

Otvoreno računarstvo

4b. Otvoreni podaci

Creative Commons



[Otvoreno računarstvo 2022/23](#) by Ivana Bosnić & Igor Čavrak, FER
is licensed under [CC BY-NC-SA 4.0](#)

Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

This license requires that reusers give credit to the creator.

It allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, for noncommercial purposes only.

If others modify or adapt the material, they must license the modified material under identical terms.

BY: Credit must be given to you, the creator.

NC: Only noncommercial use of your work is permitted.

SA: Adaptations must be shared under the same terms.

Otvoreno računarstvo

4b. Otvoreni podaci

- **Metapodaci – rječnici**
- Metapodaci – zapisi
- Modeli i okviri za vrednovanje skupova podataka

Podsjetnik na razno-razne principe otvorenih podataka

- Potpuni
 - Primarni
 - Pravovremeni
 - Pristupačni
 - Strojno čitljivi
 - ...
- Trajni
 - Provjereni
 - Dokumentirani
 - ...
 - Sveobuhvatni
 - Iskoristivi
 - Interoperabilni
 - ...

Metapodaci

- Podaci o podacima
- Podaci koji detaljno opisuju skup podataka, njegovu strukturu i pojedine dijelove
- „Strukturirani podaci koji opisuju karakteristike entiteta koji nose informacije, a koji pomažu u identifikaciji, otkrivanju, vrednovanju i upravljanju opisanim entitetima.” *(jedna od skupnih definicija)*
- Trebaju biti i ljudima čitljivi i **strojno čitljivi**
- Namjena:
 - Lakše otkrivanje i pristupačnost podataka
 - Lakše tumačenje podataka
 - Lakše povezivanje, interoperabilnost i razmjena podataka
 - Lakša analiza kvalitete podataka

Metapodaci

- **Metapodaci o skupu podataka**
 - Naziv, autor (tijelo, organizacija, osoba)
 - Opis, ključne riječi
 - Tehničke karakteristike (formati podataka)
 - Poveznice
 - Datumi objave, posljednje izmjene, učestalost ažuriranja
 - Dostupnost (licencije)
- **Metapodaci o strukturi skupa podataka**
 - Gramatika dokumenta, hijerarhija i odnosi elemenata
 - Veze među podacima, unutar i izvan skupa
- **Metapodaci o pojedinim dijelovima skupa podataka**
 - Opisivanje pojedinih elemenata – po mogućnosti računalno čitljivo
 - Rječnici

Česti problemi s metapodacima

- Poklanja im se vrlo malo pažnje, smatra ih se nevažnima
- Često se uopće ni ne objavljuju
- Objavljeni metapodaci su nepotpuni
- Objavljeni metapodaci nisu u niti jednom standardiziranom formatu
- Objavljeni metapodaci uopće nisu ili su jako loše dokumentirani
- Loši alati za rad s metapodacima

- Skupa izrada

Opet upute EU...

- Smjernice o preporučenim standardnim dozvolama, skupovima podataka i naplati ponovne uporabe dokumenata
 - [https://eur-lex.europa.eu/legal-content/HR/TXT/?uri=CELEX%3A52014XC0724\(01\)](https://eur-lex.europa.eu/legal-content/HR/TXT/?uri=CELEX%3A52014XC0724(01))
- *„Kako bi se olakšala **upotreba podataka** u javnom sektoru uz znatno povećanje vrijednosti skupova podataka za **naknadnu ponovnu uporabu**, preporučuje se da su skupovi podataka:*
 - ...
 - opisani u **bogatim metapodatkovnim formatima** i klasificirani prema **standardnim rječnicima** (DCAT, [EUROVOC](#), ADMS itd.), kako bi se olakšali pretraživanje i interoperabilnost
 - praćeni **objašnjenjima** o upotrijebljenim metapodacima i kontroliranim rječnicima, kako bi se promicala interoperabilnost baza podataka
 - ...

Slijedi:

- Rječnici

- Dublin Core / DCMI Metadata Terms
- Data Catalog Vocabulary – DCAT
- Schema.org
- ...

- Formati zapisa metapodataka:

- Microdata
- RDFa / RDFa Lite
- JSON-LD
- ...
- (Microformats)

- Sve vodi prema *Linked data*:

- RDF / RDFa
- JSON-LD
- ...

Dublin Core

- **Dublin Core (DC)**
 - DCMI – Dublin Core Metadata Initiative
- Jednostavan model metapodataka
 - Osnovni standard - 15 elemenata
 - Kasnija proširenja za pojedine svrhe ili zemlje – aplikacijski profili
- Osnovno se koristi za opis razne skupove podataka
 - izvorno ga osmislili knjižničari
- **Simple Dublin Core / Dublin Core Metadata Element Set - 15 elemenata:**
 - Creator, Contributor, Publisher, Title, Date, Language, Format, Subject, Description, Identifier, Relation, Source, Type, Coverage, Rights

DCMI Metadata Terms

- DCMI Metadata Terms = Dublin Core + dodaci, koji mogu biti:
 - Svojstvo (*property*)
 - Razred (*class*)
 - Tip podatka (*datatype – syntax encoding schemes*)
 - Shema zapisa rječnika (*vocabulary encoding schemes*)
- Ostvareni kao RDF rječnici – za povezivanje podataka
 - Ako se ne koristi RDF, mogu se koristiti npr. svojstva i slično, kroz JSON, XML, ...
- Svaki pojam (*term*) mora imati:
 - Naziv, labelu, URI, definiciju, tip (property, class, datatype, vocabulary encoding scheme)
- 4 DCMI prostora imena / konteksta / prefiksa
- <https://www.dublincore.org/specifications/dublin-core/dcmi-terms/>

Rječnik DCAT

- Data Catalog Vocabulary
- *W3C Recommendation*
- Često korišten za otvorene podatke Europske unije
- Također ostvaren kao RDF – povezivanje podataka

- Profil - posebna prilagodba određenoj zemlji ili namjeni
 - Posebne klase i svojstva
 - Prilagođena kardinalnost
 - Posebni rječnici za vrijednosti svojstava
 - Posebni načini povezivanja podataka
 - **DCAT-AP:** Application Profile for data portals in Europe
 - **DCAT-US:** Application Profile for data portals in US

Osnovne klase u rječniku DCAT

- dcat:Catalog
- dcat:Resource
- dcat:Dataset
- dcat:Distribution
- dcat:DataService
- dcat:CatalogRecord

<https://www.w3.org/TR/vocab-dcat/#dcat-scope>

- ... I pripadajuće podklase i svojstva

Schema.org

- Nezavisna specifikacija rječnika, čije se rasprave održavaju unutar W3C-a
 - Počeli zajedno Bing, Google i Yahoo :-)
- Definira tipove podataka i njihova svojstva
 - 800+ tipova
 - 1300+ svojstava
 - Dodatni tipovi koje razvija zajednica – Extensions
- Organizacija tipova podataka
 - <https://schema.org/docs/schemas.html>
 - PRIMJERI!
 - Creative works: [CreativeWork](#), [Book](#), [Movie](#), [MusicRecording](#), [Recipe](#), [TVSeries](#) ...
 - Embedded non-text objects: [AudioObject](#), [ImageObject](#), [VideoObject](#)
 - [Event](#)
 - Health and medical types: notes on the health and medical types under [MedicalEntity](#)
 - [Organization](#)
 - [Person](#)
 - [Place](#), [LocalBusiness](#), [Restaurant](#) ...
 - [Product](#), [Offer](#), [AggregateOffer](#)
 - [Review](#), [AggregateRating](#)
 - [Action](#)

Schema.org – svježí primjer

- Schema.org COVID-19: US CDC Data Table fields
 - Reprezentacija podataka američkog Centra za kontrolu i prevenciju bolesti
 - Stanje u pojedinoj bolnici – broj respiratora, broj slobodnih kreveta, ...
 - <https://schema.org/docs/cdc-covid.html>
- Schema.org Covid TestingFacility
 - <https://schema.org/CovidTestingFacility>

Schema.org

- Kako zapisati rječnik Schema.org u računalno čitljivim formatima?
 - Tj. kako serijalizirati njegove objekte?
- Microdata
- RDFa / RDFa Lite
- JSON-LD (JSON Linked Data)
 - Primjer za usporedbu: <https://en.wikipedia.org/wiki/Schema.org#Examples>

Otvoreno računarstvo

4b. Otvoreni podaci

- Metapodaci – rječnici
- **Metapodaci – zapisi**
- Modeli i okviri za vrednovanje skupova podataka

Microdata

- Dio specifikacije HTML Living Standard
 - <https://html.spec.whatwg.org/multipage/microdata.html>
- W3C Working Draft – 2018.
- Nadogradnja specifikacije HTML5 – novi atributi unutar dokumenta
- Koji rječnik koristiti?
 - Nije bitno – **Schema.org** je uobičajen, no mogu se izrađivati vlastiti
 - Mogu se koristiti neki drugi standardi, npr. vCard
 - ...
- Vrlo brzi uvod u Microdata + Schema.org:
 - <https://schema.org/docs/gs.html>

Microdata - atributi

- **itemscope** - boolean atribut u određenom elementu, identificira element koji sadrži neku stavku
- **itemtype** - popis apsolutnih URL-ova koji identificiraju tip stavke u tom rječniku
- **itemprop** - naziv svojstva, čija je vrijednost unutar elementa
- **itemid** - apsolutni URL koji globalno identificira stavku
- **itemref** - referenca identifikatora drugih elemenata, koji sadrže svojstva korištena u ovoj stavci. Zabranjene cikličke reference
- Puno primjera: HTML Living Standard: 5.1 Introduction
 - <https://html.spec.whatwg.org/multipage/microdata.html#introduction-7>

Microdata - svojstva

- Microdata nema tipove podataka, osim text i URL
 - U slučaju korištenja sa Schema.org, tamo postoje tipovi
- Nema dobru podršku za internacionalizaciju
 - RDFa je u tome puno bolji
- Primarno namijenjen web-stranicama
 - HTML-atributi!
- Za skupove podataka najprikladniji je JSON-LD

RDFa / RDFa Lite

- RDFa = Resource Description Framework in **A**tttributes
- RDFa Lite = W3C preporuka – minimalni podskup specifikacije RDF
 - Nekoliko atributa za označavanje računalno čitljivih podataka u formatima HTML/SVG/XML
 - Alternativa specifikaciji Microdata
- RDF je namijenjen povezivanju podataka, uz trojke
 - Subjekt – Predikat - Objekt
 - Više o RDF-u u poglavlju o povezanim otvorenim podacima (Linked Open Data)

RDFA Lite atributi

- vocab – korišteni rječnik (npr. Schema.org)
- typeof – tip podatka
- property - svojstvo

JSON-LD

- JavaScript Object Notation for Linked Data
 - Uobičajena JSON datoteka, uz dodatne ključne riječi
 - <https://json-ld.org/>
- W3C Recommendation
- Posebno prikladan za označavanje skupova podataka
 - Ionako ih je često moguće preuzeti u JSON-obliku
 - Metapodaci mogu biti uključeni direktno uz podatke
 - Ili navedeni u zasebnoj datoteci
- Koncept „konteksta” – povezivanje s određenim rječnikom, npr. Schema.org
 - Kao i RDF: Subjekt – Predikat – Objekt
 - Više o JSON-LD-u u poglavlju o povezanim otvorenim podacima (Linked Open Data)
- Internet media type:
 - Application/ld+json

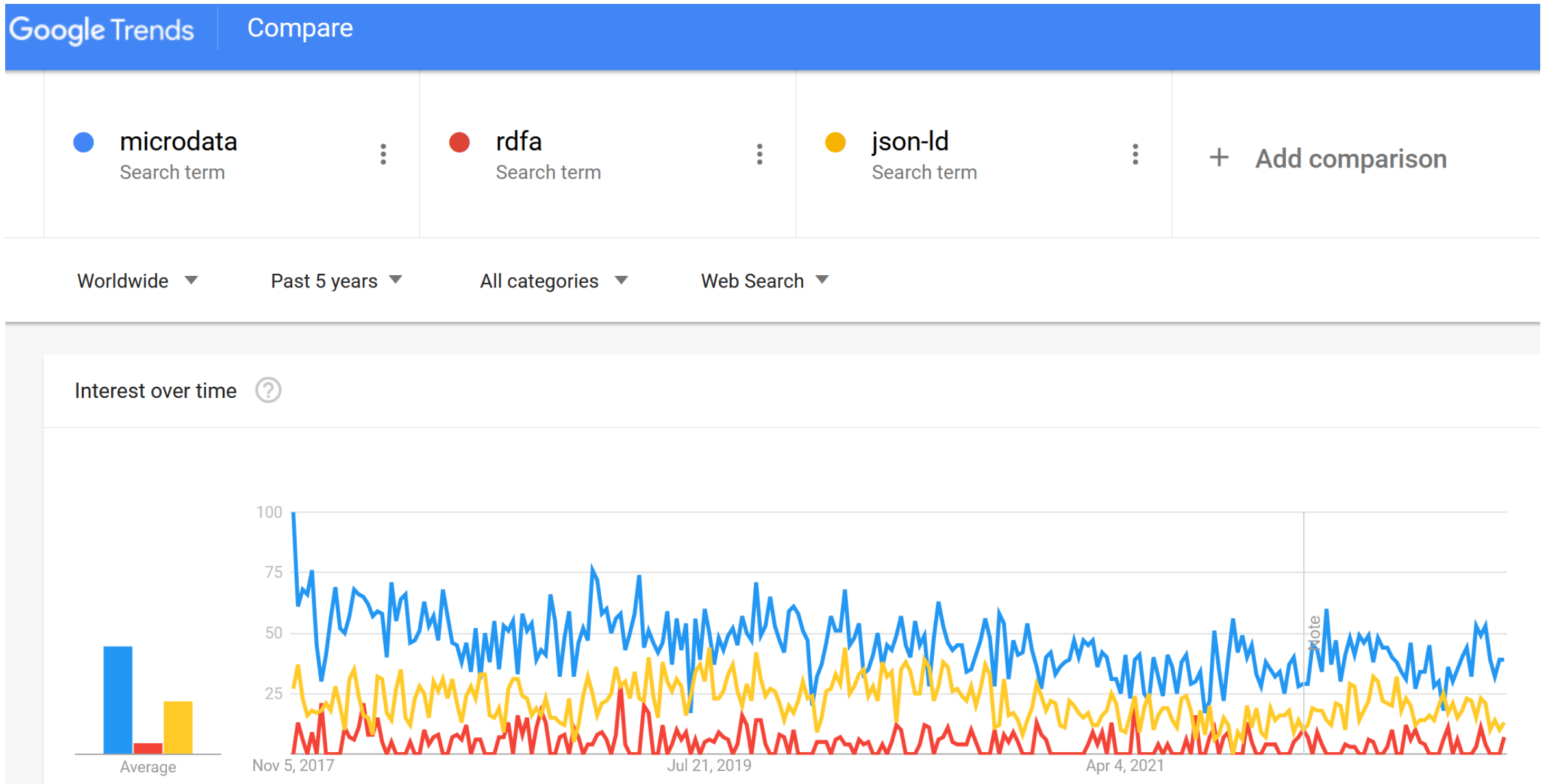
Svojstva

- @context – mapiranje na rječnik
 - @type – tip podatka, konkatena se na @context
 - @value – vrijednost podatka, ako se radi o podatku posebnog tipa (npr. datum)
 - @id – IRI identifikator
-
- Koristi IRI (Internationalized Resource Identifier) za identifikaciju podataka
 - Otprilike kao URI, uz nešto manje dostupnih znakova

Uključivanje drugog konteksta – nepovezanih shema

- Važno za metapodatke i njihovu interoperabilnost!
- Definirani prefiksi, vrijede i u Schema.org i u RDFa i u JSON-LD
- Npr.
 - dc – Dublin Core
 - dcat – Data Catalog Vocabulary (DCAT)
 - cc – Creative Commons
 - ...
- Popisi:
 - Schema.org zapisan u JSON-LD: <https://schema.org/docs/jsonldcontext.json>
 - RDFa Core Initial Context - <https://www.w3.org/2011/rdfa-context/rdfa-1.1>

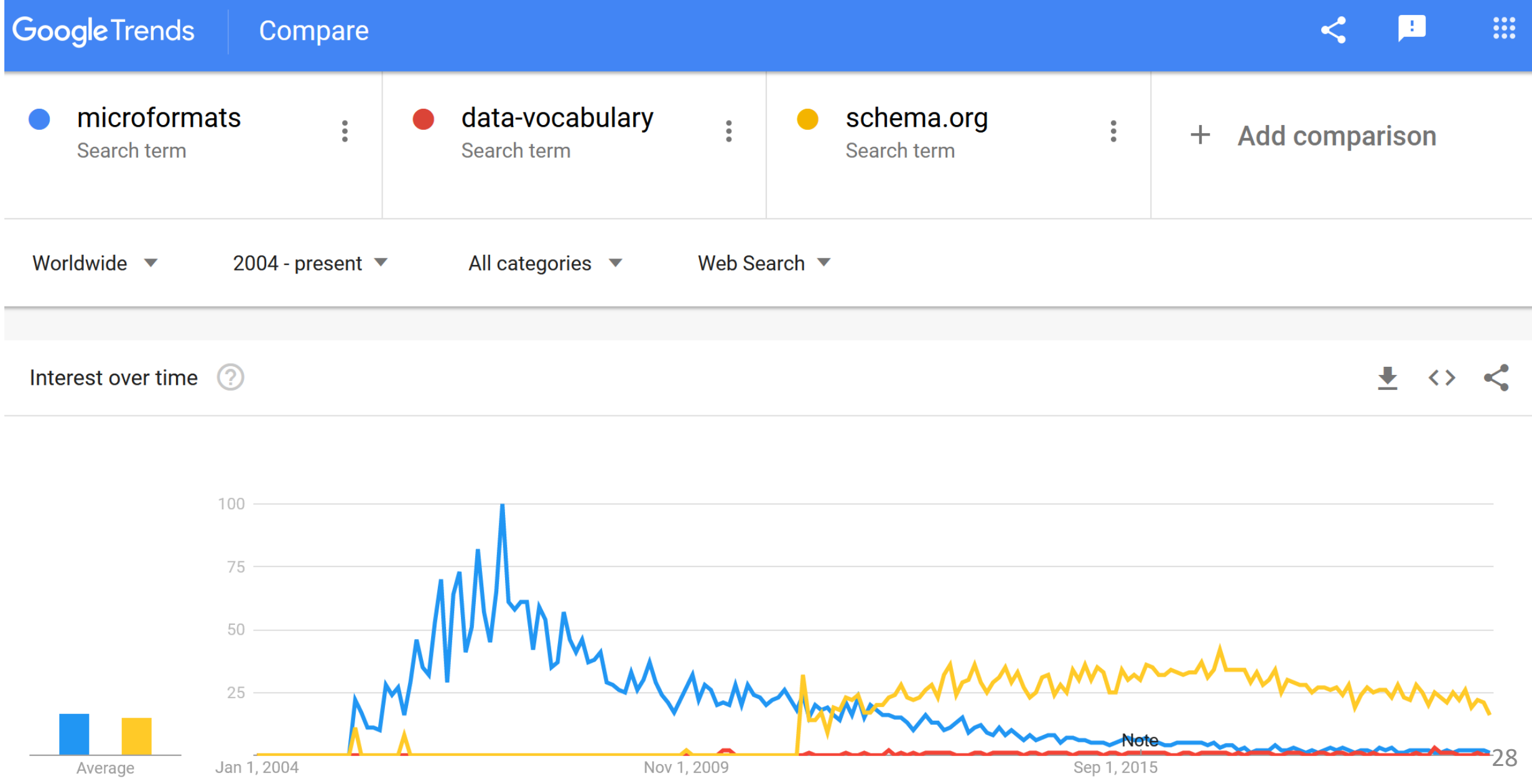
Microdata vs RDFa vs JSON-LD



Microformats

- Samo ćemo ih spomenuti, ako negdje naletite na njih
 - Nisu nam toliko važni
 - Prilično su zastarjeli
 - Ne djeluju u sklopu ikakvog normirajućeg tijela – skupina okupljena oko jednog Wikija
 - Procjena: izumrijet će – ako već nisu
- Slični su specifikaciji Microdata
 - Mali broj specifikacija, desetak
- Koriste uobičajene HTML-atribute: class i rel
- Za razliku od Microdata, korisnici dugo vremena nisu mogli razviti svoje mikro-definicije formata
 - U Microformats2 može se koristiti prefiks za vlastite definicije
- <http://microformats.org/wiki>

Usporedba microformats vs schema.org



Ali kakve veze ovo ima s otvorenim podacima?

- Ima! :-)
- Još uvijek gradimo metapodatke
 - za bolju vidljivost, dostupnost, pristupačnost
 - za lakše povezivanje
- Schema.org Dataset - opis skupa podataka
 - <https://schema.org/Dataset>
- Google Dataset Search
 - <https://datasetsearch.research.google.com/>
- Schema.org Dataset @ Google Dataset Search:
 - <https://developers.google.com/search/docs/data-types/dataset>
- Primjer: NCDC Storm Events Database <- **VIEW SOURCE**
 - <https://catalog.data.gov/dataset/ncdc-storm-events-database-336e4>

Otvoreno računarstvo

4b. Otvoreni podaci

- Metapodaci – rječnici
- Metapodaci – zapisi
- **Modeli i okviri za vrednovanje skupova podataka**

Podsjetnik na razno-razne principe otvorenih podataka

- Potpuni
 - Primarni
 - Pravovremeni
 - Pristupačni
 - Strojno čitljivi
 - ...
- Trajni
 - Provjereni
 - Dokumentirani
 - ...
 - Sveobuhvatni
 - Iskoristivi
 - Interoperabilni
 - ...

Kvaliteta podataka

- Eurostat
 - Relevantnost
 - Točnost i pouzdanost
 - Pravovremenost i točnost u objavi
 - Koherentnost i mogućnost usporedbe
 - Dostupnost i jasnoća

Modeli vrednovanja skupa otvorenih podataka

- Različite vrste modela i okvira (*framework*)
- Razlikuju se po:
 - Namjeni
 - Za državne podatke
 - Za organizacije
 - ...
 - Fokusu na određeni dio
 - Zrelost
 - Povezanost podataka
 - Dostupnost
 - Kvalitetu
 - ...
 - Metodama vrednovanja
 - Teme / Aktivnosti
 - Razine ostvarenosti
 - Ključni indikatori / metrike / dokazi
- Primjeri:
 - 5-star open data model
 - ODI Open Data Maturity Model
 - Model zrelosti otvorenih državnih podatak za prioritizaciju odluka
 - Okvir portala European Data Portal za vrednovanje zrelosti

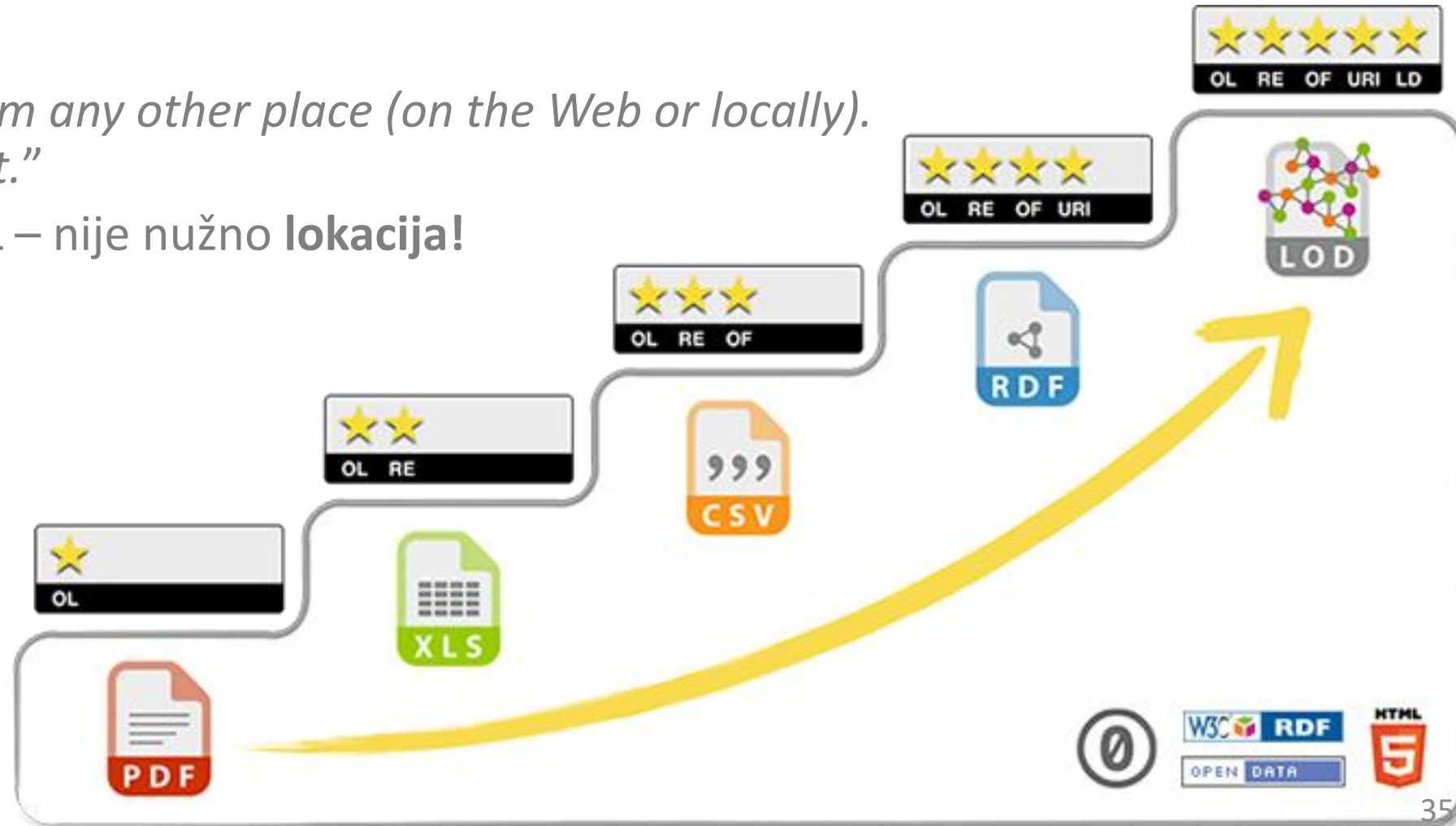
Model „5 zvjezdica otvorenih podataka” – 5-star model

- [Izvorni tekst](#) - autor: Tim Berners-Lee
 - začetnik weba i ideje povezanih podataka
- **1:** omogućite **dostupnost** podataka na webu **u bilo kojem obliku** pod **otvorenom licencijom**
- **2:** omogućite **dostupnost računalno čitljivih strukturiranih** podataka
(npr. Excel umjesto skenirane slike tablice)
- **3:** koristite **otvorene formate zapisa**
(npr. CSV umjesto XLS-a)
- **4:** koristite **otvorene norme W3C-a** za **označavanje** pojmova tako da drugi mogu izravno pristupiti vašim podacima
- **5:** **povežite svoje podatke s drugim podacima** za pružanje konteksta



5-star open data model

- Troškovi, prednosti i **primjeri - PROUČITI:**
 - <https://5stardata.info> <- dostupno i na hrvatskom jeziku
- Napomena:
 - „You can link to it from any other place (on the Web or locally).
You can bookmark it.”
 - URI nije isto što i URL – nije nužno **lokacija!**




ODI Open Data Maturity Model

- Model vrednovanja koliko dobro organizacija objavljuje i koristi otvorene podatke
- 5 većih tema podijeljenih u aktivnosti
 - Data management processes
 - Knowledge & skills
 - Customer support & engagement
 - Investment & financial performance
 - Strategic oversight
- 5 razina napretka
 1. Initial
 2. Repeatable
 3. Defined
 4. Managed
 5. Optimising
- [Vodič i okvir provedbe](#)

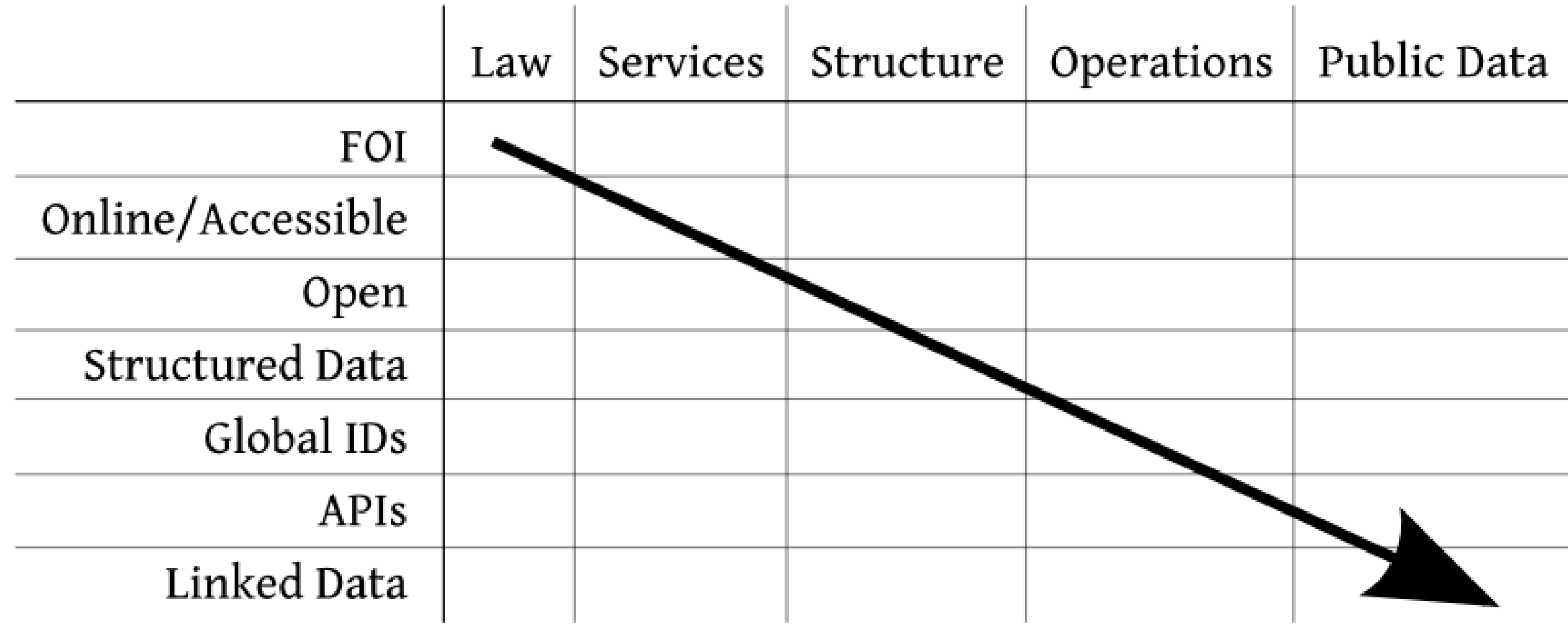
Model zrelosti otvorenih (državnih) podataka

- Razvijen s ciljem „prioritiziranja” odluka – što učiniti
 - Kako otvoriti podatke?
 - Koje državne podatke otvoriti?
- Autor: Joshua Tauberer
 - Detalji:
<https://opengovdata.io/2014/maturity-model-for-prioritization/>

	Law	Services	Structure	Operations	Public Data
FOI					
Online/Accessible					
Open					
Structured Data					
Global IDs					
APIs					
Linked Data					




The open government data maturity model



Kako čitati ovu shemu?

- Na osi X – kojim redom otvarati domene podataka
- Na osi Y – politička / tehnološka zrelost
 - Nama zanimljivija
- Redoslijed:
 - S lijeva na desno
 - Odozgor prema dolje
 - U tehnološkoj zrelosti prethodne faze su često i preduvjet za sljedeće
 - Nije nužno da je poredak faza od jeftinije prema skupljoj

	Law	Services	Structure	Operations	Public Data
FOI					
Online/Accessible					
Open					
Structured Data					
Global IDs					
APIs					
Linked Data					



Politička zrelost

- FOI – Freedom of Information

- Pravo na pristup informacijama
 - – postoji i kod nas
 - „Reaktivno” pravo – zainteresirana strana treba tražiti informacije od nadležnog tijela

- Online/Accessible

- „Aktivno” postavljanje informacija na Web - dostupnost

- Open

- // svi oni principi koji čine podatke otvorenima

	Law	Services	Structure	Operations	Public Data
FOI					
Online/Accessible					
Open					
Structured Data					
Global IDs					
APIs					
Linked Data					



Tehnološka zrelost

- **Structured data**

- Računalno čitljivi, pretraživi podaci koje je moguće analizirati i transformirati

- **Global Ids**

- Globalno jedinstveni identifikatori za (gotovo) svaki zapis, element, podatak ...
 - Omogućuju povezivanje izvora podataka, veću iskoristivost
 - Veća vrijednost: „vrijednost povezanih skupova veća od zbroja pojedinačnih vrijednosti” :-)
 - Dobar primjer: URI – Uniform Resource Identifier

- **APIs**

- Masovni zapis + puno više funkcionalnosti
 - Prvo provjeriti jesu li podaci uopće korisni – izrada i održavanje API-ja je skup posao

- **Linked Data**

- Posebni oblici povezivanja, uz globalne identifikatore postoje i odnosi - „predikati”

Primjer: European Data Portal

- Measuring open data maturity
- Dimenzije:
- **Open Data Policy**
 - Policy framework
 - Governance of open data
 - Open data implementation
- **Open Data Portal**
 - Portal feature
 - Portal usage
 - Data provision
 - Portal sustainability
- **Open Data Impact**
 - Strategic awareness
 - Political impact
 - Social impact
 - Environmental impact
 - Economic impact
- **Open Data Quality**
 - Currency and completeness
 - Monitoring and measures
 - DCAT-AP compliance
 - Deployment quality and linked data
- [Upute](#)
- [Dashboard](#), [Izvještaj za 2021.](#)

Još puno, puno modela i okvira...

- World Bank Open Data Readiness Assessment (ODRA)
 - <http://opendatatoolkit.worldbank.org/en/odra.html>
- The Open Data Barometer
 - <https://opendatabarometer.org/>
 - <https://opendatabarometer.org/leadersedition/methodology/>
- Global Open Data Index (GODI)
 - <https://index.okfn.org/>
 - <https://index.okfn.org/methodology/>

Nije dovoljno samo objaviti
BILO KAKVE
otvorene podatke!

Korišten *CreativeCommons* sadržaj
