Otvoreno računarstvo

4a. Otvoreni podaci

Creative Commons

Otvoreno računarstvo

4a. Otvoreni podaci

- **Svojstva**
- Utjecaj
- Izrada

Otvoreni podaci?

- **Open data** and content can be **freely used**, **modified**, and **shared** by **anyone** for **any purpose**
  - Open Definition, https://opendefinition.org/
  - Cjelovita definicija otvorenosti, pa tako i otvorenih podataka: https://opendefinition.org/od/2.1/en/

- Open data is **digital** data that is made available with the **technical** and **legal characteristics** necessary for it to be **freely used, reused, and redistributed** by anyone, anytime, anywhere
  - International Open Data Charter, https://opendatacharter.net

Otvorenost u otvorenim podacima

- Tri vrste otvorenosti – otvoreni podaci trebaju biti:
  - Pravno otvoreni
    - Otvorenom licencijom dozvoljeno je korištenje, izmjena i dijeljenje
  - Tehnički otvoreni
    - Bez tehničkih barijera. Dostupni u otvorenim formatima, strojno čitljivi, dostupni u masovnom/skupnom zapisu (*bulk*)
  - (Financijski otvoreni)
    - Besplatni, uz dozvoljenu „marginalnu” cijenu materijalnih troškova

8 principa otvorenih (*državnih*) podataka

1. **Potpuni** (*Complete*)
   - All public data is made available. Public data is data that is not subject to valid privacy, security or privilege limitations.
2. **Primarni** (*Primary*)
   - Primary data is data as collected at the source, with the finest possible level of granularity, not in aggregate or modified forms.
3. **Pravovremeni** (*Timely*)
   - Data are made available as quickly as necessary to preserve the value of the data.
4. **Pristupačni** (*Accessible*)
   - Data are available to the widest range of users for the widest range of purposes.

8 Principles of Open Government Data, https://opengovdata.org/

Dodatna objašnjenja: 14 Principles of Open Government Data, https://opengovdata.io/2014/principles/

8 principa otvorenih (*državnih*) podataka

5. **Strojno čitljivi** (*Machine-processible*)
   - Data is reasonably structured to allow automated processing.
6. **Nediskriminirajući** (*Non-discriminatory*)
   - Data are available to anyone, with no requirement of registration.
7. **Ne-vlasnički** (*Non-proprietary*)
   - Data are available in a format over which no entity has exclusive control.
8. **Objavljeni pod otvorenom licencijom** (*License-free*)
   - Dissemination of the data is not limited by intellectual property law such as copyright, patents, or trademarks, contractual terms, or other arbitrary restrictions.

8 Principles of Open Government Data, https://opengovdata.org/

Dodatna objašnjenja: 14 Principles of Open Government Data, https://opengovdata.io/2014/principles/

7 dodatnih principa

1. **Online i besplatni** (*Online & Free*)
   - Information is not meaningfully public if it is not available on the Internet at no charge, or at least no more than the marginal cost of reproduction. It should also be findable.
2. **Trajni** (*Permanent*)
   - Data should be made available at a stable Internet location indefinitely and in a stable data format for as long as possible.
3. **Provjereni** (*Trusted*)
   - Published content should be digitally signed or include attestation of publication/creation date, authenticity, and integrity.
4. **Presumpcija otvorenosti** (*A Presumption of Openness*)
   - Setting the default to open means that the government and parties acting on its behalf will make public information available proactively and that they'll put that information within reach of the public (online), with low to no barriers for its reuse and consumption

7 dodatnih principa

5. **Dokumentirani** (*Documented*)
   - Documentation about the format and meaning of data goes a long way to making the data useful. Government websites must provide users with sufficient information to make assessments about the accuracy and currency of legal information published on the website.
6. **Sigurni za otvaranje** (*Safe to Open*)
   - Government bodies publishing data online should always seek to publish using data formats that do not include executable content.
7. **Dizajnirani uz doprinos javnosti** (*Designed with Public Input*)
   - The public is in the best position to determine what information technologies will be best suited for the applications the public intends to create for itself. Public input is therefore crucial to disseminating information in such a way that it has value.

8 Principles of Open Government Data,
https://opengovdata.org/
Dodatna objašnjenja: 14 Principles of Open Government Data,
https://opengovdata.io/2014/principles/

I još 6 (sličnih) principa

- Međunarodna povelja o otvorenim podacima (*International Open Data Charter*)

1. **Zadano otvoreni** (*Open by Default)*
2. **Pravovremeni i sveobuhvatni** (*Timely and Comprehensive)*
3. **Pristupačni i iskoristivi** (*Accessible and Usable)*
4. **Usporedivi i interoperabilni** (*Comparable and Interoperable)*
5. **Za unaprjeđenje upravljanja i uključivanja građana** (*For Improved Governance & Citizen Engagement*)
6. **Za inkluzivni razvoj i inovaciju** (*For Inclusive Development and Innovation*)

https://opendatacharter.net/principles/

Open data = nešto poput FAIR data?

- FAIR data – znanstveni/istraživački podaci:
  - **Findable** – kako pronaći podatke? Vrlo su važni metapodaci
  - **Accessible** – kako pristupiti podacima? Otvoreni protokoli, možda uz autentifikaciju
  - **Interoperable** – i podaci i metapodaci koriste formalne standarde za integraciju s drugim podacima i aplikacijama
  - **Reusable** – ponovna uporaba podataka – potrebno jasno navesti izvor i uvjete uporabe

- Gdje su razlike?
  - Open data: „by any person”
  - FAIR data: „accessible by the defined persons,
    at the defined time
    and by the defined method”
  - FAIR data: “as open as possible, as closed as necessary”

FAIR principles, https://www.go-fair.org/fair-principles/

Vrste otvorenih podataka – po **strukturi**

- **Tabularni** (*tabular*)
- **Hijerarhijski** (*Hierarchical*)
- **Mrežni** (*Network*)
- **Podatkovne kocke** (*data cubes*)
- ...

- Nisu samo tekstualni:
  - Mape, genomi, kemijski spojevi, formule, fotografije, medicinski podaci, ...

Vrste otvorenih podataka – po **namjeni**

- Podaci sastavljeni od **zapisa** (*Record Data*)
  - Npr. popis vrtića u nekom gradu
- Podaci temeljeni na **grafovima** (*Graph-based Data*)
  - Npr. povezane web-stranice
- Podaci u **određenom poretku** (*Ordered Data*)
  - Vremenski podaci (*Temporal Data)*
    - Zapisi koji sadržavaju vremensku komponentu
    - Npr. posuđivanje knjiga iz knjižnice, kupovina u određeno vrijeme
  - Vremenski nizovi (*Time Series Data*)
    - Slijedni zapisi, često u već definiranim trenucima (npr. svaku minutu)
    - Npr. mjerenja više senzora
  - Sekvencijski podaci (*Sequence Data*)
    - Važan je poredak, bez vremenske komponente; Npr. podaci o genomima
  - Prostorni podaci (*Spatial Data*)
    - Npr. klimatski podaci

Otvoreno računarstvo

4a. Otvoreni podaci

- Svojstva
- **Utjecaj**
- Izrada

Utjecaj otvorenih podataka

- Primjeri polja djelovanja

https://data.europa.eu

- Primjeri prilika za napredak:
  - Donošenje političkih odluka temeljem dokaza u podacima
  - Suradnja među područjima
  - Praćenje javnog novca
  - Praćenje utjecaja javnih programa
  - Olakšavanje građanima donošenja odluka

https://opendatacharter.net/principles/

Način uporabe

- Data to **fact** – istraživanje pojedinih činjenica, u procesima, planiranju, ...
- Data to **information** – izrada interpretacija, vizualizacija, izvještaja...
- Data to **interface –** pružanje načina za interaktivni pristup i istraživanje
- Data to **data** – dijeljenje u izvornom ili prilagođenom, uređenom obliku, masovnog skupa podataka ili dijela podataka, ili izvedba API-ja
- Data to **service –** uporaba podataka „u pozadini” za ostvarivanje neke usluge

Open data, democracy and public sector reform -
A look at open government data use from data.gov.uk

Primjeri: otvoreni državni podaci

- data.europa.eu – The official portal for European data
  - https://data.europa.eu
- Eurostat – statistički podaci
  - https://ec.europa.eu/eurostat/
- Portal otvorenih podataka Republike Hrvatske
  - https://data.gov.hr/

Primjeri: otvoreni znanstveni podaci

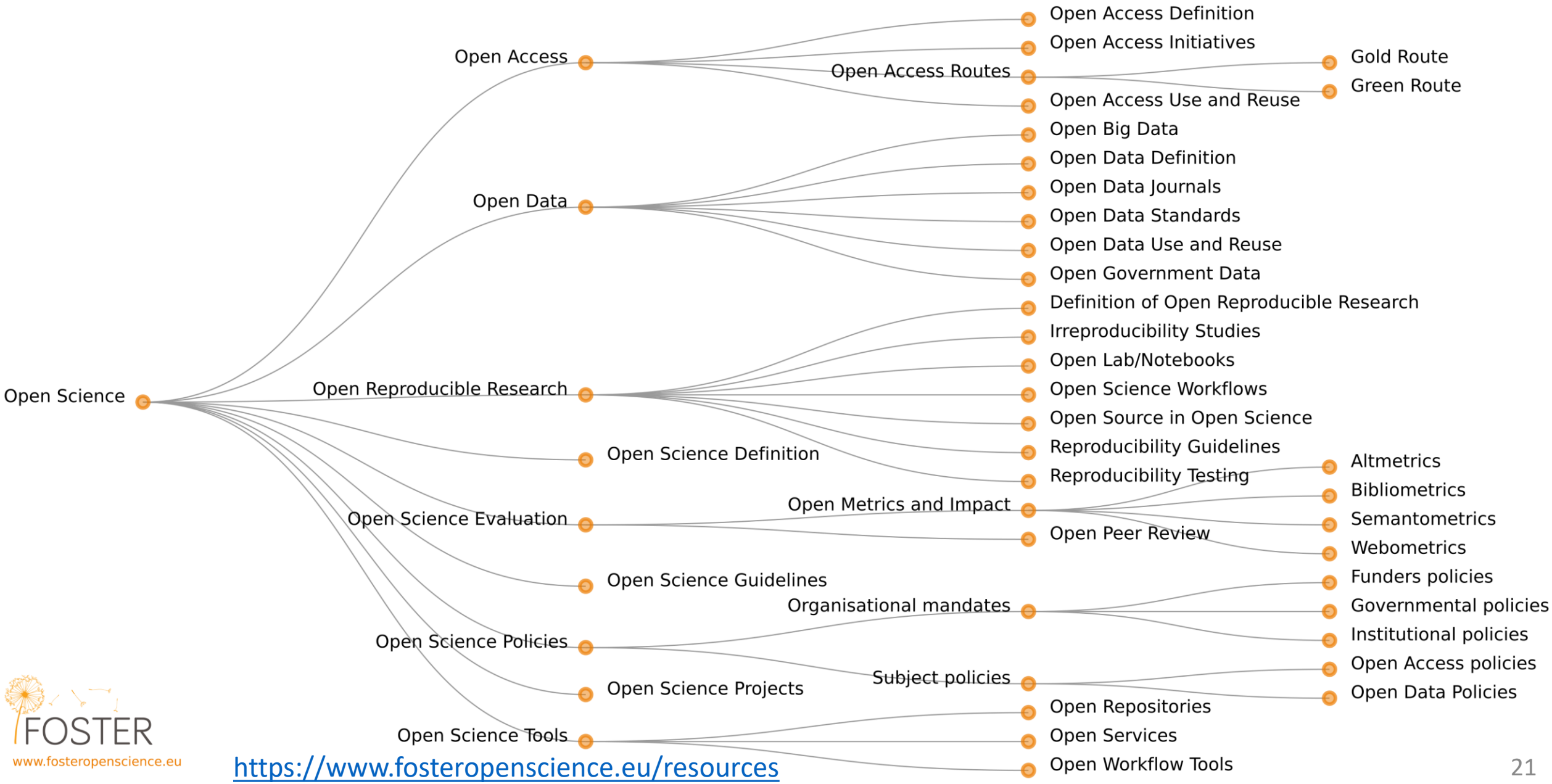
- PubChem – najveća kolekcija kemijskih informacija
  - https://pubchem.ncbi.nlm.nih.gov/

Data Collection	Live Count	Description
Compounds	111,443,036	Unique chemical structures extracted from contributed PubChem Substance records
Substances	286,824,801	Information about chemical entities provided by PubChem contributors
BioAssays	1,229,014	Biological experiments provided by PubChem contributors
Bioactivities	272,785,097	Biological activity data points reported in PubChem BioAssays
Genes	91,340	Gene targets tested in PubChem BioAssays and those involved in PubChem Pathways
Proteins	95,319	Protein targets tested in PubChem BioAssays and those involved in PubChem Pathways
Taxonomy	4,754	Organisms of targets tested in PubChem BioAssays and those involved in PubChem Pathways
Pathways	237,772	Interactions between chemicals, genes, and proteins
Literature	31,593,693	Scientific publications with links in PubChem
Patents	24,824,605	Patents with links in PubChem
Data Sources	759	Organizations contributing data to PubChem

Interdisciplinarnost

- Data Skills Framework –
  - ravnoteža među stručnjacima, kao podrška uspješnoj inovaciji u svijetu podataka
  - https://theodi.org/article/data-skills-framework

Open Science Taxonomy



FOSTER
www.fosteropenscience.eu

https://www.fosteropenscience.eu/resources

Argumenti „za” i „protiv” otvorenih podataka

- Podaci pripadaju ljudima
- Za izradu korišten javni novac
- Izrađen u državnoj instituciji
- Činjenice se ne mogu zaštititi
- Sponzori istraživanja ne dobivaju potpunu vrijednost bez otvaranja
- Podaci su potrebni za provođenje zajedničkih ljudskih aktivnosti (zdravstvo, obrazovanje, ekonomija...)
- U znanstvenim istraživanja, bolji pristup podacima ubrzava otkrića

- Javno financiranje ne smije ponavljati ili izazivati aktivnosti privatnog sektora
- Ako će izrada podataka biti korisna samo malom broju ljudi, to nije učinkovito trošenje javnog novca
- Otvoreni podaci mogu voditi brzom iskorištavanju rezultata koje provode bogate znanstvene institucije
- Naplata podataka može se iskoristiti za pokrivanje novih troškova objave
- Naplata podataka omogućuje rad neprofitnim organizacijama
- Prikupljanje, upravljanje, čišćenje podataka zahtjevaju trud i naplatu
- Nema kontrole nad korištenjem / agregacijom podataka

Wikipedia Open data – arguments for and against: https://en.wikipedia.org/wiki/Open_data#Arguments_for_and_against

Društveno zagovaranje (*advocacy*) za otvorene podatke

- *Pitch talk* – objašnjenje u 90-sekundi

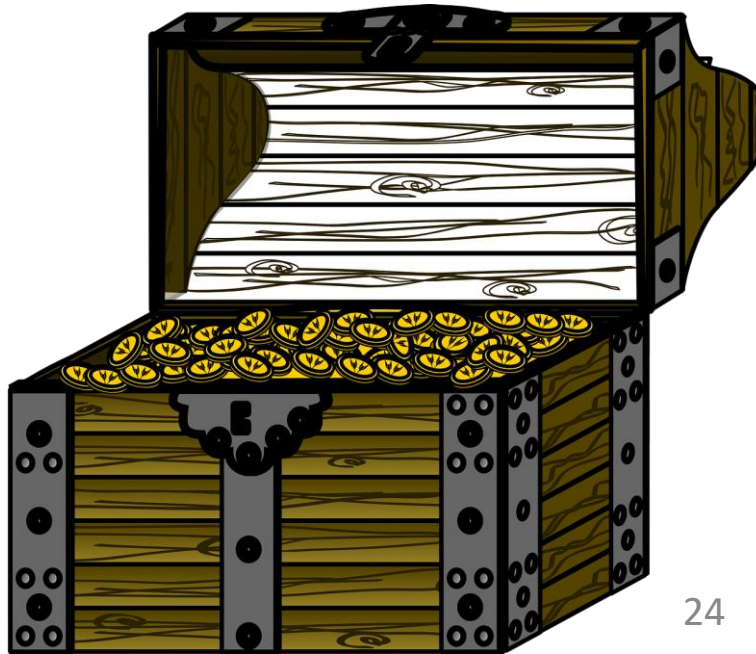
- **What it is?** Koji problem želimo riješiti korisnicima?
- **What it can do?** Prilagođeni odgovor, kako će ovo pomoći
- **Where it's helped?** Korisni i konkretni dokazi, gdje je dosad pomoglo
- **Why it's the best?** Kako se uspoređuje s drugim opcijama
- **What next?** Realna odluka, koju je moguće donijeti na licu mjesta

- RAZMISLITE SAMI! Zamislite problem kojeg otvoreni podaci mogu riješiti...

https://data.europa.eu/elearning/en/module14/#/id/co-01

Sudionici u svijetu otvorenih podataka

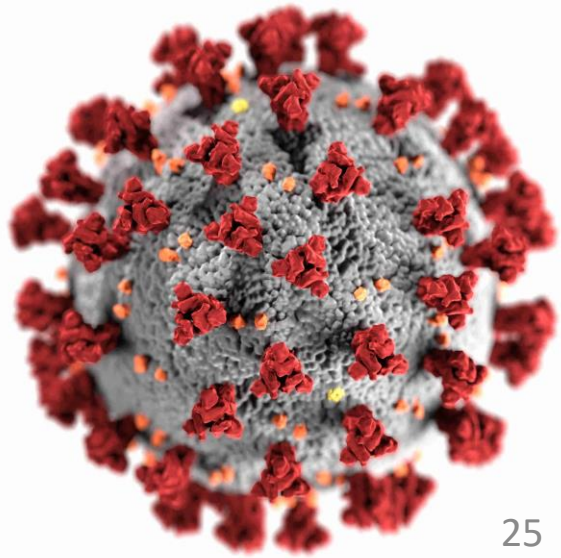
- Pojedinci
- Neprofitne organizacije, civilno društvo
- Zaklade
- Tijela javne vlasti
- Regionalna i lokalna samouprava
- Sveučilišta i istraživačke ustanove
- Tvrtke, *startupi*
- Novinari
- ...



Aktualni primjer: Covid-19

- Coronavirus: why an open future has never been more important
  - https://blog.okfn.org/2020/04/16/coronavirus-why-an-open-future-has-never-been-more-important/

- Covid-19 otvoreni skupovi podataka
  - Datahub:
    - https://datahub.io/core/covid-19
  - Our World in Data:
    - https://ourworldindata.org/coronavirus-source-data
  - Google Cloud Platform:
    - https://github.com/GoogleCloudPlatform/covid-19-open-data



Otvoreno računarstvo

4a. Otvoreni podaci

- Svojstva
- Utjecaj
- **Izrada**

Proces izrade skupa otvorenih podataka – I

1. **Izrada** / otvaranje skupa podataka

2. Odabir otvorene **licencije**

3. Omogućavanje **dostupnosti** i **pristupačnosti** skupa podataka

4. Omogućavanje **vidljivosti** skupa podataka

Proces izrade skupa otvorenih podataka - II

1. Dobavljanje podataka iz različitih izvora i tipova zapisa
2. Normalizacija, čišćenje i uređivanje podataka
   - Ovisno o vrsti podataka (numerički, tekstualni, itd.)
   - Ujednačavanje formata datuma, brisanje duplikata i suvišnih podataka, ujednačavanje numeričkih vrijednosti i opsega podataka, provjera pravopisa ...
3. Transformacija podataka
   - Promjena strukture ili sadržaja podataka, kombiniranje skupova podataka ...
4. Provjera valjanosti podataka
   - Po vlastitim pravilima
5. Pohrana skupa podataka
   - U datoteku određenog oblika, u bazu podataka, itd.

Primjer alata za pomoć pri uređivanju podataka

- OpenRefine
  - https://openrefine.org/
  - Otvoreni alat za uređivanje „neurednih” skupova podataka

The formats currently supported (in version 2.7) include:

- TSV, CSV, or values separated by a custom separator
- Line-based text files
- Fixed-width field text files
- PC-Axis text files
- MARC files
- Excel (.xls, xlsx)
- Open Document Format spreadsheets (.ods)
- XML, RDF as XML
- JSON
- Google Spreadsheets
- RDF N3 triples

- Importing
- Filtering / faceting
- Editing:
  - Editing cells, editing cells by Clustering
  - Editing columns, creating columns by Extending data
  - Editing rows
  - Understanding expressions
  - Understanding regular expressions
- Exporting

- Reconciliation - And fetching additional data with Wikidata
- Data sources
- Using Web Services to extend data

Kako odabrati otvorenu licenciju skupa podataka?

- **Licencija =** dozvola za korištenje (pod određenim uvjetima)

- **Krivo** mišljenje: **ne-**odabir licencije podataka -> javno dobro *(public domain*)
- Bez jasne definicije prava korištenja, organizacije mogu čak i zatvoriti podatke, koji su zamišljeni kao otvoreni (no bez jasne licencije)

- **Otvorene** licencije dozvoljavaju pristup, ponovnu uporabu i dijeljenje
  - S malo ili nimalo ograničenja (ukratko)

Kako odabrati otvorenu licenciju skupa podataka?

- Primjeri otvorenih licencija
  - CC0 – Creative Commons Public domain
  - CC-BY – Creative Commons, BY attribution
  - CC-BY-SA – Creative Commons, BY attribution, Share-alike

  - PDDL-1.0 – Open Data Commons Public Domain Dedication and License
  - ODC-By-1.0 – Open Data Commons Attribution License
  - ODbL-1.0 – Open Data Commons Open Database License
- Državne otvorene licencije
  - Otvorena dozvola - https://data.gov.hr/otvorena-dozvola
    - Slična CC BY
  - Open Government Licence
    - https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/
  - Dobro je znati: **Tijela javne vlasti** u RH, temeljem *Zakona o pravu na pristup informacijama* preuzimanjem *Direktive o ponovnoj uporabi informacija EU*, trebaju dati svoje informacije za ponovnu uporabu, bez nepotrebnih ograničenja
- Licensing assistant - https://data.europa.eu/en/training/licensing-assistant

Kako učiniti podatke dostupnima?

- Odabirom odgovarajućeg **formata** za željeni oblik podataka
  - CSV (TSV)
  - XML, JSON, HDF5
  - RDF, JSON-LD, drugi formati za povezivanje podataka
  - NetCDF
  - ...

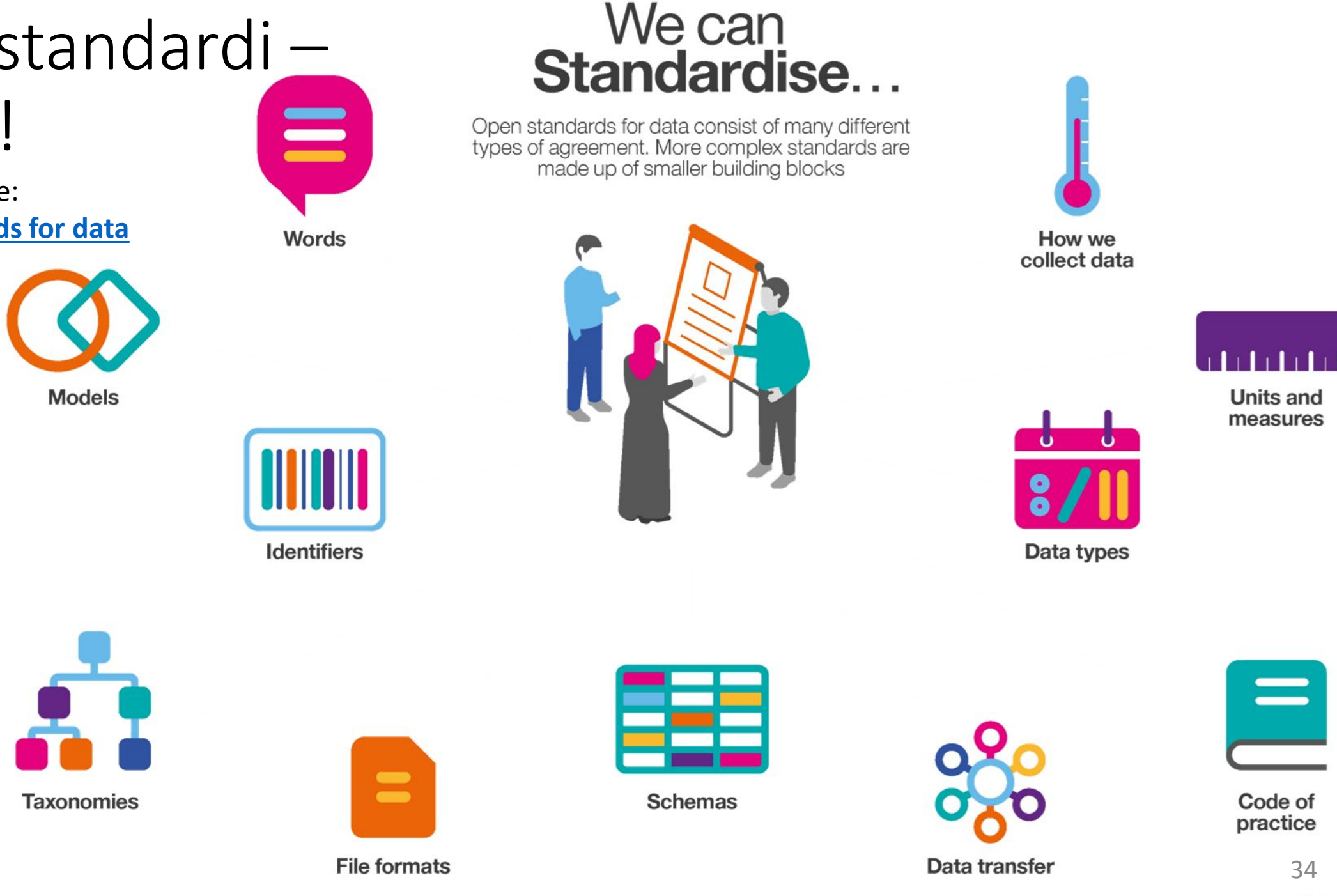
- Za veću dostupnost, dobro je korisnicima ponuditi **više formata zapisa**
- Odabrati način isporuke – protokol
  - **HTTP**
  - FTP, p2p

- **OTVORENOST!**

Otvoreni standardi – zar opet?!

The Open Data Institute:

[Types of open standards for data]



Bulk ili API?

- Masovni/skupni zapis (*bulk*)
  - Preuzimanje svih dostupnih podataka za transformaciju na strani korisnika
  - Moguće preuzimanje u više formata podataka
  - Moguć pregled podataka, ovisno o alatu
  - Prvo ostvariti ovaj način dohvata podataka, kasnije API
- Application Programming Interface - API
  - I API može „isporučiti” masovni zapis, ali nije namijenjen tome
  - Sučelje najčešće „samo za čitanje” (*read only)*
    - U svijetu otvorenih podataka postoje i API-ji pomoću kojih možemo mijenjati podatke, no oni nam ovdje nisu u interesu
  - Potreban brz odaziv, velika dostupnost, dugoročno održavanje!
  - Najčešće **RESTful API** (barem bi trebao biti)
  - Odabir podskupa podataka, uz filtriranja
  - Dostupno više formata podataka
  - Smislene poruke o pogreškama, dokumentacija
  - Verzioniranje API-ja

Kako učiniti podatke vidljivima?

- Portali otvorenih podataka – *Open Data Portals*
  - Web-aplikacije za pregled i dohvat otvorenih skupova podataka, namijenjenih ponovnoj uporabi
  - Velike mogućnosti pretrage
  - Kvalitetni metapodaci za opis skupova podataka
  - Različita sučelja, uključujući i programska (API-ji), za dohvaćanje podataka
  - Jednostavan pregled i vizualizacija podataka
  - Velike mogućnosti filtriranja
  - Jasno vidljive licencije podataka
  - Mogućnosti sudjelovanja korisnika – ocjenjivanje, komentari, dijeljenje
  - Pregled povezanih podataka (*Linked data)*

- Portali mogu sadržavati i podatke...
- ... ili biti samo *katalozi*
  - Izvor podataka je negdje drugdje, npr. pri web-sjedištu autora

Kako učiniti podatke vidljivima?

- Međunarodni portali
  - EU Open Data Portal - https://data.europa.eu
  - World Bank Open Data - https://data.worldbank.org/
  - Kaggle - https://www.kaggle.com/
- Državni portali
  - Portal otvorenih podataka - http://data.gov.hr/
  - Find open data – UK - https://data.gov.uk/
  - US Government open data - https://www.data.gov/
- Gradski / regionalni portali
  - Portal otvorenih podataka – ZG - http://data.zagreb.hr/
  - Portal otvorenih podataka – RI - http://data.rijeka.hr/

- Pretraga portala: https://dataportals.org/

Primjer platforme za otvorene podatke: CKAN

- Upravljanje otvorenim skupovima podataka i njihova objava
- Alat otvorenog kôda
- Funkcionalnosti
  - https://ckan.org/features/
- Tehnologije:
  - Python
  - PostgreSQL
  - Apache SOLR – platforma za pretraživanje

- Tko koristi CKAN?
  - Vlade, lokalna uprava, organizacije, sveučilišta, ...
  - https://ckan.org/showcase

Korišten *CreativeCommons* sadržaj

- Open science tree by FOSTER project, CC BY 4.0 International
- We can standardize by The Open Data Institute, CC BY-SA 4.0 International