Datenerfassung und Speicherung von Zwischenergebnissen

Das Ziel einer FAIRen Datenerfassung ist es, reproduzierbare und damit vertrauenswürdige Daten zu schaffen. Schon während der Datenerfassung ist es wichtig zu bedenken, wie gut sie von Personen verstanden werden können, welche die Daten nachnutzen. Die Form, in der Daten wohl am besten nachgenutzt und analysiert werden können sowie in Repositorien gespeichert werden ist als so genannte „tidy data“, [1)] so dass Spalten mit Variablen und Zeilen mit einzelnen Beobachtungen korrespondieren.

Metadaten, also die Beschreibung der erfassten Daten (genauer siehe Abschnitt 4) spielen im Stadium der Datenerfassung daher bereits eine wichtige Rolle. Durchweg sollten beschreibende Fakten wie Datum, Messgerätehersteller und –modell, sowie Umgebungsbedingungen und Besonderheiten während der Datenerfassung dokumentiert werden. Dazu gehören auch vorgenommene Umrechnungen der Daten, Transfers in weitere Programme und Formate sowie die Handhabung von Ausreißerwerten und Korrektur von Daten.[2)]

Bei der Formatwahl ist zu bedenken, wie zugänglich diese bei der Weiter- und Nachnutzung der Daten sind. Offene, textbasierte und häufig genutze Formate wie .csv, .txt, .html, .xml, .tiff, .mkv, mp4 oder PDF/A sind gut geeignet und voraussichtlich auch nach langer Zeit noch les- und verwendbar.[3)] Vor allem bei Langzeitmessungen ist dies essenziell. Bei offener Software, die jedoch ungebräuchliche Formate verwendet, ist die Archivierung der Software selbst eine Option.

Backup von Work-in-Progress: Daten unterlaufen im Forschungsprozess kontinuierlichen Veränderungen. Das gilt für berechnete Daten stärker als für Rohdaten. Ein regelmäßiges Backup des aktuellen Standes ist daher wichtig. Im Gegensatz zu einer Synchronisation von Daten, bei der eine Änderung eines Mediums die gleiche Änderung auf dem anderen Medium auslöst, ist ein Backup das Festhalten des Datenstands zu einer bestimmten Zeit.[4)] Aus einem solchen kann dann der Zustand der Daten zu ebendieser Zeit wiederhergestellt werden. Gerade wenn das Backup der Daten automatisiert erfolgt, sollte man sich informieren, wann dieses jeweils stattfindet und in welchen zeitlichen Abständen gespeichert wird. Bei Systemen, die dem Nutzenden selbst nicht zu 100 % vertraut sind, sollte der Wiederherstellungsprozess getestet werden. Bei Speicherung auf den Storage-Servern des Rechenzentrums (NAS) findet zusätzlich zu mehrmals täglich erfolgenden lokalen Snapshots jede Nacht ein dezentrales Backup auf Servern der TU Braunschweig statt. Der Wiederherstellungsprozess ist beim RZ erprobt. Diese hier beschriebene Backup-Routine gilt jedoch nicht für die Daten der TU Cloud!

Ein wirklich sicheres Backup hinterlegt den aktuellen Datenstand an bis zu drei zusätzlichen Orten – zwei davon in der Nähe und einer geographisch weiter entfernt beziehungsweise in der Cloud („nah und fern“-Prinzip).[5)] Ergänzend sei hier die 3-2-1-Regel erwähnt: 3 Versionen der Datei auf mindestens 2 unterschiedlichen Speichermedien, wovon 1 portabel oder räumlich getrennt sein sollte. Weiterhin ist es eine gute Idee, die Speicherorte auch für eine oder mehrere Personen der eigenen Einrichtung

Dieses Dokument stammt aus der Dokuwiki-Instanz mit dem Titel FDM-Wiki von DSC und EST, die vom Rechenzentrum der Technischen Universität Clausthal bereitgestellt wird.

zugänglich zu halten, da gerade in der Wissenschaft häufige Personalwechsel stattfinden.

1)
Wickham, H. 2014
2)
Aguinis, H et al. 2021
3)
Markus, K. et al. 2021
4)
Cornell University 2022
5)
Hart, E.M. et al. 2016

Direkt-Link:
https://fdm-wiki.est.tu-clausthal.de/doku.php/datenerfassung_speicherung_zwischenergebnissen

Letzte Aktualisierung: **12.05.2023 06:21**

Dieses Dokument stammt aus der Dokuwiki-Instanz mit dem Titel FDM-Wiki von DSC und EST, die vom Rechenzentrum der Technischen Universität Clausthal bereitgestellt wird.