

# Deutsch Diachron Digital: Ein diachron erschlossenes Korpus des historischen Deutschen

## Übersicht über das Projekt und die digitale Text-Aufbereitung



Worum handelt sich bei „Deutsch Diachron Digital“?

- **Name:** Deutsch Diachron Digital (kurz: DDD)
- **Organisation:** Deutsches Verbundprojekt
- **Zustand:** In den Antragsverhandlungen stehend
- **Aufgabe:** Erstellung eines multimodalen Korpus der historischen deutschen Sprachstufen (Althochdeutsch, Mittelhochdeutsch, Frühneuhochdeutsch, Altniederdeutsch, Neuhochdeutsch bis 1900)
- **Beteiligung:** 12 Universitäten zuzüglich weiterer Forschungseinrichtungen aus Deutschland
- **Zeitraum:** 7 Jahre geplant

### Ein Desiderat ...

Bisher fehlt ein zugängliches Korpus zu den Denkmalen der deutschen Sprachgeschichte, welches folgende Eigenschaften auf sich vereint:

- Großen **Umfang**
- **Ausgewogenheit** und **Repräsentativität**
- **Mehrsprachigkeit** (die deutschen Sprachstufen von den Anfängen bis 1900; darüber hinaus etwaige lateinische Paralleltexthe)
- Zuverlässige und weitmöglichst einheitliche **Annotation auf vielen Ebenen**
- **Multimodalität** (Graphik-, Text- und ggf. Klangbestandteile)
- **Diachrone Durchsuchbarkeit**

### Für welche Nutzergruppen ist DDD gedacht?

Das Korpus soll verschiedene (vorwiegend wissenschaftliche) Interessengruppen bedienen, darüber hinaus aber auch einer breiteren aufgeschlossenen Öffentlichkeit übers *Web* unentgeltlich zur Verfügung stehen. Als Nutzer im Vordergrund stehen **Sprachwissenschaftler, Philologen, Literatur- und Geschichtswissenschaftler** wie auch **Informatiker**.

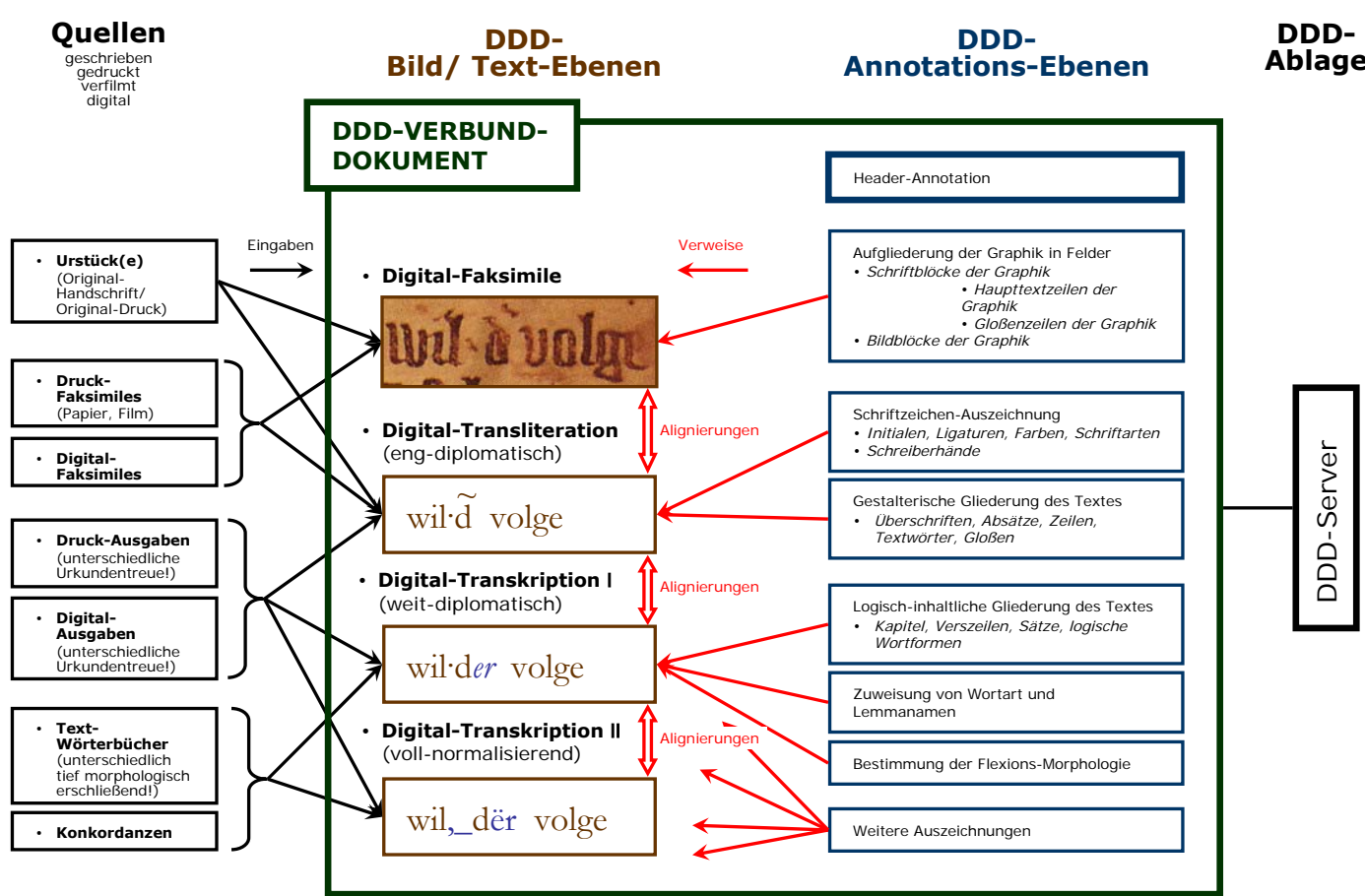
### Planung und Architektur

Dabei sei auf folgende Neu- und Besonderheiten hingewiesen:

- Während große Korpora heute meist auf Gegenwartssprachen beruhen, wird das DDD-Korpus **historische Texte** umfassen. Da die frühe Überlieferung dünn gesät ist, werden die ältesten Sprachstufen vollständig aufgearbeitet, vom Mittelhoch- und -niederdeutschen an treffen wir eine wohlhabgewagte Auswahl.
- Diese Texte werden durch **einheitliche Standards** miteinander vergleichbar gemacht und mittels Annotation abstrahierter Größen **diachron miteinander in Beziehung gesetzt**; so werden sie für sprachgeschichtliche Anfragen besser erschlossen als in vergleichbaren Korpora.
- Die **Tiefe der Annotation** der eingegliederten Texte muß **Mindestanforderungen** genügen; die geringste, konsistent durchzuhaltende Auszeichnung ist auf **Lemmatisierung** samt **Wortartbestimmung** und vollständige Aufschlüsselung der **Flexionsmorphologie** festgesetzt. Für Teile des Korpus kommen etliche weitere Annotations-Ebenen hinzu.
- Auszeichnungen unterschiedlicher Annotations-Ebenen können sich beliebig überschneiden (z.B. Auszeichnung logischer Sätzen und physikalischer Zeilen). Die Handhabung ineinander verschrankter Hierarchien wird ermöglicht, indem sowohl **Text-** als auch **Annotations-Ebenen entkoppelt** (sog. *multi-layer-stand-off*-Architektur) und **Durch Verweise aufeinander bezogen** werden.

## Quellenauswertung, digitale Aufbereitung, Darstellung und Speicherung eines DDD-Textes

➡ Erläuterung der Abbildung



- Die **Text-Ebenen** werden durch **zeichengenaue Alignierung** aneinander gekoppelt.
- Das gesamte Korpus wird nicht in Dateien, sondern in einem modernen **objekt-relationalen Datenbanksystem (Oracle)** gespeichert, welches auch Volltextsuche unterstützt.
- Um möglichst zielgenau verweisen zu können, wird als **kleinste adressierbare Korpus-text-Einheit** auch nicht – wie sonst üblich – das (graphische) Wort gewählt, sondern das **Einzelzeichen** eines Textwortes.
- Die Texte sollen **möglichst nah am Urstück** (Handschrift, Originaldruck) **digitalisiert** werden; viele Texte werden mit einem **Digital-Faksimile** ihrer Handschrift/ ihres Originaldrucks aligniert.

- Es wird sich um ein hinsichtlich sowohl der Textmenge wie auch der Ebenen **offenes Korpus** handeln, das immer weiter ausbaubar bleibt. Das heißt: die Korpusstruktur erlaubt es, **jederzeit neue Texte, Text- und Annotations-Ebenen hinzuzufügen**, solange die die Einheitlichkeit und Querdurchsuchbarkeit gewährleistenden Vorgaben eingehalten werden.
- Die **Annotation** erfolgt entweder **händisch** oder **halbautomatisch**. Dadurch können Fehlerauszeichnungen, wie sie bei vollautomatisch annotierten gegenwartssprachlichen Korpora unvermeidlich sind, weitestgehend vermieden werden. Wir hoffen, so den in ihrem Umfang vergleichsweise kurzen historischen Texten besser gerecht zu werden.

### Berücksichtigung vorhandener Leistungen und Standards

In vielerlei Hinsicht muß DDD nicht erst neue Grundlagen schaffen, sondern kann **auf vorhandene Forschung aufbauen**. Das betrifft insbesondere Standards **TEI, XCES, OLAC** und Tagsets wie das **STTS** (Stuttgart-Tübinger Tagset); in letzterem etwa sind Annotationswerte zur Wortart- und Flexionsmorphologie-Auszeichnung festgelegt, die sich entsprechend unseren Anforderungen erweitern lassen.

Auch vorhandene **Annotationswerkzeuge** und **Darstellungsoberflächen** sollen für unsere Zwecke angepaßt werden.

Das nebenstehende Schaubild stellt vereinfacht einen wesentlichen Teilbereich des Projektes genauer dar, und zwar wie bei der Eingliederung eines Textes in DDD aus den verschiedenen und unterschiedlich weit aufgearbeiteten verfügbaren Quellen ein **virtuelles Verbund-Dokument für diesen Text** entsteht und wie dieses **durch Text-Alignierung und metatextliche Annotation angereichert wird**.

Die verschiedenen, jeweils getrennt abgelegten Ebenen, lassen sich folgendermaßen einteilen:

1. Mehrere **Bild/Text-Ebenen**, die Textfassungen in unterschiedlichen Abstraktions-Zuständen widerspiegeln, werden aus den Quellen erstellt. Diese Bild/Text-Ebenen bewegen sich stufenweise in Richtung Normalisierung – vom graphischen Digital-Faksimile bis hin zur Normalisierung – und dienen jeweils als **Bezugsgrundlage für metatextliche Annotationen**. Dadurch, daß diese Bild/Text-Ebenen zeichenweise miteinander aligniert sind, kann jede von ihnen als **eindeutige Bezugs-Achse (Timeline)** dienen.
2. Die verschiedensten metatextlichen **Annotations-Ebenen** docken an die für sie jeweils **geeignetste Text-Ebene als Bezugs-Ebene an**. Von Elementen aus ihnen wird auf Zeichen (spannen) in den Bezugstext-Ebenen verwiesen. Allein die **Header-Annotationen** verweisen nicht auf einzelne Elemente in einer der Bild/Text-Ebenen, sondern **auf das gesamte Verbund-Dokument**.

Die **Ablage der Elemente** all dieser getrennten Ebenen **sowie der an sie gekoppelten Alignierungen und Auszeichnungen** erfolgt zentral in einem **DDD-Server**, an den übers Netz Anfragen herangetragen werden können.

### Fährnisse und Grenzen der Vereinheitlichung

In den einzelnen Sprachstufen werden zum Teil **unterschiedliche sprachliche Kategorien** ausgebildet, wodurch die Vereinheitlichkeit da an ihre Grenzen stößt. Auch bei **Lemmatisierung** und **Normierung** folgt jede Sprachstufe eigenen Gepflogenheiten; es gilt hier, (bereits in Entwicklung begriffene) Verfahren der **mundartübergreifenden Lemmatisierung** sowie der **sprachstufenübergreifenden Hyperlemmatisierung** fertigzustellen, welche es bei der Anfrage erlauben, auf einer gewissen Abstraktions-Ebene etymologisch verwandten Wortstämmen durch die Sprachräume und -zeiten hin nachzuspüren.

Außerdem erfordert die gegenüber heutigem Material viel größere **individuelle Vielfalt alter Texte** häufig eine **maßgeschneidertere Beurteilung und Behandlung**; oft greifen da starre, entlang moderner Texte entwickelte Standardisierungen nicht – wenn dem Sprachdenkmal keine Gewalt angetan werden soll. Wo nun ein historischer Text für einen Auszeichnungs-Teilbereich keine konsistente Aufarbeitung zuläßt, sollte die sinnvolle Abfrageverfügbarkeit eines Text(teils) in diesem Bereich eingeschränkt werden. Besonders hier ist es erforderlich, daß **parallel zur Formalisierung des DDD-Korpus Grundlagen entwickelt werden, die in Zukunft als Standard dienen können**.

Wandbild zusammengestellt von Thorwald POSCHENRIEDER, HU Berlin, 11/2005