FREYA PID Graph

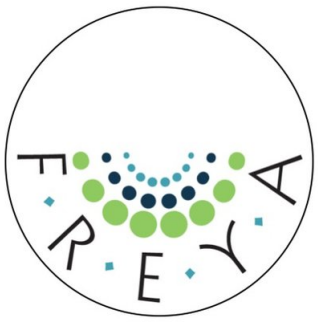
Michal Hala

Listopad 2019

Obsah

1 Úvod

Perzistentní identifikátory jsou nezbytné pro jednoznačnou identifikaci zdrojů, a to nejen online. Momentálně používané identifikátory však naráží na určitá úskalí, zejména při citování v rozsáhlých výzkumných pracích. Rovněž by bylo užitečné, kdybychom mohli zdroje propojovat na základě jejich vztahů. Takové propojení nám umožňuje PID Graph, produkt projektu FREYA, který díky metadatům dovede propojovat zdroje distribuované online, a to vše za účelem volného přístupu k datům, efektivního vyhledávání a zrychlování vědeckého pokroku.



Obrázek 1: Logo projektu FREYA[1]

2 Perzistentní identifikátory (PID)

Identifikátory jsou výrazy sloužící k jednoznačné identifikaci objektů v konkrétní doméně (např. URL pro webové stránky, ISBN pro knihy, rodná čísla pro osoby). Identifikátory nicméně časem podléhají zkáze, některé více než ostatní, zejména pak odkazy na webové stránky, u kterých není nijak garantováno, že server nebude druhý den navždy odstaven.

Jako perzistentní označujeme takové identifikátory, u kterých je dovednost identifikovat objekty zaručená ideálně věčně, prakticky ale poskytuje jen o něco silnější záruku funkčnosti (např. PURL je perzistentnější variantou URL, jelikož neodkazuje na konkrétní místo na síti, resp. na serveru, ale na meziserver, na němž se nachází udržovatelná URL adresa).

2.1 Digital Object Identifier (DOI)

DOI je centralizovaný systém PID sloužící k identifikaci digitálních objektů online. Jeho návrh pochází z roku 1996 a roku 2012 se stal mezinárodním standardem.

DOI se skládá z předpony, odkazující na instituci registrovanou u společnosti CrossRef, a koncovky, odkazující na konkrétní dokument, který musí být v dané doméně unikátní. Předpona je oddělena od přípony lomítkem, příklad: 10.1058/am23456.[2]

3 Projekt FREYA

FREYA je tříletý projekt financovaný Evropskou komisí spadající pod program Horizon 2020. Cílem projektu je budovat infrastrukturu pro používání perzistentních identifikátorů (PID) pro výzkum v nejen v Evropské unii, ale i globálně. Cílem je vylepšit objevování, odkazování, získávání vědeckých zdrojů, důraz je rovněž kladen na otevřený, volný přístup. Díky tomu bude vědecká komunita schopná lépe vyhodnocovat data, a učinit své vlastní záznamy kompletnější, spolehlivější a dohledatelnější. Projekt spolupracuje s globální komunitou prostřednictvím Research Data Alliance (RDA) za účelem plného zpřístupnění vědeckých dat. FREYA navazuje na projekt THOR (Technical and Human Infrastructure for Open Research)[3]. [4]

3.1 Pilíře projektu

- PID Graph slouží k propojování PID systémů vytvářející síť PID, na níž staví další služby.
- PID Forum slouží ke komunikaci s online komunitou na `pidforum.org`, kde se organizují konference, workshopy a další události.
- PID Commons se snaží o udržitelnost PID infrastruktury, a to i po skončení projektu.

3.2 Cíle projektu

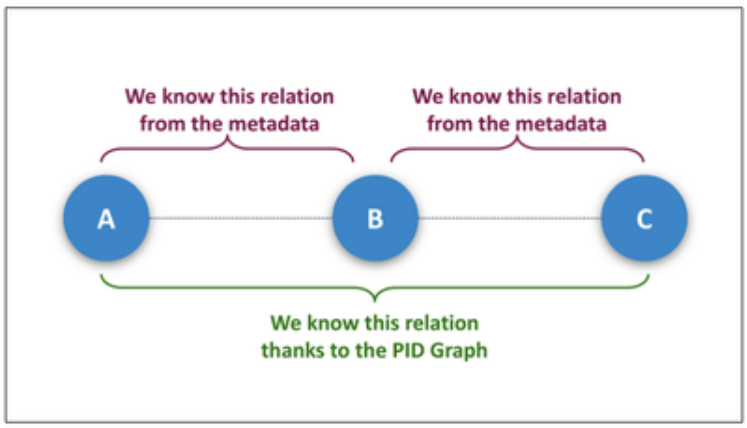
- Zlepšit vyhledávání zdrojů pomocí PID služeb a kooperaci s Crossref, DataCite, ORCID a identifiers.org.
- Zlepšovat a vyvíjet PID služby.
- Integrace PID systému do konkrétních vědeckých oborů a European Open Science Cloud (EOSC).
- Vytvářet a rozvíjet komunitu pomocí PID Fóra.
- Udržovat komunitu a PID infrastrukturu pro vědecké účely v EU i ve světě.

3.3 European Open Science Cloud (EOSC)

EOSC je projekt Evropské unie, který se snaží umožnit volný přístup a sdílení vědeckých dat. Projekt funguje od roku 2016 a od té doby spolupracuje s více než 1,7 miliony výzkumníků a 70 miliony uživatelů z EU. [5]

4 PID Graph

PID jsou sice dobrá metoda, když dojde na odkazování se na cizí zdroje, ale samy o sobě neposkytují příliš dobrou informaci o souvislostech mezi těmito zdroji. Jestliže ale PID obsahují metadata, která na jiné zdroje odkazují, pak je s jejich pomocí možné vytvořit síť odkazů mezi danými objekty (např. vazba mezi



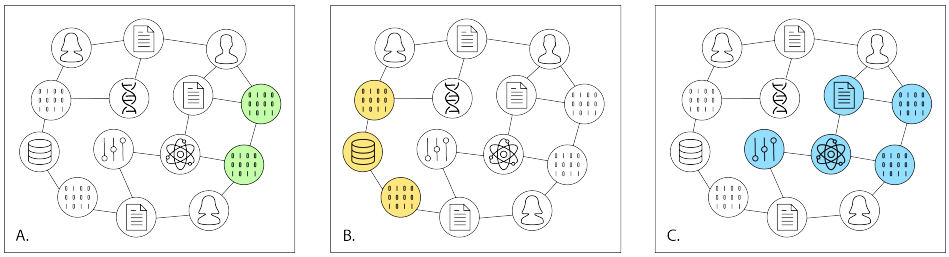
Obrázek 2: Tranzitivní vztah mezi objekty

konkrétním vědcem, zkoumanými datasety a výsledky)[6]. Tuto síť reprezentuje PID Graph.

Vytvořit takové vazby je v některých případech triviální, nicméně existuje řada případů použití, kdy je třeba použít silnější nástroj. Mezi nejčastější takové případy patří:

- Agregace citací pro všechny verze datasetu nebo zdrojového kódu. (A)
- Agregace citací pro datasety v konkrétním repozitáři. (B)
- Agregace citací pro vědecký výzkum (publikaci), tedy data produkovaná výzkumem, použité datasety, software, atd. (C)

Výsledný graf vystihující výše popsané situace je tvořen z odkazovaných objektů (uzly grafu) a jejich vzájemnými vztahy - případy použití (hrany grafu). Viz obrázek níže[7].



Obrázek 3: Ukázka grafu pro výše popsané případy použití

4.1 Případy použití

Implementaci PID grafu předcházela analýza jednotlivých uživatelských příběhů, jejíž cílem bylo zjistit nedostatky současných systémů pro koordinaci PID. Nakonec bylo identifikováno 48 případů, z nichž vyplynulo následující:

- Všechny z nich byly řešitelné PID grafem.
- Ačkoli byli uživatelé značně rozdílní (od výzkumníků po korporace), jejich požadavky byly velmi podobné.
- Vztah mezi dvěma objekty bylo možné vždy zredukovat na jednoduchou vazbu právě jednoho typu.

Následně bylo rozhodnuto o důležitosti implementovat PID graf jako standard, který bude snadno škálovatelný a udržovatelný jednotlivými PID Graph poskytovateli, mezi nimiž bude služba distribuována. Jako společný formát uchovávání dat byl zvolen RESTful JSON API, jako flexibilní jednotná varianta[8].

4.2 DataCite REST API

DataCite REST API je rozhraní navržené pro vyhledávání a zápis metadat, které jsou identifikovatelné pomocí DOI. Data ukládá a zobrazuje ve formátu JSON. Alternativami k REST API jsou MDS API, EZ API a OAI-PMH.[9]

4.2.1 Získávání přes DOI

DataCite REST API umožňuje dva hlavní způsoby získávání dat přes DOI, a to hledání jednotlivého objektu přes jeho DOI, anebo vyhledávání více zdrojů prostřednictvím dotazů směrovaným na DOI doménu. Získání zdrojů pomocí příkazové řádky může vypadat následovně:[10][11][12]

```
# vrátí zdroj s DOI 10.8438/0010
$ curl https://api.test.datacite.org/dois/10.5438/0012

# vrátí všechna DOI v doméně
$ curl https://api.test.datacite.org/dois/

# vyhledávání s použitím dotazů
$ curl https://api.test.datacite.org/dois
    ?query=climate%20change
    ?query=publicationYear:2016
    ?query=creators.familyName:mil*
```

4.2.2 Vytváření DOI

Nová DOI je možné vytvářet v doméně `https://api.datacite.org/dois`. V příkazové řádce by taková akce vypadala například:[13]

```
$ curl -X POST -H "Content-Type: application/vnd.api+json"
--user YOUR_REPOSITORY_ID:YOUR_PASSWORD -d @my_draft_doi.json
https://api.test.datacite.org/dois
```

Pro vytvoření DOI záznamu je potřeba připravit soubor ve formátu JSON obsahující konkrétní metadata (je možné pracovat i s metadaty v jiných formátech, podporovány jsou např DataCite XML nebo BibTex). Minimální obsah souboru vypadá následovně:

```
{
  "data": {
    "type": "dois",
    "attributes": {
      "doi": "10.5438/0012"
    }
  }
}
```

DOI server po přijmutí a schválení uživatelova vstupu odešle kompletní uchovaný záznam obsahující i prázná, původně nevyplněná pole. Mezi ně patří například:

```
"creators": [],
"titles": [],
"publisher": null,
"container": {},
"publicationYear": null,
"subjects": [],
"contributors": [],
"dates": [],
```

Nicméně JSON soubor v minimální formě nebude vyhledatelný, neboť nemá vyplněná nutná pole - DOI, creators, title, publisher, publicationYear, resourceTypeGeneral. Kompletní dohledatelné DOI může vypadat například takto:

```
{
  "data": {
    "id": "10.5438/0012",
    "type": "dois",
    "attributes": {
      "event": "publish",
      "doi": "10.5438/0012",
      "creators": [{
        "name": "DataCite Metadata Working Group"
      }],
      "titles": [{
        "title": "DataCite Metadata Schema Documentation"
      }],
      "publisher": "DataCite e.V.",
      "publicationYear": 2016,
      "types": {
        "resourceTypeGeneral": "Text"
      },
      "url": "https://schema.datacite.org/meta/index.html",
      "schemaVersion": "http://datacite.org/schema/kernel-4"
    }
  }
}
```

4.2.3 Získávání citací

Citace a další vztahy mezi zdroji je možné získávat pomocí DataCite Event Data. Pomocí Event Data lze specifikovat vztahy mezi zdroji, takové vztahy mohou být například verze (isVersionOf), granularita (isPartOf, isSupplementOf), financování (isFundedBy) nebo autorství (isAuthoredBy). Event Data v JSON souboru mohou vypadat následovně:[14]

```
"related_identifiers": [
    [{
      "relationType": "IsPartOf",
      "relatedIdentifier": "10.5438/0000-00ss",
      "resourceTypeGeneral": "Text",
      "relatedIdentifierType": "DOI"
    }]
]
```

4.2.4 Sledování změn metadat

DataCite REST API také umožňuje přidávat pole, jejichž účelem je dohledatelnost změn v metadatech, k nimž časem došlo. Lze zadávat například vytvoření (wasGeneratedBy), typ změny (action) nebo verze (version). Ukázka v JSON formátu:[15]

```
"attributes": {
"prov:wasGeneratedBy":
    "https://api.datacite.org/activities/a123b456",
"prov:generatedAtTime": "2019-03-28T20:58:22.251Z",
"prov:wasDerivedFrom": "https://doi.org/10.5438/jwvf-8a66",
"prov:wasAttributedTo":
    "https://api.datacite.org/providers/admin"
}
```

4.3 GraphQL

GraphQL je dotazovací jazyk vyvinutý pro vyhledávání (zejména) metadat. GraphQL API využívají data uložená online prostřednictvím `https://api.datacite.org/graphql` interface.

Ačkoli DataCite REST API dobře popisuje jednotlivé zdroje a jejich propojení pomocí DataCite Event Data, tak neumožňuje efektivní odpovídání na dotazy spojené s propojením takových zdrojů. GraphQL jako součást PID Grafu přináší řešení na tento problém.

GraphQL jako součást své funkcionality přináší dovednost specifikovat pole a vazby v dotazech, podporu PID třetích stran, řadu vývojářských nástrojů a pomocných knihoven.[16]

Ukázka GraphQL dotazu:

```
{
  funder(id: "https://doi.org/10.13039/501100000780") {
    name
    alternateName
    datasets(first: 10, after: "Mg") {
      edges {
        relationType
        source
        cursor
        node {
          id
          titles {
            title
          }
          relatedIdentifiers {
            relatedIdentifier
            relationType
          }
          fundingReferences {
            awardTitle
            awardNumber
          }
        }
      }
    }
  }
}
```

4.4 Implementace

První implementace PID grafu byla dokončena v roce 2018 organizací DataCite, která rozšířila již existující službu Event Data Service, na níž v minulosti spolupracovala s organizací Crossref.

5 Současný stav projektu

Práce týkající se návrhu a implementace jsou v projektu hotové. Hlavní starostí týmu ve vedení projektu je jeho udržitelnost a případná rozšiřitelnost. Otázky týkající se financování a spolupráci s investory, limitace prostředí a začlenění s ostatními systémy, výběr plánovací strategie a její implementace, jsou již zodpovězeny a publikovány. Stále však ještě zbývá vyřešit jak udržovat PID infrastrukturu nebo jak naložit s výsledky projektu.[17]

V březnu 2019 již projekt obsahoval přes 5,38 milionů vztahů mezi objekty, přičemž se odhadovalo, že při konstantním počtu objektů poroste až k 25 milionům. Projekt rovněž vyvíjí iniciativu směrem k ostatním organizacím směrujícím k témuž cíli a podporuje vzájemnou diskuzi.[8]

6 Závěr

Perzistentní identifikátory se jeví jako výborná cesta k správnému citování zdrojů, samy o sobě ale nemusí poskytovat takovou funkcionalitu, po jaké je v současné době poptávka. Projekt FREYA si vzal na svá bedra břímě vytvořit jednak platformu, prostřednictvím níž bude možné pracovat s citacemi a zdroji na komplexnější úrovni, a jednak komunitu, která bude výslednou platformu používat, popřípadě rozvíjet, a neskončí po vypršení grantu jako další zapomenutý neoblíbený standard. Jsou-li jimi uvedená čísla pravdivá, pak nezbývá než pogratulovat k naplnění vytyčeného cíle. Je tedy dosti pravděpodobné, že po dokončení stanoveného plánu, se minimálně v EU zrodí nová respektovaná platforma, již ocení nejen vědecká komunita.

Reference

[1] Project freya (@freya_eu) — twitter [image, online], [citováno 2019]. https://twitter.com/freya_eu.

[2] Doi [online], [citováno 2019]. https://www.crossref.cz/artkey/inf_000_0000_02_DOI.php.

[3] Project thor - technical and human infrastructure foropen research [online], [citováno 2019]. https://project-thor.eu/.

[4] The freya project - freya [online], [citováno 2019]. https://www.project-freya.eu/en/about/mission.

[5] Eosc — eosc portal [online], [citováno 2019]. https://www.eosc-portal.eu/about/eosc.

[6] The pid graph - freya [image, online], [citováno 2019]. https://www.project-freya.eu/en/pictures/slide1.png/@@images/38c67281-06ac-4d85-867b-67754c06a8c5.png.

[7] Pid graph [image, online], [citováno 2019]. https://blog.datacite.org/images/uploads/pid_graph_image.png.

[8] Amir Aryani Martin Fenner. Introducing the pid graph [online], [citováno 2019]. https://blog.datacite.org/introducing-the-pid-graph/.

[9] Datacite rest api guide [online], [citováno 2019]. https://support.datacite.org/docs/api.

[10] Retrieving a single doi [online], [citováno 2019]. https://support.datacite.org/docs/api-get-doi.

[11] Retrieving a list of dois [online], [citováno 2019]. https://support.datacite.org/docs/api-get-lists.

[12] Queries and filtering [online], [citováno 2019]. https://support.datacite.org/docs/api-queries.

[13] Creating dois with the rest api [online], [citováno 2019]. https://support.datacite.org/docs/api-create-dois.

[14] Retrieving citations and other relations [online], [citováno 2019]. https://support.datacite.org/docs/api-citations.

[15] Tracking metadata provenance [online], [citováno 2019]. https://support.datacite.org/docs/tracking-provenance.

[16] Datacite graphql api guide [online], [citováno 2019]. https://support.datacite.org/docs/datacite-graphql-api-guide.

[17] First annual report on pid commons and sustainability [pdf, online], 2018 [citováno 2019]. https://www.project-freya.eu/en/deliverables/freya_d6-1.pdf.

Seznam obrázků

Metadata

```
<dc:title>FREYA PID Graph</dc:title>
<dc:creator>Michal Hala</dc:creator>
<dc:subject>propojení perzistentních identifikátorů</dc:subject>
<dc:description>Esej obsahuje popis projektu FREYA a jeho
komponent, jehož cílem je vytvořit prostředí umožňující volné
sdílení dat a zlepšit možnosti citování komplexních zdrojů.
</dc:description>
<dc:date>2019-24-11</dc:date>
<dc:type>Text</dc:type>
<dc:format>public</dc:format>
<dc:identifier>https://www.project-freya.eu/en</dc:identifier>
<dc:language>cz</dc:language>
```