1 Projekt books4free.org

1.1 Einleitung

Im nachfolgenden Dokument wird aufgezeigt, warum das Projekt books4free.org entstanden ist. Dabei wird zunächst der historisch enge Bezug zum Projekt Gutenberg aufgezeigt, wobei ausführlich die Stärken und Schwächen von gutenberg.net aufgezeigt werden. Im zweiten Teil werden die Gründe dargelegt, weshalb wir der Überzeugung sind, dass das Projekt books4free.org weit langfristiger und einfacher in der Handhabung ist. Im dritten Teil werden einige Fragen (FAQs) beantwortet.

1.2 Project Gutenberg als Vorlage

1.2.1 1998: 1 GByte ab Projekt 'Gutenberg' über normales Modem

Im Jahre 1998 suchte unsere Firma Archivista GmbH Testdatensätze, um die damals im Entstehen begriffene Archivierungssoftware testen zu können. Gesucht waren möglichst viele ASCII-Dateien, damit die zur Verwendung gelangende Volltext-Technologie auf ihre Leistungsfähigkeit überprüft werden konnte.

Dabei stiess ich auf das Projekt Gutenberg, das damals in etwa 1 GByte an Testdaten bereitstellen konnte. Mit einem 64 KBit-Modem wurden diese Daten über einige Wochen auf den lokalen Server übertragen, um entsprechende Volltext-Indexdateien aufbauen zu können.

Bereits damals mussten wir die Text-Dateien mit erheblichem Aufwand aufbereiten, weil Probleme bei der Formatierung bestanden. Gute sechs Jahre später stellte sich die gleiche Problematik, in einer etwas anderen Grössenordnung. Gefragt waren einige Millionen Seiten für eine Demo-Datenbank, um die Leistungsfähigkeit von Archivista aufzeigen zu können.

1.2.2 2004: Enttäuschende DVD als Auslöser

Das Projekt Gutenberg hatte auf den ersten Blick enorme Fortschritte gemacht, In der Zwischenzeit waren etwa 12'000 Bücher als sogenannte EBooks erhältlich. Da es eine DVD mit den wichtigsten Werken zum Download gab, haben wir diese kurzerhand an einem Nachmittag mit einer normalen ADSL-Leitung heruntergeladen.

Bei uns angekommen, mussten wir allerdings erkennen, dass die EBooks alle gezippt über die gesamte DVD verteilt waren. Das Inhaltsverzeichnis liess sich (zumindest auf meiner Linux-Box) nicht ohne weiteres öffnen, es fehlten Links und die entzippten Dateien enthielten alle möglichen Datenformate.

So sehr ein Non-Profit-Projekt wie Gutenberg auf die Mitarbeit Freiwilliger angewiesen ist, so sehr dabei die schöpferische Freiheit zentral sein muss, so sehr war ich doch enttäuscht, dass das Projekt Gutenberg nicht mehr Konsistenz bei der Datenhaltung erreicht hatte. Aus eigenen Erfahrungen weiss ich, dass das nicht einfach zu bewerkstelligen ist, aber letztlich ist es auch eine Frage des Willens.

1.2.3 Freiheit der Formate schränkt beim Zugriff ein

Was bringt die Freiheit, jeden Buchtext in einem x-beliebigen Format beisteuern zu dürfen, wenn Jahre bzw. Jahrzehnte später diese Texte nicht mehr sauber dargestellt werden können? Was ist die Arbeit all der freiwillig Mitarbeitenden wert, wenn das Endresultat früher oder später nichts weiter als einen Datenfriedhof darstellt? Einverstanden, ganz so schlimm steht es um das Projekt Gutenberg nicht, auf der anderen Seite gibt es aber meiner Ansicht nach wesentliche Punkte, die beim Projekt Gutenberg überdacht werden sollten.

Beim Projekt Gutenberg kann jeder mitarbeiten. Es gibt keine 'Formvorschriften', wie die Texte einzureichen sind. Das ist der positive Aspekt des Projektes. Im Klartext heisst das aber, dass ich entweder ein Buch abtippe oder die Seiten scanne und mit einer Texterkennungssoftware den Text extrahiere. Danach erfolgt ein aufwändiges Verfahren, um den Text in eine akzeptable Lesequalität zu überführen. Am Ende entsteht ein sogenannter Plain-Vanilla-Text, der zwar wenig Platz benötigt, dafür aber keine Formatierung zulässt.

Weit gravierender ist, dass die in einem Buch vorhandenen Bilder nicht im Dokument gespeichert werden können, sodass diese meist entweder gar nicht oder mit einem x-beliebigen Namen versehen dem EBook-Text beigefügt werden. Weil das später wenig lesefreundlich ist, entsteht bei immer mehr EBooks derzeit entweder eine mehr oder minder sauber formatierte HTML-Datei, welche direkte Links zu den Bildern enthält oder ein Word-Dokument, eine RTF-, LaTeX- und oder PDF-Datei. Oft werden sämtliche Dateien zudem in mehreren Varianten gesichert, um der Problematik der verschiedenen Zeichensätze gerecht zu werden.

Daraus entstehen heute in der Regel vier bis zehn (oder noch mehr) EBooks, die alle mehr oder minder den gleichen Inhalt enthalten. Diese Dateien müssen, sofern sie gepflegt werden wollen, einzeln bearbeitet werden. Mag sein, dass sich das bis zu einem gewissen Grad automatisieren lässt, mag auch sein, dass die Freiheit des Formates ein wichtiger Bestandteil des Projektes ist, weil ansonsten viele Texte gar nicht erst eingereicht würden, aber letztlich verhindert die Formatvielfalt auf gutenberg.net heute einen einfachen Zugriff auf die EBooks.

1.2.4 Warum Gutenberg.net nicht komfortabel ist

Das Projekt Gutenberg ist heute nicht mehr benutzerfreundlich. So kann ich zwar nach Titeln und Autoren suchen, erhalte dann aber eine Unzahl an Treffern, die ich einzeln zu interpretieren habe. Mit der Zeit lerne ich, was 'iso 8859-1' heisst, nur interessiert mich das, wenn ich ein Buch lesen möchte? Und macht es wirklich so viel Spass, wenn ich nach dem Download eines EBook-Textes 'mein' ersehntes Buch in einem Editor vorfinde, und zunächst nachformatieren darf? Einverstanden, es gibt immer mehr EBooks auch im HTML-Format, aber wenn ich z.B. mit Konqueror surfe, dann möchte ich die eingebetteten Bilder nicht als schwarze Kasten vorfinden, nur weil diese Bilddatei-Variante nicht sauber von Konqueror unterstützt wird.

Und wie steht's eigentlich mit einer komfortablen Volltextrecherche? Es gibt mittlerweile zwar eine solche (Experimental feature), doch werden die Treffer völlig losgelöst vom Inhalt dargestellt. Irgendwie macht der Bibliotheksbesuch bei gutenberg.net nur mässig Spass. Ich gebe gerne zu, dass Fotos oder Musik- bzw. Video-Dateien sich derzeit für eine Volltextrecherche nicht eignen, aber gehören diese wirklich auf dem Gutenberg-Server? Sicher, es reizt schon, alles, was archiviert werden kann, auch zu archivieren. Was bringt ein Ausweiten der Medien (z.B. auf Musikstücke) aber, wenn nicht mal die Bücher sauber im System erfasst sind? Und wenn schon Musikdateien, müssen es unbedingt patentgeschützte MP3-Dateien sein?

1.2.5 Projekt Gutenberg leistet mehr als öffentliche Bibliotheken

Ich gebe zu, die hier geäusserte Kritik mag vielleicht etwas harsch erscheinen. Sollte dieser Eindruck entstehen, so wäre dies falsch. Ohne gutenberg.net hätten wir z.B. nie die auf dem Markt erhältlichen Volltext-Technologien testen können, was wohl dazu geführt hätte, dass wir die falsche Technologie in unsere Produkte eingeführt hätten.

Weit wichtiger aber, dank gutenberg.net wurde uns eindrücklich klargemacht, was eine Non-Profit-Organisation zu leisten im Stande ist. Etwas böse ausgedrückt, was gutenberg.net erreicht hat, habe ich bisher auf dem Netz nicht annähernd in irgendeiner Form gesehen. Wir haben mit unserer Firma jahrelang versucht, in der Schweiz ansässige Bibliotheken dazu zu ermutigen, ältere Bestände digitalisiert ins Internet zu stellen.

Der Erfolg war im sprichwörtlichsten Sinne vernichtend. Da wurden Projekte gestartet, Seminar-Arbeiten verfasst, immer dann, wenn es darum gegangen wäre, mit dem Scannen der Bestände loszulegen, hiess es, 'no budget' oder 'im nächsten Jahr' vielleicht. Das beste Argument, das mir im übrigen diesbezüglich je zu Ohren gekommen ist, lautete, das Scannen von Büchern gäbe zu wenig her. Toll zu wissen, dass das Gutenberg-Projekt bereits seit Jahren eindrücklich das Gegenteil beweist.

1.3 Vom Forschungszimmer zum Projekt books4free.org

Wahrscheinlich hätte ich die Gutenberg-DVD als normaler Benutzer beiseite gelegt, doch letztlich benötigten wir einige GByte-Textdateien und so haben wir begonnen, mit einigen Skripten die DVD zu entzippen und etwas Struktur hineinzubringen.

Leider erwies sich diese Arbeit als zeitaufwändiger als uns lieb war. Ein Wochen-Projekt weitete sich zu einigen Monaten Aufwand aus, um die Dateien einigermassen (mit Betonung auf einigermassen) sauber in eine Archivista-Datenbank einzuspeisen, um die lang ersehnten Tests durchführen zu können. Immerhin, am Ende reussierten wir (nicht zuletzt dank der tatkräftigen Unterstützung meiner Mitarbeiter) und dank den

daraus gewonnenen Erkenntnis können wir nun unsere Archiv-Produkte weiter optimieren.

Als Nebenprodukt ist dabei die Idee entstanden, die Datenbank nicht nur unserem Forschungszimmer, sondern allgemein zugänglich zu machen. Zugegeben, wenn dabei die Machbarkeit einer digital durchsuchbaren Bibliothek unter Beweis gestellt werden kann, dann sind wir diesbezüglich sicher nicht traurig, aber in erster Linie möchten wir jenen Benefit zurückgeben, den wir durch das Projekt Gutenberg erhalten haben.

Die nun realisierte Lösung unter books4free.org ist als Studie zu verstehen. Wir möchten damit aufzeigen, dass eine virtuelle Bibliothek heute mit einem bescheidenen Aufwand und '08-15'-Technologie machbar ist. Vielmehr aber noch, dass eine solche Bibliothek langfristig finanzierbar bzw. gepflegt werden kann. An dieser Stelle muss ich nochmals auf eine Problematik des Gutenberg-Projektes zu sprechen kommen.

Fast alle EBooks dürften beim Gutenberg-Projekt gescannt werden. Unbestrittenermassen wird das Scannen eines Buches weniger Zeit in Anspruch nehmen als das Abtippen des Buches. Es wird aber doch so sein, dass das Nachbereiten der Bücher einen erheblichen Aufwand erfordert, umso mehr, als heute meist verschiedene Formate generiert werden.

1.3.1 Dank books4free.org mehr Bücher in weniger Zeit online

Mit books4free.org möchten wir aufzeigen, dass dieser Aufwand vermieden werden kann, indem die gescannten Seiten direkt in eine Archiv-Datenbank eingespiesen werden. Im Unterschied zum Gutenberg-Projekt, wo am Ende nur der Text gespeichert wird, stehen bei books4free.org in erster Linie die gescannten bzw. gerasterten Seiten im Mittelpunkt. Nicht der texterkannte und korrigierte Text bildet das 'Original', sondern die Bilder werden direkt zur Verfügung gestellt. Die einzelnen Seiten können so 1:1 und ohne Formatierungsprobleme im Browser dargestellt werden, die Textextraktion erfolgt vollautomatisiert, das zeitaufwändige Nachbearbeiten der Seiten entfällt. Sowohl die Datenbank als auch die Bilddateien selber können komfortabel auf ISO9660 kompatible Datenträger ausgelagert werden.

Der grösste Vorteil der so aufgezeigten Lösung liegt darin, mehr Bücher in weniger Zeit durch die gleiche Anzahl an Personen ins Projekt einfliessen lassen zu können. Da das zeitaufwändige manuelle Nachkorrigieren entfällt, können sich alle Beteiligten wieder auf die Kernaufgabe konzentrieren, und nebenbei können die Bücher erst noch in besserer Qualität und weit einfacher betrachtet werden. Nochmals, books4free.org versteht sich nicht als Konkurrenz zu gutenberg.net, sondern möchte lediglich aufzeigen, dass mit dem gleichen Input letztlich mehr erreicht werden kann.

Books4free.org ist gegenwärtig weder perfekt noch mit gutenberg.net vergleichbar. Dadurch, dass wir 'nur' auf die nachbearbeiteten Textdateien zurückgreifen konnten, mussten wir diese zurück in eine Bilddatei konvertieren, was nach erheblichen Kompromissen verlangte. So haben wir zum Beispiel

nur die Textdateien zur Konvertierung herangezogen. Auf das Umwandeln der HTML-Dateien mussten wir aus Zeitgründen verzichten. Weiter haben wir nur 7- und 8-Bit-Dateien mit dem Zeichensatz 'ISO 8859-1' (Englisch sowie einige mitteleuropäische Sprachen) verarbeitet. Dass dem so ist, liegt primär an der verwendeten Technologie, die auf MySQL 4.0.x basiert – diese Version kennt noch keine Unicode-Zeichensätze.

1.3.2 Zukunft von books4free.org

Books4free.org versteht sich nicht als Konkurrenz zu gutenberg.net. books4free.org ist entstanden, weil es uns am Rande unserer Arbeit gefallen hat, und deshalb möchten wir es der Allgemeinheit zur Verfügung stellen. Das Projekt books4free.org wird derzeit durch die Firma Archivista GmbH gesponsert. Die Ressourcen sind beschränkt. Die Community wird aufzeigen, ob books4free.org derzeit ein Bedürfnis im Netz darstellt oder nicht.

Dass das Konzept von books4free.org langfristig überzeugt, davon bin ich selbstverständlich überzeugt, dies zu zeigen bedarf jedoch einiger Jahrzehnte, und bis dahin wünsche ich beim 'Book-Surfen' auf books4free.org viel Spass.

Urs Pfister, Archivista GmbH, August 2004

1.4 Häufig gestellte Fragen (FAQ)

1.4.1 Ist der Code von books4free.org OpenSource?

Die Lösung auf books4free.org basiert zu 100 Prozent auf OpenSource-Technologie. Der jeweils aktuelle Source-Code entspricht der OpenSource-Edition von Archivista Version 5. Der Code kann direkt unter www.archivista.ch/Av5e.exe bezogen werden.

1.4.2 Beziehung zu Projekt 'Gutenberg'

Es gibt keine direkten Beziehungen zwischen dem Projekt Gutenberg und books4free.org.

Gutenberg.net wird von der gleichnamigen Non-Profit-Organisation betrieben und hat das folgende Ziel: *Project Gutenberg is the first and largest single collection of free electronic books, or eBooks. Michael Hart, founder of Project Gutenberg, invented eBooks in 1971 and continues to inspire the creation of eBooks and related technologies today. The mission of Project Gutenberg is simple: To encourage the creation and distribution of eBooks.*

Books4free.org ist als Nebenprodukt zum Testen der Archiv-Lösung Archivista Version 5 der gleichnamigen Firma Archivista GmbH entstanden. Um die Machbarkeit einer einfach zu bedienenden virtuellen Bibliothek zu demonstrieren, wird books4free.org durch die Firma Archivista GmbH gesponsert.

1.4.3 Welche Werke umfasst books4free.org?

Books4free.org hat sämtliche verfügbaren EBooks als Plain-Vanilla-Text ab der Seite gutenberg.net übernommen und mit aufwändigen Konvertierungen in eine Archivista-Datenbank übernommen. Bei der Konvertierung wurden nur die EBooks, die entweder als 7-Bit-Datei oder im Zeichensatz 'ISO 8859-1' vorlagen, berücksichtigt. Ebenfalls nicht enthalten sind sämtliche Multimedia-Dateien (z.B. MP3-Dateien) sowie einige EBooks, die für eine Online-Recherche wenig Sinn ergeben (z.B. Genom-Projekt).

1.4.4 Kann ich Werke bei books4free.org beisteuern?

Derzeit nein. Books4free.org befindet sich momentan im Stadium einer Projekt-Studie. Die Datenbank wird im Moment nicht erweitert. Sofern ein entsprechendes Interesse der Nutzer/innen entstehen sollte, ist es zu einem späteren Zeitpunkt denkbar, dass neue Werke hinzukommen. Interessierte, die am Projekt mitarbeiten möchten, dürfen sich gerne unter der Kontakt-Adresse melden.

1.4.5 Darf ich Bücher ab books4free.org frei benützen?

Die Nutzungsrechte der Bücher ergeben sich aus den jeweiligen Hinweisen zu Beginn eines jeden Buchtextes. In diesem Umfang dürfen die EBooks frei verwendet werden.

1.4.6 Ich/wir möchte/n eine Kopie von books4free.org betreiben

Kein Problem, Sie sind herzlich dazu eingeladen, jederzeit eine 1:1 oder veränderte Kopie von books4free.org betreiben.

1.4.7 Ist kommerzieller Support erhältlich?

Ja, kommerzieller Support für die aufgezeigte Lösung ist direkt bei Archivista GmbH erhältlich. Falls Sie einen eigene virtuelle Bibliothek auf der Grundlage von books4free.org betreiben möchten, die Lösung aber nicht selber aufbauen möchten, können Sie fixfertige Archiv-Server bei Archivista GmbH erwerben.

1.4.8 Kontaktadresse

Wenn immer Sie mit uns in Kontakt treten möchten, verwenden Sie bitte die folgenden Eckdaten:

Archivista GmbH, Postfach, CH-8042 Zürich, Tel/Fax: +41(0)1 254 54 00/02, Mail: webmaster@archivista.ch