

Teil 6: Der Austausch von Metadaten

In vielen Forschungsprojekten ist das Erstellen oder Sammeln von Daten ein wichtiger Bestandteil der Arbeit. Nun sind diese Daten oft auch für andere Forschungsfragen interessant. Normalerweise wird am Anfang eines Projekts festgelegt, in welcher Form die Daten gespeichert werden sollen. Um Daten weitergeben oder Daten anderer nutzen zu können, muss definiert sein, was die Daten beschreiben.

Projektintern ist diese Frage oft leicht zu beantworten, da die Mitarbeiter wissen, wie und zu welchem Zweck die Daten gewonnen werden. Bei der Weitergabe sollten die Daten genau dokumentiert sein. Dies kann aber einen nicht zu unterschätzenden Arbeitsaufwand darstellen. Durch die Verwendung eines Standards kann diese Arbeit minimiert werden.

Standards geben vor, welche Daten wie beschrieben werden. Das kann helfen, falsche Interpretationen zu vermeiden. Im Idealfall wird eine umfassende Dokumentation mit dem Standard geliefert, die die Semantik, also die festgelegte Bedeutung der Inhalte aufschlüsselt. Auch die Struktur der Daten wird durch den Standard vorgegeben, so wird auch die Wiederverwendbarkeit von Werkzeugen und Arbeitsvorgängen erhöht.

Es gibt aber nicht einen Standard für alle Daten. Vielmehr haben sich zu verschiedenen Zwecken unterschiedliche Standards entwickelt. Es gibt Standards für den allgemeinen Datenaustausch zwischen Projekten, aber auch solche, die gezielt von Portalen entwickelt werden, die die Daten zur Verfügung stellen. Die Liste der im Folgenden vorgestellten Standards verdeutlicht beide Herangehensweisen. Viele Standards sind XML-Anwendungen, sie können also mit den üblichen XML-Technologien verarbeitet werden:

1. Der Standard der Text Encoding Initiative, kurz TEI, dient der wissenschaftlichen Auszeichnung von Texten und Textkorpora.
2. METS beschreibt Objekte in digitalen Bibliotheken und wurde von der Library of Congress und der Digital Library Federation entwickelt. Mit diesem Standard werden besonders die physische Struktur von Digitalisaten sowie deren Metadaten erfasst, ohne die eine Benutzung oft erschwert werden kann.
3. Dublin Core stellt einfache Konventionen zur Beschreibung von Dokumenten und anderen Objekten bereit. Es geht vor allem darum, die essentiellen Metadaten austauschbar zu machen, um ein Maximum der Wiederverwendbarkeit zu erreichen. Daher besteht dieser Standard in einer qualifizierten Form aus nur 15 Kernfeldern.

Im Projekt „Virtuelles Skriptorium St. Matthias“ wurden mittelalterliche Handschriften aus der Trierer Benediktinerabtei St. Matthias digitalisiert. Die Originale wurden eingescannt und alle verfügbaren Metadaten zu den Kodizes wurden in einer Datenbank festgehalten. Metadaten sind in diesem Fall vor allem beschreibende Informationen, wie man sie aus Handschriftenkatalogen kennt. Hier handelt es sich u.a. um allgemeine Angaben zu Datierung, Beschreibstoff, Inhalt, Format, Blattzahl, Sprache, Einband, Schreiberperson, Herkunft oder Buchmalerei.

Während der Projektlaufzeit ist die Anzahl der Daten ständig gewachsen. Zudem ergaben sich strukturelle Änderungen. Aus diesem Grund war es sinnvoll, die Daten nur an einer Stelle zu pflegen. Die Datenbank, in der dies geschieht, besteht aus zwei über IDs verknüpften Tabellen: Eine enthält Informationen zu jedem Kodex und die andere zu den einzelnen Texten, die in einem Kodex zusammengefasst sind.

Alle Metadaten sind unter der Creative Commons Attribution-NonCommercial (CC BY-NC) Lizenz verfügbar. Für jeden Kodex und dessen Inhalte können die erfassten Informationen angezeigt werden. Mit diesen Informationen sind auf der Homepage des Projektes die Bilddateien verknüpft, sodass diese mit dem DFG-Viewer angesehen werden können. In Zukunft soll es auch möglich sein, für jeden Kodex direkt eine XML-Datei im TEI-Format herunterzuladen.

Das Projekt „St. Matthias“ möchte außerdem seine Daten über eine sogenannte OAI/PMH-Schnittstelle verfügbar machen. Über ein Interface sollen auf diese Weise Metadaten zur Verfügung gestellt werden. Hierfür erfolgt eine Speicherung der Daten im Dublin Core-Format. Zu diesen Zwecken muss man sie aus der Datenbank auslesen und in das gewünschte Format exportieren.

Durch dieses Vorgehen liegen die Daten nicht nur in einer Datenbank mit einer Struktur vor, die an das Projekt angepasst ist und sich mehrfach verändert hat. Die Informationen können nun auch von anderen Portalen gesammelt und von anderen Forschern für ihre eigenen Zwecke verwendet werden.