



Abstimmung Digitale Bibliothek

Hintergrundinformationen

Auszug aus der öffentlichen Fassung des Projektantrags, S. 62-66:
http://www.textgrid.de/fileadmin/TextGrid/div/090804_Nachtrag_oeffentlich.pdf

Begründung für den Erwerb der Texte der Digitalen Bibliothek

Die Digitale Bibliothek stellt eine umfangreiche Sammlung von Texten vom Anfang des Buchdrucks bis zu den ersten Jahrzehnten des 20. Jahrhunderts in digitaler Form zur Verfügung. Für die Germanistische Literaturwissenschaft ist die Sammlung von besonderem Interesse, da sie nahezu alle wichtigen kanonisierten Texte und zahlreiche weitere literarhistorisch relevante Texte enthält. Ähnliches gilt für die Philosophie und die Kulturwissenschaften insgesamt. Die Texte stammen zum größten Teil aus Studienausgaben und sind daher zitierfähig; das gilt auch für den Rest, der weitgehend auf die Digitalisierung von Erstdrucken zurückgeht.

Die Texte waren lange Zeit nur als CDs verfügbar, sind jedoch zurzeit über das Portal Zeno.org zugänglich. Dort stehen sie unter einer Lizenz, die der Verlag als 'gemeinfrei' bezeichnet. Diese Lizenz erlaubt die freie Verwendung der Werke: "Dieser Inhalt darf *als Einzelwerk oder Werkbestandteil* – auch zu kommerziellen und gewerblichen Zwecken - kopiert, verbreitet, öffentlich wiedergegeben und Dritten zugänglich gemacht werden."¹ Allerdings behält sich der Verlag die Gesamtpublikation, auch auszugsweise, ausdrücklich vor: "*Die Übernahme der Gesamtheit der verfügbaren Inhalte wie auch von wesentlichen Teilen in eine andere Datenbank ist nicht gestattet.*"²

¹ <http://www.zeno.org/Zeno/-/Lizenz%3A+Gemeinfrei>

² Ebda. Hervorhebungen im Original.

Bisherige Nachnutzungsmöglichkeiten

Es stellt sich die Frage, ob die auf Zeno.org angebotenen Texte unter den bestehenden juristischen und technischen Bedingungen für eine wissenschaftliche Benutzung geeignet sind.

Dazu muss erst einmal geklärt sein, was 'wissenschaftliche Benutzung' heute genau heißt. Drei Formen der Verwendung sind unter dieser Bezeichnung zusammengefasst:

- 1) Lesen,
- 2) Suchen,
- 3) Weiterverwenden.

1) Lesen

Im Fall von Forschungsliteratur ist die wichtigste Form der Nutzung sicherlich die Lektüre, sei es am Bildschirm oder auf dem Papier. Das gilt aber nicht für die kanonischen Texte, von denen hier die Rede ist, da viele in wohlfeilen Ausgaben für eine bequeme Offline-Lektüre zur Verfügung stehen. Etwas anders stellt sich die Situation bei den Wörterbüchern und Lexika da, die auf diese Weise bequem zur Hand sind. Insgesamt aber ist Lesen sicherlich nicht die wichtigste Form der Verwendung dieser Texte.

2) Suchen

Suchen ist für die meisten Geisteswissenschaftler sicherlich die wichtigste Form des Zugriffs auf die Texte. Leider bleibt Zeno.org hierbei weit hinter den Möglichkeiten zurück,³ die die CD-ROM-Editionen bei DirectMedia aufweisen - und auch diese sind nicht wirklich zufriedenstellend.⁴ Das größte Problem ist das Fehlen einer Segmentierung der Daten nach relevanten Kategorien. Zeno.org erlaubt es nicht, die Suche auf einzelne Werke, Autoren,⁵ bestimmte Zeitabschnitte oder spezifische Gattungen einzuschränken. Das macht die Suche für viele Zwecke nur umständlich oder gar nicht brauchbar. Die Datenaufbereitung der CD-ROMs macht deutlich, dass in den Rohdaten viele Metadaten enthalten sind, die sich auf der Website gar nicht finden oder nur so, dass sie für die Suche nicht verwendet werden können. Hier gibt es also einen deutlichen Unterschied zwischen einer populären Verwendung dieser Texte, für die Zeno.org durchaus ausreichend ist, und einer wissenschaftlichen Verwendung, die einen genaueren Zugriff auf die verschiedenen Metadaten voraussetzt.

3) Weiterverwenden

Das Weiterverwenden der Daten in eigenen Forschungskontexten wird in den kommenden Jahren eine selbstverständliche Verwendung sein. In den letzten Jahren sind weltweit

³ Beschreibung der Suchfunktionen in Zeno.org: <http://wiki.zeno.org/Zeno:Suchfunktion>.

⁴ Eine Beschreibung der Suchfunktionen in den DirectMedia-Editionen findet sich hier: <http://computerphilologie.uni-muenchen.de/jahrbuch/jb1/jannidis-2.html>

⁵ Die CD-ROMs erlauben es zumindest, die Suche auf spezifische Autoren und einzelne Werke einzuschränken.

zahlreiche E-Humanities-Zentren⁶ entstanden, die inzwischen zunehmend miteinander kommunizieren und eine Infrastruktur von Diensten anbieten, deren Ziel es ist, das Erstellen und Verwenden digitaler Korpora möglichst zu vereinfachen.⁷ Schon jetzt kann ein Forscher einen Text in diesem Netzwerk statistisch auswerten, andere Dienste sind in Planung oder befinden sich bereits in der Betaphase. In diesem Kontext sind u.a. folgende Formen der Weiterverwendung zu erwarten:

- Zusammenstellung eigener Korpora für quantitative Analysen.
- Integration spezifischer Werke oder Werkblöcke in eigene Editionen oder Textverbünde, z.B. in ein Wörterbuch-Netzwerk.
- Verwendung von Texten als Grundlage einer eigenen Edition, z.B. Erstellung einer Studienausgabe durch Kommentierung von Texten.
- Kollationierung mit anderen Textzeugen / Ausgaben.
- Weitere Erschließung der vorliegenden Texte durch Eintragen von Metadaten, z.B. Geschlecht, Konfession oder soziale Herkunft der Autoren.
- Eintrag von forschungsspezifischem Markup, z.B. narratologische Auszeichnung von Erzähltexten.
- Integration aller Daten in möglichst umfassende Sammlungen, um alle Belegstellen zu finden.

Nur für die wenigsten dieser Verwendungsweisen ist die augenblickliche Situation, also das Angebot der Texte auf Zeno.org, befriedigend: die Texte sind zerstückelt und ihre Metadaten sind nicht trivial erschließbar, wenn überhaupt greifbar. Allerdings enthalten die Texte in der ursprünglichen Auszeichnung, die auf den CD-ROMs sichtbar ist und um die es bei dem Kaufangebot geht, diese und weitere Metadaten.

Typische Interessenten an diesen Texten sind:

- Historische Linguisten, die diachrone Korpora aufbauen
- Editoren digitaler Editionen (Literaturwissenschaftler, Philosophen, Historiker, Musikwissenschaftler usw.), die auf diese Weise wichtige Kontexte für ihre editierten Texte bereitstellen und ihre Texte damit verlinken können.
- Textwissenschaftler, die mit quantitativen Verfahren Texte untersuchen wollen.
- Fachwissenschaftler, die nur in einem oder wenigen Texten intensiv suchen wollen, diese aber dafür vorher noch genauer aufbereiten möchten, z.B. durch die Auszeichnung von Namen und Daten.
- Alle Wissenschaftler, die größere Textmengen zur Entwicklung von Texttechnologien benötigen

⁶ Umfassend, wenn sicherlich noch nicht vollständig:
<http://digitalhumanities.pbwiki.com/centerNet+Digital+Humanities+Centers+>

⁷ Mustergültig ist hier etwa TAPOR <http://tapor.ualberta.ca/>.

Fachwissenschaftliche Anwendungsszenarien

Im Folgenden ist eine Reihe von Anwendungen für die "Digitale Bibliothek" skizziert, die sich aus Diskussionen mit Fachwissenschaftlern entwickelt haben:

1) Die Wörterbücher in der Textsammlung könnten umgehend in das vorhandene Wörterbuchnetz eingebunden werden. Mit den Lexika und Enzyklopädien würde der Anfang eines vergleichbaren Netzes für Nachschlagewerke gelegt werden. Beides kann dann wiederum als Service zur Verfügung gestellt werden - wie es für Wörterbücher ja schon der Fall ist -, so dass Editoren in TextGrid Verweise einschließlich von Links auf Einträge in den Wörterbüchern und Enzyklopädien sehr leicht in ihren Kommentar einfügen könnten.

2) Aufbau von Textsammlungen für spezifische Zwecke.

Alle Texte würden über persistente URLs zur Verfügung gestellt werden, so dass dann stabile Verweise auf einen Großteil der kanonisierten Literatur und deutlich darüber hinaus vor dem 20. Jahrhundert für alle in TextGrid verfügbaren Korpora möglich wären. Viele Editoren scheuen wg. der Instabilität der Links z.Zt. vor solchen Verweisen noch zurück. Editoren könnten dann in TextGrid ohne großen Mehraufwand Kontexte für ihre Texte anbieten und das sehr viel schneller als früher. Hier ein Beispiel (noch aus der Zeit vor TextGrid - aber ein Import nach TextGrid ist in Arbeit), wie eine solche Kontextualisierung dann aussehen könnte:

<http://linglit193.linglit.tu-darmstadt.de:8080/exist/jgoethe/edition.xql?c=/db/jgoethe>

Bei dieser Edition des Jungen Goethe mussten die Herausgeber noch alle Kontexte, z.B. das mythologische Wörterbuch oder die Texte von Goethes Vorbildern, selbst digitalisieren.

Ähnliche Bedürfnisse haben Sprachwissenschaftler bei der Zusammenstellung ihrer Korpora. Textwissenschaftler würden also deutlich schneller solche Editionen und Korpora abschließen bzw. sehr viel mehr Material integrieren können. Schon diese Möglichkeit würde die Attraktivität von TextGrid als Arbeitsumgebung für alle Textwissenschaftler außerordentlich steigern.

3) Quantitative Analysen.

Es gibt erste Kontakte zwischen TextGrid und Projekten wie philomine, eAqua oder Monk, die mit Textmining-Verfahren auf großen Korpora arbeiten. Geplant ist eine Schnittstelle von TextGrid zu mindestens einem solchen Projekt, das die Software für philologisches Textmining entwickelt, um solche innovativen Strategien der nicht-hermeneutischen Textforschung zu ermöglichen. Hier ist ein Link zu einer Diskussion solcher Möglichkeiten; mit dem dort erwähnten Martin Mueller stehen wir wg. einer Kooperation in Kontakt:

<http://www3.isrl.uiuc.edu/~unsworth/hownot2read.html>

All das haben wir nicht beschleunigt vorangetrieben, da wir ja erst im Laufe der kommenden Jahre mit einer nennenswerten Menge von Texten rechneten. Das würde sich durch den Kauf der digitalen Bibliothek natürlich schlagartig ändern. Textwissenschaftler könnten dann in TextGrid ihre Texte für solche Untersuchungen zusammenstellen und vorbereiten (etwa

mit dem TextGrid-Tokenizer, -Lemmatizer und dem Streaming-Editor), um sie dann von den spezialisierten Werkzeugen auswerten zu lassen und die Ergebnisse in TextGrid weiterzuverarbeiten.

Idealerweise können bald einige Verfahren zur statistischen Textauswertung direkt als Service an TextGrid angegebunden werden, z.B.

die offenen Services von Tapor <http://portal.tapor.ca/portal/portal>

4) Besseres OCR

Die Erkennungsqualität des vom TUKL/DFKI in TGII entwickelten OCR-Services steigt, je mehr Text ihm für statistische Auswertungen zur Verfügung steht. Unserer Einschätzung nach ist dieser Service für viele Textwissenschaftler von größtem Interesse und seine Qualitätssteigerung ebenso.

Insgesamt wäre neben den skizzierten technischen Möglichkeiten, die sich durch den Ankauf der digitalen Bibliothek auftun, der größte Gewinn, Textwissenschaftler sehr viel schneller und von Anfang an in deutlich größerer Zahl für TextGrid interessieren zu können; es wäre also eine der wirksamsten Maßnahmen für das fachwissenschaftliche Communitybuilding.

Vereinbarte Lizenzbedingungen mit TextGrid

Das Lizenzmodell wird zur Zeit zum größtmöglichen Nutzen der Öffentlichkeit von Rechtsexperten ausgearbeitet.