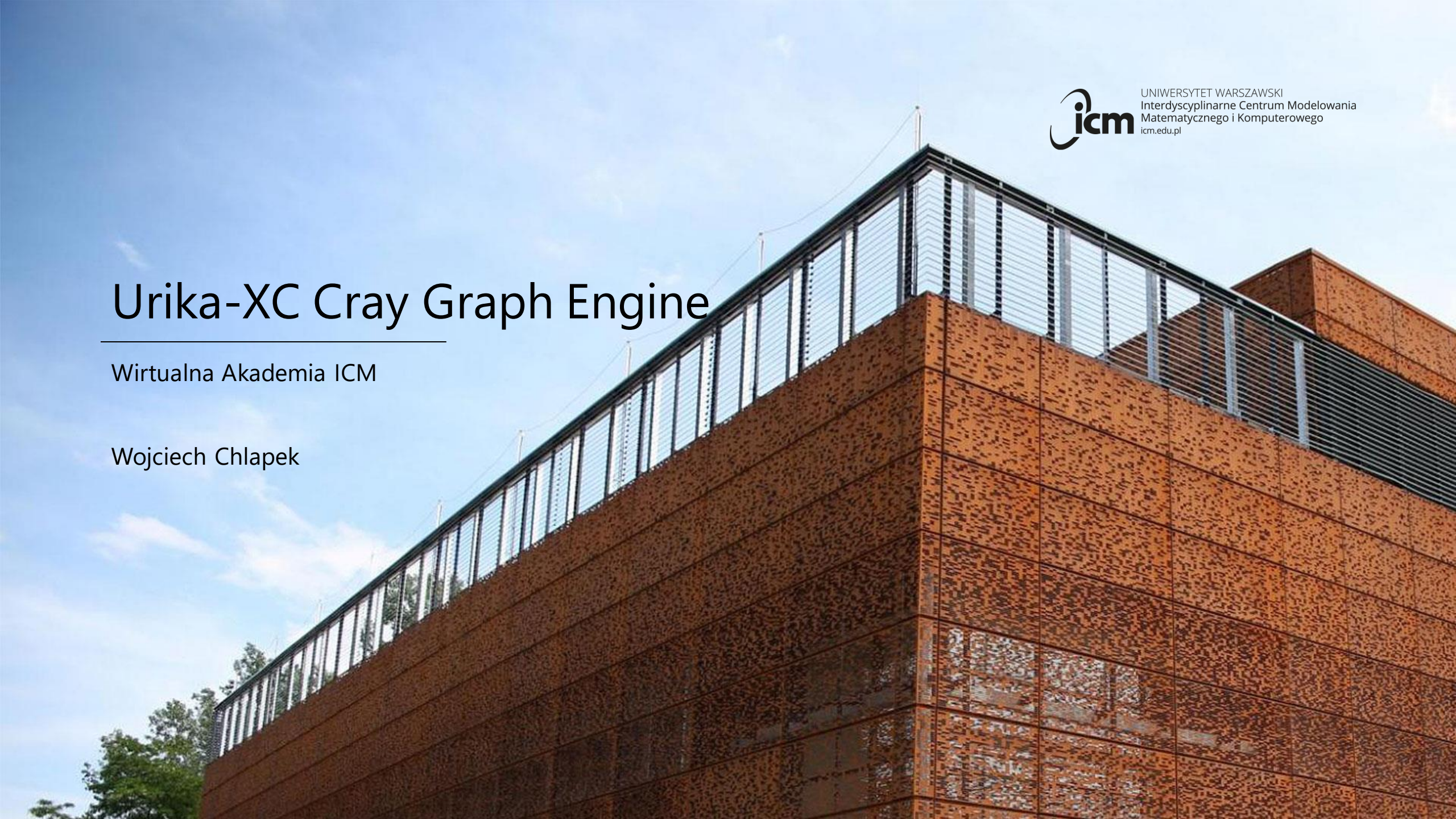


UNIWERSYTET WARSZAWSKI
Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego
icm.edu.pl

Urika-XC Cray Graph Engine

Wirtualna Akademia ICM

Wojciech Chlapek

Agenda

UNIWERSYTET WARSZAWSKI
Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego
icm.edu.pl

- Prezentacja:
  - Wprowadzenie do Semantic Web
  - Wprowadzenie do Cray Graph Engine (CGE)
- Zajęcia praktyczne:
  - Logowanie na serwer, ustawianie tuneli
  - Sposoby użycia CGE
  - SPARQL - przykłady i ćwiczenia
  - Wbudowane algorytmy grafowe - przykłady
- Prezentacja:
  - Podsumowanie
- Zajęcia praktyczne:
  - Wbudowane algorytmy grafowe - ćwiczenie dla chętnych
- Czas na eksperymenty dla uczestników

Urika-XC

UNIWERSYTET WARSZAWSKI
Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego
icm.edu.pl

- Pakiet oprogramowania do celów analityki oraz uczenia maszynowego
- Utworzony przez Cray
- Historycznie był ściśle związany ze specjalistycznym sprzętem produkowanym przez Cray (z architekturą Cray XMT)
- Obecnie dostępny na komputerach Cray z serii XC
- Komponenty:
  - Cray Graph Engine
  - Cray Programming Environment (PE) Deep Learning (DL) Plugin
  - Hyperparameter Optimization (HPO)
  - Open Source Analytics (OSA)

Dokumentacja pakietu Urika-XC

UNIWERSYTET WARSZAWSKI
Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego
icm.edu.pl

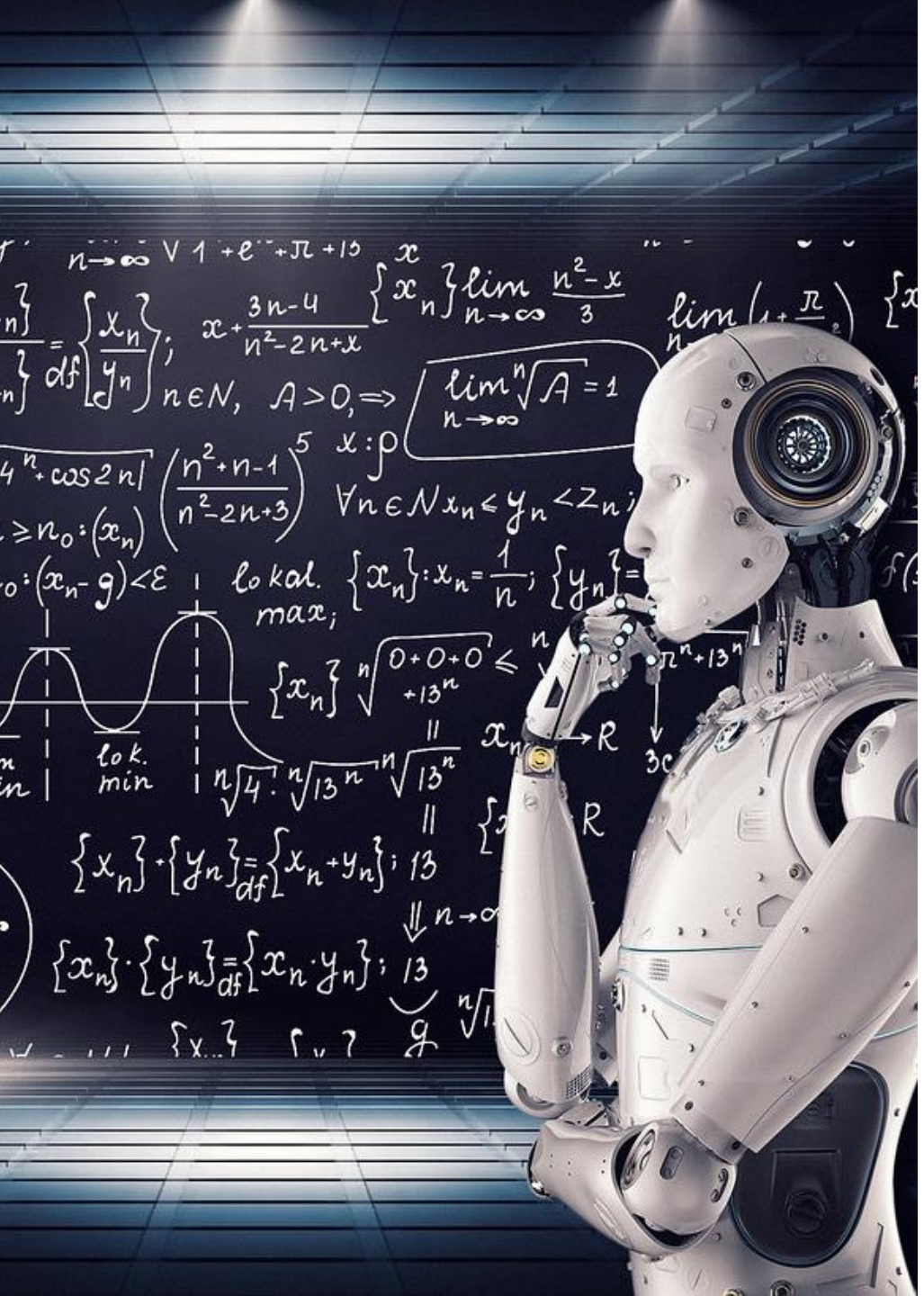
- Dostępna na stronie pubs.cray.com
- Uwaga: trudno się do niej dostać bezpośrednio ze strony cray.com
- Automatycznie generowane PDF-y mają niedoskonałości
- Prawdopodobnie w niedalekiej przyszłości strona internetowa z dokumentacją będzie jeszcze modyfikowana

Urika-XC w ICM

UNIWERSYTET WARSZAWSKI
Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego
icm.edu.pl

- ICM posiada jeden komputer z zainstalowanym pakietem Urika-XC - **Okeanos**
  - Cray XC40
  - > 1000 węzłów obliczeniowych
    - 2x 12-rdzeniowy procesor Intel w mikroarchitekturze Haswell
    - 128 GB RAM
    - Interconnect: Cray Aries
- Urika-XC jest dostępna domyślnie na wszystkich węzłach Okeanosa
- Zainstalowana wersja Cray Graph Engine: 3.2.1465

- ICM może zapewnić nie tylko dostęp do komputerów oraz oprogramowania, ale także **wsparcie techniczne**



[1]

UNIWERSYTET WARSZAWSKI
Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego
icm.edu.pl

Semantic Web

Zestaw standardów przechowywania i przetwarzania informacji, umożliwiający maszynom "zrozumienie" ludzkiej wiedzy

Fakty jako trójki

UNIWERSYTET WARSZAWSKI
Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego
icm.edu.pl

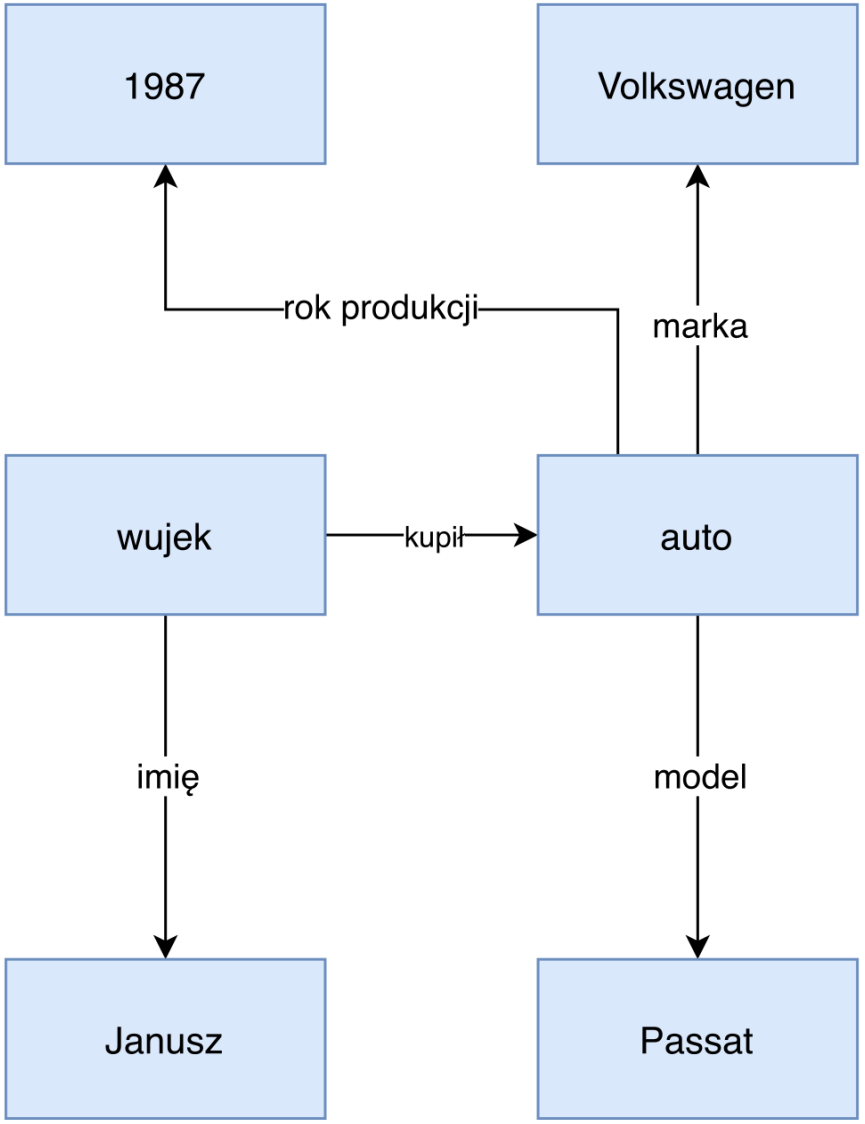
- Fakty można sprowadzić do bardzo prostych zdań składających się z trzech elementów:
- Podmiot orzeczenie przedmiot
  - Na przykład: Ala ma kota
- Zdania złożone można rozłożyć na kilka zdań prostych w identycznej konwencji
- Jak rozłożyć zdanie: "Wujek Janusz kupił Volkswagena Passata"?

Fakty jako trójki

UNIWERSYTET WARSZAWSKI
Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego
icm.edu.pl

- Fakty można sprowadzić do bardzo prostych zdań składających się z trzech elementów:
- Podmiot orzeczenie przedmiot
  - Na przykład: Ala ma kota
- Zdania złożone można rozłożyć na kilka zdań prostych w identycznej konwencji
- Jak rozłożyć zdanie: "Wujek Janusz kupił Volkswagena Passata"?
- Można np. tak:
  - Wujek kupił auto
  - Wujek ma na imię Janusz
  - Auto jest marki Volkswagen
  - Auto jest modelu Passat
- Orzeczenie nie musi być tutaj czasownikiem, jest raczej relacją między podmiotem a przedmiotem

[2]



Graf RDF

- *Resource Description Framework*
- Podmiot: wierzchołek wyjściowy
- Orzeczenie: krawędź
- Przedmiot: wierzchołek docelowy
- Z matematycznego punktu widzenia można powiedzieć, że graf RDF jest skierowanym grafem etykietowanym.

URL / URI / IRI

UNIWERSYTET WARSZAWSKI
Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego
icm.edu.pl

URL

- *Uniform Resource Locator*
- Służy do jednoznacznej identyfikacji zasobów w sieci
- Wskazuje sposób, w jaki można dostać się do danego zasobu
- Przykład: *https://icm.edu.pl/aktualnosci/*

URI

- *Uniform Resource Identifier*
- Identyfikuje w jednoznaczny sposób daną rzecz
- Nie musi wskazywać sposobu dostępu
- Pierwotnie pomyślany o zasobach dostępnych w sieci Web
- Można go wykorzystać dużo szerzej

IRI

- *Internationalized Resource Identifier*
- W przeciwieństwie do URI dopuszcza także znaki spoza ASCII
- **Używany w RDF**

Serializacja grafu RDF

UNIWERSYTET WARSZAWSKI
Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego
icm.edu.pl

- **RDF/XML**
- RDFa
- Notation3 (n3)
- **Turtle**
- **N-Triples / N-Quads**
- JSON-LD
- HexTuples
- HDT
- RDF Binary Thrift

RDF kontra relacyjne bazy danych

UNIWERSYTET WARSZAWSKI
Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego
icm.edu.pl

- RDF jest dużo bardziej elastyczny, ponieważ nie zakłada żadnego schematu danych a priori
- Dobrze nadaje się do sprowadzenia wiedzy z różnych źródeł w jedno miejsce
  - Wymagane jest wtedy ustalenie mapowania wiedzy na graf RDF
  - Konieczna również sensowna "polityka" nadawania IRI
- Zapytania na grafie RDF będą generalnie wolniejsze, niż w relacyjnej bazie danych zoptymalizowanej pod kątem takich zapytań
- Wnioski:
  - Relacyjne bazy danych będą lepsze do dobrze zdefiniowanych procesów biznesowych
  - Graf RDF sprawdzi się lepiej, jeśli nie wiadomo z góry, jakie wzorce będą sprawdzane, jakie dane będą dostępne itp.

SPARQL

UNIWERSYTET WARSZAWSKI
Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego
icm.edu.pl

- Język zapytań do grafu RDF
- Analogiczne znaczenie i zastosowanie jak dla SQL w relacyjnych bazach danych
- Głównie użyteczny do wyszukiwania wzorców
- Posiada funkcje agregujące
- Daje możliwość modyfikacji grafu (SPARUL)

Semantic Bible

UNIWERSYTET WARSZAWSKI
Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego
icm.edu.pl

- Przykładowy graf RDF
  - Nawet coś więcej: ontologia (OWL)
- Nazwane obiekty z Nowego Testamentu
- Stosunkowo niewielki
- Dobry do zastosowań demonstracyjnych

CGE: Obsługa SPARQL

UNIWERSYTET WARSZAWSKI
Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego
icm.edu.pl

- SPARQL 1.1
- Rozszerza ten język o implementacje algorytmów grafowych oraz o własne funkcje wbudowane
- Nieobsługiwane: funkcje skrótu, słowo kluczowe SERVICE
- Częściowo obsługiwane: SPARQL Property Path, funkcje UCASE i LCASE

CGE: Wbudowane algorytmy grafowe

UNIWERSYTET WARSZAWSKI
Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego
icm.edu.pl

- Bad Rank
- Betweenness Centrality
- Community Detection: Label Propagation
- Community Detection: Parallel Louvain Method
- Page Rank
- S-T Connectivity
- S-T Set Connectivity
- Triangle Counting
- Vertex Triangle Counting
- Triangle Finding

CGE: Funkcje wbudowane

UNIWERSYTET WARSZAWSKI
Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego
icm.edu.pl

- Funkcje do obliczeń na przedziałach matematycznych
- Obliczenia na Wielkim Kole (odległości między punktami na Ziemi)
- Pierwiastek kwadratowy
- Dodatkowe funkcje agregujące:
  - Wariancja
  - Odchylenie standardowe
  - Średnia geometryczna
  - Moda
  - Mediana

CGE: Format danych wejściowych

UNIWERSYTET WARSZAWSKI
Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego
icm.edu.pl

- CGE obsługuje tylko N-Triples / N-Quads
- Grafy RDF zapisane w innym formacie można skonwertować z wykorzystaniem odpowiedniego narzędzia konwersji
  - np. RIOT z pakietu Apache Jena
- Chociaż CGE obsługuje wyłącznie grafy RDF, można rozważyć konwersję grafu innego rodzaju do RDF

CGE: Ładowanie danych

UNIWERSYTET WARSZAWSKI
Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego
icm.edu.pl

- W najprostszym scenariuszu: wszystkie dane umieszczone w pliku dataset.nt
- W razie takiej potrzeby można też użyć wielu plików
- Przy pierwszym uruchomieniu baza jest kompilowana
- Następnie graf jest **ładowany do pamięci**

- Uwaga: zmiany wprowadzone po kompilacji w pliku dataset.nt bądź innych plikach "źródłowych" **nie zostaną zauważone automatycznie**
- Uwaga: zmiany wprowadzone w bazie poprzez zapytania SPARUL **nie zostaną zapisane na dysku**, jeśli nie zostanie wykonany tzw. checkpoint

Jak użyć CGE?

UNIWERSYTET WARSZAWSKI
Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego
icm.edu.pl

- cge-cli
- CGE GUI
- CGE GUI jako SPARQL Endpoint
- CGE API
  - Java
  - Python
  - Spark

Część praktyczna

- Logowanie na serwer, ustawianie tuneli
- Sposoby użycia CGE
- SPARQL - przykłady i ćwiczenia
- Funkcje wbudowane - przykłady

Apache Jena Fuseki

UNIWERSYTET WARSZAWSKI
Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego
icm.edu.pl

- Serwer SPARQL
- Narzędzie analogiczne do Cray Graph Engine
- Bardzo dobre rozwiązanie do nauki SPARQL oraz eksperymentów na (stosunkowo niewielkich) grafach RDF na własnym komputerze
- Może okazać się wystarczająco dobry, jeśli:
  - Używane grafy są małe
  - Nie ma potrzeby wykorzystania algorytmów grafowych ani funkcji wbudowanych CGE

Tematy nieomawiane na warsztatach

UNIWERSYTET WARSZAWSKI
Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego
icm.edu.pl

- Sposoby konwersji innych grafów na RDF
- CGE API
- Bezpieczeństwo CGE
- Alokacja większych / mniejszych zasobów
- Checkpointy
- SPARQL Property Path
- SPARQL Update
- Wnioskowanie
- ...

Źródła obrazków

UNIWERSYTET WARSZAWSKI
Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego
icm.edu.pl

- flickr.com
  - [1] "Artificial Intelligence & AI & Machine Learning" by mikemacmarketing
  - [2] "Volkswagen Passat Variant" by nakhon100
  - [3] "Launching an Experiment" by myfuture.com

UNIWERSYTET WARSZAWSKI
Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego
icm.edu.pl

Dziękuję za uwagę

Dane kontaktowe:

- Cray Graph Engine: Wojciech Chlapek | wc425947@icm.edu.pl
- Trovares xGT: Jakub Jałowiec | jj358817@icm.edu.pl