

Vrste formata za čuvanje podataka:

Šta je potrebno znati prilikom kreiranja planova upravljanja istraživačkim podacima?

Različite vrste podataka se prikupljaju, obrađuju i čuvaju (čuvaju i/ili skladište) na različite načine. Kada započinjete neki projekat i kreirate plan upravljanja istraživačkim podacima (eng. *Data Management Plan*, skraćeno DMP) već unapred treba odlučiti koji format datotekaće se koristiti. Mnogi vlasnički formati datoteka su tzv. kontejneri za standardne formate datoteka. Developeri smeštanjem podataka u ove kontejnere pojednostavljuje proces pružanjem dodatnih funkcija za analizu podataka stečenih na njihovoj platformi. Međutim, ovo ima negativnu posledicu jer ove podatke čini manje interoperabilnima.

Štaviše formati datoteka mogu biti sa ili bez gubitaka tj. komprimovani (JPEG za slike) kako bi se uklonili suvišni podaci i na taj način smanjila datoteka ili nekomprimovani (TIFF za slike). Obično se rade analize o gubicima podataka, ali to nužno ne znači da su to oni podaci koje treba dugoročno čuvati. U većini slučajeva, potrebno je prvi dokument, onaj koji je prvobitno snimljen instrumentom, dugoročno čuvati ili pak njegovu direktnu verziju standardnog formata datoteke bez gubitaka (pogledajte [Vodič o sirovim/neobrađenim podacima](#)).

Za H2020 projekte, uključivanje plana upravljanja podacima jeste podrazumevani uslov koji je uvela Evropska komisija:

- „Program H2020: Smernice za FAIR upravljanje podacima Horizon 2020“, https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

ZAŠTO JE OVO NEOPHODNO?

Podaci moraju da zadovolje FAIR principe, da ih odlikuje dugovečnost i interoperabilnost, a moraju se uzeti u obzir i sledeći faktori – vlasnički nasrpan standardnih formata: korišćena oprema, vrsta softvera/hardvera kao i metoda istraživanja koja je primenjena. Ne postoji garancija da će u budućnosti postojeći vlasnički formati biti dostupni. Na primer, mnogi Microsoft formati koji su danas u upotrebi kao što su Word i Excel, ne isključuju mogućnost da će nakon određenog perioda predstavljati zastarele formate. Ovo će se posebno odraziti na pojedinačne formate datoteka koji su stvoreni kao deo individualnih projekta.

KOJE JE REŠENJE PROBLEMA?

U oblasti biomedicinske slike usled postojanja velikog broja formata datoteka javila se inicijativa da se ovi formati učine interoperabilnim. Ovaj problem se rešava pomoću alata kao što je Biogormats, koji je deo OMERO projekta tako što Biogormats- programski dodatak omogućava

konverzuju vlasničkih i standardnih datoteka. Slobodan i softver otvorenog koda za analizu slika kao što je ImageJ usvojio je ovaj alat i na taj način je omogućeno korisnicima da manipulišu biomedicinskim slikama bez obzira na njihovo poreklo. Ovo ne znači da svaka disciplina poseduje ovakvo rešenje, a sami istraživači treba da čuvaju stečene podatke u standardnim formatima, kao i da razmotre upotrebu standardnih formata u planu upravljanja podacima što je pre moguće. Mnogi (većina?) komercijalnih i softverskih paketa otvorenog koda omogućavaju konverziju podataka u standardne formate i ovo treba iskoristiti.

Tokom digitalne revolucije, određen broj formata je prepoznat i preporučen jer omogućava dugotrajno čuvanje i interoperabilnost.

U nastavku se nalaze neki od izvora formata podataka za dugoročno čuvanje:

- 4TU Center for Research Data - “data description and formats”
-<https://researchdata.4tu.nl/en/publishing-research/data-description-and-formats/>
- DANS | Data Archiving and Networked Services - “file formats”
-<https://dans.knaw.nl/en/deposit/information-about-depositing-data/beforedepositing/file-formats>

Kao primer, sledeća tabela opisuje razne formate datoteka za različite discipline koji su ili preporučeni ili odabrani kao prihvatljivi (sistem podataka Velike Britanije, UK Data Service):

Vrsta podataka	Preporučeni formati	Odobreni formati
Tabelarni podaci sa obimnim metapodacima Oznake promenljivih, kodovi i vrednosti koje nedostaju	SPSS portable format (.por) tekstualne datoteke sa podacima ili programima (SPSS, Stata, SAS, etc.) strukturirani tekst ili datoteka oznaka sa informacijama o metapodacima, npr. DDI XML dokument	vlasnički formati statističkih paketa: SPSS (.sav), Stata (.dta), MS Access (.mdb/.accdb)
Tabelarni podaci sa minimalnim metapodacima Naslovi kolona, naziv promenljivih	datoteke tipa .csv ili .tab razdvojen tekst sa SQL izrazima definicije podataka	razdvojen tekst (.txt) sa znakovima koji nisu prisutni u podacima koji se koriste kao graničnici široko primenjeni formati: MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb), dBase (.dbf), OpenDocument Spreadsheet (.ods)

Geoprostorni podaci vektorski i rasterski podaci	ESRI Shapefile (.shp, .shx, .dbf, .prj, .sbx,.sbn optional) Geo-referencirani TIFF (.tif, .tiff) CAD podaci (.dwg) tabelarni podaci o GIS atributima Geography Markup Language (.gml)	ESRI Geodatabase format (.mdb) MapInfo Interchange Format (.mif) Za vektorske podatke Keyhole Mark-up Language (.kml) Adobe Illustrator (.ai), CAD podaci (.dxf or .svg) binarni formati GIS i CAD paketa
Tekstualni podaci	Rich Text Format (.rtf) plain text, ASCII (.txt) eXtensible Mark-up Language (.xml) tekst u skladu sa odgovarajućim Document Type Definition (DTD) ili šemom	Hypertext Mark-up Language (.html) Široko primenjen format: MS Word (.doc/.docx) Specifični softverski podaci: NUD*IST, NVivo and ATLAS.ti
Slikovni podaci	TIFF 6.0 nekomprimovan (.tif)	PEG (.jpeg, .jpg, .jp2) ako je original kreiran u ovom formatu GIF (.gif) TIFF druge verzije (.tif, .tiff) RAW slikovni format (.raw) Fotošoppodaci (.psd) BMP (.bmp) PNG (.png) Adobe Portable Document Format (PDF/A, PDF) (.pdf)

Audio podaci	Free Lossless Audio Codec (FLAC) (.flac)	MPEG-1 Audio Layer 3 (.mp3) ako je original kreiranu ovom formatu Audio Interchange File Format (.aif) Waveform Audio Format (.wav)
Video podaci	MPEG-4 (.mp4) OGG video (.ogv, .ogg) motion JPEG 2000 (.mj2)	AVCHD video (.avchd)
Dokumentacija i skripte	Rich Text Format (.rtf) PDF/UA, PDF/A or PDF (.pdf) XHTML or HTML (.xhtml, .htm) OpenDocument Text (.odt)	Običan tekst (.txt) široko korišćeni formati: MS Word (.doc/.docx), MS Excel (.xls/.xlsx) XML označen tekst (.xml) Prema odgovarajućem DTD Ili šema, npr. XHTML 1.0

Savetuje se da istraživači prilikom kreiranja DMP-a konsultuju ovakvu analizu formata, kako bi napravili najbolju moguću odluku pri odabiru formata datoteka i kako bi to jasno naveli.

Video sa Utreht Univerziteta na temu kako smanjiti rizik od gubitka podataka i kako sačuvati podatke na optimalan način: “Preserving research data in the optimal, technically correct way”
<https://www.youtube.com/watch?v=qENaO0Lk6eo#action=share>

LITERATURA KOJA JE KORIŠĆENA ZA SASTAVLJANJE OVOG VODIČA

Publikacije

Scaled and automated preservation planning for highly diverse digital collections: the Integrated Preservation Suite.

DOI = <https://dx.doi.org/10.7207/twr14-02>

https://zenodo.org/record/1299966/preview/ipres_ips_revised.pdf

Science Europe Guidance Document: Presenting a Framework for Discipline-specific Research Data Management (January 2018)

https://www.scienceeurope.org/wp-content/uploads/2018/01/SE_Guidance_Document_RDMPs.pdf

The FAIR Guiding Principles for scientific data management and stewardship | US National Library of Medicine National Institutes of Health

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792175/>

Drugi izvori

Best practice for file formats (Stanford Libraries)

<https://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-formats>

Consortium of European Social Sciences Data Archives (CESSDA)

Adapt your Data Management Plan: a list of Data management questions based on the Expert Tour Guide on Data management

<https://www.cessda.eu/content/download/3844/35033/file/20171117DMPQuestionsCESSDAExpertTourGuide.pdf>

Consortium of European Social Sciences Data Archives (CESSDA)

Expert Tour Guide on Data Management (on section 3 - Process | File formats and data conversion

<https://www.cessda.eu/Research-Infrastructure/Training/Expert-Tour-Guide-on-Data-Management>

Data management knowledge, tools, and training | DTL - Dutch Techcenter for Life Sciences

<https://www.dtls.nl/fair-data/research-data-management/data-management-knowledge-tools/>

Digital Preservation Coalition - Digital Preservation Handbook

<https://www.dpconline.org/handbook>

FAIR data: what it means, how we achieve it, and the role of RDA (presentation from Sarah Jones)

<https://pt.slideshare.net/sjDCC/fair-data-what-it-means-how-we-achieve-it-and-the-role-of-rda>

Storing and Preserving Data (Utrecht University)

<https://www.uu.nl/en/research/research-data-management/guides/storing-and-preserving-data>

The National Archives:

<http://www.nationalarchives.gov.uk/information-management/manage-information/digital-records-transfer/file-formats-transfer/>

UK Data Service guidelines:

<https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats>

University of Vienna (format informations and supported formats)

<https://datamanagement.univie.ac.at/en/about-phaidra/formats/>

University of Virginia Library:

<https://data.library.virginia.edu/data-management/plan/format-types/>

Prevela dokument VENKATARAMAN, Shanmugasundaram, & Moura, Paula. (2020, September 21). Data formats for preservation: What you need to know when creating a DMP. Zenodo. CC licence, <http://doi.org/10.5281/zenodo.4041512> sa engleskog na srpski Jovana Levkić u sklopu volontiranja za Serbia.RDA grupu u Beogradu aprila 2021.