

Universität Leipzig  
Fakultät für Mathematik und Informatik  
Institut für Informatik  
Abteilung Datenbanken

# Realisierung einer Datenbank zur Erfassung von PA-Fragebögen und Matching zur ICF

## Bachelorarbeit

Leipzig, März, 2013

vorgelegt von  
Simon, Chill  
Studiengang Informatik

**Betreuer** Anika Groß (Institut für Informatik, Abt. Datenbanken, Universität Leipzig)  
**Betreuer** Prof. Dr. Erhard Rahm (Institut für Informatik, Abt. Datenbanken, Universität Leipzig)  
**Betreuer** Martin Lange (Institut für Gesundheitssport und Public Health, Universität Leipzig)

# Inhaltsverzeichnis

1	Einleitung	4
1.1	Motivation . . . . .	4
1.2	Ziel der Arbeit . . . . .	5
1.3	Aufbau der Arbeit . . . . .	6
2	Grundlagen & verwandte Arbeiten	7
2.1	PA-Fragebögen . . . . .	7
2.2	Ontologie . . . . .	8
2.3	ICF . . . . .	9
2.4	Matching & Mapping . . . . .	12
2.4.1	Metriken . . . . .	13
2.4.2	Selektionsstrategien . . . . .	13
2.4.3	Mapping . . . . .	14
2.4.4	Evaluierungsmethode . . . . .	15
2.5	Verwandte Arbeiten . . . . .	15
3	Integration der Daten	17
3.1	PA-Datenbank . . . . .	18
3.1.1	Importformat . . . . .	18
3.1.2	Datenbankentwurf . . . . .	20
3.1.3	Integration der Fragebögen . . . . .	21
3.1.4	WebGUI . . . . .	22
3.2	GOMMA . . . . .	26
3.2.1	Integration der ICF . . . . .	28
3.2.2	Integration der Fragen . . . . .	28
4	Matching der Fragebögen zur ICF	29
4.1	Workflow . . . . .	29

## *Inhaltsverzeichnis*

4.2	Vorverarbeitung . . . . .	30
4.3	Matching . . . . .	31
4.4	Nachverarbeitung . . . . .	32
5	Evaluation der Ergebnisse	34
5.1	Referenzmapping . . . . .	35
5.2	Ontologie & Fragebogen Analyse . . . . .	35
5.3	Ergebnisse des Matchings . . . . .	36
5.4	Schlussfolgerung . . . . .	40
6	Zusammenfassung & Ausblick	42
	Literaturverzeichnis	44
	Abkürzungsverzeichnis	46
	Abbildungsverzeichnis	48
	Tabellenverzeichnis	49

# 1 Einleitung

## 1.1 Motivation

Fragebögen sind in der Praxis ein häufig genutztes Instrument für Umfragen. Im medizinischen Umfeld haben sich Fragebögen für klinische Studien, sowie für die individuelle Patientenbefragung etabliert. Die gewonnenen Informationen der Fragebögen geben Schlussfolgerungen auf die Gesundheit der befragten Personen. Im Rahmen dieser Arbeit wurden ausschließlich Physical Activity (PA)-Fragebögen betrachtet. Dabei handelt es sich um spezielle Fragebögen zur Erfassung der körperlichen Aktivität von Personen. PA-Fragebögen dienen der Erfassung jeglicher Art von Bewegung der Patienten im Alltag. Im Gegensatz zur Erfassung von Symptomen und Krankheiten soll hier der Status der Beweglichkeit erfasst werden. Viele PA-Fragebögen wurden für spezielle Zielgruppen entwickelt. Beispielsweise gibt es Fragebögen, welche sich ausschließlich mit körperlicher Aktivität von Menschen im Alter über 70 Jahren beschäftigen. Entwickelt werden Fragebögen meist für die eigenen Untersuchungen von Kliniken und wissenschaftlichen Institutionen. Durch die heterogene Entstehung der Fragebögen gibt es große qualitative Unterschiede. Es wurde viel Aufwand betrieben, um die Qualität der Fragebögen zu steigern. Ein Maß dafür ist die Reliabilität. Dieser Wert beschreibt die Verlässlichkeit der Ergebnisse. Er wird anhand von vielfach vorgenommenen Auswertungen des gleichen Fragebogens in verschiedenen Umfeldern ermittelt.

Aufgrund der weltweiten praktischen Anwendung der Fragebögen haben sich im Laufe der letzten 50 Jahre sehr viele Fragebögen entwickelt. In der Praxis ist die Auswahl eines passenden Fragebogens für die eigene Anwendung nicht trivial. Es gibt bisher keinen zentralen Zugriff auf die Fragebögen. Eine zeitaufwendige Recherche ist kaum vermeidbar. Auch eine semantische Suche nach PA-Fragebögen ist bisher nicht vorhanden.

## 1 Einleitung

Eine zentrale Ablage für PA-Fragebögen würde den Zugriff deutlich erleichtern. Durch umfassende Suchbedingungen kann eine weitere Verbesserung der Auswahl erfolgen.

Darin liegt der Ursprung dieser Arbeit. Eine zentrale Ablage für Fragebögen zu schaffen würde die Auswahl des richtigen Fragebogens deutlich erleichtern und beschleunigen. Um die Personen, welche mit Fragebögen arbeiten zu unterstützen und die Arbeit mit Fragebögen zu erleichtern ist eine zentrale Übersicht von großem Vorteil.

In den letzten 10 Jahren hat sich ein Werkzeug zur Klassifizierung von Funktionen und Defiziten entwickelt. Es handelt sich dabei um die Ontologie *International Classification of Functioning, Disability and Health* (ICF)[1]. Die ICF wurde 2001 von der *World Health Organisation* (WHO) empfohlen und bietet Möglichkeiten zur Klassifizierung und Messung von Defiziten. Die ICF bietet die Möglichkeit bestimmte Umstände mittels eines ICF Konzeptes zu erfassen. Dadurch wird eine weltweite Vergleichbarkeit des Gesundheitszustandes von Patienten ermöglicht.

Die ICF hat eine hohe thematische Überlappung mit PA-Fragebögen. Eine Annotation der Fragebögen mit der ICF ermöglicht die Suche nach Fragebögen über die ICF. Die ICF ist eine große Ontologie, wodurch eine manuelle Annotation sehr aufwendig ist. Eine Vollständigkeit kann dadurch nicht gewährleistet werden. Ein semi-automatisches Matching kann diesen Aufwand deutlich reduzieren. Es können Vorschläge generiert werden, welche im Anschluss durch Experten verifiziert werden.

### 1.2 Ziel der Arbeit

Diese Arbeit ist als Kooperation zwischen dem Institut für Gesundheitssport und Public Health der Universität Leipzig und dem Institut für Informatik, Abteilung Datenbanken der Universität Leipzig entstanden. In mehreren Besprechungen haben sich folgende Kernanforderungen herausgestellt:

- Es soll eine zentrale Ablage für die PA-Fragebögen erstellt werden. Dafür wird eine PA-Datenbank erstellt und die Fragebögen sollen in diese integriert werden.

## 1 Einleitung

- Der Zugriff auf die PA-Fragebögen soll online möglich sein. Suchkriterien sollen dabei helfen die Auswahl zu vereinfachen. Dies soll mit einer eigens entwickelten WebGUI realisiert werden.
- Die Fragebögen sollen mit ICF Konzepten annotiert werden, um die Suche nach geeigneten Fragebögen zu verbessern. Dazu soll ein Matching zwischen der ICF und den PA-Fragebögen erfolgen.

Es werden qualitative Aussagen über die Ergebnisse des Matchings getroffen. Dabei werden mögliche Fehlerquellen betrachtet und Verbesserungsvorschläge für zukünftige Arbeiten präsentiert. Das Mapping zwischen ICF und Fragebögen soll in die PA-Datenbank integriert werden, so dass dem Nutzer eine erleichterte Suche in der zentralen Ablage der Fragebögen ermöglicht wird. Die WebGUI soll die Verwendung des Mappings erlauben und einen nutzerfreundlichen Zugriff auf PA-Fragebögen bieten. Die in dieser Arbeit entwickelte Applikation soll am Institut für Gesundheitssport und Public Health der Universität Leipzig genutzt werden. Außerdem soll die Applikation auch externen Nutzern z.B. Sport-medizinischen Forschungseinrichtungen zugänglich gemacht werden.

### 1.3 Aufbau der Arbeit

Im folgenden Kapitel 2 werden alle wichtigen grundlegenden Begriffe erläutert und verwandte Arbeiten diskutiert. Kapitel 3 widmet sich der Integration der Fragebögen und ICF in die PA-Datenbank. Danach wird in Kapitel 4 der Matchingvorgang vorgestellt und auftretende Probleme erläutert. Dabei werden die angewandten Verfahren anhand einiger Beispiele erklärt. Anschließend folgt im Kapitel 5 eine Evaluierung und Analyse der Ergebnisse. Im letzten Kapitel befindet sich eine Zusammenfassung der Arbeit. Weiterhin werden mögliche zukünftige Arbeiten zu diesem Thema diskutiert.

## 2 Grundlagen & verwandte Arbeiten

### 2.1 PA-Fragebögen

PA-Fragebögen sind Fragebögen, welche im medizinischen Umfeld eingesetzt werden. PA-Fragebögen dienen speziell dazu, die körperliche Aktivität der Patienten zu messen. Unter PA (deutsch: körperlicher Aktivität) ist jede Form der Bewegung zu verstehen. Dazu zählt z.B. auch Garten- und Hausarbeit, sowie Einkaufen gehen. Sie werden für die umfangreiche Erfassung von medizinischen Informationen von Patientengruppen genutzt. Daraus können Schlussfolgerungen auf den Gesundheitszustand der Patienten getroffen werden und weitere Aussagen zu der allgemeinen gesundheitlichen Verfassung der betrachteten Gruppe gezogen werden.

Viele Fragebögen sind sehr ähnlich aufgebaut. Sie sind in unterschiedliche Themengebiete aufgeteilt. Zu jedem Themengebiet werden Fragen gestellt, welche der Patient beantworten muss. Die Antwortmöglichkeiten werden meist eingeschränkt oder vorgegeben. Dieses Vorgehen erhöht die Vergleichbarkeit der Ergebnisse. Außerdem können Fehler durch Homonyme, Synonyme und andere Missverständnisse vermieden werden.

In Abbildung 2.1 ist ein Auszug aus dem Priscus PAQ[13] abgebildet. Er wurde von der Universität Bochum entwickelt. Dieser Fragebogen dient speziell zur Erfassung körperlicher Aktivität von Personen im Alter von 70 Jahren und älter. In dem Auszug wird das Themengebiet *Gehfähigkeit* betrachtet. Es sind 2 Fragen abgebildet, welche beide durch Mehrfachauswahl vorgegebener Antworten beantwortet werden sollen.

## 2 Grundlagen & verwandte Arbeiten

---

**UNIVERSITÄT LEIPZIG**  
Sportwissenschaftliche Fakultät  
Institut für Gesundheitssport und Public Health

--	--	--	--	--	--

**Fragebogen**

**A) Im ersten Schritt werden wir Sie zu ihrer Gehfähigkeit befragen.**

1) *Welches Hilfsmittel wie z.B. einen Stock, eine Krücke oder einen Rollator verwendeten Sie in der vergangenen Woche, um zu gehen?*

- Es war kein Hilfsmittel notwendig.
- Es war eine Gehilfe im Sinne eines Gehstocks notwendig
- Es war eine Gehilfe im Sinne eines Rollators notwendig
- Rollstuhl
- Es war keine Fortbewegung aufgrund von Bettlägerigkeit möglich.

2) *Waren Sie in der vergangenen Woche durch ein aktuelles Ereignis wie z.B. einen Sturz oder einen Unfall in ihren Aktivitäten merklich eingeschränkt?*

- nein
- ja
- keine Angabe

**Abbildung 2.1:** Auszug aus Priscus PAQ. Dargestellt sind 2 Fragen mit vorgegebenen Antwortmöglichkeiten

## 2.2 Ontologie

Nach der Definition von Gruber: "An ontology is an explicit specification of a conceptualization." [6] ist eine Ontologie eine explizite, formale Spezifikation einer gemeinsamen Konzeptualisierung. Unter einer Ontologie ist eine Menge von Begriffen zu einem Gegenstandsbereich zu verstehen, wobei alle Beziehungen und Definitionen explizit in der Ontologie enthalten und beschrieben sind. Leichter gesagt ist es eine Sammlung von Wissen zu einem Themenbereich. Die Informationen einer Ontologie werden in Konzepten dargestellt. Formal besteht eine Ontologie  $O = (C, R, A)$  aus Konzepten  $C$ , welche durch Beziehungen  $R$  miteinander verbunden sind und Attribute  $A$  haben. Jedes Konzept hat einen eindeutigen Identifizierer, auch *accession code* genannt. Weiterhin haben Konzepte Attribute wie Name, Synonym und Definition, welche zur näheren Beschreibung dienen. Konzepte können untereinander in einer Hierarchie angeordnet sein. Es gibt Ober- und Unterklassen. Beziehungen zwischen Konzepten werden über Relationen dargestellt. Es gibt dabei verschiedene Typen von Relationen, wie z.B. *is-a* oder *part-of*. Eine Ontologie

## 2 Grundlagen & verwandte Arbeiten

dient zur strukturierten Erfassung von Wissen. Sie dient der semantischen Annotation von Objekten wie z.B. Genen und Proteinen.

In der Informatik spielt die Maschinenlesbarkeit eine große Rolle, wodurch sich einheitliche Formate für Ontologien entwickelt haben. Die für diese Arbeit verwendete ICF Ontologie liegt im *Web Ontology Language* (OWL)[9] Format vor. Es basiert auf der *Extensible Markup Language* (XML) und nutzt die *Resource Description Framework* (RDF) Syntax, wodurch sich Beziehungen zwischen Entitäten sehr genau abbilden lassen.

In dieser Arbeit wird ebenfalls das *Open Biomedical Ontologies* (OBO) Format verwendet. Im Gegensatz zu XML basierten Formaten ist OBO sowohl für Maschinen, als auch für den Menschen leicht verständlich.

### 2.3 ICF

Die *International Classification of Functioning, Disability and Health* Ontologie beschreibt Funktionen und Einschränkungen der Gesundheit des Menschen. Die ICF basiert auf der *International Classification of Diseases* (ICD). Mit der ICD werden hauptsächlich Krankheiten erfasst. Bei der ICF werden zusätzlich alle Funktionen und Einschränkungen des Gesundheitszustandes berücksichtigt. Der Fokus liegt auf der Beschreibung der Gesundheit des Menschen. Sie erweitert diese um neue Kontextfaktoren – auch äußere Umstände genannt - und bietet die Möglichkeit, Funktionen und Aktivitäten zu erfassen und zu klassifizieren. Als wichtiger Kontextfaktor sei an dieser Stelle der Arbeitsplatz genannt. Er kann großen Einfluss auf die körperlichen Funktionen haben. Unter Einschränkungen sind dabei z.B. Brillen und Gehstöcke zu verstehen. Das folgende Beispiel soll die Anwendung der ICF verdeutlichen. Man stelle sich vor, ein junger Mann und eine alte Frau haben beide ein gebrochenes Bein. Der junge Mann wird dabei weniger beeinträchtigt als die alte Frau. Er kann immer noch ohne größere Probleme einkaufen gehen und am täglichen Leben teilhaben. Die alte Frau hingegen ist durch den Beinbruch deutlich stärker eingeschränkt. Es ist ihr nicht mehr möglich, die Einkäufe zu tätigen. Eventuell wohnt Sie in einem hohen Stockwerk und hat dadurch große Probleme ihrem täglichen Tagesablauf nachzukommen. Nach ICD Klassifikation wird dabei nur das gebrochene Bein betrachtet und beide Patienten erhalten die gleichen Behandlungen. Die ICF hingegen erfasst auch alle Kontextfaktoren der alten Frau. Beispielsweise können mit der ICF ihr Wohnsitz und die Möglichkeit des

## 2 Grundlagen & verwandte Arbeiten

Einkaufens erfasst werden. Der Hintergrund der ICF ist eine einheitliche Klassifikation und daraus resultierend eine Vergleichbarkeit der Funktionen und Defizite. Diagnosen können durch die ICF genauer und umfassender erstellt werden. Resultierend aus dem erweiterten Hintergrundwissen ist eine bessere Behandlung des Patienten möglich und wünschenswert. Verwendet wird die ICF bei der Diagnose und Therapie von Patienten. Es können Behandlungen speziell für diagnostizierte ICF Faktoren angewandt werden. Ärzte können unter Verwendung der ICF genauere Diagnosen stellen. Auch Krankenkassen verwenden zur besseren Dokumentation die ICF. Die ICF wurde am 22.05.2001 erstmals von der WHO empfohlen. Zunächst war die sie als Buch verfügbar und wurde später als Ontologie zugänglich gemacht.

Die ICF beruht auf einer Hierarchie und ist in mehrere Hauptkategorien unterteilt. Der Kern besteht aus 4 Hauptkategorien, welche jeweils einen Aspekt beschreiben. Diese 4 Kategorien sind:

- *Body Functions*
- *Body Structures*
- *Activities and Participation*
- *Environmental factors*

Jedes Konzept der ICF wird durch einen eindeutigen Code identifiziert. Dieser beginnt mit einem Buchstaben. Der Buchstabe wird durch die Hauptkategorie bestimmt. Anschließend wird eine Zahl pro hierarchischer Stufe hinzugefügt. Jedes Konzept trägt zusätzlich einen Namen und enthält eine genaue Definition. Ebenso können Inklusionen und Exklusionen angegeben werden. Eine Inklusion gibt an, welche anderen Konzepte von diesem mit eingeschlossen werden. Exklusionen sind im Gegensatz dazu Konzepte, welche ausgeschlossen werden. In Abbildung 2.2 ist ein Auszug aus der deutschen Version der ICF abgebildet. Es wird eine Kategorie mit zugehörigen Unterelementen gezeigt. Jeder Code entspricht einem Konzept. Der Code b134 trägt den Namen *Funktionen des Schlafes* und enthält eine Definition. Der Code beginnt mit einem *b*. Dieses steht für die Hauptkategorie *Body Functions*. Bei diesem Konzept werden auch Inklusionen und Exklusionen angegeben. Die folgenden Konzepte stehen eine Hierarchieebene unter b134. Dadurch wird der Code um eine Stelle länger.

## 2 Grundlagen & verwandte Arbeiten

- b134 Funktionen des Schlafes**  
Allgemeine mentale Funktionen, die sich in einer periodischen, reversiblen und selektiven physischen und mentalen Lösung von der unmittelbaren Umgebung äußern, und die von charakteristischen physiologischen Veränderungen begleitet sind
- Inkl.:** □ Funktionen, die Schlafdauer, Schlafbeginn, Aufrechterhaltung des Schlafs, Schlafqualität, Schlafzyklus betreffen, wie bei Insomnie, Hypersomnie, Narkolepsie
- Exkl.:** □ Funktionen des Bewusstseins (b110); Funktionen der psychischen Energie und des Antriebs (b130); Funktionen der Aufmerksamkeit (b140); Psychomotorische Funktionen (b147)
- b1340 Schlafdauer**  
Mentale Funktionen, die an der Zeit, die im diurnalen oder circadianen Zyklus im Schlaf verbracht wird, beteiligt sind
- b1341 Schlafbeginn**  
Mentale Funktionen, die sich in einem Übergang zwischen Wachheit und Schlaf äußern
- b1342 Aufrechterhaltung des Schlafes**  
Mentale Funktionen, die sich im Durchschlafvermögen äußern

**Abbildung 2.2:** Auszug der ICF. Dargestellt sind 4 Codes mit zugehörigen Definitionen

Verfügbar ist die ICF auf mehreren Wegen. Von der WHO werden 5 verschiedene Sprachversionen bereitgestellt: Englisch, Französisch, Chinesisch, Russisch und Spanisch. Eine deutsche Übersetzung wurde vom *Deutsches Institut für Medizinische Dokumentation und Information* (DIMDI) veröffentlicht. Sie ist als Buch, PDF, online auf der Webseite der WHO<sup>1</sup>; sowie als Ontologie verfügbar. Für diese Arbeit wird die englische Version der ICF Ontologie verwendet, welche von [bioportal.org](http://bioportal.org)<sup>2</sup> zur Verfügung gestellt wird.

Die ICF ist eine neuere Ontologie. Die erste, auf Bioportal verfügbare OWL Version ist auf den 16.11.2009 datiert. Die aktuelle Ausgabe trägt die Version 1.0.2 und wurde am 8.5.2012 veröffentlicht. Bisher sind insgesamt 3 Versionen erschienen. Die derzeit aktuelle Version der ICF Ontologie enthält 1594 Klassen. Auf [bioportal.org](http://bioportal.org) existieren derzeit nur 2400 Korrespondenzen zwischen ICF und 147 anderen Ontologien. Die ICF ist somit kaum vernetzt mit anderen Ontologien. Zum Vergleich hat die ICD 585655 Korrespondenzen zu verschiedenen anderen Ontologien. Vermutlich wird die ICF Ontologie in den nächsten Jahren intensiv weiterentwickelt.

1 <http://apps.who.int/classifications/icfbrowser/>

2 <http://bioportal.bioontology.org/ontologies/1411>

## 2.4 Matching & Mapping

Ontologiematching[4] ist der Vorgang, bei welchem Korrespondenzen zwischen Konzepten von verschiedenen Ontologien gesucht werden. Die Eingabe sind dabei typischerweise 2 Ontologien. Die Domain ist die Basisontologie; die Range ist die Ontologie, zu welcher gematcht werden soll. In [11] werden verschiedene Matching Verfahren klassifiziert. Es werden dabei verschiedene Kriterien zur Einordnung der Matching Herangehensweise vorgestellt. Es wird unterschieden nach Instanz- und Schemabasiertem Matching. Bei Instanz-basierten Verfahren werden die Dateninhalte betrachtet, wobei bei Schema-basiertem Matching die Metadaten betrachtet werden. Eine weitere Unterteilung unterscheidet nach Element- und Struktur-basierten herangehen. Bei Element-basierten herangehen werden die einzelnen Elemente betrachtet, wie Attribute. Hingegen wird bei Struktur-basierten herangehen die Struktur von mehreren Elementen untersucht. Weiterhin kann nach linguistisch oder constraint-basierten Verfahren unterschieden werden. Bei linguistisch-basierten Verfahren werden die Namen und Texte auf Übereinstimmungen untersucht. Ein constraint basiertes Verfahren stützt sich auf Schlüssel und Beziehungen. Das letzte Kriterium unterscheidet zwischen der Kardinalität. Es gibt dabei 1:1, 1:n und n:m Beziehungen. Bei 1:1 wird jeder Frage genau ein ICF Konzept zugeordnet. Bei 1:n werden jeder Frage mehrere ICF Konzepte zugeordnet. Bei n:m können mehreren Fragen mehrere ICF Konzepte zugeordnet werden. Die in dieser Arbeit genutzte Herangehensweise lässt sich als Schema-basiertes, Element Matching einordnen. Verwendet werden Name, Synonym und Definition der Konzepte um Korrespondenzen zu finden. Diese werden mit linguistischen Techniken auf Ähnlichkeiten untersucht, da es sich bei diesen Daten um Strings handelt. Dabei werden verschiedene Verfahren zur Bestimmung der Ähnlichkeit zwischen Konzepten angewandt. Es können einfache String Vergleiche vorgenommen werden, aber auch Strukturmatcher angewandt werden. Diese berücksichtigen die Hierarchie der Konzepte innerhalb der Ontologie.

Weiterhin kann zwischen direkten und indirekten Matching unterschieden werden. Beim direkten Matching werden die Ontologie ohne weiteres Hintergrundwissen gematcht. Beim indirekten Ansatz werden die Eingabeontologien zunächst gegen eine weitere Ontologie gematcht, welche weitere Hintergrundinformationen enthält. Dadurch werden die beiden Eingabeontologien mit weiteren Informationen annotiert. Anschließend wird ein Matching zwischen den annotierten Ontologien vorgenommen.

## 2 Grundlagen & verwandte Arbeiten

### 2.4.1 Metriken

In dieser Arbeit werden 2 Metriken für das Matching verwendet. Es handelt sich dabei um Trigram[14] und Tf-idf[12] Verfahren. Beide sind Token basiert, dabei werden die Strings in einzelne Token aufgeteilt. Hierbei werden Trennzeichen genutzt, um Token zu finden. Typisch ist z.B. das Leerzeichen. Das Trigram Verfahren teilt die Token des Strings in Tripel auf. Beispielsweise werden aus dem Wort *Garten* folgende Trigramme: ##g, #ga, gar, art, rte, ten, en#, n##. Die # wird dabei als Füllzeichen benutzt. Die Trigramme werden für alle Strings erstellt, welche miteinander verglichen werden sollen. Danach werden die Anzahl der gleichen Trigramme für die zu vergleichenden Strings ermittelt. Dieser Wert wird in Relation zur Gesamtzahl der Trigramme gesetzt. Das Ergebnis ist der Ähnlichkeitswert. Dieser wird über die Dice Metrik folgendermaßen berechnet:

Gegeben sind zwei Strings a und b, wobei  $Q(a)$  die Menge der Trigramme aus a ist.  $Q(b)$  ist die Menge aller Trigramme aus String b.

$$sim_{Trigram}(a, b) = \frac{2 \cdot |Q(a) \cap Q(b)|}{|Q(a)| + |Q(b)|} \quad (2.1)$$

Der Tf-idf Algorithmus zählt die Vorkommen der Token. Tauchen in beiden Strings identische Token auf, ist dies ein Match. Dieser Wert wird in Relation zur Gesamtzahl der Token gesetzt. Das Resultat ist ein Ähnlichkeitswert. Für diese Berechnung wird die Jaccard Metrik genutzt. Beide Verfahren sind in *Generic Ontology Matching and Mapping Management* (GOMMA) implementiert und nutzen mehrere Parameter für eine genaue Konfiguration.

### 2.4.2 Selektionsstrategien

Nach Bestimmung der Ähnlichkeiten zwischen zwei Konzepten können verschiedene Selektionsstrategien zur Filterung der Korrespondenzen angewendet werden. Die erste Strategie für die Selektion der Korrespondenzen ist der Threshold. Der Threshold gibt einen Mindestwert für eine Ähnlichkeit an. Nur Korrespondenzen, welche mindestens die Ähnlichkeit des Threshold aufweisen, werden als Korrespondenz in das Mapping übernommen. Alle Korrespondenzen mit einer Ähnlichkeit unterhalb des Threshold

## 2 Grundlagen & verwandte Arbeiten

werden gefiltert. GOMMA stellt Funktionen für die Selektion anhand des Thresholds bereit.

Eine weitere Art der Selektion ist das MaxDelta. Die Idee hinter diesem Verfahren liegt darin, dass nicht alle Korrespondenzen zwischen 2 Konzepten behalten werden, sondern nur die besten. In Bezug auf die ICF würde dies bedeuten, dass ein ICF Konzept auf mehrere Fragen gematcht werden kann. Aber auch umgedreht eine Frage mehrere ICF Konzepte. Der Zahlenwert dieses Parameters, das delta, gibt dabei die maximale, erlaubte Abweichung von der besten Ähnlichkeit an. Hat ein Konzept eine Korrespondenz mit der besten Ähnlichkeit von 0,5, so werden bei einem delta Wert = 0,05, auch alle Ähnlichkeiten bis zu der Ähnlichkeit = 0,45 als weitere Korrespondenzen mit eingeschlossen. Es empfiehlt sich den Threshold bei Verwendung von MaxDelta zu verringern. Dadurch werden mehr Korrespondenzen gefunden, aber es werden durch MaxDelta nur die jeweils besten Korrespondenzen eines Konzeptes beachtet. MaxDelta kann aus zwei Richtungen angewendet werden: Domain zu Range und Range zu Domain, sowie beide Richtungen.

Die letzte Selektionsstrategie ist MaxN. Diese funktioniert ähnlich wie das MaxDelta. Dabei wird die Anzahl der Korrespondenzen pro Konzept limitiert. Hat der Parameter N den Wert 3, so werden die 3 besten Korrespondenzen für jedes Konzept im Mapping gespeichert. Dabei werden die 3 Korrespondenzen für das Mapping genommen, welche die höchste Ähnlichkeit haben. Auch diese Strategie kann ähnlich wie MaxDelta in 2 Richtungen angewendet werden: Domain zu Range, Range zu Domain und beide.

### 2.4.3 Mapping

Das Ergebnis ist das Mapping. Dieses enthält die gefundenen Korrespondenzen zwischen den Konzepten der Ontologien. Ein Mapping zwischen den Ontologien  $O_1$  und  $O_2$  ist folgendermaßen definiert:  $M_{O_1, O_2} = \{(c_1, c_2, sim) | c_1 \in O_1, c_2 \in O_2, sim \in [0, 1]\}$ . *sim* bezeichnet die Ähnlichkeit zwischen zwei Konzepten und liegt im Bereich zwischen  $[0, 1]$ . Je größer die Zahl ist, desto ähnlicher sind die Korrespondenzen. Das Mapping wird genutzt, um die Korrespondenzen zwischen den Konzepten zu speichern und in die PA-Datenbank zu integrieren.

## 2 Grundlagen & verwandte Arbeiten

	<b>korrekte Korrespondenz</b>	<b>falsche Korrespondenz</b>
von Matcher als korrekt vorhergesagt	true positive (tp)	false positive (fp)
von Matcher als falsch vorhergesagt	false negative (fn)	true negative (tn)

**Tabelle 2.1:** Einteilung der Korrespondenzen in 4 Kategorien.

### 2.4.4 Evaluierungsmethode

Die Qualität des Mappings wird anhand der Precision, Recall und F-Measure gemessen. Diese Werte können präzise bestimmt werden, weil ein Referenzmapping vorhanden ist. Dieses wurde vom Institut für Gesundheitssport und Public Health der Universität Leipzig erstellt und für diese Arbeit zur Verfügung gestellt. Die gefundenen Korrespondenzen werden in 4 Kategorien unterteilt. Tabelle 2.1 zeigt die Einteilung.

Folgende Formeln liegen der Berechnung der einzelnen Werte zugrunde:

$$Precision : \frac{tp}{tp + fn} \quad (2.2)$$

$$Recall : \frac{tp}{tp + fn} \quad (2.3)$$

$$F - Measure : \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (2.4)$$

## 2.5 Verwandte Arbeiten

Bisher existieren nur wenige Arbeiten im Bereich des Matching bzw. Annotation von Fragebögen zu Konzepten der ICF. In einer Arbeit [7] wurden manuell Fragen aus dem XSMFA Fragebogen[15] zu ICF Konzepten zugeordnet. Zielstellung war es herauszufinden, welche Bereiche der ICF von dem XSMFA abgedeckt werden. Es handelt sich dabei um einen kleinen Fragebogen. Er enthält insgesamt 16 Fragen. Im Gegensatz zu dieser Arbeit, wurden dabei die Korrespondenzen zwischen ICF Konzepten und Fragen

## 2 Grundlagen & verwandte Arbeiten

ausschließlich manuell ermittelt. Da in dieser Arbeit wesentlich mehr Fragen der ICF zugeordnet werden sollen, ist es hilfreich ein semi-automatisches Matching anzuwenden.

In [3] wird die *Activities and Participation* Kategorie der ICF mit der *Suggested Upper Merged Ontology* (SUMO) Ontologie gematcht. Dabei wurden die Konzepte der ICF hinsichtlich ihrer Beziehungen untereinander untersucht. Ziel war das identifizieren von logischen Fehlern innerhalb dieser Kategorie.

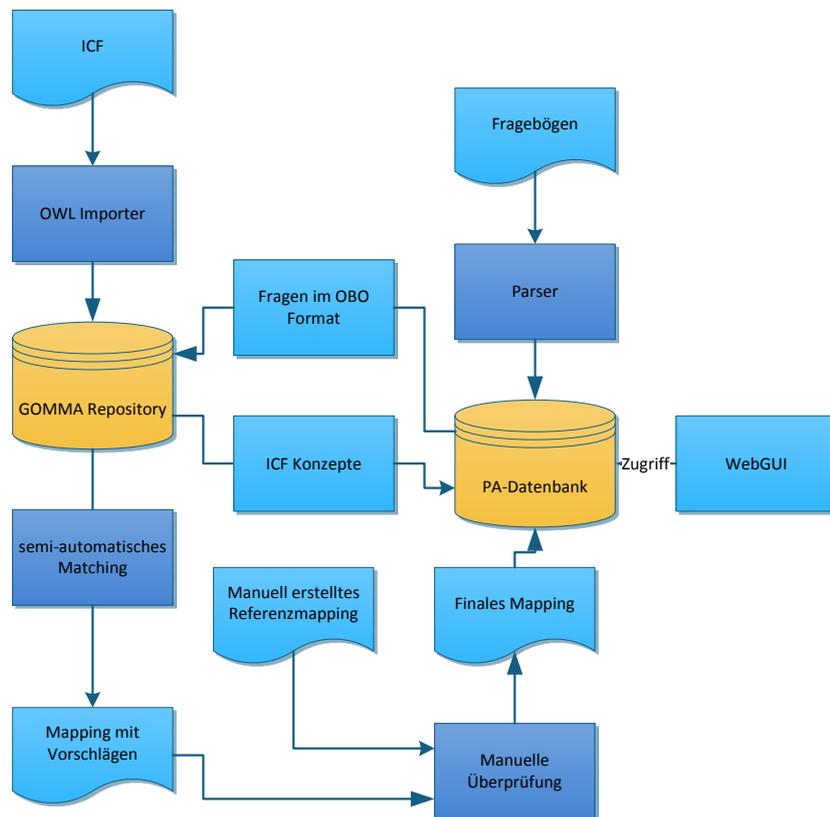
Viele aktuelle Projekte zur ICF sind auf DIMDI<sup>3</sup> gelistet. Als Beispiel sei ein Projekt genannt, dessen Ziel die Entwicklung von Core-Sets zu dem Krankheitsbild Schwindel ist. Dabei werden alle ICF Konzepte zusammengefasst, welche mit diesem Thema in Verbindung stehen - sogenannte *Core-Sets*.

---

<sup>3</sup> <http://www.dimdi.de/static/de/klassi/icf/icf-projekte.html>

# 3 Integration der Daten

Für eine zentrale Ablage der Fragebögen müssen diese in eine PA-Datenbank integriert werden. Im Folgenden wird näher auf den Vorgang der Datenintegration eingegangen.



**Abbildung 3.1:** Workflow Diagramm aller nötigen Arbeitsschritte. Der Zugriff auf die gewonnenen Informationen findet auf der PA-Datenbank statt

### 3 Integration der Daten

Abbildung 3.1 stellt den gesamten Ablauf der Integrations- und Matching Schritte dar. Es ist ein Überblick über alle nötigen Arbeitsschritte. Die Ausgangsbasis sind 2 Quellen. Bei der ersten handelt es sich um die ICF Ontologie im OWL Format. Die zweite Quelle sind die Fragebögen. Diese liegen in einem genau spezifizierten Importformat vor. Dieses wird im nächsten Abschnitt genauer vorgestellt. Die Fragebögen werden unter Verwendung eines Parsers zunächst in die PA-Datenbank für die Webseite importiert.

Für das Matching zwischen Fragebögen und ICF Ontologie wird das an der Abteilung Datenbanken der Universität Leipzig entwickelte System namens GOMMA [8] verwendet. GOMMA stellt ein Repository zur Verfügung. Für das Matching werden zunächst die Fragen der Fragebögen in das GOMMA Repository integriert. Die Fragen werden aus der PA-Datenbank extrahiert und in das GOMMA Repository importiert. Die Integration wird über das OBO Format realisiert. Die ICF Ontologie wird mit einem OWL Importer in das GOMMA Repository importiert. Sind Fragen und ICF erfolgreich in GOMMA importiert, kann das Matching erfolgen. Das Ergebnis ist ein Mapping zwischen Fragen und ICF Konzepten. Dieses Mapping dient der Vorschlagsgenerierung. Die Vorschläge werden manuell überprüft und korrekte Korrespondenzen werden in das Referenzmapping übernommen. Daraus resultiert das finale Mapping für die PA-Datenbank. Das Mapping muss für die Bereitstellung auf der Webseite in dessen PA-Datenbank importiert werden. Damit dies möglich ist, müssen zunächst die ICF Konzepte in die PA-Datenbank integriert werden. Da ein Mapping die Korrespondenzen zwischen Fragen und ICF enthält, müssen sowohl die Fragen, als auch die ICF Ontologie in der PA-Datenbank vorhanden sein. Dieser Schritt muss vor dem Import des Mappings realisiert werden, aufgrund von Fremdschlüsselbeziehungen. Es werden zunächst alle ICF Konzepte in die PA-Datenbank importiert. Anschließend kann das Mapping integriert werden. Nach der Integration ist es möglich, die Webseite mit allen Funktionen zu nutzen. Die genauen Anforderungen an die Webseite werden in einem späteren Kapitel erläutert.

## 3.1 PA-Datenbank

### 3.1.1 Importformat

Für den Import der Daten der PA-Fragebögen wird ein einheitliches, maschinenlesbares Format benötigt. Der größte Teil der Daten sind Meta-Daten zu jedem Fragebogen. Diese

### 3 Integration der Daten

wurden vom Institut für Gesundheitssport und Public Health der Universität Leipzig zusammen getragen und für diese Arbeit zur Verfügung gestellt. Sie beinhalten allgemeine Informationen zu Fragebögen z.B.: Autor, Jahr, Sprache aber auch die einzelnen Fragen zu jedem Fragebogen, sowie die Validität. Die Validität ist ein Maß für die Beständigkeit der Resultate der Fragebögen. Wobei jeder Fragebogen vielfach valide sein kann. Die Fragen wurden zudem mit Schlagwörtern versehen. Die Schlagwörter werden Synonyme genannt und geben für jede Frage den Themenbereich der Frage an. Dieses Feld spielt für das spätere Matching eine tragende Rolle. Des Weiteren enthalten Fragebögen Dimensionen. Eine Dimension ist die Zusammenfassung von Fragen eines Themengebietes unter einem Begriff. Beispielhaft können die Fragen „Treiben Sie Sport?“ und „Welche Sportart betreiben Sie am häufigsten?“ unter der Dimension Sport zusammengefasst werden. Um die gewünschten Meta-Daten der Fragebögen in die PA-Datenbank zu integrieren wurde ein Importformat, sowie eine darauf angepasste PA-Datenbank entwickelt.

Kurzform	Bezeichnung	Autor	Konstrukt	Setting	Zielgruppe	Altersgruppe MIN	Altersgruppe MAX	Ziel des Fragebogens	Begründung der Konstruktion eines deutschsprachigen Instrument	Betrachtungsz Zeitraum	maximale Bearbeitungszeit
Baecke Fragebogen	Deutsche Version des Erhebungsv erfahrens	Wagner et al.	Energieumsatz	settingunabhängig	Gesunde Erwachsene	20	65	Überprüfung			10
Frage	Antwortfor		Dimension	Cronbachs		Sprache	übersetzt		Validitätsa	Unterart	
Welche Tätigkeit üben Sie hauptberuflich aus?	Offen		Arbeit	0,864		deutsch	ja		Inhalt		
Bei der Arbeit sitze ich...	Zuordnung		Sport	0,851		englisch	ja		Kriterium	Übereinstimmung	
Bei der Arbeit gehe ich...	Zuordnung		Freizeit	0,57							

**Abbildung 3.2:** Die Excel Vorlage. Verwendung als Importvorlage für die Datenintegration

Das Importformat, dargestellt in Abbildung 3.2, ist eine Excel Vorlage mit genau festgelegten Spalten und zugehörigen Werten der Felder. So können in Spalten, wo nur Zahlen gewünscht sind, auch nur Zahlen eingetragen werden. Dadurch wird sichergestellt, dass keine falschen oder falsch formatierten Werte in der Vorlage erfasst werden können. Dieses Vorgehen minimiert die Fehleranfälligkeit für den Import beträchtlich. Genutzt wird die Vorlage vom Institut für Gesundheitssport und Public Health. Gespeichert werden die

### 3 Integration der Daten

Tabellen anschließend in einem *Comma separated Value* (CSV) Format. Dieses Format ist leicht zu verarbeiten und vereinfacht den Importprozess.

#### 3.1.2 Datenbankentwurf

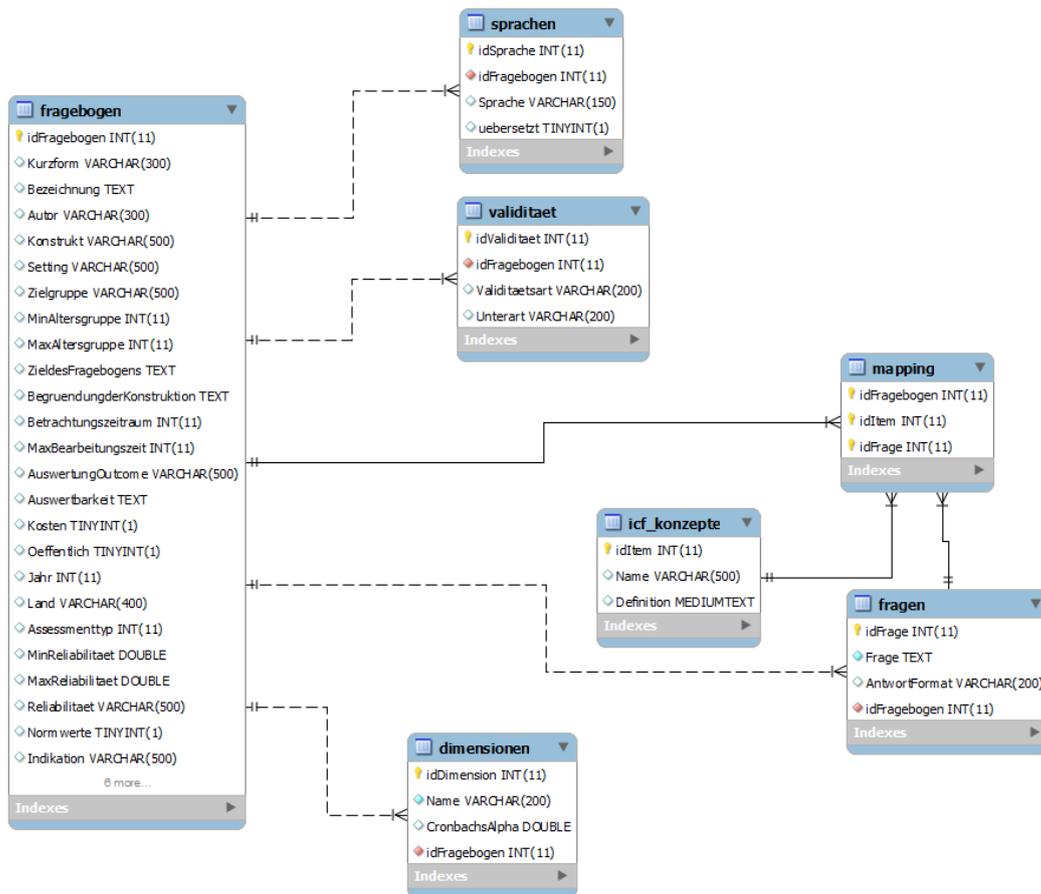


Abbildung 3.3: vollständiger Datenbankentwurf

Basierend auf dem entwickelten Importformat wurde ein Datenbankentwurf erstellt, welcher als Grundlage für die Datenbankeinstellung dient. Der Entwurf wurde unter Verwendung des Programmes *MySQL Workbench* erstellt. Bei der verwendeten Notation handelt es sich um die *Crow's Foot* Notation. In Abbildung 3.3 ist der Datenbankentwurf dargestellt. Die Hauptrelation ist *fragebogen*, in welcher alle einmal vorkommenden Informationen gespeichert werden. Es wurden den Quelldaten entsprechende Datentypen

### 3 Integration der Daten

gewählt. Ein Beispiel dafür ist die Information *Jahr*, welche nur Zahlen enthält. Dafür wurde der Datentyp INT benutzt. Jeder Fragebogen kann in mehreren Sprachen erscheinen, weshalb eine extra Relation für Sprachen erstellt wurde. Die nötige 1:n Beziehung wird über den Fremdschlüssel *idFragebogen* realisiert. Dies gilt auch für die *validitaet*, die einzelnen *fragen* und die *dimensionen*. Jeder dieser Felder kann eine beliebige Anzahl an Informationen enthalten, weshalb für jedes Feld eine separate Relation verwendet wird. Dafür wird für jede Relation ebenfalls der Fremdschlüssel *idFragebogen* benötigt.

Die Relation *icf\_konzepte* enthält alle aus der Ontologie gewonnenen ICF Konzepte (Accession, Name, Synonym, Definition). Die zweite Relation *mapping* enthält die m:n Beziehungen zwischen den ICF Konzepten und den einzelnen Fragen, sowie den zugehörigen Fragebögen. Das integrierte Mapping ermöglicht später die gezielte Suche nach Fragebögen über ICF Konzepte. Als Datenbankmanagementsystem kommt MySQL zum Einsatz. MySQL ist ein Open Source Projekt.

#### 3.1.3 Integration der Fragebögen

Die Integration der Fragebögen wurde über ein Java Programm realisiert. Es wurde im Zuge dieser Arbeit entwickelt. Die benötigten Parameter sind der Pfad der zu importierenden Fragebögen in dem vorgestellten Importformat. Außerdem werden die Parameter für die Datenbank zur Speicherung der gewonnenen Daten benötigt. Für jede gefundene Datei im Quellordner wird ein Importprozess gestartet. Dabei werden nur csv Dateien beachtet. Jede Datei wird anschließend zeilenweise durch den Parser gelesen. Alle Werte werden in den für sich entsprechenden Datentyp konvertiert. Einige weitere Informationen werden zur Laufzeit berechnet. Beispielsweise muss die Anzahl der Fragen und der Dimensionen nicht explizit genannt werden, sondern wird anhand der Anzahl der Werte berechnet. Sind alle Werte vorhanden und valide werden sie anschließend in die PA-Datenbank importiert. Zur Dokumentation wird eine .log Datei erstellt. Darin wird für jeden Fragebogen festgehalten, ob er erfolgreich importiert wurde. Im Fehlerfall wird die Fehlerursache ebenfalls ins logfile geschrieben. War die Integration eines Fragebogens erfolgreich, erhält die Importdatei die Dateiendung .done. Bei Auftreten eines Fehlers wird anstelle von .done, .error angehängt. Dadurch wird vermieden, die gleichen Fragebögen noch einmal zu bearbeiten, falls der Importvorgang erneut gestartet wird.

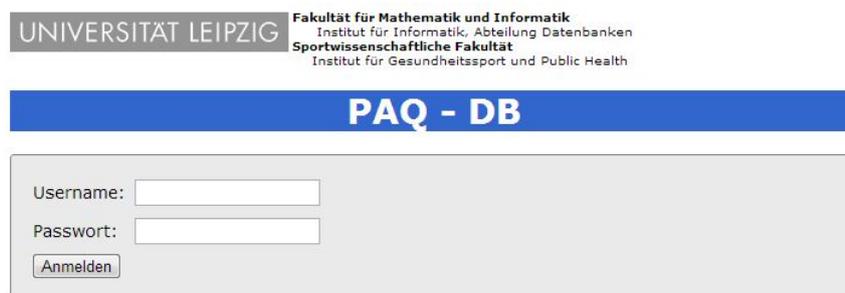
### 3 Integration der Daten

#### 3.1.4 WebGUI

Die *Graphic User Interface* (WebGUI) ist ein zentraler Bestandteil dieser Arbeit. Sie dient der Präsentation der gewonnenen Informationen und stellt diese dem Endnutzer in einer aufbereiteten Form dar. Daraus resultieren hohe Anforderungen an Bedienbarkeit und Funktionalität. Im Zuge der vorausgegangenen Besprechungen mit Mitarbeitern vom Institut für Gesundheitssport und Public Health der Universität Leipzig haben sich folgende Anforderungen ergeben:

- Login-Bereich mit Passwort und Nutzername
- Übersicht aller Fragebögen mit Filter- und Sortierfunktion
- Ausgabe aller Informationen zu jedem Fragebogen auf einer eigenen Seite
- PDF Ausgabe eines Fragebogens mit allen Informationen
- Suche nach Fragebögen durch Angabe von ICF Codes

Die WebGUI nutzt 4 verschiedene Seiten für die Darstellung der Informationen und den Zugriff auf die Seite. Ein Login Bereich ist für die Autorisierung des Nutzers vorhanden. Nur mit korrekten Anmeldeinformationen ist es möglich auf die folgenden Seiten zuzugreifen. In



UNIVERSITÄT LEIPZIG Fakultät für Mathematik und Informatik  
Institut für Informatik, Abteilung Datenbanken  
Sportwissenschaftliche Fakultät  
Institut für Gesundheitssport und Public Health

**PAQ - DB**

Username:

Passwort:

**Abbildung 3.4:** Login Bereich der WebGUI. Benötigt wird ein Nutzername und ein Passwort.

Abbildung 3.4 ist die Loginseite zu sehen. Für eine erfolgreiche Autorisierung werden ein Nutzername und ein Passwort benötigt. Es gibt eine Startseite, auf der sich eine

### 3 Integration der Daten

Übersichtstabelle für alle Fragebögen befindet. Weiterhin gibt es eine Detailseite, auf welcher alle Informationen eines Fragebogens dargestellt werden. Zuletzt gibt es noch eine PDF Ausgabe der Informationen der Fragebögen.

Die WebGUI wurde mittels *Hypertext Preprocessor* (PHP) und JavaScript entwickelt. Es handelt sich um eine Webseite mit dynamischen Inhalten. Alle Informationen werden zur Laufzeit aus der PA-Datenbank gelesen und anschließend auf der Webseite aufbereitet dargestellt. Die Zugriffe auf die PA-Datenbank wurden mit PHP Funktionen realisiert. Es werden SQL Abfragen generiert und an die PA-Datenbank gesendet. Die erhaltenen Daten werden in tabellarischer Form auf der Webseite ausgegeben. Für Nutzer der Webseite ist ausschließlich lesender Zugriff vorgesehen. Änderungen an den Daten können nur über entsprechende Datenbank Tools und Berechtigung vorgenommen werden, nicht aber über die Webseite. JavaScript findet Anwendung, um eine übersichtlichere Darstellung der Inhalte zu realisieren. Es wird genutzt, um die Übersichtstabelle auf der Startseite zu sortieren und die Detailseite mit Menüpunkten zu unterteilen.

Für die Erzeugung der PDF Ausgabe wird ein PHP Plugin genutzt. Es handelt sich um TCPDF<sup>1</sup>. Das Plugin ist ein Open Source Projekt und steht unter der *GNU Lesser General Public License* (LGPL) zur Verfügung. Der HTML Code der Webseite kann direkt an das Plugin übergeben werden. Daraus wird ein PDF Dokument erzeugt und im Browser ausgegeben.

In Abbildung 3.5 ist ein Screenshot der fertigen Webseite. Der Menüpunkt *Suchkriterien* ist aufgeklappt und alle Suchkriterien werden angezeigt. Es gibt verschiedene Arten von Suchkriterien. Es wird nach Texten, Zahlen und Boolean Werten unterschieden. Suchkriterien können dabei beliebig kombiniert werden.

In Abbildung 3.6 ist die Übersichtstabelle der Fragebögen zu erkennen. Es handelt sich hierbei um die Startseite mit der Übersicht aller Fragebögen. Es werden zunächst nur die wichtigsten Informationen angezeigt, um die Übersichtlichkeit zu erhalten. Es ist möglich nach Kurzform, Bezeichnung, Autor und Jahr zu sortieren. Das wird intuitiv über einen Klick auf den Kopf der Tabelle realisiert. Durch einen Klick auf den Namen eines Fragebogens, wird die Detailseite geöffnet. Diese stellt alle Informationen zu einen Fragebogen dar.

---

1 <http://www.tcpdf.org/>

### 3 Integration der Daten

UNIVERSITÄT LEIPZIG **Fakultät für Mathematik und Informatik**  
Institut für Informatik, Abteilung Datenbanken  
 Sportwissenschaftliche Fakultät  
 Institut für Gesundheitssport und Public Health

**PAQ - DB**

▼ Suchkriterien

Kurzform

Autor

Betrachtungszeitraum (Tage)

Jahr

Itemanzahl

Anzahl Dimensionen

MaxAltersgruppe

ICF Codes

Normwerte  ja  nein

Kosten  ja  nein

Öffentlich zugänglich  ja  nein

Indikation

Reliabilität

Kurzform ▲	Bezeichnung ▲	Autor ▲	Jahr ▲
7 Day PAR	7 Day Recall Physical Activity Recall	Sallis et al.	1985
ADL Q	Activity of daily living Questionnaire	Johnson et al.	2004
APAFOP	Assessment of Physical Activity in Frail Older People	Hauer et al.	2011

**Abbildung 3.5:** Übersichtsseite der WebGUI. Darstellung der Suchkriterien

In Abbildung 3.7 wird die Detailseite dargestellt. Die Unterteilung in die Kategorien *Kurzbeschreibung*, *Weitere Informationen* und *Literatur* ist oben zu erkennen. Unter jedem Menüpunkt werden dabei zugehörige Informationen angezeigt.

Die Abbildung 3.8 zeigt die PDF Ausgabe der Webseite. Alle Informationen werden in einer zweispaltigen Tabelle dargestellt. Die linke Spalte enthält den Namen, die rechte die zugehörigen Informationen. Jedes PDF Dokument wird dynamisch beim Aufruf erzeugt. Die enthaltenen Daten sind immer auf dem aktuellen Stand.

### 3 Integration der Daten

UNIVERSITÄT LEIPZIG **Fakultät für Mathematik und Informatik**  
 Institut für Informatik, Abteilung Datenbanken  
**Sportwissenschaftliche Fakultät**  
 Institut für Gesundheitssport und Public Health

## PAQ - DB

▸ Suchkriterien

Kurzform ▲	Bezeichnung	Autor	Jahr
7 Day PAR	7 Day Recall Physical Activity Recall	Sallis et al.	1985
ADL Q	Activity of daily living Questionnaire	Johnson et al.	2004
APAFOP	Assessment of Physical Activity in Frail Older People	Hauer et al.	2011
AQuAA	Activity Questionnaire for Adults and Adolescents	Chinapaw et al.	2009
Baecke	The Questionnaire of Baecke et al. for the Measurement of a Person 's Habitual Physical Activity	Baecke, J.A.H.	1982
Baecke Fragebogen	Deutsche Version des Erhebungsverfahrens von Baecke/Burema/Fritjers zur Erfassung der habituellen körperlichen Aktivität	Wagner et al.	2003
BRFSS	Behavioral Risk Factors Surveillance System	Center for Disease Control Collaboration	2011
CHAMPS	Community Healthy Activities Model Program for Seniors	Stewart et al.	2001
Dijon PAS	Dijon Physical Activity Score	Robert et al.	2004
EPIC Questionnaire SF	European Prospective Investigation into Cancer and Nutrition Physical Activity Questionnaire - Short Form	Pols et al.	1997
EPIC-Norfolk (EPAQ2)	European Prospective Investigation into Cancer and Nutrition Physical Activity Questionnaire from the Norfolk Cohorte	EPIC Collaboration	2001
FFKA	Freiburger Fragebogen zur körperlichen Aktivität	Frey et al.	1999
German PAQ 50+	German Physical Activity Questionnaire 50+	C. Huy et al.	2008
GLT EQ (english)	Godin Leisure Time Exercise Questionnaire	Godin & Shephard	1985

**Abbildung 3.6:** Übersichtsseite der WebGUI. Darstellung der tabellarischen Übersicht der Fragebögen

Für die Suche nach Fragebögen mit ICF Codes wird das Mapping zwischen ICF Konzepten und Fragen genutzt, welches durch den semi-automatischen Matchingvorgang generiert wird.

### 3 Integration der Daten

UNIVERSITÄT LEIPZIG **Fakultät für Mathematik und Informatik**  
Institut für Informatik, Abteilung Datenbanken  
**Sportwissenschaftliche Fakultät**  
Institut für Gesundheitssport und Public Health

**PAQ - DB**

**7 Day Recall Physical Activity Recall**

Kurzbeschreibung	Weitere Informationen	Literatur
<b>Kurzform</b>	7 Day PAR	
<b>Zielgruppe</b>	Gesunde Erwachsene mittleren und höheren Erwachsenenalters.	
<b>Autor</b>	Sallis et al.	
<b>Jahr</b>	1985	
<b>Bearbeitungszeit</b>	15 min	

[Übersicht als .pdf herunterladen](#)  
[zurück zur Übersicht](#)

**Abbildung 3.7:** Detailseite der WebGUI. Auflistung aller Informationen zu jedem Fragebogen.

## 3.2 GOMMA

GOMMA ist ein System zur Analyse und Management von Ontologien. Das Hauptaugenmerk liegt dabei auf Ontologien, welche sich mit Lebenswissenschaften befassen. Es bietet Funktionen zum effizienten Verwalten von Ontologieversionen und stellt eigene Komponenten für das Matching zwischen Ontologien bereit. Es werden außerdem Komponenten für den Import und die Analyse bereitgestellt. Damit werden alle Werkzeuge für das Matching zwischen Fragen und ICF von GOMMA zur Verfügung gestellt. GOMMA stellt ein eigenes Repository zur Verfügung, in welches die zu matchenden Quellen integriert werden können. Es ist sinnvoll die ICF, sowie die Fragebögen zunächst in das GOMMA Repository zu integrieren.

GOMMA wurde im Zuge des *Ontology Alignment Evaluation Initiative* (OAEI) getestet und hat sehr gute Ergebnisse erzielt.[5]

### 3 Integration der Daten

UNIVERSITÄT LEIPZIG	
Fakultät für Mathematik und Informatik Institut für Informatik, Abteilung Datenbanken Sportwissenschaftliche Fakultät Institut für Gesundheitssport und Public Health	
<b>7 Day Recall Physical Activity Recall</b>	
<b>Kurzform</b>	7 Day PAR
<b>Zielgruppe</b>	Gesunde Erwachsene mittleren und höheren Erwachsenenalters.
<b>Autor</b>	Sallis et al.
<b>Jahr</b>	1985
<b>Bearbeitungszeit</b>	15 min
<b>Konstrukt</b>	Energieumsatz
<b>Setting</b>	settingunabhängig
<b>Ziel des Fragebogens</b>	Erfassung von körperlicher Aktivität, die eine Aussage im Dosis-Wirkungsgefüge zulässt.
<b>Begründung der Konstruktion</b>	Nur wenige Fragebögen erfassen körperliche Aktivität und Aktivitätsverhalt systematisch und lassen somit kaum Rückschlüsse auf Gesundheitseffekte zu.
<b>Betrachtungszeitraum</b>	7 Tage
<b>Anzahl Dimensionen</b>	5
<b>Anzahl Fragen</b>	10
<b>MinAltersgruppe</b>	20
<b>MaxAltersgruppe</b>	74
<b>Auswertung Outcome</b>	Berechnung von MET x minutes x day-1
<b>Sprachen</b>	deutsch, englisch
<b>Auswertbarkeit</b>	Einfach und schnelle Auswertung durch Akkumulation von MET-Werten
<b>Kosten</b>	keine
<b>öffentlich zugänglich</b>	nein

Universität Leipzig  
Sportwissenschaftliche Fakultät  
Jahnallee 59  
04109 Leipzig  
www.uni-leipzig.de

Page 1/2

Abbildung 3.8: PDF Darstellung der WebGUI

#### 3.2.1 Integration der ICF

Die ICF liegt im OWL Format vor. Für dieses Format existiert in GOMMA bereits ein Importer. Dieser musste an einigen Stellen angepasst werden, damit alle relevanten Informationen der ICF berücksichtigt werden. So wurden zunächst nicht die Definitionen der ICF Konzepte erfasst. Dafür musste ein bisher nicht beachtetes Attribut erfasst werden. Unter Verwendung dieses Importers konnte die ICF Ontologie auf leichte Weise in GOMMA importiert werden. Der gesamte Importvorgang benötigte eine Zeit von ca. 3 min. Die Ontologie wurde dabei direkt aus der Webressource<sup>2</sup> geladen.

#### 3.2.2 Integration der Fragen

Die Fragen liegen aufgrund der Integration der Fragebögen bereits in der PA-Datenbank der Webseite. Die Fragen werden GOMMA intern als eigene Quelle repräsentiert. Dadurch ist es möglich die Matchingkomponente zu verwenden. Das GOMMA Repository nutzt ein komplexes internes System zur Verwaltung der Daten. Es ist sinnvoll die Integration über einen Importer zu realisieren. Aus diesem Grund werden die Fragen über das OBO Format in GOMMA importiert. Für dieses Format existiert ebenfalls schon ein Importer. Die Fragen müssen dafür in das OBO Format konvertiert werden. Dieses ist sehr übersichtlich. Jede Frage wird als eigenes Konzept betrachtet. Jedes Konzept wird mit einem *[Term]* Tag deklariert. Danach folgt eine eindeutige Id. Dafür bieten sich die Primärschlüssel der Datenbank an. Zuletzt wird der Text der Frage als Name gespeichert. Sind ICF und Fragen in GOMMA importiert, kann das Matching zwischen beiden erfolgen.

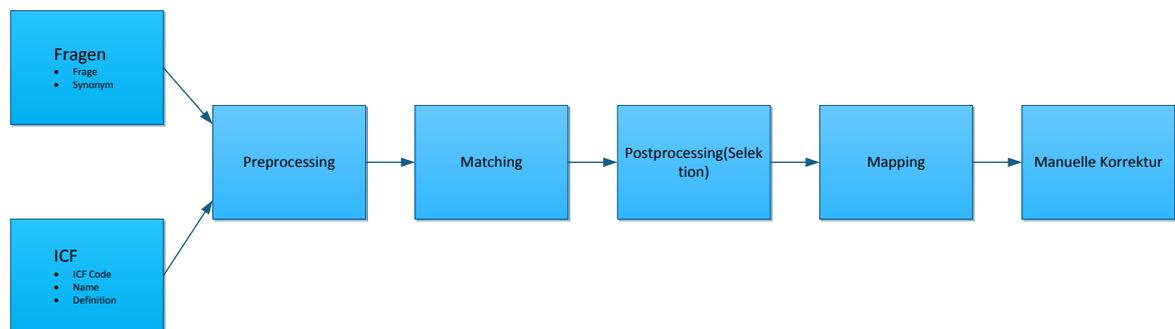
---

<sup>2</sup> <http://rest.bioontology.org/bioportal/ontologies/download/47404>

# 4 Matching der Fragebögen zur ICF

Im folgenden Kapitel wird der Ablauf des Matchings beschrieben. Die nötigen Arbeitsschritte werden vorgestellt und erläutert.

## 4.1 Workflow



**Abbildung 4.1:** Matching workflow

In Abbildung 4.1 wird der verwendete Matching Workflow abgebildet. Links stehen die Eingabedaten: die Menge aller Fragen und die ICF Ontologie. Beide sind bereits in das GOMMA Repository importiert und stehen für den Matchingvorgang zur Verfügung. Die Daten von jedem Konzept der ICF und jeder Frage der Fragebögen werden zunächst vorverarbeitet. Mit den aufbereiteten Daten wird der Matchingvorgang gestartet. Dabei werden für alle möglichen Konzeptkombinationen Ähnlichkeitswerte ermittelt. Die Ergebnisse werden in der anschließenden Nachverarbeitung ausgewertet. Aus der

## 4 Matching der Fragebögen zur ICF

Ergebnismenge können Korrespondenzen selektiert werden, welche eine angegebene Mindestähnlichkeit aufweisen.

Das Ergebnis ist das Mapping zwischen den Fragen und der ICF. Es wird genutzt, um weitere, bisher nicht gefunden Konzepte der ICF zu finden. Das Institut für Gesundheitssport und Public Health der Universität Leipzig, hat bereits ein Referenzmapping zwischen Fragen und ICF Ontologie erstellt. Es wurde in manueller Arbeit für jede Frage die zugehörigen ICF Konzepte ermittelt. Allerdings kann eine Vollständigkeit aufgrund der großen Anzahl von Konzepten nicht garantiert werden. Das automatisch erstellte Mapping wird genutzt um weitere, bisher nicht gefundene Korrespondenzen zu finden. Es wird ebenfalls von diesen Experten überprüft und eventuell fehlende Korrespondenzen können übernommen werden.

### 4.2 Vorverarbeitung

Für dieses Matchingproblem werden 2 Vorverarbeitungsschritte vorgenommen. Bei den Informationen der Konzepte handelt es sich ausschließlich um die Attribute: Name, Synonym und Definition. Jeder Vorverarbeitungsschritt muss für jedes Attribut vorgenommen werden. Im ersten Schritt wird für jeden Attributwert eine LowerCase Transformation vorgenommen. Dabei wird jeder gefundene Großbuchstabe durch seinen entsprechenden Kleinbuchstaben ersetzt, so dass eine einheitliche Formatierung der Strings erreicht wird.

Der zweite Schritt nimmt eine Delimiter & Stoppwortentfernung vor. Bei der Delimiterentfernung werden alle Satzzeichen entfernt. Satzzeichen haben keine inhaltliche Bedeutung. Die Stoppwortentfernung entfernt Wörter, die kaum inhaltliche Relevanz haben. Dazu zählen bestimmte und unbestimmte Artikel, Präpositionen und Konjunktionen. Zur Erkennung der Stoppwörter werden Stoppwortlisten genutzt. Diese enthalten Sammlungen der betreffenden Wörter für eine Sprache. In Abbildung 4.2 werden alle Vorverarbeitungsschritte anhand eines Beispiels demonstriert. Diese Frage stammt aus einem PA-Fragebogen. Durch die LowerCase Transformation werden zuerst alle Buchstaben kleingeschrieben. Im nächsten Schritt wird das Fragezeichen entfernt, sowie die Wörter *an*, *bei* und *sie*.

#### 4 Matching der Fragebögen zur ICF

Durch die Reduzierung der Strings auf inhaltlich relevante Wörter, wird die Ergebnisqualität gesteigert. Die Ähnlichkeit würde andernfalls stark verfälscht. Stoppwörter kommen in praktisch jedem Satz vor, wodurch eine hohe Ähnlichkeit entsteht. Die inhaltlich wichtigen Wörter verlieren dadurch an Bedeutung.



**Abbildung 4.2:** Vorverarbeitungsschritte. Anwendung einer LowerCase Transformation. Anschließend findet Delimiter & Stopword Normalisierung statt

### 4.3 Matching

Damit ein Matching vorgenommen werden kann, müssen alle möglichen Kombinationen für die Vergleiche ermittelt werden. Es werden folgende Daten verwendet:

#### 4 Matching der Fragebögen zur ICF

- Die Fragen der Fragebögen, sowie die Synonyme
- für das Matching relevante Daten der ICF Konzepte, dazu zählen Name und Definition

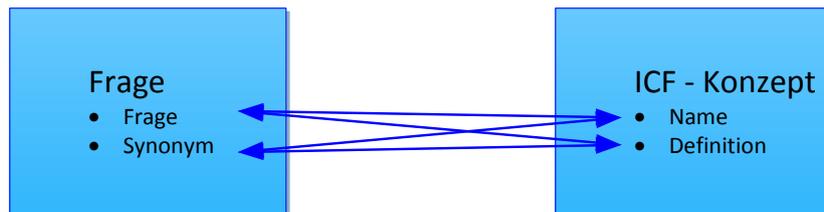


Abbildung 4.3: Matching Vergleiche

Bei allen Ausgangsdaten handelt es sich um Attributwerte mit einem oder mehreren Wörtern. In der Abbildung 4.3 sind die nötigen Vergleiche abgebildet. Es wird dabei das kartesische Produkt zwischen den Konzepten gebildet. Um die Ähnlichkeit zwischen 2 Konzepten zu ermitteln werden String Vergleiche durchgeführt. Die Ähnlichkeit wird in einem Zahlenwert zwischen 0 und 1 abgebildet, wobei 1 die maximale Ähnlichkeit darstellt. Der Wert 1 wird nur bei identischen Strings erreicht.

Die Fragen sollen mit ICF Konzepten annotiert werden. Es wird zu jeder Frage ein oder mehrere passende ICF Konzepte gesucht. Das bedeutet, es werden n:m Beziehungen gesucht.

Es gibt verschiedene Metriken zur Ähnlichkeitsermittlung. Diese Arbeit konzentriert sich auf 2 Metriken: Tf-idf und Trigram. Beides sind Token basierte Verfahren. Die Ergebnisse beider Verfahren sollen vergleichend evaluiert werden.

#### 4.4 Nachverarbeitung

Im Anschluss an das Matching erfolgt die Nachverarbeitung oder Postprocessing. Dieser Vorgang filtert die erhaltenen Korrespondenzen anhand ihrer Ähnlichkeiten. Dabei werden die drei im Kapitel 2 vorgestellten Selektionsstrategien verwendet. Threshold,

#### *4 Matching der Fragebögen zur ICF*

MaxDelta und MaxN. Das Ziel der Nachverarbeitung ist die Erstellung des fertigen Mappings.

## 5 Evaluation der Ergebnisse

Das Resultat des Matchings ist ein Mapping zwischen der ICF und den Fragebögen. Es sollen verschiedene Matchingverfahren angewendet und vergleichend evaluiert werden. Dadurch kann für dieses Matchingproblem das beste Verfahren ermittelt werden. Bei der Evaluierung sollen außerdem verschiedene Selektionsverfahren: Threshold, MaxDelta und MaxN analysiert werden. Anhand der verschiedenen Evaluationsanalysen können Schlussfolgerungen auf die Qualität der Mappings, welche mit verschiedenen Metriken erstellt wurden, gezogen werden. Um die Qualität zu berechnen, kann das vom Institut für Gesundheitssport und Public Health erstellte manuelle Referenzmapping genutzt werden. Ziel ist die Ermittlung des besten Mappings für die Verwendung der Vorschlags-generierung, so dass eventuell fehlende Korrespondenzen im Referenzmapping ergänzt werden können.

Das ausgewählte Mapping wurde für die Generierung von Vorschlägen verwendet, um weitere ICF Konzepte zu den Fragen zu finden. Bei der manuellen Evaluation wurden einige Vorschläge, welche bisher nicht im Mapping zu finden waren, in das Referenzmapping übernommen. Das daraus resultierende Mapping wurde in die PA-Datenbank integriert und findet für die Suche in der WebGUI Verwendung.

Eine exakte Ermittlung von Recall, Precision und F-Measure konnte durch Verwendung des manuell erstellten Referenzmappings erfolgen. Dieses wurde in GOMMA importiert und dient als Grundlage für die Überprüfung der Qualität des automatisch generierten Mappings.

## 5 Evaluation der Ergebnisse

Frage	ICF Konzept	korrekt
Leisure activities like Leisurely Sitting	Maintaining a sitting position	ja
Brisk Walking (10+mins in duration)	Walking long distances	ja
Do light work around the house (such as sweeping or vacuuming)	Household tasks	ja
Elimination	Elimination of faeces	nein

**Tabelle 5.1:** Korrespondenzen zwischen Fragen und ICF Konzepten.

### 5.1 Referenzmapping

Das Referenzmapping wurde manuell vom Institut für Gesundheitssport und Public Health der Universität Leipzig erstellt und zur Verfügung gestellt. Dieses enthielt 1158 Korrespondenzen zwischen ICF Ontologie und Fragen bevor die Vorschläge aus dem automatisch erstellten Mapping ergänzt wurden. Zur Verbesserung des Referenzmappings wurden die 100 besten Korrespondenzen aus dem automatisch erstellten Mapping ausgewählt. Es wurden dabei die 100 besten, als false-positiv klassifizierten Korrespondenzen verwendet. Dabei handelt es sich um Korrespondenzen, welche vom Matcher gefunden wurde aber nicht im Referenzmapping enthalten sind. Dieses sind die Kandidaten für eine Erweiterung des Referenzmappings. Diese 100 besten Korrespondenzen wurden bei Experten des Instituts für Gesundheitssport und Public Health überprüft. Als Ergebnis wurden 58 Korrespondenzen in das Referenzmapping ergänzt. Das daraus resultierende Mapping enthält 1216 Korrespondenzen. Die F-Measure steigt durch die ergänzten Korrespondenzen von 51,91% auf 56,67%. Zukünftig können noch weitere Korrespondenzen hinzukommen. Alle weiteren Analysen verwenden das erweiterte Mapping als Basis. Tabelle 5.1 zeigt einige Beispiele für gefundene Korrespondenzen. Die Texte zeigen deutlich, wie unterschiedlich die Quelldaten sind. Häufig wird nur ein Match für ein Token gefunden.

### 5.2 Ontologie & Fragebogen Analyse

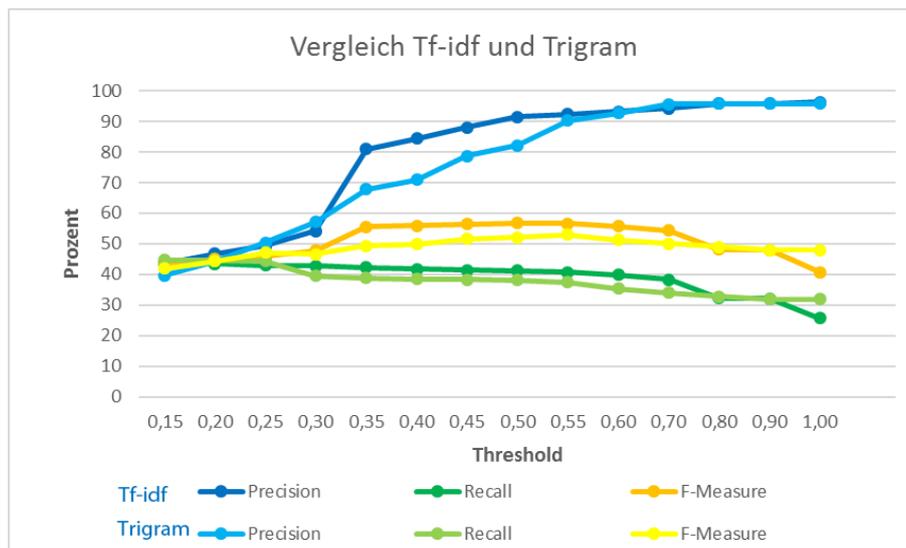
Für das Referenzmapping ergeben sich folgende statistische Auswertungen. Die Fragebögen enthalten insgesamt 805 Fragen. Das Mapping enthält Korrespondenzen für 677 Fragen. Das ergibt eine domain coverage von 84%. Die ICF Ontologie enthält 1594 Konzepte. Für 116 ICF Konzepte wurden Korrespondenzen gefunden. Die range coverage

## 5 Evaluation der Ergebnisse

beträgt 0,07%. Die Ursache liegt in der Spezialisierung der Fragebögen auf Physical Activity. Die Fragen wurden gegenüber der gesamten ICF gematcht. Eine Einschränkung auf Teilbereiche, welche mit körperlicher Aktivität in Verbindung stehen, würde die range coverage signifikant steigern. Dies würde zudem die Anzahl der Vergleiche reduzieren und somit die Laufzeit positiv beeinflussen. Ziel ist möglichst viele Korrespondenzen für jede Frage zu finden. Eine Einschränkung des Suchraumes – Blocking genannt - kann ebenfalls die Qualität des Mappings positiv beeinflussen.

### 5.3 Ergebnisse des Matchings

In diesem Abschnitt werden die Mappingergebnisse betrachtet. Dabei werden für alle Parameter die besten Werte empirisch ermittelt. Am Schluss soll die beste Konfiguration für dieses Matchingproblem festgestellt werden. Beachtung finden besonders das F-Measure und der Recall. Die Abbildung 5.1 zeigt einen Vergleich von Tf-idf und Trigram

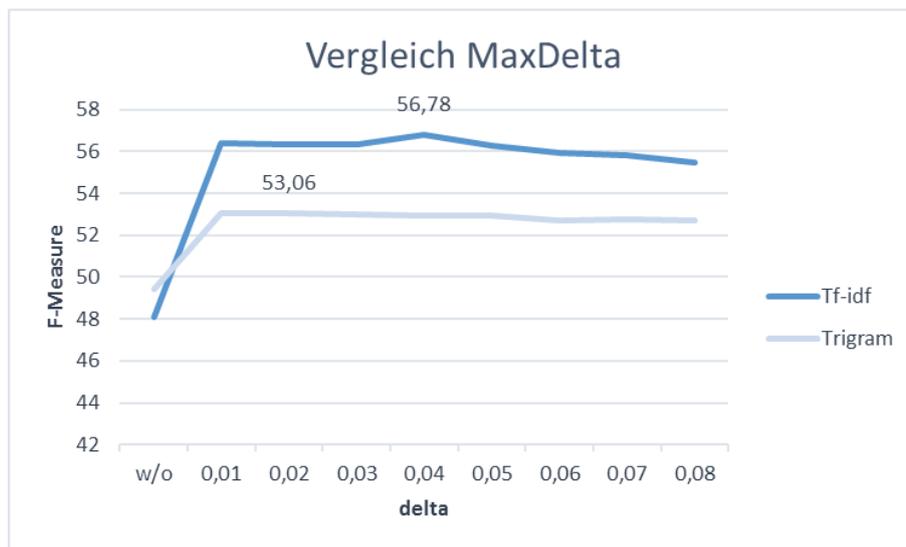


**Abbildung 5.1:** Vergleich Tf-idf mit Trigram. Betrachtung der Entwicklung der F-Measure bezüglich des Threshold

Verfahren bei gleichen Parametern. Es wurde ein MaxDelta von 0,04 gewählt. Diese Werte werden im Anschluss evaluiert. Tf-idf ist im Durchschnitt 2% besser als Trigram bezüglich der F-Measure. Die maximale Differenz zwischen beiden Matchern beträgt 6%. Beide

## 5 Evaluation der Ergebnisse

Verfahren liegen somit sehr dicht beieinander. Beide Matcher liefern die größte F-Measure bei einem Threshold von 0,55. Tf-idf mit einem Wert von 56,78%. Das Trigram Verfahren liefert eine F-Measure von 52,94%. Für die kommenden Betrachtungen wird ein Threshold von 0,55 zu Grunde gelegt.



**Abbildung 5.2:** Vergleich MaxDelta zwischen Tf-idf und Trigram. Bei w/o ist MaxDelta abgeschaltet.

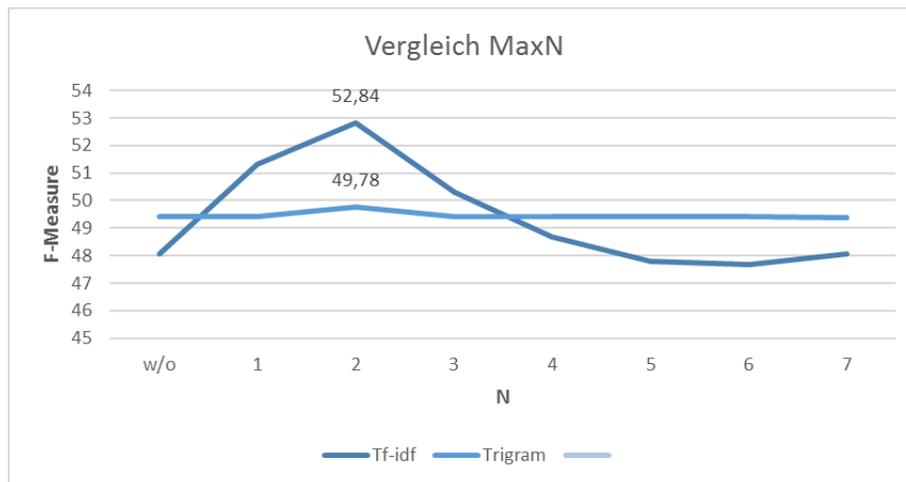
Abbildung 5.2 zeigt den Vergleich zwischen Tf-idf und Trigram Matching bei verschiedenen delta Werten. Der delta Wert w/o bedeutet, dass MaxDelta abgeschaltet ist. Die höchste F-Measure wird bei Tf-idf mit einem delta von 0,04 erreicht. Trigram erreicht seine maximale F-Measure von 53,06% bei einem delta von 0,02.

In Abbildung 5.3 ist ein Vergleich von MaxN bei beiden Verfahren zu sehen. Auch hier liefert Tf-idf stets die besseren Ergebnisse. Tf-idf hat die beste F-Measure bei  $N = 2$  mit 52,84%. Trigram erreicht seine beste F-Measure von 49,78 ebenfalls bei  $N = 2$ .

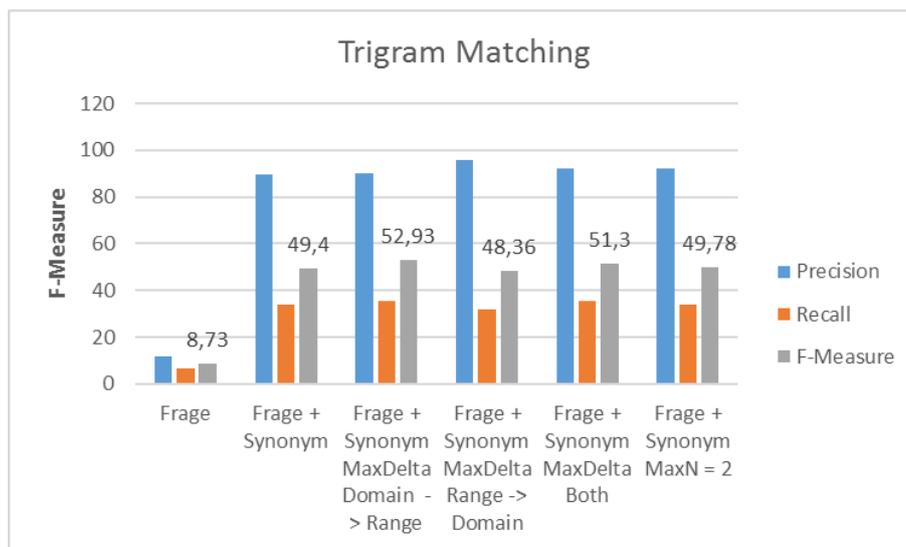
Abbildung 5.4 vergleicht 6 verschiedene Konfigurationen unter Verwendung der Trigram Metrik. Die beste F-Measure wird bei MaxDelta mit Richtung Domain Range erreicht. Die F-Measure ist 3,5% schlechter verglichen zu Tf-idf.

Die Abbildung 5.5 vergleicht verschiedene 6 verschiedene Konfigurationen unter Anwendung von Tf-idf Matching. Es werden jeweils die besten Ergebnisse untersucht. Beim ersten Versuch wurden nur die Fragen für das Matching verwendet. Synonyme wurden

## 5 Evaluation der Ergebnisse



**Abbildung 5.3:** Vergleich MaxN zwischen Tf-idf und Trigram. Bei 0 ist MaxN abgeschaltet.



**Abbildung 5.4:** Vergleich der Parameter unter Verwendung der Trigram Metrik

nicht beachtet. Eine Selektion durch MaxDelta und MaxN wurde nicht angewendet. Die F-Measure beträgt 7,38%. Das Problem ist der geringe Recall von 4,44%. Außerdem liegt die Precision bei nur 21,86%. Das bedeutet, es werden nur sehr wenige Korrespondenzen gefunden. Für den zweiten Versuch werden zusätzlich die Synonyme betrachtet. Daraus resultiert eine F-Measure von 48,08%. Die Synonyme haben somit einen sehr großen Einfluss auf die Qualität des Matchings. Die nächsten 3 Versuche untersuchen die Ver-

## 5 Evaluation der Ergebnisse

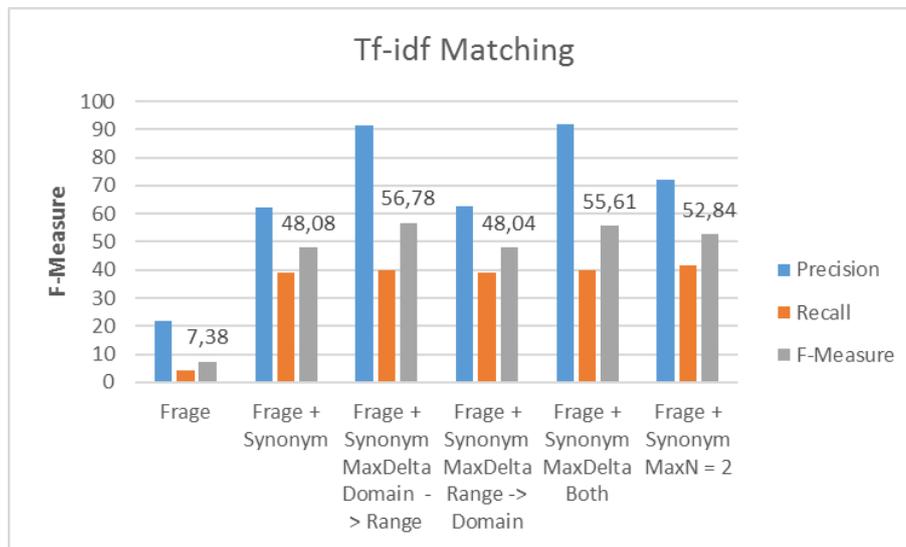


Abbildung 5.5: Vergleich der Parameter unter Verwendung von Tf-idf Matching

wendung von MaxDelta. Es wird dabei die Richtung von der Domain (Fragen) zur Range (ICF) beachtet. Die F-Measure beträgt bei dieser Richtung 56,78%. Bei Versuch 4 wird die andere Richtung untersucht: Range zu Domain. Dieser Wert ist 8,74% geringer als bei der entgegengesetzten Richtung. Versuch 5 betrachtet MaxDelta Both. Dies entspricht der Erwägung, dass jeder Frage ein oder weniger ICF Konzepte zugeordnet sind. Allerdings können ICF Konzepte durchaus zu vielen verschiedenen Fragen zugeordnet werden. Der letzte Versuch zeigt die Werte bei Verwendung von MaxN. Die beste F-Measure ist 56,78% und wird bei Verwendung von MaxDelta mit der Richtung Domain zu Range erreicht. Es zeigt sich, dass beide Verfahren die besten Ergebnisse unter Verwendung von Fragen und Synonymen mit MaxDelta haben.

Beide Verfahren liefern stabile Ergebnisse. Im Mittel ist der Unterschied zwischen Tf-idf und Trigram unter Verwendung der besten Konfiguration bei 2,05%. Die beste F-Measure beträgt 56,78%. Sie wird erreicht unter Verwendung von Tf-idf mit MaxDelta (delta = 0,04) aus Richtung Domain zu Range. Allgemein ist die F-Measure mit einem Maximum von 56,78% relativ schlecht. Die Ursache dafür liegt in den großen Unterschieden der Quelldaten. Der größte Anteil der nicht gefundenen Korrespondenzen sind Strings ohne übereinstimmende Token. Ein Match ist dadurch mit den verwendeten Verfahren nicht zu erreichen. Eine weitere Ursache sind die verschiedenen Sprachen der Fragebögen. Der größte Teil der Fragen sind in englischer Sprache. Die verwendete ICF Ontologie liegt

## 5 Evaluation der Ergebnisse

ebenfalls in englischer Sprache vor. Ein kleiner Teil, ca. 10%, der Fragen sind allerdings in deutscher Sprache verfasst. Ein Match zwischen diesen Daten wird dadurch erschwert. Um diesen Fehler zu beseitigen muss eine vorherige Übersetzung der Fragen erfolgen. Dieses könnte in zukünftigen Arbeiten realisiert werden.

Eine Übersetzung der Quellen ist nicht vorgesehen, da der Originallaut der Fragen erhalten bleiben soll. Dadurch können die Fragebögen exakt aus der Datenbank rekonstruiert werden.

### 5.4 Schlussfolgerung

Tf-idf hat sich für dieses Matchingproblem unter Betrachtung von Tf-idf und Trigram Verfahren als besser erwiesen. Der Threshold beträgt dabei 0,55. Ein höherer Wert erzeugt eine bessere Precision zulasten des Recall. Dieser spielt aber eine wichtige Rolle für das Mapping, da es der Vorschlagsgenerierung dienen soll.

MaxDelta verbessert die F-Measure weiter. Dieser große Unterschied entsteht durch die hohe Anzahl an Korrespondenzen mit geringer Ähnlichkeit. Es ist anzunehmen, dass viele Korrespondenzen mit geringen Ähnlichkeiten gefunden wurden, da nur wenige Wörter zwischen Fragen und ICF überlappen. Durch Verwendung von MaxDelta werden diese Korrespondenzen ebenfalls erkannt und in das Mapping aufgenommen. Wie sich zeigt, spielt dieser Parameter eine wichtige Rolle für die Ergebnisqualität. Das beste Ergebnis wurde unter Betrachtung der Richtung, Domain zu Range erreicht. MaxN hat die Ergebnisse verbessert, konnte die Qualität von MaxDelta allerdings nicht erreichen.

Das Hauptproblem ist, dass nur wenige Token zwischen Fragen und ICF überlappen. Dieser Umstand kann in zukünftigen Arbeiten durch Verwendung von Hintergrundwissen wie Synonym-Datenquellen z.B. *Unified Medical Language System (UMLS)*<sup>1</sup> minimiert werden. Eine mögliche Fehlerquelle zeigt sich unter Betrachtung der Ausgangsdaten. 10% der Fragen sind in deutscher Sprache verfasst. Für das Matching wird hingegen die englische Version der ICF verwendet. Eine Übersetzung der Fragen in die englische Sprache könnte diese Fehlerquelle beseitigen. Multilingualität ist für viele Matchingvorgänge ein großes Problem. Die OAEI[10] hat einen eigenen Wettbewerb für Multilingualität. Daran lässt sich

---

1 <http://www.nlm.nih.gov/research/umls/>

## *5 Evaluation der Ergebnisse*

erkennen, dass die Verbesserung der Ergebnisse für Multilinguales Matching vorhanden ist und es noch Forschungsbedarf auf diesem Feld gibt.

## 6 Zusammenfassung & Ausblick

Das Ziel dieser Arbeit war die Bereitstellung einer zentralen Ablage für PA-Fragebögen. Dafür wurden ICF Ontologie und PA-Fragebögen in einer PA-Datenbank integriert. Ein weiterer Schwerpunkt lag im öffentlichen Zugriff auf die erfassten Informationen. Es wurde eine WebGUI entwickelt. Diese Webseite nutzt alle Informationen der PA-Datenbank. Sie bietet Zugriff auf alle Fragebögen. Dabei hatten Suchkriterien und Sortierfunktionen einen besonderen Stellenwert. Es wurden Suchmasken nach gängigen Informationen implementiert. Dazu zählen Autor, Veröffentlichungsdatum, Sprache und einige weitere Kriterien. Zur weiteren Verbesserung der Suche wurde eine Annotation der Fragebögen mit Konzepten der ICF Ontologie vorgenommen. Es wurde ein semi-automatisches Matching durchgeführt und die Ergebnisse im Anschluss manuell evaluiert. Das resultierende Mapping wurde anschließend in die PA-Datenbank übernommen. Dadurch besteht die Möglichkeit, Fragebögen über die ICF zu finden. Die entwickelte WebGUI kommt bereits am Institut für Gesundheitssport und Public Health der Universität Leipzig zum Einsatz. Sie erleichtert die Arbeit mit Fragebögen und stellt die Basis für weitere Forschung mit Fragebögen bereit. Durch die Annotation der Fragen mit der ICF wird die Suche innerhalb der WebGUI beschleunigt. Die Erstellung der WebGUI ist die Grundlage für zukünftige wissenschaftliche Arbeiten mit den PA-Fragebögen. Es bietet sich an, den Umfang der Datenbasis deutlich zu erweitern. Fragebögen mit Schwerpunkten auf anderen medizinischen Bereichen können integriert werden.

Für dieses Matchingproblem hat sich das Tf-idf Verfahren als bestes erwiesen. In[2] wurde ebenfalls eine Evaluierung zwischen verschiedenen Metriken für String Ähnlichkeiten vorgenommen. Auch dort schneidet Tf-idf am besten ab. Durch Verwendung von MaxDelta wurden die Ergebnisse weiter verbessert. Aber auch an den verwendeten Matchingverfahren können weitere Verbesserungen vorgenommen werden. Dabei sollte das Hauptaugenmerk zunächst auf eine einheitliche Sprache der Fragen gelegt werden. Eine Übersetzung der deutschen Fragen in englische Sprache kann die Matchingqualität verbessern. Weiter-

## *6 Zusammenfassung & Ausblick*

hin können andere Matchingverfahren angewandt werden. Ein Strukturmatcher, welcher die Hierarchie der ICF berücksichtigt, kann das Matchingergebnis verbessern. Aber auch ein Matchingverfahren, welches Hintergrundwissen aus anderen Ontologien verwendet, bietet sich an. Dabei handelt es sich um indirekte Matchingverfahren wobei ICF und die Fragen zunächst gegen eine andere Ontologie gematcht werden. Dadurch können beide Ontologien mit weiteren Hintergrundinformationen angereichert werden und ein Matching zwischen beiden könnte deutlich mehr Übereinstimmungen finden.

# Literaturverzeichnis

- [1] P.-. CAD. International classification of functioning, disability and health (ICF). 2001.
- [2] W. W. Cohen, P. Ravikumar, S. E. Fienberg, et al. A comparison of string distance metrics for name-matching tasks. In *Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web (IIWeb-03)*, pages 73–78, 2003.
- [3] V. Della Mea, A. Simoncello, et al. An ontology-based exploration of the concepts and relationships in the activities and participation component of the international classification of functioning, disability and health. *Journal of Biomedical Semantics*, 3(1):1–9, 2012.
- [4] J. Euzenat and P. Shvaiko. *Ontology matching*, volume 18. Springer Berlin, 2007.
- [5] A. Groß, M. Hartung, T. Kirsten, and E. Rahm. GOMMA results for OAEI 2012. In *Proc. of Ontology Matching Workshop, International Semantic Web Conference (ISWC), November*, volume 11, 2012.
- [6] T. R. Gruber et al. A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2):199–220, 1993.
- [7] S. Karstens and I. Froböse. ICF-Kategorien des XSMFA im Abgleich zum Core-Set Arthrose. *18. Reha-Wissenschaftliches Kolloquium*, 2009.
- [8] T. Kirsten, A. Gross, M. Hartung, E. Rahm, et al. GOMMA: a component-based infrastructure for managing and analyzing life science ontologies and their evolution. *Journal of biomedical semantics*, 2(6), 2011.

### Literaturverzeichnis

- [9] D. L. McGuinness, F. Van Harmelen, et al. OWL web ontology language overview. *W3C recommendation*, 10(2004-03):10, 2004.
- [10] C. Meilicke, R. Garcia-Castro, F. Freitas, W. Robert van Hage, E. Montiel-Ponsoda, R. Ribeiro de Azevedo, H. Stuckenschmidt, O. Šváb-Zamazal, V. Svátek, A. Taminin, et al. Multifarm: A benchmark for multilingual ontology matching. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2012.
- [11] E. Rahm and P. Bernstein. A survey of approaches to automatic schema matching. *the VLDB Journal*, 10(4):334–350, 2001.
- [12] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.
- [13] U. Trampisch, P. Platen, I. Burghaus, A. Moschny, S. Wilm, U. Thiem, and T. Hinrichs. Reliabilität des PRISCUS-PAQ. *Zeitschrift für Gerontologie und Geriatrie*, 43(6):399–406, 2010.
- [14] E. Ukkonen. Approximate string-matching with q-grams and maximal matches. *Theoretical Computer Science*, 92(1):191–211, 1992.
- [15] N. Wollmerstedt, S. Kirschner, D. Böhm, H. Faller, A. König, et al. Design and evaluation of the Extra Short Musculoskeletal Function Assessment Questionnaire XSMFA-D. *Zeitschrift für Orthopädie und ihre Grenzgebiete*, 141(6):718, 2003.

# Abkürzungsverzeichnis

**CSV** Comma separated Value

**DIMDI** Deutsches Institut für Medizinische Dokumentation und Information

**GOMMA** Generic Ontology Matching and Mapping Management

**ICD** International Classification of Diseases

**ICF** International Classification of Functioning, Disability and Health

**LGPL** GNU Lesser General Public License

**OAEI** Ontology Alignment Evaluation Initiative

**OBO** Open Biomedical Ontologies

**OWL** Web Ontology Language

**PA** Physical Activity

**PHP** Hypertext Preprocessor

**RDF** Resource Description Framework

**SUMO** Suggested Upper Merged Ontology

**UMLS** Unified Medical Language System

**WebGUI** Graphic User Interface

## *Abkürzungsverzeichnis*

**WHO** World Health Organisation

**XML** Extensible Markup Language

# Abbildungsverzeichnis

2.1	Auszug aus Priscus PAQ . . . . .	8
2.2	Auszug der deutschen ICF . . . . .	11
3.1	Workflow Diagramm aller Arbeitsschritte . . . . .	17
3.2	Excel Vorlage für den Import . . . . .	19
3.3	Datenbankentwurf . . . . .	20
3.4	Login Bereich der WebGUI . . . . .	22
3.5	Übersicht der WebGUI . . . . .	24
3.6	Tabellarische Übersicht der Fragebögen . . . . .	25
3.7	Detailseite der WebGUI . . . . .	26
3.8	PDF Darstellung der WebGUI . . . . .	27
4.1	Matching Workflow . . . . .	29
4.2	Vorverarbeitungsschritte . . . . .	31
4.3	Matching Vergleiche . . . . .	32
5.1	Vergleich Tf-idf mit Trigram gegenüber Threshold . . . . .	36
5.2	Vergleich MaxDelta zwischen Tf-idf und Trigram . . . . .	37
5.3	Vergleich MaxN zwischen Tf-idf und Trigram . . . . .	38
5.4	Vergleich der Parameter unter Verwendung von der Trigram Metrik . . . . .	38
5.5	Vergleich der Parameter unter Verwendung von Tf-idf Matching . . . . .	39

# Tabellenverzeichnis

2.1	Einteilung der Korrespondenzen in 4 Kategorien. . . . .	15
5.1	Korrespondenzen zwischen Fragen und ICF Konzepten. . . . .	35

# Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, dass alle Stellen der Arbeit, die wörtlich oder sinngemäß aus anderen Quellen übernommen wurden, als solche kenntlich gemacht und dass die Arbeit in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegt wurde.

Ort, Datum

Unterschrift