

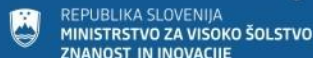
Pridobivanje in obdelava inštrumentalnih raziskovalnih podatkov v luči načel FAIR

dr. Uroš KUNAVER, Centralna tehniška knjižnica Univerze v Ljubljani

Petra DURINI, Centralna tehniška knjižnica Univerze v Ljubljani

mag. Miro PUŠNIK, Centralna tehniška knjižnica Univerze v Ljubljani

Ljubljana in online, Centralna tehniška knjižnica Univerze v Ljubljani,
Usposabljanje podatkovnih strokovnjakov, 14. – 17. 10. 2024





Vsebina predavanja

- Značilnosti pridobivanja (in obdelave) raziskovalnih podatkov iz inštrumentov (in simulacij).
- Formati podatkov.
- Metapodatkovni standardi in sheme.
- Geslovniki, šifranti, taksonomije, ontologije...
- Izvor podatkov ali provenienca.



Značilnosti pridobivanja (in obdelave) raziskovalnih podatkov iz inštrumentov (in simulacij)

- Zelo veliko število merskih metod, znotraj vsake različni inštrumenti različnih proizvajalcev.
- Zelo veliko formatov podatkov, zelo veliko različnih metapodatkov o parametrih meritev, ki niso standardizirani/dogovorjeni in odprti.



Značilnosti pridobivanja podatkov... 2

- Na marsikaterem področju prevladujejo lastniški (ang. proprietary) formati datotek, ki jih je možno odpreti le na inštrumentu istega proizvajalca (včasih le na isti generaciji inštrumentov), izvoz v kakšno od standardnih oblik ni možen ali pa ni popoln (npr. slike je možno izvoziti, metapodatkov pa ne).

Ti podatki so v binarni obliki brez dokumentacije o strukturi (ki bi omogočala izgradnjo lastne programske opreme za pretvorbo v druge formate – pri tem je tudi vprašanje, ali licenca za uporabo inštrumenta/programske opreme to sploh dovoljuje).

Raziskovalci morajo včasih celo prepisovati (meta)podatke na roke.



Značilnosti pridobivanja podatkov... 3

- Ogromne količine podatkov.
Že čisto “običajni” inštrumenti (npr. različni spektrometri) lahko “ustvarijo” ogromne količine podatkov. Včasih smo dobili na papir spektrogram (graf), ki ga je bilo potrebno ročno odčitati, sedaj pa dobimo podrobno analizo oblike vrhov ali kakršno koli dodatno analizo za vsako točko primarnega spektra/analize (npr. povezava tekočinske kromatografije in masne spektrometrije).
Mnogokrat skupaj z rezultatom dobimo tudi kompletne kalibracijske podatke (kar je dobro).
Še posebno v zadnjem času, ko se cena pomnilniškega prostora znižuje in se povečuje hitrost diskov/omrežij je pri novejših inštrumentih opazen trend “neznosne lahkote shranjevanja podatkov”.



Značilnosti pridobivanja podatkov... 4

- Podobno kot za inštrumente velja za različne vrste simulacij, kjer se producira ogromne količine podatkov. V nekaterih primeri je tok podatkov lahko kontinuiran in težko se je odločiti v katerem koraku shraniti določeno stanje.



Razlike v primerjavi z drugimi področji

Raziskovalci/podatkovni strokovnjaki drugih področij imajo drugačne probleme in v nekaterih primerih drugačen pogled na hranjenje, obdelavo in objavljane raziskovalnih podatkov.

- Humanistika
 - Del humanistike (zgodovina, filozofija) se niti ne ukvarja s podatki.
 - Del humanistike, ki se ukvarja z analizo besedil in jezika pa prav tako operira z velikimi količinami podatkov in programsko opremo za njihovo obdelavo in je v tem pogledu njihova problematika zelo blizu naravoslovju in tehniki (še posebej področju simulacij).



Razlike v primerjavi z drugimi področji 2

- Družboslovje
 - Kot rečeno – z inštrumenti se “ustvarja” ogromna količina podatkov. Nek inštrument lahko pri eni meritvi “ustvari” več podatkov, kot je na primer celoletni prirast Arhiva družboslovnih podatkov (kar ne pomeni, da so ti podatkov manj vredni). Zato se morajo raziskovalci v marsikaterem primeru odločiti, kaj hraniti/objaviti in česa ne. O tem na okrogli mizi.
 - Pri večini raziskav ne delamo s podatki o ljudeh (ZVOP, GDPR) – moramo pa se zavedati, da smo še vedno zavezani etičnim pravilom (še posebej pri delu z živalmi) in pa različnim postopkovnih pravilom (pri certificiranih analizah, pri razvoju/analizi zdravil in hrane...).



Razlike v primerjavi z drugimi področij 3

- Medicina

- Velik del raziskovalnih podatkov na področju medicine prihaja iz inštrumentov in zanje veljajo podobne značilnosti kot za ostala področja naravoslovja (količina podatkov pri npr. MR), specifika medicine pa je, da so velik del podatkov izvira od ljudi, zato je poseben poudarek na etiki in ravnanju z osebnimi podatki.
- Prav tako je mnogo bolj kot pri ostalih panogah naravoslovja in tehnike pomembno dobro obvladovanje procesov in upoštevanje postopkovnih pravil.
- Navedeno pa velja tudi za interdisciplinarne raziskave na področju naravoslovja in tehnike, ki se navezujejo na medicino in operirajo s podatki o ljudeh. To še posebej velja za farmacijo.



Formati podatkov

- Vsi raziskovalni podatki bodo vloženi v repozitorij v **digitalni obliki** in sicer v obliki datotek.
- **Datoteke** so najmanjši skupki (računalniških) podatkov, s katerimi upravlja operacijski sistem računalnika in tudi različna spletna mesta in platforme.
- Datoteke v osnovi vsebujejo binarne podatke, osnovni gradnik je bajt (ang. byte), ki lahko vsebuje 256 različnih vrednosti (kar je malo). Vsebina enega bajta je lahko zelo različna – lahko pomeni eno črko/znak, lahko pomeni majhno številko, lahko pa je del večje strukture. Večje strukture (skupki več bajtov) spet lahko pomenijo znake, večje/preciznejše številke ali zapise v neki bazi podatkov.
- Kaj je v datoteki in kakšno strukturo imajo podatki je definirano s **formatom datoteke**.



Značilnosti različnih formatov

- Nekateri formati so zelo enostavni (npr. TXT), nekateri bolj zapleteni (npr. slike v BMP formatu) ali pa tudi zelo zapleteni (npr. slike v JPG formatu ali relacijske baze podatkov).
- Nekateri formati so povsem jasno in enolično določeni (npr. SHP, GPX), nekateri formati obstajajo v različnih verzijah (npr. PDF), nekateri pa niso dovolj natančno določeni (npr. TXT in nekateri video formati), spet drugi dopuščajo možnost nadgradnje (večina formatov na osnovi XML, npr. GPX).
- Nekateri formati vsebujejo/podpirajo metapodatke (npr. JPG, TIFF, SHP, GPX), drugi formati pa ne (npr. TXT, BMP) ali pa zelo omejeno (XLS, XLSX, PPT, PPTX, DOC, DOCX).



Končnice oziroma ekstenzije datotek

- Vrsto formata običajno prepoznamo po končnici (ekstenziji) oziroma **tipu datoteke** (vse kar je za zadnjo piko v imenu datoteke), vendar pa vse končnice ne pomenijo nujno samo enega jasno določenega formata.
- V večini primerov se formati razvijajo in znotraj ene končnice imamo lahko opraviti z različnimi **verzijami formatov**, ki so običajno navzgor združljivi (podatke iz starejšega formata se da brez izgub shraniti v novejšem formatu; novejša programska oprema zna brati starejše formate), obratno pa običajno ne gre.
- Včasih se pod isto končnico oz. tipom lahko skrivajo tudi razlike v licenciranju (npr. različne verzije PDF formatov so v lasti podjetja Adobe, medtem ko je format PDF-A, ki ima prav tako končnico PDF, standardiziran in licenciran za javno rabo).



Končnice oziroma ekstenzije datotek 2

- Po drugi strani pa lahko več različnih (vendar običajno podobnih) končnic pomeni **isti format**. Taki primeri so JPG in JPEG, TIF in TIFF, HTM in HTML... Različno poimenovanje izhaja iz dejstva, da so bile včasih na Microsoftovih operacijskih sistemih končnice omejene na 3 znake, na Unix/Linux sistemih pa ne. Formati, ki so bili najprej razviti ali razširjeni na Unix/Linux sistemih so imeli originalno daljše končnice, ki so jih kasneje za rabo v Microsoftovih (in nekaterih drugih) sistemih skrajšali.



Značilnosti, pomembne za načela FAIR

- Nekateri formati omogočajo ob shranjevanju podatkov v datoteko različne vrste kompresije (brezizgubne in izgubne). V določenih primerih lahko shranimo tudi podatke, stisnjene z izgubnimi načini kompresije (npr. JPG, če je to izvorni format slike iz inštrumenta), vendar je nadaljnja obdelava takih podatkov otežkočena.
- Pomembne podatke shranjujemo v **surovi (nestisnjeni) obliki** ali z eno od **brezizgubnih kompresijskih metod** (npr. določene oblike TIFF formata).



Lastništvo, licence in odprtost

Zaprto ali odprto formatov neposredno vpliva na **FAIR načelo A (Accessibility)**, saj je lahko dostopnost podatkov bolj ali manj omejena glede na lastništvo formata. Po drugi strani pa mnogi lastniški formati in pripadajoča programska oprema omogočajo bolj bogate in popolne zapise kot odprti formati. Na primer marsičesa, kar je zapisano v XLSX formatu (formule, grafi...) ni možno zapisati v CSV formatu. Zato je velikokrat priporočljivo, da podatke objavimo tako v izvornem lastniškem formatu, kot tudi v enem od čimbolj odprtih formatov.



Lastniški formati

Lastniški (ang. proprietary) formati – so formati, katerih specifikacija je v lasti nekega podjetja, ki povsem razpolaga z avtorsko pravico nad tem formatom in pogoji uporabe. Ti formati so lahko:

- **povsem zaprti formati** (običajno binarni) – prebrati jih je možno samo s programsko opremo proizvajalca (inštrumenta).
Ti formati so seveda **najmanj FAIR**, vendar običajno vsebujejo vse podatke, ki jih inštrument zabeleži.
- pravno gledano **zaprti formati**, pri katerih je **znana specifikacija** in kupec inštrumenta lahko podatke s svojo programsko opremo izvozi (v teh primerih je zelo dobro preveriti drobni tisk pogodbe za uporabo programske opreme, če je to res dovoljeno in pod kakšnimi pogoji).
Poleg podatkov v lastniškem formatu lahko v tem primeru objavimo tudi podatke v drugem, po možnosti bolj **odprtem formatu**.



Lastniški formati 2

- pravno gledano **zaprti**, vendar je uporaba za razvoj neodvisnih (ang. third party) aplikacij, ki uporabljajo ta format, možna proti **plačilu licenčnine**. Posledično je na voljo mnogo programske opreme (tudi brezplačne), ki omogoča branje (in urejanje) podatkov v teh formatih. Mnogokrat sam lastnik formata nudi brezplačne programe za branje (in konverzijo) dokumentov v teh formatih. Taki so na primer formati Microsoftovih pisarniških orodij in nekateri Adobovi formati (PDF). Podobno kot v prejšnjih primerih objavimo podatke v lastniškem in še enem od **odprtih formatov**. Če bistvene razlike v funkcionalnosti med formati ni, objavimo samo v odprtem formatu.



Lastniški formati 3

- Podobno kot pri prejšnjem primeru, le da je **uporaba formata** (in njegove specifikacije) **prosta (brezplačna)**. Specifikacija je dobro znana (objavljena) in vsi jo uporabljajo, vendar je avtorska pravica vseeno zadržana in ni odprte trajne licence za uporabo formata. Težava je v tem, da se lahko lastnik kadarkoli premisli in začne uporabo formata zaračunavati. Tak je SHP format za opisovanje vektorskih prostorskih podatkov. Glede objave velja enako kot pri prejšnjem primeru - če gre za enega od splošno znanih in uporabljenih formatov (npr. SHP), lahko podatke objavimo kar v tem.



Odprti formati

Odprti podatkovni formati so tisti, kjer je uporaba trajno dovoljena z ustrezno odprto licenco (npr. CC ali eno od odprtih licenc za programsko opremo) ali pa so historično v javni domeni (npr. TXT).

Običajno jih **vzdržuje** kakšna organizacija, v nekaterih primerih tudi posamezna podjetja. Pogosto so tudi **standardizirani**. Taki so npr. OpenDocument (<https://en.wikipedia.org/wiki/OpenDocument>) formati (ODT, ODS, ODP...).

Te formate večinoma podpira tudi komercialna programska oprema (npr. MS Word podpira tudi format ODT).

Odprti formati so za objavo **v skladu z načeli FAIR** najprimernejši.



Standardi na področju formatov

Zaželeno je, da podatke objavimo v enem od »**standardnih**« formatov. Ta beseda lahko pomeni dvoje:

- Lahko gre za format, ki je dejansko **standardiziran**, to pomeni, da je v skladu z enim od evropskih ali mednarodnih standardov. Taki formati so na primer:
 - **PDF-A**, ki ga definirata standarda ISO 19005 in ISO 32000 (<https://en.wikipedia.org/wiki/PDF/A>) in pa
 - **OpenDocument** formati (<https://en.wikipedia.org/wiki/OpenDocument>), ki jih definira standard ISO 26300.



Standardi na področju formatov 2

- Lahko pa gre za format, ki je tako uveljavljen in že tako dolgo v uporabi, da velja kot »**de facto**« **standard**, čeprav specifikacija ni zares standardizirana. Tak je na primer SHP format. Podobno velja tudi za manj poznane formate, ki pa so dogovorjeni oz. uveljavljeni znotraj neke znanstvene panoge in jih dosledno uporabljajo vsi za shranjevanje določenih vrst podatkov. Tak je FASTQ format za shranjevanje nukleotidnih sekvenc v biologiji (https://en.wikipedia.org/wiki/FASTQ_format).



Kako izbrati format

Mnogokrat formata niti ne moremo izbrati, ampak je odvisen od različnih dejavnikov.

- Format izvornih podatkov je običajno **odvisen od inštrumenta** s katerim merimo.
 - Večina inštrumentov meritev najprej zapiše v svojem (lastniškem) formatu.
 - V primeru, da inštrument ponuja izvoz podatkov tudi v drugih, bolj odprtih formatih, shranimo podatke v lastniškem in v odprtem formatu.
 - Če neposreden izvoz v odprtem formatu ni možen, poskusimo poiskati programsko opremo, s katero lahko pretvorimo podatke iz lastniškega v odprti format. Pri tem nam lahko pomaga tudi sam proizvajalec inštrumenta.

Podatke objavimo seveda v **čimbolj odprtem formatu**, pri čemer pazimo, da nismo pri konverziji izgubili katerih (meta)podatkov.



Kako izbrati format 2

- Po drugi strani pa format objavljenih podatkov mnogokrat **določa** sam **repozitorij**. V tem primeru gre običajno za odprt ali vsaj široko sprejet format, ki ga uporabljajo vsi zainteresirani.
 - Če hočemo objaviti v izbranem repozitoriju, moramo podatke pretvoriti v izbrani format.
 - Lahko se nam zgodi, da bomo morali kakšen metapodatek, ki je vgrajen v sam podatkovni format, ročno prepisati iz lastniškega formata (oz. iz inštrumenta) ali pa ga celo posebej izmeriti (kar je dobro načrtovati vnaprej).



Kako izbrati format 3

- Pri objavi v splošnem repozitoriju običajno nismo omejeni na določen(e) formate. V tem primeru objavimo v čimbolj odprtem formatu, zraven pa priložimo tudi podatke v originalnem format (če so ti bolj popolni) ali pa manjkajoče (meta)podatke opišemo posebej v opisu provenience.

Dobro znan pregled splošno uporabnih podatkovnih formatov in njihove primernosti za objavljanje podatkov v skladu z načeli FAIR najdemo na straneh UK Data Service: <https://ukdataservice.ac.uk/learning-hub/research-data-management/format-your-data/recommended-formats/>.



UK Data Service Recommended formats

Recommended formats

File formats recommended by the UK Data Service

The table contains guidance on file formats recommended and accepted by the UK Data Service for data sharing, reuse and preservation.

You may need to convert your data files to a preservation file format.

We welcome queries from researchers about appropriate file formats for working and preservation, particularly early in the research process. If you are unsure of the suitability of your file formats for the data you want to deposit with the UK Data Service, please [get in touch](#).

Type of data	Recommended formats	Other acceptable formats
Quantitative tabular data with extensive metadata. A dataset with variable labels, code labels, and defined missing values, in addition to the matrix of data.	Proprietary formats of statistical packages e.g. SPSS (.sav), Stata (.dta), .sas7bdat. Delimited text and command ('setup') file (SPSS, Stata, SAS, etc.) containing metadata information. Some structured text or mark-up file containing metadata information, e.g. DDI XML file.	SPSS portable format (.por). MS Access (.mdb/.accdb).
Quantitative tabular data with minimal metadata. A matrix of data with or without column headings or variable names, but no other metadata or labeling.	Comma-separated values (CSV) file (.csv). Tab-delimited file (.tab). Including delimited text of given character set with SQL data definition statements where appropriate.	Delimited text of given character set – only characters not present in the data may be used as delimiters (.txt). Widely-used formats: MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb), OpenDocument Spreadsheet (.ods).

Vir: <https://ukdataservice.ac.uk/learning-hub/research-data-management/format-your-data/recommended-formats/>



Drugi (specifični) formati

Biologija:

Name	Link	Description	Content
VCF			genomics
BAM			genomics
GFF			gene model
GTF			gene model
Snakemake, CWL			software
SBGN-ML			systems biology model
SBML	http://sbml.org/	Systems Biology Mark-up Language	systems biology model
SED-ML	https://sed-ml.org/	Simulation Experiment Description Markup Language	systems biology simulation experiment
OMEX		Open Modeling EXchange format	systems biology
COMBINE Archive			systems biology
OrthoXML	https://www.orthoxml.org/xml/Main.html		orthology
SeqXML	https://www.orthoxml.org/xml/Main.html		biological sequences
FASTA	https://en.wikipedia.org/wiki/FASTA_format	Text-based format for representing nucleotide sequences	biological sequences
CellML	https://www.cellml.org/	XML markup language to store and exchange computer-based mathematical models	systems biology
NeuroML	https://neuroml.org/	XML-based description language that provides a common data format for defining and exchanging descriptions of neuronal cell and network models	



Nekaj primerov formatov: TXT

TXT format je en najbolj enostavnih in jasnih formatov – ali res?

- Končnica TXT običajno pomeni enostaven besedilni format, vendar se dejanski zapis besedila med različnimi TXT formati razlikuje. Tradicionalno TXT pomeni besedilo v ASCII formatu.
 - ASCII (American Standard Code for Information Interchange) format je zapis, kjer vsak bajt predstavlja 1 znak (pri čemer je izkoriščenih le 128 vrednosti oziroma le 7 bitov). ASCII format vsebuje velike in male črke ameriške abecede, cifre in posebne znake (ločila, znake za dolar, procent...) ter posebne kontrolne znake za upravljanje tiskalnikov. Slovenskih diakritičnih znakov ni v naboru ASCII znakov.



Razširitve ASCII nabora znakov

- Potreba po zapisu znakov **drugih abeced** (diakritični znaki, cirilica ... vse do kitajskih pismenk) in drugih **posebnih znakov** (matematika, finance...) je pripeljala do razširitev definicije tekstovnega formata. Nobeno besedilo v slovenščini torej ni zapisano z ASCII naborom znakov, ampak uporablja eno od razširitev tega zapisa.
- Najprej je prišlo do razširitve na 256 znakov, nato do uporabe različnih “kodnih tabel” (ang. code page), ki znotraj teh 256 možnih vrednosti predstavljajo znake različnih abeced. Primer je kodna tabela Windows-1252 oziroma CP-1252 (<https://en.wikipedia.org/wiki/Windows-1252>), ki jo lahko še vedno uporabljajo starejše naprave.



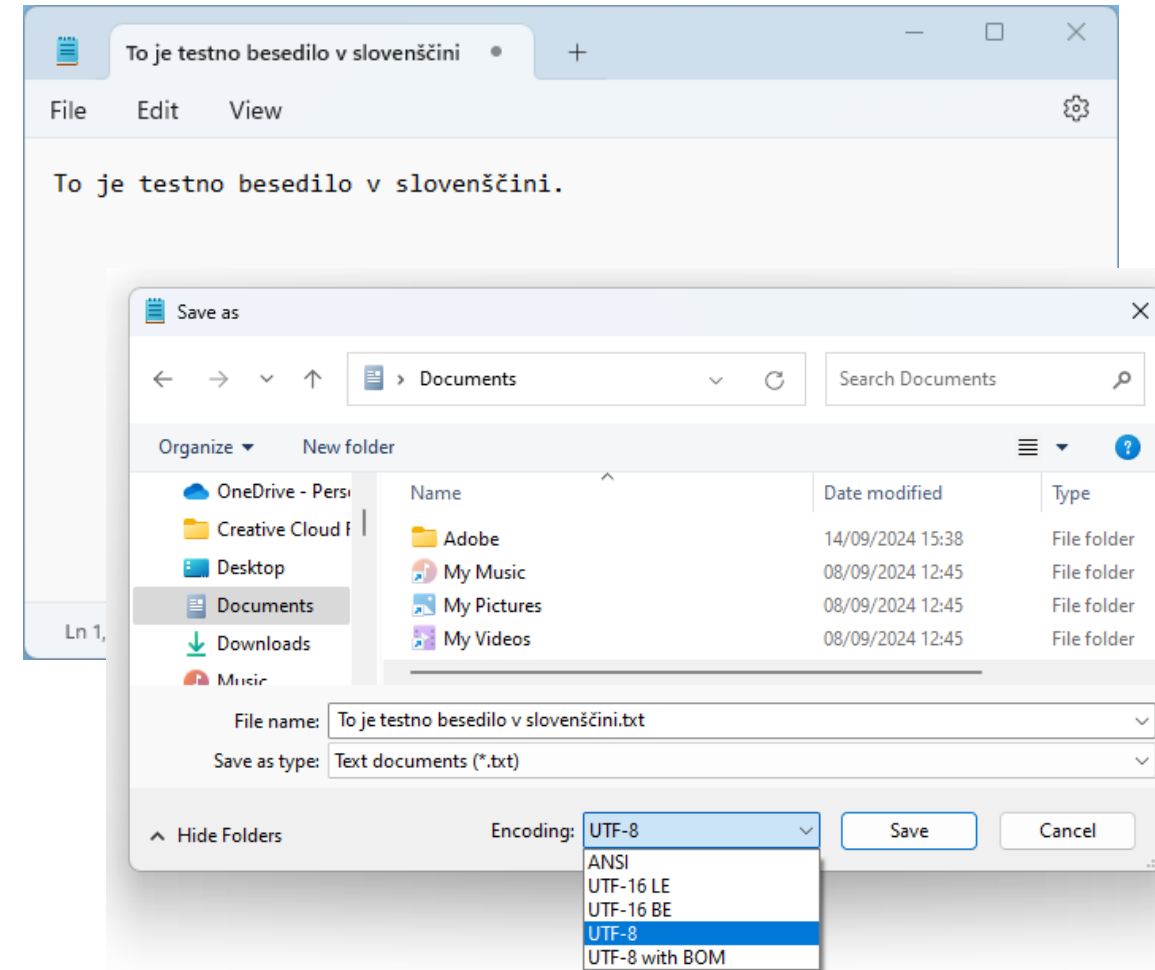
Unicode

Kasneje je prišlo do razvoja novih standardov, ki uporabljajo do štiri bajte za kodiranje. Daleč najbolj pogosto uporabljen standard je Unicode s podvariantama UTF-8 in UTF-16 (<https://en.wikipedia.org/wiki/Unicode>). Unicode uporablja večina sodobnih operacijskih in datotečnih sistemov (tudi za poimenovanje datotek).

- UTF-8 je najbolj razširjen univerzalen besedilni format/kodiranje. Z njim je možno zapisati 1.112.064 različnih znakov, kar pokrije vse znane pisave sveta. Znaki so zapisani z od enim do štirimi bajti in je združljiv z ASCII zapisom. Skoraj vse spletne strani uporabljajo UTF-8 in tudi nekateri specifični formati (XML) in metapodatkovne sheme priporočajo UTF-8 kodiranje. Več o UTF-8: <https://en.wikipedia.org/wiki/UTF-8>.
- UTF-16 je podoben UTF-8, le da je manj uporabljan. Za kodiranje uporablja eno ali dve 16 bitni enoti, pri čemer je pomemben še vrstni red obeh bajtov v enoti – dva podformata. Več o UTF-16: <https://en.wikipedia.org/wiki/UTF-16>.

Pretvarjanje med različnimi načini kodiranja

- Najenostavnejši način, kako pogledamo kateri način kodiranja uporablja nek besedilni dokument je, da ga odpremo v za to namenjenem programu (v MS Windows je to Notepad/Beležnica). Način kodiranja vidimo v spodnjem desnem kotu.
- Če hočemo spremeniti kodiranje in dokument shraniti z drugim kodiranjem izberemo "Save as"/"Shrani kot" in v spustnem seznamu na dnu okna izberemo drug način.
- Za avtomatsko pretvorbo več dokumentov uporabimo namensko programsko opremo (o tem več v tretjem sklopu predavanj).





XML format

XML (Extensible Markup Language) je format, ki ga definira nabor pravil in je namenjen shranjevanju in izmenjavi podatkov v obliki ki je berljiva ljudem in strojem. (definicija: <https://www.w3.org/TR/xml/>).

- XML se je razvil iz HTML zaradi potrebe po ločitvi med vsebino (XML) in obliko (CSS).
- Za specifikacijo XML formata skrbi World Wide Web Consortium (W3C) <https://www.w3.org/>.
- XML je v bistvu besedilna datoteka s posebno strukturo (podobno kot HTML). Besedilo je kodirano v skladu z Unicode standardom, pri čemer je priporočeno UTF-8 kodiranje.

```
<?xml version="1.0" encoding="utf-8" standalone="yes"?>
<gpx version="1.0" creator="CanWay -
http://www.canwaygps.com"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="http://www.topografix.com/GPX/1/0"
xsi:schemaLocation="http://www.topografix.com/GPX/1/0
http://www.topografix.com/GPX/1/0/gpx.xsd">
<time>2021-02-27T12:01:13Z</time>
<bounds minlat="46.020289" minlon="14.636936"
maxlat="46.040156" maxlon="14.686744" />
<trk>
<name>2001-04-29 13-17</name>
<trkseg>
<trkpt lat="46.038765" lon="14.640090">
<ele>502.5</ele>
<time>2001-04-29T11:17:33Z</time>
<speed>0.0</speed>
</trkpt>
<trkpt lat="46.038716" lon="14.640190">
<ele>493.9</ele>
<time>2001-04-29T11:17:43Z</time>
<speed>0.0</speed>
</trkpt>
</trkseg>
</trk>
</gpx>
```

Izpeljani formati (npr. GPX, KML...)

- XML sestavljajo (poenostavljeno):
 - značke (ang. tag) – npr. `<time>` in `</time>`
 - Elementi – npr. `<time>2021-02-27T12:01:13Z</time>`
 - lastnosti (atributi). `<gpx version="1.0" ... >`
- XML je okvir, dejansko vsebino pa določajo različne XML sheme, ki definirajo različne podatkovne formate. XML sheme so formalno zapisane v obliki posebnega XSD formata. Pravila za oblikovanje shem: <https://www.w3.org/TR/xmlschema-1/>.
- Primer: format GPX (namenjen podatkom, zajetim z GPS inštrumenti) <https://www.topografix.com/gpx.asp>.
- Primer: KML (Keyhole Markup Language) <https://www.ogc.org/standard/kml/>.

Global Schema Components

Element: **gpx**

Name	gpx
Type	gpxType
Nilable	no
Abstract	no
Documentation	GPX is the root element in the XML file.

☐ XML Instance Representation

```
<gpx
  version="1.1 [1] ?"
  creator="xsd:string [1] ?">
  <metadata> metadataType </metadata> [0..1] ?
  <wpt> wptType </wpt> [0..*] ?
  <rte> rteType </rte> [0..*] ?
  <trk> trkType </trk> [0..*] ?
  <extensions> extensionsType </extensions> [0..1]
  ?
</gpx>
```

☐ Schema Component Representation

```
</trkseg>
</trk>
</gpx>
```



Razširitve sheme

- Mnogi formati (scheme) omogočajo tudi razširitve sheme. V primeru GPX zapisa na desni je glavi z uporabo lastnosti (atributa) "xmlns" navedeno, kje se nahaja definicija razširitve. Znotraj posameznega elementa (npr. <wpt>) pa je uporabljen podelement <extensions> in znotraj tega ena od razširitev, ki so navedene v definiciji razširitve.
- Težava: kreatorji razširitev radi pozabijo, da bi morala biti definicija na trajno dostopnem mestu, zato se lahko zgodi, da definicije razširitev ni več na voljo.

```
<?xml version="1.0" encoding="UTF-8"?>
<xsd:schema xmlns="http://www.garmin.com/xmlschemas/TrackPointExtension/v1"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema"
  targetNamespace="http://www.garmin.com/xmlschemas/TrackPointExtension/v1"
  elementFormDefault="qualified">
  <xsd:annotation>
    <xsd:documentation> This schema defines Garmin extensions to be used
      with the GPX 1.1 schema. The root element defined by this schema is
      intended to be used as a child element of the "extensions" elements in
      the trkpt element in the GPX 1.1 schema. The GPX 1.1 schema is
      available at http://www.topografix.com/GPX/1/1/gpx.xsd. This is a
      replacement for TrackPointExtension in
      http://www.garmin.com/xmlschemas/GpxExtensions/v3 </xsd:documentation>
    </xsd:annotation>
    <xsd:element name="TrackPointExtension" type="TrackPointExtension_t"/>
  </xsd:annotation>
  <xsd:complexType name="TrackPointExtension_t">
    <xsd:annotation>
      <xsd:documentation> This type contains data fields that cannot be
        represented in track points in GPX 1.1 instances. </xsd:documentation>
    </xsd:annotation>
    <xsd:sequence>
      <xsd:element name="atemp" type="DegreesCelsius_t" minOccurs="0"/>
      <xsd:element name="wtemp" type="DegreesCelsius_t" minOccurs="0"/>
      <xsd:element name="depth" type="Meters_t" minOccurs="0"/>
      <xsd:element name="hr" type="BeatsPerMinute_t" minOccurs="0"/>
      <xsd:element name="cad" type="RevolutionsPerMinute_t" minOccurs="0"/>
      <xsd:element name="Extensions" type="Extensions_t" minOccurs="0"/>
    </xsd:sequence>
  </xsd:complexType>
  <xsd:simpleType name="DegreesCelsius_t">
    <xsd:annotation>
      <xsd:documentation> This type contains a temperature value measured in
        degrees Celsius. </xsd:documentation>
    </xsd:annotation>
    <xsd:restriction base="xsd:double"/>
  </xsd:simpleType>
</xsd:schema>
```



FASTA in FASTQ

- Sta besedilna formata, ki sta namenjena shranjevanju DNA/RNA sekvenc:
<https://compgenomr.github.io/book/fasta-and-fastq-formats.html>
- FASTA je starejši in vsebuje samo sekvence (začne se z >) – npr.
 >Mus_musculus_tRNA-Ala-AGC-1-1 (chr13.trna34-AlaAGC)
 GGGGGTGTAGCTCAGTGGTAGAGCGCGTGCTTAGCATGCACGAGGcCCTGGGTTCGATCCCCAGCACCTCCA
- FASTQ je novejši format, ki vsebuje tudi podatke o kvaliteti analize (Phred quality score) – začne se z @.

```
Identifier | @HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
Sequence  | TTAATTGGTAAATAAATCTCCTAATAGCTTAGATNTTACCTTNNNNNNNNNTAGTTTCTTGAGA
+ sign & identifier | +HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
Quality scores | efcfffffcfeefffcfffffdddf`feed]`_]_Ba_^__[YBBBBBBBBBBRTT\]] [] dddd`
```

Base T
phred Quality] = 29



Metapodatkovni standardi in sheme

Metapodatki so podatki, ki opisujejo (raziskovalne) podatke. Njihov glavni namen je iskanje po repozitoriju (ali drugi zbirki podatkov).

Metapodatkovni standardi so večinoma dogovorjena pravila, ki jih vzdržujejo in podpirajo mednarodne organizacije – le nekateri (npr. ISO 19115) so v obliki mednarodnih standardov.

Metapodatkovne sheme predstavljajo ključno sestavino metapodatkovnih standardov in opisujejo strukturo metapodatkov. V praksi marsikdo ta dva izraza uporablja kot sinonima. Metapodatkovne sheme so lahko definirane v običajnim besedilom, lahko pa so zapisane tudi v XSD formatu (XML schema definition) - [https://en.wikipedia.org/wiki/XML_Schema_\(W3C\)](https://en.wikipedia.org/wiki/XML_Schema_(W3C)).

Z metapodatkovnimi standardi so pogosto povezane tudi ontologije/taksonomije/slovarji, ki definirajo nabor ustreznih izrazov.



Metapodatkovni standardi in povezane sheme

CIF - Crystallographic Information Framework

A well-established standard file structure for the archiving and distribution of crystallographic information, CIF is in regular use for reporting crystal structure determinations to Acta Crystallographica and other journals.

Sponsored by the International Union of Crystallography, the current standard dates from 1997. As of July 2011, a new version of the CIF standard is under consideration.

Related Vocabularies	IUCr Recommended abbreviations within data names; CIF dictionaries
Specification	http://www.iucr.org/resources/cif/spec
Standard's website	http://www.iucr.org/resources/cif

ABBREVIATIONS USED IN CIF NAMES

abbrev	abbreviation
abs	absolute (configuration, not structure)
absorpt	absorption

alt
amp
AN
anal
aniso
anisotrop
anom
ASTM
asym

CifCore (CORE_DIC) CIF dictionary

Version 3.2.0 (2023-04-04)

The CIF_CORE dictionary records all the CORE data items defined and used with in the Crystallographic Information Framework (CIF).

Category tree view of data-item definitions

^ DIFFRACTION

The DICTIONARY group encompassing the CORE DIFFRACTION data items defined and used with in the Crystallographic Information Framework (CIF).

CIF ☐

^ DIFFRN

The CATEGORY of data items used to describe the diffraction experiment.

CIF ☐

> _diffrn.ambient_environment
> _diffrn.ambient_pressure
> _diffrn.ambient_pressure_su
> _diffrn.ambient_pressure_gt
> _diffrn.ambient_pressure_lt
> _diffrn.ambient_temperature
> _diffrn.ambient_temperature_su
> _diffrn.ambient_temperature_details



Objava metapodatkovnega standarda v reviji

- Metapodatkovni standard lahko predlaga/vzpostavi tudi skupina zainteresiranih raziskovalcev. Primer: standard BioPAX (Biological Pathway eXchange standard language for integration, exchange and analysis of biological pathway data) je bil vzpostavljen z objavo v reviji (<https://www.nature.com/articles/nbt.1666>).

nature biotechnology

Explore content ▾ About the journal ▾ Publish with us ▾ Subscribe

[nature](#) > [nature biotechnology](#) > [perspectives](#) > article

Perspective | Published: 09 September 2010

The BioPAX community standard for pathway data sharing

[Emek Demir](#), [Michael P. Cary](#), [Suzanne Paley](#), [Ken Fukuda](#), [Christian Lemer](#), [Imre Vastrik](#), [Guanming Wu](#), [Peter D'Eustachio](#), [Carl Schaefer](#), [Joanne Luciano](#), [Frank Schachner](#), [Irma Martinez-Flores](#), [Zhenjun Hu](#), [Veronica Jimenez-Jacinto](#), [Geeta Joshi-Tope](#), [Kumaran Kandasamy](#), [Alejandra C. Lopez-Fuentes](#), [Huaiyu Mi](#), [Elgar Pichler](#), [Igor Rodchenkov](#), [Andrea Splendiani](#), [Sasha Tkachev](#), [Jeremy Zucker](#), [Gopal Gopinath](#), ... [Gary D. Bader](#)  [+ Show authors](#)

Nature Biotechnology **28**, 935–942 (2010) | [Cite this article](#)

10k Accesses | 9 Altmetric | [Metrics](#)

Abstract

Biological Pathway Exchange (BioPAX) is a standard language to represent biological pathways at the molecular and cellular level and to facilitate the exchange of pathway data. The rapid growth of the volume of pathway data has spurred the development of databases and computational tools to aid interpretation; however, use of these data is hampered by the current fragmentation of pathway information across many databases with incompatible formats. BioPAX, which was created through a community process, solves this problem by making pathway data substantially easier to collect, index, interpret and share. BioPAX can represent metabolic and signaling pathways, molecular and genetic interactions and gene regulation networks. Using BioPAX, millions of interactions, organized into thousands of pathways, from many organisms are available from a growing number of databases. This large amount of pathway data in a computable form will support visualization, analysis and biological discovery.



Vrste metapodatkov 1

Za naše potrebe lahko metapodatke razdelimo na:

- **Splošne oziroma bibliografske metapodatke** – to so podatki o avtorju, njegovi inštituciji, letu objave/izdaje, vrsti dokumenta, jeziku, pa tudi o načinu financiranja, številki projekta, avtorskih pravicah in licencah in podobno.

Nabor teh metapodatkov je zelo podoben za različne objekte od tiskanih del preko monografij in člankov v elektronski obliki do raziskovalnih podatkov.

Večina metapodatkovnih standardov, ki so v uporabi v različnih repozitorijih, arhivih in knjižnicah uporablja zelo podoben nabor teh metapodatkov.



Vrste metapodatkov 2

- **Vsebinske oziroma faktografske metapodatke** – to so tisti, ki vsebuje podatke o sami meritvi/eksperimentu. V to skupino sodijo podatki o vzorcu/snovi/organizmu (ime, klasifikacije...), o metodi in inštrumentu (naziv, proizvajalec), o parametrih meritve (npr. temperatura...) in podobno.

Ti podatki že spadajo v področje **provenience** in so pomembni za **ponovno uporabljivost in interoperabilnost** naših podatkov.

V metapodatkovnih shemah so običajno vključeni tisti podatki, ki posamezne meritve ločijo med seboj v skupine z namenom, da lahko (v repozitoriju) hitro in učinkovito poiščemo objavljene podatke z določenimi lastnostmi, ki nas zanimajo.



Kako izbrati metapodatkovni standard

- Večinoma nam **minimalni/obvezni** metapodatkovni standard določi kar **repozitorij**, v katerem želimo objaviti podatke. Zato je dobro že pred začetkom raziskovanja ali vsaj pred začetkom meritev načrtovati, katere metapodatke bomo potrebovali.



Kje najdemo podatke o metapodatkovnih shemah

- Najbolj enostaven način, da izvemo, katere podatke bomo morali zbrati, je da pogledamo obstoječe objave v izbranem repozitoriju ali pa poizkusimo, katere podatke bi morali vnesti ob dejanski oddaji podatkov v izbrani repozitorij.
- Še bolje pa je, da na spletni strani repozitorija poiščemo navodila, kakšne podatke in v kakšni obliki je potrebno vnesti ob vnosu samih raziskovalnih podatkov.
- Če pa nas zanima kaj več, še posebej če načrtujemo novo podatkovno shemo ali programsko opremo za pretvorbo metapodatkov iz ene sheme v drugo, pa je najbolje, da preučimo podatke o metapodatkovnih standardih, ki so običajno objavljeni na njihovih matičnih spletnih straneh.



Katalogi metapodatkovnih standardov

- Pregled metapodatkovnih standardov najdemo v Katalogu podatkovnih standardov - **Metadata Standards Catalog** (<https://rdamsc.bath.ac.uk/>) , ki je nadaljevanje istoimenskega kataloga iz spletnega mesta RDA (<https://rd-alliance.github.io/metadata-directory/>).
- Brskamo lahko po imenih metapodatkovnih standardov ali po področjih uporabe. Pri vsakem od standardov lahko dobimo povezave na matično spletno stran standarda, pri večini pa tudi na definicijo standarda v besedilni obliki in v XSD formatu.



Metadata Standards Catalog

Metadata Standards Catalog

Search Sign in

Index of metadata standards

ABCD (Access to Biological Collection Data)
ABCD Zoology
ABCD DNA
ABCEFG (Access to Biological Collection Databases Extended for Geosciences)
HISPID (Herbarium Information Standards and Protocols for Interchange of Data)
AgMES (Agricultural Metadata Element Set)
AGRI Application Profile
AVM (Astronomy Visualization Metadata)
Brain Imaging Data Structure (BIDS)
CEDAR Template Model
CERIF (Common European Research Information Format)
OpenAIRE Guidelines
CF (Climate and Forecast) Metadata Conventions
COARDS Conventions

Metadata Standards Catalog

Search Sign in

MARC (Machine-Readable Cataloging)

MARC is a standard and serialization format for representing bibliographic metadata, originally designed as a way of exchanging bibliographic records between library catalogs. Various different versions have been defined, mostly with national or regional scope, of which MARC 21 is probably the most widely used. There also exists an XML serialization of MARC 21, known as MARCXML.

Used in

Bibliographic standards

Information/library standards

Documentation

Visit website

View specification

View specification in XSD format

Responsible organizations

- Maintainer: [Library of Congress](#)

View website

```
<xsd:schema xmlns="http://www.loc.gov/MARC21/slim"
xmlns:xsd="http://www.w3.org/2001/XMLSchema"
targetNamespace="http://www.loc.gov/MARC21/slim"
elementFormDefault="qualified" attributeFormDefault="unqualified"
version="1.1" xml:lang="en">
  <xsd:annotation>
    <xsd:documentation> MARCXML: The MARC 21 XML Schema Prepared by
    Corey Keith May 21, 2002 - Version 1.0 - Initial Release
    ***** Changes. August 4,
    2003 - Version 1.1 - Removed import of xml namespace and the use
    of xml:space="preserve" attributes on the leader and
    controlfields. Whitespace preservation in these subfields is
    accomplished by the use of xsd:whiteSpace value="preserve" May 21,
    2009 - Version 1.2 - in subfieldcodeDataType the pattern "[\da-
    z!#$%&'()*+,-./:;<=>?]{_}^~\[\]\{1}" changed to: "[\da-Za-
    z!#$%&'()*+,-./:;<=>?]{_}^~\[\]\{1}" i.e "A-Z" added after "
    [\d" before "a-z" to allow upper case. This change is for
    consistency with the documentation.
    ***** This
    schema supports XML markup of MARC21 records as specified in the
    MARC documentation (see www.loc.gov). It allows tags with
    alphabetic and subfield codes that are symbols, neither of which
    are as yet used in the MARC 21 communications formats, but are
    allowed by MARC 21 for local data. The schema accommodates all
    types of MARC 21 records: bibliographic, holdings, bibliographic
    with embedded holdings, authority, classification, and community
    information. </xsd:documentation>
  </xsd:annotation>
  <xsd:element name="record" type="recordType" nillable="true"
  id="record.e">
    <xsd:annotation>
      <xsd:documentation>record is a top level container element for
      all of the field elements which compose the
      record</xsd:documentation>
    </xsd:annotation>
  </xsd:element>
```

Digital Curation Centre (DCC)

Nekateri drugi katalogi metapodatkovnih standardov so še:

- Katalog metapodatkovnih standardov na Digital Curation Centre (DCC) - <https://www.dcc.ac.uk/guidance/standards/metadata>

ABCD - Access to Biological Collection Data
The Access to Biological Collections Data (ABCD) Schema is an evolving comprehensive standard for the access to and exchange of data about specimens and observations (a.k.a. primary biodiversity data). The ABCD Schema attempts to be comprehensive and highly structured, supporting data from a wide variety of databases. It is compatible with several existing data standards. Parallel structures exist so that either (or both) atomised data and free-text can be accommodated.

Sponsored by Biodiversity Information Standards TDWG - the Taxonomic Databases Working Group, the current specification was last modified in 2007.

AgMES - Agricultural Metadata Element Set
A semantic standard developed by the Food and Agriculture Organization (FAO) of the United Nations, AgMES enables description, resource discovery, interoperability and data exchange of different types of information resources in all areas relevant to food production, nutrition and rural development.

Sponsored by the UN AIMS - Agricultural Information Management Standards, the current standard was issued in November 2010.

CIF - Crystallographic Information Framework

A well-established standard file structure for the archiving and distribution of crystallographic information, CIF is in regular use for reporting crystal structure determinations to Acta Crystallographica and other journals.

Sponsored by the International Union of Crystallography, the current standard dates from 1997. As of July 2011, a new version of the CIF standard is under consideration.

Related Vocabularies	IUCr Recommended abbreviations within data names; CIF dictionaries
Specification	http://www.iucr.org/resources/cif/spec
Standard's website	http://www.iucr.org/resources/cif



FAIRsharing.org

- Katalog standardov za metapodatke, ontologije, formate... na FAIRsharing.org - <https://fairsharing.org/search?fairsharingRegistry=Standard>.

The screenshot shows the FAIRsharing.org search interface. At the top, there's a search bar with 'datacite metadata' entered. Below the search bar, it says 'Displaying 1 to 16 of 16.' The main result is for 'DataCite Metadata Schema'. It includes a description: 'The DataCite Metadata Schema is a list of core metadata properties chosen for accurate and consistent identification of...'. There are three tabs: 'Subject Ag...', 'Citation', and 'All'. At the bottom, there are statistics: 'Linked Collections: 8', 'Linked Databases: 89', and 'Linked Policies: 18'.

The screenshot shows the details page for the 'DataCite Metadata Schema'. It includes a 'GENERAL INFORMATION' section with the following details:

- Type: Model and format
- Registry: Standard
- Description: The DataCite Metadata Schema is a list of core metadata properties chosen for accurate and consistent identification of a resource for citation and retrieval purposes, with recommended use instructions in the documentation. The resource that is being identified can be of any kind, but it is typically a dataset. We use the term 'dataset' in its broadest sense. We mean it to include not only numerical data, but any other research objects in keeping with DataCite's mission (<https://datacite.org/value.html>). The metadata schema properties are presented and described in detail in the section DataCite Metadata Properties in this document.
- Homepage: <https://schema.datacite.org/>
- Year of Creation: 2009
- Maintainers: DataCite
- Countries developing this resource: Worldwide
- Subjects: Subject Agnostic
- Domains: Citation, Resource Metadata, Annotation, Data Retrieval, Centrally Registered Identifier
- Taxonomic Range: All
- User Defined Tags: Metadata Standardization

At the bottom, there is a 'VIEW RELATION GRAPH' button.



re3data.org

Zelo priročno orodje za raziskovanje metapodatkovnih standardov/shem je re3data.org (<https://www.re3data.org/>). Gre sicer za orodje za iskanje repozitorijev, s pomočjo katerega pa lahko vidimo, kateri metapodatkovni standard (ali standarde) uporablja določen repozitorij.

Nekateri repozitoriji imajo navedeno več kot en standard/shemo, kar pomeni, da uporabljajo lastno shemo, ki pa je združljiva z navedenimi shemami (torej vsebuje najmanj vse metapodatke, ki jih zahteva vsaka od teh shem).



Primeri: MARC

- MARC (<https://www.loc.gov/marc/>) gre za enega najstarejših metapodatkovnih standardov, ki je znan predvsem knjižničarjem, saj tudi slovenski knjižnični katalog COBISS temelji na izvedenki tega standarda, imenovani COMARC.

Table of Contents

- [Introduction](#) [Full | Concise]
- [Format Summary](#)
- [Leader](#) [Full | Concise]
- [Directory](#)
- [00X: Control Fields](#)
- [01X-09X: Numbers and Code Fields](#)
- [Heading Fields - General Information](#)
- [1XX: Main Entry Fields](#)
- [20X-24X: Title and Title-Related Fields](#)
- [25X-28X: Edition, Imprint, Etc. Fields](#)
- [3XX: Physical Description, Etc. Fields](#)
- [4XX: Series Statement Fields](#)
- [5XX: Note Fields](#)
- [6XX: Subject Access Fields](#)
- [70X-75X: Added Entry Fields](#)
- [76X-78X: Linking Entry and Description Fields](#)
- [80X-83X: Series Added Entry Fields](#)
- [841-88X: Holdings, Location, Alternate Graphics, Etc. Fields](#)
- [Appendix A: Control Subfields](#)
- [Appendix B: Full Level Record Examples](#)
- [Appendix C: Minimal Level Record Examples](#)
- [Appendix D: Multiscript Records](#)
- [Appendix E: Alphabetical List of Ambiguous Headings](#)
- [Appendix F: Initial Definite and Indefinite Articles](#)
- [Appendix G: Format Changes for Update No. 38 \(June 2024\)](#)
- [Appendix H: Local Data Elements](#)
- [Appendix I: Organization Code Sources](#)
- [Appendix J: Data Provenance Subfields](#)



Dublin Core

- Dublin Core (<https://www.dublincore.org/>) je zelo znan in razširjen standard, uporaben tako za klasične knjižnične kataloge kot za repozitorije. V kombinaciji z DataCite standardom ga uporablja tudi Zenodo.

DublinCore DCMI-2024 Conference Search

Specifications Events Community News Resources About DCMI

Home / DCMI Specifications

Dublin Core™

The original Dublin Core™ of thirteen (later fifteen) elements was first published in the [report of a workshop in 1995](#). In 1998, this was formalized in the Internet Engineering Task Force standard [RFC 2413](#), and [discussions began](#) about making it a standard of the (US) National Information Standards Organization (NISO). The European Committee for Standardization (CEN) published the Dublin Core Metadata Element Set as [CWA 13874](#). The element set was then published as the US standard ANSI/NISO Z39.85-2001 and as an international standard, ISO 15836-2003. The most recent updates of these standards are [RFC 5791 \(2010\)](#), [Z39-85-2012](#), and [ISO 15836-1:2017](#). [ISO 15836 Part 2](#), covering several dozen properties and classes that have been added to DCMI namespaces since 1999, was published in 2019.

Starting in 2002, DCMI grew into the role of "de facto" standards agency by maintaining its own, updated documentation for [DCMI Metadata Terms](#). The [DCMI Usage Board](#) currently serves as the maintenance agency for ISO 15836.

In addition to these semantic specifications, DCMI working groups have developed specifications on other topics of relevance to metadata, such as encoding syntaxes, usage guidelines, and metadata models. In the rapidly evolving environment of the World Wide Web, most of these specifications have been superseded over time, sometimes after influencing subsequent work by other technical communities, notably the World Wide Web



DataCite

- DataCite (<https://schema.datacite.org/>) je standard pogosto uporabljen v repozitorijih, ki so v skladu s priporočili OpenAIRE za podatkovne arhive (<https://guidelines.openaire.eu/en/latest/>). Uporabljajo ga slovenski univerzitetni repozitoriji in DIRROS. V kombinaciji z Dublin Core standardom ga uporablja tudi Zenodo.

🏠 / DataCite Metadata Properties

DataCite Metadata Properties

- [Overview](#)
 - [Conventions](#)
 - [Levels of obligation](#)
 - [Naming and numbering](#)
 - [Occurrences](#)
 - [XML schema](#)
 - [Mandatory Properties](#)
 - [Table 1: DataCite Mandatory Properties](#)
 - [Recommended and Optional Properties](#)
 - [Table 2: DataCite Recommended and Optional Properties](#)
- [1. Identifier](#)
 - [1.a identifierType](#)
- [2. Creator](#)
 - [2.1 creatorName](#)
 - [2.1.a nameType](#)
 - [2.2 givenName](#)
 - [2.3 familyName](#)
 - [2.4 nameIdentifier](#)
 - [2.4.a nameIdentifierScheme](#)
 - [2.4.b schemeURI](#)
 - [2.5 affiliation](#)
 - [2.5.a affiliationIdentifier](#)



DDI (Data Documentation Initiative)

- DDI (Data Documentation Initiative) (<https://ddialliance.org/Specification/>) je standard, ki je pogosto uporabljan na področju družboslovja. Uporablja ga tudi Arhiv družboslovnih podatkov (ADP).



DDI Specification

DDI has two development lines, described here.

DDI-Lifecycle

Encompassing all of the DDI-Codebook specification and extending it, [DDI-Lifecycle](#) is designed to document and manage data across the entire life cycle, from conceptualization to data publication and analysis and beyond. Based on XML Schemas, DDI-Lifecycle is modular and extensible.

Users new to DDI are encouraged to use this DDI-Lifecycle development line as it incorporates added functionality. Use DDI-Lifecycle if you are interested in:

- Metadata reuse across the data life cycle
- Metadata-driven survey design
- Question banks
- Complex data, e.g., longitudinal data
- Detailed geographic information
- Multiple languages
- Compliance with other metadata standards like ISO 11179
- Process management and automation

The current version of the DDI-L Specification is Version 3.1. [DDI 3.1](#) was published in October 2009, superseding DDI 3.0 (published in April 2008).

DDI-Codebook

[DDI-Codebook](#) is a more light-weight version of the standard, intended primarily to document simple survey data. Originally DTD-based, DDI-C is now available as an XML Schema.

The current version of DDI-C is 2.*.

Drugi metapodatkovni standardi

Biologija

Name	Link	Description
4DN-BINA-OME	https://github.com/WU-BIMAC/NBOMicroscopyMetadataSpecs	minimum Information guidelines for fluorescence microscopy OME - Open Microscopy Environment model, a specification for storing data on biological imaging - https://docs.openmicroscopy.org/ome-model/5.6.3/index.html Biological PATHway eXchange standard language for integration, exchange and analysis of biological pathway data
BioPAX	https://www.biopax.org/	Describes how semantic annotations should be made about mathematical models encoded in the CellML (www.cellml.org) format, and their elements.
CellML	https://cellml-metadata-spec-20.readthedocs.io/en/latest/	Low Entropy Model Specification
LEMS	https://lems.github.io/LEMS/	Minimum Information About a Cellular Assay
MIACA	https://miaca.sourceforge.net/	A standardized description of cell-based functional assay projects
MIAME	http://fged.org/projects/miame/	Minimum Information about a Microarray Experiment
MIAPA	https://www.ebi.ac.uk/ols4/ontologies/miapa	Minimum Information About a Phylogenetic Analysis
MIAPPE	https://www.miappe.org/	Minimum information about a proteomics experiment
MIARE	https://mibbi.sourceforge.net/projects/MIARE.shtml	Minimum Information About Plant Phenotyping Experiments
MIASE	http://co.mbine.org/standards/miase	Minimum Information About an RNAi Experiment
MIFlowCyt		A checklist describing the information that should be reported for an RNA interference experiment.
minSCe		Minimum Information about a Simulation Experiment
MINSEQE	https://www.fged.org/projects/minseqe/	Minimum Information about a Flow Cytometry Experiment
MIQE		Minimum Information about a Single-Cell Experiment
MIRAGE		Minimum Information About a Next-generation Sequencing Experiment
MIRIAM	http://co.mbine.org/standards/miriam	Minimum Information for Quantitative Real-Time PCR Experiments
MISFISHIE		Minimum Information Required About a Glycomics Experiment
MixS	https://www.gensc.org/pages/standards-intro.html	Minimum Information Required in the Annotation of Models
MixS - MIMARKS		Minimum information specification for in situ hybridization and immunohistochemistry experiments
NeuroML	https://neuroml.org/	Minimum Information about any Sequence
REMBI	Recommended Metadata for Biological Images—enabling reuse of microscopy data in biology	Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MixS) specifications
SBGN	http://www.sbgm.org/	Language for describing neuronal systems
SBML	http://sbml.org/	Systems Biology Graphical Notation
SBOL	https://sbolstandard.org/	Systems Biology Mark-up Language
		Synthetic Biology Open Language



Združljivost z OpenAIRE

Za to, da je omogočeno harvestiranje repozitorijev in s tem gradnja drevesne strukture OpenAIRE infrastrukture morajo biti repozitoriji in njihove metapodatkovne sheme združljive z zahtevami OpenAIRE in protokolom OAI-PMH. Zahteve (oziroma navodila) so zbrane na strani **OpenAIRE Guidelines** (<https://guidelins.openaire.eu/en/latest/>).



Metapodatki v podatkovnih datotekah

Nekateri formati datotek omogočajo hranjenje določenih metapodatkov v samih podatkovnih datotekah (ali v ločenih, vendar povezanih datotekah):

- EXIF – metapodatkovni standard za shranjevanje metapodatkov v multimedijskih (slikovnih in zvočnih) formatih kot so JPG, TIF, PNG, WEBP in WAV. EXIF podatki zajemajo tip in nastavitve kamere, podatke o objektivu, podatke o sliki (velikost, barve), čas in lokacijo posnetka, avtorstvo... EXIF metapodatkovni standard je razširljiv z lastnimi podatki proizvajalcev fotografske in programske opreme.
- GEOTIFF – metapodatkovni standard, ki omogoča zapis georeferenčnih podatkov (lokacije) v samo fotografijo v TIFF formatu. Alternativa temu je TFW (World File) format, kjer so podatki o položaju in orientaciji fotografije shranjeni v istoimenski datoteki s končnico TFW.



Geslovniki, šifranti, taksonomije, ontologije...

Geslovniki, slovarji, šifranti, registri, tezavri, nomenklature, taksonomije in ontologije so sistemi, ki omogočajo enotno in nedvoumno rabo strokovnih izrazov in različnih identifikatorjev ter odnosov med njimi v znanstveni komunikaciji.

Zato je njihova uporaba bistvena za doseganje **interoperabilnosti** objavljenih raziskovalnih podatkov.

Medtem ko metapodatkovne sheme določajo **vrste metapodatkov**, prej naštetih sistemi (geslovniki, šifranti ... ontologije) določajo **vsebino metapodatkov**.

Če na primer obstaja metapodatkovno polje „ime vrste“, potem biološka klasifikacija (taksonomija) določa, kako se tvori to ime (latinskega imena rodu in vrste). Če obstaja polje „ime spojine“, potem kemijska nomenklatura določa, kako se to zapiše na enoten in nedvoumen način.



Geslovniki, terminološki slovarji, leksikoni

Geslovnik / terminološki slovar / leksikon – Seznam pojmov iz nekega strokovnega področja (s kratko razlago).

sl monografija

1. navadno znanstvena publikacija, ki obravnava eno vprašanje ali eno temo
2. monografska publikacija

V različnih virih imamo nekoliko različne definicije teh pojmov. Po Fran-u je geslovnik „abecedni seznam besednih enot, določenih za gesla ali podgesla v slovarju, enciklopediji“ (<https://fran.si/iskanje?Query=geslovnik>).

primeri geslovnikov, terminoloških slovarjev in leksikonov :

- Bibliotekarski terminološki slovar (<https://www.termania.net/?searchIn=Linked&Id=85>).
- Leksikon računalništva in informatike (tiskan).
- Splošni geslovnik COBISS.SI (<https://www.cobiss.si/sgc/>).



Šifranti in registri

Šifrant – seznam, ki povezuje pojme z identifikacijskimi številkami, ki v neki zbirki podatkov nadomestijo te pojme.

Register – seznam objektov ali pojmov (mnogokrat opremljen z identifikacijskimi številkami), ki vsebuje podatke, na podlagi katerih je možno enolično določiti posamezen objekt/pojem v registru.



Primer: CAS Registry Number

CAS Registry Number je številka, ki enolično določa neko kemijsko spojino – v samem registru, pa so poleg imena v skladu s kemijsko nomenklaturo tudi podatki o objavah, kjer so spojina in njene lastnosti nedvoumno opisane.

- Primeri registrov so še register prebivalstva, CONOR, ORCID...

CAS Common Chemistry

← Return to Results

Caffeine

CAS Registry Number®
58-08-2

CAS Name
Caffeine

Molecular Formula
 $C_8H_{10}N_4O_2$

Molecular Mass
194.19

Discover more in Scifinder®

Cite this Page
Caffeine. CAS Common Chemistry. CAS, a division of the American Chemical Society, n.d. https://commonchemistry.cas.org/detail?cas_rn=58-08-2 (retrieved 2022-01-24) (CAS RN: 58-08-2). Licensed under the Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Compound Properties

Melting Point (1)
238 °C

Density (2)
1.23 g/cm³ @ Temp: 18 °C

Source(s)
(1) International Chemical Safety Cards data were obtained from the National Institute for Occupational Safety and Health (US)
(2) Hazardous Substances Data Bank data were obtained from the National Library of Medicine (US)

Other Names and Identifiers

InChI
InChI=1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3

InChIKey
InChIKey=RYYVLZVUVJVGH-UHFFFAOYSA-N

SMILES
O=C1C2=C(N(C)C(=O)N1C)N=CN2C

Canonical SMILES
O=C1C2=C(N=CN2C)N(C)=O/N1C/C

Other Names for this Substance

- 1H-Purine-2,6-dione, 3,7-dihydro-1,3,7-trimethyl-
- Caffeine
- 3,7-Dihydro-1,3,7-trimethyl-1H-purine-2,6-dione
- Guanine
- Methyltheobromine

View All

Deleted or Replaced CAS Registry Numbers
95789-13-2, 71701-02-5

CAS Common Chemistry is provided under the Creative Commons Attribution-NonCommercial 4.0 International License, or CC BY-NC 4.0 license. By using CAS Common Chemistry, you agree to the terms and conditions of this license. To use or license CAS Common Chemistry for commercial purposes, contact us.



Tezavri

Tezaver – organiziran in kontroliran geslovník v katerem so odnosi med pojmi jasno predstavljeni na standardiziran način (npr. širši/ožji pojem, sinonim, nasprotje...).

primer:

MeSH (Medical Subject Headings)

(<https://meshb.nlm.nih.gov/>) je en najbolj znanih tezavrov. Organiziran je v obliki drevesa s 16 vejami, ki pokrivajo različne aspekte medicine in se med seboj prepletajo. Glavne veje so označene s črkami, podrejeni pojmi pa s številkami. Npr. pojem Foot [A01.378.610.250] v veji Anatomy [A] je povezan s pojmom Foot Diseases [C17.800.321] iz veje Diseases [C].

Foot MeSH Descriptor Data 2024

Details	Qualifiers	MeSH Tree Structures	Concepts
MeSH Heading	Foot		
Tree Number(s)	A01.378.610.250		
Unique ID	D005528		
RDF Unique Identifier	http://id.nlm.nih.gov/mesh/D005528		
Annotation	vertebrates only; for invertebrates use EXTREMITIES: dis of foot = FOOT DISEASES; st		

Sc
Ent
Entry Co
Date E
Da
Rev

Foot Diseases MeSH Descriptor Data 2024

Details Qualifiers MeSH Tree Structures Concepts

Musculoskeletal Diseases [C05]
Bone Diseases [C05.116] +
Cartilage Diseases [C05.182] +
Fasciitis [C05.321] +
Foot Deformities [C05.330] +
Foot Diseases [C05.360] -
Fasciitis, Plantar [C05.360.350]
Fibromatosis, Plantar [C05.360.375]
Heel Spur [C05.360.400]
Metatarsalgia [C05.360.500] +
Posterior Tibial Tendon Dysfunction [C05.360.750]
Hand Deformities [C05.390] +
Jaw Diseases [C05.500] +



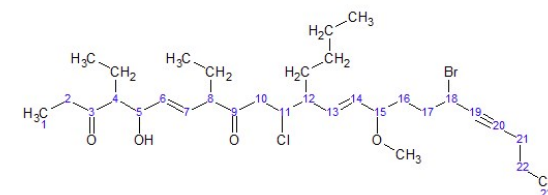
Nomenklature

Nomenklatura – sistem poimenovanja objektov ali pojmov, ki vključuje pravila za tvorjenje imen teh objektov ali pojmov.



Primer: IUPAC nomenklatura kemijskih imen

- IUPAC nomenklatura kemijskih imen – razdeljena je na več delov (barv), najbolj pomembna sta Nomenklatura anorganske kemije (Red Book) in Nomenklatura organske kemije (Blue Book) – vse najdemo na <https://iupac.org/what-we-do/books/>, bolj enostavno in kratko pa so principi predstavljeni na Wikipediji (https://en.wikipedia.org/wiki/IUPAC_nomenclature_of_organic_chemistry in https://en.wikipedia.org/wiki/IUPAC_nomenclature_of_inorganic_chemistry_2005).
- Podobno je nomenklatura imen rastlinskih vrst je del biološke sistematike.



Now, following the above steps:

- The parent **hydrocarbon** chain has 23 carbons. It is called tricos-.
- The functional groups with the highest precedence are the two ketone groups.
 - The groups are on carbon atoms 3 and 9. As there are two, we write 3,9-dione.
 - The numbering of the molecule is based on the ketone groups. When numbering from left to right, the ketone groups are numbered 3 and 9. When numbering from right to left, the ketone groups are numbered 15 and 21. 3 is less than 15, therefore the ketones are numbered 3 and 9. The smaller number is always used, not the sum of the constituents numbers.
- The side chains are: an ethyl- at carbon 4, an ethyl- at carbon 8, and a butyl- at carbon 12.

Note: the $-O-CH_3$ at carbon atom 15 is not a side chain, but it is a methoxy functional group.

 - There are two ethyl- groups. They are combined to create, 4,8-diethyl.
 - The side chains are grouped like this: 12-butyl-4,8-diethyl. (But this is not necessarily the final grouping, as functional groups may be added in between to ensure all groups are listed alphabetically.)
- The secondary functional groups are: a hydroxy- at carbon 5, a chloro- at carbon 11, a methoxy- at carbon 15, and a bromo- at carbon 18. Grouped with the side chains, this gives 18-bromo-12-butyl-11-chloro-4,8-diethyl-5-hydroxy-15-methoxy.
- There are two double bonds: one between carbons 6 and 7, and one between carbons 13 and 14. They would be called "6,13-diene", but the presence of alkynes switches it to 6,13-dien. There is one triple bond between carbon atoms 19 and 20. It will be called 19-yne.
- The arrangement (with punctuation) is: 18-bromo-12-butyl-11-chloro-4,8-diethyl-5-hydroxy-15-methoxytricos-6,13-dien-19-yne-3,9-dione
- Finally, due to **cis-trans isomerism**, we have to specify the relative orientation of functional groups around each double bond. For this example, both double bonds are trans isomers, so we have (6E,13E)

The final name is (6E,13E)-18-bromo-12-butyl-11-chloro-4,8-diethyl-5-hydroxy-15-methoxytricos-6,13-dien-19-yne-3,9-dione.

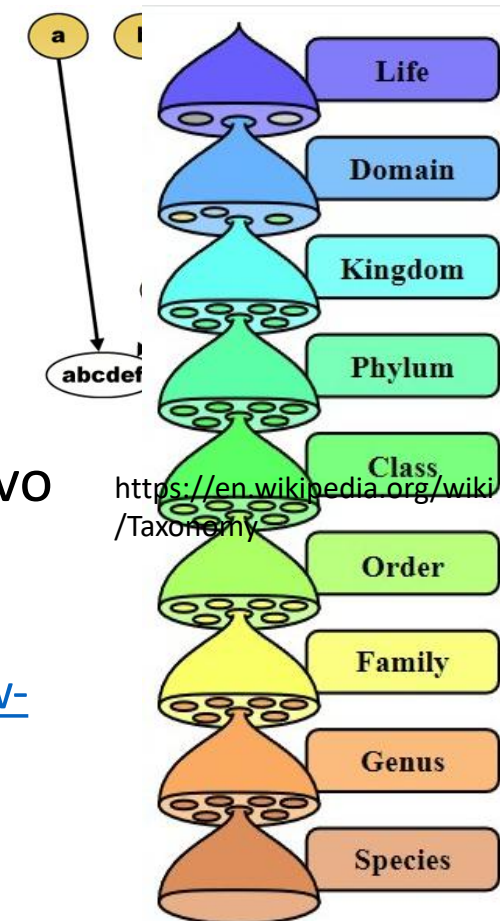


Taksonomije

Taksonomija – hierarhična klasifikacija, pri kateri so stvari organizirane v skupine oziroma tipe v obliki drevesne strukture.

Primeri:

- Klasifikacija (taksonomija) živih bitij je hierarhično urejeno drevo skupin živih bitij na podlagi njihove sorodnosti. Vsak taksonomski nivo vsebuje skupine imenovane taksoni. Vsak takson ima en nadrejeni takson in enega ali več podrejenih taksonov. Na najnižjem nivoju so posamezne vrste živih bitij (seznam: <https://www.enago.com/academy/how-to-write-scientific-names-in-a-research-paper-animals-plants/>).
- UDK (Univerzalna decimalna klasifikacija) klasificira literaturo glede na strokovna in literarna področja s pomočjo številčnega identifikatorja različnih dolžin.



<https://en.wikipedia.org/wiki/Taxonomy>

[https://en.wikipedia.org/wiki/Taxonomy_\(biology\)](https://en.wikipedia.org/wiki/Taxonomy_(biology))



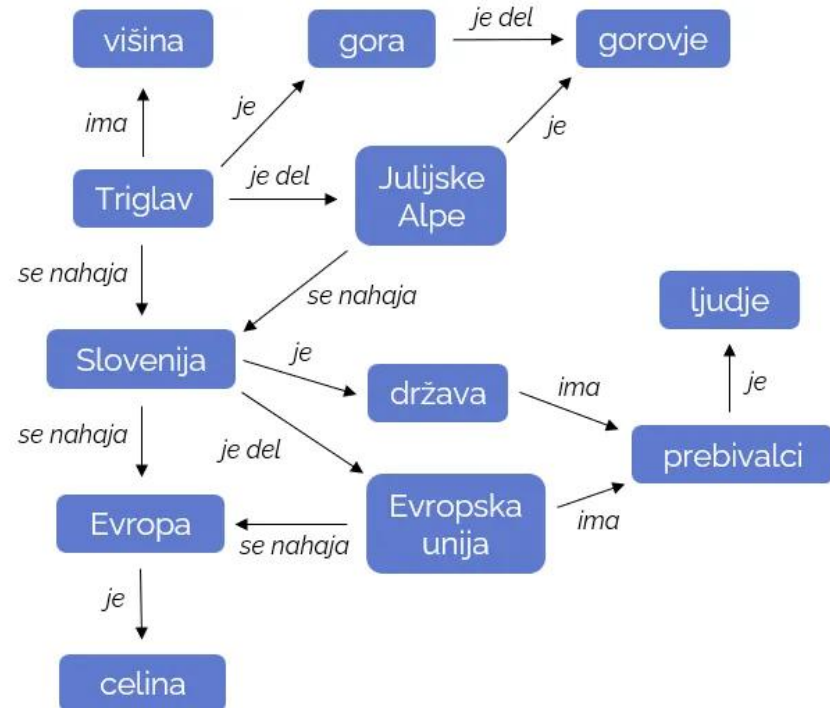
Ontologije

Ontologija – v širšem pomenu sistem organizacije znanja – gre za okvir oziroma strukturo v kateri so organizirani katalogi, taksonomije, terminologije...; v ožjem pomenu pa se nanaša na razlago skupine idej oziroma konceptov v določeni domeni, ki določa medsebojno povezanost teh idej oziroma konceptov (npr. v umetni inteligenci, biologiji).

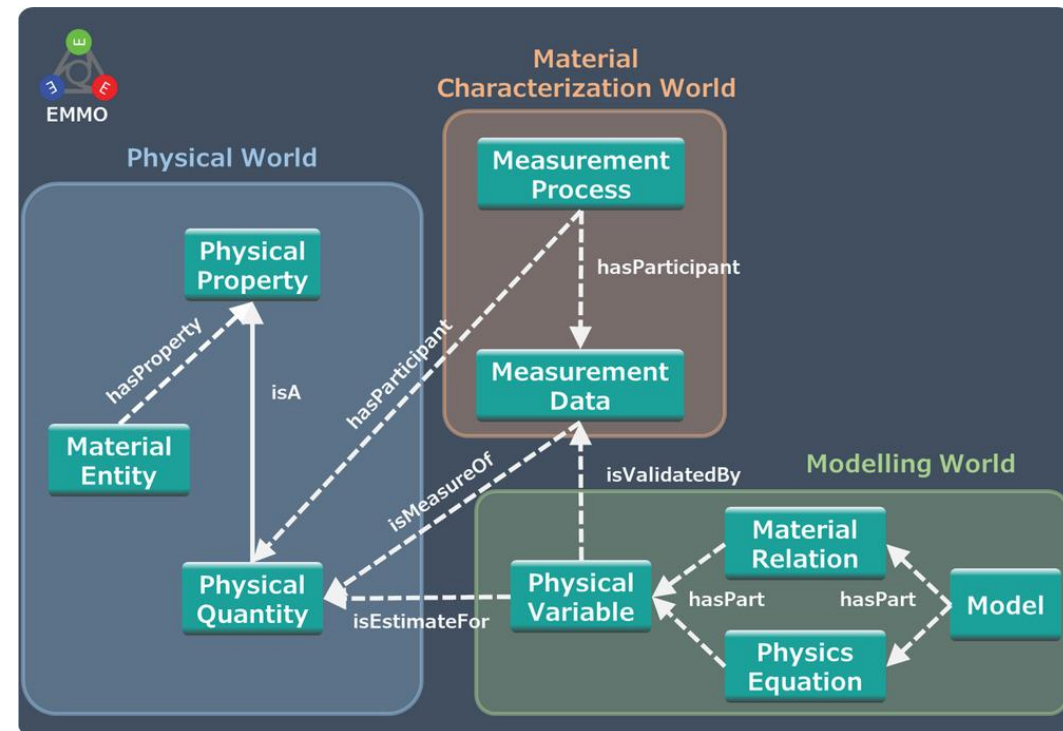
- Ontologija torej povezuje **različne vrste pojmov z različnimi vrstami povezav/relacij** in je zato najbolj splošna oblika klasifikacijskega sistema.
- Relacije lahko zapišemo v obliki **semantičnih trojčkov** subjekt – odnos – objekt (na primer Triglav – je – gora ali Triglav – je del – Julijskih Alp).
- Semantična mreža povezuje semantične trojčke med seboj.

Ontologije in semantične mreže

Semantično mrežo lahko narišemo v obliki semantičnega grafa, ki predstavlja grafično reprezentacijo odnosov med objekti v ontologiji:



Vir: <https://dirrosdata.ctl.uni-lj.si/metapodatki/metapodatkovne-ontologije/>



Vir: <https://onlinelibrary.wiley.com/doi/10.1002/aenm.202102702>



FAIRsharing.org

- Katalog standardov za metapodatke, ontologije, formate... na FAIRsharing.org - <https://fairsharing.org/search?fairsharingRegistry=Standard>.

The screenshot shows the FAIRsharing.org search interface. At the top, there's a search bar with 'datacite' entered. Below the search bar, it says 'Displaying 1 to 24 of 24.' The main result is for 'DataCite Ontology'. It includes a description: 'The DataCite Ontology (DataCite) is an ontology that enables the metadata properties of the DataCite Metadata Sche...'. There are three buttons: 'Subject Ag...', 'Citation', and 'Not applic...'. Below the description, there are links for 'Linked Collections' (0), 'Linked Databases' (5), and 'Linked Policies' (9).

DataCite Ontology (DataCite)

URL

<http://purl.org/spar/datacite> (alternative at w3id.org)

DOI

10.25504/FAIRsharing.c06f1e

Documentation

<http://purl.org/spar/datacite.html>

Source

<http://purl.org/spar/datacite.xml> (RDF/XML)

<http://purl.org/spar/datacite.ttl> (Turtle)

<http://purl.org/spar/datacite.nt> (N-triples)

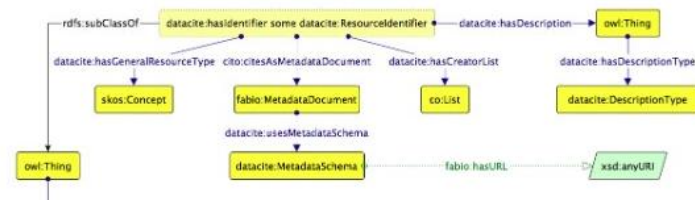
<http://purl.org/spar/datacite.json> (JSON-LD)

Repository

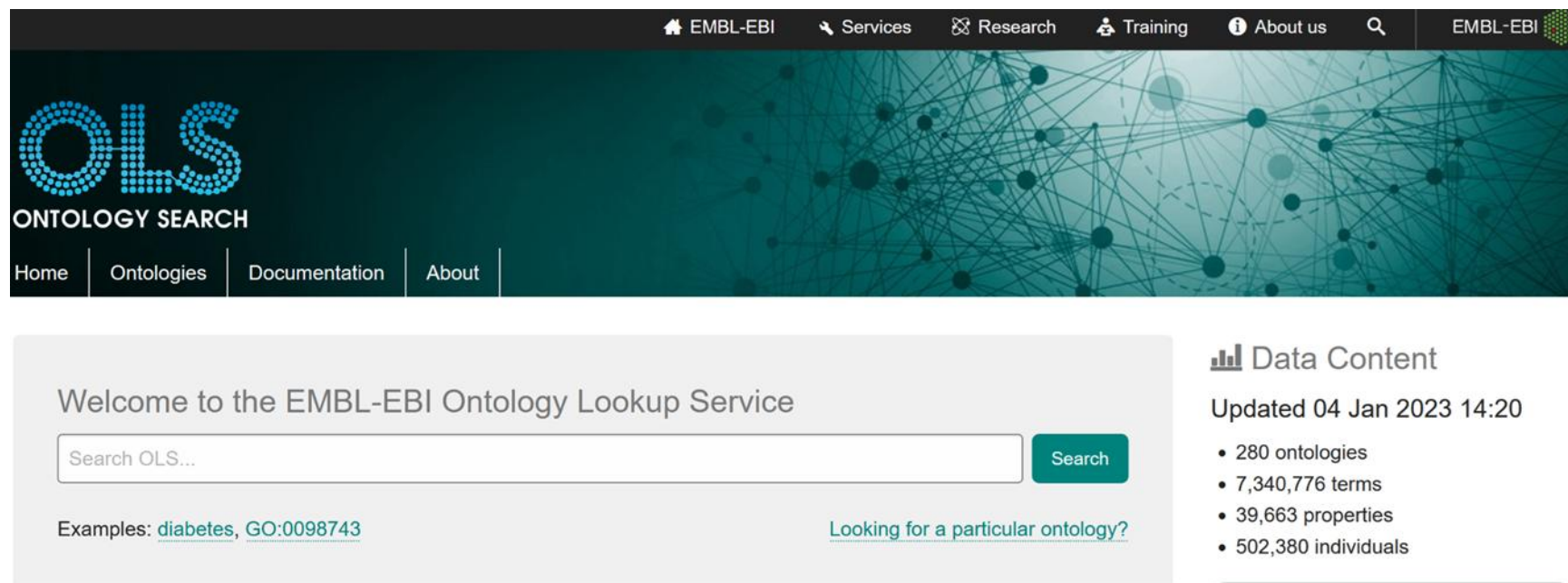
<https://github.com/sparontologies/datacite>

The *DataCite Ontology* (or simply *DataCite*) is an ontology written in OWL 2 DL to enable the metadata properties of the DataCite Metadata Schema version 3.1 to be described in RDF.

The main intent of the DataCite Ontology (shown in the following diagram) is to provide a flexible mechanism to define identifier for bibliographic resources (e.g., papers and datasets) and related entities (e.g., authors). To this end, we use the object property `datacite:hasIdentifier` that has as its object a member of the class `datacite:Identifier` or of one of its two sub-classes, `datacite:ResourceIdentifier` or `datacite:AgentIdentifier`, while the latter one is sub-classed by three additional classes, i.e., `datacite:PersonalIdentifier`, `datacite:OrganizationIdentifier`, and `datacite:FunderIdentifier`.



Ontology Lookup Service (OLS)



The screenshot shows the EMBL-EBI Ontology Lookup Service (OLS) homepage. At the top, there is a navigation bar with links for EMBL-EBI, Services, Research, Training, and About us. Below this, the OLS logo is displayed, followed by the text 'ONTOLOGY SEARCH'. A navigation menu includes links for Home, Ontologies, Documentation, and About. The main content area features a search bar with the placeholder text 'Search OLS...' and a 'Search' button. Below the search bar, examples are provided: 'diabetes' and 'GO:0098743'. A link 'Looking for a particular ontology?' is also present. On the right side, a 'Data Content' section shows the update date 'Updated 04 Jan 2023 14:20' and a list of statistics: 280 ontologies, 7,340,776 terms, 39,663 properties, and 502,380 individuals.

Welcome to the EMBL-EBI Ontology Lookup Service

Search OLS... **Search**

Examples: [diabetes](#), [GO:0098743](#) [Looking for a particular ontology?](#)

Data Content
Updated 04 Jan 2023 14:20

- 280 ontologies
- 7,340,776 terms
- 39,663 properties
- 502,380 individuals

Iskalnik po ontologijah na področju biomedicine

<https://www.ebi.ac.uk/ols/index>.



Primer ontologije: Gene Ontology

<https://www.geneontology.org/>

Ena najbolj znanih ontologij. Opisuje znanje na področju biologije iz treh aspektov:

- delovanja molekul,
- delov celic in
- bioloških procesov.

Definicijo ontologije si lahko prenesemo v dveh formatih:

- OWL (<https://purl.obolibrary.org/obo/go.owl>) – definicija formata: <https://www.w3.org/TR/owl-ref/>.
- OBO (<https://purl.obolibrary.org/obo/go.obo>) – definicija formata: <https://owcollab.github.io/oboformat/doc/obo-syntax.html>.

GENE ONTOLOGY
Unifying Biology

```
property_value: http://purl.org/dc/elements/1.1/description "The Gene Ontology (GO) provides a framework and set of concepts for describing the functions of gene products from all organisms." xsd:string
property_value: http://purl.org/dc/elements/1.1/title "Gene Ontology" xsd:string
property_value: http://purl.org/dc/terms/license http://creativecommons.org/licenses/by/4.0/
property_value: owl:versionInfo "2024-09-08" xsd:string

[Term]
id: GO:0000001
name: mitochondrion inheritance
namespace: biological_process
def: "The distribution of mitochondria, including the mitochondrial genome, into daughter cells after mitosis or meiosis, mediated by interactions between mitochondria and the cytoskeleton." [GOC:mcc, PMID:10873824, PMID:11389764]
synonym: "mitochondrial inheritance" EXACT []
is_a: GO:0048308 ! organelle inheritance
is_a: GO:0048311 ! mitochondrion distribution

[Term]
id: GO:0000002
name: mitochondrial genome maintenance
namespace: biological_process
def: "The maintenance of the structure and integrity of the mitochondrial genome; includes replication and segregation of the mitochondrial chromosome." [GOC:ai, GOC:vw]
is_a: GO:0007005 ! mitochondrion organization
```

biological process

Homo sapiens ▾ Examples Launch >

Hint: can use UniProt ID/AC, Gene Name, Gene Symbols, MOD IDs



Drugi primeri ontologij

- Elementary Multiperspective Material Ontology (EMMO)
<https://github.com/emmo-repo> skupina ontologij na področju materialov in tehnologij. Vsebuje ontologije kot so:
 - Crystallography Domain Ontology <https://github.com/emmo-repo/domain-crystallography>.
 - Battery Domain Ontology <https://github.com/emmo-repo/domain-battery>.
- The Environment Ontology <https://www.ebi.ac.uk/ols4/ontologies/envo>.
- Chemical Entities of Biological Interest (ChEBI) <https://www.ebi.ac.uk/chebi/>.
- Plant Ontology <https://www.ebi.ac.uk/ols4/ontologies/po>.



Druge ontologije

Biologija:

Accronym	Name	Link	Ontology describes ...
CLO	Cell Line Ontology		Standardize and integrate cell line information.
ChEBI	Chemical Entities of Biological Interest CropOntology		Classification of molecular entities of biological interest focusing on 'small' chemical compounds. Agronomic, morphological, physiological, quality, and stress traits
ERO	eagle-i resource ontology		Research resources such as instruments, protocols, reagents, animal models and biospecimens.
EDAM-BIOIMAGING	EDAM-BIOIMAGING		Bioimage analysis, bioimage informatics, and bioimaging (extension to EDAM)
EFO	Experimental Factor Ontology	https://www.ebi.ac.uk/efo/	Experimental variables, combines parts of several biological ontologies
FBbi	FBbi	https://www.ebi.ac.uk/ols/ontologies/FBbi	Biological Imaging Methods Ontology
GO	Gene Ontology NCBI taxonomy		Gene and gene product attributes Curated nomenclature for all of the organisms in the public sequence databases
PATO	Phenotype and Attribute Ontology		Phenotypic qualities (properties, attributes or characteristics)
PSO	Plant Stress Ontology		Biotic and abiotic stresses that a plant may encounter
PO	PlantOntology		
PROV	PROV Ontology schema.org		Data Provenance Structured data on the Internet
SBO	Systems Biology Ontology		Terms commonly used in Systems Biology, and in particular in computational modeling
BAO	The BioAssay Ontology		Biological assays
EFO	The Experimental Factor Ontology	https://www.ebi.ac.uk/efo/	Aspects of disease, anatomy, cell type, cell lines, chemical compounds and assay information.
IAO	The Information Artifact Ontology		Information entities
OBI	The Ontology for Biomedical Investigations		Life-science and clinical investigations.
EXACT	The ontology of experimental actions		Experimental actions
EXPO	The ontology of experiments		Scientific experiments.
UBERON	Uberon multi-species anatomy ontology		Cross-species anatomy, covering animals and bridging multiple species-specific ontologies.



Izvor podatkov ali provenienca

En najpomembnejših predpostavk v znanosti je, da so raziskave ponovljive. To velja za vse - od enostavne meritve temperature do zapletenih eksperimentov na področju fizike osnovnih delcev.

Pri tem se vsakemu raziskovalcu poraja vprašanje ali so in koliko so podatki zanesljivi. To še posebej velja za podatke, ki jih pridobimo iz drugih virov oziroma od drugih raziskovalcev – od kod so podatki, kako so bili pridobljeni in procesirani, na kakšen način so bile meritve kalibrirane in preverjene in kateri dodatni podatki so bili pri tem uporabljeni.

Vse naštetu s tujko imenujemo provenienca oziroma izvor podatkov (*)

*) Besedna zveza „izvor podatkov“ je precej široka in lahko v kakšnem drugem kontekstu pomeni tudi kaj drugega, tukaj pa pomeni metapodatke s katerimi opišemo izvor, kontekst in način pridobitve določenega seta raziskovalnih podatkov.



Zakaj je provenienca tako zelo pomembna

Poglejmo naslednjo tabelo:

1	2	5,3	30	-10
2	3	3,2	269	25
3	4	2,1	92	-80
3	5	2,9	16	2
5	6	4,6	350	-12

Kaj nam pove taka tabela?

Brez konteksta, podatkov o kolonah in merskih enot nam ne pove NIČ!

Vse kar lahko ugotovimo je, da so v prvih dveh kolonah naravna števila med 1 in 6, v tretji pozitivna decimalna števila manjša od 5,3, v četrti naravna števila manjša od 350 in v peti cela števila med -80 in +25.



Zakaj je provenienca tako zelo pomembna

Že če bi tabelo postavili v kontekst (da gre za meritve poligona v neki jami), bi lahko izkušen opazovalec uganil pomen stolpcev in merske enote:

Izhodišče	Cilj	Razdalja [m]	Azimut [°]	Naklon [°]
1	2	5,3	30	-10
2	3	3,2	269	25
3	4	2,1	92	-80
3	5	2,9	16	2
5	6	4,6	350	-12

Še vedno pa ne vemo niti od česa so te meritve, niti kdaj, kdo in s kakšnim inštrumentom jih je izmeril. In tukaj nastopi pomen provenience.



Provenienca in odprto objavljane

Provenienca je zelo pomembna pri odprtem objavljanju, saj zagotavlja, da lahko kdorkoli (z ustreznim znanjem) objavljene raziskovalne podatke razume in pozna njihov izvor, kar omogoča njihovo nadaljnjo uporabo.

Objavljeni raziskovalni podatki morajo biti čimbolj skladni s principom FAIR.

Če ponovimo: princip FAIR pomeni, da so podatki:

- F** findable - najdljivi
- A** accessible - dostopni
- I** Interoperable - interoperabilni
- R** Reusable - ponovno uporabni

Provenienca je ena od zahtev, da so podatki ponovno uporabni (**točka R**)



Kaj vse spada pod provenienco?

Znanstvene metode in načini pridobivanja podatkov se med različnimi znanstvenimi panogami močno razlikujejo, zato seveda ni enotnega recepta, kaj vse morajo vsebovati (meta)podatki o provenienci.

Vsekakor pa velja, da je potrebno navesti **vse podatke**, ki omogočajo **ponovitev** raziskave oziroma meritve **pod enakimi pogoji**, kot je to izvedel raziskovalec, ki je podatke pridobil. To omogoča tudi, da lahko ugotovimo, ali je možno **podatke** različnih raziskav **primerjati med seboj**.

V svetu je že sedaj na voljo ogromna količina raziskovalnih podatkov, ki pa so slabo dokumentirani in zato tudi precej neuporabni (brez posvetovanja z avtorji). Ustrezni podatki o provenienci lahko uporabnost raziskovalnih podatkov močno povečajo.



Kaj vse spada pod provenienco? 2

Med podatke o provenienci je potrebno vključiti (vsaj):

- **Kdo, kdaj in kje** je izvedel meritev/raziskavo/anketo/kalibracijo/obdelavo podatkov...
- Katera **metoda** je bila uporabljena (npr. masna spektroskopija, anketiranje...).
- S kakšnim **inštrumentom** (proizvajalec, tip, naziv) je bila meritev izvedena.
- Pod kakšnimi (zunanji) **pogoji** je bila izvedena meritev/raziskava (temperatura, tlak...).
- Kakšni **parametri/nastavitve/predpostavke** so bile uporabljene (npr. izbira anketirancev, izbira topila...).
- Kako je bil inštrument/metoda **kalibriran** (standardi, postopek, kdo, kje, kdaj...).
- Kako je bila pravilnost rezultatov **verificirana** (slepe probe, kontrolne skupine, ozadje...).
- Kakšna je **natančnost, točnost in občutljivost meritev**.
- Kako so bili podatki **obdelani** (algoritmi in parametri, programska oprema...).



Standardi za provenienco

Če je možno, provenienco dokumentiramo z uporabo katerega od standardov oziroma priporočil, kot so:

- W3C PROV,
- WGDC,
- DataCite,
- ECRIN,
- PCCS.



W3C Provenance Data Model

W3C Provenance Data Model ([PROV-DM](https://www.w3.org/2001/sw/wiki/PROV)) je del širšega standarda (<https://www.w3.org/2001/sw/wiki/PROV>), ki vključuje podatkovni model, ontologijo, prevedbo v Dublin Core...

Obravnava odnose med podatki, ustvarjalci in uporabniki („WasDerivedFrom“, WasAttributedTo“...).

Prednost: celotna infrastruktura (ontologija, iskanje...).

Slabost: precej komplicirano za razumevanje.

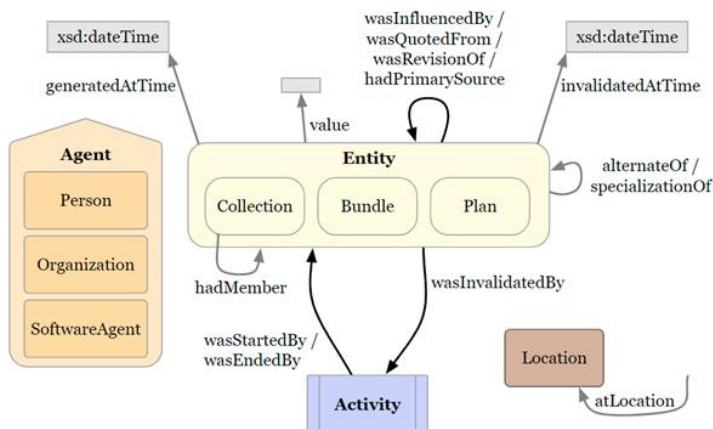
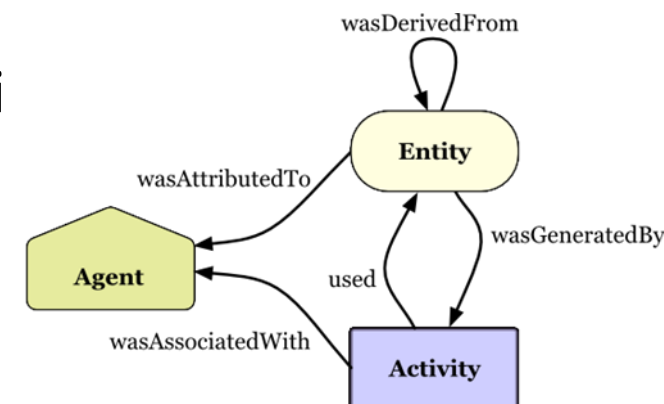


Table 1: Prefix and Namespaces used in this specification

prefix	namespace IRI	definition
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#	The RDF namespace [RDF-CONCEPTS]
xsd	http://www.w3.org/2000/10/XMLSchema#	XML Schema Namespace [XMLSCHEMA11-2]
owl	http://www.w3.org/2002/07/owl#	The OWL namespace [OWL2-OVERVIEW]
prov	http://www.w3.org/ns/prov#	The PROV namespace [PROV-DM]
(others)	(various)	All other namespace prefixes are used in examples only. In particular, IRIs starting with "http://example.com" represent some application-dependent IRI [IRI]



WGDC (Working Group on Data Citation)

WGDC je delovna skupina pri Research Data Alliance (RDA) <https://www.rd-alliance.org/rationale/data-citation-wg/>.

- Gre za priporočila, razdeljena v štiri kategorije:

- Priprava podatkov in podpore za iskanje
- Trajno identificiranje
- Razreševanje PIDov in pridobivanje podatkov
- Upravljanje s spremembami in migracijami

Priporočila so usmerjena tudi v podporo organizaciji podatkov, ki omogoča lažje iskanje.

Opis priporočil: https://rd-alliance.org/system/files/documents/RDA-DC-Recommendations_151020.pdf (ni dostopno).

Podrobno poročilo: [10.5281/zenodo.4048304](https://doi.org/10.5281/zenodo.4048304)

Goals of this WG are to create identification mechanisms that:

- allows us to identify and cite arbitrary views of data, from a single record to an entire data set in a precise, machine-actionable manner
- allows us to cite and retrieve that data as it existed at a certain point in time, whether the database is static or highly dynamic
- is stable across different technologies and technological changes

Solution: The WG recommends solving this challenge by:

- ensuring that data is stored in a versioned and timestamped manner.
- identifying data sets by storing and assigning persistent identifiers (PIDs) to timestamped queries that can be re-executed against the timestamped data store.



Drugi standardi za provenienco

DataCite (ponudnik DOI) uporablja PROV-DM model za sledenje sprememb v atributih DOI (<https://support.datacite.org/docs/tracking-provenance>)

ECRIN (javna, neprofitna organizacija, ki povezuje raziskovalne organizacije s področja medicine).

Dokument „Data Certification Standards“ pokriva vse od IT infrastrukture do trajne hrambe podatkov (<https://ecrin.org/sites/default/files/inline-files/Requirements%20for%20Certification%20of%20ECRIN%20Data%20Centres%2C%20with%20Explanation%20and%20Elaboration%20of%20Standards%2C%20Version%205.0.pdf>).



Standardi za provenienco

Primerjava med standardi:

Jauer in Deserno, Nemčija

Criteria	DataCite	ECRIN	WGDC	DQC
<i>Persistent identifier (PID)</i> : Each data object is assigned a unique, persistently stored identifier. Ideally, a DOI is assigned.	✓	✓	✓	✗
<i>Data origin</i> : The project or event that generated the data.	✓	✓	✗	✓
<i>Data creator</i> : A person or institution to be credited for.	✓	✓	✗	✓
<i>Data timestamp</i> : The time of dataset creation/modification.	✓	✓	✓	✗
<i>Data versioning</i> : Each transformation result of the data object is stored. Earlier versions are retrievable.	✓	✓	✓	✓
<i>Query PID</i> : If (sub-)sets of data are generated or cited, the query is stored with a persistent ID for reproducibility.	✗	✗	✓	✓

Vir: <https://ebooks.iospress.nl/pdf/doi/10.3233/SHTI200380>



Provenienca po standardu PCCS

Kot primer bomo pogledali kako podatke o provenienci opišemo po standardu **PCCS** ([The Provenance and Context Content Standard](#)).

- Standard je leta 1998 vzpostavil U.S. Global Change Research Program (USGCRP - <https://www.globalchange.gov/>) - združenje 13 ameriških zveznih agencij s področja raziskav sprememb na Zemlji.
- Standard uporablja tudi **NASA** v okviru svoje specifikacije za hranjenje podatkov na področju geo-znanosti (<https://earthdata.nasa.gov/standards/preservation-content-spec>)
- Več o tem: <https://eos.org/opinions/the-importance-of-data-set-provenance-for-science>



PCCS – pregled kategorij podatkov

Kategorija	Vsebina
Priprave na meritve	Opis in priprava inštrumentov za merjenje; podatki o kalibraciji
Nivo podatkov	Nivo 1: surovi podatki (neprocesirani podatki neposredno iz merilnega inštrumenta); Nivo 2: izpeljani podatki (npr. georeferencirani...); Nivo 3: normirani podatki (podatki normirani na enotno skalo...); Nivo 4: analizirani podatki (analizirani, klasificirani ali interpretirani podatki).
Dokumentacija o pridobljenem podatkovnem setu	Člani merilne ekipe; zahteve za pridobljene podatke; metapodatki o izvedenem zajemu podatkov (npr. čas snemanja, višina, datum...); zgodovina obdelave podatkov (kdo, kdaj in na kakšen način je iz surovih podatkov pridobil prečiščene/izvedene podatke različnih nivojev; uporabljeni algoritmi (kateri algoritmi so bili uporabljeni pri preračunavanju/prečiščevanju izvedenih podatkov; ocena kvalitete (na kakšen način je bila izvedene ocena kvalitete zajetih podatkov in kasnejših obdelav); reference in povratne informacije uporabnikov podatkov.



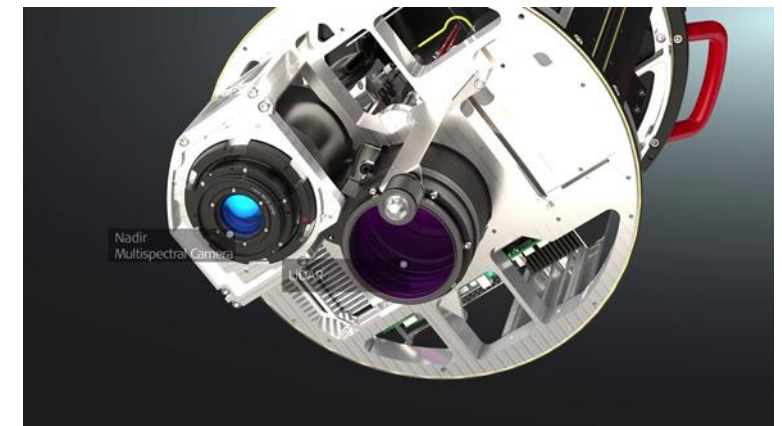
PCCS – podroben pregled standarda

Item Number	Category	Content Name	Definition/ Description	Rationale (Why content is needed)	Criteria (How good the content should be)	Priority	Source	Project Phase for Capture
1,1	Preflight/Pre-operational	Instrument Description	Documentation of Instrument/sensor characteristics including pre-flight or pre-operational performance measurements (e.g., spectral response, instrument geometric calibration (geo-location offsets), noise characteristics, etc.). Components include: instrument specifications, vendor calibration reports, user guides, spectral and radiometric calibration.	Needed for users to understand how instrument operates. Documentation of measurements made before deploying instruments in space (or in situ) will help establish a baseline and help users understand changes that may have occurred over time while in operation.	At a level of detail sufficient for reviewers to verify that the instrument is capable of supplying input data of sufficient quality. Researchers should be able to determine when significantly better knowledge of atmospheric transmittance or improved computational methods could be used to improve calibration. Detailed criteria should be	H	Instrument teams and vendors	Preflight/pre-operational
1,2	Preflight/Pre-operational	Preflight/Pre-operational calibration data	Numeric (digital) files of Instrument/sensor characteristics including pre-flight or pre-operational performance measurements (e.g., spectral response, instrument geometric calibration (geo-location offsets), noise characteristics, etc.). Components include: instrument specifications, vendor calibration reports, user guides, spectral and radiometric calibration.	Measurements made before deploying instruments in space (or in situ) will help establish a baseline and help users understand changes that may have occurred over time while in operation.	Sufficient to maintain knowledge of the geophysical reference and sources of errors for comparing calibrated values to other long term observational data sets. Detailed criteria for Algorithm Theoretical Basis Documents (ATBD).	H	Instrument developer	Pre-mission (during early part of mission, at the latest; certainly before development contracts expire)

Praktičen primer - LIDAR

LIDAR (Light Detection And Ranging) je metoda daljinskega zaznavanja s pomočjo laserskega merjenja razdalj, ki se uporablja za mapiranje zemeljskega površja, arheoloških najdišč, arhitekturnih spomenikov ipd.

Metoda deluje na osnovi laserskega merjenja oddaljenosti od merilne naprave do objekta. Enostavne laserske merilnike razdalje se uporablja v gradbeništvu, LIDARski instrumenti pa so nameščeni na stojalu (za skeniranje manjših objektov) ali na letalu (za skeniranje zemeljskega površja).

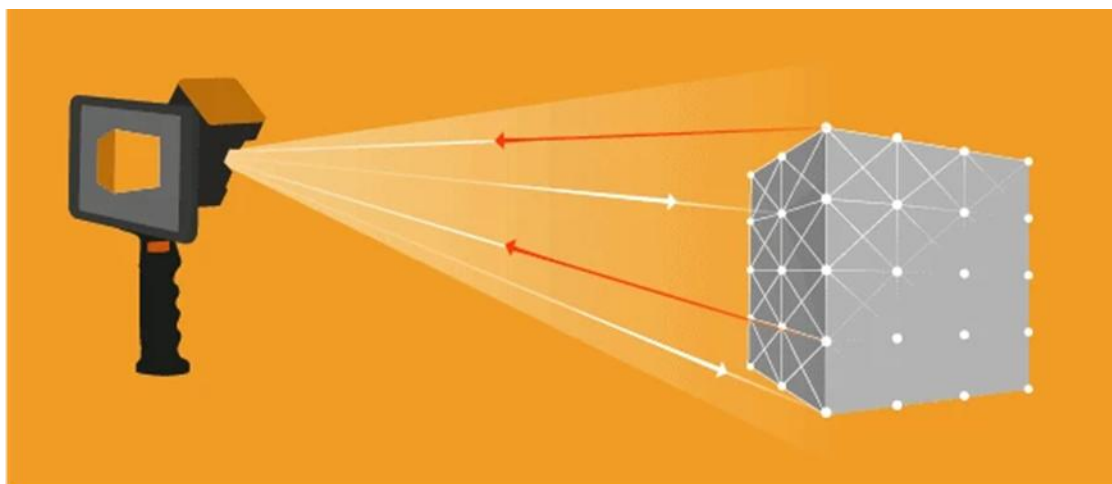




Praktičen primer – LIDAR 2

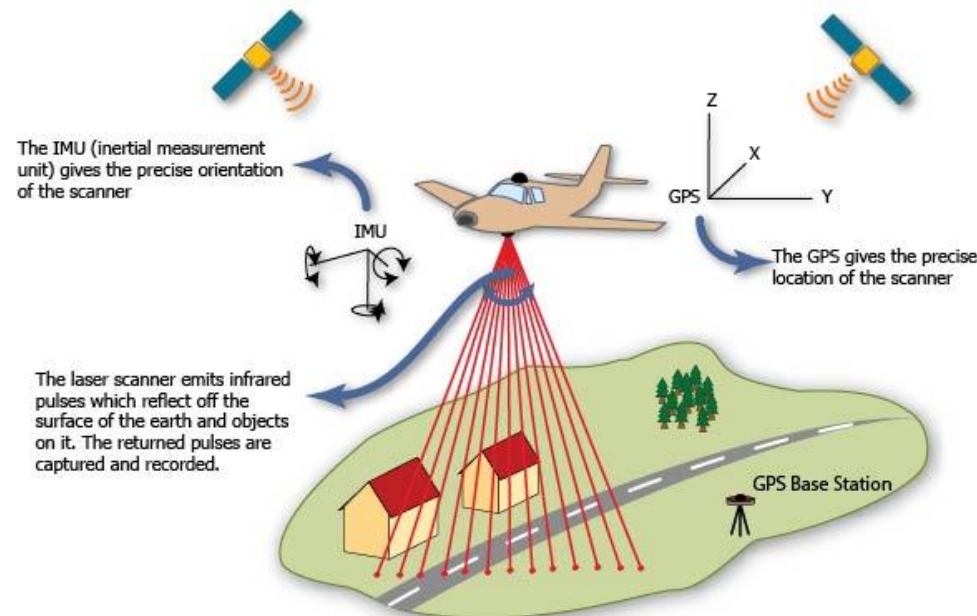
Primer LIDARskih podatkov je zanimiv, ker lahko zelo veliko stvari vpliva na rezultat in ker podatki od surovih preidejo več faz obdelave, za kar vse mora biti na voljo dobra dokumentacija (=provenienca).

Pri terestričnem LIDARju naprava z laserskim žarkom avtomatsko skenira teren, tako da spreminja kot (tako horizontalni, kot vertikalni) žarka in pri vsakem položaju posname oddaljenost (kar naredi večkrat na sekundo).



Praktičen primer – LIDAR 3

- Pri snemanju zemeljskega površja je naprava nameščena na letalu, ki se premika, laserski žarek pa se odklanja pravokotno na smer letenja. Pri tem je bistveno, da vsak trenutek na milimetre točno poznamo lokacijo, za kar poskrbi zelo natančen GPS, ki meri v diferencialnem načinu.





Praktičen primer – LIDAR 4

Kaj torej dejansko merimo, kaj vse vplivana meritev, kaj moramo o meritvi vedeti?

- Merimo položaj inštrumenta v prostoru (dejansko merimo čas, ki ga radijski signal potrebuje od GPS satelitov do GPS sprejemnika) – vplivi: temperatura in tlak ozračja, položaj satelitov, položaj referenčnega GPS...
- Merimo orientacijo inštrumenta v prostoru (dejansko merimo majne spremembe v silah na žiroskopsko vezje iz katerih izračunamo kotne premike) – vplivi: gravitacija (višina), temperatura, začetni položaj... - kalibrirati je treba vsaj začetno orientacijo
- Merimo kot laserskega žarka (to je opto-mehanska meritev položaja laserja v napravi) – vplivi: mehanski, temperatura merilne naprave...
- Merimo oddaljenost objekta (dejansko merimo čas, potreben za prelet laserskega žarka do objekta in nazaj) – vplivi: temperatura in tlak ozračja, lažni odboji...



Praktičen primer – LIDAR 5

Kaj dobimo kot surove podatke (**nivo 1**)?

- Položaj in orientacijo inštrumenta v času vsake meritve, kot laserja in oddaljenost do površja.
- Vsi ti podatki so posneti z določeno stopnjo točnosti in natančnosti, ki vpliva na končni rezultat. Brez obdelave so ti podatki neuporabni.

V prvi fazi obdelave iz položaja, kotov in oddaljenosti izračunamo položaj objekta (v bistvu položaj odboja) v 3-D prostoru. Nato te koordinate še pretvorimo v geografske koordinate (upoštevamo, da je model zemlje v obliki elipsoida) in po potrebi v pravokotne koordinate v skladu z izbrano projekcijo.

Dobimo georeferenciran oblak točk s koordinatami, ki so že uporabne za nadaljnje izračune (**nivo 2**), pri čemer moramo vedeti, na kakšen način (npr. izbrani elipsoid, projekcija) so bili obdelani.



Praktičen primer – LIDAR 6

Naslednja faza v obdelavi podatkov je klasifikacija. Po posebnem postopku (ki bi moral biti prav tako opisan) se za vsako točko določi ali gre za tla, vegetacijo, stavbe ali kaj drugega. Tako dobimo klasificirane podatke (**nivo 3**).

Iz klasificiranih podatkov pa lahko izračunamo marsikaj – npr. digitalni model reliefa, sliko senčenja terena, območja poplav, identifikacija arheoloških najdišč... (**nivo 4**). Seveda spet potrebujemo dokumentacijo o algoritmih in programski opremi, ki so bili uporabljeni za izdelavo končnega produkta.

Če na primer pri kateri koli od opisanih faz uporabimo drugačno metodo (npr. drugačno projekcijo) lahko dobimo tako drugačne končne podatke, da jih brez poznavanja postopkov obdelave podatkov ni mogoče primerjati med seboj.

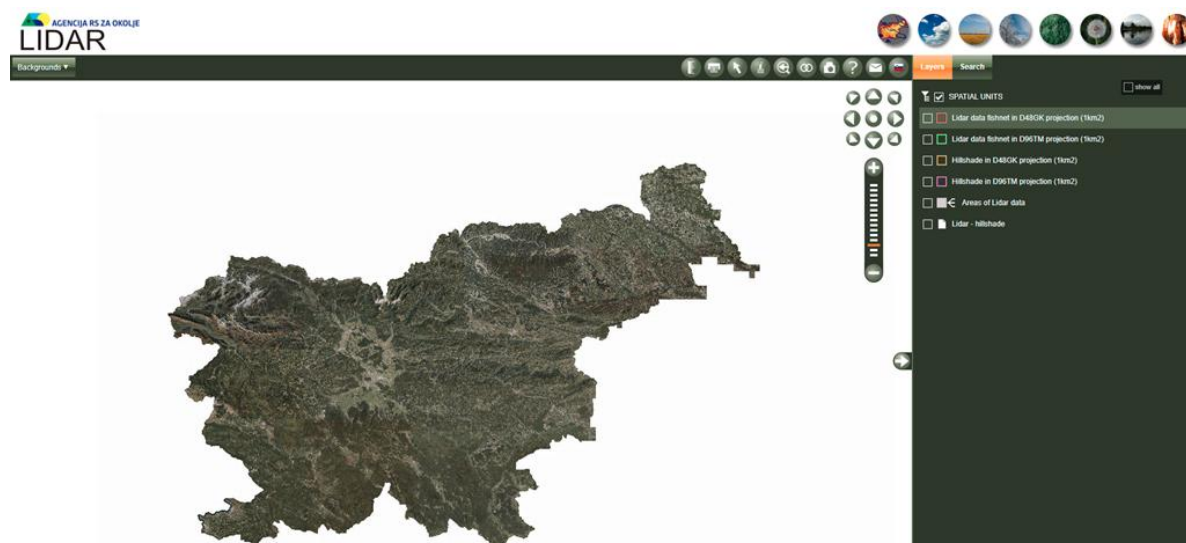


Praktičen primer – LIDAR 7

LIDARski podatki snemanja površja Slovenije ne spadajo med raziskovalne podatke, so pa javno dostopni. Na njihovi osnovi lahko izvajamo različne izračune in raziskave, kjer potrebujemo podatke o njihovi provenienci.

Podatki so na voljo na naslovu:

https://gis.arso.gov.si/evode/profile.aspx?id=atlas_voda_Lidar%40Arso&initialExtent=402591.76%2C39904.09%2C2.64583



Centralna tehniška knjižnica Univerze v Ljubljani
Central Technical Library at the University of Ljubljana
Trg republike 3, SI-1000 Ljubljana
Slovenija / Slovenia



Zahvala:

Goran Dražić in podatkovni skrbniki KI,
Urša Krašovec Opara, Mojca Kotar, UL,
Brane Leskošek, Marko Vidak, Bojan Kverh in Polonca Ferk, IBMI, ELIXIR-SI, MF, UL,
Barbara Malič, Jan Jona Javoršek, Andrej Filipčič in Boštjan Kokot, IJS,
Maja Zagorščak in Carissa Bleker, NIB,
Zavod za gradbeništvo.



Univerza v Ljubljani

