědeckých dat. V rámci této eseje pojednám o historii iniciativy FAIR, jejích principech a dopadu na kvalitu vědeckých dat.

Stručná historie

Iniciativa FAIR historicky vychází z principů strojové zpracovatelnosti vědeckých dat formulovaných iniciativou Concept Web Alliance[1] v roce 2009 a z principů citovatelnosti vědeckých dat formulovaných občanským sdružením the Future of Research Communications and E-scholarship (FORCE11) v roce 2014 [1].

V roce 2014 uspořádal správce nizozemské datové infrastruktury ELIXIR, institut Dutch Techcentre for Life Sciences (DTL), konferenci „Jointly designing a Data FAIRPORT" týkající se vyhledatelnosti, přístupnosti, použitelnosti a citovatelnosti vědeckých dat. Konference se konala v nizozemském Leidenu a zúčastnila se jí jak odborná veřejnost, tak zástupci grantových agentur a společností poskytujících datové repositáře a nástroje pro práci s daty. Zde byly ustanoveny principy FAIR[2] a stejnojmenná pracovní skupina [2] občanského sdružení FORCE11.

[1] `https://conceptweblog.wordpress.com/`.
[2] `https://www.force11.org/fairprinciples`.

V současnosti je iniciativa FAIR významnou součástí evropského projektu European Open Science Cloud (EOSC) [3], který má do roku 2020 sjednotit přístup k datovým výstupům evropských výzkumných projektů.

Principy FAIR

Klíčovým pojmem v terminologii FAIR je *datový objekt*, který je zobecněním *digitálního objektu* Kahna a Wilenského [4] a sestává z identifikátoru, metadat a dat. Zúžením pojmu je tzv. *FAIR datový objekt*, který je vymezený jednotlivými principy FAIR. Těmito principy jsou *vyhledatelnost*, *přístupnost*, *interoperabilita* a *znovupoužitelnost* datových objektů, přičemž každý se skládá z několika dílčích požadavků.

Vyhledatelný datový objekt je neformálně takový datový objekt, který má dostatečně bohatá metadata k tomu, aby jej bylo možné konzistentně vyhledat na základě jeho parametrů. Konkrétně by měl objekt obsahovat identifikátor, který je globálně unikátní a trvalý a totéž by mělo platit pro další identifikátory obsažené v datech a metadatech objektu. Metadata by měla být trvale dostupná i v případě smazání dat a dostatečně bohatá, aby bylo možné dva datové objekty strojově odlišit.

Přístupný datový objekt je neformálně takový objekt, jehož metadata a data lze stáhnout, a to ideálně v množství strojově i lidsky čitelných formátů. Konkrétně by měl být objekt dostupný skrz standardní komunikační protokol.

Interoperabilní datový objekt je neformálně takový objekt, který je strojově zpracovatelný. Konkrétně by měl objekt obsahovat data a metadata, která jsou předvídatelně strukturovaná a používají standardní ontologie.

Znovupoužitelný datový objekt by měl být vyhledatelný, přístupný a interoperabilní. Metadata a data by měla být dostatečně bohatá, aby byla strojově zpracovatelná ve spojení s dalšími daty. Metadata by měla umožnit publikovaný datový objekt citovat podle Joint Declaration of Data Citation Principles (JDDCP) [1].

Výsledky

Během nizozemského předsednictví Evropské unii v roce 2016 spolupublikoval vedoucí nizozemské infrastruktury ELIXIR, Barend Mons, impaktovaný článek v časopisu Nature [5], ve kterém popsal principy FAIR a ilustroval je na příkladech existujících datových repozitářů Dataverse, FAIRDOM, Open PHACTS [6], Worldwide Protein Data Bank (wwPDB) [7, 8] a UniProt [9]. V témže roce Mons navrhnul Evropské komisi z pozice člena expertní skupiny EOSC, aby byly principy FAIR povinně uplatňovány na veškeré datové výstupy evropských výzkumných projektů. Dopad článku a význam strojové zpracovatelnosti dat ilustruje i veřejné vyjádření podpory iniciativě FAIR formulované leadry G20 na zářijovém summitu v Číně.

V reakci na spuštění projektu EOSC vznikla v roce 2017 iniciativa Global Open FAIR

(GO FAIR), která navrhla praktickou implementací EOSC. Návrh zahrnuje vzdělávání evropských akademiků a datových expertů, využití existujících infrastruktur na úrovni členských států a rozvoj standardů, protokolů a služeb, které umožní federativní propojení národních infrastruktur do funkčního celku. V březnu téhož roku vydali státní tajemníci Nizozemska a Německa prohlášení [10], ve kterém nabádají k okamžité podpoře GO FAIR v obavách, že prodlevy povedou ke snížení konkurenceschopnosti. Kromě Nizozemska a Německa se do iniciativy GO FAIR na národní úrovni zapojila i Francie, a to formou nově vzniknuvší pracovní skupiny International Support and Coordination Office (ISCO).

Pro Čechy je zajímavá informace, že v roce 2017 pracovníci české infrastruktury ELIXIR ve spolupráci s nizozemským institutem DTL vyvinuli program „Data Stewardship Wizard“, který usnadňuje návrh plánu managementu dat pro výzkumné projekty.

Z publikační činnosti v roce 2017, která se týká iniciativy FAIR, se jeví jako významný článek od Wilkinsona a spol. [11], kteří se zabývají návrhem postupů, které by umožnily kvantitativně vyhodnotit naplnění jednotlivých principů FAIR. Stojí však za zmínku, že srovnatelné postupy popisují již existující normy ISO 16363:2012, DIN 31644 a pracovní verze doporučení W3C [12] o anotaci dat.

Metadata v Dublin Core (DC)

```
<dc:title>
  FAIR data: Principy pro správu
  výzkumných dat
</dc:title>
<dc:creator>Vít Novotný</dc:creator>
<dc:date>2017-11-28</dc:date>
<dc:type>Text</dc:type>
<dc:format>public</dc:format>
<dc:language>cz</dc:language>
```

Zkratky

DC Dublin Core 3

DTL Dutch Techcentre for Life Sciences 1, 3

EOSC European Open Science Cloud 2, 3

FAIR Findability, Accessibility, Interoperability, Reusability 1–3

FORCE11 the Future of Research Communications and E-scholarship 1

GO FAIR Global Open FAIR 2, 3

ISCO International Support and Coordination Office 3

JDDCP Joint Declaration of Data Citation Principles 2

wwPDB Worldwide Protein Data Bank 2

Odkazy

1. FORCE11. *Data Citation Synthesis Group: Joint Declaration of Data Citation Principles* [online]. San Diego CA, 2014 [cit. Nov. 27, 2018]. Dostupné z: https://www.force11.org/datacitation.
2. MARTONE, Maryann E. FORCE11: Building the Future for Research Communications and e-Scholarship. *BioScience*. 2015, vol. 65, no. 7, s. 635. Dostupné z DOI: 10.1093/biosci/biv095.
3. COMMISSION, European. Implementation Roadmap for the European Open Science Cloud [online]. 2018 [cit. Nov. 28, 2018]. Dostupné z: https://ec.europa.eu/research/openscience/pdf/swd_2018_83_f1_staff_working_paper_en.pdf.
4. KAHN, Robert; WILENSKY, Robert. A framework for distributed digital object services [online]. 1995 [cit. Nov. 27, 2018]. Dostupné z: http://www.cnri.reston.va.us/k-w.html.
5. WILKINSON, Mark D et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*. 2016, vol. 3.
6. HARLAND, Lee. Open PHACTS: A semantic knowledge infrastructure for public and commercial drug discovery research. In: *International Conference on Knowledge Engineering and Knowledge Management*. 2012, s. 1–7.
7. BERMAN, Helen M; WESTBROOK, John; FENG, Zukang; GILLILAND, Gary; BHAT, Talapady N; WEISSIG, Helge; SHINDYALOV, Ilya N; BOURNE, Philip E. The protein data bank. *Nucleic acids research*. 2000, vol. 28, no. 1, s. 235–242.
8. BERMAN, Helen; HENRICK, Kim; NAKAMURA, Haruki. Announcing the worldwide protein data bank. *Nature Structural and Molecular Biology*. 2003, vol. 10, no. 12, s. 980.
9. CONSORTIUM, UniProt. UniProt: a hub for protein information. *Nucleic acids research*. 2014, vol. 43, no. D1, s. D204–D212.
10. GERMANY; NETHERLANDS, the. Joint Position Paper on the European Open Science Cloud [online]. 2017 [cit. Nov. 27, 2018]. Dostupné z: https://www.dtls.nl/wp-content/uploads/2017/05/DE-NL-Joint-Paper-FINAL.pdf.
11. WILKINSON, Mark D; SANSONE, Susanna-Assunta; SCHULTES, Erik; DOORN, Peter; SILVA SANTOS, Luiz Olavo Bonino da; DUMONTIER, Michel. A design framework and exemplar metrics for FAIRness. *Scientific data*. 2018, vol. 5.
12. GRAY, Alasdair J.G.; BARAN, Joachim; MARSHALL, M. Scott; DUMONTIER, Michel, eds. *Dataset Descriptions: HCLS Community Profile* [online]. 2015 [cit. Nov. 28, 2018]. Dostupné z: https://www.w3.org/TR/hcls-dataset/.