Das Deutsche Textarchiv (DTA) im Kontext der NFDI

Dr. Marius Hug (BBAW)

15. September 2022

Im Kontext der Nationalen Forschungsdateninfrastruktur (NFDI) bringt die Berlin-Brandenburgische Akademie der Wissenschaften (BBAW) mit dem Zentrum Sprache eine etablierte Infrastruktur in das Konsortium Text+ ein. Diese Infrastruktur umfasst einerseits Entwicklungen aus dem Akademienvorhaben »Digitales Wörterbuch der deutschen Sprache« (DWDS) und dem sich im Aufbau befindenden »Zentrum für digitale Lexikographie der deutschen Sprache« (ZDL). Andererseits fungiert das »Deutsche Textarchiv« (DTA) als Archiv für Sammlungen und strukturierte, v. a. deutschsprachige Texte aus dem Zeitraum von ca. 1650 bis 1900.

Unser Vortrag möchte drei Schlaglichter dieser Infrastruktur herausgreifen:

1. Datenbereitstellung gemäß der FAIR-Prinzipien am Beispiel des DTA

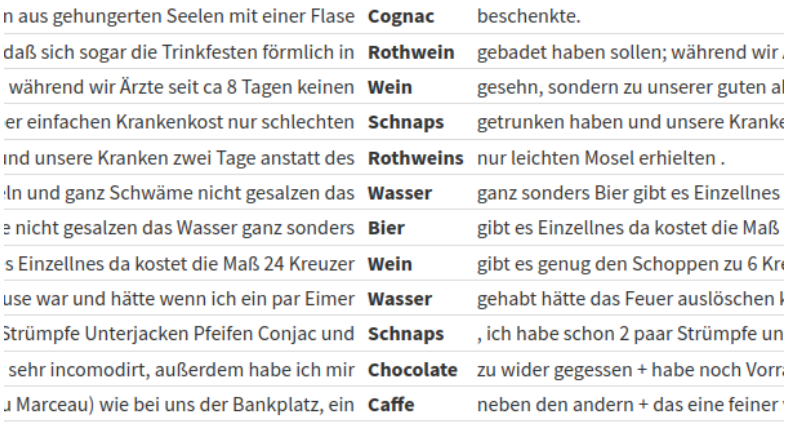
Die Anfänge des DTA als DFG-gefördertes Projekt mit dem Ziel der Erstellung einer Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache reichen 15 Jahre zurück. Das DTA wurde aber früh schon als Aktives Archiv interpretiert und ermöglichte es Forschenden, digitalisierte Werke der entsprechenden Sprachstufe innerhalb der Infrastruktur »fair« zu veröffentlichen.

2) Bedeutung von Standards am Beispiel des DTA-Basisformats

Den Kern der Kurationsbemühungen des DTA bildet das DTA-Basisformat, ein mittlerweile von der DFG empfohlenes TEI-P5-Subsets. Dieses ist die Voraussetzung für eine kohärente Textauszeichnung. Die Weiterentwicklung des DTABf folgt den von einer eigenen Steuerungsgruppe herausgegebenen Leitlinien.

3) Korpusanalyse historischer Texte am Zentrum Sprache

Ein Alleinstellungsmerkmal der Infrastruktur des Zentrums Sprache besteht in der Aufbereitung der historischen Texte aus dem DTA durch (computer-)linguistische Methoden. Diese sind die Voraussetzung für eine komplexe Suche, die u. a. historische Schreibweisen, Wortarten und Textstrukturen berücksichtigt (s. Abb. 1).



n aus gehungerten Seelen mit einer Flase	**Cognac**	beschenkte.
daß sich sogar die Trinkfesten förmlich in	**Rothwein**	gebadet haben sollen; während wir
während wir Ärzte seit ca 8 Tagen keinen	**Wein**	gesehn, sondern zu unserer guten al
er einfachen Krankenkost nur schlechten	**Schnaps**	getrunken haben und unsere Kranke
ınd unsere Kranken zwei Tage anstatt des	**Rothweins**	nur leichten Mosel erhielten .
In und ganz Schwäme nicht gesalzen das	**Wasser**	ganz sonders Bier gibt es Einzellnes
e nicht gesalzen das Wasser ganz sonders	**Bier**	gibt es Einzellnes da kostet die Maß
s Einzellnes da kostet die Maß 24 Kreuzer	**Wein**	gibt es genug den Schoppen zu 6 Kr
use war und hätte wenn ich ein par Eimer	**Wasser**	gehabt hätte das Feuer auslöschen
Strümpfe Unterjacken Pfeifen Conjac und	**Schnaps**	, ich habe schon 2 paar Strümpfe un
sehr incomodirt, außerdem habe ich mir	**Chocolate**	zu wider gegessen + habe noch Vorr
u Marceau) wie bei uns der Bankplatz, ein	**Caffe**	neben den andern + das eine feiner

Abbildung 1: Trefferansicht einer thesaurusbasierten Suche nach „Getränk“ in einem DTA-Korpus.

Seit dem Ende des initialen DFG-Projekts war das DTA an verschiedenen nationalen Infrastrukturprojekten (CLARIN-D, CLARIAH-DE, Text+) beteiligt. Der Vortrag soll also nicht zuletzt auch genutzt werden, die sich im Laufe der Zeit verändernde Rolle des Archivs zu reflektieren. Was waren die Erwartungen an das DTA vor 10 Jahren? Wonach fragt die Community heute, aber auch morgen?