David Buschhüter[1]
Jannis Zeller[2]
Stefan Oltmanns[3]
Andreas Borowski[1]
Christoph Kulgemeyer[3]
Josef Riese[2]
Christoph Vogelsang[4]

[1]Universität Potsdam
[2]RWTH Aachen
[3]Universität Bremen
[4]Universität Paderborn

Forschungsdatenmanagement erleichtern durch relationale Datenbanken: Ein Datenmodell für naturwissenschaftsdidaktische Forschung

Hintergrund

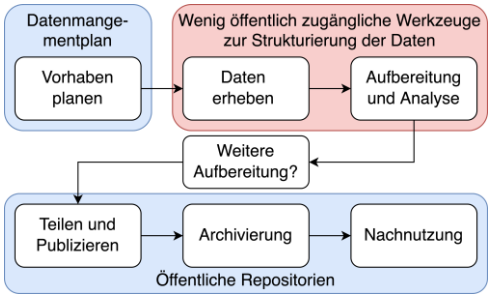

Abb. 1 Darstellung angelehnt an gängige Forschungsdatenzyklen (z. B. Jensen, 2012)

Es existieren heute vielerlei Unterstützungsangebote zur Bereitstellung von Forschungsdaten mit dem Ziel ihrer sekundären Nutzbarkeit entsprechend der FAIR-Prinzipien (Findable, Accessible, Interoperable, Reusable, Wilkinson et al., 2016). So helfen Dienstleister (z. B. GESIS) dabei, dass Forschungsdaten in entsprechenden Repositorien unter Anderem zur Sekundärnutzung nach Projektende zentral bereitgestellt werden (s. Abb. 1). Demgegenüber fehlt es an einheitlichen und frei zugänglichen Datenstrukturen (häufig implementiert in Form von Datenbanken), die lokal (z. B. an Lehrstühlen) eingesetzt werden können. Eine solche Datenstruktur wäre insbesondere hilfreich,

- um bereits früh eine hohe Datenqualität zu gewährleisten und so die mehrfache und sehr aufwändige (Perry & Netscher, 2022; Press, 2016) Datenaufbereitung möglichst zu vermeiden (s. Abb. 1) und
- um innerhalb einer datenauswertenden Einheit (zum Beispiel Lehrstuhl, Institut, Projektverbund) möglichst schnell auf unterschiedliche verschiedene Forschungsdaten zuzugreifen.

Zusammen mit standardisierten Datenerhebungsplänen (DDP Team, 2022; VerbundFDB, 2022) könnte eine solche Struktur dazu beitragen, eine Nutzung entsprechend der FAIR-Prinzipien von Daten bereits von Beginn an technisch zu implementieren (Wilkinson et al., 2016). Langfristig könnte dies positive Effekte auf Produktivität sowie Datenqualität haben.
Nach aktuellem Stand gibt es kaum veröffentlichte Datenmodelle, die sich für die naturwissenschaftsdidaktische Forschung eignen. Erste Anhaltspunkte zur Strukturierung liefern sogenannte Ontologien (Kudryavtsev, Gavrilova & Begler, 2020) oder bereits bestehende Datenmodelle für spezifische Forschungsdaten (Sanalan & Irving, 2007). In Anlehnung an (Sanalan & Irving, 2007) schlagen auch wir die Verwendung einer relationalen Datenbank vor.

Relationale Datenbanken können in einer ersten Form als eine Menge von verbundenen Tabellen verstanden werden (genauer s. Schicker, 2017). Dies ermöglicht eine hohe Standardisierung sowie eine explizite Verbindung von Daten und ihren Metainformationen.
Im Gegensatz zur lokalen Speicherung von z. B. CSV-Dateien kann auch sichergestellt werden, dass mehrere Personen hochfrequent auf die Daten zugreifen können, diese verändern können und Konsistenz dabei trotzdem gewährleistet wird (Schicker, 2017).

Kontext

Das in dieser Studie konstruierte Datenmodell wurde im Zuge der Aufbereitung der Projektdaten für das Projekt Profile-P+ entwickelt (Vogelsang et al., 2018). Der entsprechende Datensatz eignete sich aufgrund seiner Komplexität (mehrere Messzeitpunkte, Standorte, Kohorten, Kodierungen, Instrumente, Datentypen) und der vorhandenen Metainformationen (z. B. Manuale, Protokolle), um eine substantiell verallgemeinerbare Datenbank zu entwickeln.

Ziel

Ausgehend von diesem Datensatz wurde folgende Zielsetzung generiert: Entwicklung einer relationalen Datenbank der Daten des Profile-P+-Projekts mit möglichst hoher Generalisierbarkeit für andere empirische Studien der Naturwissenschaftsdidaktiken.

Entwicklung des Datenmodells

Abb. 2 stellt dar, wie ausgehend von der Zielsetzung der Archivierung der Paper-und-Pencil-Daten des Profile-P+ Projekts ein möglichst verallgemeinerbares Modell für klassische naturwissenschaftsdidaktische Studien entwickelt wurde. Als Zielkriterium galt es dabei, ein Gleichgewicht zu bewahren aus Einfachheit (z. B. über eine möglichst geringe Anzahl an Tabellen) und Generalisierbarkeit über möglichst viele Studiendesigns hinweg (z. B. Berücksichtigen von mehreren Testheften des gleichen Instruments). Das physikalische

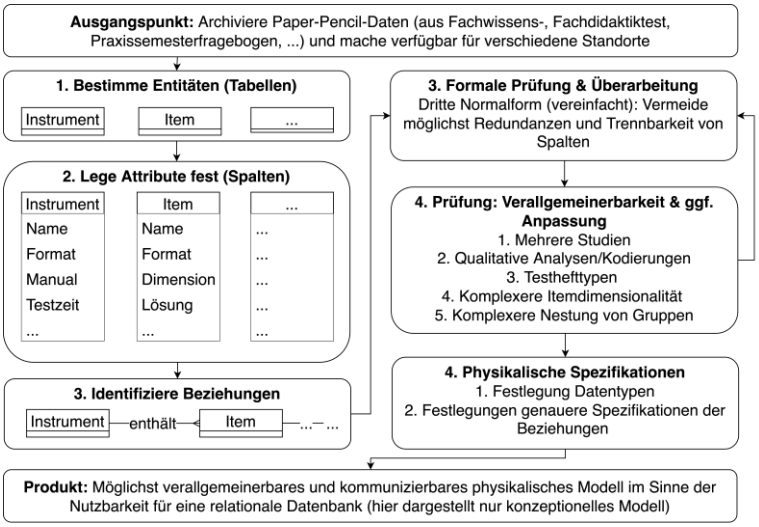

Abb. 2 Darstellung des Prozesses der Entwicklung des Datenmodells

Datenmodell wurde mit MySQL-Workbench 8.0 erstellt. Es ist zu beachten, dass im Folgenden lediglich die Tabellen (Typen von Entitäten) nicht aber die Spalten (Attribute) dargestellt werden. Das entsprechende physikalische Modell (inkl. Attribute und deren Datentypen) steht zur Verfügung unter https://github.com/dbuschhue/rdb_ser_data.

Ergebnis: Datenmodell

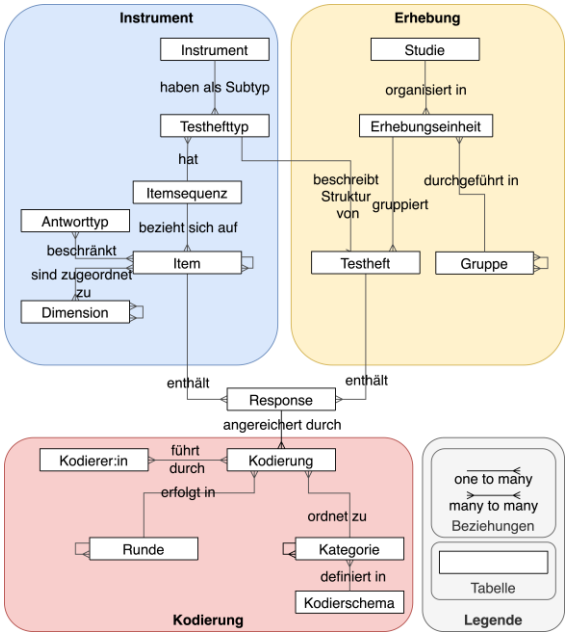

Abb. 3. Darstellung des Datenmodells zur Speicherung von Daten naturwissenschaftsdidaktischer Untersuchungen

In Abb. 3 ist das Datenmodell dargestellt. Dabei wurde die sogenannte Barker-Notation verwendet. Jedes der Rechtecke beschreibt eine Tabelle und steht in Beziehung zu einer anderen Tabelle.
Im Mittelpunkt steht die Response (z. B. die Antwort einer Studierenden auf ein Fachwissensitem). Es können die folgenden drei Bereiche unterschieden werden: Instrument (erlaubt die Speicherung von Informationen zu Erhebungsinstrumenten, Items und Informationen zur Dimensionalität), Erhebung (enthält Informationen zur Studie, sowie zum Studiendesign und die Testhefte), Kodierung (enthält Informationen, die zum Beispiel durch menschliche Kodierung die Response anreichern).

Diskussion

Das oben dargestellte Datenmodell ermöglicht eine einheitliche Speicherung über viele Studien hinweg. Dabei ist jedoch zu beachten, dass die Entwicklung einer Datenbank nicht zu einer eindeutigen Lösung führt und Datenbanken im Zuge neuer Anforderungen häufig weiterentwickelt werden. Die Datenstruktur des Modells ist in den Grundzügen ähnlich zum Modell von Sanalan und Irving (2007).
Zum jetzigen Zeitpunkt ist das Datenmodell nur theoretisch geprüft. Die nächsten Schritte umfassen insbesondere das Importieren der Daten und Testen der Datenbank. Des Weiteren ist denkbar, die Vorgaben des sogenannten STAMP (DDP Team, 2022; VerbundFDB, 2022) explizit zu implementieren. Aufbauend darauf können Front-End-Applikationen, wie Archivierungs-, Löschungs-, und Eingabetools oder Dashboards konstruiert werden, die auf die Datenbank zugreifen. Herausforderungen umfassen insbesondere die Kosten für die Überwachung und Wartung der entsprechenden Datenbank. Mittlerweile gibt es hierzu allerdings bereits Unterstützungsangebote von Seiten vieler Rechenzentren oder Bibliotheken, was die Implementation vielerorts substantiell beschleunigen sollte.

Literatur

DDP Team. (2022). Informationsveranstaltung: Stamp – Standardisierter Datenmanagementplan für die Bildungsforschung. Zugriff am 3.9.2022. Verfügbar unter: https://ddp-bildung.org/2022/05/12/information-events/

Jensen, U. (2012). Leitlinien zum Management von Forschungsdaten. Sozialwissenschaftliche Umfragedaten. Nr. 2012, 7. GESIS-Technical Reports (S. 82). Köln: GESIS – Leibniz-Institut für Sozialwissenschaften. Zugriff am 3.9.2022. Verfügbar unter: http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2012/TechnicalReport_2012-07.pdf

Kudryavtsev, D., Gavrilova, T. & Begler, A. (2020). Creating core ontology for social sciences empirical data integration (Poster). Zugriff am 28.10.2022. Verfügbar unter: https://www.researchgate.net/publication/345362424_Creating_core_ontology_for_social_sciences_empirical_data_integration

Perry, A. & Netscher, S. (2022). Measuring the time spent on data curation. Journal of Documentation, 78(7), 282–304. Zugriff am 28.10.2022. https://doi.org/10.1108/JD-08-2021-0167

Press, Gil. (2016, März 23). Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says. forbes. Zugriff am 27.10.2022. Verfügbar unter: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/

Sanalan, V. & Irving, K. (2007). Database Development for a Large Scale Educational Research Project. In R. Carlsen, K. McFerrin, J. Price, R. Weber & D.A. Willis (Hrsg.), Proceedings of Society for Information Technology & Teacher Education International Conference 2007 (S. 1672–1676). San Antonio, Texas, USA: Association for the Advancement of Computing in Education (AACE). Zugriff am 28.10.2022. Verfügbar unter: https://www.learntechlib.org/p/24808

Schicker, E. (2017). Datenbanken und SQL (Informatik & Praxis Lehrbuch) (5. Auflage). Berlin: Springer Vieweg.

Verbund Forschungsdaten Bildung (VerbundFDB). (2022). Stamp nutzen – Standardisierter Datenmanagementplan für die Bildungsforschung. Zugriff am 27.10.2022. Verfügbar unter: https://www.forschungsdaten-bildung.de/stamp-nutzen

Vogelsang, C., Borowski, A., Kulgemeyer, C., Riese, J., Buschhüter, D., Enkrott, P. et al. (2018). Profile-P + : Entwicklung von Kompetenz und Performanz im Physiklehramt. In C. Maurer (Hrsg.), Qualitätsvoller Chemie- und Physikunterricht- normative und empirische Dimensionen. Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Regensburg 2017 (S. 875–878). Regensburg: Universität Regensburg. Zugriff am 28.10.2022. Verfügbar unter: https://gdcp-ev.de/wp-content/tb2018/TB2018_875_Vogelsang.pdf

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A. et al. (2016). Comment: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 1–9. Zugriff am 28.10.2022. https://doi.org/10.1038/sdata.2016.18