Otwarte Dane Połączone

Dominik Tomaszuk

Agenda

1. Dane
2. Otwarte Dane
3. Dane Połączone
4. Semantyczny Internet
5. Opisywanie zasobów w RDF
6. Linkowanie
7. Przypadek użycia: konwersja tabel do RDF
8. Podsumowanie z FAQ

Dane

Dane

Słownik języka polskiego odwołuje się do dwóch znaczeń.

Dane to:

- fakty, liczby, na których można się oprzeć w wywodach,
- informacje przetwarzane przez komputer.

Dane strukturalne i dane niestrukturalne

- **Dane niestrukturalne** to dane (zazwyczaj tekstowe lub graficzne), których struktura nie jest ustalona i nie może być w prosty sposób rozpoznana czy przetwarzana.
- **Dane strukturalne** to dane uporządkowane (zazwyczaj listy lub tabele), których struktura jest ustalona.

Otwarte Dane

Otwarte Dane

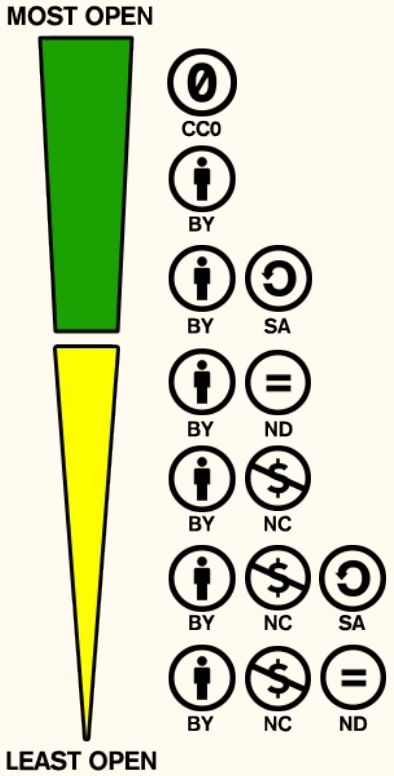
Otwarte Dane to idea, że niektóre dane powinny być swobodnie dostępne dla każdego do użytku i ponownego publikowania, jak chcą, bez ograniczeń praw autorskich, patentów lub innych mechanizmów kontroli.

Otwarte Licencje

Otwarta Licencja to rodzaj licencji, która bezpośrednio zezwala na kopiowanie oraz modyfikowanie utworu przez wszystkich, bez ograniczeń do poszczególnych osób (prawnych lub fizycznych), przy zachowaniu szczegółowych warunków wymaganych przez tę konkretną licencję.

Otwarte Licencje - przykłady

- **Licencje z rodziny Creative Commons**
- Design Science License
- Licencja FreeBSD Documentation
- Licencja GNU Free Documentation
- Free Art License (Licencja Licencja Wolnej Sztuki)
- Licencje do oprogramowania np. GNU GPL, GNU LGPL, Licencja X11 (MIT), BSD



Otwarte Formaty

Otwarty Format to format plików komputerowych, które w odróżnieniu od formatu zamkniętego, posiadają jawną, ogólnodostępną specyfikację oraz strukturę, która nie jest ograniczona w żaden sposób przez prawo związane z licencjonowaniem, patentami, znakami towarowymi lub w inny sposób powodując, że każdy może wykorzystać je bezpłatnie w dowolnym celu.

- Przykład formatu otwartego: CVS
- Przykład formatu zamkniętego: XLS

Dane które nie spełniają definicji otwartości

- Udostępnianie danych za opłatą,
- Dostęp do danych mają tylko zarejestrowani członkowie lub klienci,
- Korzystanie z zastrzeżonej lub zamkniętej technologii, które stanowią barierę dostępu,
- Patent zabraniający ponownego wykorzystywania danych,
- Ograniczony czasowo dostęp do zasobów.

Dane Połączone

Dane Połączone

Dane Połączone (ang. Linked Data) to **strukturalne dane**, które są powiązane z innymi danymi, dzięki czemu stają się bardziej przydatne dzięki zapytaniom semantycznym.

- Dane Połączone opierają się na standardowych technologiach internetowych, RDF i IRI (np. URL).
- Częścią wizji Danych Połączonych jest przekształcenie Internetu w globalną bazę danych.

Otwarte Dane Połączone

★ udostępnianie materiałów w Internecie (niezależnie od formatu) na podstawie otwartej licencji,

★★ jeśli to możliwe udostępnianie materiałów jako dane strukturalne (np. XLS zamiast zeskanowanego obrazu tabeli),

★★★ udostępnianie materiałów w otwartym formacie (np. CSV zamiast XLS),

★★★★ używanie IRI (np. adresów URL) do oznaczania materiałów, aby inni mogli wskazywać na nie,

★★★★★ używanie odnośników z danych do innych danych aby zapewnić lepszy kontekst.

Semantyczny Internet

Semantyczny Internet

Semantyczny Internet (ang. Semantic Web) to projekt, który ma przyczynić się do stworzenia i rozpowszechnienia standardów opisywania treści w Internecie w sposób, który umożliwi maszynom i programom przetwarzanie informacji w sposób odpowiedni do ich znaczenia.

- Wśród standardów Semantycznego Internetu znajdują się m.in. RDF.
- Znaczenia zasobów informacyjnych określa się za pomocą tzw. ontologii lub słowników.

RDF

RDF (Resource Description Framework) to model danych.

- Za pomocą RDF możemy tworzyć zdania opisujące dowolną rzecz.
- Te zdania są w postaci trójek.
- Trójka składa się z tematu, predykatu i obiektu.
- RDF jest podobny do języka naturalnego.

```
:Ala :ma :kota .
```

Formaty danych zgodne z RDF

RDF to nie format danych, tylko model danych.

RDF ma kilka formatów np. RDF/XML, Turtle, JSON-LD, RDFa i każdy z nich może opisać dokładnie to samo.

Każdy z tych formatów może być konwertowany do innego, bez żadnej straty.

Niezależnie od formatu liczba trójek (zdań) opisująca to samo będzie taka sama.

Ontologie i Słowniki

Ontologia lub **słownik** to formalna reprezentacja pewnej dziedziny wiedzy, na którą składa się zapis zbiorów pojęć i relacji między nimi.

- **Ontologia** zwykle zawiera bardziej dokładną i formalną reprezentację pewnej wiedzy i zwykle wyrażana jest w OWL.
- **Słownik** zwykle zawiera mniej dokładną i mniej formalną reprezentację i zwykle wyrażony jest w RDFS.
- Zarówno RDFS jak i OWL wykorzystują RDF i są częścią Semantycznego Internetu i Danych Połączonych.

Opisywanie zasobów

Metadane w RDF vs. Dane w RDF

Metadane to dane o danych czyli ustrukturalizowane informacje stosowane do opisu zasobów informacji lub obiektów informacji.

Metadane mogą być używane zarówno do opisu danych strukturalnych jak i niestrukturalnych.

Jednak dane niestrukturalne nie mogą być przetransformowane na RDF.

- Przykłady:
  - dane tabelaryczne (np. CSV) możemy w pełni przetransformować na RDF bo są strukturalne.
  - obrazy rastrowe (np. JPG) możemy tylko opisać metadanymi.

Gdzie szukać ontologii i słowników?

- AberOWL repozytory: http://aber-owl.net/#/
- Ontobee: http://www.ontobee.org/
- Ontology Lookup Service: https://www.ebi.ac.uk/ols/ontologies
- AmiGo2: http://amigo.geneontology.org/amigo
- i wiele wiele innych

Linkowanie

Piąta gwiazdka - `rdfs:seeAlso` i `owl:sameAs`

W języku RDF do linkowania z zewnętrznymi zasobami używamy głównie dwóch predykatów:

- `rdfs:seeAlso` - opisuje odnośnik czytelny dla człowieka np. https://pl.wikipedia.org/wiki/Dom
- `owl:sameAs` - opisuje odnośnik czytelny dla maszyny w formacie RDF np. http://www.wikidata.org/entity/Q255708

Gdzie szukać odnośników do pięciu gwiazdek?

`rdfs:seeAlso`:

- każda strona internetowa,
- zasoby międzydziedzinowe np. Wikipedia.

`owl:sameAs`:

- zasoby specyficzne dla danej domeny np. MeSH, PubMed,
- zasoby międzydziedzinowe np. Wikidata.

Przypadek użycia:
dane tabelaryczne do RDF

Instalacja

Wymagania

- Python: https://www.python.org/downloads/
- LibreOffice (lub MS Excel): https://pl.libreoffice.org/pobieranie/

Instalacja cow_csvw

- Pobrać narzędzie cow_csvw ze strony https://csvw-converter.readthedocs.io/en/latest/
- Wejść do linii poleceń (cmd) i wydać następujące komendy:
  - `virtualenv .`
  - `source bin/activate`
  - `pip install cow_csvw`

Konwersja XLS do CSV

Ten krok nie dotyczy przypadku gdy już mamy plik CSV.

1. Uruchomić plik XLS w LibreOffice Calc
2. [Krok opcjonalny] Zmienić strukturę tabeli zgodnie ze swoimi preferencjami
3. Zapisać jako plik CSV (Plik | Zapisz jako... | Tekst CSV)

Konwersja CSV do Turtle RDF

1. W linii poleceń wydać komendę: `cow_tool build PLIK.csv --base=http://uniwersytet.pl/`
   gdzie, `PLIK.csv` to plik CSV do konwersji, a `http://uniwersytet.pl/` to adres URL do danej instytucji
2. [Krok opcjonalny] W tym samym folderze stworzy się plik `PLIK.csv-metadata.json` gdzie można dodać swoje słowniki i ontologie
3. W linii poleceń wydać komendę: `cow_tool convert PLIK.csv --format=turtle`
4. W tym samym folderze pojawi się plik `PLIK.csv.ttl`
5. [Krok opcjonalny] można dodać nowe informacje w RDF w pliku `PLIK.csv.ttl` (np. autor itp)

Podsumowanie rekomendacji w odniesieniu do formatów plików

Podsumowanie

Typ	Format rekomendowany	Format niepreferowany	Rekomendowanie działania
Tabular data	CSV, TSV, SPSS portable	Excel	Przetransformować do RDF (jeśli się da)
Tekst	Plain text, HTML, RTF PDF/A only if layout matters	Word	HTML - można otagować za pomocą formatu RDFa (system to obsługuje) lub użyć tylko metadanych Reszta - tylko metadane
Media	Container: MP4, Ogg Codec: Theora, Dirac, FLAC	Quicktime H264	Tylko metadane
Obrazy	TIFF, JPEG2000, PNG	GIF, JPG	Tylko metadane
Dane ustrukturyzowane	XML, RDF	RDBMS	Są same w sobie opisane (będą w RDF), brak działań dodatkowych

Pytania i odpowiedzi

- **Pytanie 1:** Niektóre formaty plików pozwalają wstawiać metadane? Czy muszę to robić do każdego pliku ręcznie?

  Odpowiedź: Nie. Jeśli dane są zdeponowane w systemie to mają metadane generowane automatycznie.

- **Pytanie 2:** Czy dane strukturalne muszą być przetransformowanie do RDF?

  Odpowiedź: Tak. Dane strukturalne albo będą transformowane przez system automatycznie (np. XML) albo należy je przetransformować ręcznie (np. niektóre CSV)

Pytania i odpowiedzi

- **Pytanie 3:** Czy dane w CSV muszą być zamienione na RDF aby spełnić wymóg 4*?

  Odpowiedź: Tak, jeśli plik zawiera kontekst (np. ma zdefiniowane nagłówki, które coś oznaczają).

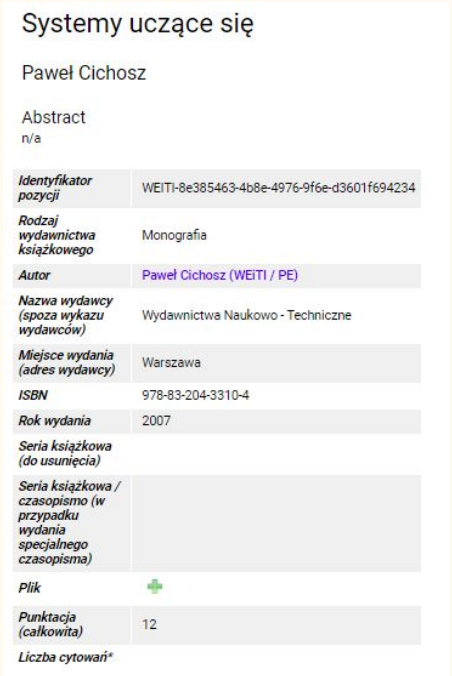
- **Pytanie 4:** CSV przekształcony do RDF ale bez odnośników wystarczy, jeśli metadane są w 5*?

  Odpowiedź: Tak. Dodatkowo warto pamiętać, że w tabelach są czasami adresy URL, co dobrze wzbogaca transformację (ale nie jest konieczne). Oczywiście zawsze można dodać nowe odnośniki ręcznie (ale nie jest to konieczne).

Pytania i odpowiedzi

- **Pytanie 5:** Dlaczego metadane bibliograficzne będą 5* z automatu i Partnerzy “nic nie muszą robić”?

  Odpowiedź: Warto podkreślić, że “metadane bibliograficzne” właściwie są danymi strukturalnymi (a nie metadanymi). Te dane mają strukturę, czyli wiemy, które słowo to imię autora, które słowo to nazwisko autora, która fraza to tytuł itp. dlatego system bez problemu będzie mógł wygenerować RDF z tych danych i z ich struktury.



Systemy uczące się

Paweł Cichosz

Abstract

n/a

Identyfikator pozycji	WEITI-8e385463-4b8e-4976-9f6e-d3601f694234
Rodzaj wydawnictwa książkowego	Monografia
Autor	Paweł Cichosz (WEITI / PE)
Nazwa wydawcy (spoza wykazu wydawców)	Wydawnictwa Naukowo - Techniczne
Miejsce wydania (adres wydawcy)	Warszawa
ISBN	978-83-204-3310-4
Rok wydania	2007
Seria książkowa (do usunięcia)	
Seria książkowa / czasopismo (w przypadku wydania specjalnego czasopisma)	
Plik	
Punktacja (całkowita)	12
*Liczba cytowań**	

```
@prefix schema: <https://schema.org/> .
@prefix fabio: <http://purl.org/spar/fabio/> .
@prefix xsd:   <http://www.w3.org/2001/XMLSchema#> .
@prefix rdfs:  <http://www.w3.org/2000/01/rdf-schema#> .
@prefix ite:   <http://w3id.org/sparql-generate/iter/> .
@prefix ns2:   <http://ii.pw.edu.pl/lib> .
@prefix prism: <http://prismstandard.org/namespaces/basic/2.0/> .
@prefix fun:   <http://w3id.org/sparql-generate/fn/> .
@prefix dc:    <http://purl.org/dc/elements/1.1/> .

<urn:example:WEITI-8e385463-4b8e-4976-9f6e-d3601f694234>
        a               schema:Book , fabio:Book ;
        dc:publisher    "Wydawnictwa Naukowo - Techniczne" ;
        dc:title        "Systemy uczące się" ;
        schema:isbn     "978-83-204-3310-4" .
```

Dziękuję. Pytania?