Make Metadata relational again: Entwicklung einer Metadatenstruktur innerhalb des UMG-MeDIC

Autoren: Caroline Bönisch[1], Dorothea Kesztyüs, Tibor Kesztyüs

Einleitung: Im Rahmen der klinischen Datenerhebung werden zu den jeweiligen Daten auch die Metadaten, d.h. Metainformationen wie beispielsweise ihr klinischer Herkunftsort, erhoben und gesichert abgelegt. Respektive Anforderungen an diese Metadaten können unter anderem aus den FAIR Prinzipien [1] und Standards wie Dublin Core abgeleitet werden [2].
Innerhalb des Betriebs des Medizinischen Datenintegrationszentrums der Universitätsmedizin Göttingen (UMG-MeDIC) werden Metadaten ein grundlegender Wert für die Qualitätssicherung der Daten, zur Weiterverwendung in der Forschung beigemessen. Die identifizierten Metadaten umfassen dabei Informationen über Quellsysteme, Data Owner und Consent, sowie Informationen über die Datengüte.
Methodik: In der ersten Aufbauphase des UMG-MeDICs wurden Metadaten innerhalb der Extract-Transform-Load Prozesse in Form von JSON Objekten in eine CouchDB übertragen. Die Umstrukturierung des UMG-MeDIC und die kontinuierliche Weiterentwicklung der Prozesse resultierte in der Umstellung der Datenhaltung hin zu einer relationalen Datenstruktur. Um dieser Änderung Rechnung zu tragen, erarbeitet das vorliegende Konzept eine generische Datenstruktur zur Speicherung der Metadaten aller klinischen Primär- und Sekundärsysteme der UMG. Dafür wurde vorab die bisherige Datenstruktur des MeDICs betrachtet und mittels Expertengesprächen geprüft, wie die bereits extrahierten Metadaten in die neue Struktur einzuordnen sind. Weiterhin wurden zusätzliche Metadaten auf Basis früherer Anforderungsanalysen identifiziert und eingeordnet.
Ergebnis: Eine generische, relationale Struktur für die Ablage der Metadaten wurde entwickelt. Bereits vorhandene Metadaten zu klinischen Systemen, aus der vorherig genutzten Datenbank (in Form von JSON-Objekten) konnten übernommen werden. Weitere Metadaten, welche bisher nicht automatisch aus den Datenbeständen ausgelesen wurden, werden nun mittels angepasster ETL-Strecken ebenfalls extrahiert und befüllen die neuen Relationen. Die relationale Struktur zur Erfassung der Metadaten umfasst dabei jeweils eine Tabelle für den Metadaten-Typ (Consent, Qualität, Lizenz), eine für die spezifische Ausprägung des Metadatums, und eine für die Metadatenbeschreibung inklusive der Verknüpfung zu anderen Source-Relationen.
Diskussion: Die vorliegende Arbeit zeigt die Entwicklung einer relationalen Metadatenstruktur zur Speicherung gemäß spezifischer Anforderungen im UMG-MeDIC. Die Anforderungen an die Relationen ergeben sich dabei aus dem derzeitigen Betrieb und den Prozessen innerhalb des UMG-MeDICs und der UMG. Die skalierbare Struktur unterliegt kontinuierlichen Weiterentwicklungen und Anpassungen an neue Anforderungen, die eine Datenhaltung mit klinischen Massendaten beinhalten. Innerhalb der Struktur sind demgemäß Erweiterungen vorgesehen und mitbedacht.

[1] Corresponding Author: caroline.boenisch@med.uni-goettingen.de

Referenzen:

[1] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18

[2] ISO. (2019) *ISO 15836-2:2019, Information and documentation — The Dublin Core metadata element set — Part 2: DCMI Properties and classes.* Retrieved from https://www.iso.org/standard/71341.html