

Daten & XML

Christoph Rinne*

05. Januar 2022

Inhaltsverzeichnis

1	Grundlegende Information	1
1.1	Entwicklung und Forschungsgeschichte	1
1.2	XML Nutzen	2
1.3	XML in R / RStudio	2
2	Software	2
2.1	BaseX	2
3	XML Grundlagen	3
3.1	XML Beispieldaten	4
3.2	XPath	4
3.3	XQuery	4
	Zitierte Literatur	4

1 Grundlegende Information

Das gesamte Thema im Kontext dieser Handreichungen auch nur einführend vorzustellen ist nicht möglich. Es werden also diverse Aspekte nur angerissen oder erwähnt oder stehen nur als Schlagworte im Text. Ziel ist es, Ihnen das Thema “Daten & XML” im Kontext der Archäologie verkürzt vorzustellen und mit wenigen Beispielen Nutzungsszenarien aufzuzeigen.

1.1 Entwicklung und Forschungsgeschichte

Daten sammeln und nachnutzen, zur Recherche oder der Analyse, ist in den archäologischen Fächern ein alltäglicher Vorgang. Meist sind es einzelne Tabellen, gelegentlich auch Tabellen mit Referenzen untereinander oder auch etwas komplexere relationale Datenbanken die hier zum Einsatz kommen. Daneben stehen aber auch andere Techniken zur Strukturierung von Informationen zur Verfügung, eine sehr grundlegende ist die Auszeichnungssprache [XML](#).

Von dem eher technischen Aspekt der Sprache, z.B. XML oder SQL, ist das Konzept zur Strukturierung der Informationen zu trennen. Bedingt durch die jeweilige Entwicklungsgeschichte stehen die vorgenannten Sprachen XML und SQL für sehr unterschiedliche Konzepte von Datenmodellen. SQL steht an den Anfängen der Datenbankentwicklung und für relationale Datenbankmodelle. XML ist die jüngere und mit dem Internet eng verbundene Entwicklung, die zu semistrukturierten Daten in einer hierarchischen Baumstruktur führt.

Zu Beginn der 2000er wurde von Schloen ArchaeoML, ein auf XML basierendes und auf Objekte (*item*) zentriertes Datenmodell präsentiert ([Schloen, 2001](#)). Zeitgleich wird CIDOC-CRM entwickelt, welches ebenfalls auf XML und hieraus mit [RDF](#) als auch [OWL](#) entwickelten Relationen basierend, eine komplexe Ontologie

*Christian-Albrechts Universität zu Kiel, crinne@ufg.uni-kiel.de

für Informationen der Kulturwissenschaften darstellt (Doerr u. a., 2003). Beide sind in ihrer Entstehung eng mit der Entwicklung des Internet verbunden und basieren im Kern auf XML, sind darüber hinaus aber sehr unterschiedlich (Kansa, 2005). ArchaeoML ist in OCHRE aufgegangen und bildet auch eine Grundlage von Open Context Heritage Bytes, z.B. bei DINAA. Die weitere Entwicklung von Open Context erfolgt auf GitHub wobei GeoJSON-LD und damit der Raumbezug und die Referenzen untereinander zunehmend an Bedeutung gewinnen Kansa 2014. Auch CIDOC-CRM wird weiter entwickelt (Oldman, 2014) und vielfach als Grundlage für die Erfassung von Datenbeständen genutzt. Während CIDOC-CRM anfänglich stark an den Relationen orientiert war tritt nun auch hier das Objekt oder *item*, als Ding (*thing*), Objekt (*objekt*) oder (*feature*) verstärkt in das Zentrum.

Ausgehend von XML steht also das Element zunehmend im Zentrum der Strukturierung weniger die Relation von definierten Klassen wie es bei SQL eher der Fall ist. Weder XML noch SQL sind entgegen dieser sehr unterschiedlichen genuinen Konzeption aber auf die jeweiligen Modelle eingeschränkt.

1.2 XML Nutzen

Zur Nutzung von XML-Datenbanken wurde XQuery entwickelt. Grundlegendes Datenmodell ist eine Liste, mit keinem oder einem Element oder auch mehreren Elementen. Hierbei kann ein Element auch wiederum eine Liste sein. Diese als Sequenz bezeichnete Liste von Elementen entspricht damit der Struktur eines XML-Dokumentes. Bei XQuery spielen FLOWER-Ausdrücke ([flou-er], Blume) eine zentrale Rolle. FLOWER steht für *for*, *let*, *where*, *order by*, *return* und ist damit erkennbar an SQL orientiert (*SELECT*, *FROM*, *WHERE*, *ORDER BY*).

Eine Programm zur Nutzung von XML-Datenbanken und XQuery ist BaseX (s.u.).

1.3 XML in R / RStudio

Da ich diesen Text in R-Markdown mit RStudio schreibe verwende ich sogenannte *code chunks*, um die Befehle darzustellen. Mit den Paketen `xml` oder `xml2` können XML-strukturierte Daten in R genutzt werden, dabei handelt es sich allerdings um sogenannte *wrapper* mit einer typischen R-Syntax und nicht um die Implementierung einer XML-Sprache wie z.B. XQuery. Beide Pakete haben jeweils eigene Dokumentationen (`xml`, `xml2`). Die Pakete bieten einige unterschiedliche Funktionen, z.B. `xmlToDataFrame()` bei `xml` und unterscheiden sich weitgehend in den Funktionsnamen: “`xml_<Beschreibung>`” bei `xml2`.

XQuery wird in R nicht umgesetzt, die nachfolgenden *code chunks* mit XQuery sind demnach funktionsloses Layout ohne hervorgehobene Syntax. Für die Nutzung der Pakete `xml` und `xml2` sind einführende Beispiele in der oben genannten Dokumentation vorhanden.

2 Software

XML-Daten und deren Nutzung sind eng mit dem Internet verbunden (s.o.). Die zugehörige Software setzt deshalb meist einen Server (z.B. `localhost`) und PHP voraus.

2.1 BaseX

BaseX ist plattformunabhängig, wird unter Open-Source-Lizenz (BSD) angeboten und für Windows gibt es ein eigenes Installationsprogramm. BaseX wird auf GitHub weiter entwickelt. Die Website des Projektes bietet eine umfassende Dokumentation, darunter auch eine knappe Einführung zu der nachfolgend genutzte GUI die für mich der hervorzuhebende Aspekt dieser Software ist. Für den ersten Kontakt mit XML, XPath und XQuery empfehle ich auch den Text *Basex for dummies von Paul Swennenhuis*.

Die GUI von BaseX erlaubt die Recherche in den Daten mit XPath und XQuery, das Editieren von Daten ist hier nicht möglich.

3 XML Grundlagen

XML besteht aus Elementen (s.o. Sequenz) die auch als Knoten eines Baumes bezeichnet und dargestellt werden. Jedes Element kann Attribute besitzen und einen Wert oder weitere Elemente beinhalten. Ein Beispiel ist:

```
<text autor="Sepp Herberger">
  <satz>
    <wort typ="artikel">Der</wort>
    <wort typ="substantiv">Ball</wort>
    <wort typ="verb">ist</wort>
    <wort typ="adjektiv">rund</wort>
  </satz>
</text>
```

Das Wurzel-Element des Beispiels ist `<text>...</text>` mit den weiteren Elementen “satz” und “wort” die neben den *start tags* ebenfalls zugehörige *end tags* besitzen. Das Element “text” besitzt das Attribut “autor” und das Element “wort” das Attribut “typ”. Dabei ist das Attribut “typ” für “wort” gut gewählt, da es eine eindeutige Zusatzinformation ist während wir für einen Text auch mehrere Autoren erwarten können und deshalb mit den Elementen “autoren” und “autor” statt einem Attribut arbeiten sollten.

```
<?xml version="1.0" encoding="UTF-8"?>
```

```
<kml xmlns="http://www.opengis.net/kml/2.2" xmlns:gx="http://www.google.com/kml/ext/2.2" xmlns:kml="http://www.opengis.net/kml/2.2">
```

Wenn Sie wohlgeformte komplexere XML-Dateien in einem Editor öffnen, z.B. KML-Dateien, sehen Sie als erstes eine Deklaration und auch Verweise auf anzuwendende Regelwerke. Der folgende Beispieldatensatz ignoriert dies.

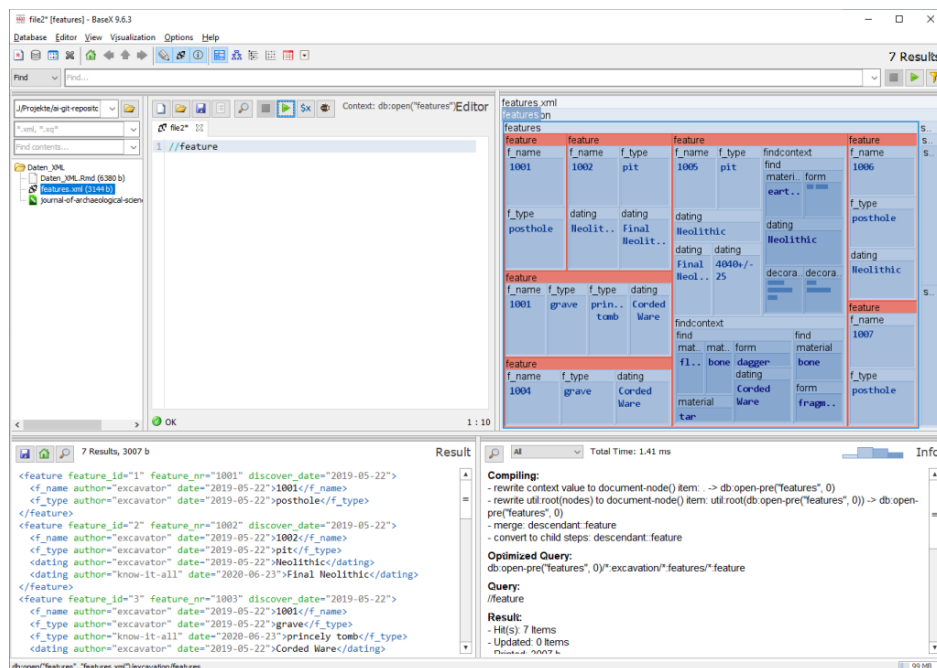


Abbildung 1: Die GUI von BaseX mit der geöffneten XML-Datenbank features

3.1 XML Beispieldaten

Starten Sie für die Arbeit mit den Beispieldaten das Programm **BaseX**. Mit “Database > New ...” öffnen Sie einen Ordner mit XML-Dateien oder eine einzelne XML-Datei und weisen dieser Datenbank einen Namen zu. Mit “Databases > Open & Manage ...” können einmal angelegte Datenbanken erneut geladen werden. Ist die Datenbank “features” in BaseX geladen sollten Sie annähernd vorangehende Bild sehen.

Unterhalb der üblichen Menü- und Icon-Leiste sehen Sie folgende Fenster: Projekt-Ordner, Editor für Syntax, *Map* der Datenbank, *Result* mit aktuell dem gesamten Datenbestand und *Info*. Diese Fenster können über das Menü unter *View* und *Visualization* auch an und ausgeblendet werden.

3.2 XPath

3.3 XQuery

Zitierte Literatur

- Doerr, M., Hunter, J., Lagoze, C., 2003. Towards a Core Ontology for Information Integration. *Journal of Digital Information* 4.
- Kansa, E., 2005. A Community Approach to Data Integration: Authorship and Building Meaningful Links across Diverse Archaeological Data Sets. *Geosphere* 1, 97–109. <https://doi.org/10.1130/GES00013.1>
- Oldman, D., 2014. The CIDOC Conceptual Reference Model (CIDOC-CRM): PRIMER.
- Schloen, J.D., 2001. Archaeological Data Models and Web Publication Using XML. *Computers and the Humanities* 35, 123–152. <https://doi.org/10.1023/A:1002471112790>