

Neobrađeni podaci, rezervne kopije i verzije

Šta je potrebno da znate prilikom čuvanja istraživačkih podataka?

Neobrađeni/sirovi podaci (eng. *raw data*): osnovne pretpostavke

Neobrađeni podaci predstavljaju originalne podatke koji su prikupljeni iz nekog izvora, ali koji nisu obrađeni ili analizirani. Oni čine osnovu za dalju analizu. U većini slučajeva usnimljeni ili prikupljeni podaci su jedinstveni i njih je nemoguće reprodukovati. Primeri takvih podataka su vremenska prognoza u određenim vremenskim intervalima ili intervjui. Kao takvi oni moraju biti zaštićeni od bilo kakvih gubitaka. Prema tome, neobrađene podatke treba čuvati u datotekama koje su nekomprimovane, poput TIFF formata jer su bezbedne od gubitaka, za razliku od JPG formata datoteka koje su komprimovane i nisu bezbedne od gubitaka. U nekim slučajevima, neobrađeni podaci mogu sadržati dodatne informacije koje su specifično povezane sa vrstom instrumenata koji je korišćen za prikupljanje podataka. Na primer, Leica mikroskopi poseduju zaštićeni vlasnički format podataka ali, poseduju i kontejnere za nekomprimovane podatke koji sadrže metapodatke specifične za ove mikroskope koji omogućuju čitanje, pisanje i analiziranje putem Leica softvera.

Za više o formatima, podataka pogledajte [Vodič o formatima](#).

Kada se govori o obrađenim podacima misli se na podatke koji su već podvrgnuti nekoj vrsti manipulacije. Na primer, to su podaci koji su digitalizovani, komprimovani, prevedeni, transkribovani, očišćeni, potvrđeni, provereni i/ili anonimizirani.

Analizirani podaci su podaci koji su već obrađeni, interpretirani i analizirani. Oni mogu sadržati prikaz podataka u vidu tabela, teksta, ili grafika, sa ciljem lakšeg i boljeg razumevanja podataka.

U većini slučajeva neobrađeni podaci se mogu smatrati i zvaničnim podacima, odnosno glavne/master kopije bilo kog zapisa (pogledati tekst o [zlatnoj kopiji](#)). Dakle, osim što predstavljaju polaznu tačku za dalju analizu, oni mogu biti korišćeni i u druge svrhe. Iz ovih razloga, u tipičnoj proceduri rada tj. toku analize, preporučuje se kreiranje kopije neobrađenih podataka koja bi dalje bila korišćena kao radna kopija. Originalne podatke zatim treba arhivirati na odgovarajući način duži vremenski period. Radna kopija se onda može koristiti za obradu i analizu bez opasnosti da dođe do slučajnog brisanja ili presnimavanja originalnih fajlova.

Za više informacija o formatima podataka, pogledajte sledeći [Vodič](#).

Bekap ili kreiranje rezervne kopije

Izrada rezervnih kopija (eng. *backup*) ne podrazumeva isto što i čuvanje i/ili skladištenje. Kada obrada podataka dostigne konačnu fazu, čuvanje takvih podataka skladištenjem u npr. repozitorijume omogućava njihovu laku dostupnost. Ukoliko postoje bilo kakvi podaci, bilo da

su oni obrađeni ili ne ili ako su to finalni podaci, koji se smatraju dovoljno važnim i potrebno ih je čuvati na bezbedan način tokom dužeg vremenskog perioda, onda njih treba arhivirati. Ovi podaci mogu biti preuzeti, ali ne tako jednostavno kao što su isti i sačuvani.

Skladištenje podataka se vrši zbog: dalje analize i istraživanja, potencijalne vrednosti koje podaci mogu imati njihovom ponovnom upotrebom, zbog nacionalnog/internacionalnog statusa i kvaliteta, njihove originalnosti i/ili jedinstvenosti, zbog nastajanja troškova prilikom formiranja novih podataka, zbog njihovog značaja za (naučnu) istoriju kao i zbog važnosti upotrebe podataka u neakademske svrhe kao što je kulturno nasleđe ili na zahtev finansijera. Sa druge strane, izrada rezervnih kopija se vrši radi sprečavanja gubitka podataka tokom njihove analize. Skladištenjem i pravljenjem rezervnih kopija se bave istraživači tokom svog aktivnog rada sa podacima. Preporučuje se da se za podatke od značaja naprave najmanje tri kopije i da se pritom koriste najmanje dva medija za skladištenje kao i da najmanje jedan medij bude na različitoj lokaciji. Takođe, preporučuje se da se prilikom rada koriste automatske rezervne kopije ukoliko digitalni servisi institucija to poseduju. Komercijalni i nekomercijalni načini skladištenja zasnovani na klad tehnologiji kao što je Dropbox koji iako je popularan način ne daju garanciju o njihovom postojanju i one takođe pokreću pitanje vlasništva.

U narednom tekstu izlistane su stavke koje mogu pomoći istraživačima za definisanje strategije za izradu rezervnih kopija podataka:

1. Najpre potrebno je proveriti da li je institucija sprovela strategiju izrade rezervnih podataka i da li ona ispunjava potrebe projekta. Poželjno je uspostaviti kontakt sa IT sektorom ako postoje neke nedoumice i kako bi istraživači dobili direktan savet. Veza sa IT sektorom je značajna, jer i oni na taj način dobijaju saznanje kako istraživači njihovih institucija koriste njihove usluge. Ukoliko data institucija ne poseduje gore pomenutu strategiju onda je potrebno da:
2. Utvrdite da li postoje alati nezavisnih proizvođača koji se mogu koristiti za automatsku izradu rezervnih podataka. Ovakvim pristupom postiže se bolja garancija redovne izrade kopija i njihovo čuvanje pravom mestu. Na ovaj način se umanjuje greška koja može nastati ljudskim radom. [Microsoft](#) i [Apple](#) imaju softverske alate koji podržavaju automatsku izradu rezervni kopija. Postoji i metoda skladištenja na tzv. oblaku kao što su [CloudStor](#), [Figshare](#), [SpiderOak](#) i [StoneFly](#). Svakako je preporučljivo da provera izrade rezervnih kopija bude deo rutine prilikom rada istraživača sa podacima.
3. Potrebno je utvrditi vrste izrade rezervne kopije:
 - a. Potpuna izrada rezervne kopije sistema i datoteka;
 - b. Diferencijalna rezervna kopija, gde se beleži sve što je izmenjeno nakon poslednje potpune izrade rezervne kopije. Ukoliko je potreban oporavak podataka potrebne su poslednja potpuna, ali i poslednja diferencijalna kopija;
 - c. Postepena izrada rezervne kopije, koja predstavlja čuvanje poslednjih izmena od poslednje napravljene rezervne kopije. Kako bise povratili podaci i/ili sistem

potrebna je potpuna poslednja rezervna kopija i čitava serija postepenih tj. inkrementalnih kopija.

- d. Postepene i diferencijalne rezervne kopije ubrzavaju proces izrade jer one ažuriraju izmenjene podatke štedeći prostor i vremena disku prilikom svakodnevnog pokretanja potpune izrade rezervne kopije.
4. Potrebno je planirati koliko često i koliko rezervnih kopija je potrebno napraviti. Preporuka je da se naprave tri rezervne kopije kako bi se umanjila mogućnost gubitka podataka, čak i ako je jedna od rezervnih kopija oštećena ili izgubljena. Ukoliko je veličina datoteka velika ili ukoliko se upravlja osetljivim podacima, tada se preporučuje rad sa manje kopija. U tom slučaju, potrebno je konsultovati se sa institucijom ili finansijerom.
5. Potrebno je definisati gde će se rezervne kopije čuvati. Preporučuje se da to budu fizički različite lokacije. Pravljenje rezervnih kopija može biti na mrežnim upravljačkim programima, skladišta u tzv. oblaku ili lokalnim i prenosnim uređajima. Koju će se opcija izabrati zavisi od količine podataka, njihove učestanosti, nivoa automatizacije i njihove osetljivosti.
6. Potrebno je predvideti količinu prostora koja je potrebna za izradu rezervnih kopija podataka i njihove dokumentacije, a zatim je potrebna i procena kapaciteta na određenom mediju.
7. Potrebno je odrediti koliko dugo je potrebno čuvati određene kopije i kako će one biti obrisane. Nije preporučljivo zameniti jednu rezervnu kopiju drugom. Ukoliko postoji potreba za rezervnim kopijama velikog broja podataka i to često, verovatno neće biti moguće zadržati sve rezervne kopije tokom trajanja datog projekta. Za osetljive podatke, morate biti sigurni da su podaci uspešno obrisani i da se isti ne mogu povratiti na bilo koji način.
8. Potrebno je uveriti se da će lični podaci biti zaštićeni i da li su rezervne kopije koje sadrže lične podatke zaštićene od neovlašćenog pristupa.
9. Potrebno je napraviti plan oporavka u slučaju da dođe do gubitka podataka. On treba da sadrži instrukcije za period pre i nakon gubitka podataka kao i kontakt osobe odgovorne za pružanje podrške u ovom slučaju.
10. Neophodno je da se odgovornosti u potpunosti identifikuju u pogledu performansi ručnih sigurnosnih kopija i verifikacije automatskih rezervnih kopija, kao i performansi testova za oporavak podataka i obnavljanje izgubljenih podataka.

Verzisionisanje

Prilikom analiziranja podataka, dolazi do kreiranja derivata radnih kopija, ponekad čak i automatski putem skripti. Od vlasnika podataka zavisi da li je potrebno dugoročno sačuvati neku od ovih verzija. Preporučuje se da se verzije često nadgledaju kako bi se odbacile one kojima verifikacija ili transparentnost nije potrebna. Verzije mogu podrazumevati velike datoteke koje zauzimaju vredan prostor. Potrebno je uvesti odgovarajuće mere kako bi se kreirale verzije koje su ključan deo svakog procesa, bilo da je reč samo o promeni imena datoteka ili korišćenje

namenskih alata za njihovu kontrolu. Kontrola omogućava da se kod velikih projekata prijavi i odjavi više korisnika koristeći isti dokument nakon izvršenih promena. Ovo takođe omogućava čuvanje podataka o poreklu, odnosno dokumentovanje ili pregled istorije promena. Baš kao i kod rezervnih kopija, prvi korak je saznati šta pruža organizacija.

Literatura

<https://researchdata.nl/en/services/data-management/selecting-research-data/>

<https://blog.zenodo.org/2017/05/30/doi-versioning-launched/>

<http://help.zenodo.org/#versioning>

<https://www.ru.nl/rdm/archiving-data/what-data-should-archived-minimum/>

<https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/4.-Store/Backup>

Prevela dokument VENKATARAMAN, Shanmugasundaram, & Moura, Paula. (2020, September 21). Raw data, backup and versioning: What you need to know to preserve your research data. Zenodo. CC licence, <http://doi.org/10.5281/zenodo.4041557> sa engleskog na srpski Jovana Levkić u sklopu volontiranja za Serbia.RDA grupu u Beogradu aprila 2021.