Informatik
Hauptcampus

H O C H
S C H U L E
T R I E R

# gedcom7.js

Realisierung einer JavaScript-Bibliothek für das genealogische Austauschformat FamilySearch GEDCOM Version 7

Marius Müller & David Gruber

Bachelor-Projektarbeit

Betreuer: Christian Bettinger

Trier, 28.02.2023

# Kurzfassung

In der Kurzfassung soll in kurzer und prägnanter Weise der wesentliche Inhalt der Arbeit beschrieben werden. Dazu zählen vor allem eine kurze Aufgabenbeschreibung, der Lösungsansatz sowie die wesentlichen Ergebnisse der Arbeit. Ein häufiger Fehler für die Kurzfassung ist, dass lediglich die Aufgabenbeschreibung (d.h. das Problem) in Kurzform vorgelegt wird. Die Kurzfassung soll aber die gesamte Arbeit widerspiegeln. Deshalb sind vor allem die erzielten Ergebnisse darzustellen. Die Kurzfassung soll etwa eine halbe bis ganze DIN-A4-Seite umfassen.

Hinweis: Schreiben Sie die Kurzfassung am Ende der Arbeit, denn eventuell ist Ihnen beim Schreiben erst vollends klar geworden, was das Wesentliche der Arbeit ist bzw. welche Schwerpunkte Sie bei der Arbeit gesetzt haben. Andernfalls laufen Sie Gefahr, dass die Kurzfassung nicht zum Rest der Arbeit passt.

# Abstract

The same in English.

# Inhaltsverzeichnis

# Abbildungsverzeichnis

# 1

# Einleitung und Problemstellung

In dieser Ausarbeitung ...

## 1.1 Anforderungsanalyse & Ziele

Folgende Anforderungen werden an die Bibliothek gestellt:

- AF01: Dateien oder Strings im Format Gedcom7 sollen eingelesen werden können
- AF02: Die Syntax von Dateien oder Strings soll gemäß der Gedcom7-Spezifikation überprüfbar sein
- AF03: Eingelesene Dateien sollen gemäß der Gedcom7-Spezifikation verändert und erweitert werden können
- AF04: Die in der Gedcom7-Spezifikation definierten Datentypen sollen unterstützt werden
- AF04: Dateien sollen im Gedcom7 Format ausgegeben werden können
- AF05: Die Bibliothek soll erweiterbar sein

# 2

# Theoretische Grundlagen

In diesem Kapitel...

## 2.1 Genealogie und FamilySearch

Genealogie ist ein Überbegriff für die Familien- und Ahnenforschung und beschäftigt sich mit der historischen Herkunft und der Geschichte von Menschen weltweit [Ahn]. Dabei sind insbesondere Abstammungs- und Verwandtschaftsverhältnisse von besonderer Bedeutung, die anhand von Beiweisen aus validen Quellen in Stammbäumen zusammengefasst werden, die aufzeigen, wie eine Generation mit der nächsten verbunden ist. Auf Basis der so erlangten Erkenntnisse kann eine Familiengeschichte erstellt werden, die eine biographische Studie einer genealogisch nachgewiesenen Familie und der Gemeinde in der sie lebten, darstellt [Gen].

Das Aufkommen des Internets stellte einen Wendepunkt in der Genealogie dar.

## 2.2 GEDCOM Version 7

Das Datenformat FamilySearch GEDCOM 7.0 wurde 2021 von der Kirche Jesu Christi der Heiligen der Letzten Tage entwickelt und stellt ein einheitliches, flexibles Format für den Austausch von genealogischen Daten bereit. Das Ziel besteht darin, eine langfristige Speicherung von genealogischen Informationen zu ermöglichen, die für zukünftige Genealogen und die von ihnen verwendeten System zugänglich und verständlich ist [Fam22]. Die im Rahmen dieser Arbeit verwendete Version 7.0.11 wurde am 01.11.2022 veröffentlicht und stellt die aktuelleste[1] Version des Standards dar.

GEDCOM ist ein UTF-8 kodiertes hierarchisches Containerformat, das die Dateinamenserweiterung *.ged* verwendet. Der erste Character einer GEDCOM-Datei sollte das Byte-Order-Mark (U+FEFF) sein. Der Inhalt einer GEDCOM-Datei ist in sog. *Structures* unterteilt, die aus einem *Structure Type* und einem optionalen *Payload* bestehen und mehrere Substrukturen besitzen können. Hat eine *Structure* eine *Substructure*, dann ist die *Structure* die *Superstructure* der *Structure*. Jede

[1] Stand 31.01.2023

*Substructure* hat genau eine *Superstructure* und ist so in der Gesamtstruktur eindeutig zugeordnet. Eine *Structure*, die keine *Supertructure* besitzt, heißt *Record*. Alle Records zusammen mit einer *Header*- und einer *Trailer*-Struktur bilden ein *Dataset*, das den Inhalt einer GEDCOM-Datei darstellt. [Fam22]

Der *Payload* einer *Structure* ist eine Zeichenkette eines bestimmten Datentyps, die entweder Informationen für die *Supertructure* bereithält, oder einen Zeiger auf eine andere *Structure* repräsentiert und somit auf diese verweist. GEDCOM v7 definiert 11 verschiedene Datentypen in [Fam22] mit denen Namen, Daten, Uhrzeiten, Texte und vieles mehr dargestellt werden können. Der *Structure Type* ist eindeutig definiert durch eine URI und gibt an, welche Bedeutung und welchen Datentyp die *Structure* besitzt, welche *Substructures* enthalten sein können und mit welcher Kardinalität diese auftreten können. [Fam22]

Kodiert wird der Inhalt einer GEDCOM-Datei in sog. *Lines*, die eine Zeichenkettenrepräsentation einer Struktur (bzw. eines Teils einer Struktur) darstellen und wie folgt aufgebaut sind (eckige Klammern repräsentieren optionale Inhalte):

Level D [Xref D] Tag [D LineVal] EOL

- Level: Eine Line beginnt mit einem Level, das die Verhältnisse der *Structures* untereinander beschreibt. Alle *Structures* mit dem kleinstmöglichen Level 0 sind Records - Level $\geq 1$ repräsentieren *Substructures*. Eine *Structure* mit dem Level $x$ ist also die *Superstructure* aller folgenden *Structures* mit dem Level $x + 1$.
- D: *D* steht für *Delimiter*, was englisch für Trennzeichen ist und repräsentiert in diesem Fall das Leerzeichen mit dem Unicode $u + 0020$.
- Xref: Xref ist die Abkürzung für *Cross-Reference Identifier* und fungiert als Adresse für eine *Structure*. Möchte man von einer *Structure* auf eine andere *Structure* verweisen, kann dies über einen Zeiger-Payload auf die entsprechende *Structure* realisiert werden.
- Tag: Der *Tag* kodiert den *Structure Type* einer *Structure*.
- LineVal: Im *LineVal* einer Struktur ist der Payload kodiert.
- EOL: EOL steht für End-Of-Line und kodiert das Ende einer Line. Im Format GEDCOM v7 kann dies entweder durch einen Carriage-Return (Unicode U+000D), Line-Feed(Unicode U+000A) oder einen Carriage-Return gefolgt von einem Line-Feed repräsentiert werden.

Ein Ausschnitt aus einer GEDCOM-Datei ist in **Abbildung XY** dargestellt. Dieser Ausschnitt zeigt einen *Record* vom Typ *Individual*, in dem Informationen über ein Individuum gespeichert werden können. Dem Individuum Cross-Reference Identifier *@I1@* zugewiesen, sodass im Dokument auf dieses verwiesen werden kann. In diesem Fall handelt es sich um ein männliches Individuum mit dem Namen John Doe. Über die *Structure* mit dem *Tag BIRT* kann das Geburtsdatum (Birthdate) angegeben werden, das in diesem Fall auf den 1.März 1951 datiert ist. Mit der *Structure FAMS* wird eine Zugehörigkeit zur Family mit dem Cross-Reference Identifier *@F2@* ausgedrückt.

0 @I1@ INDI 1 NAME John Doe 1 SEX M 1 BIRT 2 DATE 1 MAR 1951 1 FAMS @F2@

Detaillierte Erklärungen, alle Informationen zu *Structure Types*, Datentypen, usw. und viele weitere Beispiele können in [Fam22] nachgelesen werden.

## 2.3 Nearley

## 2.4 Mocha

# 3

# Related Work

In diesem Kapitel...

# 4

# Konzept

Die Bibliothek *gedcom7.js* lässt sich wie in Abbildung 4.1 dargestellt in vier logische Teile gliedern. Das zentrale Element ist der Gedcom Parser, mit dem Dateien oder Strings im Format Gedcom7 eingelesen werden und mit Hilfe von *Nearley* auf Korrektheit der Syntax überprüft werden können. Die dafür zugrundeliegende Grammatik wird mit Hilfe eines *Grammar Generators* generiert, der die in [Fam22] definierte Spezifikation in eine nearley-konforme Syntax überführt. Die so eingelesenen Informationen werden in Gedcom Strukturen gespeichert, die verändert und erweitert werden und anschließend im Format Gedcom7 ausgegeben werden können. In den folgenden Abschnitten werden die vier Teile und das Zusammenspiel dieser in detaillierter Form vorgestellt.

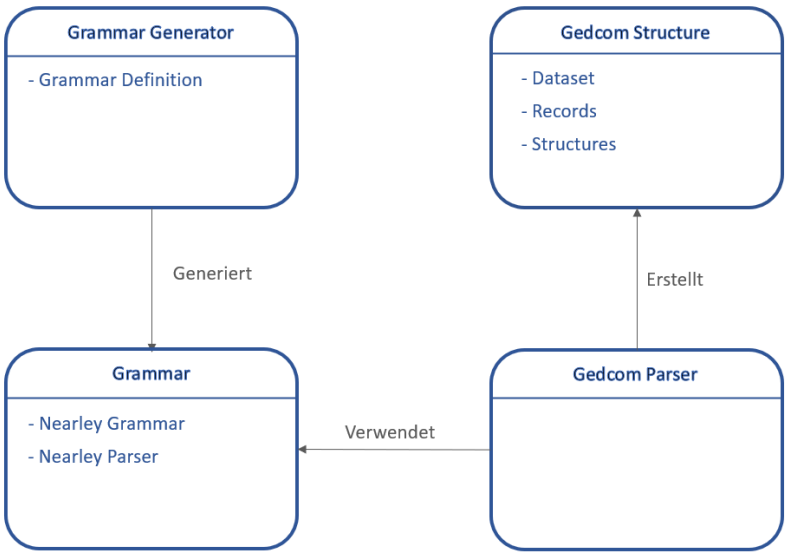


Abbildung 4.1: Allgemeiner Aufbau

## 4.1 Grammatik Generator

asdasd

## 4.2 Gedcom Grammatik

asdasd

## 4.3 Gedcom Struktur & Parser

asdasd

# 5

# Implementierung und Test

In diesem Kapitel...

# 6

# Zusammenfassung und Ausblick

In dieser Arbeit wurde ...

# Literaturverzeichnis


Ahn. Ahnenforschung: *Genealogie*. Abgerufen am 02.02.2022 von `https://www.ahnenforschung.de/themen/genealogie/`.

Fam22. Family History Department: *The FamilySearch GEDCOM Specification 7.0.11*. The Church of Jesus Christ of Latter-day Saints, 15 East South Temple StreetSalt Lake City, UT 84150 US, 7.0.11 Auflage, November 2022.

Gen. Genealogists, Society Of: *Genealogy or Family History*. Abgerufen am 02.02.2022 von `https://www.sog.org.uk/learn/hints-tips/genealogy-or-family-history`.

# A

# Glossar

| | |
|---|---|
| GEDCOM | GEnealogical Data COMunication |
| URI | Uniform Resource Identifier |

# B

## Selbstständigkeitserklärung

Ich erkläre hiermit, dass ich die vorliegende Seminararbeit ohne fremde Hilfe verfasst und nur die im Literaturverzeichnis angegebenen Quellen verwendet habe.

| | |
|---|---|
| Datum | Unterschrift der Kandidatin/des Kandidaten |