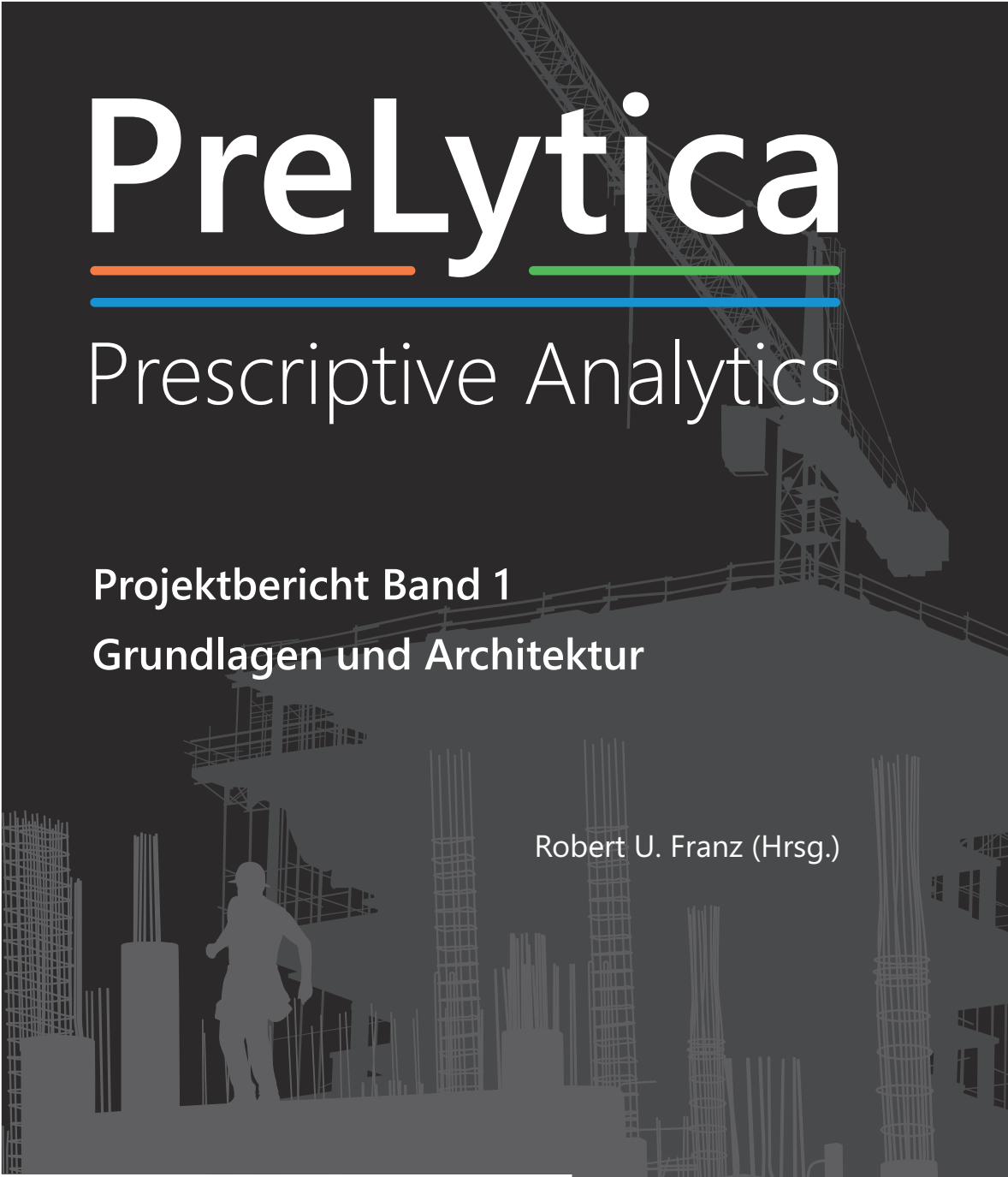


# PreLytica



## Prescriptive Analytics

### Projektbericht Band 1 Grundlagen und Architektur

Robert U. Franz (Hrsg.)

Berichte aus der Informationstechnik

**Robert U. Franz (Hrsg.)**

# **PreLytica**

Band 1  
Grundlagen, Architektur

Shaker Verlag  
Aachen 2019

**Bibliografische Information der Deutschen Nationalbibliothek**

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

Copyright Shaker Verlag 2019

Alle Rechte, auch das des auszugsweisen Nachdruckes, der auszugsweisen oder vollständigen Wiedergabe, der Speicherung in Datenverarbeitungsanlagen und der Übersetzung, vorbehalten.

Printed in Germany.

ISBN 978-3-8440-6446-9

ISSN 1610-9406

Shaker Verlag GmbH • Postfach 101818 • 52018 Aachen

Telefon: 02407 / 95 96 - 0 • Telefax: 02407 / 95 96 - 9

Internet: [www.shaker.de](http://www.shaker.de) • E-Mail: [info@shaker.de](mailto:info@shaker.de)

# Vorwort

Die Technische Hochschule Brandenburg (THB) hat sich in ihrem Strukturentwicklungsplan 2020 dem Motto „Wissen für die Region“ verschrieben. Die anwendungsorientierte Forschung mit dem Ziel einen Wissenstransfer aus der Hochschule in die Praxis hinein zu realisieren, hat daher zu zahlreichen Projekten geführt, die eine enge Kooperation und Vernetzung mit den Unternehmen der Region vorantreiben. Ebenso führen Forschungsergebnisse immer wieder zu Ausgründungen, die die Region wirtschaftlich stärken und zukunftsfähig machen.

Bei der Realisierung von Forschungsprojekten ist es an den Hochschulen immer wieder erforderlich, Drittmittel einzuwerben, um neue Themen mit ausreichend Mitarbeitern und Sachmitteln angehen zu können. Auf der Grundlage der Richtlinie des Ministeriums für Wissenschaft, Forschung und Kultur (MWFK) des Landes Brandenburg konnte zur Förderung der „Stärkung der technologischen und anwendungsnahen Forschung an Wissenschaftseinrichtungen im Land Brandenburg“ (StaF-Richtlinie) das Projekt „Präskriptive Analytik für KMU in Brandenburg (PreLytica)“ am Fachbereich Wirtschaft der Technischen Hochschule Brandenburg am Lehrstuhl „Betriebswirtschaftliche Anwendungen der Informatik“ gegründet werden.

Das Projekt PreLytica soll die systematische Erfassung, qualitätsgesicherte Integration und entscheidungsrelevante Auswertung von Massendaten für kleine und mittelständische Unternehmen (KMU) ermöglichen. Mit Hilfe präskriptiver Analytik können große, teilweise unstrukturierte Datenmengen zusammengefasst, analysiert und für konkrete Entscheidungen nutzbar gemacht werden. Dazu werden Verfahren des maschinellen Lernens, der Signal- und Bildverarbeitung und der Statistik angewandt. Mit dem anwendungsnahen Vorhaben sollen die wissenschaftlichen Grundlagen für die Analyse und Visualisierung unterschiedlicher Datenquellen geschaffen werden, auf deren Basis KMU im Land Brandenburg ihre Geschäftsprozesse digitalisieren können.

Da das Projekt durch die Europäische Union im Rahmen des Europäischen Fonds für Regionale Entwicklung (EFRE) gefördert wird, soll dieses Buch dazu beitragen,

dass die Projektergebnisse einer möglichst breiten Öffentlichkeit zur Verfügung stehen. Es wird dargelegt, wie das Projekt organisiert ist, welche Projektziele verfolgt werden, mit welchen Verfahren Erkenntnisse gewonnen werden und die Ergebnisse sollen dokumentiert werden.

Der geplante Projektzeitraum beträgt drei Jahre. Alle Projektarbeiten finden zwischen Dezember 2016 und November 2019 statt. Damit das Projekt umfassend dokumentiert werden kann, ist daher eine zweistufige Form der Veröffentlichung gewählt worden. In dem vorliegenden Band 1 sollen vor allem Grundlagen, Architektur und Anwendungen vorgestellt werden. Dieser Band 1 entstand etwa zur Halbzeit der Projektlaufzeit. Zum Ende des Projekts ist dann ein abschließender Band 2 geplant, in dem auf die Projektergebnisse und Auswertungen eingegangen wird.

## Kontakt

**Prof. Dr. Robert U. Franz**

Technische Hochschule Brandenburg

Fachbereich Wirtschaft

Magdeburger Str. 50

14770 Brandenburg an der Havel

Internet: [www.th-brandenburg.de/mitarbeiterseiten/fbw/robert-franz/](http://www.th-brandenburg.de/mitarbeiterseiten/fbw/robert-franz/)

E-Mail: [robert.franz@th-brandenburg.de](mailto:robert.franz@th-brandenburg.de)

# Inhaltsverzeichnis

|                                                           |     |
|-----------------------------------------------------------|-----|
| Einleitung – Projekt PreLytica .....                      | 1   |
| Screen Scraping .....                                     | 17  |
| Sports Performance Insight                                |     |
| Systemarchitektur .....                                   | 69  |
| Spielerakte und Datenschutz .....                         | 141 |
| Web-Portal .....                                          | 161 |
| Mobil-App .....                                           | 181 |
| Solar Potential Analytics                                 |     |
| Leistungsfähigkeit von Solarstromanlagen .....            | 193 |
| Nutzen, Vorteile und Optimierung des Energieertrags ..... | 205 |
| Timeline von Publikationen und Pressemitteilungen .....   | 227 |
| Beiträge zur 19. NWK .....                                | 229 |
| Pressemitteilungen .....                                  | 239 |
| Transfer in die Lehre .....                               | 251 |
| Zusammenfassung der Dissertation von André Nitze .....    | 253 |
| Zusammenfassung und Ausblick .....                        | 255 |



# Projekt PreLytica

## Informationen zu Projektzielen und -organisation

Robert U. Franz

### Inhalt

|                                                                    |    |
|--------------------------------------------------------------------|----|
| Projektmotivation .....                                            | 2  |
| Projektziele .....                                                 | 3  |
| Projektorganisation .....                                          | 5  |
| Teilprojekt: Sport Performance Insight (SPI)                       |    |
| Integration und Auswertung sportmedizinischer Leistungsdaten ..... | 8  |
| Teilprojekt: Solar Potential Analytics (SPA)                       |    |
| Auswertung heterogener Datenquellen .....                          | 8  |
| Projekt Sport Performance Insight – Hintergründe .....             | 10 |
| Projektpartner .....                                               | 10 |
| Konzept der Spielerakte .....                                      | 13 |
| Quellen .....                                                      | 15 |



## Projektmotivation

Im Rahmen der Digitalisierung aller wirtschaftlichen und gesellschaftlichen Bereiche werden immer schneller neue strukturierte und unstrukturierte Daten erzeugt. Daten werden dabei als eine formalisierte Repräsentation von Informationen verstanden, die zum Treffen von Entscheidungen genutzt werden können. Gleichzeitig werden Konsumenten von Daten (über die damit verbundenen Dienste) zunehmend auch zu Produzenten („Prosumers“). Und die Daten selbst werden immer mehr zum integralen Bestandteil von Produkten und Dienstleistungen. Diese beiden Entwicklungen führen nicht nur zur Erweiterung bestehender Geschäftsmodelle, sondern ermöglichen auch völlig neue digitale Geschäftsmodelle.

Die Nutzbarkeit vieler Daten hängt jedoch stark davon ab, ob und inwieweit sie automatisiert verarbeitet werden können. In vielen Organisationen können Informationsbedarfe auf individueller oder gruppenspezifischer Ebene nur selten mit vertretbarem Aufwand befriedigt werden. In den meisten Fällen wären dazu Integrationsprojekte nötig, die gleichermaßen technische Fachkompetenz und Domänenwissen erfordern. Zielstellung ist dabei oft die Anbindung externer Systeme mit strukturierten Zugriffsmöglichkeiten auf bestimmte Informationsressourcen. Durch die üblicherweise sehr heterogenen Anwendungslandschaften eröffnen sich hier zahlreiche Herausforderungen in Bezug auf Formate, Strukturiertheit, Persistenz und Automatisierbarkeit. Querschnittsthema ist dabei immer die Sicherheit. Diesen Herausforderungen soll im Rahmen des hier geschilderten Vorhabens begegnet werden.

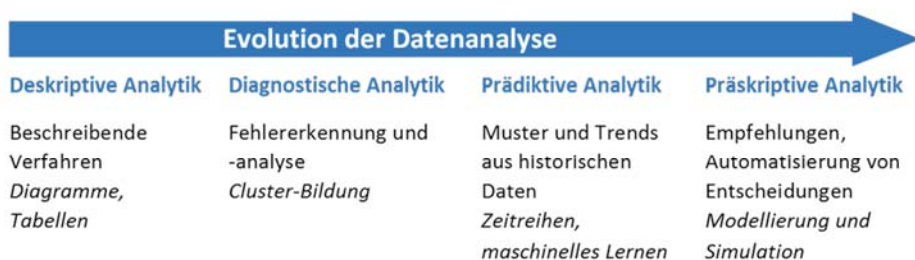


Abb. 1: Evolution der Datenanalyse von deskriptiver bis präskriptiver Analytik mit jeweiligen Verfahren

Neben der Verfügbarkeit von Datenquellen und der Qualitätssicherung der eigentlichen Daten spielt die Art der Datenanalyse eine wichtige Rolle für die Wertschöpfung, die damit erzeugt werden kann. Es hat in den vergangenen Jahren eine Evolution der Datenanalyse und seiner Werkzeuge stattgefunden (vergl. Abbildung 1). Grundlage ist die beschreibende Analytik, in der Tabellen und Diagramme zur Darstellung von Rohdaten genutzt werden. Ein erster Evolutionsschritt wurde in Richtung diagnostischer Analytik vollzogen, mit der Fehlererkennung durch Cluster-Bildung möglich wurde. Mit dem Wunsch der Vorhersage zukünftiger Ereignisse anhand historischer Daten gewann die prädiktive Analytik an Bedeutung, in der beispielsweise Zeitreihen und neuronale Netze eingesetzt wurden, um Trends und Muster vorherzusagen, zum Beispiel in den Anrufzahlen von Call-Centern zur besseren Personalbedarfsplanung oder die Stauvermeidung durch Verkehrsleitsysteme mittels Daten von Sensoren aus Autos und Straßen. Heute stehen Forschung und Wirtschaft vor der Herausforderung, sehr große und nur teilweise strukturierte Datenmengen in einer prädiktiven Analytik zu untersuchen, in der Fotos, historische Datenbanken und Echtzeitdaten zusammen zu Aussagen führen, die für Entscheidungen im Rahmen betrieblicher Prozesse eingesetzt werden können. Die verschiedenen Formen der Datenanalyse unterscheiden sich außerdem in ihrem Level der Schwierigkeit was die einzelnen Methoden und Schritte angeht, wobei aber auch die Aussagekraft der abgeleiteten Erkenntnisse steigt.

## Projektziele

Das Projekt „Präskriptive Analytik für KMU in Brandenburg“ (PreLytica) soll die systematische Erfassung, qualitätsgesicherte Integration und entscheidungsrelevante Auswertung von Massendaten für kleine und mittelständische Unternehmen (KMU) ermöglichen. Mit Hilfe präskriptiver Analytik können große, teilweise unstrukturierte Datenmengen zusammengefasst, analysiert und für konkrete Entscheidungen nutzbar gemacht werden. Dazu werden Verfahren des maschinellen Lernens, der Signal- und Bildverarbeitung und der Statistik angewandt. Mit dem anwendungsnahe Vorhaben sollen die wissenschaftlichen Grundlagen für die Analyse und Visualisierung unterschiedlicher Datenquellen geschaffen werden, auf deren Basis KMU im Land Brandenburg ihre Geschäftsprozesse digitalisieren können.



Abb. 2: Thematische Ausrichtung des Forschungsvorhabens

Das Konzept ist prototypisch in zwei Szenarien umgesetzt, in denen Daten integriert, unter Nutzung präskriptiver Analyseverfahren aufbereitet und visualisiert werden. Das erste Szenario behandelt die Entwicklung eines Analyse-Systems für die individuelle Auswertung und das Benchmarking personenbezogener sportmedizinischer Daten (vergl. Kapitel SPI, ab S. 69). Das zweite Szenario beinhaltet die Entwicklung eines Prognosemodells für die Installation von dezentralen Photovoltaik (PV)-Anlagen auf beliebigen Dachflächen im Land Brandenburg unter Berücksichtigung von Milieudaten (vergl. Kapitel SPA, ab S. 193).

Herausforderungen des Projekts sind zum einen die Erforschung passender Analyseverfahren, die auf die verschiedenartigen Datenquellen anwendbar sind und zum anderen die Entwicklung geeigneter Technologien zur Verarbeitung großer Datenmengen.

Ziel des Vorhabens ist die Nutzbarmachung unterschiedlichster Datenquellen in Form von strukturierten Diensten (Webservices) und deren automatisierte Auswertung und Visualisierung auf Web-basierten Informationsportalen. Diese sollen mit einer möglichst einfach zu bedienenden grafischen Oberfläche ausgestattet sein, die den Nutzer durch den Prozess der Auswahl, Bereinigung und Auswertung der gewünschten Datenquellen führt (ähnlich zu Installationsassistenten oder „Wizards“). Dadurch soll das Erstellen von Daten-Analysen auch für Fachabteilungen

und andere Personen ohne technisches Fachwissen möglich sein. Diese strukturierte Bereitstellung ermöglicht die Weiternutzung der Daten in allen IT-Systemen, die standardisierte Datenaustauschformate unterstützen (vgl. Abbildung 3). Die Systeme werden dadurch zudem voneinander entkoppelt. Systeme, die im Paradigma einer Service-orientierten Architektur (SOA) entwickelt wurden, profitieren direkt von diesem Ansatz. Die meisten anderen können mit geringem Aufwand dafür geöffnet werden.

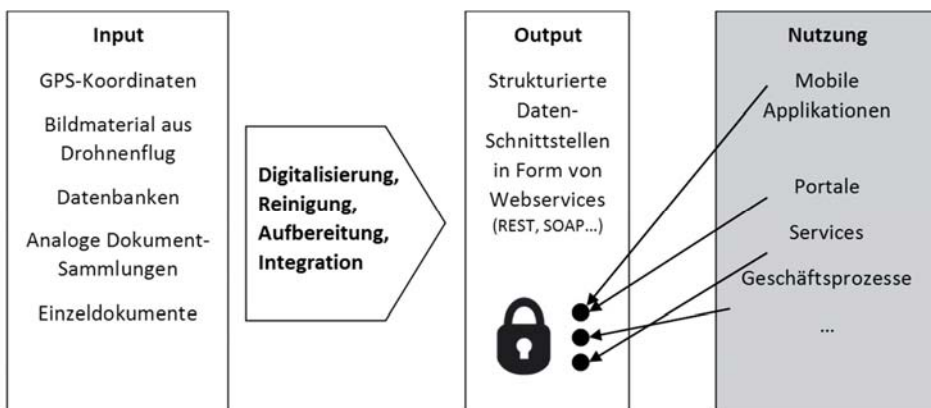


Abb. 3: Integration und sichere Bereitstellung verschiedenartiger Datenquellen

## Projektorganisation

Das Projekt PreLyrica wird am Lehrstuhl "Betriebswirtschaftliche Anwendungen der Informatik" am Fachbereich Wirtschaft der Technischen Hochschule Brandenburg durchgeführt. Es wird finanziert durch Drittmittel, die auf der Grundlage der Richtlinie des Ministeriums für Wissenschaft, Forschung und Kultur (MWFK) des Landes Brandenburg zur Förderung der "Stärkung der technologischen und anwendungsnahen Forschung an Wissenschaftseinrichtungen im Land Brandenburg" (StaF-Richtlinie) eingeworben wurden. Diese Förderung wird aus Mitteln des Europäischen Fonds für regionale Entwicklung (EFRE) kofinanziert. Die Durchführung des Projekts ist für den Zeitraum vom 01.12.2016 bis zum 30.11.2019 vorgesehen.

Das Projekt ist eng in den Lehr- und Forschungsbetrieb der Technischen Hochschule Brandenburg eingebunden. In dem Projekt arbeiten Professoren, akademische Mitarbeiter, der Technische Dienst und studentische Hilfskräfte. Aber auch die Studenten aus den Bachelor- und Masterstudiengängen der Studiengänge Wirtschaftsinformatik, Betriebswirtschaftslehre und Informatik werden über Lehrveranstaltungen, Praktika und Abschlussarbeiten an die Forschungsthemen des Projekts herangeführt.

Mit dem geschilderten Vorhaben sollen die Grundlagen für die Entwicklung höherwertiger Geschäftsmodelle geschaffen werden, die zum Beispiel die Aggregation, Aufbereitung oder Anreicherung von Informationsdiensten und deren Handel beinhalten können. Daher ist eine Technologieplattform das erste konkrete Projektziel.

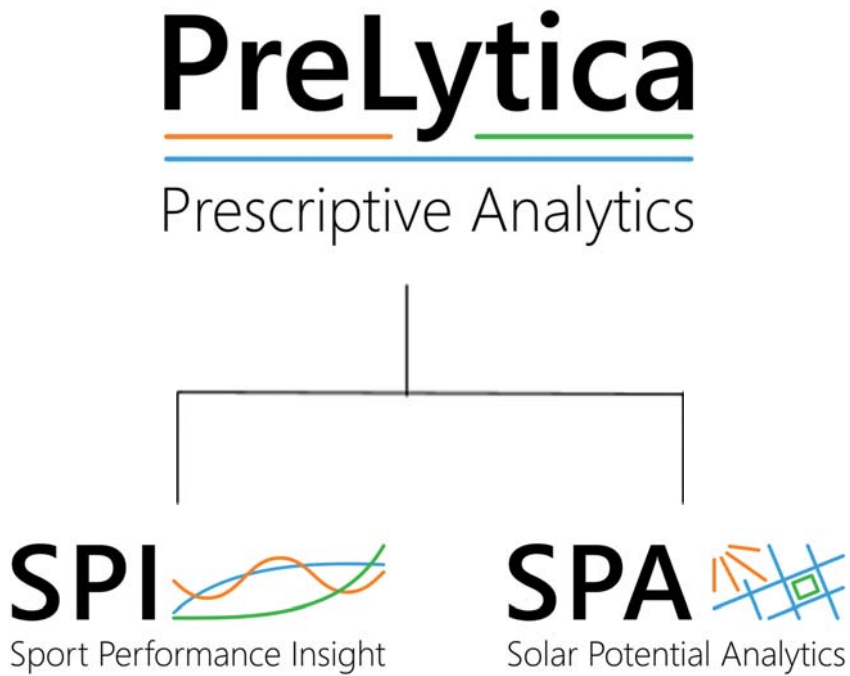
Auf Basis der Technologieplattform können Anwendungsszenarien aufgebaut werden, die die Vorteile des Systems greifbar machen. Im Rahmen des Projekts sollen beispielhaft zwei prototypische Anwendungsszenarien entwickelt werden, die die grundsätzliche Machbarkeit des Konzepts demonstrieren und konkreten Nutzen stiften. In beiden Szenarios werden Portale entwickelt, die eine Auswertung, Visualisierung und weiterführende Nutzung der integrierten Datenquellen durch Anwender ermöglichen.

Entsprechend der Cluster-Strategie des Landes Brandenburgs und um möglichst unterschiedliche Anwendungsszenarios abzudecken, wurden zwei Themengebiete herausgesucht:

- a) Gesundheitswirtschaft
- b) Energietechnik

Durch den Charakter des Projekts selbst ist zusätzlich das Themenfeld Informations-/Kommunikationstechnologie (IKT) abgedeckt. Durch diese Wahl der Themenfelder in der prototypischen Projektumsetzung werden die sonstigen Aktivitäten der Wirtschaftsförderung im Land Brandenburg unterstützt, da angestrebt wird, dass Projektergebnisse dieses Forschungsprojekts zukünftig möglichst Anwendung in der Praxis finden werden.

Aufgrund der unterschiedlichen zu realisierenden Prototypen wurde das Projekt PreLytics in die Teilprojekte "Sport Performance Insight" (SPI) und "Solar Potential Analytics" (SPA) unterteilt:



## Teilprojekt: Sport Performance Insight (SPI)

### Integration und Auswertung sportmedizinischer Leistungsdaten

Gerade im Sport- und Gesundheitsbereich werden heute durch eine Vielzahl von Institutionen Daten erfasst und durch individuelle Datensätze z. B. von Fitness-Trackern, Schrittzählern etc. ergänzt. Diese Daten sind dabei nicht nur im Sinne der Heilung von eingetretenen Gesundheitsschäden interessant, sondern sollen vielmehr der Vorsorge dienen. Weiterhin kann mit Hilfe solcher Daten ein Analyse-System entwickelt werden, welches zum Benchmarking personenbezogener sportlicher wie medizinischer Daten herangezogen werden kann.

Die Erfassung personenbezogener Daten zum Zwecke der automatisierten algorithmischen oder gar manuellen Auswertung birgt dabei die besondere Herausforderung in Bezug auf Datenschutz und Datensicherheit. Dieser Aspekt tritt insbesondere beim anonymisierten Vergleich von Merkmalsausprägungen im Sinne eines Benchmarkings beispielsweise nach Altersgruppe, Geschlecht und körperlicher Konstitution hervor. Hier gilt es, technologisch skalierbare Lösungen zu finden, die den Schutz der personenbezogenen Daten sicherstellen und gleichzeitig dem europäischen bzw. deutschen Rechtsrahmen entsprechen.

## Teilprojekt: Solar Potential Analytics (SPA)

### Auswertung heterogener Datenquellen

Dezentrale Energieversorgung rückt immer stärker in den Mittelpunkt von Betrachtungen im Kontext nachhaltiger Energiewirtschaft. Photovoltaik (PV) gilt in diesem Zusammenhang noch immer als eine der wachstumsstärksten alternativen Energiequellen.

Das Feld bietet viele Ansatzpunkte für den Einsatz präskriptiver Analytik. Bereits die Verknüpfung von Kartenmaterial und Wetterdaten lässt Prognosen über den zu erwartenden Wirkungsgrad einzelner PV-Anlagen in bestimmten Regionen zu. Werden weitere strukturierte und unstrukturierte Datenquellen hinzugezogen, können beliebige Gebäude in Hinblick auf die Eignung für PV-Anlagen bewertet werden.

Dazu ist eine umfangreiche Datenbasis notwendig, die aus verschiedenen Quellen gespeist werden kann.

Je nach Qualität der Datenbasis sind hier Echtzeit-Visualisierungen und Effizienz-Prognosen einzelner PV-Anlagen denkbar. Auch der Vergleich von PV-Anlagen, deren Nutzleistung und Verschleiß nach Region, Modell und Hersteller können verglichen werden. Die Umsetzung des Szenarios trägt zur Markttransparenz bei und ermöglicht die Optimierung von Angebot und Nachfrage.

Die Aufzeichnung von Leistungsdaten ermöglicht die Prognose von Instandhaltungsbedarfen, die Vorhersage von Ausfällen und unterstützt Nutzer daher bei der individuellen Beauftragung von Dienstleistern. Umgekehrt können die brandenburgischen Handwerksbetriebe, denen die Ergebnisse zur Verfügung gestellt werden können, gezielt und mit wenigen Streuverlusten auf potentiell interessierte Eigentümer zuzugehen. Das Forschungsprojekt leistet damit einen Beitrag zu Sicherung und Ausbau der Arbeitsplätze im brandenburgischen Handwerk.

Weiterhin hat die Entwicklung eine hohe Bedeutung für die Verteilnetzbetreiber, denen wertvolle Informationen über den zu erwartenden Ausbau dezentraler Energieerzeugungsanlagen zur Verfügung gestellt werden können, die eine zuverlässigere Prognose für die Auslastung des Netzes und damit die Identifizierung zu erwartenden knapper Netzkapazitäten ermöglichen. Damit kann zur Sicherung der Versorgungsqualität im Land Brandenburg beigetragen werden.



# Projekt Sport Performance Insight – Hintergründe

Dem prototypischen Ansatz des Gesamtprojekts folgend, musste für das SPI-Teilprojekt ein passendes Szenario gefunden werden. Schon seit Jahren wird in Deutschland an der Patientenakte oder elektronischen Gesundheitsakte gearbeitet (vergl. [1], [2]). Für das konkrete Projekt der präskriptiven Analyse, bei dem verschiedene Datenquellen kombiniert werden sollen, um neues Wissen zu generieren, eignet sich dieser Bereich des Gesundheitswesens jedoch nicht. Zum einen gibt es aufgrund des Datenschutzes hier nur wenig Möglichkeiten offene Datenquellen zu finden und die Technische Hochschule Brandenburg selbst hat keinen medizinischen Fachbereich, so dass auch keine eigenen Datenquellen in diesem Bereich existieren.

Da der Sport als wichtiger Bestandteil der öffentlichen Gesundheitsförderung einzuordnen ist (vergl. [3]) und in diesem Bereich gerade über die Veröffentlichung von Sportergebnissen zahlreiche Datenquellen verfügbar sind, wurde für das Projekt der Sport im allgemeinen und ganz konkret der Golfsport im speziellen ausgewählt. Die Sportart Golf hat sich dabei durch die Suche nach Projektpartnern ergeben, grundsätzlich sind die Verfahren und der umgesetzte Prototyp aber unabhängig von der gewählten Sportart.

## Projektpartner

Damit das Projekt nicht an rein theoretischen Fragestellungen arbeitet, sondern die Projektergebnisse möglichst direkt an konkreten Herausforderungen der Realität gemessen werden können, wurde nach Anwendern mit realen Fragestellungen gesucht. Sehr schnell ergab sich, dass die Sportart Golf ein guter Forschungsgegenstand sein könnte. Zum einen lässt sich Golf vom Kindes- bis Seniorenalter betreiben, so dass man umfangreiches Datenmaterial finden und erzeugen kann. Zum anderen gibt es gerade auf der amerikanischen Profigolftour umfangreiche Statistiken,

die fast alle Aspekte des Spiels umfassen<sup>1,2</sup>. Weiterhin gibt es beim Golfsport den Trend, dass sich Spieler immer mehr selbst über Apps vermessen, um spätere Daten zur Analyse des eigenen Stärken-/Schwächenprofils zu haben, mit denen sich dann das eigene Spiel verbessern lässt (vergl. [4]).

Aufgrund der Analyse des vorhandenen Datenmaterials ergab sich, dass auf Ebene des professionellen Golfs viele Daten existieren. Sportler können sich so umfassend in fast allen Bereichen des Spiels mit anderen Spielern auch jenseits der reinen Turnierergebnisse vergleichen. Die Datenlage gerade im Nachwuchsbereich ist jedoch sehr dünn. Hier gibt es nur vereinzelt Daten zu einzelnen Spielern oder von einzelnen Clubs.

Als Teilprojektziel wurde daher identifiziert: Wie kann man Benchmarks im Nachwuchsbereich erzeugen, so dass Spieler sinnvolle Vergleiche mit anderen Spielern haben. Dabei ist wichtig zu beachten, dass mehr als nur der Parameter "Alter" zu berücksichtigen ist, da jedes Kind sehr individuell unterschiedliche Entwicklungsstufen durchläuft, so dass trotz gleichen biologischen Alter sehr große Unterschiede in der körperlichen und geistigen Entwicklung vorliegen können.

Auf der Suche nach Projektpartnern wurde also insbesondere Ausschau nach Vereinen gehalten, die eine große Anzahl von Jugendlichen haben. Die Statistik des Deutschen Golf Verband e.V. (DGV) wies 2016 insgesamt 643.158 Golfspieler (vergl. [3]) aus, davon allerdings nur 43.953 Spieler unter 18 Jahren. Diese Mitgliedszahlen lassen den DGV auf Platz 8 in der Liste der Olympischen Spitzenverbände des Deutschen Olympischen Sportbund (DOSB) 2016 erscheinen, hinter den Sportarten Fußball, Turnen, Tennis, Schützenbund, Leichtathletik, Handball und Reiten.

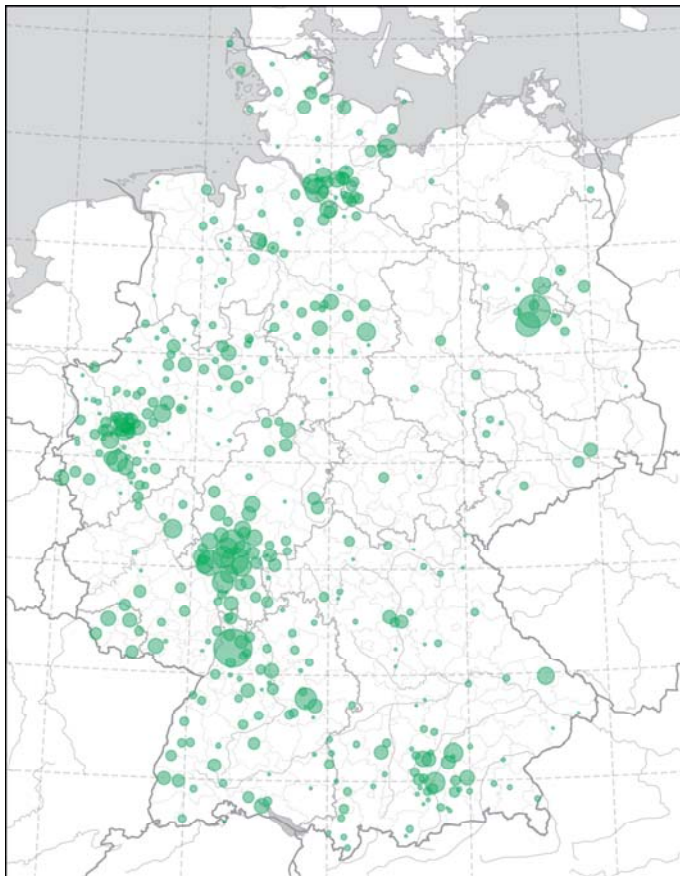
Da für das Projekt vor allem aktive Nachwuchssportler interessant sind, die am Turniergeschehen teilnehmen, wurden Datenbanken der Portale golf.de und gjgr.de ausgewertet. Wie Abbildung 4 zeigt, ergab die Suche nur eine geringe Anzahl potentieller Projektpartner im Umfeld der Hochschule.

---

<sup>1</sup> <http://www.pgatour.com>

<sup>2</sup> <http://www.golfstats.com>

Der Golf- und Land-Club Berlin-Wannsee e.V. (GLCBW) zählt nicht nur im Hochschulumfeld, sondern auch deutschlandweit zu den Vereinen mit einer sehr ausgeprägten Jugendarbeit. Vom DGV wurde der GLCBW wiederholt mit dem Qualitätsmanagement-Zertifikat in Gold und dem Bundespreis “Zukunft Jugend” für die beste Nachwuchsförderung Deutschlands ausgezeichnet, weiterhin erhielt er bereits zwei Mal die Ehrung mit dem “Grünen Band für vorbildliche Talentförderung” vom Deutschen Olympischen Sportbund.



**Abb. 4: Golfclubs mit aktiven Jugendspielern unter 18 Jahren**  
(je größer der Punkt in der Karte, desto mehr aktive Spieler hat der entsprechende Verein)

Der GLCBW hatte großes Interesse an dem Projekt der Technischen Hochschule Brandenburg, so dass der Prototyp zusammen mit den Anforderungen des Vereins realisiert werden konnte.

## Konzept der Spielerakte

Analog des Prinzips der elektronischen Patientenakte im Gesundheitswesen, wurde als Prototyp im Projekt SPI ein System zur Verwaltung von Spielerakten aufgebaut. Dabei sollen alle Daten, die für die Förderung und Entwicklung eines Spielers als Entscheidungsgrundlage notwendig sind, zentral zugänglich sein. Bisher mussten die Trainer und Sportdirektoren auf verschiedenste Systeme zugreifen:

- Clubverwaltungssoftware für Kontaktdaten, Geburtsdatum etc.
- Verschiedene Portale für die Ergebnisse von nationalen und internationalen Turnieren
- Individuelle Dokumente der Trainer für bestimmte Leistungsdaten (z. B. Schlägerkopfgeschwindigkeiten, Testergebnisse in Bereichen wie Putten, Chippen, Pitchen usw.)
- Individuelle Dokumente des Clubs zu Trainingsanwesenheiten, Teilnehmerlisten etc.
- Sport-Gesundheitszeugnisse
- Daten von Apps, die der Spieler selbst zur Zwecke der Rundenanalyse u.ä. nutzt
- Daten von stationären Messsystemen wie TrackMan, SAM PuttLab etc.
- u.v.m.

Diese Vielzahl an Systemen, die die Informationen verwalten, erzeugt einige Probleme: Unterschiedliche Datenformate, Zugang zu den Informationen, Aktualität und Verfügbarkeit. Der Trainer des Spielers kann sich somit nicht auf seine Kernkompetenz der Spielerentwicklung konzentrieren, sondern verliert sich häufig in administrativen Tätigkeiten. Die Unterstützung des Trainers und des Spielers bei der Entwicklung seiner sportlichen Potentiale ist daher die Vision der Spielerakte, die schematisch in Abbildung 5 dargestellt ist.

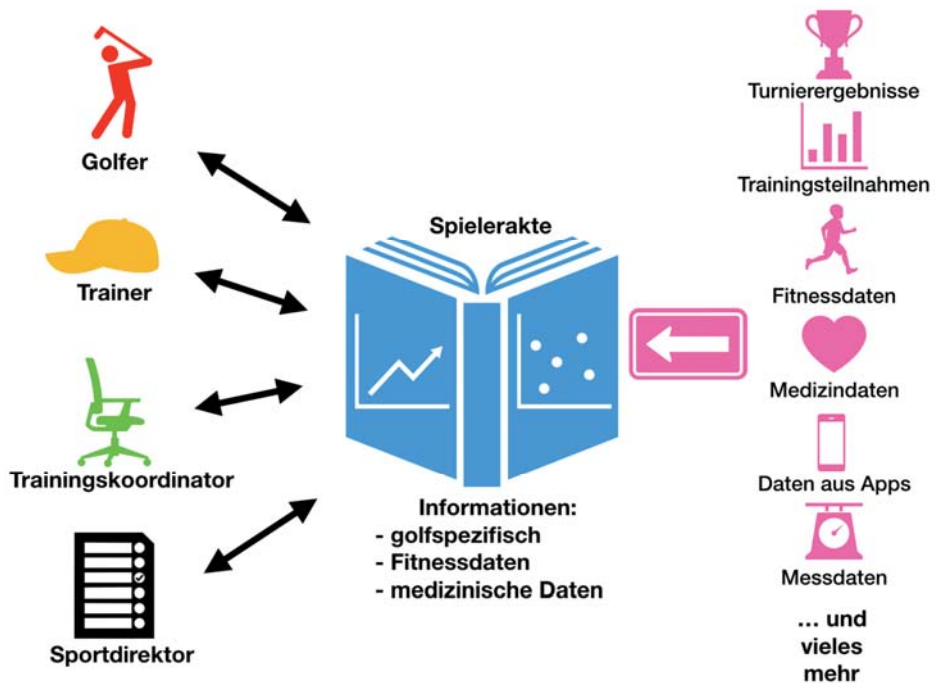


Abb. 5: Konzept der Spielerakte

(Zentraler Zugang zu allen relevanten Informationen, die den Spieler betreffen. Alle berechtigten Personen wie der Spieler, Trainer, Trainingskoordinator und Sportdirektor haben lesenden wie schreibenden Zugriff auf die Informationen in der Spielerakte, die ansonsten auch automatisch mit den aktuellen Turniresultate etc. gefüllt wird.)

Für die Trainer wurde neben dem Portal zur Einsicht der Daten aus der Spielerakte auch eine App für Smartphones mit iOS und Android Betriebssystem entwickelt. Über die App können die Trainer golfspezifische Messwerte schnell und direkt beim Training erfassen. Ebenso können über diese App die Anwesenheiten der Spieler aller Trainingsgruppen gepflegt werden. Durch die App ist also eine Pflege der Daten in der Spielerakte möglich. Durch diese Realisierung ohne Medienbrüche und mittels einfacher und schneller Bedienoberfläche ist ein hoher Anreiz gegeben, dass die Daten immer aktuell sind und kein administrativer Mehraufwand gefühlt wird. Aktuelle Daten sind die Grundvoraussetzung aller Projekte zur Datenanalyse.

## Quellen

- [1] WARDA, F. : *Elektronische Gesundheitsakten – Möglichkeiten für Patienten, Ärzte und Industrie*. 2005, Rheinware-Verlag. — ISBN 978-3938975008
- [2] MOHR, F. : *Empirische Untersuchung und Metaanalyse zur Akzeptanz der elektronischen Gesundheitskarte für Deutschland*. 2010, Verlag Dr. Kovac. — ISBN 978-3830052807
- [3] KNOLL, M. : *Sport and Health – The German Perspectives*. In: Int. Journal of Physical Education 37. 2000/2, p. 50-57
- [4] BROADIE, M. : *Every Shot Counts – Using Revolutionary Strokes Gained Approach to Improve Your Golf Performance and Strategy*. 2014, Avery-Verlag. — ISBN 978-1592407507



# Screen Scraping

Integration heterogener Daten aus Webseiten in einen unternehmenseigenen Datenbestand

Tobias Kiertscher

## Inhalt

|                                          |    |
|------------------------------------------|----|
| Einleitung .....                         | 19 |
| Zielstellung.....                        | 19 |
| Begriffe.....                            | 20 |
| Kriterien zur Optimierung.....           | 21 |
| Navigationsstrategie .....               | 21 |
| Reverse Engineering von URLs .....       | 21 |
| Explorative Traversierung.....           | 23 |
| Erkenntnisse .....                       | 24 |
| Abfragestrategie.....                    | 24 |
| Verzögerung von HTTP-Anfragen.....       | 25 |
| Abruf von referenzierten Ressourcen..... | 26 |
| Parallelisierung.....                    | 27 |
| HTTP-Header .....                        | 27 |
| Abfragezeitpunkt .....                   | 28 |
| Erkenntnisse .....                       | 28 |
| Analysestrategie.....                    | 29 |
| Elementtypen.....                        | 30 |
| HTML-IDs.....                            | 30 |
| CSS-Klassen.....                         | 30 |
| XPath oder CSS-Selektoren.....           | 30 |



|                                                      |    |
|------------------------------------------------------|----|
| Mustererkennung .....                                | 31 |
| Erkenntnisse .....                                   | 31 |
| Herausforderungen .....                              | 33 |
| De-normalisierte Präsentation .....                  | 33 |
| Pagination.....                                      | 38 |
| Abfragebegrenzungen .....                            | 40 |
| Passwortgeschützte Bereiche .....                    | 43 |
| Fehlende Benachrichtigung über Datenänderungen ..... | 44 |
| Aktive Inhalte.....                                  | 45 |
| Identifikatoren .....                                | 47 |
| Fehlerquellen.....                                   | 48 |
| Netzausfälle und Server-Fehler.....                  | 48 |
| Änderungen an der Struktur der Website.....          | 49 |
| Architektur.....                                     | 50 |
| Module .....                                         | 51 |
| Fremdsysteme.....                                    | 59 |
| Allgemeine Aspekte .....                             | 60 |
| Implementierung.....                                 | 62 |
| Empfehlungen .....                                   | 64 |
| Quellen.....                                         | 66 |

## Einleitung

Das Internet umfasst eine große Anzahl von Websites mit frei verfügbaren Daten, die aber in vielen Fällen nur für die Nutzung durch einen menschlichen Benutzer gestaltet sind. Erst in den letzten Jahren nimmt die zusätzliche Bereitstellung von Schnittstellen, die für die Maschine-zu-Maschine-Kommunikation optimiert sind, zu. Unter dem Stichwort *Open Data* entstehen vermehrt Bemühungen, Daten nicht nur frei zu Verfügung zu stellen, sondern auch über eine Schnittstelle, mit der sie sich leicht in neue oder bestehende Anwendungen integrieren lassen.

Diese Bemühungen werden jedoch auf absehbare Zeit nicht alle interessanten Websites erfassen. Es existieren weiterhin eine Vielzahl von Websites mit wertvollen Informationen, wo entweder die notwendigen Mittel fehlen, um neben der menschenfreundlichen auch noch eine maschinenfreundliche Schnittstelle anzubieten, oder wo seitens der Betreiber schlicht kein Interesse an einer solchen besteht.

In der Integration der Informationen aus diesen nicht maschinenfreundlichen Websites, mit den unternehmenseigenen Daten, steckt großes Potential. Um dieses Potential auszuschöpfen müssen die Informationen aus den Websites extrahiert und strukturiert gespeichert werden. Diesen Zweck erfüllt das Screen Scraping.

Das *Screen Scraping* ist eine Vorgehensweise, bei der automatisiert strukturierte Informationen aus Webseiten gewonnen werden. Dabei sind die Webseiten als grafische Schnittstelle für den Menschen gestaltet und wurden nicht dafür optimiert, dass Programme die enthaltenen Informationen automatisch extrahieren und weiterverarbeiten. Beim Screen Scraping werden Webseiten nicht wie üblich durch einen Browser abgerufen und grafisch dargestellt, sondern durch ein spezielles Programm abgerufen, das den HTML-Quelltext, i. d. R. ohne grafische Ausgabe, nach spezifischen Mustern durchsucht und Informationen extrahiert.

## Zielstellung

Beim Erfassen von Informationen durch *Screen Scraping* ist das Ziel, strukturierte Informationen in einer Datenbank abzulegen, um diese anschließend effizient durchsuchen und/oder weiterverarbeiten zu können. Insbesondere sollen tabellarische Informationen, potentiell mit verlinkten Detailseiten, in einer Website iden-

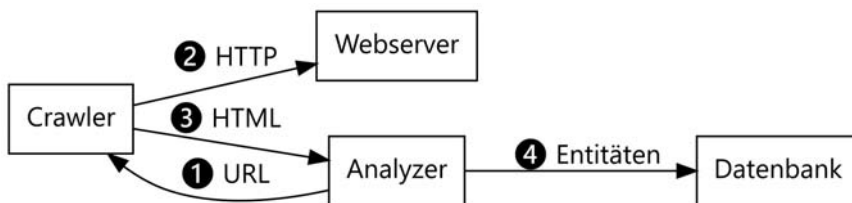
tifiziert, extrahiert, bereinigt, zusammengeführt und dann in einer Datenbank abgelegt werden. Diese Daten sollen anschließend mit unternehmenseigenen Daten kombiniert und in Folge ein Erkenntnisgewinn erreicht werden, der einen Wettbewerbsvorteil ermöglicht.

## Begriffe

Als *Website* wird eine Gruppe von *Webseiten* verstanden, welche über eine Navigationsstruktur, z. B. ein Menü, miteinander verknüpft sind. Die Webseiten einer Website sind in HTML kodiert und werden mittels eines *Webserver*s über HTTP bzw. HTTPS bereitgestellt. Eine Website wird über einen *Domännennamen* und optional einen Pfad identifiziert. Der Domänenname und der Pfad werden in einer *URL* zusammengefasst.

Die Informationen, welche aus einer Website extrahiert werden sollen, sind in Entitäten gruppiert. Eine *Entität* enthält Informationen über eine identifizierbare Sache. Die Sache kann gegenständlich oder abstrakt sein. Eine Entität besitzt einen *Identifikator* mit dem sie von anderen Entitäten eindeutig unterschieden werden kann. Wenn Informationen über eine *Entität* auf mehrere Webseiten verteilt sind, werden diese partiellen Informationen *Facetten* genannt.

Ein Screen-Scraping-System besteht mindestens aus einem Crawler und einem Analyzer. Der *Crawler* ruft die HTML-kodierten Webseiten vom Webserver ab. Der *Analyzer* extrahiert die Informationen aus den Webseiten und speichert sie in einer *Datenbank*.



Grundlegender Ablauf des Screen Scraping

## Kriterien zur Optimierung

Ein Screen-Scraping-System kann nach unterschiedlichen Kriterien optimiert werden. Dabei bilden diese Kriterien oft Gegensätze, welche manchmal nur die Parametrisierung, aber manchmal auch die Architektur des Systems beeinflussen. Das hat zur Folge, dass u.U. bereits während der Entwicklung des Systems entschieden werden muss, welche Kriterien Priorität haben.

- Navigationsstrategie:  
Effiziente Erfassung  $\Leftrightarrow$  Unauffällige Erfassung
- Abfragestrategie:  
Schnelle Erfassung  $\Leftrightarrow$  Unauffällige Erfassung
- Analysestrategie:  
Einfache Implementierung  $\Leftrightarrow$  Robustheit gegen Änderungen des Seitenaufbaus

## Navigationsstrategie

Die Navigationsstrategie steuert die Erzeugung bzw. Ermittlung der URLs für den Seitenabruf. Folgen die URLs einem festen Schema und enthalten sie eindeutige Bezeichner oder IDs, die im Voraus bekannt sind, können sie oft direkt gebildet werden und müssen nicht aus einer Navigationsschaltfläche extrahiert werden. Dies wird als *Reverse Engineering von URLs* bezeichnet. Stehen die URLs jedoch nicht in einer direkten Beziehung zu den Informationen auf den Webseiten, müssen die entsprechenden Link-Elemente auf den Webseiten identifiziert und die im href-Attribut referenzierten URLs verwendet werden. Dies wird als *explorative Traversierung* bezeichnet.

Dieser Abschnitt gliedert sich wie folgt:

- Reverse Engineering von URLs
- Explorative Traversierung
- Erkenntnisse

## Reverse Engineering von URLs

URLs für den Webseitenabruf besitzen üblicherweise den folgenden Aufbau:

*Protokoll* : // *Server* [ : *Port* ] / *Pfad* [ ? *Abfrageparameter* ] [ # *Anker* ]

Der *Pfad* und die *Abfrageparameter* sind üblicherweise die variablen Teile beim Screen Scraping. Das *Protokoll*, der *Server* und, falls erforderlich, der *Port* stehen i. d. R. fest. Der *Anker* spielt für das Screen Scraping keine Rolle, da er angibt, welchen Teil einer Seite der Browser als erstes darstellen soll. Beim Screen Scraping wird die Seite jedoch normalerweise nicht gerendert, sondern lediglich der HTML-Quelltext der Seite analysiert.

Die Analyse einer URL kann unterschiedlich kompliziert sein und soll hier nicht im Detail erläutert werden. Es lassen sich jedoch häufig die folgenden Teile in Pfad oder Abfrageparametern identifizieren.

- Kategorie
- Seitenname
- Entitätenklasse
- Entitäten-ID
- Filterkriterium
- Paginationsparameter
- Anzeigeparameter
- Session-ID

An dieser Stelle ist das Ausprobieren von manuell veränderten URLs hilfreich, um die Möglichkeiten des Web-Servers zu erproben. Falls die Website ein Web-Framework oder ein Content-Management-System verwendet, dessen Dokumentation öffentlich zugänglich ist, kann es helfen diese nach Hintergrundinformationen und Systematiken zu durchsuchen.

### Vorteile

- Unabhängiger von konkreter Gestaltung der Navigationsschaltflächen
- Effizienter in der Abfrage da Seiten mit gesuchten Inhalten direkt abgefragt werden können

### Nachteile

- Abhängig vom konkreten Aufbau der URLs
- Auffälliger im Abfragemuster

Unter der Annahme, dass sich der technische Unterbau einer komplexen Website nicht so schnell ändert wie das Design bzw. die konkrete Beschriftung von Navigationsschaltflächen, ist das Reverse Engineering der URLs die robustere Variante.

Ob sich der technische Unterbau einer Website ändert, hängt natürlich von einer ganzen Reihe von Faktoren ab, die sich nicht immer sicher ermitteln bzw. voraussehen lassen. Dazu zählen z. B. die Reife der aktuellen Implementierung, die strukturelle Wandlung der präsentierten Informationen, die Kapazitäten des Website-Betreibers und zukünftige technische Entwicklungen.

## Explorative Traversierung

Bei der explorativen Traversierung werden, ausgehend von der Startseite, Navigationsschaltflächen gesucht, welche unmittelbar oder mittelbar zu der Seite mit den gesuchten Informationen führen. Aus diesen Schaltflächen wird dann die URL der Zielseite extrahiert. Normalerweise werden Navigationsschaltflächen im HTML durch a-Elemente repräsentiert. Das href-Attribut enthält dann die absolute oder relative URL der Zielseite.

Die Herausforderung bei dieser Strategie ist das algorithmische Beschreiben von Navigationspfaden. In einem einfachen Fall kann eine Liste von Identifikatoren ausreichen, um eine Zielseite zu beschreiben. Dabei beschreibt jeder Identifikator eindeutig eine Navigationsschaltfläche im HTML-Baum der aktuellen Seite. Als Identifikator kann z. B. eine ID, eine CSS-Klasse, ein XPath oder ein CSS-Selektor dienen. Die Liste wird ausgehend von der Startseite abgearbeitet, indem mit dem jeweiligen Identifikator die Navigationsschaltfläche und mit ihr die URL der nächsten Seite ermittelt wird. Diese URL wird aufgerufen und auf der geladenen Seite wird der nächste Identifikator verwendet. Ist die Liste zu ende, sollte die zuletzt geladene Seite die Zielseite sein. Zusätzliche Prüfkriterien, wie zu erwartende Elemente in der Seite oder Seitentitel können dazu dienen, frühzeitig zu erkennen, ob ein Navigationsschritt erfolgreich war oder nicht.

### Vorteile

- Unabhängiger von technischer Umsetzung (URLs)
- Entspricht mehr dem menschlichen Verhalten und ist damit unauffälliger

### Nachteile

- Komplexer in der Implementierung
- Abhängiger von konkreter Ausprägung des Designs
- Es müssen öfter Zwischenseiten aufgerufen werden, um zu den eigentlich gesuchten Informationen zu gelangen

## Erkenntnisse

Das Reverse Engineering von URLs eignet sich besser für das effiziente Erfassen von Inhalten. Da die URLs für die Webseiten mit den interessanten Inhalten u.U. direkt gebildet werden können, ist es nicht erforderlich, Zwischenseiten ausschließlich für den Zweck aufzurufen, über Navigationsschaltflächen die nächste URL zu erhalten. Des Weiteren können Abfragen häufig parallel getätigt werden, was den Durchsatz der Datenerfassung erhöht.

Die explorative Traversierung eignet sich eher für das unauffällige Erfassen von Inhalten. Denn das Traversieren der Navigationspfade stellt sicher, dass man Webseiten in einer Reihenfolge aufruft, die auch von einem menschlichen Benutzer gewählt werden könnte. Auch der Zwang eine Seite nach der anderen abzurufen, entspricht dem menschlichen Verhalten. Durch die serielle Natur der Traversierung ist jedoch die Geschwindigkeit und Effizienz der Datenerfassung begrenzt.

## Abfragestrategie

Die Art wie URLs mittels HTTP abgerufen werden, kann ein kritischer Punkt für den Erfolg für das Screen Scraping sein. Je nach Kooperationsbereitschaft des Website-Betreibers und Kapazitäten des Web-Servers kann die Abfrage auf Geschwindigkeit und Effizienz oder auf Unauffälligkeit optimiert werden.

Dieser Abschnitt gliedert sich wie folgt:

- Verzögerung von HTTP-Anfragen
- Abruf von referenzierten Ressourcen
- Parallelisierung
- HTTP-Header
- Abfragezeitpunkt
- Erkenntnisse

Ein menschlicher Benutzer ruft i. d. R. im Browser nur einen Link einer Seite auf und wartet, bis diese aufgebaut ist. Der Browser ruft die Seite über HTTP ab und schickt dabei eine Reihe von HTTP-Header-Werten mit. HTML-Seiten referenzieren üblicherweise eine Vielzahl von zusätzlichen Ressourcen (CSS-Styles, JavaScript-Dateien, Bilder, ...), welche von einem Browser abgerufen werden müssen. Erst dann

stellt der Browser die Seite grafisch dar und der Benutzer kann evtl. erneut auf einen Link klicken. Menschen besuchen Webseiten üblicherweise zu bestimmten Tageszeiten und selten mehrfach zum exakt gleichen Zeitpunkt.

Ein performantes Vorgehen beim Screen Scraping wäre hingegen das frühzeitige Erzeugen von mehreren URLs und deren parallele Abfrage. Es werden nur die HTTP-Header-Werte mitgeschickt, die unbedingt erforderlich sind, damit der Web-Server die gewünschte HTML-Seite ausliefert. Weitere Ressourcen werden nicht abgerufen, da dies die Implementierung verkompliziert und unnötige Übertragungskapazität verbraucht. Eine wesentliche Verzögerung von aufeinanderfolgenden Abrufen ist auch nicht notwendig. Sobald URLs erzeugt oder extrahiert wurden, die abgerufen werden sollen, kann dies unmittelbar geschehen, um die Übertragungskapazität optimal auszulasten. Sollen Informationen regelmäßig abgerufen werden, bietet es sich an, das Screen Scraping jede Nacht zu einem geplanten Zeitpunkt z. B. mit einem Cron-Job durchzuführen.

## Verzögerung von HTTP-Anfragen

Eine Verzögerung von HTTP-Anfragen senkt die Last auf dem Web-Server und kann das Zugriffsmuster dem eines menschlichen Benutzers angleichen. Als Nachteil senkt sie naturgemäß auch die Abfragegeschwindigkeit. Die Verzögerung kann einfach implementiert werden, indem vor jeder HTTP-Anfrage eine künstliche Pause eingeführt wird. Sollten im HTML referenzierte Ressourcen abgerufen werden, brauchen diese Abfragen nicht verzögert zu werden, da auch ein Browser referenzierte Ressourcen so schnell wie möglich und i. d. R. parallel abrufen.

Jede Art von Pause senkt die Last auf dem Web-Server. Die Länge der Pause kann jedoch wichtig sein, wenn es darum geht einen menschlichen Benutzer nachzuahmen. Möglich sind z. B. eine feste Länge, eine zufällige Länge aus einem Intervall, oder eine Länge, die aus Eigenschaften der zuvor abgerufenen Web-Seite abgeleitet wird.

Eine feste Länge ist gut geeignet, um die Last auf dem Web-Server zuverlässig zu senken. Auch dadurch fallen die Zugriffe des Screen-Scraping-Programms schon weniger in den Statistiken des Web-Servers auf. Sie ist jedoch kaum geeignet, das Zugriffsmuster eines menschlichen Benutzers nachzuahmen.



Bei einer zufälligen Länge kann auch die statistische Verteilung eine Rolle spielen. Die üblichen Pseudozufallsgeneratoren erzeugen eine Gleichverteilung. Eine Normalverteilung entspricht jedoch eher einem natürlichen Prozess und ahmt damit vermutlich das menschliche Verhalten besser nach (vergl. [1]).

Soll die Länge der Pause aus der zuvor abgerufenen Web-Seite abgeleitet werden, kann z. B. die Größe der Web-Seite – Anzahl der Zeichen im HTML-Quelltext, Anzahl der HTML-Elemente innerhalb des body-Elements, o.ä. – als Multiplikator für eine minimale Länge genutzt werden. Dies spiegelt die Annahme wieder, dass ein menschlicher Benutzer bei großen Seiten länger braucht, um auf einen Folgelink zu klicken, als bei kleinen Seiten. Eine Berücksichtigung der zuvor abgerufenen Web-Seite ist nur dann sinnvoll, wenn die Navigationsstrategie der explorativen Traversierung eingesetzt wird.

Die letzten beiden Ansätze lassen sich auch gut kombinieren.

## Abruf von referenzierten Ressourcen

Für die Darstellung einer Web-Seite verwendet der Browser ein *Cascading Stylesheet* (CSS), welches für jedes HTML-Element vorgibt, wie es formatiert werden soll. Nahezu jede Web-Seite enthält eine oder mehrere Referenzen auf CSS-Dateien, welche das Aussehen der Web-Seite steuern. Zusätzlich referenzieren HTML-Seiten auch noch weitere Ressourcen, wie JavaScript-Code, Bilder und Web-Fonts. Der Browser muss diese mittels HTTP-Abrufen, bevor er die Web-Seite korrekt rendern kann. Um dem Benutzer möglichst schnell eine korrekte grafische Darstellung bieten zu können, lädt der Browser alle im HTML-Code referenzierten Ressourcen parallel herunter. Zusätzlich versucht der Browser bei dem Wechsel von einer Web-Seite zur nächsten, bereits heruntergeladene Ressourcen wieder zu verwenden. Dazu speichert er diese in einem lokalen Cache.

Folglich entsteht auf dem Web-Server ein typisches Muster für einen Website-Besuch: Bei Abruf der ersten Seite einer Website wird zunächst die HTML-Quelle und anschließend parallel alle referenzierten Ressourcen angefragt. Bei einem Seitenwechsel, z. B. durch das Anklicken eines internen Links, wird wieder zunächst die HTML-Quelle und anschließend parallel nur noch jene referenzierten Ressourcen angefragt, die noch nicht im lokalen Cache zwischengespeichert sind.

Für das Screen Scraping sind die referenzierten Ressourcen i. d. R. nicht von Bedeutung, da die gesuchten Informationen im HTML-Code enthalten sind. Deshalb müssen diese gar nicht abgerufen werden, wenn der Abruf effizient und schnell geschehen soll. Soll der Screen Scraper jedoch möglichst unauffällig sein, muss er derart implementiert werden, dass er das Verhalten eines Browsers nachahmt, der von einem Menschen bedient wird.

## Parallelisierung

Um einen möglichst hohen Durchsatz beim Abruf von HTML-Seiten zu erreichen, können die HTTP-Anfragen parallelisiert werden. Das bedeutet, dass z. B. mittels Threads, mehrere HTTP-Anfragen gleichzeitig an den Web-Server geschickt werden. Hat der Web-Server ausreichende Rechenkapazitäten, evtl. mehrere Prozessoren, oder werden mittels Load-Balancer mehrere Web-Server eingesetzt, kann die Abfragegeschwindigkeit damit erheblich gesteigert werden. Besonders wirksam kann die Parallelisierung bei der Navigationsstrategie mittels Reverse Engineering von URLs eingesetzt werden.

Man sollte sich jedoch bewusst sein, dass das Screen-Scraping-Programm bei parallelisiertem Abruf das Vielfache an Kapazitäten auf der Seite des Website-Betreibers nutzt, als ein einzelner menschlicher Benutzer. Es sollte sichergestellt werden, dass ausreichend Kapazitäten für gewöhnliche Besucher der Website verbleiben.

Soll das Abfragemuster dem eines menschlichen Benutzers angeglichen werden, kann die Parallelisierung jedoch nicht beliebig zur Durchsatzsteigerung eingesetzt werden. Vielmehr sollte sie nur zum Abruf von im HTML referenzierten zusätzlichen Ressourcen verwendet werden, um das Verhalten eines Web-Browsers nachzubilden.

## HTTP-Header

Bei jeder HTTP-Anfrage hat der Client, i. d. R. der Browser, die Möglichkeit neben der URL zusätzliche Informationen an den Web-Server zu übermitteln. Dies geschieht im HTTP-Header. Mit den Web-Entwickler-Werkzeugen von modernen Browsern kann man leicht ermitteln, welche Header-Felder der Browser bei einem typischen Web-Seiten-Abruf mitsendet. Dazu gehören u. a. User-Agent, Accept, Accept-Encoding, Accept-Language und evtl. Cookie.

Um sicherzugehen, dass der Abruf zuverlässig funktioniert, können einfach alle Header-Felder einer typischen Browser-Anfrage auch im Screen-Scraping-Programm verwendet werden. Dies ist auch erforderlich wenn das Screen Scraping unauffällig erfolgen soll. Zu beachten ist dabei, dass das User-Agent-Feld mit den üblichen Update-Zyklen der Browser aktualisiert werden muss.

Ist eine offene Absprache mit dem Website-Betreiber möglich und dieser mit dem Screen Scraping einverstanden, könnte das User-Agent-Feld jedoch auch genutzt werden, um das Screen-Scraping-Programm eindeutig als solches zu kennzeichnen. Dann hätte der Website-Betreiber die Möglichkeit, die Anfragen des Screen Scrapings mit entsprechenden Filtern nach seinen Erfordernissen zu priorisieren, oder zumindest die anteilige Last des Screen Scrapings in der Überwachung zu erkennen.

## Abfragezeitpunkt

Der Zeitpunkt zu dem ein Screen Scraping erfolgt, kann aus zweierlei Gründen wichtig sein. Wenn die Kooperation des Website-Betreibers vorausgesetzt werden kann, könnte ein Zeitpunkt gewählt werden, an dem üblicherweise wenige Besucher den Web-Server frequentieren. Dadurch werden Kapazitäten genutzt, die anderenfalls brachliegen würden und der normale Betrieb der Website wird nicht behindert. Soll der Abruf allerdings unauffällig erfolgen, muss das Screen Scraping eben genau dann durchgeführt werden, wenn viele Menschen die Website besuchen. Dann ist die Wahrscheinlichkeit höher, dass die Abfragen des Screen-Scraping-Programms nicht auffallen.

Bei der Wahl des Abfragezeitpunktes ist die Zeitzone der üblichen Besucher zu beachten. Hat die Website internationales Publikum, fällt die Zeitzone nicht so schwer ins Gewicht, wie wenn sich das Publikum z. B. auf ein Land beschränkt.

Wenn der Abruf regelmäßig aber unauffällig geschehen soll, muss der Zeitpunkt des Abrufs jeweils um eine zufällige und natürlich wirkende Zeitspanne verschoben werden (vergl. [1]).

## Erkenntnisse

Das Abrufen von Webseiten und anderen Ressourcen von einem Web-Server verbraucht Kapazitäten (Übertragungskapazität, Rechenkapazität) und hinterlässt

Spuren in Form von Log-Dateien. Ein Website-Betreiber mit entsprechenden Kenntnissen kann leicht erkennen, ob ein einfaches Screen-Scraping-Programm so schnell wie möglich Informationen sammelt oder ob ein menschlicher Benutzer mit einem Browser die Website abrufen. Als Folge kann ein IP-Filter konfiguriert werden, der einen potentiell unerwünschten Client ausschließt. Es existieren auch Application-Level-Firewalls, welche die Muster von HTTP-Anfragen permanent überwacht und den Administrator warnt oder Anfragen sogar selbstständig blockiert, wenn unübliche Muster auftreten.

Je nach Szenario ist es notwendig, das Verhalten des Screen-Scraping-Programms dem eines menschlichen Benutzers anzugleichen. Die Motivation dafür kann sehr unterschiedlich aussehen, darf jedoch nicht darin bestehen die Interessen des Website-Betreibers zu verletzen. In vielen Fällen kann es technisch notwendig sein, da der Web-Server nur Browser-typische Anfragen zuverlässig verarbeiten kann.

Wird das Screen Scraping nur ein einziges Mal durchgeführt, rechtfertigt sich der zusätzliche Aufwand für das Nachahmen von menschlichem Verhalten nur in Maßen. Wird es jedoch regelmäßig durchgeführt oder sogar als Datenquelle für ein angebundenes System genutzt, hängt es von der Kooperationsbereitschaft des Website-Betreibers ab, wie weit die Abfrage auf Performance optimiert werden darf. Ein auf Performance optimiertes Screen-Scraping-Programm kann bei schwachen Web-Servern oder einer geringen Übertragungskapazität zu einem deutlichen Anstieg der Antwortzeit führen und kommt damit einem Denial-of-Service-Angriff nahe.

## Analysestrategie

Die Analysestrategie steuert das Auffinden von einzelnen Informationen in einer HTML-Seite. Um eine bestimmte Information in einer HTML-Seite zu finden, können verschiedene Techniken verwendet werden, die von festen IDs über CSS-Klassen und XPath bzw. CSS-Selektoren bis hin zu komplexer Mustererkennung reichen.

Zentrales Element in der Analyse einer Webseite ist der HTML-Quelltext, der beim Einlesen, mit einer darauf spezialisierten Programmierbibliothek, in eine baumartige Datenstruktur überführt wird. Diese Datenstruktur wird auch *DOM* (Document Object Model) genannt. Im DOM kann mit verschiedenen Techniken nach Elementen gesucht werden.

Dieser Abschnitt gliedert sich wie folgt:

- Elementtypen
- HTML-IDs
- CSS-Klassen
- XPath oder CSS-Selektoren
- Mustererkennung
- Erkenntnisse

## Elementtypen

HTML definiert eine ganze Reihe von Elementtypen, mit denen die Struktur einer Webseite gegliedert wird. Beispiele sind `div`, `a`, `table`, `footer`. In einigen Fällen reicht allein der Typ aus, um Elemente mit gesuchten Informationen oder Navigations-schaltflächen für die Traversierung eindeutig zu identifizieren.

## HTML-IDs

Ein Element im DOM anhand seiner ID zu finden, ist die einfachste und dabei eine recht robuste Technik. In HTML können einzelne Elemente mit einem Identifikator versehen werden, der innerhalb der Seite eindeutig ist. Dazu dient das Attribut `id` (vergl. [2]). HTML-IDs können u. a. dafür genutzt werden CSS-Styles anzuwenden, deshalb werden sie regelmäßig in Webseiten genutzt. Allerdings eignet sich die ID naturgemäß nur für Elemente die in jedem Fall nur ein einziges Mal in einer Webseite vorkommen.

## CSS-Klassen

Soll in einer Webseite eine Gruppe von Elementen in einer gleichen Weise formatiert werden, können CSS-Klassen genutzt werden um beliebige Untermengen von Elementen in einer Webseite mit CSS-Attributen zu verknüpfen (vergl. [3]). Diese CSS-Klassen können auch beim Screen Scraping zur Identifikation von Elementen im DOM genutzt werden.

## XPath oder CSS-Selektoren

Viele Programmierbibliotheken zur Analyse und Manipulation von HTML-Seiten stellen eine *DSL* (Domain Specific Language) zur Verfügung, um kontextabhängig

Elemente im DOM auszuwählen. Dazu erlauben sie die Kombination von allen bisher beschriebenen Identifikatoren in einem Pfadmuster. Typische Vertreter sind XPath (vergl. [4]), welches aus der Welt der XML-Techniken stammt und CSS-Selektoren (vergl. [5]), welche aus dem Sprachstandard für Web-Seiten-Formatierung stammt.

## Mustererkennung

Ist die Struktur einer Webseite vor dem Screen Scraping nicht detailliert bekannt, müssen andere Analysemethoden in Betracht gezogen werden. Das Navigationsmenü einer Webseite z. B. kann mit Hilfe von verschachtelten `div`-Elementen, durch ein `ul`-Element oder auch durch ein `table`-Element kodiert werden. Optional kann die Navigationsstruktur in ein HTML5 `nav`-Element eingebettet werden. Um auf einer unbekannten Website das Hauptmenü zu finden, reichen XPath- und CSS-Selektoren deshalb nicht aus.

Auf modernen Websites, die mit dem Ziel der Barrierefreiheit entworfen wurden, können evtl. WIA-ARIA-Attribute helfen, Strukturen in Webseiten zu erkennen (vergl. [6]). Diese Attribute haben u. a. das Ziel, Webseiten für Menschen mit Sehbehinderungen zugänglich zu machen. Für diese Zielgruppe bereiten spezielle Programme oder Browser-PlugIns eine Webseite derart auf, dass sie z. B. mit Hilfe einer Braille-Zeile navigiert werden kann. Eine ähnliche Analyse könnte auch beim Screen Scraping eingesetzt werden.

Die Informationsextraktion aus nahezu beliebigen Websites ist ein aktives Forschungsfeld. Es existieren eine ganze Reihe Veröffentlichungen in diesem Gebiet (vergl. [7–10]). Derartige Algorithmen versuchen typische Strukturen in Webseiten zu identifizieren, ohne ihre genaue Kodierung zu kennen. Auf eine detaillierte Betrachtung dieser Möglichkeiten wird an dieser Stelle verzichtet.

## Erkenntnisse

Programmierbibliotheken für die HTML-DOM-Analyse können helfen, die Implementierung einfach zu halten und dennoch einen hohen Grad an Robustheit zu erzielen.

Die folgende Tabelle enthält eine Auswahl an beliebten Programmierbibliotheken für das Parsen und Analysieren von HTML in verschiedenen Programmiersprachen.

| Name              | Sprache     | Website                                                                                                     |
|-------------------|-------------|-------------------------------------------------------------------------------------------------------------|
| Beautiful Soup    | Python      | <a href="https://www.crummy.com/software/BeautifulSoup/">https://www.crummy.com/software/BeautifulSoup/</a> |
| DOM               | PHP         | <a href="http://php.net/manual/en/book.dom.php">http://php.net/manual/en/book.dom.php</a>                   |
| QueryPath         | PHP         | <a href="http://querypath.org/">http://querypath.org/</a>                                                   |
| jsdom             | JavaScript  | <a href="https://www.npmjs.com/package/jsdom">https://www.npmjs.com/package/jsdom</a>                       |
| TagSoup           | Java        | <a href="http://vrici.lojban.org/~cowan/XML/tagsoup/">http://vrici.lojban.org/~cowan/XML/tagsoup/</a>       |
| clj-tagsoup       | Clojure     | <a href="https://github.com/nathell/clj-tagsoup">https://github.com/nathell/clj-tagsoup</a>                 |
| Html Agility Pack | C# / VB.NET | <a href="http://html-agility-pack.net/">http://html-agility-pack.net/</a>                                   |
| CsQuery           | C# / VB.NET | <a href="https://github.com/jamietre/CsQuery">https://github.com/jamietre/CsQuery</a>                       |
| Tidy              | C++         | <a href="http://www.html-tidy.org/">http://www.html-tidy.org/</a>                                           |
| Xerces            | C++         | <a href="http://xerces.apache.org/xerces-c/">http://xerces.apache.org/xerces-c/</a>                         |
| tagsoup           | Haskell     | <a href="https://hackage.haskell.org/package/tagsoup">https://hackage.haskell.org/package/tagsoup</a>       |

Bei CSS- und XPath-Selektoren sollte eine Balance zwischen spezifisch und generisch gefunden werden. Um ein Element in einer Webseite mit einem Selektor zu beschreiben, kann entweder der gesamte Pfad der Elternelemente evtl. mit Element-Typ und CSS-Klasse spezifiziert werden, was zu einem sehr spezifischen Selektor führt. Oder aber es werden nur die unbedingt notwendigen Pfadelemente im Selektor spezifiziert. In dem einen Fall führen schon kleine Änderungen in der Webseite dazu, dass der Selektor das gesuchte Element nicht mehr findet. In dem anderen Fall kann es schnell passieren, dass der Selektor unbeabsichtigt nicht gesuchte Elemente als Ergebnis liefert. Es muss also das richtige Mittelmaß bei der Formulierung der Selektoren gefunden werden.

Es hat sich bewährt die CSS- oder XPath-Selektoren, die bei der Analyse eingesetzt werden, mit aussagekräftigen Namen zu versehen und in eine Konfigurationsdatei auszulagern. Dies hat den Vorteil, dass im Falle einer kleinen Veränderung in der Webseitenstruktur, lediglich die Konfigurationsdatei angepasst werden muss, und nicht der Programm-Code.

Wurde eine Webseite erkennbar durch ein Server-seitiges Framework generiert, kann es helfen, sich deren Komponenten und die üblicherweise produzierten HTML-Strukturen genauer anzusehen. Es kann nützlich sein, spezialisierte Funktionen zu implementieren, welche sich die durch das Framework produzierten Muster

in der Webseitenstruktur zunutze machen. Z. B. könnte eine Funktion implementiert werden, welche eine immer wieder verwendete Struktur für Pagination-Links erkennt und die URLs der Einzelseiten zurückgibt (vergl. Abschnitt Pagination).

## Herausforderungen

Beim Screen Scraping begegnet man einer Reihe von Herausforderungen, die dadurch entstehen, dass man eine Schnittstelle “missbraucht”. Konkret verwendet man ein Datenaustauschformat (HTML) und eine Informationsstruktur (Aufbau und Navigationsstruktur der Website), die für die grafische Präsentation von Inhalten für einen menschlichen Konsumenten und die interaktive Exploration optimiert sind, aber nicht unbedingt für den Austausch von strukturierten Informationen zwischen maschinellen Systemen.

Viele Aspekte des Datenaustauschs bei Websites sind daher suboptimal für die automatisierte maschinelle Weiterverarbeitung der enthaltenen Informationen. Diese fangen mit der Kommunikation von Fehlersituationen an und erstrecken sich bis hin zur Einbettung eigentlich strukturierter Informationen in unstrukturiertem Fließtext.

- De-normalisierte Präsentation
- Pagination
- Abfragebegrenzungen
- Passwortgeschützte Bereiche
- Fehlende Benachrichtigung über Datenänderungen
- Aktive Inhalte
- Identifikatoren

## De-normalisierte Präsentation

Eine Darstellung, bei der jede Entität mit all ihren Attributen auf mindestens einer Seite vollständig präsentiert wird, soll als *normalisierte Darstellung* bezeichnet werden. Die Erfassung von Informationen auf einer Website, welche eine normalisierte Darstellung verwendet, ist verhältnismäßig einfach, weil für die Rekonstruktion einer Entität immer nur genau eine Webseite erforderlich ist.



Bei der Präsentation von Daten in einer Website, kommt es jedoch häufig vor, dass eine Entität auf mehreren Seiten in unterschiedlichem Detailgrad oder mit einer variierenden Teilmenge von Attributen dargestellt wird. Die Ursache dafür sind üblicherweise wahrnehmungspsychologische Aspekte und Erkenntnisse im Bereich der Gestaltung von Mensch-Maschine-Schnittstellen, die den Aufbau der Website beeinflussen. Auch Optimierungen um Ressourcen zu sparen, können zu einer Aufteilung von Entitäten führen. Wenn z. B. das Ausliefern einer Detailseite wesentlich mehr Server-Leistung in Anspruch nimmt als das einer Übersichtsseite. Solch eine Darstellung soll als *de-normalisierte Präsentation* bezeichnet werden.

Bei der Erfassung von Entitäten, die de-normalisiert präsentiert werden, kann man unterschiedliche Strategien verfolgen.

**1. Erfassung von Teilsichten der Entitäten, die der Darstellung in den Webseiten entsprechen**

Hierbei wird jede Webseite individuell betrachtet und die enthaltenen Informationen direkt in einer Datenbank gespeichert. Ist die Datenbank relational, wird für jede *Facette* (Teilsicht einer Entität) eine eigene Tabelle als Vorstufe verwendet. Erst wenn alle Facetten einer Entität vollständig zusammengetragen wurden, wird die Entität in ihrer Zieltabelle zusammenhängend gespeichert.

**2. Erfassung von vollständigen Entitäten durch die Kombination der Inhalte mehrerer Webseiten**

Bei dieser Strategie werden alle Webseiten, die für die Rekonstruktion einer Entität notwendig sind, direkt hintereinander abgerufen. Die extrahierten Informationen werden im Arbeitsspeicher gehalten und zu einer Entität zusammengefügt. Ist eine Entität vollständig zusammengetragen, wird sie in einer Datenbanktabelle gespeichert.

Der Vorteil der ersten Strategie ist, dass der Prozess jeder Zeit abgebrochen und später wieder fortgeführt werden kann. Sie ist dementsprechend robust gegen Ausfälle der Netzwerkverbindungen oder Störungen auf dem Server oder dem Client. Eine Website muss i. d. R. nur ein einziges Mal abgerufen werden, da alle Informationen die extrahiert werden können sofort in der Datenbank gesichert werden. Der Nachteil sind die zusätzlichen Tabellen, die u.U. redundante Informationen enthalten.

Der Vorteil der zweiten Strategie ist, dass für jeden Typ von Entität nur eine Tabelle benötigt wird. Nachteilig ist jedoch, dass Webseiten mehrfach aufgerufen werden

müssen, sollte die Erfassung durch eine Störung unterbrochen werden. Denn bei einer Unterbrechung der Erfassung, die einen Neustart des erfassenden Programms notwendig macht, gehen die im Arbeitsspeicher vorgehaltenen Informationen von unvollständigen Entitäten verloren.

Bei einer geringen Menge von zu erfassenden Informationen, ist die zweite Strategie wahrscheinlich die effizientere Wahl. Sollen jedoch größere Mengen an Informationen erfasst werden, und das womöglich in regelmäßigen Abständen über einen längeren Zeitraum, ist die erste Strategie empfehlenswert.

## Beispiel

Eine Liste von Unternehmen wird in drei Darstellungsformen auf einer Website präsentiert. Als Einstieg dient eine tabellarische Übersicht aller Unternehmen. Wobei jedoch nur der Name und das Bundesland aufgeführt sind. Ein Link je Unternehmen, dessen Zieladresse eine eindeutige ID enthält, führt zu einer Detailseite. Die Detailseiten enthalten die meisten interessanten Attribute wie *PLZ*, *Ort*, *Anschrift*, *Telefon*. Allerdings führt von jeder Detailseite ein Link auf eine weitere Seite mit dem Unternehmensprofil, das u. a. eine Kontaktperson für das Unternehmen aufführt. Eine Entität (ein Unternehmen) wird also in drei unterschiedlichen Teilsichten de-normalisiert präsentiert.

Die Zieltabelle für Unternehmen könnte wie folgt aussehen:

- unternehmen
  - id
  - zeitstempel
  - name
  - plz
  - ort
  - anschrift
  - bundesland
  - telefon
  - kontaktperson

Mit der ersten Strategie werden drei zusätzliche Datenbanktabellen für die Teilsichten benötigt:

- denorm\_unternehmen\_info
  - id

- url
- zeitstempel
- name
- bundesland
- denorm\_unternehmen\_detail
  - id
  - url
  - zeitstempel
  - plz
  - ort
  - anschrift
  - telefon
- denorm\_unternehmen\_kontakt
  - id
  - url
  - zeitstempel
  - kontaktperson

Zunächst wird die Webseite mit der Übersichtstabelle abgerufen und jedes gelistete Unternehmen in `denorm_unternehmen_info` gespeichert. Dabei wird mit einem Zeitstempel der Zeitpunkt festgehalten, wann die Informationen abgerufen wurden. Für die Spalte `id` wird eine neue ID gebildet. In der Spalte `url` wird dokumentiert, mit welcher URL die Informationen abgerufen wurden. Anschließend wird die Tabelle `denorm_unternehmen_info` nach allen bekannten Unternehmen befragt. Für jedes Unternehmen, für das noch kein Eintrag in `denorm_unternehmen_detail` existiert, wird die Detailseite abgerufen. Die daraus extrahierten Informationen werden in der Tabelle `denorm_unternehmen_detail` gespeichert. Dabei wird für das Feld `id` die generierte ID aus der Tabelle `denorm_unternehmen_info` verwendet. Auch hier wird URL und Zeitpunkt in `url` und `zeitstempel` dokumentiert. Für die dritte Teilsicht `denorm_unternehmen_kontakt` wird genauso verfahren. Als letzten Schritt werden alle drei Tabellen zusammengeführt und in die Tabelle `unternehmen` übertragen. Dabei wird mit einem Zeitstempel der Zeitpunkt festgehalten, wann die Informationen zusammengeführt wurden.

Bei dieser Strategie kann der Prozess im Prinzip jederzeit unterbrochen und neugestartet werden, ohne dass Webseiten mehrfach abgerufen werden müssen. Der Ver-

lauf des Seitenabrufs kann über die einzelnen Zeitstempel genau nachverfolgt werden. Und die Tabelle `unternehmen` beinhaltet zu jedem Zeitpunkt nur vollständig erfasste Unternehmen.

Es stellt sich die Frage, warum für die erste Strategie zusätzliche Tabellen benötigt werden. Die Tabelle `unternehmen` könnte ja auch schrittweise vervollständigt werden. Das klingt verlockend bringt jedoch Probleme mit sich. I. d. R. gibt es keine Garantie dafür, dass für alle Entitäten jede Teilsicht verfügbar ist. So könnte z. B. ein Unternehmen eine Profilseite ohne Kontaktperson besitzen. Würde man nur die Tabelle `unternehmen` verwenden, müsste man aus einem leeren Feld `kontaktperson` schließen, dass die Profilseite noch nicht abgerufen wurde. Der Abruf der Profilseite steht also noch aus. Enthält die Profilseite aber keine Kontaktperson, gibt es keine elegante Möglichkeit, zu dokumentieren, dass die Profilseite sehr wohl schon abgerufen wurde, jedoch kein Ergebnis lieferte. Durch die zusätzliche Tabelle `denorm_unternehmen_kontakt` ist es möglich einen Eintrag mit `web_id` und `zeitstempel` zu speichern, der ein leeres Feld `kontaktperson` besitzt. Damit ist eindeutig dokumentiert, dass die Profilseite abgerufen wurde, es aber keine Kontaktperson gibt. Ein weiterer Nachteil würde darin bestehen, dass für einen Konsumenten der Tabelle `unternehmen` nicht ersichtlich ist, welche Unternehmen vollständig erfasst wurden und welche nicht.

Mit der zweiten Strategie wird zunächst die Webseite mit der Übersichtstabelle abgerufen und die Liste der Unternehmen im Arbeitsspeicher gehalten. Für jedes Unternehmen werden sowohl die Detailseite als auch das Unternehmensprofil abgerufen und der Datensatz für das Unternehmen anschließend mit einer neuen eindeutigen ID in die Tabelle `unternehmen` gespeichert. Dabei wird mit einem Zeitstempel der Zeitpunkt festgehalten, wann die Informationen zusammengeführt wurden.

Wird bei dieser Strategie der Prozess unterbrochen und neugestartet, muss in jedem Fall die Übersichtsseite erneut abgerufen werden und evtl. auch die Detailseite für ein Unternehmen. Der genaue Verlauf des Seitenabrufs kann nicht nachverfolgt werden, da nur der Zeitpunkt für das Zusammenführen der Informationen gespeichert wird. Auch bei dieser Strategie beinhaltet die Tabelle `unternehmen` zu jedem Zeitpunkt nur vollständig erfasste Unternehmen.

## Pagination

Pagination ist eine Technik, bei der eine Menge von Elementen auf mehrere Seiten verteilt wird. Das hat den Zweck, dass z. B. bei der Beantwortung einer Suchanfrage nicht sofort alle Ergebnisse gefunden, übertragen und dargestellt werden müssen, sondern erst wenn der Benutzer tatsächlich an einer größeren Anzahl von Ergebnissen interessiert ist. In vielen Szenarien wird auch eine Sortierung der Ergebnisse vorgenommen, welche die Wahrscheinlichkeit erhöht, dass sich die Elemente, für die sich der Benutzer interessiert, unter den zuerst präsentierten Ergebnissen befinden.

Üblicherweise wird eine fest vorgegebene oder vom Benutzer wählbare maximale Anzahl von Elementen verwendet, um die Elemente in Untermengen aufzuteilen. Der Benutzer bekommt, wenn er zu einer Ansicht mit mehreren Elementen navigiert oder eine Suchanfrage abschickt, nur die erste Ergebnisseite, bzw. Untermenge von Elementen, präsentiert und muss Navigationsschaltflächen wie z. B. *Weiter* oder *Nächste Seite* verwenden, um zur nächsten Untermenge von Elementen zu navigieren.

In vielen Fällen werden neben *Zurück* und *Weiter* auch nummerierte Navigationsschaltflächen angezeigt, die es dem Benutzer erlauben, direkt zur N-ten Untermenge zu springen. Dabei sind jedoch wiederum nicht immer alle Untermengen mit einer nummerierten Navigationsschaltfläche vertreten, sondern oft nur eine begrenzte Anzahl vor und nach der aktuell angezeigten Untermenge.

Um alle Elemente einer paginierten Menge mit Screen Scraping zu erfassen, gibt es zwei mögliche Strategien: Den Kettenabruf und den Direktabruf.

## Kettenabruf

Der Kettenabruf folgt der Navigationsstrategie der explorativen Traversierung. Auf jeder aktuell vorliegenden Seite, beginnend mit der ersten, wird nach der Navigationsschaltfläche gesucht, welche zu einer Seite mit der nachfolgenden Untermenge von Elementen führt. Ist die Schaltfläche vorhanden und aktiv, kann davon ausgegangen werden, dass noch weitere Elemente existieren. Es wird demzufolge die verlinkte Seite abgerufen und gleichsam untersucht.

## Vorteile

- Einfacher Algorithmus
- Unabhängigkeit vom konkreten Aufbau der URL

## Nachteile

- Abhängigkeit von Navigationsschaltflächen

Der Kettenabruf ist i. d. R. besser geeignet als der Direktabruf, wenn die Navigationsschaltfläche leicht identifiziert werden kann, die Gesamtanzahl der Elemente nicht bekannt ist und es nicht auf eine hohe Erfassungsgeschwindigkeit ankommt.

## Direktabruf

Der Direktabruf folgt der Navigationsstrategie des Reverse Engineering von URLs. Die URL der paginierten Ergebnismenge wird auf Parameter für die Pagination untersucht und es werden direkt abgeleitete URLs für die verschiedenen Untermengen gebildet. Sollte ein Parameter für die Anzahl der Elemente je Untermenge existieren, bietet es sich an, diesen Parameter zu maximieren, um die Anzahl der notwendigen Aufrufe, für das Erfassen der gesamten Menge von Elementen, zu minimieren.

## Vorteile

- Parallelisierung möglich
- Mehr Kontrolle über die Abfrage-Parameter (z. B. mehr Elemente pro Seite)

## Nachteile

- Abhängigkeit vom konkreten Aufbau der URL
- Ohne Informationen über Gesamtanzahl von Elementen ineffizient

Der Direktabruf kann sowohl verwendet werden, wenn die Gesamtanzahl der gesuchten Elemente bekannt ist, als auch wenn diese unbekannt ist. Wenn die Gesamtanzahl der Elemente bekannt ist, kann, unter Zuhilfenahme der maximalen Anzahl von Elementen je Seite, direkt die benötigte Menge von URLs gebildet werden. Ist die Gesamtanzahl jedoch nicht bekannt, müssen die URLs inkrementell gebildet und iterativ abgerufen werden. Als Abbruchkriterium kann der Erhalt einer Fehlerseite, der HTTP-Status-Code 404 oder der Erhalt einer Seite ohne Elemente dienen.

Der Direktabruf ist i. d. R. besser geeignet als der Kettenabruf, wenn die URL einfach aufgebaut ist, sich über die URL-Parameter mehr Möglichkeiten als die Navigations-schaltflächen bieten und wenn eine besonders effiziente Abfrage wichtig ist.

Wenn das Screen-Scraping-System auf Effizienz und Erfassungsgeschwindigkeit optimiert werden soll, ist es wichtig, die Abfrage mit dem Betreiber der Webseite abzustimmen, da z. B. durch eine parallelisierte und hochfrequente Abfrage der Betrieb bzw. die Erreichbarkeit der Webseite beeinträchtigt werden kann.

## Abfragebegrenzungen

Auf einigen Websites steht eine Suchfunktion zur Verfügung, welche jedoch nur eine begrenzte Anzahl von Ergebnissen liefert. Erfolgt die Suche nur mit einem einzigen Kriterium, kann man gegen diese Begrenzung nichts unternehmen. Stehen jedoch mehrere Kriterien zur Einschränkung der Suche zur Verfügung, ist es prinzipiell möglich, den Suchraum durch geschickte Wahl von zusätzlichen Kriterien derart zu partitionieren, dass mit mehreren Suchanfragen doch alle Ergebnisse zurückgegeben werden.

Mathematisch betrachtet, wird dabei ein mehrdimensionaler Raum derart in Partitionen aufgeteilt, dass die Vereinigung aller Partitionen den gesamten Raum abdeckt. Die Ergebnismenge für jede Partition muss dabei kleiner als die Abfragegrenze sein. Werden die Partitionen jedoch zu klein gewählt, werden unnötig viele Suchanfragen erzeugt.

Bei einigen Arten von Kriterien sind überlappende Partitionen notwendig, um den gesamten Suchraum abzudecken. Um solche Abfragen effizient zu gestalten, ist es erstrebenswert, die Überlappung der Partitionen zu minimieren. Des Weiteren müssen bei Ergebnissen aus überlappenden Partitionen mögliche Duplikate entfernt werden.

## Beispiel

Das folgende Beispiel soll den Einsatz von Partitionen veranschaulichen. Ein Suchformular für Veranstaltungen in Deutschland besitzt die folgenden Auswahlfelder:

- *Name*
- Zeitraum mit *Datum von* und *Datum bis*
- *PLZ* und *Umkreis* in Kilometern

Zielstellung ist es, alle Veranstaltungen zu erfassen, die im Juli 2017 stattgefunden haben. Wird als Zeitraum 01.07.2017 bis 31.07.2017 eingegeben, ergibt die Suche genau 200 Treffer. Das Datum der ersten 50 Veranstaltungen ist jedoch einheitlich der 01.01.2017. Das legt die Vermutung nahe, dass nicht alle Suchergebnisse ausgegeben werden, sondern die Trefferanzahl immer auf 200 begrenzt ist.

Nun kann der Suchraum mit Hilfe der Suchkriterien partitioniert und für jede Partition eine eigene Abfrage gestartet werden. In diesem Fall bietet sich das Datum als erste Stufe für eine Partitionierung an. Anstatt eine Suchanfrage für den gesamten Zeitraum zu verwenden, kann für jeden Tag eine eigene Suchanfrage gestellt werden. Bei der Partitionierung mittels des Datums sind die Partitionen disjunkt solange sich die als Kriterium angegebenen Zeiträume nicht überlappen. Zunächst scheint das Problem der Abfragebegrenzung gelöst. Um alle Veranstaltungen im Juli zu erfassen, wurden 31 Abfragen durchgeführt. Doch eine nähere Prüfung zeigt, dass es drei Tage gibt, an denen die Suche wieder genau 200 Ergebnisse liefert. Folglich reicht die Partitionierung mittels des Datums noch nicht aus.

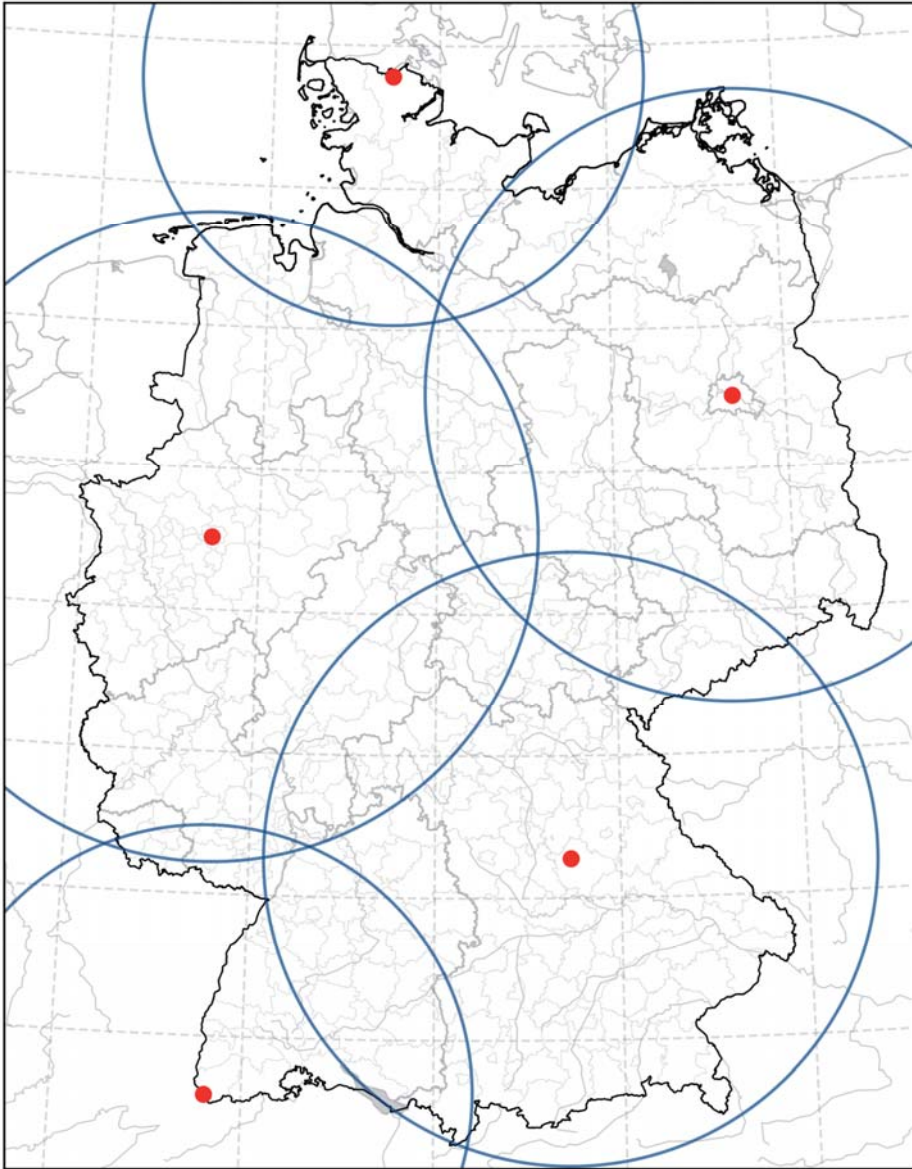
Als zweite Stufe für eine Partitionierung bieten sich die Felder *PLZ* und *Umkreis* an. Da mit einer Postleitzahl als Zentrum und einem Radius für die Ausdehnung nur eine kreisförmige Region beschrieben werden kann, muss eine Menge von Kreisen gefunden werden, die Deutschland vollständig abdeckt – in der sich die Kreise jedoch möglichst wenig überlappen (vergl. Abbildung 1). Zusätzlich muss das jeweilige Zentrum der Kreise an einem geografischen Ort liegen, der mit einer Postleitzahl beschrieben werden kann. Die angestrebte Größe der Kreise ergibt sich aus der Anzahl von möglichen Suchergebnissen in einer Region. Als Ausgangspunkt könnte Deutschland durch fünf Kreise abgedeckt werden, die jeweils durch eine Kombination von *PLZ* und *Umkreis* beschrieben sind.

| PLZ   | Ort                       | Radius |
|-------|---------------------------|--------|
| 10115 | Berlin                    | 240    |
| 24939 | Flensburg                 | 195    |
| 44147 | Dortmund                  | 255    |
| 79576 | Weil am Rhein             | 210    |
| 92318 | Neumarkt in der Oberpfalz | 240    |

Sollte der Fall eintreten, dass auch mit dieser Partitionierung die Abfragegrenze erreicht wird, muss eine Aufteilung in kleinere Regionen gefunden werden. Da durch



die Natur der kreisförmigen Partitionen Überlappungen notwendig sind, müssen die Suchergebnisse, nach dem Erfassen aller Partitionen, von Duplikaten befreit werden.



Fünf kreisförmige Regionen, die Deutschland vollständig abdecken

## Passwortgeschützte Bereiche

Wenn Inhalte von Seiten erfasst werden sollen, welche durch eine Zugangskontrolle geschützt sind, sind die folgenden Aspekte zu berücksichtigen:

Zunächst ist die Legalität des Erfassens besonders zu hinterfragen. Der Schutz der Inhalte durch einen Authentifizierungsschritt kommuniziert die Intention, dass die Inhalte nur einem begrenzten Personenkreis zugänglich gemacht werden sollen. Daher ist zu prüfen, ob das Erfassen der Daten oder die beabsichtigte Weiterverwendung der Daten den berechtigten Interessen desjenigen zuwiderläuft, der die Daten über die Website bereitstellt.

Des Weiteren ist die technische Umsetzung der Zugangskontrolle im Einzelfall zu berücksichtigen. Es existiert ein breites Spektrum an technischen Lösungen für die Zugangskontrolle in Websites. Grundsätzlich kann man zustandsbehaftete und zustandslose Lösungen unterscheiden.

Zustandsbehaftete Lösungen erzeugen bei erfolgreichem Login einen Datensatz auf dem Server, welcher dokumentiert, dass dem Benutzer Zugang gewährt wurde. Dieser Datensatz ist verknüpft mit der aktuellen Sitzung (*Session*) des Benutzers. Die Sitzung wird dem Benutzer durch eine temporäre ID zugeordnet, welche bei jeder Anfrage mitgeschickt werden muss.

Zustandslose Lösungen stellen dem Benutzer bei erfolgreichem Login ein Token aus, welches Informationen beinhaltet, die ausreichen, damit der Server ein valides Login wiedererkennen kann. Tokens arbeiten i. d. R. mit einer digitalen Unterschrift und einem Verfallsdatum. Auch Tokens müssen nach dem Login bei jeder nachfolgenden Anfrage mitgeschickt werden.

Zustandslose Lösungen haben eine Reihe von Vorteilen für die Server-Architektur, machen aber im Kontext des Screen Scrapings keinen wesentlichen Unterschied aus. In beiden Fällen muss dem Erfassen von Informationen eine Login-Anfrage vorangestellt werden. Im Falle eines erfolgreichen Logins muss ein Teil der Antwort aufbewahrt und bei jeder nachfolgenden Anfrage mitgeschickt werden. Der einzige Unterschied besteht darin, dass bei zustandsbehafteten Lösungen nach dem Erfassen der Daten eine Logout-Anfrage geschickt werden sollte, damit der Server die Möglichkeit bekommt die Sitzung zu zerstören. Dies ist bei zustandslosen Lösungen nicht notwendig, da das Verwerfen des Tokens einem Logout gleichkommt.

Es gibt eine Vielzahl von Varianten bei der Übermittlung der Sitzungs-ID bzw. dem Token bei Anfragen. U. a. sind die folgenden Wege üblich:

- Übergabe als Abfrageparameter in der URL
- Übergabe in einem Cookie
- Übergabe als standardisiertes oder proprietäres HTTP-Header-Feld (z. B. *Authorization*)
- Übergabe als Formular-kodiertes Feld in POST-Anfragen
- Übergabe als Teil einer kodierten Datenstruktur im Körper der Anfrage (z. B. XML oder JSON)

## Fehlende Benachrichtigung über Datenänderungen

Soll ein Datenbestand für eine Anzahl von Entitäten für die Weiterverarbeitung vorgehalten werden, können die Informationen der Entitäten mittels Screen Scraping aus einer Website extrahiert werden. Werden die Informationen der Entitäten auf der Website jedoch aktualisiert, veraltet der durch Screen Scraping abgerufene Datenbestand.

Ein einfacher, wenn auch nicht immer praktikabler Weg ist es, das Screen Scraping immer nur dann auszuführen, wenn die Daten in aktueller Form benötigt werden. Das wird als *synchrones Screen Scraping* bezeichnet. Ist der Datenbestand entsprechend groß, das Screen Scraping zeitaufwändig und werden die Daten häufig weiterverwendet, muss das Screen Scraping von der Weiterverarbeitung zeitlich entkoppelt werden. Dann sind regelmäßige Abrufe der Website erforderlich, um die Daten aktuell zu halten. Das wird als *asynchrones Screen Scraping* bezeichnet. Welches Intervall angebracht ist und welche Webseiten abgerufen werden müssen, richtet sich dabei nach der Natur der Entitäten.

## Beispiel

Eine Website stellt eine Liste mit Veranstaltungen zur Verfügung. Sie präsentiert dabei bereits bekannte Veranstaltungen in der Zukunft, und die Veranstaltungen aus der Vergangenheit, die nicht länger als 12 Monate zurückliegen. Die Veranstaltungen sind auf mehrere Webseiten verteilt, wobei jede Webseite die Veranstaltungen für jeweils einen Monat enthält. Es kommt regelmäßig vor, dass Veranstaltungen abgesagt werden und dann aus der Liste verschwinden.

Soll diese Liste durch asynchrones Screen Scraping für die Weiterverarbeitung erfasst werden, können die Webseiten für die verschiedenen Monate mit unterschiedlichen Intervallen abgefragt werden. Monate die in der Vergangenheit liegen, brauchen nur einmal abgefragt werden. Denn es ist nicht zu erwarten, dass sich die Informationen zu Veranstaltungen in der Vergangenheit ändern. Anders sieht es mit dem aktuellen Monat und den Monaten in der Zukunft aus. Monate die weit in der Zukunft liegen, müssen nur relativ selten aktualisiert werden. Der aktuelle Monat und die in naher Zukunft liegenden Monate hingegen sollten jeden Tag aktualisiert werden.

## Aktive Inhalte

Viele moderne Websites verwenden aktive Inhalte, wie z. B. JavaScript, um die Interaktivität und den Komfort für menschliche Besucher zu steigern. Dabei kann die Aufgabe der aktiven Anteile der Website von einfachen Animationen bis hin zu wesentlichen Funktionalitäten der Website reichen. Eine Extremform sind die sogenannten Single-Page-Websites (vergl. [11]).

Immer dann, wenn gesuchte Informationen nur dann angezeigt werden, wenn aktive Inhalte ausgeführt werden, sind sie für das Screen Scraping relevant. Sobald eine Website von JavaScript Gebrauch macht, ist es deshalb i. d. R. notwendig die Rolle des JavaScript-Codes in der Website zu untersuchen. Ein einfacher Test ist, JavaScript im Browser zu deaktivieren und anschließend zu überprüfen, ob die Website noch die gesuchten Informationen darstellt (vergl. [12], [13]).

Sollte JavaScript-Code verantwortlich dafür sein, dass die gesuchten Informationen angezeigt werden, kann das unterschiedliche Konsequenzen für das Screen Scraping haben. Im Folgenden sollen einige mögliche Fälle beschrieben werden.

1. Sichtbarkeitsänderungen
2. AJAX mit HTML-Inhalten
3. AJAX mit strukturierten Inhalten (z. B. JSON oder XML)
4. Berechnung von Inhalten im Browser

In einigen Fällen enthält der HTML-Code bereits die gesuchten Informationen, er ist jedoch mittels CSS ausgeblendet und wird erst nach dem Betätigen einer Schaltfläche, die JavaScript-Code ausführt, sichtbar. In diesen Fällen kann der JavaScript-

Code ignoriert werden, da die Informationen beim Screen Scraping, unabhängig vom referenzierten CSS, direkt aus dem HTML extrahiert werden können.

Wird das JavaScript dazu eingesetzt, dynamisch HTML-Code-Blöcke nachzuladen (AJAX), welche die gesuchten Informationen enthalten, können die URLs untersucht und nachgebildet werden, mit denen die HTML-Schnipsel nachgeladen werden. Anschließend werden die HTML-Schnipsel zum Ziel für das Screen Scraping.

Wird das JavaScript dazu eingesetzt, dynamisch strukturierte Daten nachzuladen (AJAX), welche die gesuchten Informationen enthalten, verlässt die Lösung beinahe den Bereich des Screen Scraping. Denn offensichtlich stellt der Web-Server die gesuchten Informationen über entsprechende URLs in strukturierter Form, z. B. JSON oder XML, zur Verfügung und das Analysieren des HTML-Codes entfällt.

Eine Sonderrolle nimmt der Fall ein, dass die gesuchten Informationen gar nicht vom Web-Server geladen werden, sondern durch ein JavaScript-Programm im Browser berechnet werden. In einem solchen Fall könnte es möglich sein, den JavaScript-Code zu extrahieren und offline, vollständig ohne den Kontakt zum Web-Server, auszuführen. Das Format der Datenausgabe kann in diesem Fall durch Anpassung des JavaScript-Codes selbst bestimmt werden. Wird der Quellcode des JavaScript-Programms in einem solchen Szenario eingesetzt und evtl. angepasst, müssen die Lizenzbestimmungen beachtet werden, unter denen das JavaScript-Programm veröffentlicht wird.

Ist der Aufbau der Website komplex und wird JavaScript an vielen Stellen eingesetzt, so dass ein Reverse Engineering der Abläufe bis hin zum Ursprung der gesuchten Informationen zu aufwändig erscheint, kann sich der Einsatz von *PhantomJS* lohnen (vergl. [14]). PhantomJS ist ein Browser der keine grafische Ausgabe besitzt, jedoch durch eigenen JavaScript-Code gesteuert werden kann. Mit PhantomJS ist es möglich, Webseiten aufzurufen und wie in einem gewöhnlichen Browser mit dem Nachladen von Ressourcen und dem Ausführen von JavaScript-Code aufzubauen. Anschließend kann eigener JavaScript-Code mit der Website interagieren, den dynamisch entstehenden DOM analysieren und die gesuchten Informationen extrahieren.

## Identifikatoren

Die Entitäten die beim Screen Scraping erfasst werden, müssen eindeutig identifizierbar sein. Das ist vor allem wichtig, wenn das Screen Scraping wiederholt durchgeführt und dabei ein bereits existierender Datenbestand aktualisiert wird. Dabei ist entscheidend, in welchem Kontext die verwendeten Identifikatoren eindeutig sind.

Werden auf einer Website Detailseiten für die Entitäten präsentiert, existiert mit hoher Wahrscheinlichkeit bereits ein Identifikator – z. B. eine Zahl oder eine alphanumerische Zeichenkette – der in der URL für die Detailseite enthalten ist. Es ist jedoch zu beachten, ob der Identifikator nur in einer Klasse von Entitäten oder für alle Entitäten eindeutig ist.

Weiterhin ist wichtig zu erkennen, ob ein Identifikator *natürlich* oder *künstlich* ist. Natürliche Identifikatoren leiten sich direkt aus den wesentlichen Eigenschaften einer Entität ab. Besitzen die gesuchten Entitäten z. B. eine Eigenschaft, in der sie sich eindeutig unterscheiden und die naturgemäß unveränderlich ist, kann diese Eigenschaft als *natürlicher Identifikator* genutzt werden. *Künstliche Identifikatoren* sind semantisch unabhängige Daten, deren wichtigste Eigenschaft die Eindeutigkeit in einem definierten Kontext ist. Das können inkrementierte Zahlen sein, zufällige Zahlen deren Eindeutigkeit bei der Vergabe gesichert wird, zufällige alphanumerische Zeichenfolgen oder UUIDs bzw. GUIDs (vergl. [15]). Ist es schwierig für die Entitäten einen Identifikator aus den Eigenschaften zu bilden, werden i. d. R. künstliche Identifikatoren eingesetzt.

Natürliche Identifikatoren besitzen den Reiz, dass sie direkt aus einer Entität abgeleitet werden können, egal woher deren Informationen stammen. Sie bergen jedoch auch das Risiko, dass sie aus einer Eigenschaft abgeleitet werden, von der angenommen wurde, dass sie sich nicht ändern kann, sich in der Zukunft aber doch ändert.

Künstliche Identifikatoren besitzen den Reiz, dass sie Entitäten stabil identifizieren, auch wenn sich deren Eigenschaften wesentlich verändern, ziehen jedoch zusätzlichen Aufwand nach sich, wenn Dubletten erkannt und evtl. zusammengeführt werden müssen. Des Weiteren ist bei künstlichen Identifikatoren unbedingt der Kontext zu beachten in dem sie eindeutig sind. In diesem Sinne gibt es schwache Identifika-

toren, die nur in einem genau definierten Kontext eindeutig sind – z. B. inkrementierte Ganzzahlen – und starke Identifikatoren wie UUIDs, die nahezu uneingeschränkt eindeutig sind.

Die im Folgenden beschriebene Strategie bietet eine hohe Sicherheit bei verhältnismäßig geringem Aufwand:

Beim Erfassen der Rohdaten werden, soweit möglich, Identifikatoren verwendet, die auch in der Website und deren URLs verwendet werden. Beim Zusammenführen der Teilinformationen zu vollständigen Entitäten werden neue künstliche Identifikatoren vergeben und in den Rohdaten nachträglich ergänzt. Dadurch ist ein Abgleich der Rohdaten mit neu erfassten Daten von der Website leicht möglich, da die Identifikatoren der Website verwendet werden können. Und auch der Abgleich der Rohdaten mit den Entitäten ist leicht möglich. Der neue künstliche Identifikator hat zusätzlich den Vorteil, dass Entitäten aus verschiedenen Quellen – z. B. mehreren Websites mit ähnlichen Inhalten – leichter zusammengeführt werden können, da keine Kollision von Identifikatoren aus unterschiedlichen Quellen auftreten können.

## Fehlerquellen

Es gibt eine ganze Reihe von Fehlerquellen, die sich bereits gut während des Entwurfs oder der Implementierung eines Screen-Scraping-Programms berücksichtigen lassen. Im Folgenden sollen die wesentlichen Fehlerquellen genannt und Möglichkeiten zur Begegnung beschrieben werden.

- Netzausfälle und Server-Fehler
- Änderungen an der Struktur der Website

## Netzausfälle und Server-Fehler

Bei länger laufenden Screen-Scraping-Abfragen kann es immer wieder vorkommen, dass auf der Netzwerkebene Unterbrechungen auftreten oder der Web-Server einen internen Fehler meldet. Deshalb ist es wichtig, ein Screen-Scraping-Programm so zu konzipieren, dass es mit Störungen bei HTTP-Anfragen umgehen kann. Zunächst muss sichergestellt werden, dass Fehler aus der Netzwerkebene korrekt abgefangen werden. Dazu gehören z. B. Namensauflösungsfehler, Zeitüberschreitungen

beim Verbindungsaufbau, unerwartete Verbindungsabbrüche, unvollständige Antworten, leere Antworten und HTTP-Statuscodes im Bereich 5xx. Dann muss eine Strategie für den Umgang mit dem Fehler implementiert werden. Zunächst sollte der Fehler in irgendeiner Form protokolliert werden. Dies ist für eine spätere Fehlersuche unerlässlich.

Anschließend kommen verschiedene mögliche Reaktionen auf eine fehlgeschlagene HTTP-Abfrage in Frage, z. B.:

1. Programm sofort beenden
2. Fehler ignorieren und zur Abfrage der nächsten Seite übergeben, falls eine weitere abzurufende URL bekannt ist
3. Sofort erneut versuchen
4. Einige Zeit warten und anschließend erneut versuchen

Wenn ein erneuter Versuch unternommen wird, kann ein Zähler inkrementiert, und der Vorgang bei einer maximalen Anzahl an Versuchen abgebrochen werden.

Wichtig ist, dass ein Fehler auf der Netzwerkebene oder ein interner Server-Fehler nicht derart interpretiert wird, dass eine abzurufende Entität nicht mehr existiert. Vielmehr ist ihr Zustand solange unbekannt, bis eine erfolgreiche Abfrage durchgeführt wurde. Eine Antwort mit einem HTTP-Statuscode 404 kann hingegen oft so interpretiert werden, dass die Entität nicht mehr existiert. Als Folge kann sie im lokalen Datenbestand gelöscht oder als gelöscht markiert werden.

## Änderungen an der Struktur der Website

Wird die Struktur oder das Design einer Website durch den Website-Betreiber geändert, ist es sehr wahrscheinlich, dass das Screen Scraping die gesuchten Informationen nicht mehr extrahieren kann. Grundsätzlich lässt sich sagen, dass Änderungen an der Struktur einer Website unvorhersehbar umfangreiche Anpassungen am Screen-Scraping-Programm erforderlich machen.

Es wäre möglich, dass die Navigationsstruktur der Website geändert wird und Unterseiten über einen anderen Pfad erreichbar werden. Dabei ändert sich nicht zwangsläufig aber möglicherweise auch die URL einer Web-Seite mit gesuchten Informationen. Als Konsequenz müssen zukünftig lediglich die URLs für das Screen Scraping anders gebildet werden.



Es wäre aber auch möglich, dass der HTML-Code neu strukturiert wird, ohne dass sich das sichtbare Design der Website wesentlich ändert – z. B. in einem sog. Refactoring. In diesem Fall kann es notwendig sein, die ganze Analysestrategie neu zu konzipieren und evtl. wesentliche Teile des Screen-Scraping-Programms neu zu schreiben.

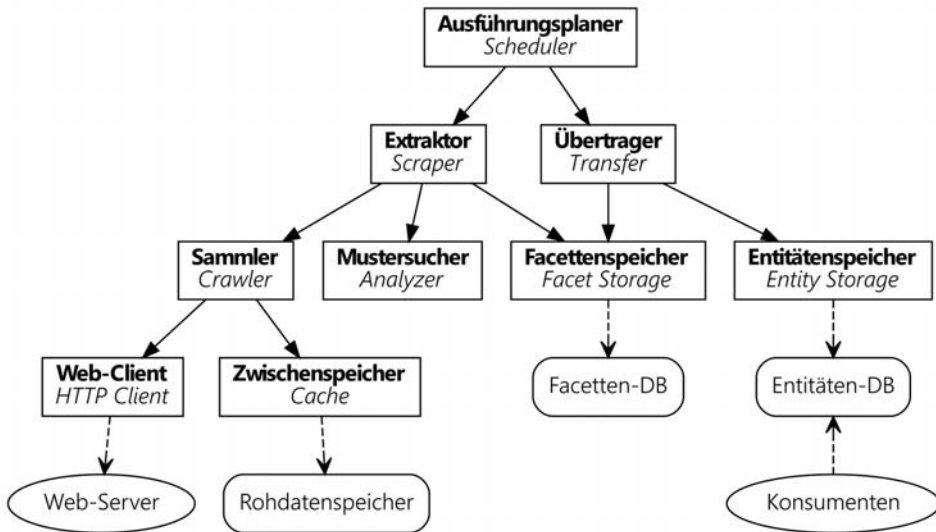
Die wichtigste Erkenntnis in diesem Kontext ist, dass Screen Scraping als regelmäßige Datenquelle nur zuverlässig funktioniert, wenn permanent ein Softwareentwickler mit den erforderlichen Kenntnissen zur Verfügung steht, um das Screen-Scraping-Programm bei Bedarf an Änderungen der Website anzupassen.

## Architektur

In diesem Abschnitt wird eine beispielhafte Architektur vorgestellt, welche die Erkenntnisse aus den vorangegangenen Abschnitten berücksichtigt. Sie kann als Vorlage verwendet werden, unabhängig von der gewählten Navigations-, Abfrage- und Analysestrategie.

Die Architektur gliedert sich in *Module* und *Fremdsysteme*. Einige Module sind generischer Natur und können in mehreren Screen-Scraping-Systemen wiederverwendet werden. Andere Module müssen für jedes Screen-Scraping-System individuell entworfen und implementiert werden, da ihre Logik stark von der Struktur der Website und den gesuchten Informationen abhängt.

- Module
- Fremdsysteme
- Allgemeine Aspekte



Architekturübersicht

## Module

Im Folgenden werden die einzelnen Module der Architektur beschrieben. Ziel ist es, dass jedes Modul im Wesentlichen nur eine Aufgabe wahrnimmt.

| Name              | Englisch       | Komponente           | Typ                     | Strategie  |
|-------------------|----------------|----------------------|-------------------------|------------|
| Ausführungsplaner | Scheduler      | Scheduler            | generisch               | Abfrage    |
| Extraktor         | Scraper        | Scraper              | individuell             | Navigation |
| Sammler           | Crawler        | Scraper              | generisch               | Abfrage    |
| Web-Client        | HTTP Client    | Scraper              | generisch               | Abfrage    |
| Zwischenspeicher  | Cache          | Scraper              | generisch               |            |
| Mustersucher      | Analyzer       | Scraper              | individuell             | Analyse    |
| Facettenspeicher  | Raw Storage    | Scraper,<br>Transfer | individuelles<br>Schema |            |
| Übertrager        | Transfer       | Transfer             | individuell             |            |
| Entitätspeicher   | Entity Storage | Transfer             | individuelles<br>Schema |            |

## Ausführungsplaner

Der *Ausführungsplaner* startet die Extraktion der Informationen aus den Web-Seiten und die Zusammenführung von Entitäten in den Entitätenspeicher. Je nach Anwendungsfall kann dies zu regelmäßigen oder zufälligen Zeitpunkten erfolgen. Seine Aufgabe beschränkt sich auf das Anstoßen des Screen Scrapings. Er implementiert den Aspekt des Abfragezeitpunkts aus der Abfragestrategie und begegnet damit der Herausforderung von fehlenden Benachrichtigungen bei Datenänderungen.

Der Ausführungsplaner ist unabhängig von den Spezifika einer Website und daher ein *generisches* bzw. wiederverwendbares Modul.

### Merkmale

- Modul: Scheduler
- Systemkomponente: Scheduler
- Typ: generisch
- Aufgabe: Zeitsteuerung für den Extraktor und den Übertrager
- Strategie: Abfragestrategie

## Extraktor

Der *Extraktor* bildet aus der Konfiguration und dem Zustand des Facettenspeichers eine Menge von Ziel-URLs. Diese Ziel-URLs übergibt er dem Sammler, um die HTML-Quellen abzufragen. Das Ergebnis dieser Abfragen wird dem Mustersucher übergeben, der entweder Facetten von Entitäten findet oder, der Navigationsstrategie der explorativen Traversierung folgend, weitere URLs. Gefundene Informationen werden im Facettenspeicher abgelegt. Nicht mehr existierende Informationen, die während einer früheren Abfrage extrahiert und im Facettenspeicher abgelegt wurden, werden aus dem Facettenspeicher entfernt oder als gelöscht markiert. Er implementiert die Navigationsstrategie und muss den meisten Herausforderungen begegnen.

Der Extraktor ist stark von der Navigationsstrategie und dem Aufbau der Website abhängig, daher muss er für jedes Screen-Scraping-System *individuell* entworfen und implementiert werden.

## Merkmale

- Modul: Scraper
- Systemkomponente: Scraper
- Typ: individuell
- Aufgabe: Abarbeitung der Ziel-URLs unter Nutzung der Schnittstellen von Sammler und Mustersucher
- Strategie: Navigationsstrategie

## Sammler

Der *Sammler* nimmt URLs entgegen und liefert die HTML-Quelle oder eine Fehlermeldung zurück. Er greift zunächst auf den Zwischenspeicher zu, um die HTML-Quelle für eine URL abzufragen. Erst wenn dieser keinen Treffer liefert, übergibt er die URL dem Web-Client. Erhält er vom Web-Client die HTML-Quelle, speichert er sie für spätere Anfragen im Zwischenspeicher. Er kann auch das Abfragen von referenzierten Ressourcen übernehmen, um Browser-typische Abfragemuster nachzuahmen. Er implementiert auch die Aspekte Verzögerung von HTTP-Anfragen und Parallelisierung der Abfragestrategie.

Der Sammler wird lediglich durch die Abfragestrategie beeinflusst und kann als *generisches* Modul implementiert werden. Dadurch wird der Sammler für mehrere Screen-Scraping-Systemen wiederverwendbar. Die notwendigen Freiheitsgrade zur Anpassung an die jeweilige Abfragestrategie lassen sich leicht durch entsprechend flexible Parametrisierung unterstützen.

## Merkmale

- Modul: Crawler
- Systemkomponente: Scraper
- Typ: generisch
- Aufgabe: Organisation von HTTP-Anfragen
- Strategie: Abfragestrategie

## Web-Client

Der *Web-Client* nimmt URLs entgegen, führt die HTTP-Kommunikation mit dem Web-Server durch, und gibt die HTML-Quelle zurück. Wenn es zu Störungen beim

Abwurf der Web-Seite kommt, gibt er stellvertretend eine Fehlermeldung zurück. Er implementiert den Aspekt HTTP-Header der Abfragestrategie.

Der Web-Client ist unabhängig von den Spezifika der Website und kann als *generisches* Modul implementiert werden. Die zu sendenden HTTP-Header zur Anpassung an die jeweilige Abfragestrategie lassen sich leicht durch entsprechende Parametrisierung unterstützen.

## Merkmale

- Modul: HTTP Client
- Systemkomponente: Scraper
- Typ: generisch
- Aufgabe: HTTP-Kommunikation mit dem Web-Server
- Strategie: Abfragestrategie

## Zwischenspeicher

Der *Zwischenspeicher* nimmt eine URL entgegen und liefert die dazugehörige HTML-Quelle oder einen Fehlschlag zurück. Er kann die Effizienz und die Abfragegeschwindigkeit steigern, wenn URLs mehrfach abgefragt werden. Wichtiger ist seine Aufgabe während der Entwicklungszeit und wenn Anpassungen am Mustersucher vorgenommen werden müssen.

Mit einem Zwischenspeicher ist es prinzipiell möglich alle erforderlichen URLs nur einmal abzufragen und anschließend verschiedenen Varianten bzw. Versionen von Extraktoren und Mustersuchern durchlaufen zu lassen ohne erneut Abfragen an den Web-Server schicken zu müssen. Das erhöht die Entwicklungsgeschwindigkeit maßgeblich und schont die Kapazitäten des Web-Server. Wird im produktiven Betrieb ein Fehler bei der Mustersuche entdeckt, kann dieser behoben, die Extraktion erneut durchgeführt und damit der Rohdatenbestand bereinigt werden.

Der Zwischenspeicher sollte neben der URL und der HTML-Quelle auch einen Zeitstempel speichern, der angibt zu welchem Zeitpunkt die Web-Seite abgerufen wurde. Es sollten auch HTTP-Status-Code und Zugriff- bzw. Übertragungszeit gespeichert werden. Damit enthält der Zwischenspeicher ein vollständiges Protokoll der HTTP-Kommunikation mit dem Web-Server, was im Fehlerfall die Analyse und Fehlerbehebung erleichtert. Zusätzlich sind auch aussagekräftige Statistiken bei einer Überwachung des Screen Scraping möglich.

Der Zwischenspeicher ist von den Spezifika der Website und den gesuchten Informationen unabhängig und kann als *generisches* Modul implementiert werden. Für die eigentliche Speicherung (*Rohdatenspeicher*) kann entweder direkt das Dateisystem oder ein dafür geeignetes Datenbanksystem genutzt werden.

## Schema

- URL
- Zeitstempel
- HTTP-Status-Code
- HTML-Quelle

## Merkmale

- Modul: Cache
- Systemkomponente: Scraper
- Typ: generisch
- Aufgabe: HTML-Quellen für mehrfache Abfragen zwischenspeichern

## Mustersucher

Der *Mustersucher* analysiert den DOM einer HTML-Quelle mit dem Ziel die gesuchten Informationen aus einer Web-Seite zu extrahieren. Je nach Navigationsstrategie und aktuell zu analysierender Web-Seite sucht er nach Facetten von Entitäten, oder nach Navigationsschaltflächen, um weitere URLs für die Abfrage zu ermitteln. Er implementiert damit die Analysestrategie.

Der Mustersucher ist stark vom Aufbau der Web-Seiten und den gesuchten Informationen abhängig und muss daher *individuell* für jedes Screen-Scraping-System implementiert werden. Teilaspekte der Mustersuche können jedoch gut in wiederverwendbare Programmierbibliotheken ausgelagert werden.

## Merkmale

- Modul: Analyzer
- Systemkomponente: Scraper
- Typ: individuell
- Aufgabe: Facetten von Entitäten und URLs in HTML-Quellen finden
- Strategie: Analysestrategie

## Facettenspeicher

Der *Facettenspeicher* nimmt vom Extraktor die extrahierten Informationen entgegen und speichert sie, damit sie vom Übertrager später abgerufen und zusammengeführt werden können. Der Facettenspeicher ermöglicht es, das Extrahieren der Informationen aus den Web-Seiten vom Bilden konsistenter Entitäten zu entkoppeln.

Facetten werden durch Identifikatoren identifiziert, die aus dem Kontext der Website stammen. Sobald aus den Facetten konsistente Entitäten gebildet und im Entitätenspeicher abgelegt wurden, wird zu jeder Facette der Identifikator der abgeleiteten Entität gespeichert.

Es ist ratsam, zu jeder Facette die URL jener Web-Seite zu speichern, aus der ihre Informationen extrahiert wurden. Verschwinden die Informationen einer Entität von der Website, müssen die bereits gespeicherten Facetten der Entität als gelöscht markiert werden. Dazu kann ein Indikator-Feld an den Facetten dienen. Des Weiteren sind zwei Zeitstempel je Facette hilfreich. Einer der den Zeitpunkt der ersten Abfrage festhält, bei der die Informationen der Facette extrahiert wurden. Und ein weiterer für den Zeitpunkt der letzten Abfrage, aufgrund welcher die Facette zuletzt aktualisiert oder bestätigt wurde.

Der Facettenspeicher ist zwar von den Spezifika der Website unabhängig, muss jedoch ein individuelles Schema für die Facetten-Typen unterstützen. Solange das Schema der Facetten flexibel konfiguriert werden kann, ist der Facettenspeicher in mehreren Screen-Scraping-System wiederverwendbar. Für die eigentliche Speicherung ist es sinnvoll, ein Datenbanksystem zu verwenden (*Facetten-DB*).

### Schema

- Facetten-Typ
- Entitäten-Typ
- Identifikator (Website)
- Identifikator (Entitätenspeicher)
- URL
- erster Abfragezeitpunkt
- letzter Abfragezeitpunkt
- Indikator für Löschung

- *Facettendaten*

## Merkmale

- Modul: Raw Storage
- Systemkomponente: Scraper, Transfer
- Typ: individuelles Schema
- Aufgabe: Facetten von Entitäten vom Extraktor entgegennehmen, speichern und wieder an den Übertrager ausliefern

## Übertrager

Der *Übertrager* durchsucht den Facettenspeicher nach Facetten von Entitäten und überträgt vollständige und konsistente Entitäten in den Entitätenspeicher. Findet er Entitäten im Entitätenspeicher, deren Facetten im Facettenspeicher fehlen oder als gelöscht markiert sind, entfernt er diese auch aus dem Entitätenspeicher. Als Ergebnis enthält der Entitätenspeicher ein konsistentes Abbild der aus der Website extrahierten Entitäten.

Beim Zusammenführen der Facetten zu vollständigen Entitäten vergibt er bei Bedarf neue Identifikatoren. Diese neuen Identifikatoren vermerkt der Übertrager auch an den Facetten im Facettenspeicher um sich den erneuten Abgleich zwischen den Facetten und den Entitäten zu erleichtern.

Prinzipiell sollte es jederzeit möglich sein, den Inhalt des Entitätenspeichers zu verworfen und diesen mittels des Übertragers aus dem Facettenspeicher wiederherzustellen. Dies kann sehr hilfreich sein, wenn bei der Zusammenführung von Entitäten Fehler auftreten. Dann kann der Fehler im Übertrager behoben und der Entitätenspeicher neu aufgebaut werden, ohne dass Web-Seiten vom Web-Server erneut abgerufen werden müssen.

Der Übertrager ist abhängig von der Aufteilung der Entitäten in einzelne Facetten. Auch die Art der Identifikatoren von Facetten und Entitäten und die Konsistenzigenschaften der Entitäten beeinflussen die Logik des Übertragers. Daher muss der Übertrager für jedes Screen-Scraping-System *individuell* implementiert werden.

## Merkmale

- Modul: Transfer
- Systemkomponente: Transfer



- Typ: individuell
- Aufgabe: Facetten von Entitäten aus dem Facettenspeicher zusammenführen und konsistente Daten im Entitätenspeicher ablegen

## Entitätenspeicher

Der *Entitätenspeicher* nimmt vom Übertrager konsistente Entitäten entgegen, speichert sie und stellt sie Konsumenten für weitere Verarbeitung zur Verfügung. Bei der Ablage im Entitätenspeicher erhalten neue Entitäten einen künstlichen Identifikator. Werden Facetten einer Entität als gelöscht markiert, wird die Entität aus dem Entitätenspeicher entfernt.

Wird eine Entität aktualisiert, weil sich deren zugrundeliegende Facetten geändert haben, sollte der Zeitpunkt der Aktualisierung festgehalten werden. Dies gilt auch für die erste Zusammenführung einer Entität. Zu jeder Entität sollte ihr Ursprung gespeichert werden. Das ist mindestens dann sinnvoll, wenn Entitäten gleichen Typs aus verschiedenen Websites extrahiert werden.

Der Entitätenspeicher muss ein individuelles Schema für die Entitäten-Typen unterstützen. Solange das Schema der Entitäten flexibel konfiguriert werden kann, ist der Entitätenspeicher in mehreren Screen-Scraping-Systemen wiederverwendbar. Werden mehrere Screen-Scraping-Systeme eingesetzt, um Entitäten aus mehreren Websites zu sammeln, kann ein einzelner Entitätenspeicher als gemeinsame Schnittstelle zwischen den Screen-Scraping-Systemen und den Konsumenten verwendet werden. Für die eigentliche Speicherung ist es sinnvoll, ein Datenbanksystem zu verwenden (*Entitäten-DB*).

Sind die Konsumenten während des Entwurfs eines Screen-Scraping-Systems bekannt, kann bei der Gestaltung der öffentlichen Schnittstelle des Entitätenspeichers auf deren technische Bedürfnisse Rücksicht genommen werden. Andernfalls ist eine standardisierte Schnittstelle wie SQL oder HTTP-REST eine gute Option.

## Schema

- Entitäten-Typ
- Identifikator
- Ursprung
- Zeitpunkt der letzten Aktualisierung
- *Entitätsdaten*

## Merkmale

- Modul: Entity Storage
- Systemkomponente: Transfer
- Typ: individuelles Schema
- Aufgabe: Entitäten vom Übertrager entgegennehmen, speichern und Konsumenten zur Verfügung stellen

## Fremdsysteme

Ein Screen-Scraping-System kommuniziert mit mindestens zwei Fremdsystemen – einem Web-Server und einem Konsumenten.

### Web-Server

Der *Web-Server* ist der Ursprung der extrahierten Informationen. Er liefert die Website über das HTTP-Protokoll aus, in deren Web-Seiten die gesuchten Informationen enthalten sind. Der Web-Server besitzt eine bestimmte Kapazität – Rechenkapazität, Übertragungsrate, Zugriffszeiten, maximale Anzahl paralleler Anfragen – und steht unter der Verantwortung eines Betreibers.

Bei dem Web-Server muss es sich nicht zwingend um einen Hardware-Server oder eine Programminstanz handeln, vielmehr ist er der logische Endpunkt für eine Domain oder einen Dienst. Der Betreiber eines Web-Servers hat viele Möglichkeiten die Kapazitäten eines Web-Servers mit Hilfe von Load-Balancern weit über die Kapazitäten eines einzelnen Hardware-Servers hinaus zu steigern. Umgekehrt kann ein Hardware-Server aber auch eine Vielzahl von Web-Servern ausführen, wobei sich alle Web-Server die Kapazitäten eines Hardware-Servers teilen müssen.

Bei der Entwicklung eines Screen-Scraping-Systems besteht i. d. R. nicht die Möglichkeit auf technische Eigenschaften des Web-Servers Einfluss zu nehmen. Daher muss das Screen-Scraping-System sich in seiner Implementierung und Konfiguration immer am Web-Server und den von ihm ausgelieferten Web-Seiten ausrichten.

### Konsumenten

Die *Konsumenten* sind Systeme, welche über den Entitätenspeicher an das Screen-Scraping-System angebunden sind. Sie sind an konsistenten Entitäten in einem

möglichst aktuellen Zustand interessiert und können diese direkt aus dem Entitätenspeicher abfragen. Konsumenten können die Entitäten beliebig weiterverarbeiten. Sie können entweder im Pull-Prinzip, bei Bedarf oder regelmäßig, auf den Entitätenspeicher zugreifen. Oder sie lassen sich über eine zusätzliche Schnittstelle des Entitätenspeichers über neue Entitäten, Aktualisierungen und Löschungen im Push-Prinzip informieren.

## Allgemeine Aspekte

Bei der Implementierung der Module sollte auf einige allgemeine Aspekte Rücksicht genommen werden.

## Protokollierung

Generell sollte ein Screen-Scraping-System eine ausführliche Protokollierung ermöglichen. Besonders wichtig ist dies für Systeme, die unter Umständen lange benötigen, um alle gesuchten Informationen aus der Website zu extrahieren.

Die Protokollierung dient primär zwei Zielen:

1. Die Ursache für ein Fehlverhalten des Systems ermitteln
2. Die Aktivität des Systems überwachen

Es ist sinnvoll, alle Teilschritte in unterschiedlicher Granularität zu protokollieren. Üblich sind Detailgrade wie z. B. Verbose, Info, Warning und Error. Alle Code-Abschnitte die mit externen Systemen kommunizieren oder unmittelbar mit den abgefragten HTML-Quellen zu tun haben, müssen eine sorgfältige Fehlerbehandlung enthalten. Auftretende Fehler sollten zunächst sofort protokolliert werden (Warning oder Error), bevor entsprechend der Konfiguration und der logischen Möglichkeiten fortgefahren wird.

Jeder Protokolleintrag sollte mit einem präzisen Zeitstempel, dem Detailgrad bzw. Typ der Meldung und dem Namen des Ursprungsmodul versehen sein. Dadurch lässt sich ein Protokoll im Nachhinein leichter filtern und die Vorgänge, die zu einem Fehler geführt haben, sauber rekonstruieren.

## Konfiguration

Es ist sinnvoll, alle Stellgrößen die für die Algorithmen in den Modulen benötigt werden, als Konfigurationsparameter zu implementieren. So dass sie in einer ausgelagerten Konfigurationsdatei leicht verändert werden können und für deren Anpassung keine Änderungen im Quelltext des Screen-Scraping-Systems notwendig sind. Besonders komfortabel ist eine Konfiguration, wenn sie aus mehreren priorisierten Quellen zusammengesetzt wird – z. B. Konfigurationsdatei, Umgebungsvariablen und Befehlszeilen-Parameter.

Eine detaillierte Konfiguration ist sowohl während der Entwicklungszeit des Screen-Scraping-Systems hilfreich, als auch während des produktiven Einsatzes. Während der Entwicklungszeit kann das Verhalten des Systems so schnell für bestimmte Testsituationen oder zur Fehlersuche angepasst werden. Es können auch verschiedene Konfigurationsdateien für verschiedene Testszenarien verwendet werden. Und auch während des produktiven Einsatzes ist es immer wieder erforderlich, einzelne Stellgrößen anzupassen.

Im Folgenden werden einige Beispiele für mögliche Konfigurationsparameter genannt.

`Common.LogFile`

*Dateiname für die Protokolldatei*

`Common.LogLevel`

*Detailgrad von Protokolleinträgen*

`Scheduler.ScrapeInterval`

*Intervall für das Starten des Screen Scraping*

`Scraper.CancelOnError`

*Schalter für das Verhalten, wenn bei der Extraktion Fehler auftreten*

`Scraper.BaseUrl`

*Domainname und Basis-Pfad für alle zu bildenden URLs*

`Scraper.<FacetType>.Url`

*URL-Teil für das Erfassen eines bestimmten Facetten-Typs*

`Analyzer.<FacetType>.<Field>`

*CSS-Selektor für das Auffinden eines Datums in der Web-Seite*

`Crawler.Offline`

*Schalter für den Offline-Betrieb – Anfragen werden nur aus dem Zwischenspeicher beantwortet*

`HttpClient.userAgentString`

*Zeichenfolge für das HTTP-Header-Feld UserAgent*

`HttpClient.Throttle`

*Schalter für das Aktivieren von Pausen zwischen Anfragen*

`HttpClient.MinThrottle`

*Minimale Pausenzeit*

`HttpClient.MaxThrottle`

*Maximale Pausenzeit*

`HttpClient.Proxy`

*Ein möglicherweise erforderlicher Proxy-Server*

`Cache.Directory`

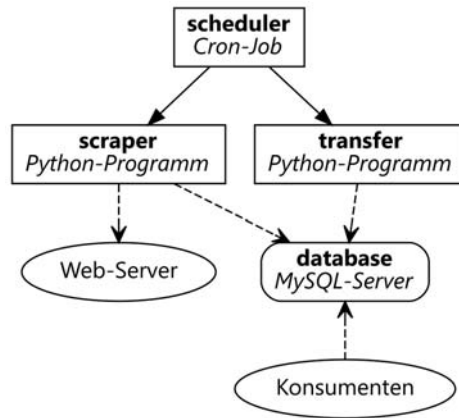
*Verzeichnis für das Speichern der Rohdaten*

`Cache.RetentionTime`

*Zeitspanne nach der gespeicherte Web-Seiten verworfen werden*

## Implementierung

Im Folgenden wird kurz eine mögliche Implementierung eines Screen-Scraping-Systems beschrieben. Das System wird auf einem Linux-Betriebssystem ausgeführt und verwendet den Cron-Scheduler des Betriebssystems als *Scheduler*-Modul. Die übrigen Module werden in Python 3 implementiert und sind in zwei Prozesse aufgeteilt.



### Übersicht über eine mögliche Implementierung

Der erste Prozess ist für das Erfassen und Extrahieren der Informationen zuständig und wird hier mit *scraper* bezeichnet. Er umfasst die Module *Scraper*, *Crawler*, *HTTP Client*, *Cache*, *Analyzer* und *Facet Storage*. Der zweite Prozess ist für das Zusammenführen der Facetten in konsistente Entitäten zuständig und wird hier mit *transfer* bezeichnet. Er umfasst die Module *Transfer*, *Facet Storage* und *Entity Storage*.

Da das Module *Facet Storage* in beiden Prozessen benötigt wird, wird es in einer Bibliothek implementiert, welche von beiden Prozessen verwendet werden kann. Allgemeine Aspekte wie Protokollierung und Konfiguration werden ebenfalls in einer Bibliothek implementiert und von beiden Prozessen referenziert.

Das Modul *HTTP Client* verwendet die Python-Bibliothek *requests* für die HTTP-Kommunikation mit dem Web-Server (vergl. [16]). Das Modul *Cache* speichert die Rohdaten als Dateien in einem Verzeichnis. Dabei wird der Dateiname für eine HTML-Quelle aus dem Zeitstempel des Abfragezeitpunktes, einem SHA-1-Hash der URL und dem HTTP-Status-Code gebildet (vergl. [17]). Das Modul *Analyzer* verwendet die Python-Bibliothek *bs4* (Beautiful Soup 4) für die DOM-Repräsentation und das Selektieren von HTML-Elementen mit der Bibliothek *lxml* für das Parsen von HTML (vergl. [18], [19]). Die Module *Facet Storage* und *Entity Storage* kommunizieren mit einem MySQL-Server und setzen dabei die Python-Bibliothek *sqlalchemy* für das OR-Mapping mit *pymysql* als SQL-Adapter ein (vergl. [20], [21]). Für die Facetten und die Entitäten werden zwei getrennte Datenbanken auf dem MySQL-Server verwendet. Konfigurationsdateien werden im INI-Format gespeichert und mit der Klasse *SafeConfigParser* aus der Python-Bibliothek *configparser* eingelesen

(vergl. [22]). Für das Verarbeiten von Befehlszeilen-Parametern wird die Python-Bibliothek `argparse` verwendet (vergl. [23]). Für die Protokollierung wird die Python-Bibliothek `logging` verwendet (vergl. [24]).

Konsumenten greifen direkt auf die Entitäten-Datenbank auf dem MySQL-Server zu.

## Empfehlungen

Eine gute Grundlage zur Implementierung eines Screen-Scraping-Programms ist das Python-Framework *Scrapy* (vergl. [25]). Scrapy ist ein Framework, welches weite Teile des Screen Scraping unterstützt. Es löst die Herausforderungen, die bei den HTTP-Anfragen entstehen, sehr gut und stellt eine solide Basis für die Analyse mittels Selektoren zur Verfügung. Darüber hinaus bietet es definierte Schnittstellen für Erweiterungen und Anpassungen.

Scrapy bietet u. a. die folgenden Funktionalitäten:

- Robuster und konfigurierbarer Crawler
- Parallele Ausführung mehrerer HTTP-Abfragen
- Statistiken
- Middleware für
  - Authentifizierung
  - Cookies
  - Cache
  - Abruf über Proxy
  - Verfolgen von Umleitungen
  - User-Agent überschreiben
  - Webseiten nach Fehler erneut Abfragen
  - URL-Filter
  - HTTP-Status-Code-Filter
- Analyse mittels
  - CSS-Selektoren
  - XPath-Selektoren
- Einstellungen über
  - Befehlszeilenargumente
  - Python-Code
- Ausgabe extrahierter Facetten

- XML
- CSV
- Pickle (kompakte Python-spezifische Serialisierung)
- pprint
- JSON/JSON-Lines
- E-Mail-Versand für Benachrichtigungen
- Interaktive Python-Shell über eine Telnet-Konsole

Die folgenden Aspekte werden von Scrapy nicht direkt unterstützt:

- Speichern der extrahierten Facetten in einer relationalen Datenbank (nur über Erweiterung möglich)
- Zusammenführen von Facetten in Entitäten (nachgelagerter Transfer-Schritt notwendig)
- Verteilung der Ausführung auf mehrere Rechner (Ausführung von Crawlern auf einem Server mit *Scrapyd* möglich, vergl. [26])

Erfüllt Scrapy eine besondere Anforderung nicht, sollte zunächst geprüft werden ob diese mittels einer Erweiterung erfüllt werden kann, oder Scrapy als Teilkomponente einer größeren Architektur genutzt werden kann. Ist auch dies nicht der Fall, kann die Implementierung eines spezialisierten Programms in Erwägung gezogen werden. Der Aufwand dafür sollte jedoch nicht unterschätzt werden.

Wird die Implementierung eines spezialisierten Programms in Betracht gezogen, oder wird Scrapy als Teilkomponente in eine größere Architektur eingebettet, könnte zur Orchestrierung der einzelnen Teilaufgaben (Webseitenabruf, Datenextraktion, Bereinigung) das Python-Framework *Luigi* genutzt werden (vergl. [27]). Das Framework unterstützt eine klare Gliederung der Teilaufgaben mit ihren Abhängigkeiten und ermöglicht die Koordination von verteilt ausgeführten Aufgaben. Dafür bietet es einen zentralen Dienst, welcher die Ausführung der Teilaufgaben überwacht und in einer Web-Oberfläche visualisiert.

Bei großen Datenmengen eignet sich der *PostgreSQL*-Server u.U. besser für die Speicherung von Facetten und Entitäten als der *MySQL*-Server bzw. dessen freier Fork *MariaDB*. Der *PostgreSQL*-Server bietet mächtigere Datentypen und seine SQL-Syntax ist mit der von *Amazon Redshift* kompatibel und eröffnet damit, falls nötig, eine leichtere Migration in die Cloud (vergl. [28], [29]). Ausschlaggebend für die Wahl des Datenbanksystems ist natürlich auch die Kompatibilität mit den Konsumenten der Entitäten.



## Quellen

- [1] HEMMERICH, WANJA A. : *Normalverteilung*. 2017. — URL <https://mathe-guru.com/stochastik/normalverteilung.html>
- [2] APSEL, MATTHIAS ; SCHWARWIES, MATTHIAS : *selfhtml – HTML / Universalattribute / id*. 2017. — URL [https://wiki.selfhtml.org/wiki/HTML/ ... Universalattribute#id](https://wiki.selfhtml.org/wiki/HTML/...Universalattribute#id)
- [3] APSEL, MATTHIAS : *selfhtml – CSS / Selektoren / Klassenselektor*. 2017. — URL <https://wiki.selfhtml.org/wiki/CSS/Selektoren/Klassenselektor>
- [4] *XPath*. Wikipedia, 2017. — URL <https://de.wikipedia.org/wiki/XPath>
- [5] APSEL, MATTHIAS ; SCHWARWIES, MATTHIAS : *selfhtml – CSS / Selektoren*. 2017. — URL <https://wiki.selfhtml.org/wiki/CSS/Selektoren>
- [6] DIGGS, JOANMARIE ; MCCARRON, SHANE ; COOPER, MICHAEL ; SCHWERTDFEGER, RICHARD ; CRAIG, JAMES : *Accessible Rich Internet Applications (WAI-ARIA) 1.1: W3C Recommendation 14 December 2017*. W3C, 2017. — URL [https://www.w3 ... .org/TR/wai-aria-1.1/](https://www.w3.org/TR/wai-aria-1.1/)
- [7] CHANG, CHIA-HUI ; LUI, SHAO-CHEN : IEPAD: Information Extraction Based on Pattern Discovery. In: *Proceedings of the 10th International Conference on World Wide Web, WWW '01*. New York, NY, USA : ACM, 2001 — ISBN 1-58113-348-0, S. 681–688
- [8] CHANG, CHIA-HUI ; HSU, CHUN-NAN ; LUI, SHAO-CHENG : Automatic information extraction from semi-structured Web pages by pattern discovery. In: *Decision Support Systems* Bd. 35 (2003), Nr. 1, S. 129–147. — Web Retrieval and Mining
- [9] AMITAY, EINAT ; CARMEL, DAVID ; DARLOW, ADAM ; LEMPEL, RONNY ; SOFFER, AYA : The Connectivity Sonar: Detecting Site Functionality by Structural Patterns. In: *Proceedings of the Fourteenth ACM Conference on Hypertext and Hypermedia, HYPERTEXT '03*. New York, NY, USA : ACM, 2003 — ISBN 1-58113-704-4, S. 38–47
- [10] CHAKRABARTI, DEEPAYAN ; MEHTA, RUPESH : The Paths More Taken: Matching DOM Trees to Search Logs for Accurate Webpage Clustering. In: *Proceedings*

- of the 19th International Conference on World Wide Web, WWW '10*. New York, NY, USA : ACM, 2010 — ISBN 978-1-60558-799-8, S. 211–220
- [11] *Single-Page-Webanwendung*. Wikipedia, 2018. — URL <https://de.wikipedia.org/wiki/Single-Page-Webanwendung>
  - [12] PODMANICKI, TONI : *Wie Sie JavaScript in Ihrem Browser einschalten und wofür es gebraucht wird*. 2018. — URL <https://www.enable-javascript.com/de/>
  - [13] MAONE, GIORGIO : *NoScript – JavaScript/Java/Flash blocker for a safer Firefox experience*. 2018. — URL <https://noscript.net>
  - [14] HIDAYAT, ARIYA : *PhantomJS – Full web stack, No browser required*. 2018. — URL <http://phantomjs.org>
  - [15] *Universally Unique Identifier*. Wikipedia, 2018. — URL [https://de.wikipedia.org/wiki/Universally\\_Unique\\_Identifier](https://de.wikipedia.org/wiki/Universally_Unique_Identifier)
  - [16] REITZ, KENNETH : *Requests: HTTP for Humans*. 2017. — URL <http://python-requests.org>
  - [17] *Secure Hash Algorithm*. Wikipedia, 2017. — URL [https://de.wikipedia.org/wiki/Secure\\_Hash\\_Algorithm](https://de.wikipedia.org/wiki/Secure_Hash_Algorithm)
  - [18] RICHARDSON, LEONARD : *Beautiful Soup*. 2017. — URL <https://www.crummy.com/software/BeautifulSoup/>
  - [19] BEHNEL, STEFAN ; U.A. : *lxml – XML and HTML with Python*. 2017. — URL <http://lxml.de/>
  - [20] BAYER, MICHAEL ; KIRTLAND, JASON ; DE, GAETAN ; CLARKE, DIANA ; TRIER, MICHAEL ; JENVEY, PHILIP ; AASMA, ANTS ; JOHNSTON, PAUL ; U. A. : *SQLAlchemy*. 2018. — URL <http://www.sqlalchemy.org>
  - [21] NAOKI, INADA ; BLACK, DANIEL ; RODRIGUES, MARCEL ; HUNT, PETE ; U.A. : *PyMySQL – Pure Python MySQL Client*. 2017. — URL <https://github.com/PyMySQL/PyMySQL>
  - [22] FOUNDATION, PYTHON SOFTWARE : *configparser – Configuration file parser*. 2017. — URL <https://docs.python.org/3.5/library/configparser.html>
  - [23] FOUNDATION, PYTHON SOFTWARE : *argparse – Parser for command-line options, arguments and sub-commands*. 2017. — URL <https://docs.python.org/3.6/library/argparse.html>

- [24] FOUNDATION, PYTHON SOFTWARE : *logging – Logging facility for Python*. 2018.  
— URL <https://docs.python.org/3/library/logging.html>
- [25] EVANS, SHANE ; U.A. : *Scrapy – Extracting the data you need from websites*. 2018.  
— URL <https://scrapy.org/>
- [26] HOFFMAN, PABLO ; U.A. : *Scrapyd – A service daemon to run Scrapy spiders*. 2018.  
— URL <https://scrapyd.readthedocs.io/en/stable/>
- [27] BERNHARDSSON, ERIK ; FREIDER, ELIAS ; U.A. : *Luigi – Python package for building complex pipelines and batch jobs*. 2018. — URL <https://luigi.readthedocs.io/en/stable/>
- [28] POSTGRESQL GLOBAL DEVELOPMENT GROUP : *PostgreSQL*. 2018.  
— URL <https://www.postgresql.org/>
- [29] AMAZON WEB SERVICES : *Redshift*. 2018. — URL <https://aws.amazon.com/de/redshift/>

# Systemarchitektur

## Die Architektur des SPI-Systems mit Datenhaltung, Datenintegration, Datenanalyse und Benutzerschnittstellen

Tobias Kiertscher

### Inhaltsverzeichnis

|                                |     |
|--------------------------------|-----|
| Allgemein .....                | 71  |
| Übersichten .....              | 72  |
| Datenquellen .....             | 76  |
| Golf.de .....                  | 76  |
| GJGR .....                     | 77  |
| Meteomatics API .....          | 77  |
| Datenbanken für Rohdaten ..... | 78  |
| Golf.de DB .....               | 78  |
| GJGR DB .....                  | 78  |
| Wetter DB .....                | 78  |
| Datensammler .....             | 81  |
| Golf.de Crawler .....          | 81  |
| GJGR Crawler .....             | 86  |
| Meteomatics Weather .....      | 90  |
| SPI Anwendung .....            | 93  |
| SPI DB .....                   | 93  |
| SPI Backend .....              | 102 |
| SPI Backend Jobs .....         | 113 |

|                                           |     |
|-------------------------------------------|-----|
| SPI Frontend Web.....                     | 116 |
| SPI Frontend Mobile.....                  | 117 |
| Datenintegration .....                    | 117 |
| Golf.de Transfer.....                     | 118 |
| GJGR Transfer.....                        | 119 |
| Datenanalyse.....                         | 120 |
| Analyzer.....                             | 121 |
| Intermediate Golf DB .....                | 127 |
| Golf Analytics.....                       | 128 |
| Golf Analytics DB.....                    | 129 |
| Jupyter Notebook.....                     | 132 |
| Aufgabenplanung .....                     | 132 |
| Cron .....                                | 132 |
| Prochestra.....                           | 133 |
| Aufbau einer Python-Komponente .....      | 134 |
| Package-Struktur .....                    | 134 |
| CLI-Skripte.....                          | 136 |
| Konfiguration von Python-Komponenten..... | 137 |

## Allgemein

Das SPI-System führt mehrere externe Datenquellen zusammen und pflegt in einer Datenbank einen konsistenten Datenbestand, der aus eigenen und externen Daten besteht. Diese Datenbank wird mit einem Online-Portal verknüpft, über welches zum einen die zusammengeführten Daten eingesehen werden können und zum anderen direkt eigene Daten eingegeben oder geändert werden können.

Es steht damit beispielhaft für eine Vielzahl von Systemen, welche externe heterogene Datenquellen mit eigenen Daten verknüpfen und dadurch neue Erkenntnisse ermöglichen.

In Abbildung 1 wird deutlich, wie sich die generelle Entwicklung des SPI-Systems vollzieht. Ein *Analyst* (oder auch *Data Scientist*) erschließt sich *Datenquellen* und entwickelt eine prototypische Software zur *Datenanalyse*. Seine Erkenntnisse über die Datenquellen fließen anschließend in die *Datenintegration* ein, die der *Entwickler* implementiert, um die externen Daten in die *SPI-Anwendung* einzubinden. Die vom *Analysten* entwickelten Algorithmen zur Analyse, fließen nach einer Evaluation in die Entwicklung der *SPI-Anwendung* ein. Der *Benutzer* interagiert mit der SPI-Anwendung lediglich über die Web- oder Mobil-Schnittstellen.

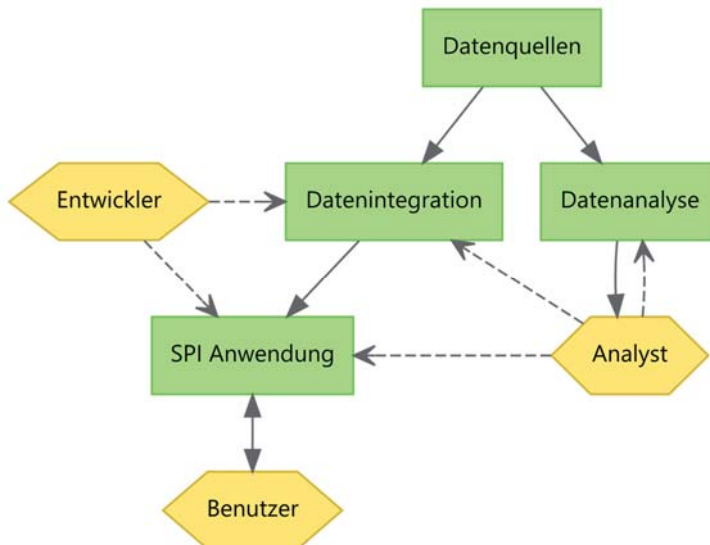


Abbildung 1: Systemübersicht mit Rollen

## Übersichten

Die folgenden Abbildungen veranschaulichen den Aufbau des SPI-Systems aus unterschiedlichen Perspektiven.

Zum aktuellen Stand des Forschungsprojektes (September 2018), existieren die in Abbildung 2 dargestellten Komponenten. Das Diagramm veranschaulicht den Weg der Daten durch das System.

Einige Teile des Systems werden durch Abfragen oder Eingaben eines Benutzers aktiv, andere werden automatisch und regelmäßig ausgeführt. Einige Komponenten werden auch nur manuell durch einen Analysten genutzt, um weitere Erkenntnisse aus den Daten zu gewinnen oder Algorithmen zu testen, bevor neue Funktionalitäten im System implementiert werden. Die Art der Ausführung für die wesentlichen Komponenten im System wird in Abbildung 3 dargestellt.

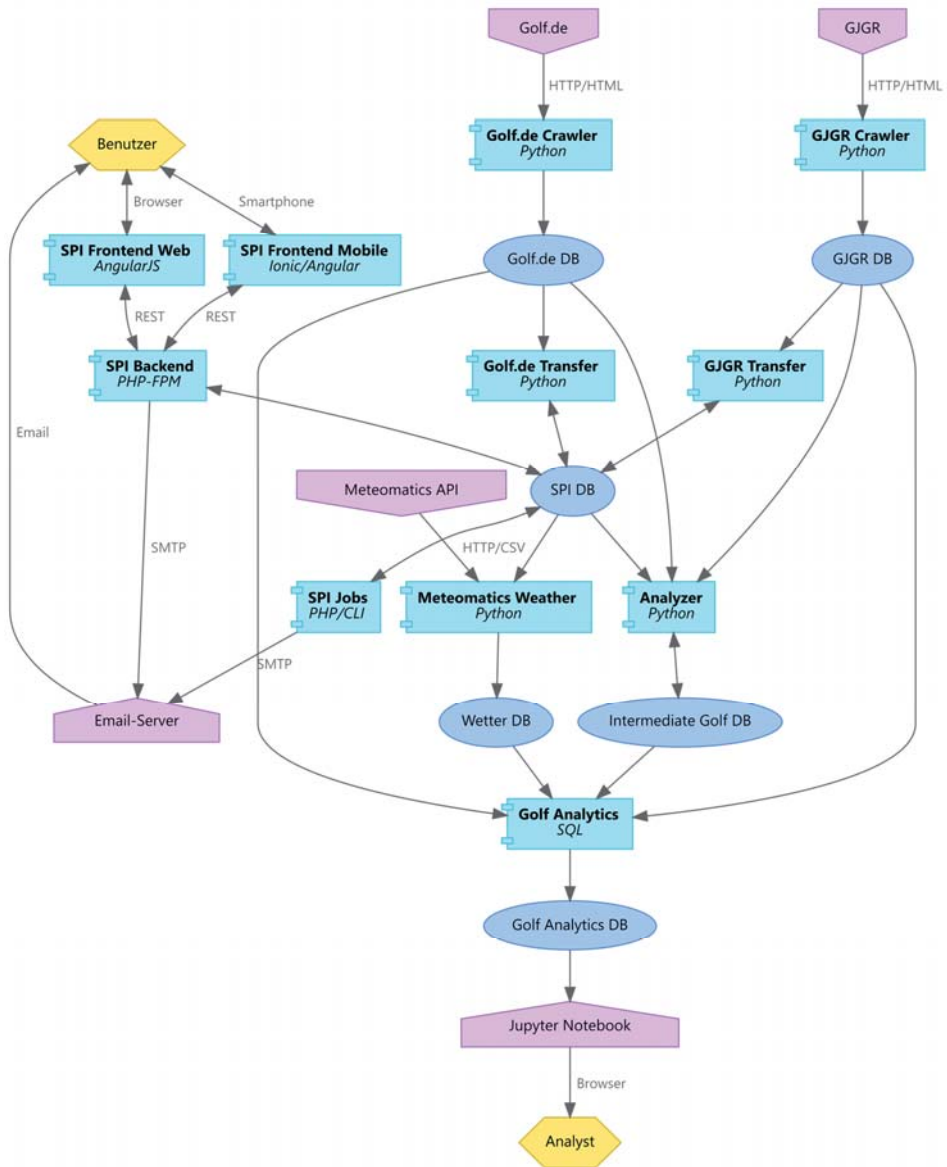


Abbildung 2: Datenfluss



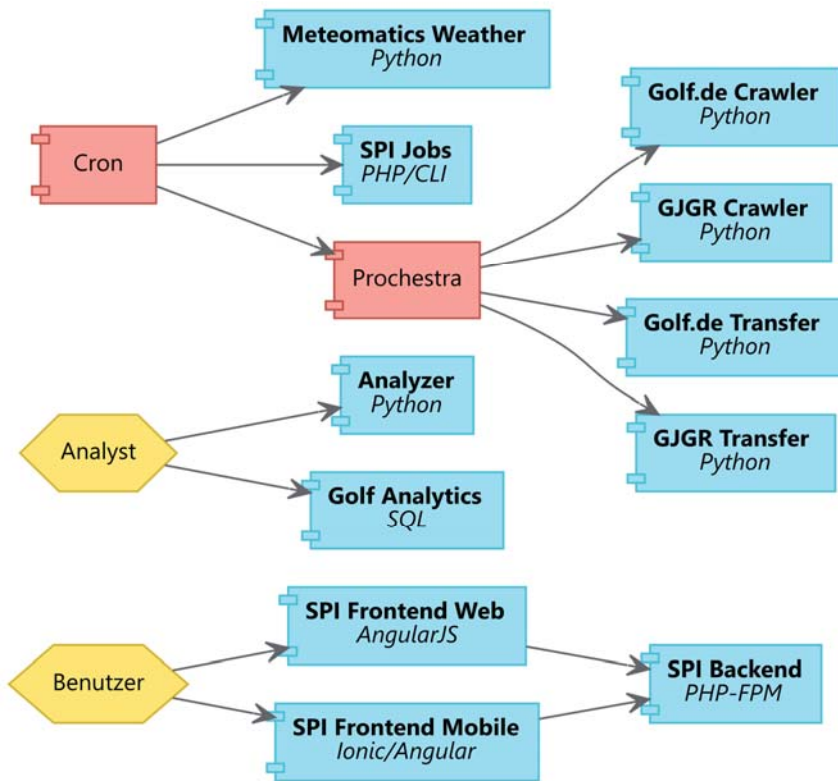


Abbildung 3: Steuerung

Die folgende Tabelle bietet eine grobe Übersicht über die eingesetzten Technologien und listet für jede Komponente einen oder mehrere Ansprechpartner auf, die an der Implementierung der Komponente beteiligt waren.

| Komponente           | Ansprechpartner                    | Typ          | Technologie           |
|----------------------|------------------------------------|--------------|-----------------------|
| Golf.de              | Deutsche Golf Online GmbH, München | Website      | HTTP, HTML            |
| GJGR                 | subsist GmbH, Dresden              | Website      | HTTP, HTML            |
| Meteomatics API      | Meteomatics GmbH, Berlin           | API          | HTTP, CSV             |
| Golf.de DB           | Tobias Kiertscher                  | Datenbank    | MySQL                 |
| GJGR DB              | Tobias Kiertscher                  | Datenbank    | MySQL                 |
| Wetter DB            | Tobias Kiertscher, Markus Lampert  | Datenbank    | MySQL                 |
| Golf.de Crawler      | Tobias Kiertscher                  | Programm     | Python                |
| GJGR Crawler         | Tobias Kiertscher                  | Programm     | Python                |
| Meteomatics Weather  | Tobias Kiertscher, Markus Lampert  | Programm     | Python                |
| SPI DB               | Tobias Kiertscher, André Nitze     | Datenbank    | MySQL                 |
| SPI Backend          | Tobias Kiertscher, André Nitze     | Webanwendung | PHP                   |
| SPI Backend Jobs     | Tobias Kiertscher                  | Programm     | PHP                   |
| SPI Frontend Web     | Tobias Kiertscher, André Nitze     | Web-UI       | JavaScript, HTML, CSS |
| SPI Frontend Mobile  | André Nitze, Kilian Friedrich      | Mobil-App    | TypeScript, Ionic     |
| Golf.de Transfer     | Tobias Kiertscher                  | Programm     | Python                |
| GJGR Transfer        | Tobias Kiertscher                  | Programm     | Python                |
| Analyzer             | Tobias Kiertscher                  | Programm     | Python                |
| Intermediate Golf DB | Tobias Kiertscher                  | Datenbank    | MySQL                 |
| Golf Analytics       | Tobias Kiertscher                  | Script       | SQL                   |
| Golf Analytics DB    | Tobias Kiertscher                  | Datenbank    | MySQL                 |
| Jupyter Notebook     | Tobias Kiertscher                  | Script       | Jupyter               |
| Prochestra           | Tobias Kiertscher                  | Programm     | Python                |

## Datenquellen

In diesem Abschnitt werden die externen Datenquellen für das SPI-System beschrieben. Das SPI-System umfasst beispielhaft die Einbindung einer für M2M-Kommunikation entworfene Schnittstelle, die Extraktion von Daten aus einer einfach aufgebauten Website und die Extraktion von Daten aus einer komplizierter aufgebauten Website.

### Golf.de

- Typ: Website
- Technologie: HTTP, HTML
- Ansprechpartner: Deutsche Golf Online GmbH, München

Die Website <http://www.golf.de> ist die Hauptquelle für Turnierergebnisse von Golf-Spielern in Deutschland. Sie enthält einen Turnierkalender, in dem mit Hilfe von mehreren Kriterien nach Turnieren gesucht und gefiltert werden kann. Die Suchfunktion liefert maximal 250 Turniere als Ergebnis. Die Ergebnisse werden mit Pagination auf maximal 10 Seiten mit je 25 Turnieren angezeigt.

Golf.de steht beispielhaft für eine kompliziertere Website. Für das SPI-System interessante Informationen befinden sich auf den folgenden Seitentypen:

- Turnierliste (Ergebnis einer Suche)  
Z. B.: [https://www.golf.de/publish/turnierkalender/alle-turniere ...  
?breite=&laenge=&lgv=&clubnameClubnr=&begdat=29.09.2018 ...  
&enddat=29.09.2018&begdat\\_mobil=&enddat\\_mobil=&plzOrt=10115 ...  
&radius=240&turniername=&filter\\_lochanzahl=18&filter\\_geschlecht=W ...  
&filter\\_offen=&filter\\_vorgwirk=1&modus=E](https://www.golf.de/publish/turnierkalender/alle-turniere...?breite=&laenge=&lgv=&clubnameClubnr=&begdat=29.09.2018...&enddat=29.09.2018&begdat_mobil=&enddat_mobil=&plzOrt=10115...&radius=240&turniername=&filter_lochanzahl=18&filter_geschlecht=W...&filter_offen=&filter_vorgwirk=1&modus=E)
- Club-Details  
Z. B.: [http://www.golf.de/publish/plaetze/national/1109/golf-und- ...  
landclub-semmlin-am-see-e-v](http://www.golf.de/publish/plaetze/national/1109/golf-und-...landclub-semmlin-am-see-e-v)
- Turnierdetails  
Z. B.: [https://www.golf.de/publish/turnierkalender/alle-turniere ...  
/kalender/info?tournamentid=490001251460&clubid=3005&modus=E](https://www.golf.de/publish/turnierkalender/alle-turniere.../kalender/info?tournamentid=490001251460&clubid=3005&modus=E)
- Turnierergebnisse  
Z. B.: [https://www.golf.de/publish/turnierkalender/alle-turniere ...  
/ergebnisse/scoreList?tournamentid=490001251460&clubid=3005 ...  
&playingRound=1](https://www.golf.de/publish/turnierkalender/alle-turniere.../ergebnisse/scoreList?tournamentid=490001251460&clubid=3005...&playingRound=1)

- Score-Karten  
Z. B.: [https://www.golf.de/publish/turnierkalender/alle-turniere ... /ergebnisse/scorecard?scorecardId=1\\_149343351,2\\_149343391](https://www.golf.de/publish/turnierkalender/alle-turniere.../ergebnisse/scorecard?scorecardId=1_149343351,2_149343391)

## GJGR

- Typ: Website
- Technologie: HTTP, HTML
- Ansprechpartner: subsist GmbH, Dresden

Die Website <http://www.gjgr.de> ist eine zusätzliche Quelle für Turnierergebnissen der Golf-Jugend in Deutschland, einschließlich einiger internationaler Turniere. Besonders wertvoll aus dieser Quelle sind die veröffentlichten Geburtsjahre der Spieler.

Für das SPI-System interessante Informationen befinden sich auf den folgenden Seitentypen:

- Turnierliste  
Z. B.: <http://www.gjgr.de/de/tournaments/f>
- Turnierergebnisse, evtl. mit Score-Karten  
Z. B.: <http://www.gjgr.de/de/listTournament/4195>

## Meteomatics API

- Typ: API
- Technologie: HTTP, CSV
- Ansprechpartner: Meteomatics GmbH, Berlin

Der Wetterdatendienst von Meteomatics ist ein kommerzielles Produkt, welches für die Dauer des Projektes in Anspruch genommen wurde, weil die vom DWD kostenfrei veröffentlichten Wetterdaten nicht den Ansprüchen der Analyseziele des Projektes gerecht wurden.

Die API von Meteomatics liefert über eine URL Zeitreihen für ausgewählte Wetterparameter und GPS-Koordinaten u. a. im CSV-Format aus.

## Datenbanken für Rohdaten

Die im Folgenden beschriebenen Datenbanken nehmen die Rohdaten aus den externen Datenquellen auf bevor sie vorverarbeitet und zusammengeführt werden. Diese Datenbanken sind wichtige Schnittstellen, welche die Entkopplung von Datensammlung und -auswertung ermöglicht.

### Golf.de DB

- Typ: Datenbank
- Technologie: MySQL
- Ansprechpartner: Tobias Kiertscher

Die *Golf.de Datenbank* ist eine MySQL-Datenbank welche die Rohdaten vom Screen-Scraping von *golf.de* aufnimmt. In Abbildung 4 ist das Tabellenschema der Datenbank dargestellt. Alle Tabellen in der *Golf.de Datenbank* besitzen das Präfix *scraped\_golf\_*.

### GJGR DB

- Typ: Datenbank
- Technologie: MySQL
- Ansprechpartner: Tobias Kiertscher

Die *GJGR Datenbank* ist eine MySQL-Datenbank welche die Rohdaten vom Screen-Scraping von *gjgr.de* aufnimmt. In Abbildung 5 ist das Tabellenschema der Datenbank dargestellt. Alle Tabellen in der *GJGR Datenbank* besitzen das Präfix *scraped\_gjgr\_*.

### Wetter DB

- Typ: Datenbank
- Technologie: MySQL
- Ansprechpartner: Tobias Kiertscher, Markus Lampert

Die *Wetter Datenbank* speichert den stündlichen Wetterverlauf für alle in der SPI DB vorkommenden Golf-Clubs. Die Datenbank enthält nur eine Tabelle, welche in Abbildung 6 dargestellt ist.

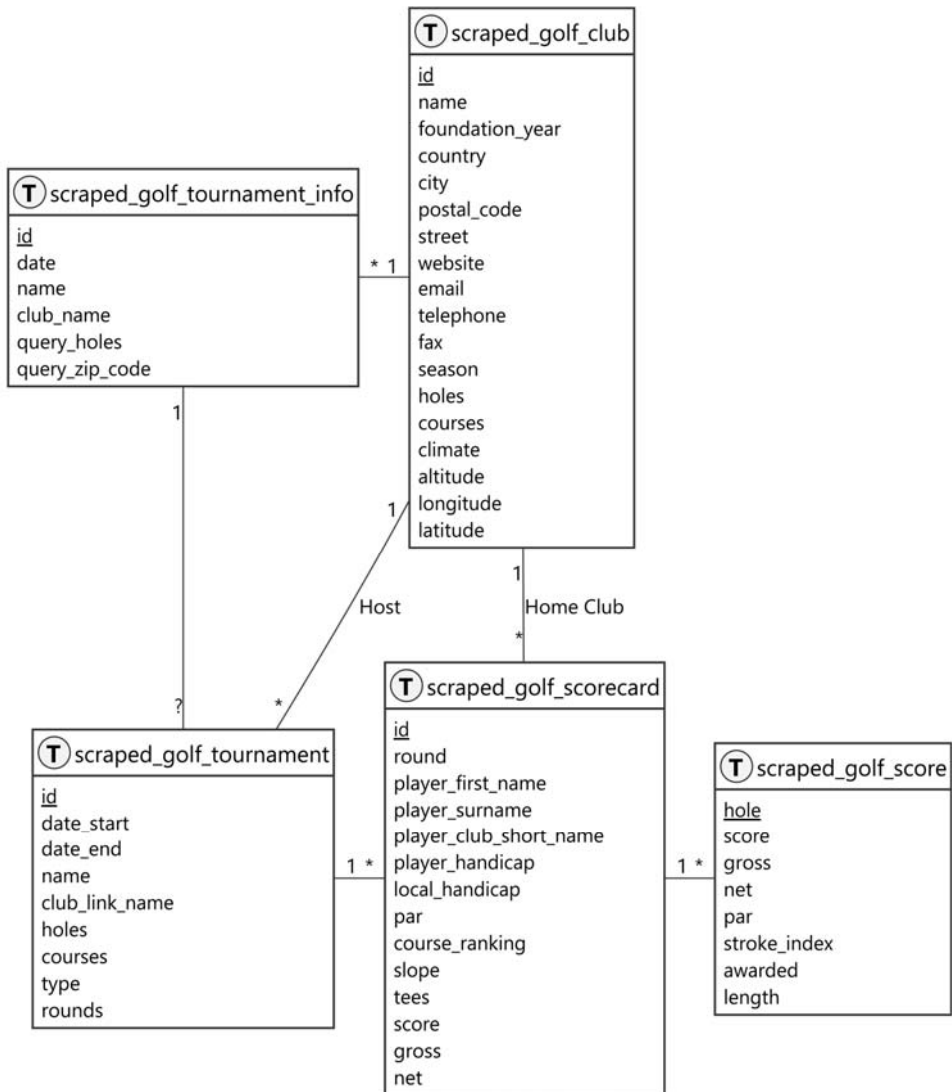


Abbildung 4: Datenbank für Golf.de Rohdaten

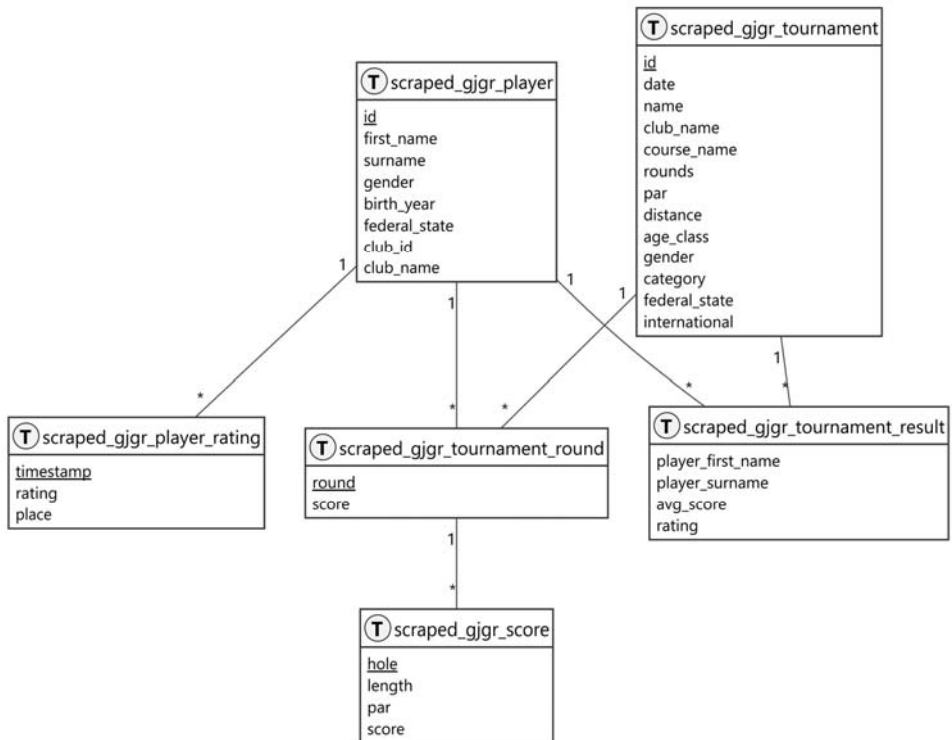


Abbildung 5: Datenbank für GJGR Rohdaten



Abbildung 6: Datenbank für Meteomatics Rohdaten

# Datensammler

Die im Folgenden beschriebenen Komponenten sind für das Einsammeln bzw. Abfragen der externen Datenquellen verantwortlich. Sie sind alle in Python implementiert und basieren zum größten Teil auf einer gemeinsamen Architektur.

## Golf.de Crawler

- Typ: Programm
- Technologie: Python
- Ansprechpartner: Tobias Kiertscher

Der *Golf.de Crawler* ist ein Python-Modul, welches das Screen-Scraping für golf.de implementiert. Es ist als Befehlszeilenwerkzeug konzipiert und lässt sich flexibel konfigurieren, um verschiedenartige Abfragen auf golf.de durchzuführen.

## Package-Struktur

Die Python-Komponente *Golf.de Crawler* basiert auf der gemeinsamen Architektur für Python-Komponenten und ist zusätzlich in den folgenden Packages und Klassen implementiert:

- `spi.crawler.golf`
  - `GolfDbSessionWrapper`  
Eine Klasse, die von `spi.db.DbSessionWrapperBase` erbt und den objektorientierten Zugriff auf die Golf.de Rohdaten ermöglicht.
  - `GolfCrawler`  
Eine Klasse, die von `spi.web.CrawlerBase` erbt und die Abfragen von golf.de organisiert.
  - `web`
    - `SkyScraper`  
Eine Klasse, die von `spi.web.ScraperBase` erbt und das Extrahieren von Daten aus der Webseiten von golf.de implementiert.
    - `ReverseGeoCoder`  
Eine Klasse, die das Ermitteln von GPS-Koordinaten für eine postalische Adresse ermittelt. Sie nutzt dafür eine Web-API von *OpenStreetMap*.
  - `model`  
Ein Package, das die Datenklassen für die Tabellen der Golf.de DB enthält.
    - `Club`



- TournamentInfo
- Tournament
- Scorecard
- Score
- ...

## Befehlszeile

Das CLI-Skript für *Golf.de Crawler* heißt `golfcrawler.py`. Der folgende Hilfetext, der mit der Befehlszeilenoption `--help` ausgegeben werden kann, erklärt den Aufruf der Komponente detailliert.

```
usage: golfcrawler.py [-h] [-v] [-d TARGET_DATE] [-i INTERVAL] [-s START_DATE]
                    [-e END_DATE] [-r {b,backwards,f,forwards}]
                    [-c CONFIG_FILES] [-o OPTIONS [OPTIONS ...]]
```

SPI Golf Crawler

optional arguments:

- h, --help show this help message and exit
- v, --version print the program version and exit
- d TARGET\_DATE, --date TARGET\_DATE  
Crawles one specific date, given in the format yyyy-mm-dd.
- i INTERVAL, --interval INTERVAL  
Crawles the specified number of days in the past.
- s START\_DATE, --start-date START\_DATE  
Specifies the earliest date to crawl, in the format yyyy-mm-dd.
- e END\_DATE, --end-date END\_DATE  
Specifies the latest date to crawl, in the format yyyy-mm-dd.
- r {b,backwards,f,forwards}, --direction {b,backwards,f,forwards}  
Specifies the direction to crawl. Default is backwards.
- c CONFIG\_FILES, --config-file CONFIG\_FILES  
A path to a configuration file in UTF-8 encoded INI format. This argument can be used multiple times.
- o OPTIONS [OPTIONS ...], --options OPTIONS [OPTIONS ...]  
One or more configuration options, given in the format <section>.<option>=<value>.

## Optionen

Die Konfiguration von *Golf.de Crawler* umfasst die folgenden Sektionen und Optionen.

### Sektion **user**

#### **home\_club**

Der vollständige Name des Heimatclubs, wie er auf golf.de genannt wird. Wenn der Schalter `behavior.load_all_scores` nicht aktiviert ist, werden Score-Karten nur von Spielern abgerufen, welche dem Heimatclub zugeordnet sind.

Beispiel = Golf- und Land-Club Berlin-Wannsee e.V.

#### **home\_club\_short**

Die Kurzschreibweise des Heimatclubs, wie er auf golf.de genannt wird.

Standardwert = Berlin-Wannsee, G&LC

#### **start\_date**

Das erste abzufragende Datum.

Standardwert = 2000-01-01

#### **end\_date**

Das letzte abzufragende Datum oder leer.

### Sektion **logging**

#### **store\_responses**

Ein Schalter, der das Speichern der vom HTTP-Server empfangenen HTML-Dateien aktiviert.

Standardwert = yes

#### **store\_responses\_with\_timestamp**

Ein Schalter, der dem Dateinamen der gespeicherten HTML-Dateien einen Zeitstempel anfügt.

Standardwert = no

#### **response\_storage\_dir**

Ein relativer oder absoluter Pfad zu dem Verzeichnis, in dem die HTML-Dateien gespeichert werden sollen. Ist der Pfad relativ, bezieht er sich auf das aktuelle Arbeitsverzeichnis.

Standardwert = golf\_responses

### Sektion **behavior**

#### **exit\_on\_first\_error**

Ein Schalter zum Steuern des Verhaltens im Fehlerfall. Ist der Schalter aktiv, bricht

das Programm ab, sobald der erste Fehler auftritt.

Standardwert = no

#### `load_tournament_infos`

Gibt an, ob nach Turnieren gesucht werden sollen.

Standardwert = yes

#### `reload_tournament_infos`

Gibt an, ob erneut nach Turnieren gesucht werden sollen, auch wenn für den jeweiligen Tag bereits welche in der Datenbank existieren.

Standardwert = no

#### `delete_obsolete_tournament_infos`

Gibt an, ob Turniere mit ihren Ergebnissen aus der Datenbank gelöscht werden sollen, wenn das dazugehörige Turnier auf [golf.de](http://golf.de) nicht mehr gefunden wird.

Standardwert = yes

#### `load_club`

Gibt an, ob die Detailinformationen zu den Clubs abgerufen werden sollen.

Standardwert = yes

#### `reload_club`

Gibt an, ob die Detailinformationen zu den Clubs erneut abgerufen werden sollen, auch wenn für den Club bereits Informationen in der Datenbank gespeichert sind.

Standardwert = no

#### `load_tournament`

Gibt an, ob Detailinformationen zu den Turnieren abgerufen werden sollen.

Standardwert = yes

#### `reload_tournament`

Gibt an, ob Detailinformationen zu den Turnieren erneut abgerufen werden sollen, auch wenn für das Turnier bereits Informationen in der Datenbank gespeichert sind.

Standardwert = no

#### `load_tournament_results`

Gibt an, ob die Ergebnislisten für die Turniere abgerufen werden sollen.

Standardwert = yes

#### `reload_tournament_results`

Gibt an, ob die Ergebnislisten für die Turniere erneut abgerufen werden sollen, auch wenn für das jeweilige Turnier bereits Ergebnisse in der Datenbank gespeichert sind.

Standardwert = no

#### `delete_obsolete_tournament_results`

Gibt an, ob Turnierergebnisse aus der Datenbank gelöscht werden sollen, wenn sie

in der Ergebnisliste eines Turniers nicht mehr auftauchen.

Standardwert = yes

#### **load\_scores**

Gibt an, ob Score-Karten für die Turnierergebnisse abgerufen werden sollen.

Standardwert = yes

#### **reload\_scores**

Gibt an, ob Score-Karten für die Turnierergebnisse erneut abgerufen werden sollen, auch wenn für das jeweilige Turnierergebnis bereits Scores in der Datenbank gespeichert sind.

Standardwert = no

#### **load\_all\_scores**

Gibt an, ob die Score-Karten von allen Spielern, und nicht nur von den Spielern des Heimatclubs abgerufen werden sollen.

Standardwert = no

#### **request\_timeout**

Gibt den Timeout für HTTP-Verbindungen in Sekunden an.

Standardwert = 10.0

#### **throttle**

Ein Schalter, der eine zusätzliche Pause zwischen HTTP-Anfragen aktiviert. Die Wartezeit wird zufällig aus einem Intervall gewählt.

Standardwert = no

#### **throttle\_min**

Die minimale Wartezeit zwischen HTTP-Anfragen. Wirkt nur, wenn **throttle** aktiv ist. Standardwert = 1

#### **throttle\_max**

Die maximale Wartezeit zwischen HTTP-Anfragen. Wirkt nur, wenn **throttle** aktiv ist.

Standardwert = 10

#### **progress\_interval**

Die Anzahl von Sekunden die zwischen Fortschrittsmeldungen verstreichen dürfen.

Standardwert = 300

### **Sektion target**

#### **url**

Die Basis-URL für HTTP-Anfragen.

Standardwert = <http://www.golf.de/>

### user-agent

Die Zeichenkette, welche bei jeder HTTP-Anfrage im Header-Feld **User-Agent** mitgesendet wird.

Standardwert = SPI PreLytica WebCrawler

### Sektion system

#### html\_parser

Der Parser für das Parsen der HTML-Seiten. Möglich sind alle Parser, die von *Beautiful Soup 4* unterstützt werden: `html.parser`, `lxml` und `html5lib`.

Standardwert = `html5lib`

Für die Konfiguration der MySQL-Datenbankverbindung wird die Sektion `database.golf` verwendet. Der Aufbau der Sektion ist in Konfiguration von Python-Komponenten beschrieben.

## GJGR Crawler

- Typ: Programm
- Technologie: Python
- Ansprechpartner: Tobias Kiertscher

Der *GJGR Crawler* ist ein Python-Modul, welches das Screen-Scraping für `gjgr.de` implementiert.

## Package-Struktur

Die Python-Komponente *GJGR Crawler* basiert auf der gemeinsamen Architektur für Python-Komponenten und ist zusätzlich in den folgenden Packages und Klassen implementiert:

- `spi.crawler.gjgr`
  - `GjgrDbSessionWrapper`

Eine Klasse, die von `spi.db.DbSessionWrapperBase` erbt und den objektorientierten Zugriff auf die GJGR Rohdaten ermöglicht.
  - `GjgrCrawler`

Eine Klasse, die von `spi.web.CrawlerBase` erbt und die Abfragen von `gjgr.de` organisiert.
  - `web`

- SkyScraper  
Eine Klasse, die von `spi.web.ScraperBase` erbt und das Extrahieren von Daten aus der Webseiten von `gjgr.de` implementiert.
- `model`  
Ein Package, das die Datenklassen für die Tabellen der `Golf.de` DB enthält.
  - `Tournament`
  - `TournamentResult`
  - `Player`
  - `PlayerRating`
  - `Score`
  - ...

## Befehlszeile

Das CLI-Skript für *GJGR Crawler* heißt `gjgrcrawler.py`. Der folgende Hilfetext, der mit der Befehlszeilenoption `--help` ausgegeben werden kann, erklärt den Aufruf der Komponente detailliert.

```
usage: gjgrcrawler.py [-h] [-v] [-c CONFIG_FILES] [-o OPTIONS [OPTIONS ...]]
```

SPI GJGR Crawler

optional arguments:

```
-h, --help            show this help message and exit
-v, --version         print the program version and exit
-c CONFIG_FILES, --config-file CONFIG_FILES
                     A path to a configuration file in UTF-8 encoded INI
                     format. This argument can be used multiple times.
-o OPTIONS [OPTIONS ...], --options OPTIONS [OPTIONS ...]
                     One or more configuration options, given in the format
                     <section>.<option>=<value>.
```

## Optionen

Die Konfiguration von *GJGR Crawler* umfasst die folgenden Sektionen und Optionen.

### Sektion `logging`

#### `store_responses`

Ein Schalter, der das Speichern der vom HTTP-Server empfangenen HTML-Dateien aktiviert.

Standardwert = `yes`

#### `store_responses_with_timestamp`

Ein Schalter, der dem Dateinamen der gespeicherten HTML-Dateien einen Zeitstempel anfügt.

Standardwert = no

#### `response_storage_dir`

Ein relativer oder absoluter Pfad zu dem Verzeichnis, in dem die HTML-Dateien gespeichert werden sollen. Ist der Pfad relativ, bezieht er sich auf das aktuelle Arbeitsverzeichnis.

Standardwert = `gjgr_responses`

### Sektion `behavior`

#### `exit_on_first_error`

Ein Schalter zum Steuern des Verhaltens im Fehlerfall. Ist der Schalter aktiv, bricht das Programm ab, sobald der erste Fehler auftritt.

Standardwert = no

#### `load_tournaments`

Gibt an, ob die Turnierlisten abgerufen werden sollen.

Standardwert = yes

#### `reload_tournaments`

Gibt an, ob die Turnierlisten abgerufen werden sollen, auch wenn bereits Turniere in der Datenbank gespeichert sind.

Standardwert = yes

#### `delete_obsolete_tournaments`

Gibt an, ob Turniere mit Ergebnissen aus der Datenbank gelöscht werden sollen, wenn sie nicht mehr in der Turnierliste aufgeführt werden.

Standardwert = no

#### `load_tournament_results`

Gibt an, ob Turnierergebnisse und Score-Karten abgerufen werden sollen.

Standardwert = yes

#### `reload_tournament_results`

Gibt an, ob Turnierergebnisse und Score-Karten erneut abgerufen werden sollen, auch wenn bereits Turnierergebnisse für das jeweilige Turnier in der Datenbank gespeichert sind.

Standardwert = no

#### `load_players`

Gibt an, ob Detailinformationen zu den Spielern abgerufen werden sollen.

Standardwert = yes

#### `reload_players`

Gibt an, ob Detailinformationen zu den Spielern erneut abgerufen werden sollen,

auch wenn bereits Informationen zu dem Spieler in der Datenbank gespeichert sind.  
Standardwert = no

#### `update_player_ratings`

Gibt an, ob bei jeder Ausführung der aktuelle Rang der Spieler abgerufen werden soll.  
Standardwert = no

#### `request_timeout`

Gibt den Timeout für HTTP-Verbindungen in Sekunden an.  
Standardwert = 10.0

#### `throttle`

Ein Schalter, der eine zusätzliche Pause zwischen HTTP-Anfragen aktiviert. Die Wartezeit wird zufällig aus einem Intervall gewählt.  
Standardwert = no

#### `throttle_min`

Die minimale Wartezeit zwischen HTTP-Anfragen. Wirkt nur, wenn `throttle` aktiv ist. Standardwert = 1

#### `throttle_max`

Die maximale Wartezeit zwischen HTTP-Anfragen. Wirkt nur, wenn `throttle` aktiv ist.  
Standardwert = 10

#### `progress_interval`

Die Anzahl von Sekunden die zwischen Fortschrittsmeldungen verstreichen dürfen.  
Standardwert = 60

### Sektion `target`

#### `url`

Die Basis-URL für HTTP-Anfragen.  
Standardwert = `http://www.gjgr.de/`

#### `user-agent`

Die Zeichenkette, welche bei jeder HTTP-Anfrage im Header-Feld `User-Agent` mitgesendet wird.  
Standardwert = `SPI Prelytica WebCrawler`

### Sektion `system`

#### `html_parser`

Der Parser für das Parsen der HTML-Seiten. Möglich sind alle Parser, die von *Beautiful Soup 4* unterstützt werden: `html.parser`, `lxml` und `html5lib`.  
Standardwert = `html5lib`



Für die Konfiguration der MySQL-Datenbankverbindung wird die Sektion `database.gjgr` verwendet. Der Aufbau der Sektion ist in Konfiguration von Python-Komponenten beschrieben.

## Meteomatics Weather

- Typ: Programm
- Technologie: Python
- Ansprechpartner: Tobias Kiertscher, Markus Lampert

Das Python-Programm *Meteomatics Weather* ruft Wetterdaten von der Meteomatics API ab und speichert sie in einer Datenbank. Dabei werden nur jene Zeitreihen abgerufen, die noch nicht in der Datenbank gespeichert sind.

*Meteomatics Weather* hält sich nicht an die gemeinsame Architektur für Python-Komponenten.

## Dateistruktur

Das Programm ist in mehrere Python-Module aufgeteilt.

- `config.py` Parser für die Konfigurationsdatei
- `database.py` Schnittstelle zur SPI Datenbank
- `log.py` Methoden zur formatierten Ausgabe von Protokolleinträgen
- `meteomatics.py` Schnittstelle zur Web-API unter Verwendung der Python-Bibliothek *requests*
- `meteomatics_to_mysql.py` CLI-Skript mit Parser für Befehlszeilenargumenten
- `timerange.py`
  - `TimeRange` Klasse zur Repräsentation eines Zeitbereichs
- `weather.py`
  - `WeatherDownloader` Anwendungslogik für das Herunterladen neuer Wetterdaten
  - `MeteomaticsCsvFile` Parser für das Übersetzen vom CSV-Format in Objekte
  - `WeatherTransferManager` Anwendungslogik für das Ermitteln, Herunterladen und Speichern fehlender Wetterdaten
- `bin`
  - `meteomatics_to_mysql.bat` CMD-Wrapper
  - `meteomatics_to_mysql.sh` Bash-Wrapper

## Befehlszeile

Das CLI-Skript für *Meteomatics Weather* heißt `meteomatics_to_mysql.py`. Der folgende Hilfetext, der mit der Befehlszeilenoption `--help` ausgegeben werden kann, erklärt den Aufruf der Komponente detailliert.

```
usage: meteomatics_to_mysql.py [-h] [-v] [-sd START_DATE] [-m {sync,check}]
                               [-c CONFIG_FILES] [-o OPTIONS [OPTIONS ...]]

meteomatics_to_mysql

optional arguments:
  -h, --help            show this help message and exit
  -v, --version          Print the program version and exit
  -sd START_DATE, --start-date START_DATE
                        Get weather data from meteomatics since this UTC date
                        until 12 hours in the past. Given in the format
                        "2015-01-01T12:00"
  -m {sync,check}, --mode {sync,check}
                        Specifies the mode of operation. Mode "sync" is the
                        default mode.
  -c CONFIG_FILES, --config-file CONFIG_FILES
                        A path to a configuration file in UTF-8 encoded INI
                        format.
  -o OPTIONS [OPTIONS ...], --options OPTIONS [OPTIONS ...]
                        One or more configuration options, given in the format
                        <section>.<option>=<value>.
```

## Optionen

Die Konfiguration von *Meteomatics Weather* umfasst die folgenden Sektionen und Optionen.

### Sektion `database_spi`

#### `host`

IP oder Hostname des Datenbankservers mit der SPI Datenbank  
Standardwert = `localhost`

#### `user`

Benutzername für die Anmeldung am Datenbankserver  
Standardwert = `root`

#### `password`

Passwort für die Anmeldung am Datenbankserver

#### `database`

Name der Datenbank  
Standardwert = `spi`

## table

Quelltabelle mit den Golf-Clubs Standardwert = golf\_clubs

## Sektion database\_weather

### host

IP oder Hostname des Datenbankservers mit der Wetter Datenbank  
Standardwert = localhost

### user

Benutzername für die Anmeldung am Datenbankserver  
Standardwert = root

### password

Passwort für die Anmeldung am Datenbankserver

### database

Name der Datenbank  
Standardwert = meteomatics

### table

Zieltabelle für die Wetterdaten Standardwert = golf\_club\_weather\_hourly

## Sektion behavior

### time\_chunk\_size

Die maximale Anzahl von Tagen, die in einer Abfrage verwendet werden dürfen  
Standardwert = 30

### max\_chunk\_size

Die maximale Anzahl von GPS-Positionen, die in einer Anfrage verwendet werden dürfen  
Standardwert = 100

### max\_query\_days

Die maximale Anzahl von Abfragetagen (Tage \* Positionen), die in einer Anfrage heruntergeladen werden dürfen  
Standardwert = 1000

### exit\_on\_first\_error

Gibt an, ob das Programm beim ersten Fehler abbricht, oder fortfährt  
Standardwert = yes

### query\_attempts

Gibt an, wie oft eine Anfrage wiederholt werden soll, wenn sie nicht erfolgreich war  
Standardwert = 2

## Sektion `logging`

### `log_file`

Ein Pfad zu der Protokolldatei

Standardwert = ``

### `log_to_console`

Gibt an, ob die Protokollmeldungen auf der Standardausgabe ausgegeben werden sollen  
Standardwert = yes

## Sektion `meteomatics`

### `user`

Der Benutzername für die Meteomatics-API

### `password`

Das Passwort für die Meteomatics-API

# SPI Anwendung

Die *SPI Anwendung* umfasst alle Komponenten und Programme, mit denen Endbenutzer direkt oder indirekt interagieren. Dazu zählen die SPI Datenbank, das SPI Backend, die beiden Frontends Web und Mobil und zuletzt die Hintergrunddienste.

## SPI DB

- Typ: Datenbank
- Technologie: MySQL
- Ansprechpartner: Tobias Kiertscher, André Nitze

Die *SPI Datenbank* speichert alle notwendigen Daten über die Mandanten (Sportvereine), welche die SPI-Anwendung nutzen. Daten werden zum einen von Hintergrundprozessen in die Datenbank eingepflegt und zum anderen direkt durch Eingaben der Benutzer in eines der Frontends.

Die *SPI Datenbank* enthält die folgenden Tabellengruppen.

## Benutzerverwaltung

Die Mandanten- und Benutzerverwaltung ist in den Tabellen `clubs`, `user`, `user_settings` und `password_tokens` abgebildet. Die Attribute und Beziehungen dieser Tabellen sind in Abbildung 7 dargestellt. Die Mandanten sind in der Tabelle

clubs gespeichert. Alle Benutzer, ob Administrator, Koordinator, Trainer oder Spieler besitzen einen Eintrag in der Tabelle user. Jeder Benutzer besitzt eine UUID als systeminternes Identifikationsmerkmal. Die Rolle eines Benutzers wird durch die Spalte role unterschieden. Benutzerdefinierte Einstellungen werden in der Tabelle user\_settings gespeichert.

In der Tabelle password\_tokens werden zeitlich begrenzt gültige Tokens für das Zurücksetzen von Passwörtern gespeichert.

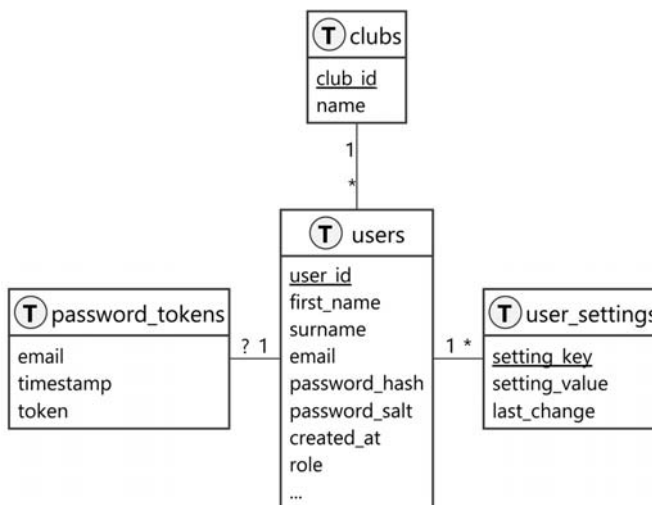


Abbildung 7: Tabellen für Mandanten- und Benutzerverwaltung

## Trainingsplanung

Die Trainingsplanung ist in den Tabellen schedules, rest\_periods, trainings und training\_dates\_canceled abgebildet. Die Attribute und Beziehungen dieser Tabellen sind in Abbildung 8 dargestellt. Jeder Trainingsplan (schedules) ist für einen bestimmten Zeitraum gültig und kann trainingsfreie Zeiten (Feiertage, Ferien, ...) enthalten (restperiods). Der Trainingsplan enthält Trainingsgruppen (trainings), welche sich regelmäßig an einem bestimmten Wochentag zu einer bestimmten Uhrzeit treffen. Wenn eine Trainingsgruppe sich nicht trifft, also das Training an einem bestimmten Tag ausfällt, z. B., weil der Trainer verhindert ist, wird das Datum des Ausfalls in training\_dates\_canceled gespeichert.

Jeder Trainingsgruppe sind genau ein Trainer und mehrere Spieler zugeordnet (user). Ein Trainer kann mehrere Trainingsgruppen betreuen und ein Spieler kann an mehreren Trainingsgruppen teilnehmen. Die M:N-Beziehung zwischen Training und Spieler ist in der Hilfstabelle `users_trainings` abgebildet.

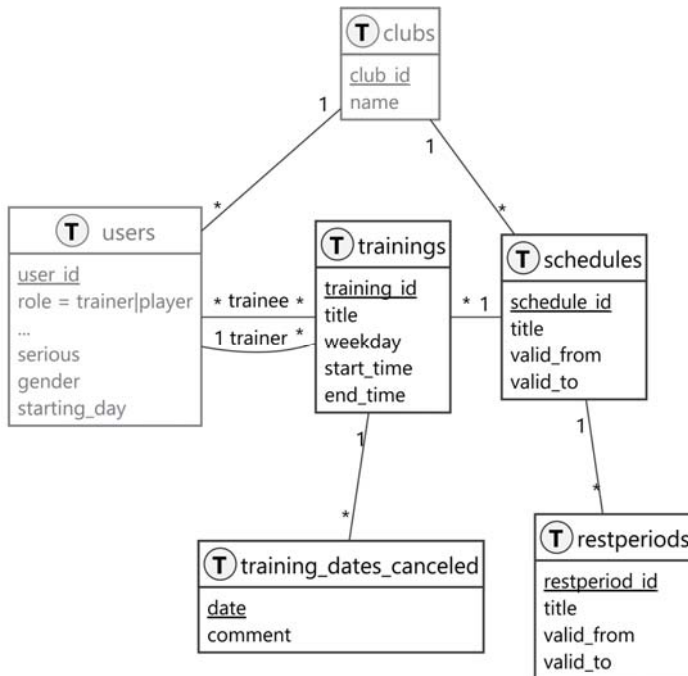


Abbildung 8: Tabellen für Trainingsplanung

## Trainingsaktivität

Die Trainingsaktivität der Spieler wird in den Tabellen `focus_type`, `attendance`, `attendance_focus`, `practice_trainings` und `practice_trainings_focus` erfasst. Die Attribute und Beziehungen dieser Tabellen sind in Abbildung 9 dargestellt. Ein Spieler kann entweder im Rahmen einer betreuten Trainingsgruppe trainieren oder selbstständig. Die Anwesenheit eines Spielers in einer Trainingsgruppe zu einem bestimmten Termin wird in der Tabelle `attendance` gespeichert. Es wird also explizit gespeichert, ob ein Spieler bei einem betreuten Training anwesend oder abwesend war. Es können auch Spieler an einem betreuten Training teilnehmen,

die der Trainingsgruppe laut Trainingsplan nicht zugeordnet sind. Das selbstständige Training wird in der Tabelle `practice_trainings` gespeichert.

Sowohl im betreuten als auch im selbstständigen Training kann der Spieler die Trainingszeit mit unterschiedlichen Trainingsschwerpunkten verbringen (`focus_types`). Die jeweilige Dauer, die der Spieler mit einem Trainingsschwerpunkt verbringt, wird in den Tabellen `attendance_focus` und `practice_trainings_focus` gespeichert.

Die Tabelle `focus_types` ist nicht mandantenfähig, d.h. alle Mandanten in einer Installation des SPI-Systems müssen die gleichen Trainingsschwerpunkte verwenden. Das macht für den Forschungscharakter des Systems durchaus Sinn, da nur so eine mandantenübergreifende Vergleichbarkeit der Trainingsdaten entsteht.

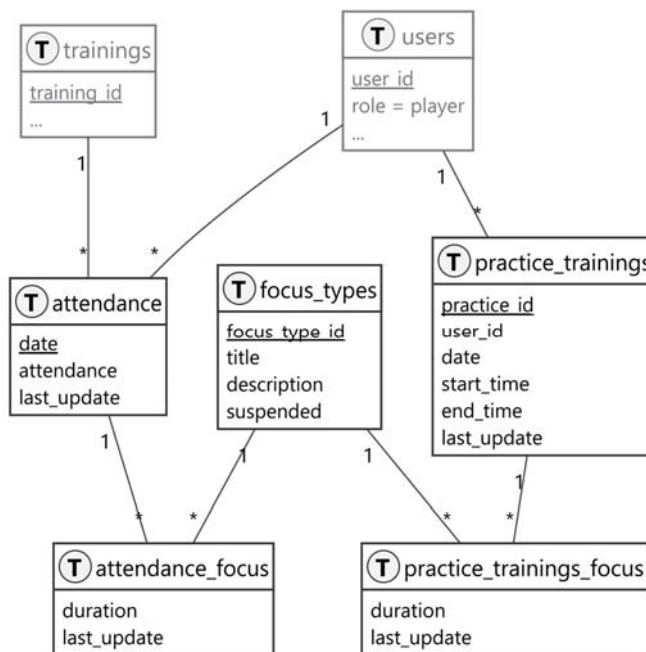


Abbildung 9: Tabellen für Trainingsaktivität

## Messwerterfassung

Messwerte zu den Spielern werden in den Tabellen `measurement_types`, `measurements` und `measurement_series` gespeichert. Die Attribute und Beziehungen dieser Tabellen sind in Abbildung 10 dargestellt. Für jeden Spieler können für jeden Tag Messwerte unterschiedlichen Typs erhoben werden (`measurements`). Für jeden Tag kann für einen Messwerttyp nur ein Messwert gespeichert werden. Das können physische Eigenschaften des Spielers wie Größe oder Gewicht sein, aber auch sportliche Leistungswerte. Setzt sich ein Messwert aus einer Messreihe zusammen kann diese in der Tabelle `measurement_series` und der aggregierte Wert dann in `measurements` gespeichert werden.

Die verfügbaren Messwerttypen sind in der Tabelle `measurement_types` abgelegt. Die Messwerttypen können nach Kategorien und Unterkategorien geordnet werden. Wird ein Messwert immer aus einer Messreihe abgeleitet, kann im Messwerttyp hinterlegt werden, wie viele Einzelwerte für die Messreihe erforderlich sind.

Das SPI-System bildet durch einen Hintergrundprozess selbstständig Messwerte. Das sind statistische Größen die aus anderen Daten erhoben werden. Z. B. die prozentuale Trainingsanwesenheit pro Monat und Jahr. Die Messwerttypen deren Messwerte durch das SPI-System gebildet werden sind als *berechnet* gekennzeichnet.

Bei der Anonymisierung eines Spielers, werden standardmäßig alle Messwerte unverändert aufbewahrt, es sei denn der Messwerttyp ist als *sensibel* gekennzeichnet, dann werden sie gelöscht.



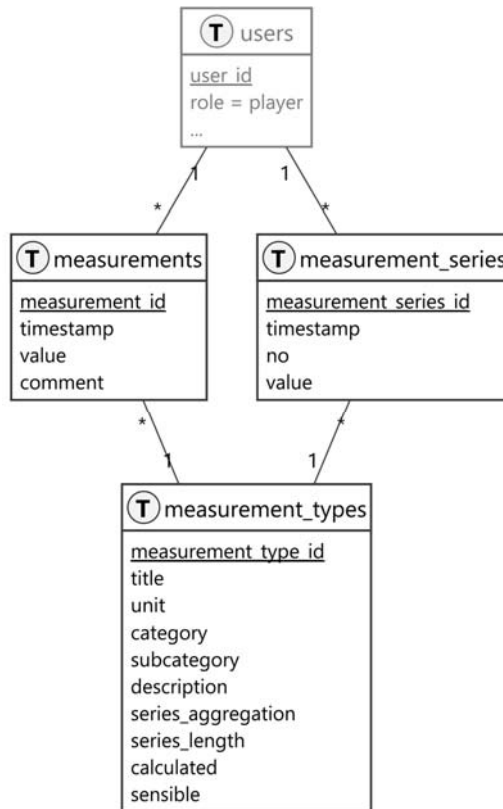


Abbildung 10: Tabellen für Messwerterfassung

## Golf-Turnierergebnisse

Die Turnierergebnisse der Golf-Spieler werden in den Tabellen `golf_clubs`, `golf_club_associations`, `golf_tournaments`, `golf_scorecards` und `golf_scores` gespeichert. Die Attribute und Beziehungen dieser Tabellen sind in Abbildung 11 dargestellt. Jeder Golf-Club, der Turniere veranstaltet, wird in der Tabelle `golf_clubs` gespeichert. Auch der jeweilige Golf-Club des Mandanten wird dort hinterlegt. Die Turniere werden in der Tabelle `golf_tournaments` gespeichert. Zu jedem Turnier gehören die dort erzielten Score-Karten die in `golf_scorecards` gespeichert werden. Im SPI-System werden nur Score-Karten gespeichert, die Spielern im SPI-System (`users`) gehören. Sind für eine Score-Karte auch die Ergebnisse der einzelnen Löcher bekannt, werden diese in `golf_scores` gespeichert.

Damit Score-Karten aus unterschiedlichen Turnieren vergleichbar werden, werden sie normiert. Ist eine Score-Karte in einem Stableford-Spiel entstanden, werden zunächst die gestrichenen Löcher geschätzt. Dazu wird die Anzahl von Schlägen verwendet, bei welcher das Loch 0 Stableford-Punkte erzielt hätte. Bei 9-Loch-Spielen wird die Schlagzahl verdoppelt. Diese geschätzte Schlagzahl wird in `golf_scorecards.estimated_score` gespeichert. Bei 9-Loch-Spielen wird auch das Kurs-Par verdoppelt so dass es mit der geschätzten Schlagzahl kompatibel ist. Als letzten Schritt wird die Schlagzahl anhand des Kurs-Par auf 72 normiert und das Ergebnis in `golf_scorecards.norm_score` gespeichert.

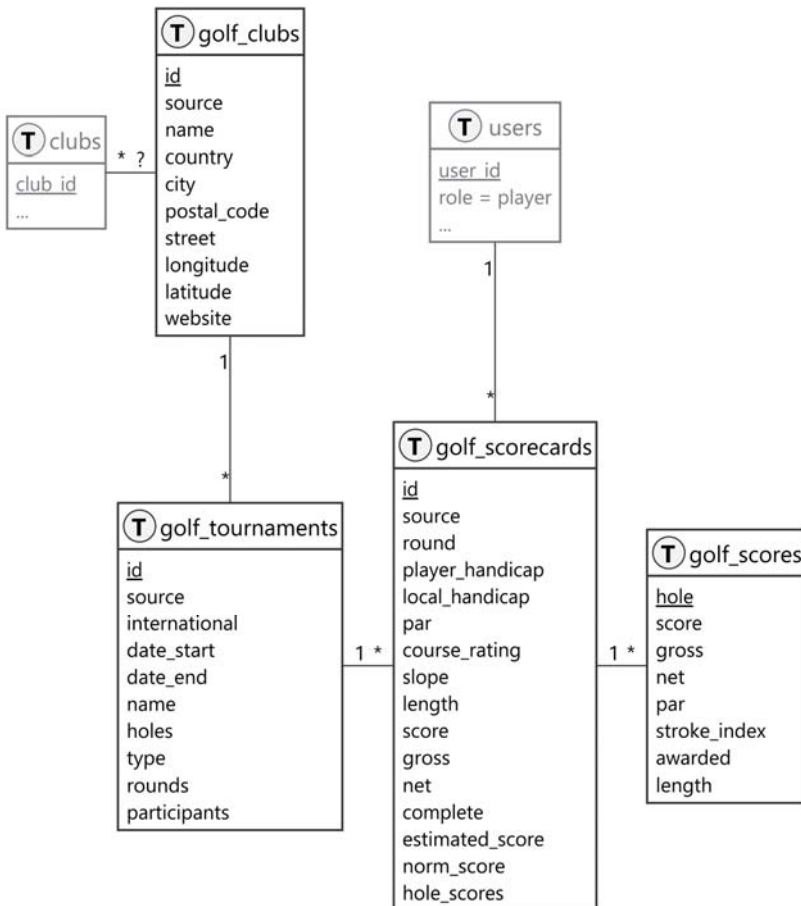


Abbildung 11: Tabellen für Golf-Turnierergebnisse

## Aggregierte Daten

Zur schnellen Abfrage und zur Anonymisierung von Spielern aggregierte Daten werden in den Tabellen `attendance_aggregated`, `training_aggregated` und `golf_aggregated` gespeichert. Die Attribute und Beziehungen dieser Tabellen sind in Abbildung 12 dargestellt. Wenn ein Spieler den Golf-Club eines Mandanten verlässt, darf dieser die personenbezogenen Daten des Spielers nicht länger aufbewahren. Um dennoch Möglichkeiten zur Datenanalyse zu behalten, werden Spieler die den Club verlassen nicht aus dem System gelöscht, sondern anonymisiert. Dazu werden ID, Namen, Email-Adresse und weitere Informationen gelöscht oder zufällig verändert. Einige Daten im SPI-System sind jedoch so individuell, dass sie ein Wiederherstellen der eigentlichen Identität sehr leicht machen würden. Dazu zählt z. B. die Teilnahme an bestimmten Golf-Turnieren. Deshalb werden die folgenden Daten-gruppen jeweils für den Zeitraum eines Quartals aggregiert:

- Trainingsanwesenheit und Pflegezustand der Trainingsanwesenheit
- Betreute und selbstständige Trainingsaktivität
- Golf-Turnielergebnisse

Die aggregierten Daten sind nicht nur nützlich bei anonymisierten Spielern. Da die Aggregation in einem Hintergrundprozess regelmäßig für alle nicht anonymisierten Spieler durchgeführt wird, können die aggregierten Daten für Zusammenfassungen genutzt werden und entlasten damit Backend und Datenbank. Während der Anonymisierung eines Spielers wird die Aggregation ein letztes Mal ausgeführt, damit die aggregierten Daten bei anonymisierten Spielern so aktuell wie möglich sind.

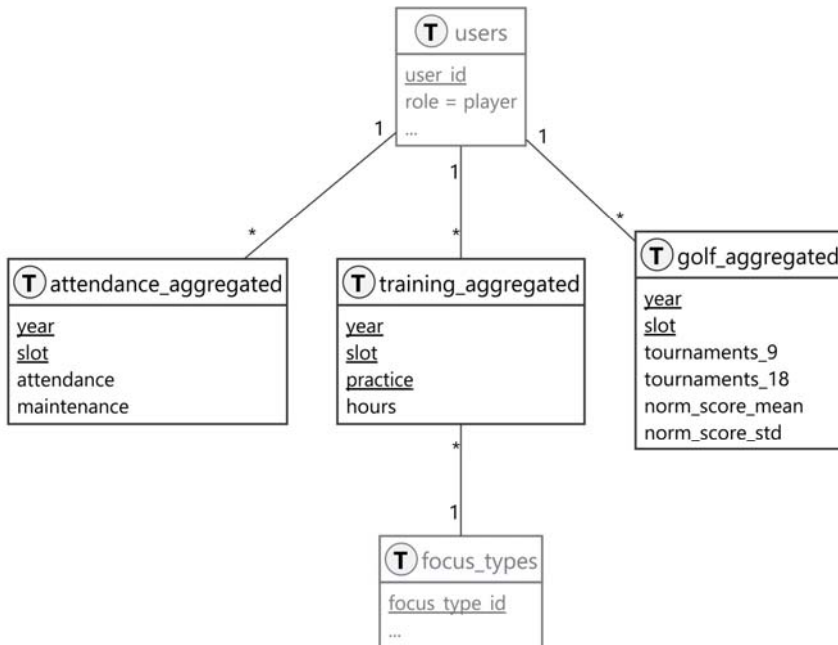


Abbildung 12: Tabellen für Aggregierte Daten

## Email-Berichte

Die Informationen über verschickte oder fehlgeschlagene Email-Berichte werden in den Tabellen `user_reports` und `user_report_values` gespeichert.

Administratoren, Koordinatoren und Trainer haben die Möglichkeit in ihren Benutzereinstellungen einen täglichen und/oder wöchentlichen Bericht zu aktivieren. Dieser wird durch einen Hintergrundprozess versendet. Der Bericht enthält eine Reihe von statistischen Größen und hebt Veränderungen dieser Größen hervor.

Um die Veränderungen zwischen dem letzten in der Vergangenheit verschickten Bericht und dem aktuellen zu ermitteln, wird für jeden Bericht ein Eintrag in `user_reports` und für jede Größe ein Eintrag in `user_report_values` gespeichert. Kann der Bericht aus technischen Gründen nicht versendet werden, wird das in `user_reports.send_error` vermerkt.

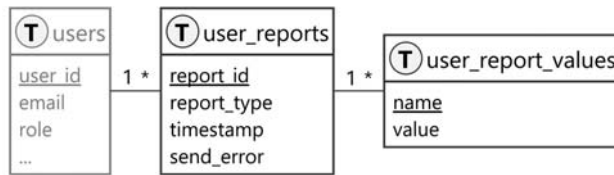


Abbildung 13: Tabellen für Email-Berichte

## SPI Backend

- Typ: Webanwendung
- Technologie: PHP
- Ansprechpartner: Tobias Kiertscher, André Nitze

Das *SPI Backend* ist eine PHP-Anwendung und kann mit einem PHP-FPM-Server oder als (F)CGI-Anwendung ausgeführt werden. Es benötigt einen Webserver (z. B. Apache oder nginx), um die Anwendung über einen HTTP(S)-Port verfügbar zu machen.

Das *SPI Backend* bietet eine REST-API über HTTP an, mit der Daten aus der SPI DB abgerufen und geändert werden können. Des Weiteren stellt es verschiedene Sichten mit statistischen Auswertungen für Spieler, Trainings, Clubs u. a. zur Verfügung. Der Zugriff ist über ein Rollen- und Rechtekonzept geschützt. Der Austausch der Daten über HTTP erfolgt fast ausschließlich in JSON. Die Authentifikation erfolgt zustandslos über JSON-Web-Tokens (JWT).

Das Backend kennt die folgenden Rollen, wobei jeder Benutzer immer genau eine Rolle besitzt:

- Administrator `admin`
- Koordinator `manager`
- Trainer `trainer`
- Spieler `player`

Das Backend ist begrenzt mandantenfähig. Das bedeutet, dass mehrere Golf-Clubs ihre Spieler in das System einpflegen können und es ist sichergestellt, dass nicht ein Benutzer eines Clubs die Daten eines anderen Clubs abrufen oder verändern kann. Einige Konfigurationsdaten gelten jedoch für alle Clubs gemeinsam und können nur durch einen Benutzer mit der Rolle Administrator geändert werden. Dazu zählen

die Trainingsschwerpunkte (*Focus Types*) und die Messwerttypen (*Measurement Types*).

## Verarbeitung einer Anfrage

Als Framework für die Abwicklung von Anfragen wird *Slim* in der Version 2.6 verwendet. Slim erlaubt es sprechende URLs für eine RESTful-API zu bilden. Die HTTP-Anfragen durchlaufen eine Reihe von Middleware-Objekten, welche die Möglichkeit haben, die Anfrage und die Antwort zu manipulieren. Die eigentliche Antwort wird in den Handler-Funktionen der Routen gebildet. Die Kette der Middleware-Objekte ist in Abbildung 14 dargestellt.

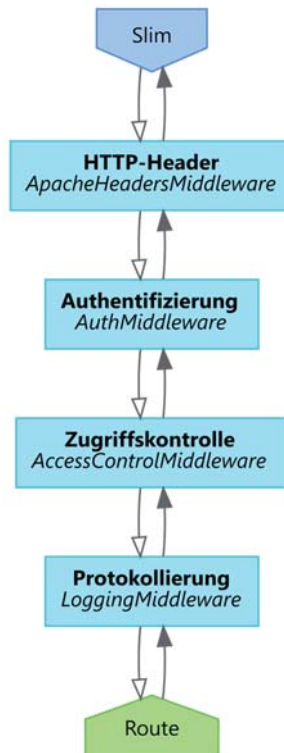


Abbildung 14: Slim Middleware-Objekte

## Slim

Das Backend bindet das Framework *Slim* durch Vererbung ein: Die zentrale Klasse *App* erbt von der Klasse *Slim\Slim*. Slim verpackt die HTTP-Anfrage in ein *Request*-Objekt und leitet sie durch die *Middleware*-Objekte. Das *Request*-Objekt enthält bereits die Information, an welche Route die Anfrage geleitet werden soll. Das hat Slim anhand der Pfadmuster der registrierten Routen bestimmt. Haben alle *Middleware*-Objekte die Anfrage weitergeleitet, übergibt Slim das *Request*-Objekt an die Route und nimmt ein *Response*-Objekt zurück. Dieses wird dann in umgekehrter Reihenfolge wieder durch alle *Middleware*-Objekte geleitet und zuletzt, falls noch erforderlich, serialisiert und an den Client geschickt.

## HTTP-Header

Einige Header-Informationen für die Authentifizierung werden von Slim nicht automatisch in das *Request*-Objekt verpackt, wenn der HTTP-Server ein Apache ist. Die *HTTP-Header-Middleware* prüft, ob die erforderlichen Informationen fehlen und ruft sie ggf. über eine Apache-spezifische Funktion ab.

## Authentifizierung

Für die zustandslose Authentifizierung werden JWTs und die gleichnamige Bibliothek von Luís Cobucci in Version 3.1 genutzt. Die Authentifizierungs-Middleware prüft zunächst, ob für die aktuelle Route eine Authentifizierung erforderlich ist. Diese Information fragt sie von der Klasse *AccessControllist* ab. Anschließend entnimmt sie dem *Request*-Objekt das JWT und prüft ob es gültig ist und mit den Benutzerdaten in der Datenbank übereinstimmt. Dann speichert es die Token-Informationen in der *App*-Klasse, wo sie von der *Session*-Klasse genutzt werden können, um das Datenobjekt des für die aktuelle Anfrage authentifizierten Benutzers zu laden. Ist das JWT ungültig oder nicht vorhanden, obwohl für die aktuelle Route eine Authentifizierung erforderlich ist, reicht die *Middleware* die Anfrage nicht zur nächsten *Middleware* weiter, sondern antwortet mit einem HTTP-Status-Code 401 für *“Nicht autorisiert”*.

## Zugriffskontrolle

Jede Route im SPI Backend hat neben ihrem Pfadmuster auch einen eindeutigen Namen. Die Middleware für die Zugriffskontrolle lädt mit der `Session`-Klasse das Benutzerobjekt des aktuell authentifizierten Benutzers. Anschließend prüft es mit der Klasse `AccessControllList`, ob dem Benutzer Zugriff auf die aktuelle Route gestattet ist.

Die Zugriffskontrolle erlaubt die Steuerung in drei Ebenen: `unrestricted`, `club`, `personal`. Für jede Route muss definiert werden, welche Benutzerrolle in welcher Ebene Zugriff erhält. So haben auf die Route für das Laden der Spielerakte z. B. Administratoren unbeschränkten Zugriff, Trainer nur im eigenen Club (Mandant) und Spieler nur auf ihre eigene Akte.

Ist der Zugriff für den authentifizierten Benutzer erlaubt, wird die Anfrage durchgereicht, andernfalls wird mit einem HTTP-Status-Code 403 für *“Zugriff verweigert”* geantwortet.

## Protokollierung

Die Protokollierung der Anfragen ist in der Klasse `LoggingMiddleware` implementiert. Aus der Anfrage werden Anfrageeigenschaften wie HTTP-Methode und Pfad entnommen und zwischengespeichert. Dann wird die Anfrage an die nächste Middleware weitergeleitet. Kommt die Antwort zurück, werden ihr weitere Eigenschaften wie der HTTP-Status-Code und der authentifizierte Benutzer entnommen. Zuletzt werden die Eigenschaften in eine Textzeile kodiert und an die Logging-Infrastruktur übergeben.

Für die Serialisierung von Objekten für die HTTP-Antworten wird *JMS-Serializer* in der Version 1.6 verwendet.

## Routen

Die Routen sind thematisch gruppiert und in einer Reihe von PHP-Dateien implementiert. Jede Route wird bei ihrer Definition bei dem globalen `App`-Objekt und damit im Slim-Framework registriert. Beim Aufruf erhalten sie sowohl eine Referenz auf das `App`-Objekt als auch das aktuelle `Request`-Objekt. Sie können verschiedene Dienste nutzen, auf die sie über das `App`-Objekt Zugriff erhalten.



## Anwendungskern

Die Middleware-Objekte und die Routen haben Zugriff auf das App-Objekt und damit auf alle Kerndienste der PHP-Anwendung. Die App-Klasse stellt die Kerndienste über den Inversion-of-Control-Mechanismus (IoC) der Slim-Klasse zur Verfügung. Dadurch werden Dienstobjekte erst bei Bedarf als Singletons initialisiert. Die Beziehungen der Dienstobjekte und des App-Objekts sind in Abbildung 15 dargestellt.

Die Kerndienste haben die folgenden Aufgaben:

**log: Slim\Log**

für die Protokollierung

**ormCache: Doctrine\Common\Cache\CacheProvider**

für das Zwischenspeichern von ORM-Metadaten

**em: Doctrine\ORM\EntityManager**

für den direkten Zugriff auf die Datenbank über das OR-Mapping

**serializer: JMS\Serializer\Serializer**

für die vorkonfigurierte De-/Serialisierung von Objekten in und aus JSON

**model: DataModel**

für einen Großteil der mehrfach verwendeten Anwendungslogik im Umgang mit Datenobjekten

**session: Session**

für das Laden des Benutzerobjekts für den authentifizierten Benutzer

**stats: Statistics**

für statistische Berechnung auf Benutzerdaten  
(z. B. Trainingsdaten, Messwerten, Turnierergebnissen)

**conv: DataConversion**

für die Umwandlung und die Kombination von Datenobjekten aus der Datenbank-sicht in die Client-Sicht (z. B. Trainerprofil, Spielerakte)

**out: ResponseFormatter**

für das komfortable Erzeugen von Response-Objekten

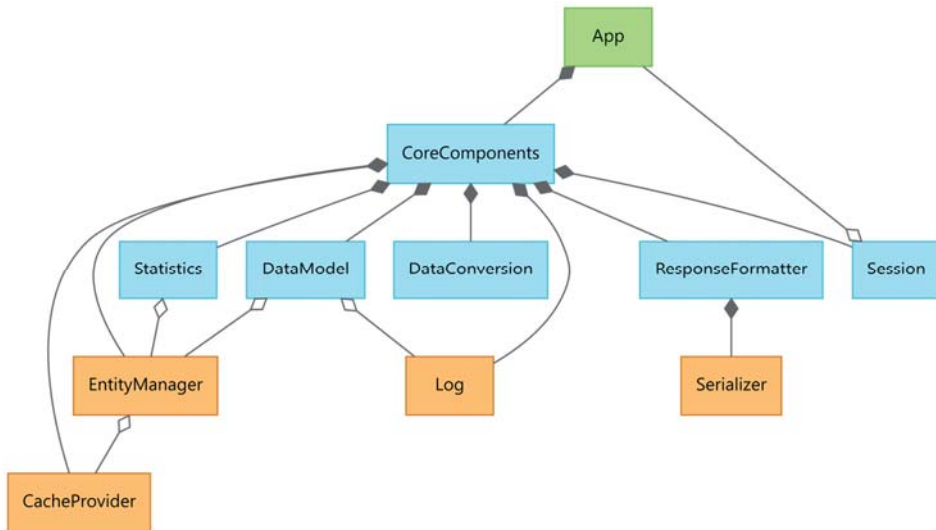


Abbildung 15: Kerndienste des SPI Backend

## Datenbankzugriff

Als Framework für das Mapping zwischen relationaler Datenbank und Objekten wird *Doctrine* in der Version 2.6 verwendet. Der *EntityManager* erlaubt den generischen, objektorientierten Zugriff auf die Datenbank. Alle Daten werden in Form von Datenobjekten ausgetauscht. Die dazugehörigen Klassen besitzen Annotation in den PHP-Doc-Kommentaren, welche die Zuordnung der Klassenattribute zu den Tabellen und Spalten beschreiben. Auch die Beziehungen zwischen den Objekten (1:1, 1:N, M:N, Vererbung) werden in Annotationen definiert.

Komfortabler und im Quellcode besser lesbar ist die Verwendung der auf dem *EntityManager* aufbauenden Klasse *DataModel*. Sie besitzt eine Vielzahl von spezialisierten Methoden für typische Szenarien in den Routen. Z. B. das Abrufen aller Spieler für einen Trainer oder das Ermitteln des aktuellen Trainingsplans für den Club des authentifizierten Benutzers.

## Dateien und Namensräume

Die folgende Übersicht zeigt in Auszügen den Aufbau des PHP-Projektes.

- `index.php` (Einsprungspunkt der Web-Anwendung)
- `config.php` (Konfiguration laden, prüfen, vervollständigen)

- `backend-config.ini` (Konfiguration im INI-Format)
- `shutdown_hook.php` (Ausgabe von fatalen Fehlern im JSON-Format)
- `utility.php`
- `app` (Hauptnamensraum)
  - `Jobs` (Namensraum `Jobs` mit Klassen für Hintergrundaufgaben)
    - `IBackgroundJob.php`
    - `AutoMeasurements.php`
    - `Initialization.php`
    - `MailReport.php`
    - ...
  - `Report` (Namensraum `Report` mit Klassen für Email-Berichte)
    - `Delta.php`
    - `IReportSection.php`
    - `Report.php`
    - `ReportMailer.php`
    - `StatisticsTable.php`
    - ...
  - `App.php`
  - `AccessControllList.php`
  - `AccessControlMiddleware.php`
  - `AuthMiddleware.php`
  - `CoreComponents.php`
  - `DataConversion.php`
  - `DataModel.php`
  - `InputParsing.php`
  - `Session.php`
  - ...
- `models` (Namensraum `Models` mit Datenklassen)
  - `Administrator.php`
  - `Attendance.php`
  - `Club.php`
  - `FocusType.php`
  - ...
  - `Client` (Namensraum `Models\Client` mit Datenklassen für den Client)
    - `AnonymizedPlayerSummary.php`
    - `PlayerProfile.php`
    - `SystemStats.php`

- ...
- Golf (Namensraum Models\Golf für golfspezifische Datenklassen)
  - GolfClub.php
  - GolfTournament.php
  - GolfScorecard.php
  - ...
- routes (Verzeichnis mit PHP-Dateien, die Routen implementieren)
  - attendances.php
  - auth.php
  - clubs.php
  - managers.php
  - players.php
  - ...

## Konfiguration

Das SPI Backend (und auch die SPI Backend Jobs) werden in der Datei `backend-config.ini` konfiguriert. Es werden die folgenden Konfigurationsoptionen unterstützt.

### Datenbankverbindung

#### `db_server`

Hostname oder IP-Adresse des Datenbankservers mit der SPI Datenbank  
Standardwert = db

#### `db_name`

Der Name der SPI Datenbank  
Standardwert = spi

#### `db_user`

Der Benutzer zur Anmeldung am Datenbankserver  
Standardwert = spi

#### `db_password`

Das Passwort zur Anmeldung am Datenbankserver

#### `generate_orm_proxies`

Bestimmt ob zur Laufzeit der Anwendung die Proxy-Klassen für das OR-Mapping generiert werden sollen, oder ob davon ausgegangen werden kann, dass diese bereits existieren  
Standardwert = false

#### `orm_proxy_dir`

Das Verzeichnis in dem die generierten Proxy-Klassen für das OR-Mapping gespeichert werden

Standardwert = `/var/www/html/tmp/ormproxies`

#### `orm_cache_enabled`

Ein Schalter, der das Zwischenspeichern der OR-Metadaten aktiviert. Ist dieser Schalter auf `true` gesetzt, muss entweder ein `orm_cache_dir` oder ein `memcached_host` angegeben werden.

Standardwert = `false`

#### `orm_cache_dir`

Das Verzeichnis, in dem OR-Metadaten zwischengespeichert werden können

Standardwert = `/var/www/html/tmp/ormcache`

#### `memcached_host`

Hostname oder IP-Adresse eines Memcached-Servers

Standardwert = leer

### Initialisierung

#### `autoinit_testdata`

Ein Schalter, der das Erzeugen von Testdaten für automatisierte Tests in der SPI-Datenbank aktiviert

Standardwert = `false`

#### `jwtsecret`

Eine geheime Zeichenkette, mit der die Tokens zur Authentifizierung verschlüsselt werden

Standardwert = `0123456789ABCDEF`

#### `initial_admin_password`

Das initiale Passwort des primären Administrator-Benutzers im SPI-System

Standardwert = `spi`

### Protokollierung und Debugging

#### `debug`

Ein Schalter für das Aktivieren von Debug-Meldungen und der Ablaufverfolgung in der REST-API

Standardwert = `false`

#### `monitor_performance`

Ein Schalter der Zeitmessungen für unterschiedliche Phasen in der Anfragebearbeitung aktiviert und als HTTP-Header in der Antwort mitsendet

Standardwert = `false`

**logfile**

Ein Pfad zur Protokolldatei für Zugriffe auf das *SPI Backend*

Standardwert = /var/log/spi/backend.log

**jobs\_logfile**

Ein Pfad zur Protokolldatei für die Hintergrunddienste

Standardwert = /var/log/spi/backend-jobs.log

**anonymize\_ip**

Ein Schalter der die IP-Adressen der Zugriffe im Protokoll anonymisiert

Standardwert = true

**jobs\_log\_to\_console**

Ein Schalter der die Ausgabe von Protokollmeldungen bei der Ausführung von Hintergrunddiensten auf der Standardausgabe aktiviert

Standardwert = false

**Frontend-Optionen****frontend\_base\_url**

Die URL für die Single-Page-App des Web-Frontends

Standardwert = "http://localhost/spi/"

**Passwort-Reset-Optionen****password\_reset\_lifetime**

Gibt die Anzahl von Stunden an, die ein Link zum Zurücksetzen eines Passwortes gültig ist

Standardwert = 48

**password\_reset\_reply\_name**

Der Antwortadresse für Emails zum Zurücksetzen eines Passwortes

Standardwert = PreLytica SPI Administrator

**password\_reset\_reply\_address**

Die Antwortadresse für Emails zum Zurücksetzen eines Passwortes

Standardwert = no-reply@localhost

**password\_reset\_link**

Der relative Link im Frontend zum Formular für das Zurücksetzen eines Passwortes.

Das Reset-Token wird in den Platzhalter \$token eingesetzt.

Standardwert = #/password-reset/\$token

## Email-Berichte

### report\_storage\_dir

Ein Pfad zu einem Verzeichnis in dem der HTML-Inhalt von versendeten Email-Berichten gespeichert werden soll

Standardwert = /var/www/html/reports

### report\_reply\_name

Der Antwortname in einem Email-Bericht

Standardwert = PreLytica SPI Administrator

### report\_reply\_address

Die Antwortadresse in einem Email-Bericht

Standardwert = no-reply@localhost

### report\_user\_settings\_link

Der relative Link im Frontend zu den Benutzereinstellungen

Standardwert = #/user-settings

### report\_training\_date\_link

Der relative Link im Frontend zu einem Trainingstermin. Die Platzhalter \$trainingId und \$date werden durch die ID des Trainings und das Datum des Termins ersetzt.

Standardwert = #/training/\$trainingId/\$date

### report\_player\_link

Der relative Link im Frontend zu einem Spielerprofil. Der Platzhalter \$userId wird durch die ID des Spielers ersetzt.

Standardwert = #/player/\$userId

### report\_scorecard\_link

Der relative Link im Frontend zu einer Score-Karte. Der Platzhalter \$scorecardId wird durch die ID der Score-Karte ersetzt.

Standardwert = #/scorecard/\$scorecardId

## Email-Server

### mail\_from\_name

Der Absendername für Emails

Standardwert = THB PreLytica SPI Server

### mail\_from\_address

Die Absenderadresse für Emails

Standardwert = spi@localhost

### mail\_mode

Der Modus für das Versenden von Emails: `sendmail` oder `smtp`

Standardwert = `sendmail`

Die folgenden Optionen gelten nur, wenn der `mail_mode` auf `smtp` gesetzt ist.

**mail\_smtp\_debug**

Ein Schalter, der die Ausgabe von Debug-Informationen während dem Versenden von Emails aktiviert  
Standardwert = `false`

**mail\_smtp\_host**

Hostname oder IP-Adresse eines SMTP-Servers

**mail\_smtp\_port**

Portnummer eines SMTP-Servers  
Standardwert = `587`

**mail\_smtp\_secure**

Verschlüsselung der SMTP-Verbindung: `tls`, `ssl`, `none`  
Standardwert = `tls`

**mail\_smtp\_username**

Der Benutzername zur Anmeldung am SMTP-Server

**mail\_smtp\_password**

Das Passwort zur Anmeldung am SMTP-Server

## SPI Backend Jobs

- Typ: Programm
- Technologie: PHP
- Ansprechpartner: Tobias Kiertscher

Die *SPI Backend Jobs* benutzen in weiten Teilen den Quellcode des SPI Backend, können jedoch als Kommandozeilenprogramm ausgeführt werden. Ihre Aufgabe ist es u. a., periodisch aggregierte Sichten zu generieren, automatisch berechnete Messwerte zu aktualisieren und E-Mail-Berichte zu versenden.

Die Befehlszeilenschnittstelle (CLI) für die Backend Jobs ist in der Datei `backgroundjob.php` implementiert. Die Ausführungssteuerung übernimmt die Klasse `BackgroundJobRunner`. Die einzelnen Jobs sind im Namensraum `Jobs` untergebracht.

## Jobs

Es gibt eine Anzahl unterschiedlicher Jobs und jeder Job kann mehrere Aufgaben enthalten. Die folgenden Jobs werden unterstützt.



## Initialisierung

Der Job INIT enthält nur die Aufgabe DATABASE.

Die Aufgabe INIT DATABASE überprüft, ob die Datenbank die für das SPI-System benötigte Tabellenstruktur enthält. Falls nicht, werden die Tabellen mit Hilfe der Datei `schema.sql` angelegt.

## Aufräumen

Der Job CLEANUP enthält nur die Aufgabe ORMCACHE.

Die Aufgabe CLEANUP ORMCACHE leert den konfigurierten ORM-Cache. Dies ist erforderlich, wenn die Datenklassen z. B. durch ein Update geändert wurden und ein persistenter ORM-Cache (dateibasiert oder Memcache) konfiguriert ist.

## Aggregationen

Der Job AGGREGATE enthält für jeden Spieler eine Aufgabe, die durch die ID des Spielers bezeichnet wird.

Bei der Aggregation für einen Spieler werden alle diesen Spieler betreffenden Daten in den Tabellen `attendance_aggregated`, `training_aggregated` und `golf_aggregated` aktualisiert.

## Berechnete Messwerte

Der Job AUTOMEASURE enthält für jeden Spieler eine Aufgabe, die durch die ID des Spielers bezeichnet wird.

Für jeden Spieler werden die folgenden Messwerte automatisch berechnet:

- Trainingsanwesenheit (jährlich, monatlich)
- Selbstständiges Training (jährlich, monatlich)
- Anzahl der Score-Karten (jährlich, monatlich)
- Schlagzahl (für jede Score-Karte)
- Handicap (für jede Score-Karte die eine neue Stammvorgabe enthält)

## Email-Berichte

Der Job MAILREPORT enthält für jeden Benutzer, der in seinen Benutzereinstellungen einen Email-Bericht aktiviert hat, eine Aufgabe, die durch die ID des Benutzers bezeichnet wird.

Wird die MAILREPORT-Aufgabe für einen Benutzer ausgeführt, werden die Benutzereinstellungen des Benutzers geladen und geprüft ob und welcher Bericht in der aktuellen Stunde gesendet werden soll. Es wird also der Berichtstyp (wöchentlich, täglich), der Wochentag und die Uhrzeit berücksichtigt.

Damit die Email-Berichte wie beabsichtigt versendet werden, muss der Job MAILREPORT jede Stunde einmal ausgeführt werden.

## Ausführung

Das PHP-Skript `backgroundjob.php` für die Ausführung von Hintergrundaufgaben hat die folgende Befehlszeilensyntax:

```
php ./backgroundjob.php [<JOB> [<AUFGABE>]]
```

Eine konkrete Aufgabe eines Jobs wird mit der folgenden Befehlszeile ausgeführt:

```
php ./backgroundjob.php <JOB> <AUFGABE>
```

Wird nur der Job aber keine Aufgabe angegeben, werden die Bezeichner der Aufgaben ausgegeben:

```
php ./backgroundjob.php <JOB>
```

Wird die CLI ohne Argumente ausgeführt, gibt sie die Namen der Jobs aus:

```
php ./backgroundjob.php
```

Das Bash-Skript `backgroundjob.sh` besitzt die gleiche Befehlszeilensyntax, ändert die Semantik aber wie folgt: Wird nur der Job angegeben aber keine Aufgabe, werden alle Aufgaben des Jobs ausgeführt. Wird kein Job angegeben werden alle Jobs mit allen Aufgaben ausgeführt. Wird der Job und die Aufgabe angegeben, verhält sich das Bash-Skript genauso wie das PHP-Skript.

### Beispiele

- Initialisieren der Installation:  
`./backgroundjob.sh INIT`

- Stündlicher Aufruf für das Versenden der Email-Berichte:  
`./backgroundjob.sh MAILREPORT`
- Die berechneten Messwerte eines Spielers aktualisieren:  
`./backgroundjob.sh AUTOMEASURE 13b53078-fb32-13e5-8a5d-bc305bd31271`  
(Die UUID muss durch die tatsächliche ID des Spielers ersetzt werden.)
- Leeren des ORM-Caches:  
`php ./backgroundjob.php CLEANUP ORMCACHE`
- Auflisten aller Spieler, für die eine Aggregation durchgeführt werden kann:  
`php ./backgroundjob.php AGGREGATE`

## Konfiguration

Die Konfiguration für die SPI Backend Jobs ist die gleiche wie für das SPI Backend. Und sie wird ebenfalls aus der Datei `backend-config.ini` gelesen.

## SPI Frontend Web

- Typ: Web-UI
- Technologie: JavaScript, HTML, CSS
- Ansprechpartner: Tobias Kiertscher, André Nitze

Das *SPI Frontend Web* (Web-Portal) ist eine Single-Page-Anwendung, implementiert mit AngularJS 1.4 und Bootstrap 4. Sie wird in Form von statischen Ressourcen (HTML, CSS, JavaScript) durch einen Web-Server ausgeliefert und kommuniziert per AJAX mit der REST-API des SPI Backend.

Das Frontend umfasst die folgenden Funktionen:

- unterschiedliche Dashboards für Benutzer unterschiedlicher Rollen
- eine Liste aller Spieler in einem Club mit Sortier- und Filterfunktionen
- eine Spielerakte für jeden Spieler
  - Zusammenfassung
  - Trainingsaktivität (Anwesenheit, Schwerpunkte)
  - Messdatenübersicht
  - Turnierergebnisse (im Verlauf und Detail)
- komfortabler Editor für Trainingspläne
- Übersicht des Pflegezustands der Trainingsanwesenheit für alle Trainer
- Einzelsicht für jeden Trainer mit Trainingsgruppen und zugeordneten Spielern
- Einzelsicht für jede Trainingsgruppe
- Erfassung von Messwerten

- für einzelne Spieler
  - tabellarisch für Spielergruppen
- Benutzerverwaltung
- Administrationsoberfläche
  - Mandantenverwaltung
  - Trainingsschwerpunkte
  - Messwerttypen

Eine ausführliche Beschreibung des Web-Portals [1] ist im gleichnamigen Kapitel ab Seite 161 zu finden.

## SPI Frontend Mobile

- Typ: Mobil-App
- Technologie: TypeScript, Ionic
- Ansprechpartner: André Nitze, Kilian Friedrich

Das *SPI Frontend Mobile* ist eine Mobil-App für iOS und Android. Es ist in TypeScript mit Angular implementiert und benutzt das Ionic-Framework.

Wie das SPI Frontend Web kommuniziert es per AJAX mit der REST-API des SPI Backend.

Das Frontend umfasst u. a. die folgenden Funktionen für Trainer:

- Erfassen von Trainingsanwesenheit und Trainingsschwerpunkten
- Erfassen von Messwerten für einzelne Spieler

Für Spieler umfasst es u. a. die folgenden Funktionen:

- Erfassen von freiem Training mit Trainingsschwerpunkten
- Erfassen von eigenen Messwerten

Eine ausführliche Beschreibung der Mobile-App [2] ist im gleichnamigen Kapitel ab Seite 181 zu finden.

# Datenintegration

Die Rohdaten aus den externen Datenquellen müssen mit den internen Daten zusammengeführt werden, damit im Online-Portal die verknüpften Informationen angezeigt werden können. Diese Aufgabe übernehmen die beiden im Folgenden beschriebenen Transferkomponenten.

## Golf.de Transfer

- Typ: Programm
- Technologie: Python
- Ansprechpartner: Tobias Kiertscher

Das Programm *Golf.de Transfer* wird immer dann ausgeführt, wenn die Rohdaten in Golf.de DB durch den Golf.de Crawler aktualisiert wurden. Es aktualisiert die Clubs, Turniere und Turnierergebnisse der Spieler in der SPI DB.

## Package-Struktur

Die Python-Komponente *Golf.de Transfer* basiert auf der gemeinsamen Architektur für Python-Komponenten und ist zusätzlich in den folgenden Packages und Klassen implementiert:

- `spi.transfer.golf`
  - `GolfTransfer`  
Eine Klasse, welche von `spi.transfer.TransferBase` erbt und die Übertragung der Daten von Golf.de DB nach SPI DB organisiert.
  - `conversion`  
Ein Modul mit Funktionen zur Konvertierung von Datenklassen der Golf.de DB aus `spi.crawler.golf.model` zur SPI DB aus `spi.transfer.model`.

## Befehlszeile

Das CLI-Skript für *Golf.de Transfer* heißt `golftransfer.py`.

```
usage: golftransfer.py [-h] [-v] [-c CONFIG_FILES] [-o OPTIONS [OPTIONS ...]]
```

SPI Golf Transfer

optional arguments:

```
-h, --help            show this help message and exit
-v, --version         print the program version and exit
-c CONFIG_FILES, --config-file CONFIG_FILES
                        A path to a configuration file in UTF-8 encoded INI
                        format. This argument can be used multiple times.
-o OPTIONS [OPTIONS ...], --options OPTIONS [OPTIONS ...]
                        One or more configuration options, given in the format
                        <section>.<option>=<value>.
```

## Optionen

Für die Konfiguration der MySQL-Datenbankverbindungen werden die Sektionen `database.golf` und `database.spi` verwendet. Der Aufbau der Sektionen ist in Konfiguration von Python-Komponenten beschrieben.

Darüber hinaus unterstützt diese Komponente die allgemeinen Optionen der Sektion `logging`.

## GJGR Transfer

- Typ: Programm
- Technologie: Python
- Ansprechpartner: Tobias Kiertscher

Das Programm *GJGR Transfer* wird immer dann ausgeführt, wenn die Rohdaten in GJGR DB durch den GJGR Crawler aktualisiert wurden. Es aktualisiert die Clubs, Turniere und Turnierergebnisse der Spieler in der SPI DB.

Die Daten des Golf.de Transfer haben eine höhere Priorität als die des *GJGR Transfer*. Das bedeutet, dass Turnierergebnisse von GJGR nur dann in die SPI DB übertragen werden, wenn noch keine vergleichbaren Daten aus Golf.de übertragen wurden. Das gilt insbesondere für die internationalen Turniere, welche von Golf.de nicht veröffentlicht werden.

## Package-Struktur

Die Python-Komponente *GJGR Transfer* basiert auf der gemeinsamen Architektur für Python-Komponenten und ist zusätzlich in den folgenden Packages und Klassen implementiert:

## ■ spi.transfer.gjgr

### □ GjgrTransfer

Eine Klasse, welche von `spi.transfer.TransferBase` erbt und die Übertragung der Daten von GJGR DB nach SPI DB organisiert.

### □ conversion

Ein Modul mit Funktionen zur Konvertierung von Datenklassen der GJGR DB aus `spi.crawler.gjgr.model` zur SPI DB aus `spi.transfer.model`.

## Befehlszeile

Das CLI-Skript für *GJGR Transfer* heißt `gjgrtransfer.py`.

```
usage: gjgrtransfer.py [-h] [-v] [-c CONFIG_FILES] [-o OPTIONS [OPTIONS ...]]
```

SPI GJGR Transfer

optional arguments:

```
-h, --help            show this help message and exit
-v, --version          print the program version and exit
-c CONFIG_FILES, --config-file CONFIG_FILES
                        A path to a configuration file in UTF-8 encoded INI
                        format. This argument can be used multiple times.
-o OPTIONS [OPTIONS ...], --options OPTIONS [OPTIONS ...]
                        One or more configuration options, given in the format
                        <section>.<option>=<value>.
```

## Optionen

Für die Konfiguration der MySQL-Datenbankverbindungen werden die Sektionen `database.gjgr` und `database.spi` verwendet. Der Aufbau der Sektionen ist in Konfiguration von Python-Komponenten beschrieben.

Darüber hinaus unterstützt diese Komponente die allgemeinen Optionen der Sektion `logging`.

## Datenanalyse

Die *Datenanalyse* umfasst Programme, welche Daten aus externen Quellen bereinigen und zusammenführen. Diese Programme werden zum aktuellen Projektstand nur manuell durch einen Analysten ausgeführt, um aus den gesammelten Daten neue Erkenntnisse zu gewinnen und die Implementierung von weiteren Funktionalitäten in der SPI Anwendung vorzubereiten.

## Analyzer

- Typ: Programm
- Technologie: Python
- Ansprechpartner: Tobias Kiertscher

Der *Analyzer* ist eine Gruppe von Programmen die Rohdaten aus Golf.de und GJGR vorverarbeiten und Gemeinsamkeiten identifizieren. Sie werden manuell durch einen Analysten ausgeführt, um eine konsistente Datenbasis zu schaffen, die weitere Analysen auf den Turnierdaten ermöglicht.

Die folgenden Verarbeitungsschritte sind implementiert:

- Golf.de Analyzer  
In Abbildung 16 ist der Datenfluss dargestellt.
  - Score-Karten in Golf.de normalisieren
  - Spieler in Golf.de identifizieren
  - Score-Karten den identifizierten Spielern zuordnen
- GJGR Analyzer  
In Abbildung 17 ist der Datenfluss dargestellt.
  - Score-Karten in GJGR normalisieren
- Common Analyzer  
In Abbildung 18 ist der Datenfluss dargestellt.
  - Übereinstimmende Golf-Clubs in Golf.de und GJGR finden
  - Übereinstimmende Spieler in Golf.de und GJGR finden

Eine Besonderheit der *Analyzer*-Programme ist die Unterstützung für mehrere Prozessorkerne. Für das normalisieren der Score-Karten werden optional mehrere Prozesse gestartet, um die Verarbeitung zu beschleunigen.



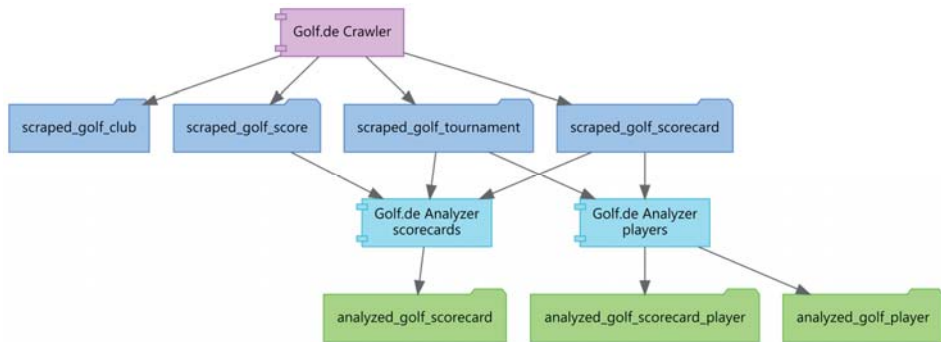


Abbildung 16: Datenfluss der Golf.de-Daten im Analyzer

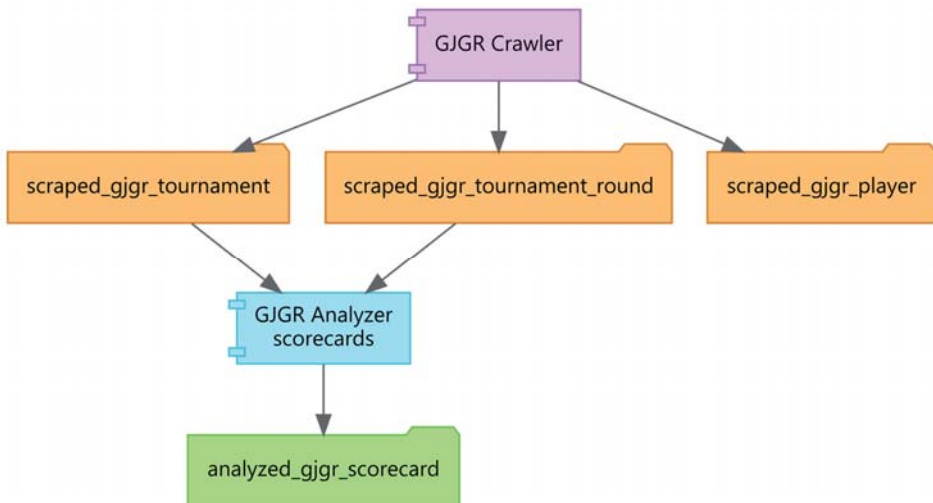


Abbildung 17: Datenfluss der GJGR-Daten im Analyzer

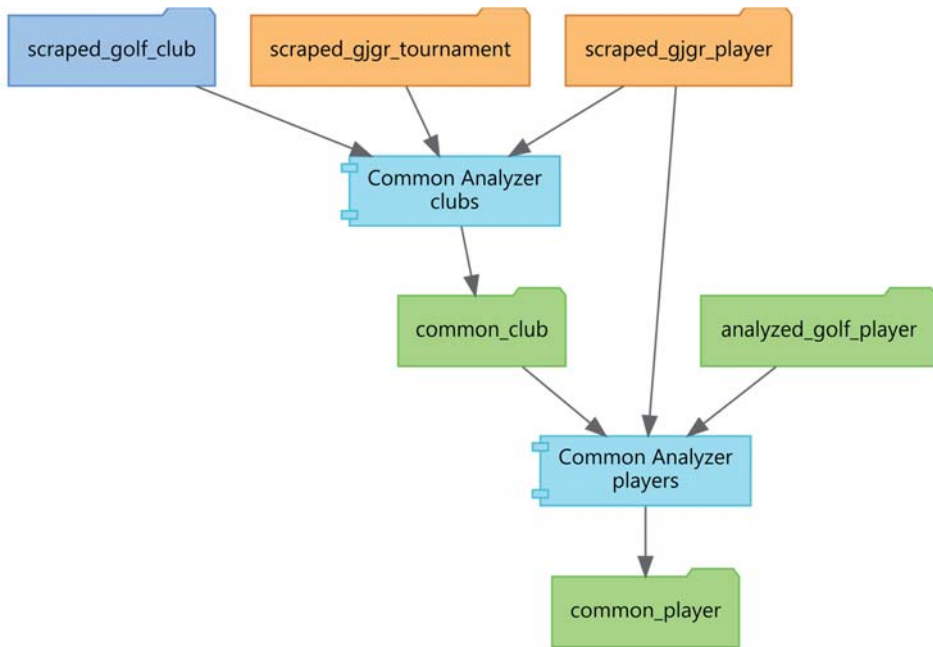


Abbildung 18: Verknüpfung von Golf.de- und GJGR-Daten im Analyzer

## Package-Struktur

Die Python-Komponente *Analyzer* basiert auf der gemeinsamen Architektur für Python-Komponenten und ist zusätzlich in den folgenden Packages und Klassen implementiert:

- `spi.analyze`
  - `golf`
    - `GolfAnalyzer`  
Eine Klasse, welche die Aufbereitung der Rohdaten aus Golf.de DB implementiert.
  - `gjgr`
    - `GjgrAnalyzer`  
Eine Klasse, welche die Aufbereitung der Rohdaten aus GJGR DB implementiert.
  - `common`
    - `CommonAnalyzer`  
Eine Klasse, welche das Zusammenführen von Entitäten aus unterschiedlichen Datenquellen organisiert.

## Befehlszeilen

Die drei Analyzer-Programme besitzen jeweils ein eigenes CLI-Skript.

### Golf.de Analyzer

Das CLI-Skript für *Golf.de Analyzer* heißt `golfanalyzer.py`.

```
usage: golfanalyzer.py [-h] [-v] [-s START_DATE] [-e END_DATE]
                        [-m {all,scorecards,players}] [-p] [-t THREADS]
                        [-d CHUNK_SIZE] [-c CONFIG_FILES]
                        [-o OPTIONS [OPTIONS ...]]
```

#### SPI Golf Analyzer

##### optional arguments:

```
-h, --help                show this help message and exit
-v, --version             print the program version and exit
-s START_DATE, --start-date START_DATE
                          Specifies the earliest date to analyze, in the format
                          yyyy-mm-dd.
-e END_DATE, --end-date END_DATE
                          Specifies the latest date to analyze, in the format
                          yyyy-mm-dd.
-m {all,scorecards,players}, --mode {all,scorecards,players}
                          The mode to operate in: all, scorecards, players.
-p, --parallel            A flag to activate parallel processing.
-t THREADS, --threads THREADS
                          The number of threads for parallel processing
-d CHUNK_SIZE, --chunk-size CHUNK_SIZE
                          The maximal number of days to process at once during
                          parallel processing.
-c CONFIG_FILES, --config-file CONFIG_FILES
                          A path to a configuration file in UTF-8 encoded INI
                          format. This argument can be used multiple times.
-o OPTIONS [OPTIONS ...], --options OPTIONS [OPTIONS ...]
                          One or more configuration options, given in the format
                          <section>.<option>=<value>.
```

## GJGR Analyzer

Das CLI-Skript für *GJGR Analyzer* heißt `gjgranalyzer.py`.

```
usage: gjgranalyzer.py [-h] [-v] [-s START_DATE] [-e END_DATE] [-p]
                        [-t THREADS] [-d CHUNK_SIZE] [-c CONFIG_FILES]
                        [-o OPTIONS [OPTIONS ...]]
```

SPI GJGR Analyzer

optional arguments:

```
-h, --help            show this help message and exit
-v, --version          print the program version and exit
-s START_DATE, --start-date START_DATE
                        Specifies the earliest date to analyze, in the format
                        yyyy-mm-dd.
-e END_DATE, --end-date END_DATE
                        Specifies the latest date to analyze, in the format
                        yyyy-mm-dd.
-p, --parallel         A flag to activate parallel processing.
-t THREADS, --threads THREADS
                        The number of threads for parallel processing
-d CHUNK_SIZE, --chunk-size CHUNK_SIZE
                        The maximal number of days to process at once during
                        parallel processing.
-c CONFIG_FILES, --config-file CONFIG_FILES
                        A path to a configuration file in UTF-8 encoded INI
                        format. This argument can be used multiple times.
-o OPTIONS [OPTIONS ...], --options OPTIONS [OPTIONS ...]
                        One or more configuration options, given in the format
                        <section>.<option>=<value>.
```

## Common Analyzer

Das CLI-Skript für *Common Analyzer* heißt `commonanalyzer.py`.

```
usage: commonanalyzer.py [-h] [-v] [-m {all,match-clubs,match-players}]
                        [-c CONFIG_FILES] [-o OPTIONS [OPTIONS ...]]
```

SPI Common Analyzer

optional arguments:

```
-h, --help            show this help message and exit
-v, --version          print the program version and exit
-m {all,match-clubs,match-players}, --mode {all,match-clubs,match-players}
                        The mode to operate in: all, match-clubs, match-
                        players.
-c CONFIG_FILES, --config-file CONFIG_FILES
                        A path to a configuration file in UTF-8 encoded INI
                        format. This argument can be used multiple times.
-o OPTIONS [OPTIONS ...], --options OPTIONS [OPTIONS ...]
                        One or more configuration options, given in the format
                        <section>.<option>=<value>.
```

## Optionen

Die Analyzer-Programme unterstützen die folgenden Konfigurationsoptionen.

### Sektion `behavior`

#### `exit_on_first_error`

Ein Schalter zum Steuern des Verhaltens im Fehlerfall. Ist der Schalter aktiv, bricht das Programm ab, sobald der erste Fehler auftritt.

Standardwert = no

#### `progress_interval`

Die Anzahl von Sekunden die zwischen Fortschrittmeldungen verstreichen dürfen.

Standardwert = 300

### Sektion `common.clubs`

#### `banned_club_ids`

Eine kommaseparierte Liste mit Club-IDs aus `golf.de`, die bei der Analyse ignoriert werden sollen. Dies ist z. B. nützlich, wenn es an einem Standort einen Verein als Club für die Spieler und eine GmbH für die Platzverwaltung gibt und beide gleich oder sehr ähnlich heißen. Dann kann die ID der GmbH angegeben werden um Verwechselungen bei der Zuordnung des Heimatclubs zu vermeiden.

Beispiel = 6660, 8976, 7766, 4585, 4583, 4588

Für die Konfiguration der MySQL-Datenbankverbindungen werden die Sektionen `database.golf`, `database.gjgr`, `database.spi` und `database.intermediate` verwendet. Der Aufbau der Sektionen ist in Konfiguration von Python-Komponenten beschrieben.

Darüber hinaus unterstützen diese Komponenten die allgemeinen Optionen der Sektion `logging`.

## Intermediate Golf DB

- Typ: Datenbank
- Technologie: MySQL
- Ansprechpartner: Tobias Kiertscher

Die *Intermediate Golf Datenbank* speichert die Ergebnisse des Analyzers und bildet die Grundlage für eine quellenübergreifende Golf-Turnieranalyse. In der gesamten Architektur stellt sie lediglich einen Zwischenschritt dar und wird nicht direkt durch einen Anwender genutzt. In Abbildung 19 sind die Tabellen mit ihren Attributen und Beziehung dargestellt. Tabellen mit dem Präfix `scraped_` sind nur zur Veranschaulichung mit aufgeführt. Sie gehören in Wirklichkeit zu den Rohdatenbanken Golf.de DB und GJGR DB.

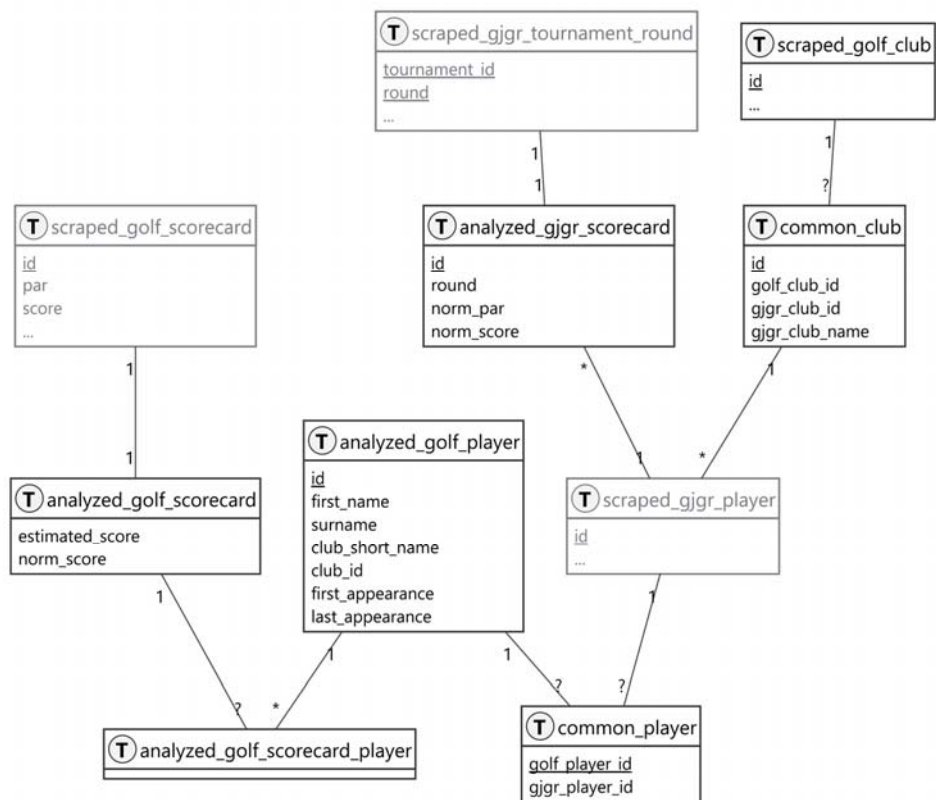


Abbildung 19: Tabellen der Intermediate Golf Datenbank

## Golf Analytics

- Typ: Script
- Technologie: SQL
- Ansprechpartner: Tobias Kiertscher

Das Script *Golf Analytics* erstellt aus den Datenbanken Golf.de DB, GJGR DB, Intermediate Golf DB und Wetter DB eine einheitliche bereinigte Datenbasis mit vorberechneten Aggregationen und Sichten für die weitere Analyse. In Abbildung 20 wird die Verarbeitungskette der Daten dargestellt. Es wird deutlich, aus welchen Roh- und Zwischendaten die Tabellen der Golf Analytics DB aufgebaut werden.

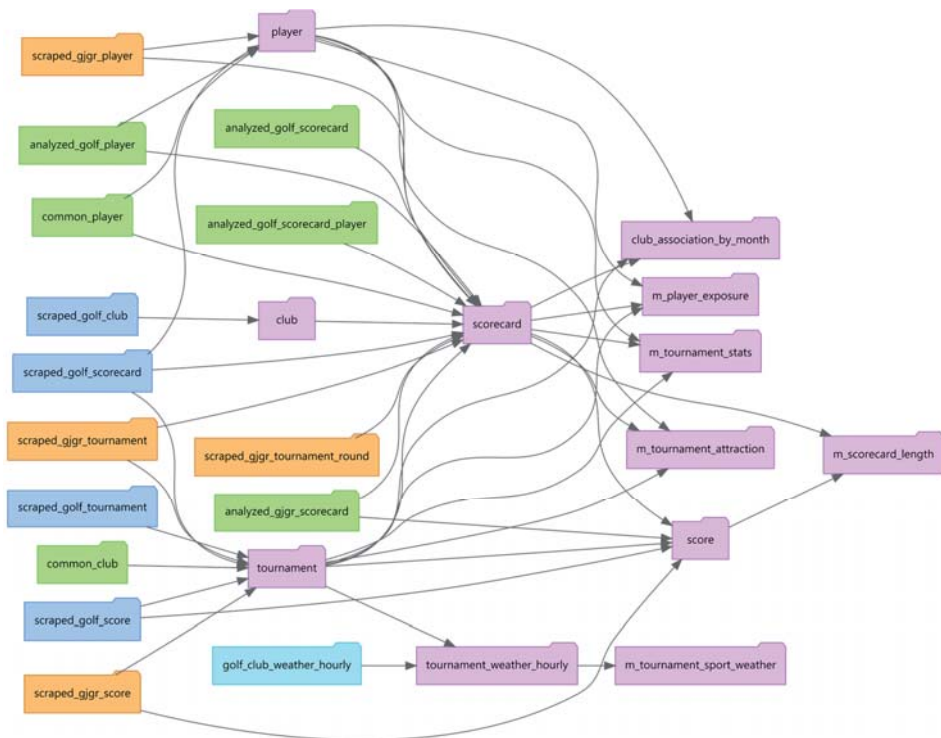


Abbildung 20: Datenfluss zum Aufbau der Golf Analytics Datenbank

## Golf Analytics DB

- Typ: Datenbank
- Technologie: MySQL
- Ansprechpartner: Tobias Kiertscher

Die *Golf Analytics DB* ist eine bereinigte und konsistente Datensammlung mit den wesentlichen Informationen aus Golf.de DB, GJGR DB und Wetter DB. Sie ermöglicht es aussagekräftige Analysen über die Golf-Turniere in Deutschland im Ganzen und den Turnierergebnissen der Golf-Jugend im Besonderen zu erstellen.

Die Datenbank umfasst u. a. die folgenden Tabellen und Sichten:

- Basisdaten
  - club
  - player
  - tournament
  - tournament\_weather\_hourly



- ☐ scorecard
- ☐ score
- Vorberechnete Aggregationen
  - ☐ m\_player\_exposure
  - ☐ m\_tournament\_attraction
  - ☐ m\_tournament\_sport\_weather
- Sichten mit zusammengeführten Daten
  - ☐ v\_club\_stat
  - ☐ v\_player
  - ☐ v\_tournament
  - ☐ v\_scorecard
  - ☐ v\_scorecard\_youth

In Abbildung 21 wird dargestellt, aus welchen Tabellen die Daten für eine Sicht kombiniert werden.

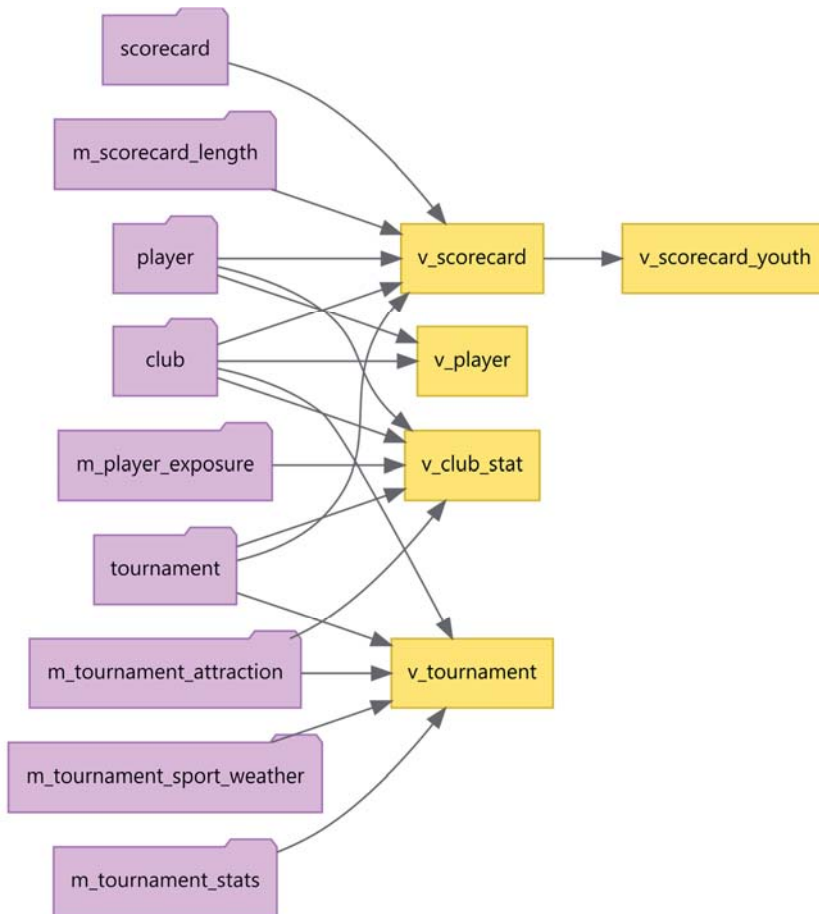


Abbildung 21: Sichten in der Golf Analytics Datenbank

## Jupyter Notebook

- Typ: Script
- Technologie: Jupyter
- Ansprechpartner: Tobias Kiertscher

Die *Jupyter Notebooks* zur Datenanalyse sind ein Werkzeug für den Analysten, um Statistiken und Analysen auf den bereinigten und konsistenten Daten der Golf Analytics DB durchzuführen.

Ein *Jupyter Notebook* erlaubt die Umformung, Filterung, Kombination, Aggregation und Visualisierung von Daten in der Programmiersprache Python z. B. unter Zuhilfenahme der Bibliothek *Pandas*.

## Aufgabenplanung

Wiederkehrende Aufgaben im SPI-System werden durch eine Aufgabenplanung gesteuert. Da alle wiederkehrenden Aufgaben durch einen Befehlszeilenaufruf auf der Linux-Shell ausgeführt werden können, bietet sich der Scheduler Cron an.

Sollen mehrere Aufgaben in einem Aufruf zusammengefasst werden, wird Prochestra verwendet. Ein einfaches Bash-Script würde diesen Zweck ähnlich gut erfüllen.

## Cron

*Cron* ist ein Systemdienst in Linux, der automatisch, entsprechend einer einfachen Zeitplanung, Befehle ausführt. Im SPI-System führt *Cron* die SPI Backend Jobs aus. Für die Datensammlung und -integration führt *Cron* Prochestra aus, welches wiederum die verschiedenen Python-Komponenten aufruft.

## Prochestra

- Typ: Programm
- Technologie: Python
- Ansprechpartner: Tobias Kiertscher

Das Programm *Prochestra* ist ein Hilfsmittel um mehrere Prozesse zu starten, die Abhängigkeiten untereinander besitzen. Endet ein Prozess mit einem Exit-Code ungleich 0, werden Prozesse die von ihm abhängen nicht ausgeführt. *Prochestra* führt alle Prozesse nacheinander aus. Welche Prozesse mit welchen Parametern und welchen Abhängigkeiten ausgeführt werden sollen, kann in einer JSON- oder YAML-Datei definiert werden.

## Befehlszeile

*Prochestra* ist als Python-Bibliothek mit Befehlszeilenwerkzeug auf PyPI veröffentlicht und kann mit `pip install prochestra` installiert werden.

## Beispiel

Eine Job-Liste für die Crawler- und Transfer-Komponenten könnte wie folgt aussehen:

```
jobs:
- id: golf-crawler
  cmd: "golfcrawler"
  args: ["-c", "/usr/etc/spi/golfcrawler.ini", "-i", "30"]
- id: golf-transfer
  cmd: "golftransfer"
  args: ["-c", "/usr/etc/spi/golftransfer.ini"]
  dependencies:
  - golf-crawler
- id: gjgr-crawler
  cmd: "gjgrcrawler"
  args: ["-c", "/usr/etc/spi/gjgrcrawler.ini"]
- id: gjgr-transfer
  cmd: "gjgrtransfer"
  args: ["-c", "/usr/etc/spi/gjgrtransfer.ini"]
  dependencies:
  - golf-transfer
  - gjgr-crawler
```

Der dazugehörige Cron-Job könnte dann so aussehen:

```
# m h dom mon dow user  command
0 18 * * * spi    prochestra -l /var/spi/logs/tasks.log ...
/usr/lib/spi/tasks.yaml 2>&1
```

## Aufbau einer Python-Komponente

Mehrere Komponenten im SPI-System sind mit Python implementiert und folgen einem einheitlichen Architekturmuster. In diesem Abschnitt soll dieses Muster kurz vorgestellt werden.

Das SPI-System setzt Python in der Version 3.6 voraus.

### Package-Struktur

Die Business-Logik der Komponente wird in einem oder mehreren Packages implementiert. Alle Packages sind Kinder des Wurzel-Packages `spi`. Funktionalitäten, die von mehr als einer Komponente genutzt wird, sind als Package ausgelagert und somit durch alle Komponenten nutzbar.

Auszug aus der Package-Struktur der Python-Komponenten im SPI-System:

- `spi`
  - `Configuration`

Eine Klasse, welche die Konfiguration für eine Komponente aus ihrer Standardkonfiguration, evtl. mehreren Konfigurationsdateien und Befehlszeilenargumenten zusammenführt und mit einer einheitlichen Schnittstelle zugänglich macht.
  - `log`
    - `Logger`

Eine Klasse, welche die Protokollierung implementiert. Die Konfiguration wird direkt aus einer `spi.Configuration`-Instanz übernommen.
  - `db`
    - `DbConnection`

Eine Klasse, welche eine Datenbankverbindung aufbaut und eine Fabrik für Session-Klassen implementiert. Die Konfiguration für die Datenbankverbindung wird direkt aus einer Sektion in einer `spi.Configuration`-Instanz übernommen. Für die Datenbankverbindung und das OR-Mapping wird *SQLAlchemy* verwendet. Als Datenbanktreiber wird die Python-Bibliothek *pymysql* genutzt.
    - `DbSessionWrapperBase`

Eine Basisklasse für Session-Klassen, die den Zugriff auf eine bestimmte Datenbank bzw. Tabellengruppe kapseln.
  - `web`
    - `HttpClient`

Eine Klasse, welche das Abfragen von HTTP-Ressourcen implementiert. Sie unterstützt das Parsen von HTML-Inhalten in einen *Beautiful Soup 4*-DOM und von JSON-Inhalten in Python-Datenstrukturen. Sie kümmert sich um das Decoding von

Textdaten und implementiert die zusätzlichen Wartepausen zwischen HTTP-Anfragen. Die Konfiguration wird direkt aus einer `spi.Configuration`-Instanz übernommen. Für die HTTP-Anfrage wird die Python-Bibliothek `requests` genutzt.

- **ScraperBase**  
Eine Basisklasse für Scraper-Klassen, die das Extrahieren von Daten aus einer Webseite implementieren. Sie implementiert eine Anzahl von Hilfsmethoden um den Umgang mit dem *Beautiful Soup 4*-DOM zu vereinfachen.
- **CrawlerBase**  
Eine Basisklasse für Crawler-Klassen, die den Abruf und die Datenextraktion von einer Website organisieren. Sie implementiert im Wesentlichen eine spezialisierte Art der Fehlerbehandlung.
- **crawler**  
Ein Package, welches alle Crawler-Komponenten enthält.
  - **golf**  
Das Package welches die Komponente Golf.de Crawler implementiert.
  - **gjgr**  
Das Package, welches die Komponente GJGR Crawler implementiert.
- **transfer** Ein Package, welches alle Transfer-Komponenten enthält.
  - **SpiDbSessionWrapper**  
Eine Klasse, die von `spi.db.DbSessionWrapperBase` erbt und den objektorientierten Zugriff auf die SPI DB implementiert.
  - **TransferBase**  
Eine Basisklasse für Transfer-Klassen, die das Übertragen von Daten aus einer Datenbank in eine andere organisieren. Im Wesentlichen implementiert diese Klasse das Erfassen und Protokollieren von Statistiken für die Übertragung.
  - **model**  
Ein Package mit den Datenklassen für die Tabellen der SPI DB.
  - **golf**  
Das Package, welches die Komponente Golf.de Transfer implementiert.
  - **gjgr**  
Das Package, welches die Komponente GJGR Transfer implementiert.
- **analyze** Ein Package, welches alle Analyzer-Komponenten enthält
  - **IntermediateDbSessionWrapper**  
Eine Klasse, die von `spi.db.DbSessionWrapperBase` erbt und den objektorientierten Zugriff auf die Intermediate Golf DB implementiert.
  - **model**  
Ein Package mit den Datenklassen für die Tabellen in Intermediate Golf DB.

Die Klassen in Python-Komponenten des SPI-Systems sind grob in zwei Gruppen aufgeteilt: Datenklassen und Klassen zur Implementierung von Algorithmen. Datenklassen besitzen nach Möglichkeit keine eigene Logik.

Klassen die einen Algorithmus implementieren, haben i. d. R. einen Konstruktor, der als erstes Argument eine Instanz von `spi.Configuration` und als zweites Argument eine Instanz von `spi.log.Logger` erwartet. Durch dieses Muster haben alle wichtigen Klassen direkten Zugriff auf die Konfiguration und können Protokollmeldungen absetzen.

*Hinweis: Die Implementierung einer eigenen Logging-Infrastruktur sollte nach Möglichkeit vermieden werden. Es wäre eine bessere Lösung, auf das Python-Modul `Logging` aus der Standardbibliothek zurückzugreifen. Es ist flexibel genug, alle erdenklichen Szenarien abzudecken. Das im Rahmen dieses Projektes eine eigene Lösung implementiert wurde, ist auf die begrenzte Erfahrung des Entwicklers zurückzuführen.*

## CLI-Skripte

Jede Python-Komponente besitzt ein CLI-Skript, mit dem es ausgeführt werden kann. Das Skript lädt die notwendigen Python-Packages und führt im Wesentlichen bei jeder Komponente die gleiche Folge von Schritten aus.

1. Befehlszeilenargumente parsen
2. Konfiguration laden
3. Protokollierung initialisieren
4. Komponente ausführen

Das Parsen der Befehlszeilenargumente übernimmt das Python-Modul `argparse` der Standardbibliothek. Für die Python-Skripte welche die CLI einer Komponente repräsentieren, existiert jeweils ein CMD-Wrapper für Windows und ein Bash-Wrapper für Linux. Der Wrapper hat bis auf die Dateiendung den gleichen Namen wie das Python-Skript.

CMD-Wrapper:

```
@ECHO OFF
SetLocal
CALL python "%~dpn0.py" %*
SET RESULT=%ERROR_LEVEL%
EXIT /B %RESULT%
```

Bash-Wrapper:

```
#!/usr/bin/env bash
script="$(basename $(realpath $0))"
exec python3 "$(dirname $(realpath $0))/${script%.*}.py" $@
```

*Hinweis: Im Rückblick erscheint es sinnvoller, alle Python-Komponenten, die in einer gemeinsamen Code-Basis implementiert sind, durch ein einzelnes CLI-Skript zu repräsentieren und für das Parsen der Befehlszeilenargumente von `argparse` auf die Python-Bibliothek `click` zu wechseln. Das würde es erlauben auch die CLI-Einsprungspunkte der Komponenten in den Packages zu implementieren, statt sie in Skripte auszulagern. Zusätzlich würde der Aufwand für die Implementierung weiterer Komponenten (z. B. experimenteller Analysen) sinken.*

## Konfiguration von Python-Komponenten

Die Konfiguration einer Komponente wird von der Klasse `spi.Configuration` implementiert. Diese verwendet das Python-Standardmodul `configparser` zum Lesen der Konfigurationsdateien. Konfigurationsdateien haben das INI-Format. Es werden die folgenden Konfigurationsebenen unterstützt:

1. Standardkonfiguration  
Eine Konfigurationsdatei mit Standardwerten für alle von der Komponente genutzten Optionen. Diese Standardkonfigurationsdatei ist als Ressource im Package der Komponente enthalten.
2. Konfigurationsdateien  
Mit dem Befehlszeilenargument `-c/--config-file` kann der Pfad zu einer zusätzlichen Konfigurationsdatei angegeben werden. Das Argument `-c/--config-file` kann mehrmals verwendet werden. Die Konfigurationen werden in der angegebenen Reihenfolge übereinandergelegt, so dass die Einträge in der zuletzt angegebenen Konfigurationsdatei die höchste Priorität erhalten.
3. Befehlszeilenooptionen  
Mit dem Befehlszeilenargument `-o/--options` können ein oder mehrere Optionen direkt angegeben werden. Dabei wird das Argument `-o/--options` nur einmal angegeben, gefolgt von einer oder mehrerer Optionen in der Form `section.option=value`.

Die folgenden Sektionen und Optionen werden komponentenübergreifend verwendet:



## Sektion logging

In der Sektion `logging` wird die Protokollierung konfiguriert. Die Sektion enthält die folgenden Optionen:

- Protokollierung mit der Standardausgabe

### `console_log_levels`

Der Detailgrad der Meldungen in der Standardausgabe.

Standardwert = `INFO,WARNING,ERROR`

### `console_log_details`

Ein Schalter zum Aktivieren von Zusatzinformationen zu Meldungen. (Dateipfade, Stack-Traces, ...)

Standardwert = `yes`

### `console_log_format`

Die Formatzeichenfolge für Meldungen.

Standardwert = `$(time) [$(level)] ${text}`

### `console_detail_format`

Die Formatzeichenfolge für Zusatzinformationen.

Standardwert = `| ${text}`

- Protokollierung in einer Datei

### `log_file`

Ein relativer oder absoluter Pfad zu der Protokolldatei. Ist der Pfad relativ, bezieht er sich auf das aktuelle Arbeitsverzeichnis. Der Pfad kann die Platzhalter `date` und `time` enthalten.

Standardwert = `golfcrawler_$(date)_$(time).log`

### `file_log_levels`

Der Detailgrad der Meldungen in der Protokolldatei.

Standardwert = `DEBUG,INFO,WARNING,ERROR`

### `file_log_details`

Ein Schalter zum Aktivieren von Zusatzinformationen zu Meldungen. (Dateipfade, Stack-Traces, ...)

Standardwert = `yes`

### `file_log_format`

Die Formatzeichenfolge für Meldungen.

Standardwert = `[$(date)T$(time)] $(level): ${text}`

### `file_detail_format`

Die Formatzeichenfolge für Zusatzinformationen.

Standardwert = `> ${text}`

- Protokollierung der Datenbankkommunikation

#### `log_sql`

Ein Schalter zur Ausgabe von SQL-Befehlen im Protokoll.

Standardwert = no

Jede Protokollmeldung besitzt ein Log-Level. Es werden die folgenden Log-Level DEBUG, INFO, WARNING und ERROR unterstützt. Der Detailgrad für die Protokollierung wird als kommaseparierte Zeichenfolge von Log-Levels angegeben, welche im Protokoll erscheinen sollen.

Zeichenfolgenformate (z. B. `file_log_format`) können Platzhalter enthalten. Unterstützt werden die Platzhalter `date`, `time`, `level` und `text`.

#### Sektion `database.*`

Jede Datenbankverbindung wird durch eine eigenständige Sektion konfiguriert. So wird z. B. die Datenbankverbindung zur Datenbank mit den GJGR-Rohdaten in der Sektion `database.gjgr` konfiguriert.

Eine Sektion zur Konfiguration einer Datenbankverbindung enthält die folgenden Optionen:

##### `host`

Der Hostname oder die IP-Adresse der Datenbankservers.

Standardwert = localhost

##### `schema`

Der Name der Datenbank.

Standardwert abhängig von der konkreten Datenbankverbindung, z. B. `gjgr` oder `spi`.

##### `user`

Der Benutzername für die Anmeldung am Datenbankserver.

Standardwert = root

##### `password`

Das Passwort für die Anmeldung am Datenbankserver.

Standardwert ist leer.

## Quellen

- [1] KIERTSCHER, TOBIAS : Aufbau der Browser-basierten Benutzerschnittstelle für das SPI-System. In: FRANZ, ROBERT U. (Hrsg.): *PreLytica – Projektbericht Band 1*. Shaker, 2019. p. 161.
- [2] NITZE, ANDRÉ : Konzeption und Implementierung einer plattformübergreifenden mobilen Applikation zur Erfassung von individuellen und gruppenbezogenen Leistungsdaten. In: FRANZ, ROBERT U. (Hrsg.): *PreLytica – Projektbericht Band 1*. Shaker, 2019. p. 181.

# Spielerakte und Datenschutz

## Personenbezogene Inhalte im SPI-System und Maßnahmen für den Datenschutz

Tobias Kiertscher

### Inhalt

|                                               |     |
|-----------------------------------------------|-----|
| Einleitung .....                              | 142 |
| Spielerakte .....                             | 142 |
| Anforderungen .....                           | 142 |
| Datenquellen .....                            | 143 |
| Inhalte .....                                 | 144 |
| Technische Umgebung .....                     | 148 |
| Personenbezogene Daten.....                   | 149 |
| Personenbezogene Daten im Wettkampfsport..... | 150 |
| Datennutzung .....                            | 151 |
| Datenschutz.....                              | 152 |
| Anforderungen .....                           | 152 |
| Maßnahmen.....                                | 152 |
| Zusammenfassung.....                          | 159 |
| Quellen.....                                  | 160 |

## Einleitung

Im Rahmen des Teilprojektes SPI (*Sport Performance Insight*) im Forschungsprojekt *PreLytics* wird ein System betrieben, welches personenbezogene Daten von Golf-Spielern und -Trainern erfasst, speichert und analysiert. Personenbezogene Daten sind durch den Gesetzgeber besonders geschützt und dürfen nur unter einhalten enger Grenzen gespeichert und verarbeitet werden.

In den folgenden Abschnitten werden zunächst die personenbezogenen Daten beschrieben, die im SPI-System gespeichert werden. Anschließend werden die gesetzlichen Anforderungen aufgelistet, die bei der Speicherung und Verarbeitung eingehalten werden müssen. Abschließend werden die konkreten organisatorischen und technischen Maßnahmen erläutert, die im Projekt zum Datenschutz ergriffen werden.

## Spielerakte

Das SPI-System ist eine Infrastruktur zum Erfassen, Verknüpfen und Auswerten von Informationen über jugendliche Leistungssportler. Beispielhaft konzentriert sich das System dabei auf den Golf-Sport.

Die Informationen über die Leistungssportler, im Folgenden Spieler genannt, und das Training werden vom System aufbereitet und in einem Web-Portal präsentiert. Dabei werden für die unterschiedlichen Rollen der Benutzer unterschiedliche Sichten erzeugt. Denn ein Trainer benötigt z. B. einen Überblick über alle von ihm betreuten Spieler, während ein Spieler nur Zugriff auf seine eigenen Daten erhalten soll.

Ein sehr großer Anteil der Daten die im SPI-System gespeichert werden, sind jeweils direkt mit einem Spieler verknüpft. Die aufbereitete Darstellung dieser Daten wird unter dem Begriff *Spielerakte* zusammengefasst.

## Anforderungen

Jede Benutzerrolle stellt ihre eigenen Anforderungen an das Web-Portal und insbesondere die Spielerakte von SPI.

Der **Spieler** möchte einen unkomplizierten Überblick über seine persönlichen Fortschritte aber auch seine Niederlagen bzw. Schwachstellen erhalten. Der **Trainer** ist zusätzlich daran interessiert, wie viel Zeit der Spieler im Training mit welchen Trainingsschwerpunkten verbringt und möchte erkennen wie sich das Training konkret auf die Entwicklung des Spielers auswirkt. Der **Koordinator** möchte neben der Leistung der Spieler auch ihr Engagement wie z. B. die Teilnahme an betreuten Trainingsgruppen einschätzen können. Für Trainer und Koordinator ist es wichtig, das Potential von jungen Spielern einschätzen zu können, um die zur Verfügung stehenden Ressourcen im Club effizient einzusetzen und vielversprechende Spieler angemessen zu fördern.

## Datenquellen

Um die Anforderungen der verschiedenen Benutzergruppen zu erfüllen, müssen mehrere Datenquellen verknüpft werden. Zunächst können die Trainer und Spieler selbst Daten in das SPI-System eingeben, um die Trainingsanwesenheit und Trainingsinhalte zu dokumentieren. Koordinator und Trainer möchten ggf. auch zu einzelnen Spielern Anmerkungen oder strategische Attribute speichern. Das kann in strukturierter Form oder im Freitext geschehen. Auch regelmäßig durch Trainer oder Spieler erhobene Messwerte können im SPI-System gespeichert werden. Alle Daten die der Golf-Club selbst in das SPI-System eingibt, werden als **interne Daten** bezeichnet. Die tatsächliche Leistung eines Spielers misst sich jedoch letztendlich an seinem Erfolg in den Turnieren. Die Turnierergebnisse im Golf-Sport werden auf frei zugänglichen Webseiten<sup>1,2</sup> veröffentlicht. Das SPI-System fragt die Turnierergebnisse mittels Screen-Scraping ab und verknüpft sie mit den im System bekannten Spielern. Daten aus solchen öffentlichen Quellen werden als **externe Daten** bezeichnet.

---

<sup>1</sup> <https://www.golf.de/>

<sup>2</sup> <https://www.gjgr.de/>

## Inhalte

Die Spielerakte umfasst die folgenden Inhalte.

### Personenstammdaten

Die Personenstammdaten sind interne Daten.

#### **Primärschlüssel / ID**

Als primäres Identifikationsmerkmal für einen Spieler wird im System eine UUID verwendet. Das ist eine 32 Stellen umfassende hexadezimale Zahl, welche üblicherweise in 5 Gruppen wie folgt dargestellt wird (vergl. [1], [2]):

XXXXXXXX-XXXX-XXXX-XXXX-XXXXXXXXXXXX

Jeden Benutzer und Spieler der im System angelegt wird, erhält als erstes eine UUID, für die garantiert ist, dass sie kein zweites Mal verwendet wird. Eine UUID ist eine der sichersten Methoden in der Softwaretechnik, um ein eindeutiges Identifikationsmerkmal zu erzeugen.

Die UUID eines Spielers hat nur im Kontext des SPI-Systems eine Bedeutung und lässt außerhalb des Systems keinerlei Rückschlüsse auf den Spieler zu.

#### **Name**

Für jeden Spieler werden der Vorname und der Nachname gespeichert. Der Vorname kann aus mehreren Namen bestehen.

#### **Geburtsdatum**

Das Geburtsdatum eines Spielers dient zum einen dazu, Verwechslungen von Spielern mit gleichen Vor- und Nachnamen zu verhindern, zum anderen ist es wichtig, um das Alter eines Spielers zu einem gegebenen Zeitpunkt bestimmen zu können. Anhand des Alters wird der Spieler im Training und in Turnieren in Altersklassen eingeordnet. Und auch für den konkreten Vergleich von Messwerten ist das Alter des Spielers von höchster Bedeutung.

#### **Geschlecht**

Das Geschlecht des Spielers wird erfasst, um ihn Trainingsgruppen und Mannschaften zuzuordnen, die nach Geschlecht getrennt aufgestellt werden.

**Email-Adresse**

Für alle Spieler, welche die Mobil-App oder das Web-Portal nutzen möchten, wird eine Email-Adresse gespeichert. Die Email-Adresse dient als Identifikationsmerkmal (Benutzername) beim Login. Bei Benutzern, denen eine Berichtsfunktion zur Verfügung steht, wird die Email-Adresse auch als Empfänger für Email-Berichte genutzt.

**Passwort**

Passwörter werden nicht im Klartext gespeichert, sondern kryptografisch abgesichert (vergl. [3]).

## Training

Die Trainingsdaten sind interne Daten.

**Gruppenzugehörigkeit**

Für jeden Spieler wird gespeichert, zu welchen Trainingsgruppen er gehört. Diese Zuordnung wird für alle Trainingspläne gespeichert. Trainingspläne werden i. d. R. nicht gelöscht, so dass eine Historie über die Trainingsgruppen eines Spielers entsteht.

Über die Gruppenzugehörigkeit können auch der betreuende Trainer und die Spieler ermittelt werden, mit denen gemeinsam trainiert wurde.

**Trainingsteilnahme**

Der Trainer erfasst für jeden Trainingstermin die Anwesenheit der Spieler. So wird für jeden Spieler gespeichert, bei welchen Trainingsterminen er anwesend war.

Bei Terminen von Gruppen denen der Spieler zugeordnet ist, wird sowohl erfasst, ob er anwesend war oder auch nicht. Bei Terminen von Gruppen denen der Spieler nicht zugeordnet ist, kann er trotzdem zusätzlich anwesend sein.

**Freies Training**

Ein Spieler kann über die Mobil-App erfassen, wann er wie lange selbstständig trainiert. Diese Informationen können dem Trainer helfen, das Verhältnis zwischen betreutem und selbstständigem Training einzuschätzen.

Üblicherweise ist der Umfang des selbstständigen Trainings eines leistungsorientierten Spielers um ein Vielfaches höher als der des betreuten Trainings.



### **Trainingsaktivität**

Sowohl für das betreute Training als auch für das freie Training wird gespeichert, wieviel Zeit der Spieler mit welchem Trainingsschwerpunkt verbracht wurde. Daraus lässt sich in der Summe der Trainingsaktivität eines Spielers eine Art Trainingsprofil erstellen. Die Trainingsschwerpunkte sind in Absprache mit dem Koordinator und den Trainern einheitlich vorgegeben und umfassen z. B. Platzspiel, Range, Putten und Fitness.

## **Turniерergebnisse**

Die Turnierergebnisse sind externe Daten.

### **Turnierteilnahme**

Für jeden Spieler im System werden die veröffentlichten Turnierergebnisse aus den oben angegebenen Datenquellen abgerufen. Mindestens wird das Datum, der Austragungsort, der Turniername und die gespielte Rundenanzahl erfasst.

Falls verfügbar werden noch weitere Informationen über das Spiel und den Platz gespeichert. Dazu gehören das Par des Spiels, die Anzahl der gespielten Löcher, der Name des Tees, die Gesamtlänge des gespielten Platzes, sowie Course Ranking und Slope.

### **Score-Karten**

Falls verfügbar werden auch die Daten aus den Score-Karten in die Spielerakte übernommen. Dazu gehören für jedes gespielte Loch die tatsächliche Schlagzahl, das Par, die Vorgabe, die Länge vom Tee zum Loch, und falls das Spiel nach den Stableford-Regeln gespielt wurde, auch die Vorgabestriche, die Brutto- und Netto-Punktzahl.

## **Messwerte**

Die Messwerte sind gemischte (interne und externe) Daten.

### **Medizinische Werte**

Aus Datenschutzgründen werden zurzeit nur wenige sportmedizinischen Daten erhoben. Es werden keine Daten erhoben und gespeichert, die Rückschlüsse auf Krankheiten oder spezielle sportliche Eignungen zulassen. Bisher werden Gewicht und Größe erfasst.

Geplant waren weitere Größen wie z. B.:

Körperdepotfett (%), Vitalkapazität (ml), Expiratorisches Lungenvolumen (ml), Ruhe-Herzfrequenz (pro min) und Ruhe-Blutdruck systolisch/diastolisch (mmHG)

### **Training (automatisch)**

Aus den Daten zur Trainingsanwesenheit und dem freien Training werden monatlich und jährlich Messwerte gebildet. Für die Trainingsanwesenheit wird jeweils die Anwesenheitsquote in 0 bis 100 Prozent gespeichert. Für das freie Training wird die erfasste Stundenanzahl gespeichert.

### **Golf Abschlag**

Mit einem Messgerät für den Golf-Abschlag (z. B. TrackMan<sup>3</sup>) können eine Vielzahl von Parametern während des Abschlages aufgezeichnet werden. Bisher werden die folgenden Parameter jeweils für Eisen und Driver erfasst:

Ball Speed, Carry, Club Speed, Side Test und Smash

### **Golf Kurzspiel**

In Absprache mit den Projektpartnern wurde ein Test-Parkour für Kurzspiel entwickelt, den die Spieler regelmäßig absolvieren, und deren Ergebnisse als Messwerte festgehalten werden.

Ein Durchgang besteht aus jeweils 10 Schlägen aus einer festgelegten Entfernung zum Loch. Gemessen wird der Abstand zwischen Loch und Ball, nachdem dieser zur Ruhe gekommen ist. Die Abstände werden als Messreihe erfasst. Für den finalen Messwert wird der Durchschnitt gebildet.

- Bunker: 18 Metern zum Loch
- Chip: 9 Meter zum Loch
- Pitch: 28 Meter zum Loch

### **Golf Putten**

So wie für das Kurzspiel, wurde auch für das Putten ein einheitlicher Versuch definiert, um die Fähigkeiten der Spieler in Lochnähe zu messen. Es werden jeweils 10 Schläge aus unterschiedlichen Entfernungen geschlagen. Dabei wird

---

<sup>3</sup> TackMan Golf: <https://trackmangolf.com/>

gezählt wie oft dabei eingelocht wird, bzw. wie häufig der Ball in einem Mindestabstand zum Loch zur Ruhe kommt.

- 12 Meter: Ball muss in weniger als 80 Zentimetern Abstand zum Loch liegen bleiben oder ins Loch fallen
- 6 Meter: Ball muss ins Loch fallen
- 2,5 Meter: Ball muss ins Loch fallen

Als Messwert wird die Trefferquote zwischen 0 und 100 Prozent gespeichert.

### **Golf Turnierergebnisse (automatisch)**

Aus den externen Daten der Golf Turniere werden automatisch Messwerte gebildet. So wird für jede offizielle Score-Karte ein Messwert mit der normierten Schlagzahl gespeichert. Ebenso wird für jede offiziell auf einer Score-Karte als Startvoraussetzung angegebene Stammvorgabe (Handicap) als Messwert gespeichert.

Zusätzlich wird an jedem Monatsende ein Messwert gespeichert, der die Anzahl der in diesem Monat gespielten Turniere angibt. Auch an jedem Jahresende wird die Turnieranzahl für das ganze Jahr als Messwert gespeichert.

## Technische Umgebung

Alle Benutzerdaten im SPI-System, also auch die Daten der Spielerakten, werden in einer MySQL-Datenbank gespeichert (vergl. Abbildung 1). Diese läuft auf einem Linux-Server in der DMZ (*Demilitarisierte Zone*) des Rechenzentrums der Technischen Hochschule Brandenburg in einem Docker-Container<sup>4</sup>, der durch Watchtower<sup>5</sup> automatisch aktualisiert wird (vergl. [4]). Der Zugriff auf die Datenbank erfolgt durch die PHP-Anwendung des SPI-Backends, welche auf dem gleichen Server in einem eigenen Docker-Container läuft.

Im Backend werden die Daten mit SQL aus der Datenbank abgefragt und durch das OR-Framework Doctrine<sup>6</sup> von relationalen Datensätzen in PHP-Objekte übersetzt.

---

<sup>4</sup> Docker: <https://www.docker.com/>

<sup>5</sup> Watchtower: <https://github.com/v2tec/watchtower>

<sup>6</sup> Doctrine ORM: <https://www.doctrine-project.org>

Diese werden dann in das JSON-Format serialisiert und über HTTP an das Web-Portal im Browser des Benutzers oder an die Browser-Komponente in der Mobil-App ausgeliefert. Neue interne Daten gelangen über den umgekehrten Weg zum Backend und zur Datenbank.

Die HTTP-Verbindung verläuft innerhalb der DMZ unverschlüsselt bis zum Reverse-Proxy des Rechenzentrums. Der Reverse-Proxy terminiert die mit TLS verschlüsselte Verbindung durch das Internet zum Browser oder zur Mobil-App des Benutzers.

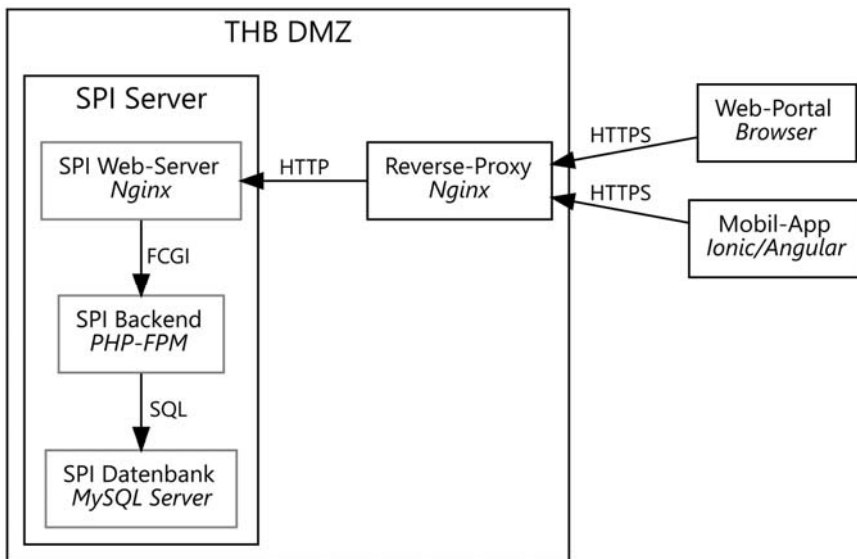


Abbildung 1: Transport der Daten im SPI-System

## Personenbezogene Daten

Die Spielerakte enthält personenbezogene Daten, nach DSGVO Art. 4 Abs. 1 Nr. 1., über den Spieler, welche einen besonderen Schutz erfordern (vergl. [5]). Daten sind personenbezogen, wenn mit ihrer Hilfe eine natürliche Person direkt oder indirekt identifiziert werden kann. Das schließt sowohl Kennungen, Namen, Standortdaten und physische oder elektronische Adressen, als auch Merkmale des Aussehens und der sozialen bzw. kulturellen Identität ein. In der aktuellen Rechtsprechung wird

der Begriff der *personenbezogenen Daten* recht weit ausgelegt, so dass z. B. auch Anwesenheitszeiten als personenbezogen eingestuft werden.

Die Spielerakte enthält **keine** personenbezogenen Daten, die als *sensibel* eingestuft werden (vergl. [6]).

Die folgenden Daten aus der Spielerakte werden als personenbezogen betrachtet.

- Primärschlüssel
  - UUID des Benutzers (vergl. [2])
  - E-Mail-Adresse
- Primäre Merkmale
  - Vorname(n)
  - Nachname
  - Geburtsdatum
- Sekundäre Merkmale
  - Messwerte über Größe und Gewicht
- Verhaltensmuster
  - Anwesenheit an Trainingsterminen
  - Selbstständige Trainingszeiten
  - Turnierteilnahmen aus externen Datenquellen

## Personenbezogene Daten im Wettkampfsport

Einige personenbezogenen Daten, die direkt mit Wettkampfsport im Bezug stehen, werden in der Sportgemeinschaft trotz ihrer Schutzwürdigkeit einvernehmlich oder durch individuelle Zustimmung veröffentlicht. Das sind z. B. die Wettkampfteilnahmen und -ergebnisse. I. d. R. werden dazu die Personen mit Vor-, Nachname und Heimatverein/-Club genannt. In manchen Fällen wird auch das Geburtsjahr angegeben. Eine zuverlässige Identifizierung aller Wettkampfteilnehmer ist dabei, z. B. wegen häufig vorkommenden Namen, nicht immer möglich – eine Identifizierung einzelner Personen jedoch sehr wohl.

Das SPI-System selbst stützt sich auf diese veröffentlichten (externen) Daten und nimmt eine Zuordnung der externen zu den internen personenbezogenen Daten vor. Durch die Zuordnung entsteht in der Spielerakte ein angereicherter Datensatz, der eine zuverlässige Identifizierung der Spieler erlaubt und weit mehr personenbezogene Daten umfasst als die ursprünglich veröffentlichten Daten.

## Datennutzung

Die Daten in der Spielerakte werden von unterschiedlichen Personenkreisen genutzt. Im Forschungsprojekt PreLytica werden die Daten zur Erforschung von Datenintegrations- und Datenanalyseverfahren genutzt. Die Benutzer von Web-Portal und Mobile-App (Spieler, Trainer, Koordinatoren), nutzen die Daten für die Organisation des Trainingsbetriebs und als Informationsquelle über die sportliche Entwicklung der Spieler.

## Forschungsprojekt

Im Rahmen des Forschungsprojektes werden die Daten der Spielerakten für die folgenden Zwecke genutzt:

- Experimente, Entwicklung von Analyseprozessen und Algorithmen
- Verknüpfung mit externen Datenquellen
- Entwicklung von Visualisierungen
- Entwicklung von Präskriptionen
- Demonstrationen und Präsentationen

## Mandanten

Die Mandanten des SPI-Systems (Golf-Clubs mit Koordinatoren, Trainern und Spielern) nutzen die Daten der Spielerakten für die folgenden Zwecke:

- Saison-Planung und Trainingsorganisation
- Inhaltliche Trainingsgestaltung
- Wettkampfstrategie / Mannschaftsaufstellung
- Einschätzung des Entwicklungspotentials einzelner Spieler
- Individuelle sportliche Entwicklung

# Datenschutz

Die Spielerakte enthält eine Reihe personenbezogener Daten, die nur entsprechend den gesetzlichen Bestimmungen erfasst, gespeichert und verarbeitet werden dürfen.

## Anforderungen

Die im Mai 2018 in Kraft getretenen Datenschutz Grundverordnung (DSGVO) legt die gesetzlichen Rahmenbedingungen für die Verarbeitung personenbezogener Daten fest. Nach Art. 5 Abs. 1 DSGVO müssen die folgenden Grundsätze beachtet werden:

- Rechtmäßigkeit, Verarbeitung nach Treu und Glauben, Transparenz
- Zweckbindung
- Datenminimierung
- Richtigkeit
- Integrität, Vertraulichkeit, Schutz vor zufälligem Verlust, zufälliger Zerstörung oder Veränderung
- Zeitliche Speicherbegrenzung

## Maßnahmen

Den durch die DSGVO entstehenden Anforderungen an die Datenverarbeitung im Projekt wird durch die im Folgenden beschriebenen Maßnahmen begegnet.

Wenn in diesem Abschnitt von *Daten* die Rede ist, sind damit solche Daten gemeint, die personenbezogene Daten der Benutzer des SPI-Systems enthalten und damit unter den Schutz der DSGVO fallen. Ausgenommen sind alle übrigen Forschungsdaten die keinen Personenbezug besitzen.

## Organisatorisch

Die *Rechtmäßigkeit* der Erfassung und Speicherung der Daten von Benutzern des SPI-Systems (Koordinatoren, Trainer und Spieler) im Projekt *PreLyrica* ist durch zwei Schritte gegeben:

1. Eine *schriftliche Einverständniserklärung der Benutzer* bzw. der Eltern der Benutzer, welche die Erfassung, Speicherung und Verarbeitung der Daten im Golf-Club und im

Rahmen des Forschungsprojektes *PreLytica* an der Technischen Hochschule Brandenburg unter Einhaltung der gesetzlichen Vorschriften erlaubt

2. Ein leicht zugänglicher *Datenschutzhinweis* mit Ansprechpartner und Verwendungszweck in den Benutzeroberflächen (Web-Portal und Mobil-App)

Die Daten dürfen daher nicht außerhalb des Projektes verwendet werden. Das bedeutet, dass die Daten nur an Personen weitergegeben werden dürfen, die als Mitarbeiter für das Projekt angestellt oder durch organisatorische Maßnahmen klar erkennbar in das Projekt eingebunden sind. Insbesondere eine Weitergabe der Daten an andere Institutionen oder Unternehmen ist ohne eine erneute schriftliche Einverständniserklärung der Benutzer nicht möglich.

Der *Zweck* der Datenverarbeitung ist durch die Forschungsziele des Projektes definiert. Die Daten dürfen nicht zu anderen Zwecken genutzt werden. Z. B. dürfen die Daten nicht im Rahmen anderer Projekte an der THB genutzt oder weiterverarbeitet werden.

Die *Minimierung* von Daten wird dadurch eingehalten, dass nur die unbedingt zur Identifizierung und sportlichen Einordnung erforderlichen personenbezogenen Daten erhoben und erfasst werden: Name, Geburtsdatum und Geschlecht. Darüber hinaus wird für aktive Nutzer des Web-Portals oder der Mobil-App eine E-Mail-Adresse gespeichert.

Die Minimierung von personenbezogenen Daten im weiteren Sinne (Aufenthaltsdaten, personenbezogene Merkmale) ist im Forschungsprojekt nur bedingt gegeben, da die Forschungsziele nur erreicht werden können, wenn Informationen erhoben werden, deren konkrete Aufgabe während der Erfassung noch nicht klar ist. Durch den Umstand, dass die Erforschung der Daten und ihrer Bedeutung selbst Teil des Forschungsziels ist (Stichwort *Data Mining*), kann das Forschungsprojekt noch keine Minimierung aller in der Spielerakte erfassten Daten durchführen. Als Teilergebnis des Projektes wird jedoch ggf. deutlich werden, welche Merkmale der Spieler für die Steuerung des Trainings und Steigerung des sportlichen Erfolgs bedeutungsvoller sind als andere. Darauf aufbauend kann in Nachfolgeprojekten oder in aus dem Projekt hervorgehenden Software-Produkten eine angemessene Minimierung der Daten vorgenommen werden.

Die Benutzer haben jederzeit die Möglichkeit die *Richtigkeit* der zu ihrer Person gespeicherten Daten zu überprüfen. Insbesondere können sie die über sich gespeich-



cherten Daten *einsehen*, *korrigieren* oder *löschen* lassen. Einen Teil dieser Möglichkeiten ist direkt über das Web-Portal des SPI-Systems gegeben. Sollte das Web-Portal die Rechte des Benutzers während der Projektlaufzeit nicht vollständig erfüllen, kann den Rechten durch eine kurzfristige Anfrage an die Projektmitarbeiter entsprochen werden.

Die Benutzer wurden mit der von ihnen unterzeichneten Einverständniserklärung schriftlich über ihre Rechte informiert und haben darüber hinaus die Möglichkeit im Web-Portal und in der Mobil-App eine in der Benutzeroberfläche leicht erreichbare Informationsseite aufzurufen, die sie ebenfalls über ihre Rechte und die Maßnahmen zum Schutz ihrer personenbezogenen Daten aufklärt.

Die Mitarbeiter im Projekt *PreLytica* wurden darauf hingewiesen, dass die personenbezogenen Daten der Benutzer nur nach den gesetzlichen Vorschriften erfolgen kann. Darüber hinaus wurden weitere technische Maßnahmen ergriffen, um den Schutz der Daten zu gewährleisten. Diese Maßnahmen werden in den folgenden Abschnitten näher erläutert.

## Verschlüsselung

Daten die auf Endgeräten im Besitz der Projektmitarbeiter gespeichert werden, sind verschlüsselt. Dazu kommen abhängig vom Betriebssystem unterschiedliche Werkzeuge zum Einsatz.

- Windows PC oder Notebook: BitLocker<sup>7</sup>
- Linux PC oder Notebook: LUKS<sup>8</sup>
- Datensicherungen auf mobilen Datenträgern: VeraCrypt<sup>9</sup>

Daten die auf den Servern im Rechenzentrum der THB gespeichert sind, werden nicht verschlüsselt.

Der Datenaustausch zwischen den Endgeräten der Benutzer (Web-Portal oder Mobil-App) erfolgt über HTTPS. Die HTTPS-Verbindungen werden durch den Reverse-

---

<sup>7</sup> BitLocker: [https://docs.microsoft.com/de-de/previous-versions/mt404675\(v=vs.85\)](https://docs.microsoft.com/de-de/previous-versions/mt404675(v=vs.85))

<sup>8</sup> LUKS: <https://gitlab.com/cryptsetup/cryptsetup/>

<sup>9</sup> VeraCrypt: <https://www.veracrypt.fr>

Proxy des Rechenzentrums als Brücke zwischen dem offenen Internet und der DMZ aufgebaut. Dabei kommen die folgenden Parameter zum Einsatz<sup>10</sup>.

- Zertifikat: RSA 4096 Bits mit SHA-265 und begrenzter Gültigkeit bis Ende des Projektes
- HTTPS
  - Stärkste: TLS 1.2 mit ECDHE-RSA (SHA-384) mit AES-256 in GCM (256 Bits)
  - Schwächste: TLS 1.0 mit DHE-RSA (SHA-1) mit 3DES (EDE) in CBC (112 Bits)

Der Datenaustausch zwischen den in das SPI-System eingebundenen Servern in der DMZ des Rechenzentrums der THB findet zurzeit unverschlüsselt statt.

Die Passwörter der Benutzer werden nicht im Klartext gespeichert, sondern zusammen mit einem Salt und mit der kryptografischen Einwegfunktion SHA-256 verrechnet (vergl. [3], [7], [8]).

## Zugriffskontrolle

Der Zugriff auf die Daten aus dem Internet ist nur über das SPI-Backend möglich. Das Web-Portal und die Mobil-App benutzen das SPI-Backend, um die Daten der Spielerakte abzurufen. Das SPI-Backend erwartet als Teil jeder Anfrage ein JSON-Web-Token (JWT), welches den Benutzer authentifiziert (vergl. [9], [10]). Dieses wird über eine Login-Anfrage an das SPI-Backend erzeugt. Dabei werden E-Mail-Adresse und Passwort verschlüsselt übertragen.

Jeder Benutzer besitzt eine Rolle im SPI-System, die ihm entsprechende Rechte für die Einsicht und die Veränderung von Daten im Backend einräumt. Im Web-Portal und in der Mobil-App stehen für den jeweils angemeldeten Benutzer nur jene Steuerelemente und Navigationsmittel zur Verfügung, die ihn zu Ansichten führen für die er eine Berechtigung besitzt. Diese Schutzmaßnahmen lassen sich jedoch technisch leicht umgehen. Deshalb werden die Zugriffsrechte zusätzlich bei jedem einzelnen Zugriff im Backend anhand des JWT geprüft. Bei der Ermittlung der Zugriffsrechte werden die Rolle des Benutzers, die ID des Benutzers und der Mandant, dem der Benutzer angehört, berücksichtigt.

---

<sup>10</sup> Die technischen Parameter der HTTPS-Verschlüsselung wurden mit dem Test der SSL-Labs von Qualys ermittelt. <https://www.ssllabs.com/ssltest/>

Projektmitarbeiter, die von ihrem Endgerät aus Zugriff auf Server in der DMZ haben, und einen Benutzer-Account auf dem SPI-Datenbank-Server besitzen, können am Backend vorbei direkt auf die SPI-Datenbank zugreifen. Diesen erweiterten Zugriff haben jedoch nur maximal zwei Mitarbeiter, die auch für die Administration und Softwareentwicklung zuständig sind.

Die übrigen Projektmitarbeiter erhalten keinen direkten Zugriff auf die SPI-Datenbank. Stattdessen erhalten sie für die Forschung Zugriff auf eine pseudonymisierte Kopie der SPI-Datenbank. Das Vorgehen bei der Pseudonymisierung wird in einem der folgenden Abschnitte beschrieben.

Die Administratoren des Rechenzentrums orientieren sich bei ihrer Arbeit an den Empfehlungen des IT-Grundschutzkatalogs vom Bundesamt für Sicherheit und Informationstechnik (BSI) (vergl. [4], [11], [12]). Das Rechenzentrum ist jedoch nicht zertifiziert und begegnet mit ihren Dienstleistungen primär den alltäglichen Anforderungen von Lehre und Forschung.

Die Endgeräte der Projektmitarbeiter sind durch ein angemessen sicheres Passwort geschützt. Verschlüsselte Sicherungskopien der Projektdaten werden nur in abschließbaren Räumen aufbewahrt.

## Sicherer Austausch

Wenn Daten zwischen Projektmitarbeitern ausgetauscht werden, geschieht das i. d. R. nur mit pseudonymisierten Daten. Dabei kommen Mobile Speicher (USB-Sticks) zum Einsatz, auf denen die Daten nur kurzzeitig für den Transport gespeichert werden. Der Datenträger wird nach Abschluss des Datentransfers wieder gelöscht. Dabei wird keine standardisierte Löschmethode angewendet, sondern der Datenträger wird mit Bordmitteln des Betriebssystems vollständig formatiert. Anschließend ist es nur noch mit hohem Aufwand und einer geringen Wahrscheinlichkeit möglich, die Daten wiederherzustellen. Die Übertragung der Daten geschieht nach dem Vier-Augen-Prinzip.

Personenbezogene Daten werden im Rahmen des Projektes nicht per E-Mail versendet und auch nicht in Cloud-Speichern abgelegt.

## Pseudonymisierung

Für die Forschungstätigkeit ist in den meisten Fällen unerheblich, welche natürliche Person hinter einer Spielerakte steht. Deshalb werden die Datensätze vor dem Austausch unter den Projektmitarbeitern pseudonymisiert um zu verhindern, dass Projektmitarbeiter sich detaillierte Informationen zu bekannten Spielern einprägen.

Daten sind dann pseudonymisiert, wenn die primären Identifikationsmerkmale so verändert werden, dass die Person mit angemessenem Aufwand nicht mehr identifiziert werden kann. I. d. R. wird bei der Pseudonymisierung jedoch eine Zuordnungstabelle erstellt, mit deren Hilfe die pseudonymisierten Datensätze wieder den ursprünglichen personenbezogenen Daten zugeordnet werden können.

Im Projekt *PreLytica* werden die personenbezogenen Daten mit den folgenden Schritten pseudonymisiert:

1. Ersetzen des Vor- und Nachnamens durch zufällige Namen aus einer Datenbank häufiger deutscher Namen
2. Ersetzen der E-Mail-Adresse durch generische Adressen, für die kein Postfach existiert
3. Verschieben des Geburtsdatums um eine zufällige Anzahl von bis zu 90 Tagen

Eine Zuordnung der pseudonymisierten Spielerakte zu dem ursprünglichen Spieler ist jedoch weiterhin über die systeminterne ID und die Verknüpfungen mit Score-Karten möglich. Ein Projektmitarbeiter müsste absichtlich Schritte unternehmen, um die Pseudonymisierung rückgängig zu machen, was nicht seinen Forschungszielen entspricht.

## Anonymisierung

Es gibt mehrere Situationen, in denen ein Benutzer aus dem SPI-System gelöscht werden muss. Das ist zunächst der Fall, wenn ein Benutzer sein durch die DSGVO zugestandenes Recht wahrnimmt, seine Daten zu löschen. Auch wenn ein Spieler den Golf-Club verlässt und damit die Einverständniserklärung zur Speicherung und Verarbeitung der Daten erlischt. Und zuletzt müssen am Ende der Projektlaufzeit alle personenbezogenen Daten aus dem SPI-System gelöscht werden.

Damit in diesen Fällen aber die Forschungsergebnisse bewahrt und weiter genutzt werden können, wird die Spielerakte eines Spielers, dessen personenbezogene Daten gelöscht werden sollen, anonymisiert. Bei der Anonymisierung werden die Daten der Spielerakte so verändert, dass eine Zuordnung zur ursprünglichen Person nur noch mit unangemessenem Aufwand möglich ist.

Bei der Anonymisierung muss also zwischen dem für die Bearbeitung der Forschungsfragen notwendigen Informationsgehalt und dem Datenschutz, konkret dem für eine De-Anonymisierung erforderlichen Aufwand, abgewogen werden.

Die Anonymisierung einer Spielerakte erfolgt mit den folgenden Schritten:

1. Ersetzen des Vor- und Nachnamens durch die Zeichenkette ANONYM
2. Löschen der E-Mail-Adresse
3. Verschieben des Geburtsdatums um eine zufällige Anzahl von bis zu 90 Tagen
4. Aggregation der Daten zu Trainingsanwesenheit und Trainingsaktivität in zeitlichen Fenstern von 3 Monaten
5. Aggregation der Golf-Turnielergebnisse in zeitlichen Fenstern von 3 Monaten
6. Löschung der Zuordnung zu Trainingsgruppen
7. Löschung der Trainingsteilnahmen und Trainingsaktivität
8. Löschung der Golf-Turnierteilnahmen und Score-Karten
9. Löschung von als *sensibel* markierten Messwerten  
(z. B. konkreten Schlagzahlen, denn diese können auf eine Score-Karte verweisen und damit zur De-Anonymisierung beitragen)
10. Ersetzen der bisherigen systeminternen ID durch eine zufällige neue ID

Bei der Löschung und Anonymisierung von Datensätzen ist zu beachten, dass weiterhin unveränderte Daten in Sicherungskopien existieren. Deshalb müssen die Sicherungskopien in einem regelmäßigen Zyklus ersetzt und gelöscht werden.

## Protokollierung

Die Zugriffe auf das SPI-Backend werden protokolliert. Dadurch wird die Betriebssicherheit des Systems erhöht, weil es möglich wird, Versuche von unbefugtem Zugriff auf die Daten zu erkennen und wenn nötig Maßnahmen zum Schutz der Daten zu ergreifen. Insbesondere werden Login-Anfragen zum Erzeugen eines JWT und Passwort-Reset-Anfragen beobachtet. In den Protokollen erfasste IP-Adressen von Endgeräten werden anonymisiert.

## Zusammenfassung

Die Spielerakte des SPI-Systems umfasst nicht nur personenbezogene Daten zu Golf-Spielern wie Name, Geburtsdatum und E-Mail-Adresse, sondern verknüpft diese mit einem detaillierten Profil von Trainingsaktivitäten, Golf-Turnielergebnissen und Messwerten. Solche Daten sind vom Gesetzgeber besonders geschützt. Im Projekt *PreLytics* werden die gesetzlichen Anforderungen durch eine Reihe von organisatorischen und technischen Maßnahmen erfüllt. Dazu gehören schriftliche Einverständniserklärungen der Trainer und Eltern der Spieler ebenso wie die Nutzung von sicheren Kommunikationsprotokollen, das Verschlüsseln von Datenträgern und eine sorgfältig implementierte Zugriffskontrolle. Ein bewusster Umgang mit den personenbezogenen Daten durch die Projektmitarbeiter wird durch Pseudonymisierung und Anonymisierung ergänzt, um das Potential für eine Verletzung des Datenschutzes auszuschließen.

## Quellen

- [1] LEACH, P. ; MEALLING, M. ; SALZ, R. : *RFC 4122: A Universally Unique Identifier (UUID) URN Namespace*. IETF, 2005. — URL <https://tools.ietf.org/html/rfc4122>
- [2] *Universally Unique Identifier*. Wikipedia, 2018. — URL [https://de.wikipedia.org/wiki/Universally\\_Unique\\_Identifier](https://de.wikipedia.org/wiki/Universally_Unique_Identifier)
- [3] SCHNEIER, BRUCE : *Secrets & lies - IT-Sicherheit in einer vernetzten Welt*. Köln : Dpunkt-Verlag, 2004 – Maßstab — ISBN 978-3-898-64302-3
- [4] *Demilitarisierte Zone (Informatik)*. Wikipedia, 2018. — URL [https://de.wikipedia.org/wiki/Demilitarisierte\\_Zone\\_%28Informatik%29](https://de.wikipedia.org/wiki/Demilitarisierte_Zone_%28Informatik%29)
- [5] *Was sind personenbezogene Daten?* Europäische Kommission, 2018. — URL [https://ec.europa.eu/info/law/law-topic/data-protection/reform/what-personal-data\\_de](https://ec.europa.eu/info/law/law-topic/data-protection/reform/what-personal-data_de)
- [6] *Sensible Daten*. Europäische Kommission, 2018. — URL [https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/legal-grounds-processing-data/sensitive-data/what-personal-data-considered-sensitive\\_de](https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/legal-grounds-processing-data/sensitive-data/what-personal-data-considered-sensitive_de)
- [7] *Salt (Kryptologie)*. Wikipedia, 2018. — URL [https://de.wikipedia.org/wiki/Salt\\_%28Kryptologie%29](https://de.wikipedia.org/wiki/Salt_%28Kryptologie%29)
- [8] EASTLAKE, D. 3RD ; HANSEN, T. ; LABS, MOTOROLA ; LABS, AT&T : *US Secure Hash Algorithms (SHA and HMAC-SHA)*. IETF, 2006. — URL <https://tools.ietf.org/html/rfc4634>
- [9] *JSON Web Token*. Auth0 Inc., 2018. — URL <https://jwt.io/>
- [10] JONES, M. ; BRADLEY, J. ; N., SAKIMURA ; MICROSOFT ; IDENTITY, PING : *RFC 7519: JSON Web Token (JWT)*. IETF, 2015. — URL <https://tools.ietf.org/html/rfc7519>
- [11] *IT-Grundschutz*. BSI, 2018. — URL <https://www.bsi.bund.de/grundschutz>
- [12] *BSI-Standard 200-2: IT-Grundschutz-Methodik*. BSI. — URL [https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Grundschutz/Kompodium/standard\\_200\\_2.pdf](https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Grundschutz/Kompodium/standard_200_2.pdf)

# Web-Portal

## Aufbau der Browser-basierten Benutzerschnittstelle für das SPI-System

Tobias Kiertscher

### Inhalt

|                             |     |
|-----------------------------|-----|
| Einleitung .....            | 162 |
| Aufbau .....                | 163 |
| Login und Passwörter .....  | 164 |
| Dashboards .....            | 165 |
| Benutzereinstellungen ..... | 168 |
| Administration .....        | 170 |
| Training .....              | 170 |
| Messwerte .....             | 173 |
| Spielerakte .....           | 176 |
| Ausblick .....              | 180 |



## Einleitung

Das Web-Portal des SPI-Systems spielt eine zentrale Rolle bei der Erfassung, Pflege und Auswertung der Daten im Teilprojekt *SPI* des Forschungsprojekts *PreLyrica*. Gemeinsam mit der Mobile-App stellt es die primäre Schnittstelle zu den Benutzern des Systems dar. Aus Sicht des Forschungsprojektes ist das Erfassen von qualitativ hochwertigen Daten aus der Domäne des Benutzers von Interesse. Deshalb ist es wichtig, die Aktualität und Dichte der Daten zu visualisieren, damit der Nutzer leicht erkennen kann, ob die Daten ausreichend gut gepflegt sind oder mehr Aufwand und Sorgfalt fordern. Die Motivation des Benutzers Daten in das System einzugeben wird gesteigert, wenn er im Gegenzug Auswertungsmöglichkeiten für die Daten erhält und ein Erkenntnisgewinn unterstützt wird. Dazu werden die vom Benutzer gepflegten *internen* Daten mit Daten aus *externen* Datenquellen verknüpft. Zusätzlichen Antrieb bietet die Darstellung der Datendichte im Vergleich mit anderen Nutzern, was eine Art Wettbewerbseffekt auslöst und damit zur Datenqualität beitragen kann.

Das Web-Portal bietet deshalb einerseits eine komfortable Möglichkeit, die domänenspezifischen Daten der Benutzer zu pflegen, und andererseits eine Palette an Übersichten, Visualisierungen und interaktiven Analysemöglichkeiten für die Auswertung der internen und externen Daten.

Benutzer aller Rollen (Administrator, Koordinator, Trainer, Spieler) können sich am Web-Portal anmelden. Abhängig von der jeweiligen Rolle werden jedoch unterschiedliche Sichten und Eingabemöglichkeiten angezeigt. Nur Benutzer mit der Administratorrolle haben die Möglichkeit, mandantenübergreifend Daten abzurufen. Koordinatoren und Trainer haben nahezu gleichberechtigte Möglichkeiten Daten abzurufen und zu pflegen, erhalten jedoch entsprechend ihres Aufgabengebiets unterschiedliche Sichten. Der Zugriff für Spieler begrenzt sich bisher auf eine aussagekräftige Darstellung ihrer eigenen Daten.

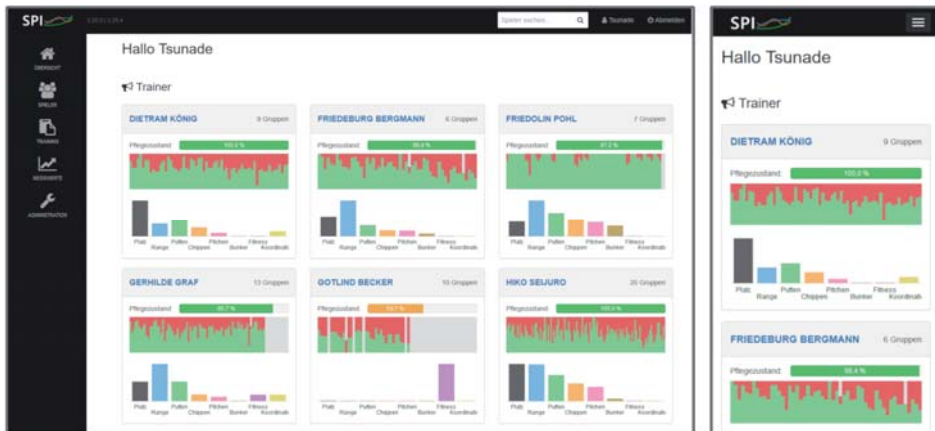


Abbildung 1: Web-Portal auf großem (links) und kleinem Bildschirm (rechts)

Das Web-Portal ist als *Single Page Application* umgesetzt und speichert selbst keine Daten. Es kommuniziert mittels asynchronen JavaScript-Aufrufen (AJAX) mit der REST-Schnittstelle des SPI-Backends. Das Layout des Web-Portals ist *responsive* umgesetzt, was bedeutet, dass sich das Layout und die Steuerelemente an unterschiedliche Bildschirmgrößen anpassen können (vergl. Abbildung 1).

## Aufbau

Das Web-Portal, im Projekt auch manchmal als "Frontend Web" bezeichnet, kann grob in die folgenden Bereiche unterteilt werden – dabei sind die Bereiche in der Navigationsstruktur nicht immer klar abgegrenzt, um eine möglichst intuitive und komfortable Bedienung zu ermöglichen.

- Login und Passwortverwaltung
- Dashboards (Startseite nach der Anmeldung)
- Benutzereinstellungen
- Administration
- Training
- Messwerte
- Spielerakte

## Login und Passwörter

Der Login erfolgt mit einer E-Mail-Adresse und einem Passwort. Wurde für einen Benutzer eine E-Mail-Adresse eingepflegt, hat er dadurch auch die Möglichkeit sich am Web-Portal anzumelden. Hat er noch kein Passwort vergeben, oder kann er sich an sein Passwort nicht mehr erinnern, so kann er auf der Login-Seite über einen Link ein Formular für das Zurücksetzen des Passwortes aufrufen. Dort gibt er seine E-Mail-Adresse an und fordert damit eine E-Mail mit einem Link für das Neusetzen des Passwortes an. Hat er die E-Mail empfangen, kann er über den enthaltenen Link ein Formular für das Neusetzen des Passwortes aufrufen. Die Sicherheit des Verfahrens wird dabei durch ein Token gewährleistet, welches mit der E-Mail verschickt wird und nur eine begrenzte Anzahl von Stunden gültig ist.

Passwörter müssen einige Mindestanforderungen erfüllen, damit sie vom SPI-System akzeptiert werden. Sie müssen mindestens 8 Zeichen lang sein, große und kleine Buchstaben enthalten und mindestens eine Zahl. Bei der Eingabe eines neuen Passwortes wird die Stärke durch einen farbigen Fortschrittsbalken visualisiert (vergl. Abbildung 2).

**Passwort zurücksetzen**

.....

gut

.....

Das neue Passwort muss mindestens 8 Zeichen lang sein, kleine Buchstaben, große Buchstaben und Zahlen enthalten.

☒ Zurücksetzen

**Abbildung 2: Das interaktive Formular für das Neusetzen eines Passwortes**

Wird eine URL aufgerufen, welche direkt zu einer Unterseite des Web-Portals führt, der Benutzer ist jedoch noch nicht angemeldet, wird er auf die Login-Seite umgeleitet und erst nach einer erfolgreichen Anmeldung weiter auf die Zielseite.

Es ist unter Umständen möglich, dass ein Benutzer durch direktes Eingeben einer URL eine Seite im Web-Portal aufruft, die nicht für seine Rolle gestaltet ist. Das hat

jedoch **nicht** zur Folge, dass der Benutzer Daten abrufen kann, für die er keine Rechte besitzt. Denn das Backend prüft anhand des übermittelten Tokens zur Authentifizierung alle Anfragen erneut auf Zulässigkeit. In dem beschriebenen Szenario würde der Benutzer zwar evtl. die Seite aufrufen können, sie würde jedoch keine Daten enthalten.

## Dashboards

Wenn sich ein Benutzer im Web-Portal anmeldet, gelangt er als erstes auf sein persönliches Dashboard. Die Inhalte des Dashboards sind abhängig von seiner Rolle.

**Hallo Sasuke**

 System

| NUTZERDATEN                    |     |
|--------------------------------|-----|
| Offene Passwort-Reset-Anfragen | 0   |
| Trainingsschwerpunkte          | 8   |
| Messwerttypen                  | 29  |
| Vereine                        | 2   |
| Benutzer Gesamt                | 402 |
| Administratoren                | 3   |
| Koordinatoren                  | 6   |
| Trainer                        | 11  |
| Aktive Spieler                 | 334 |
| Anonymisierte Spieler          | 40  |
| Trainingspläne                 | 7   |
| Trainingsgruppen               | 416 |

| SERVER              |              |
|---------------------|--------------|
| SPI Backend Version | 3.25.4       |
| PHP-Version         | 7.2.3        |
| Gateway Interface   | CGI/1.1      |
| Server Software     | nginx/1.15.1 |
| Administrator Email | Unknown      |

| ORM CACHE   |            |
|-------------|------------|
| Typ         | ArrayCache |
| Online seit | 49 Jahre   |
| Treffer     | 0          |
| Fehlschläge | 48         |

 Vereine

| GOLF- UND LAND-CLUB BERLIN-WANNSEE E.V. |       |
|-----------------------------------------|-------|
| Benutzer Gesamt                         | 336   |
| Koordinatoren                           | 6     |
| Trainer                                 | 10    |
| Spieler                                 | 320   |
| Trainingspläne                          | 7     |
| Trainingsgruppen                        | 416   |
| Trainingsteilnahmen                     | 29657 |
| Freie Trainings                         | 487   |

| GOLFCLUB DRESDEN ULLERSDORF E.V. |    |
|----------------------------------|----|
| Benutzer Gesamt                  | 15 |
| Koordinatoren                    | 0  |
| Trainer                          | 1  |
| Spieler                          | 14 |
| Trainingspläne                   | 0  |
| Trainingsgruppen                 | 0  |
| Trainingsteilnahmen              | 0  |
| Freie Trainings                  | 0  |

Abbildung 3: Dashboard eines Administrators

Ein Administrator erhält eine statistische Zusammenfassung der Daten in der SPI-Datenbank und eine Übersicht über die Mandanten (vergl. Abbildung 3).



Abbildung 4: Dashboard eines Koordinators

Ein Koordinator erhält eine Übersicht über die Trainer im Club, jeweils mit dem Pflegezustand der Trainingsaktivität und den Trainingsschwerpunkten (vergl. Abbildung 4). Zusätzlich sieht er eine Liste mit den Turnierteilnahmen der Spieler in der jüngeren Vergangenheit.

Hallo Hiko

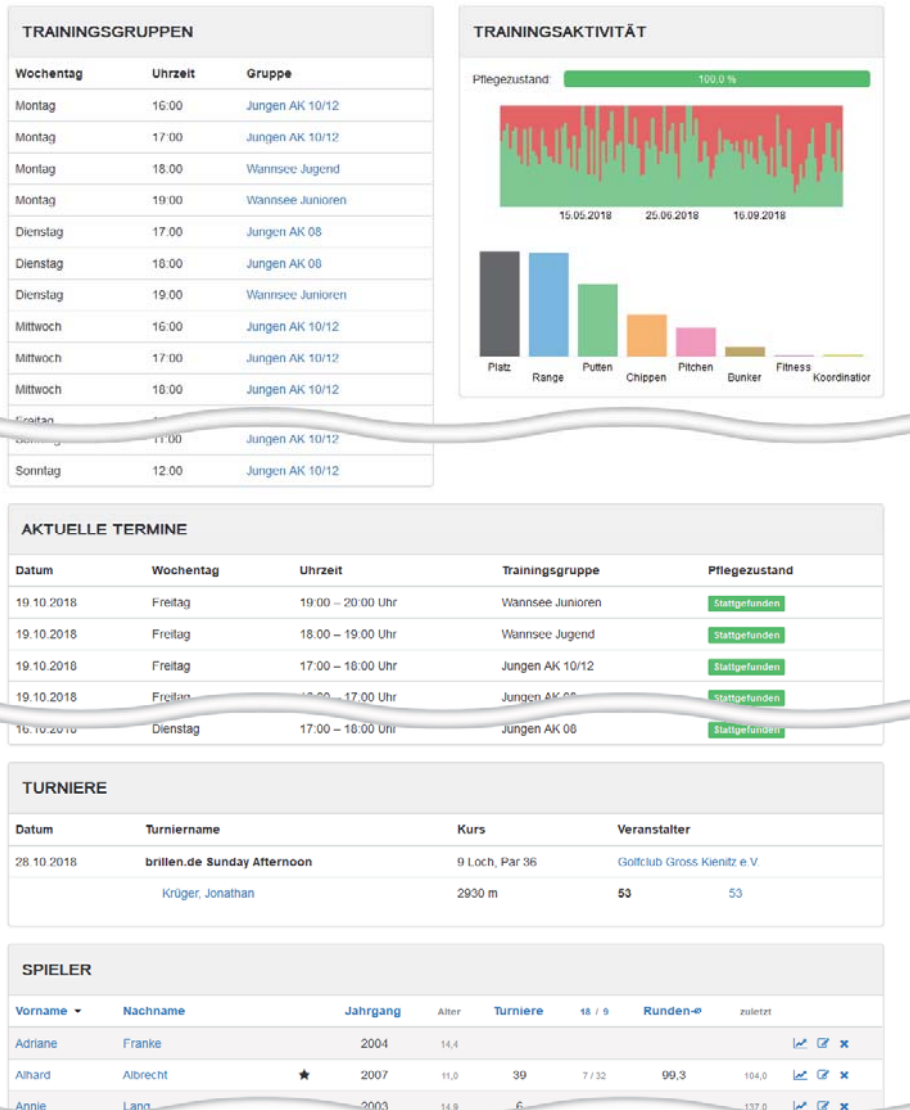


Abbildung 5: Dashboard eines Trainers

Ein Trainer erhält eine Darstellung des Pflegezustands in seinen Trainingsgruppen (vergl. Abbildung 5). Er hat eine Liste mit den aktuellen Trainingsterminen und eine Liste aller Spieler, die einer seiner Trainingsgruppen zugeordnet sind. Zusätzlich

sieht auch er eine Liste mit den kürzlich erzielten Turnierergebnissen der Spieler, die er betreut.

Hallo Naruto

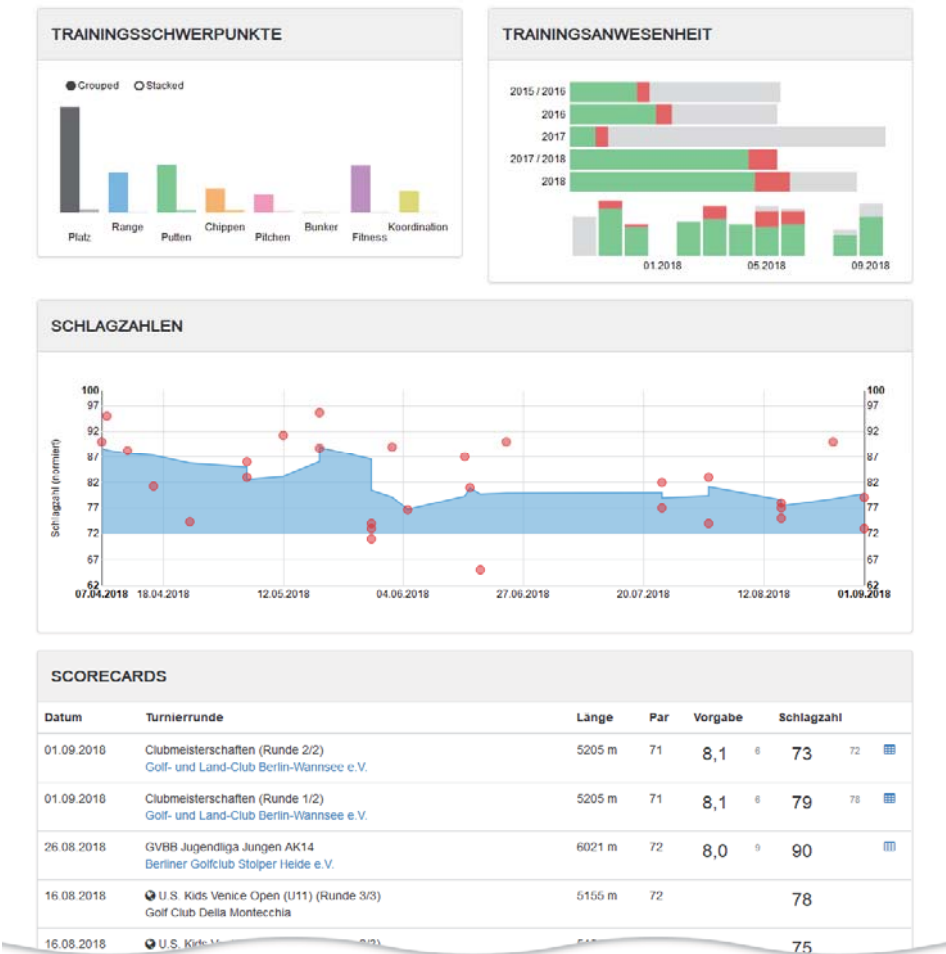


Abbildung 6: Dashboard eines Spielers

Ein Spieler erhält eine Visualisierung seiner Trainingsaktivität und seiner Anwesenheit beim Training (vergl. Abbildung 6). Er erhält eine grafische Darstellung seiner Turnierteilnahmen mit der zeitlichen Entwicklung seiner Schlagzahl und eine Liste seiner zuletzt gespielten Score-Karten.

## Benutzereinstellungen

Nach einem erfolgreichen Login kann der Benutzer eine Seite mit persönlichen Einstellungen aufrufen (vergl. Abbildung 7). Auf dieser Seite kann er sein Passwort ändern und die E-Mail-Berichte konfigurieren. Er kann festlegen ob er einen Bericht täglich und/oder wöchentlich erhalten möchte. Und er kann festlegen, an welchem Wochentag der Wochenbericht und zu welcher Uhrzeit die Berichte versendet werden sollen.

**Tsunade Jiraiya**  
manager@test

### PASSWORT ÄNDERN

**Bisheriges Passwort**

**Neues Passwort**

**Nochmal neues Passwort**

Das neue Passwort muss mindestens 8 Zeichen lang sein, kleine Buchstaben, große Buchstaben und Zahlen enthalten.

[Ändern](#)

### BERICHTSVORSCHAU

[Tagesbericht](#) [Wochenbericht](#)

### EINSTELLUNGEN

**Tagesbericht versenden**  
Wenn diese Option aktiv ist, wird täglich ein Bericht mit Ereignissen, Erkenntnissen und Erinnerungen versendet.

☒ ☐

**Zeitpunkt für Tagesbericht**  
Die Stunde in welcher der Tagesbericht versendet wird.

8

**Minimale Turnierhistorie im Tagesbericht**  
Die Anzahl der vergangenen Tage, die auf jeden Fall in der Turnierhistorie berücksichtigt werden sollen.

3

**Wochenbericht versenden**  
Wenn diese Option aktiv ist, wird wöchentlich eine Zusammenfassung der letzten sieben Tage versendet.

☒ ☐

**Tag für Wochenbericht**  
Der Tag an dem der Wochenbericht versendet wird.

**Berichte nur bei Training**  
Berichte nur versenden, wenn ein Trainingsplan aktiv und kein trainingsfreier Tag ist.

☒ ☐

[Speichern](#) [Standardwerte](#) [Zurücksetzen](#)

Abbildung 7: Benutzereinstellungen



## Administration

Der Administrationsbereich bietet unterschiedliche Möglichkeiten, je nachdem ob ein Administrator, ein Koordinator oder ein Trainer angemeldet ist.

Trainer haben die Möglichkeit weitere Trainer anzulegen und die Trainingspläne zu verwalten. Koordinatoren haben darüber hinaus die Möglichkeit weitere Koordinatoren anzulegen (vergl. Abbildung 8). Administratoren können alles, was die Trainer und Koordinatoren können, und das für alle Mandanten. Zusätzlich können sie weitere Mandanten anlegen und die Konfigurationsdaten wie Trainingsschwerpunkte und Messwerttypen verwalten.

Golf- und Land-Club Berlin-Wannsee e.V.

 Benutzer

| KOORDINATOREN                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |          |                                                                                   |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|-----------------------------------------------------------------------------------|
| Name                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | Email    |                                                                                   |
|  Development Test                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | dev@test |  |
|                                                                                                                                                                                                                                                                                                                                                                                                                             |          |                                                                                   |

sich Trainingsgruppen in einem wöchentlichen Rhythmus treffen. Der Trainingsplan hat deshalb ein Start- und ein Enddatum und enthält eine Anzahl von Trainingsgruppen. Jede Trainingsgruppe besteht aus einem Trainer und mehreren zugeordneten Spielern. Die Termine der Trainingsgruppen ergeben sich aus einem Wochentag und einer Start- und Enduhrzeit. Unterbrechungen im Trainingsplan (z. B. Feiertage oder Ferien) können durch trainingsfreie Zeiträume definiert werden. Diese besitzen lediglich ein Start- und Enddatum und müssen sich innerhalb des Zeitraums des Trainingsplans befinden. Jede Trainingsgruppe hat nur genau einen Trainer, aber ein Trainer kann mehrere Trainingsgruppen betreuen. Auch Spieler können mehreren Trainingsgruppen zugeordnet werden. Der Trainingsplan kann in einem komfortablen Editor bearbeitet werden (vergl. Abbildung 9).

**Sommertrainingsplan 2017**

Gültig von 18.04.2017 bis 21.10.2017

| Montag                      | Dienstag                  | Mittwoch                     | Donnerstag                  | Freitag                   | Samstag                        | Sonntag                     |
|-----------------------------|---------------------------|------------------------------|-----------------------------|---------------------------|--------------------------------|-----------------------------|
| Mädchen AK 12 16:00-17:00   | Mädchen AK 14 16:00-17:30 | Mädchen AK 16/18 16:00-18:00 | Jungen AK 16/18 15:00-16:00 | Mädchen AK 16 16:00-17:30 | Wannsee Kids 10:00-11:00       | Jungen AK 08/10 10:00-11:00 |
| Jungen AK 14 16:00-17:00    | Jungen AK 16 16:00-17:30  | Mädchen AK 12 16:00-18:00    | Mädchen AK 14 16:00-18:00   | Mädchen AK 12 16:00-17:30 | Jungen AK 08/10 11:00-12:00    | Jungen AK 12 11:00-13:00    |
| Jungen AK 08/10 16:00-17:00 | Mädchen AK 12 17:00-18:00 | Jungen AK 16/18 16:00-17:30  | Mädchen AK 12 16:00-17:00   | Jungen AK 16 16:00-17:00  | Wannsee Kids 12:00-13:00       | Wannsee Kids 13:00-14:00    |
| Mädchen AK 12 17:00-18:00   | Mädchen AK 14 17:30-18:00 | Jungen AK 14 16:00-17:00     | Jungen AK 18 16:00-17:00    | Jungen AK 14 16:00-17:00  | Jungen AK 12 Platz 16:00-17:00 |                             |
| Jungen AK 12 18:00-19:00    |                           | Jungen AK 12 18:30-19:30     |                             | Mädchen AK 12 17:30-19:00 |                                |                             |
| Athletik 19:00-20:00        |                           | Athletik 18:30-20:00         |                             | Jungen AK 14 18:00-19:00  |                                |                             |

Alle auf-zuklappen Trainingsfreie Tage Plan wiederherstellen Plan herunterladen Plan speichern

**Trainings**

Neues Training

- Jugendtraining
- Mo | Mädchen AK 12
- Mo | Jungen AK 14
- Mo | Mädchen AK 08/10
- Mo | Mädchen AK 12
- Mo | Jungen AK 14
- Mo | Jungen AK 08/10
- Mo | Athletik
- Mo | Jungen AK 14
- Mo | Athletik
- Mo | Jungen AK 14
- Mo | Athletik
- Di | Mädchen AK 14
- Di | Jungen AK 16
- Di | Mädchen AK 12

**Mädchen AK 12**

Bezeichnung: Mädchen AK 12

Trainer: Burga Schutz

Zeit: 16:00 – 17:00 Uhr

Teilnehmer

- Wolff, Carolin
- Richter, Rudolfina

**Teilnehmer**

Neuer Spieler

Spüler suchen...

Jahrgang von 1990 bis 2020

- Albrecht, Alhard
- Albrecht, Doraline
- Albrecht, Editha
- Albrecht, Heige
- Albrecht, Lilly

Abbildung 9: Editor für das Bearbeiten der Trainingspläne

Durch den Trainingsplan können für jeden Trainer seine individuellen Trainingstermine berechnet und eine Benutzeroberfläche für das Erfassen der Anwesenheit und der Trainingsschwerpunkte zur Verfügung gestellt werden (vergl. Abbildung 10).

## Jungen AK 16/18

Augusta Maier Mittwochs, 16:00 - 17:30 Uhr

12.07.2017

**VERANSTALTUNG**

Nicht erfasst **Stattgefunden** Ausgefallen Speichern Zurücksetzen

**SPIELER**

|          |       |  |  |  |
|----------|-------|--|--|--|
| Ditmar   | Wolff |  |  |  |
| Herger   | Klein |  |  |  |
| Siegmund | Stein |  |  |  |

Weitere Spieler hinzufügen

**Legende**

- Teilgenommen
- Gefehlt
- Nicht erfasst

**FOKUS**

|         |  |  |        |
|---------|--|--|--------|
| Bunker  |  |  |        |
| Chippen |  |  | 30 min |
| Putten  |  |  |        |
| Range   |  |  | 30 min |

Abbildung 10: Eingabemaske für die Anwesenheit und die Trainingsschwerpunkte

## Mädchen AK 16/18

Messwerte erfassen

Gottfriede Schreiber Mittwochs, 16:00 - 18:00 Uhr

**FOKUS**

**ANWESENHEIT**

**TERMINE**

|            |            |            |
|------------|------------|------------|
| 19.04.2017 | 26.04.2017 | 03.05.2017 |
| 10.05.2017 | 17.05.2017 | 24.05.2017 |
| 12.07.2017 | 06.09.2017 | 13.09.2017 |
| 20.09.2017 | 27.09.2017 | 04.10.2017 |
| 11.10.2017 | 18.10.2017 |            |

**SPIELER**

| Vorname    | Nachname  | Jahrgang | Alter | Turniere | 18 / 9 | Runden | zuletzt |
|------------|-----------|----------|-------|----------|--------|--------|---------|
| Elisabeth  | Haas      | ★        | 1999  | 19,5     |        |        | 74,4    |
| Friedlinda | Schmitt   | ★        | 2001  | 16,9     | 22 / 0 | 80,8   | 80,6    |
| Liesgret   | Winkler   | ★        | 2000  | 18,7     | 8 / 0  | 81,5   | 81,2    |
| Ulla       | Schreiber | ★        | 1999  | 19,4     | 1      |        | 73,6    |

Abbildung 11: Übersicht einer Trainingsgruppe

Aus den erfassten Daten wird für jede Trainingsgruppe eine Übersicht erzeugt, die den Trainer über Anwesenheitsverlauf, Trainingsschwerpunkte und Pflegezustand der Termine informiert (vergl. Abbildung 11). Zusätzlich erhält er über diese Seite leichten Zugriff auf die Spieler der Gruppe und auf eine angepasste Messwerterfassung, welche die Spieler der Gruppe als Vorauswahl verwendet.

## Messwerte

Um die Entwicklung der Spieler zu dokumentieren, können Messwerte für jeden Spieler aufgenommen werden. Die Messwerttypen werden durch den Administrator mandantenübergreifend vorgegeben, um eine Vergleichbarkeit zwischen den Mandanten zu ermöglichen (vergl. Administration). Messwerte können jederzeit erfasst werden. Für jeden Messwerttyp ist aber nur ein Messwert pro Tag möglich. Wird ein Messwert für einen Tag mehrfach erfasst, wird nur der zuletzt erfasste aufbewahrt. Es wird davon ausgegangen, dass sich die Messwerte eines Spielers innerhalb eines Tages nicht wesentlich verändern. Falls der Durchschnitt von mehreren Messwerten an einem Tag verwendet werden soll, kann ein Messwerttyp als Messreihe konfiguriert werden.

Messwerte können über zwei unterschiedliche Wege aufgenommen werden. Zum einen können Messwerte für einen einzelnen Spieler erfasst werden (vergl. Abbildung 12). Zum anderen gibt es eine tabellarische Eingabemaske um Messwerte für Spielergruppen zu erfassen. Messreihen können in der tabellarischen Maske nicht erfasst werden. Die tabellarische Maske wird in zwei Schritten aufgebaut. Zunächst werden die Spieler ausgewählt, für die Messwerte erfasst werden sollen. Anschließend werden die Messwerttypen ausgewählt, die erfasst werden sollen. Daraus wird dann dynamisch eine Tabelle zur Eingabe der Messwerte erzeugt (vergl. Abbildung 13). Wird die tabellarische Messwerterfassung aus einer Trainingsgruppe heraus geöffnet, entfällt die Spielerauswahl (vergl. Training).

## MESSWERTE ERFASSEN



Friedrich Thalheim

Datum

05.07.2018

**Golf***Kurzspiel*

Bunker 18m (in m)

2,3

5,8

1,2

0

4,8

4,5

2,6

0,0

0,0

0,0

Putten 2,5m

4

von 10

Putten 6m

2

von 10

*Abschlag Driver*

Driver Ball Speed

135

mph

Driver Carry

175

m

Driver Smash

0,0

*Abschlag Eisen 6***Medizinisch**

Gewicht

0,0

kg

Größe

0,0

cm

**Hinweis:** Es müssen nicht alle Felder ausgefüllt werden. Bereits erfasste Messwerte für das selbe Datum werden überschrieben.

Abbrechen

Verwerfen

Messwerte speichern

Abbildung 12: Messwerte für einen einzelnen Spieler erfassen

## Messwerterfassung

### Messwerteingabe

**UMGEBUNG**

Datum

Hinweis: Geben Sie das Datum ein, an dem die Werte gemessen wurden. Das Feld enthält standardmäßig das heutige Datum. Korrigieren Sie es bitte, wenn Sie Werte nachtragen, die nicht heute gemessen wurden.

**MESSWERTE**
Ausrichtung umschalten

| Spieler            | Driver Ball Speed                | Driver Carry                     | Driver Club Speed                | Driver Side Test                 | Driver Smash                      | Status                                                                              |
|--------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|-----------------------------------|-------------------------------------------------------------------------------------|
| Albrecht, Doraline | <input type="text" value="108"/> | <input type="text" value="152"/> | <input type="text" value="121"/> | <input type="text" value="8,5"/> | <input type="text" value="0,7"/>  |  |
| Albrecht, Heige    | <input type="text" value="90"/>  | <input type="text" value="140"/> | <input type="text" value="101"/> | <input type="text" value="4,7"/> | <input type="text" value="0,84"/> |  |
| Albrecht, Lilly    | <input type="text" value="114"/> | <input type="text" value="197"/> | <input type="text" value="133"/> | <input type="text" value="5,3"/> | <input type="text"/>              |  |

Zurücksetzen
Messwerte speichern

Abbildung 13: Tabellarische Messwerterfassung für eine Spielergruppe

Wurden ausreichend Messwerte über einen gewissen Zeitraum erfasst, können sie in einem interaktiven Diagramm analysiert und verglichen werden (vergl. Abbildung 14). Das Diagramm ist ein Korrelationsdiagramm und seine Achsen können flexibel angepasst werden. Für die vertikale Achse kann ein beliebiger Messwerttyp ausgewählt werden. Für die horizontale Achse können das Alter des Spielers zum Zeitpunkt der Messwerterhebung, der Zeitpunkt der Messwerterhebung oder ein weiterer Messwerttyp ausgewählt werden. Die Farbe der Punkte im Diagramm kann durch Geschlecht des Spielers, Altersklasse oder Spielerstatus bestimmt werden. Die Messwerte eines einzelnen Spielers können hervorgehoben werden, um die Entwicklung eines Spielers im Kontext der anderen Spieler zu beurteilen. Sind sehr viele Messwerte vorhanden, kann das Diagramm auf eine zufällige Untermenge der Messwerte reduziert werden.

## Messwerteanalyse



Abbildung 14: Interaktives Korrelationsdiagramm zur Messwertanalyse

## Spielerakte

Für jeden Spieler wird eine umfangreiche Übersichtsseite erzeugt, die möglichst alle Informationen über den Spieler übersichtlich darstellt. Diese Seite wird *Spielerakte* genannt. Sie beginnt mit einer Zusammenfassung im Kopfbereich mit einer Übersicht über verschiedene Größen, welche den Spieler charakterisieren und einer Liste der letzten 5 Score-Karten (vergl. Abbildung 15). Darunter enthält sie eine Visualisierung der Trainingsaktivität (vergl. Abbildung 16) und eine Übersicht über die letzten Messwerte und deren Alter (vergl. Abbildung 17). Den Abschluss bildet eine Darstellung der gespielten Turniere u. a. mit einer Liste aller Score-Karten (vergl. Abbildung 18).

## Naruto Uzumaki

Messwerte erfassen

Bearbeiten

Entfernen

## ZUSAMMENFASSUNG

|                                                       |       |                |
|-------------------------------------------------------|-------|----------------|
| Alter / Geburtsdatum                                  | 11,9  | 09.12.2006     |
| Trainingssteilnahme<br>im aktuellen Trainingsplan     | 84%   | 58 von 90      |
| Trainingsaktivität<br>selbstständig / betreut         | 3%    | 3.0 h / 64.0 h |
| Turnierrunden<br>12 Monate / Saison 18er / Saison 9er | 29    | 24 / 5         |
| Gespielte Distanz<br>12 Monate                        | 140,2 | km             |
| Stammvorgabe<br>Letzte Turnierrunde                   | 8,1   |                |
| Rundendurchschnitt<br>Saison / Letzte 5               | 81,7  | 79,0           |

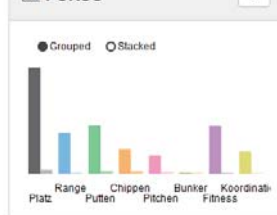
## LETZTE TURNIERRUNDEN

| Datum      | Turnierrunde                                                           | Par | Schlagzahl |
|------------|------------------------------------------------------------------------|-----|------------|
| 01.09.2018 | Clubmeisterschaften (R 2/2)<br>Golf- und Land-Club Berlin-Wannsee e.V. | 71  | 73 72      |
| 01.09.2018 | Clubmeisterschaften (R 1/2)<br>Golf- und Land-Club Berlin-Wannsee e.V. | 71  | 79 70      |
| 26.08.2018 | CVBB Jugendliga Jungen AK14<br>Berliner Golfclub Stolper Heide e.V.    | 72  | 80         |
| 16.08.2018 | U.S. Kids Venice Open (U11) (R 3/3)<br>Golf Club Delta Montecchia      | 72  | 78         |
| 16.08.2018 | U.S. Kids Venice Open (U11) (R 2/3)<br>Golf Club Delta Montecchia      | 72  | 75         |

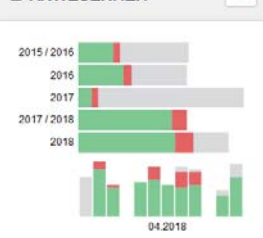
Abbildung 15: Zusammenfassung mit den letzten 5 Score-Karten

## Training

## FOKUS



## ANWESENHEIT



## TERMINE

|            |        |        |        |        |        |
|------------|--------|--------|--------|--------|--------|
| ✓ Anwesend | 12.04. | 16.04. | 17.04. | 18.04. | 19.04. |
|            | 19.04. | 23.04. | 24.04. | 25.04. | 26.04. |
|            | 26.04. | 30.04. | 01.05. | 02.05. | 03.05. |
|            | 14.05. | 16.05. | 17.05. | 17.05. | 22.05. |
|            | 23.05. | 24.05. | 24.05. | 06.06. | 07.06. |
|            | 07.06. | 11.06. | 12.06. | 13.06. | 18.06. |
|            | 20.06. | 21.06. | 21.06. | 25.06. | 26.06. |
|            | 13.09. | 17.09. | 18.09. | 19.09. | 20.09. |
| x Gefehit  | 15.05. | 28.05. | 29.05. | 30.05. | 31.05. |
|            | 31.05. | 04.06. | 05.06. | 14.06. | 14.06. |
|            | 19.06. |        |        |        |        |

Abbildung 16: Übersicht über die Trainingsaktivität



 Messwerte

| ÜBERSICHT              |               |                                                                                              |        |                    |                    |     |  |
|------------------------|---------------|----------------------------------------------------------------------------------------------|--------|--------------------|--------------------|---------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
| Bezeichnung            | Erste Messung | Letzte Messung                                                                               | Anzahl | Durchschnitt       | Letzter Wert       |                                                                                       |                                                                                     |
| <b>Golf</b>            |               |                                                                                              |        |                    |                    |                                                                                       |                                                                                     |
| <i>Abschlag Driver</i> |               |                                                                                              |        |                    |                    |                                                                                       |                                                                                     |
| Driver Ball Speed      | 11.04.2018    | 09.10.2018  | 3      | 114 mph / 184 km/h | 113 mph / 182 km/h |    |                                                                                     |
| Driver Carry           | 11.04.2018    | 09.10.2018  | 3      | 153,30 m           | 154,40 m           |    |                                                                                     |
| Driver Club Speed      | 11.04.2018    | 09.10.2018  | 3      | 77 mph / 125 km/h  | 77 mph / 124 km/h  |    |                                                                                     |
| Pitcher 2,0m           | 17.09.2018    | 17.09.2018  | 1      | 0,40               | 0,40               |    |                                                                                     |
| Putten 6m              | 17.09.2018    | 17.09.2018  | 1      | 0,20               | 0,20               |    |                                                                                     |
| <i>Turnier</i>         |               |                                                                                              |        |                    |                    |                                                                                       |                                                                                     |
| Schlagzahl             | 26.05.2015    | 02.09.2018                                                                                   | 113    | 93,14              | 73,01              |    |                                                                                     |
| Stammvorgabe           | 02.04.2016    | 02.09.2018                                                                                   | 85     | 18,00              | 8,10               |    |                                                                                     |
| Turniere (Jahr)        | 31.12.2015    | 31.12.2018                                                                                   | 4      | 28,25              | 29,00              |    |                                                                                     |
| Turniere (Monat)       | 31.05.2015    | 30.09.2018                                                                                   | 23     | 4,91               | 2,00               |    |                                                                                     |
| <b>Medizinisch</b>     |               |                                                                                              |        |                    |                    |                                                                                       |                                                                                     |
| Gewicht                | 11.04.2018    | 18.09.2018  | 4      | 49,95 kg           | 52,00 kg           |    |                                                                                     |
| Größe                  | 11.04.2018    | 18.09.2018  | 4      | 151,50 cm          | 153,00 cm          |    |                                                                                     |
| <b>Training</b>        |               |                                                                                              |        |                    |                    |                                                                                       |                                                                                     |
| Anwesenheit (Jahr)     | 31.12.2015    | 31.12.2018                                                                                   | 4      | 83%                | 84%                |   |                                                                                     |
| Anwesenheit (Monat)    | 30.11.2015    | 30.09.2018                                                                                   | 20     | 83%                | 100%               |  |                                                                                     |

Abbildung 17: Übersicht über die letzten Messwerte

Turniere

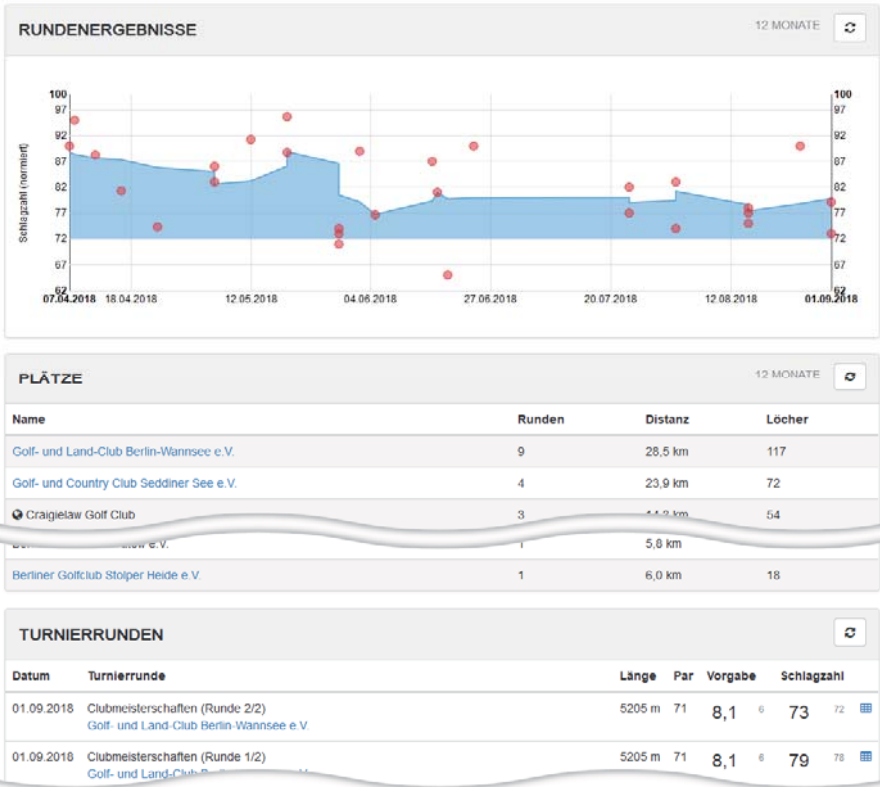


Abbildung 18: Visualisierung der Teilnahme an Golf-Turnieren

## Ausblick

Als Entwicklungsmöglichkeiten für das Web-Portal im weiteren Projektverlauf sind die folgenden Punkte angedacht.

Die Bildung von anonymen Spielergruppen für den Vergleich eines einzelnen Spielers mit seiner potentiellen Konkurrenz (auch Benchmark-Gruppen genannt) ist einer der nächsten Ziele, um die aktuelle Leistung und seine Entwicklung besser einschätzen zu können. Darauf aufbauend könnte in der Spielerakte eine Einordnung des Spielers in seine Benchmark-Gruppe nach verschiedenen Leistungsgrößen visualisiert werden.

Zurzeit werden in der Spielerakte nur die jeweils letzten Messwerte eines Messwerttyps dargestellt. Es wäre wünschenswert, die Entwicklung der Messgrößen als Trend oder kleines Liniendiagramm darzustellen. Dadurch wären Fortschritt, Stagnation oder sogar Rückschritte in der Entwicklung des Spielers leichter erkennbar.

Die Messwertanalyse kann in mehrfacher Hinsicht verbessert werden. Zunächst sollten die Benchmark-Gruppen auch in der interaktiven Messwertanalyse zur Verfügung stehen. Des Weiteren kann es helfen nicht nur einen Spieler hervorzuheben, sondern gezielt zwei Spieler zu vergleichen. Zuletzt ist es interessant die Messwerte einer Untermenge der Spieler darzustellen. Als Untermenge sind z. B. die durch einen Trainer betreuten Spieler, eine konkrete Trainingsgruppe, oder eine Altersklasse denkbar.

# Mobile-App

Konzeption und Implementierung  
einer plattformübergreifenden mobilen Applikation  
zur Erfassung von individuellen  
und gruppenbezogenen Leistungsdaten

**André Nitze**

## **Inhalt**

|                                                |     |
|------------------------------------------------|-----|
| Einleitung .....                               | 182 |
| Aufbau und Funktionen .....                    | 182 |
| Vorgehensweise und Anwendungsarchitektur ..... | 185 |
| Ergebnisse und Ausblick .....                  | 188 |
| Quellen .....                                  | 192 |

## Einleitung

Im Rahmen des Forschungsprojekts “PreLytics” an der Technischen Hochschule Brandenburg wurde ein Informationssystem entwickelt, das die im Projektverlauf gefundenen Lösungsansätze prototypisch demonstriert. Das Projekt beschäftigt sich mit der Verarbeitung heterogener Daten, wobei offene, proprietäre und kommerzielle Datenquellen integriert und gemeinsam zur automatisierten Erstellung von spezifischen Entscheidungshilfen (Präskriptionen) verwendet werden sollen.

Zum Aufbau eines eigenen (proprietären) Datenbestands wurde anhand des Anwendungsfalls des ersten Projektpartners eine mobile Applikation (App) konzipiert und implementiert. Der Zweck der mobilen App sollte die Erfassung individueller und gruppenbezogener Leistungsdaten sein. Insbesondere sollten die Anwesenheit in geplanten Trainings und individuelle Trainings mit eigenen Schwerpunkten, aber auch spezifische Leistungskennzahlen wie standardisierte Schlagweiten und Ergebnisse von standardisierten Lauf- und Fitnesstests erfasst werden können.

In diesem Beitrag werden die Ausgangssituation, die technologische Basis und der inhaltliche Aufbau der mobilen App beschrieben. Weiterhin wird der API-First-Ansatz beschrieben, der dem Entwicklungsprozess zugrunde liegt. Der Beitrag schließt mit einem Ausblick auf die Entwicklungspotenziale im weiteren Verlauf des Projekts.

## Aufbau und Funktionen

In der inhaltlichen Konzeptionsphase ließen sich schnell die zwei großen funktionalen Bereiche “Training” und “Statistik” ableiten, die zur flexiblen Erfassung und Darstellung von Daten dienen sollten.

Im Bereich *Training* können Trainer ihre eigenen Trainingsgruppen einsehen und dafür die in einem Zeitplan festgelegten wöchentlichen Trainings erfassen (vergl. Abbildung 1). Die Trainingsteilnahme einzelner Spieler war die wesentliche zu erfassende Kennzahl, die Aufschluss über die Trainingsintensität und damit Einfluss auf die Leistungsfähigkeit der Spieler geben sollte. Ein rot markierter Hinweis mit der Anzahl nicht erfasster Trainings signalisiert zudem Handlungsbedarf. Eine wesentliche projektinterne Zielgröße war hier der Pflegezustand der Trainings. Je

mehr Trainings erfasst wurden, desto belastbarer sollten die davon abgeleiteten Präskriptionen sein. Daher mischt sich für den Trainer hier die fachliche Information mit der Meta-Information über die Vollständigkeit der Daten.

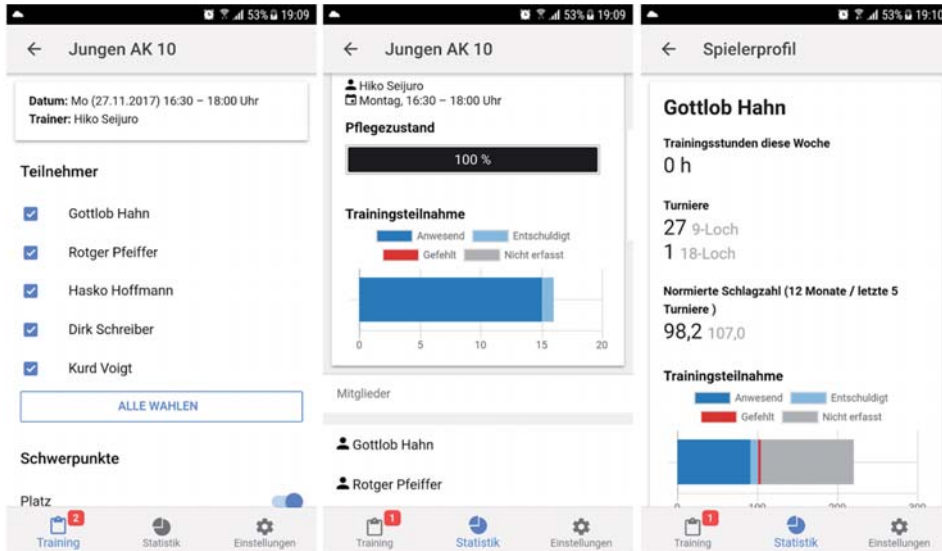


Abbildung 1: Trainer-Bildschirme in der mobilen App (v.l.n.r.: Training erfassen, Details einer Trainingsgruppe anzeigen, Spielerprofil anzeigen)

Im Bereich *Statistik* können Trainer auch die Trainings der anderen Trainer einsehen und auch bearbeiten. Dies ist ein gewünschtes Verhalten, um Trainingssitzungen untereinander tauschen zu können. Über die Statistik lassen sich auch alle Trainingstermine in der aktuellen Saison als Listenansicht anzeigen, um Daten schnell und komfortabel nachtragen zu können. Die Detailansicht eines Trainings zeigt nicht nur die stattgefundenen und zukünftigen Trainings in der Saison an, sondern gibt auch einen Überblick über die Spieler der betroffenen Trainingsgruppe.

Über die Detailinformation der Trainingsgruppe ist auch ein Wechsel in die *Spielerakte* einzelner Spieler möglich. In der Spielerakte sollen möglichst viele relevante Informationen zum Spieler gesammelt und strukturiert dargestellt werden (vergl. Abbildung 2). Dazu zählen neben den Grunddaten des Spielers auch Trainingsinten-

sität, Anwesenheit und gespielte Turniere. Für Trainer und Spieler besteht gleichermaßen die Möglichkeit, Leistungskennzahlen zu erfassen. Spieler können diese nur für sich selbst erfassen, während Trainer Daten für alle Spieler erfassen können.

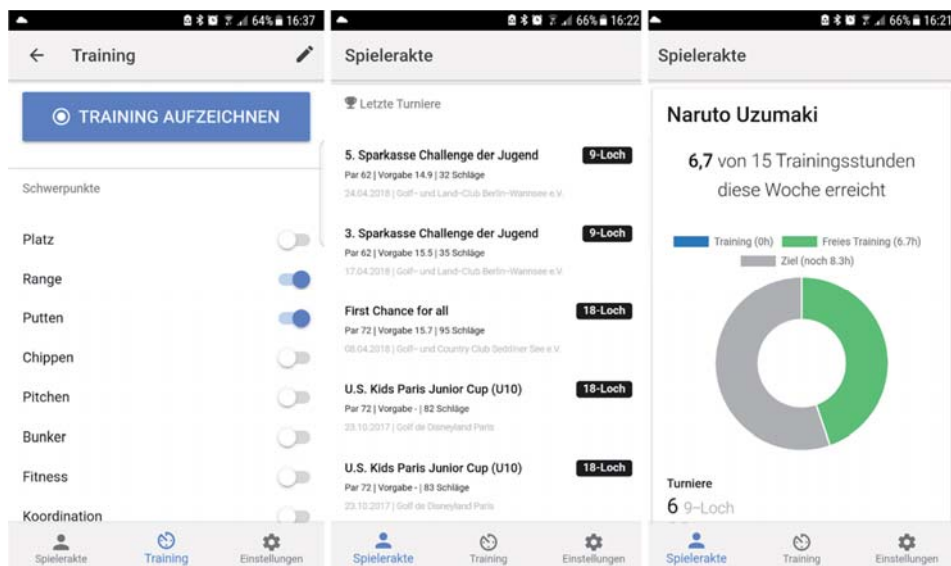


Abbildung 2: Spieler-Bildschirme in der mobilen App (v.l.n.r.: Training mit Schwerpunkten aufzeichnen, Spielerakte - Turniere, Spielerakte - Grunddaten)

Die App war zunächst nur für Trainer konzipiert und ist erst später auf Spieler-Funktionen ausgeweitet worden. Spieler konnten als Grundfunktionalität zunächst ihre eigene Spielerakte einsehen, wodurch ein neuer Bereich in der Navigation anstelle der Statistik entstanden ist. Die Spielerakte enthält aus Spielersicht fast vollständig dieselben Informationen, die auch die Trainer einsehen können.

Nach der Einführung der eigenen Spielerakte konnten die Spieler im Bereich *Training* nicht nur die von den Trainern für sie erfassten Trainings einsehen, sondern auch eigene Trainings mit individuellen Schwerpunkten, sowie körperliche (Größe, Gewicht) und leistungsbezogene Daten (Laufzeiten, Fitness-Tests, Schlagweiten etc.) erfassen. Technisch werden hier zunächst nur andere API-Endpunkte abge-

fragt. Im Backend werden in der Rolle des Spielers jedoch keine schreibenden Zugriffe auf die nur für Trainer vorgesehenen Ressourcen zugelassen, um Manipulationen auszuschließen.

Zur Verbesserung des *Nutzungserlebnisses* enthält die App einige Komfortfunktionen, die die Validierung von Eingaben oder Hinweise in verständlichen Dialogen betreffen. Die Interaktion zur Datenerfassung wurde mehrfach überarbeitet, um Fehlerquellen auszuschließen und den Nutzer bei der schnellen Lösung seiner Aufgabe zu unterstützen. Es wurden zudem Funktionen ergänzt, die das Korrigieren und Löschen von Eingaben erlauben, um dem Nutzer ein höchstmögliches Kontrollgefühl über die App zu geben. Kleine Animationen von Diagrammen und Icons unterstützen subtil den sportlich-spielerischen Charakter der App. Die blaue Akzentfarbe weist auf allen Bildschirmen auf die wesentlichen Informationen hin. Zusätzliche Farben für Diagramme und andere Visualisierungen sind in einer standardisierten Farbpalette vorgegeben.

## Vorgehensweise und Anwendungsarchitektur

Die technische Entwicklung der mobilen Anwendung wurde mit dem “Ionic” Framework<sup>1</sup> realisiert. Das Framework basiert auf dem weit verbreiteten AngularJS-Framework<sup>2</sup> und stellt grundlegende Elemente des User Interface (UI) wie Eingabefelder, Schaltflächen und Listen für alle unterstützten mobilen Betriebssysteme bereit, was den Entwicklungsaufwand neuer Funktionen wesentlich verringert. Alle UI-Komponenten werden mit den Basistechnologien HTML, (S)CSS und TypeScript realisiert. Der TypeScript-Code wird zur Laufzeit zu einfachem JavaScript umgewandelt und kompiliert, um eine maximale Kompatibilität mit den jeweiligen Plattformen zu erreichen.

Durch die Nutzung nativer Erweiterungen über Cordova<sup>3</sup> können bei Bedarf auch native Bedienelemente, wie zum Beispiel zur Datumsauswahl oder zum Hochladen von Fotos, eingesetzt werden. Dadurch lässt sich die resultierende hybride App

---

<sup>1</sup> <https://ionicframework.com/>

<sup>2</sup> <https://angularjs.org/>

<sup>3</sup> <https://cordova.apache.org/>



kaum von einer nativen Anwendung der jeweiligen Plattform unterscheiden. Die Möglichkeit zur plattformübergreifenden (hybriden) Entwicklung für mehrere mobile Plattformen, insbesondere iOS und Android, war ein wesentliches Argument für die Auswahl des Frameworks (vergl. [1]).

Der Quellcode kann, wie alle entwicklungsrelevanten Projektergebnisse, in der zentralen Versionsverwaltung (*Version Control System*, VCS) “GitLab” abgerufen werden<sup>4</sup>. Das VCS wird auf einem virtuellen Server im Rechenzentrum der Technischen Hochschule Brandenburg betrieben. Die Projektstruktur orientiert sich am Standard für AngularJS-Projekte. Die komplette App wird als *Single-Page-Application* (SPA) umgesetzt. Das Routing zwischen den Bildschirmmasken wird demnach über klassische Links abgebildet, wodurch auch tiefe Verweise (deep links) wie auf herkömmlichen Webseiten möglich sind.

Die technische Projektstruktur ist dem Fachkonzept nachempfunden, sodass alle einzelnen Seiten bzw. “Bildschirme” mit ihrer Geschäftslogik (TypeScript), Struktur (HTML) und Stylesheets (SCSS) im `pages`-Ordner zu finden sind. Mehrfach verwendete Elemente wie der Fortschrittsbalken oder das Diagramm zur Trainingsanwesenheit sind unter `components` abgelegt. Die zentrale Konfiguration der App findet sich im Ordner `app` wieder. Grafiken, Icons und Lokalisierungsdateien können, insofern sie im Ordner `assets` liegen, mit den Framework-eigenen Funktionen in Templates referenziert werden.

Durch die parallele Entwicklung des Web-Frontends konnten Teile der existierenden oder parallel benötigten Geschäftslogik wiederverwendet werden. Als Architekturstil kam *REpresentational State Transfer* (REST) zum Einsatz. Die Programmierschnittstelle (Application Programming Interface, API) im SPI Backend wurde demgemäß REST-basiert implementiert und von den mobilen Clients konsumiert. Interaktionen sind also über die üblichen HTTP-Verben (POST, GET, UPDATE, DELETE, PATCH) auf den bestehenden Ressourcen (Spieler, Gruppen, Trainings, Kennzahlen etc.) möglich. Als Medium für den Nachrichtentransport wurde die JavaScript Object Notation (JSON) gewählt, da sie von den JavaScript-basierten Clients sowohl im Web als auch mobil nativ und ohne Deserialisierung verarbeitet werden kann.

---

<sup>4</sup> <https://fbwgitlab.th-brandenburg.de/prelytica-spi/frontend-mobile>

Der Entwicklungsprozess der App beginnt nach der Anforderungserhebung mit der Definition der zentralen API, was auch als “API-First”-Ansatz bezeichnet wird. Zwischen dem jeweiligen Frontend- und Backend-Entwickler werden dafür zunächst das für die neu zu implementierende Funktion benötigte Datenmodell und die zu verwendenden Endpunkte abgestimmt. Diese Vereinbarung wird in einer OpenAPI-Spezifikation<sup>5</sup> formalisiert und mit konkreten Beispielen ergänzt, sodass die Erwartungen eindeutig sind und Fehlentwicklungen minimiert werden. Ausgehend von diesem “Vertrag” zwischen Frontend und Backend können beide Seiten mit der jeweiligen Implementierung beginnen.

Für das mobile Frontend (die App) können aus der OpenAPI-Spezifikation Modelle im Sinne des Entwurfsmusters *Model-View-Controller* (MVC) generiert werden. Mit Hilfe eines Generators<sup>6</sup> lassen sich nicht nur diese Modelle, sondern auch die notwendigen Zugriffsmethoden generieren. Dadurch ist eine sehr schnelle und strukturierte Entwicklung möglich, die sogar parallel zu einer noch in der Entwicklung befindlichen Backend-Funktionalität stattfinden kann. Die Projektstruktur wird durch diesen Ansatz um die Ordner `models` mit den Interfaces der Datenmodelle bzw. `Models` und die den Ordner `api` mit den Zugriffsmethoden ergänzt. Über die Einbindung des ebenfalls generierten API-Providers im entsprechenden Abschnitt der Konfigurationsdatei `app.module.ts` lassen sich die nach den Routen benannten API-Zugriffsmethoden in allen Controllern der Anwendung nutzen. Die Datenmodelle lassen sich wiederum direkt in den HTML-Templates bzw. Views referenzieren. Dadurch lässt sich ein drastischer Zugewinn in der Entwicklungsgeschwindigkeit und der Robustheit des Produkts erzielen.

Zur Vermeidung von Regressionsfehlern sollten die generierten Methoden und Modellen nicht manuell verändert, sondern neu aus der API-Beschreibung generiert und überschrieben werden. Die gesamte Herangehensweise wird auch als vertragsbasierte Entwicklung (*contract-driven development*) bezeichnet. Weitere Details zum Entwicklungsprozess der API und deren vertragsbasierter Beschreibungsform finden sich in [2].

---

<sup>5</sup> <https://github.com/OAI/OpenAPI-Specification>

<sup>6</sup> <https://github.com/swagger-api/swagger-codegen>

Technisch interessant ist noch die Aufteilung von Geschäftslogik, die nicht eindeutig dem Frontend- oder Backend-Bereich zuzuordnen ist, wie zum Beispiel Sortierfunktionen. Als grundsätzliche Richtlinie hat sich hier die natürliche Struktur des Datenmodells etabliert, sodass natürliche Reihenfolgen wie ein Zeitverlauf oder einander bedingende Handlungsschritte auch in ihrer impliziten Reihenfolge wiedergegeben werden. Ist dies nicht gegeben, dürfen von Frontends keine Annahmen über die Reihenfolge von Ressourcensammlungen getroffen werden.

In Bezug auf Sicherheit basiert die gesamte Architektur des Informationssystems auf dem RFC-7519-Standard<sup>7</sup> der Internet Engineering Task Force (IETF), auch bekannt als *JSON Web Token* (JWT). Nach dem mit TLS 1.2 verschlüsselten Versand der Zugangsdaten wird ein manipulationsgeschützter Token (JWT) ausgestellt und vom Client in jeder darauffolgenden Kommunikation mit dem Server genutzt. Eine böswillige Veränderung der Token-Inhalte (Name, Rollen, Rechte, Zeitstempel etc.) führt bei der Verifizierung mit der enthaltenen Signatur zu einem Fehler, wodurch zahlreiche Angriffsvektoren erschwert werden. Im Token können beliebige Informationen zur Nutzerin wie die E-Mail-Adresse, Profildaten und Rollen (Spieler/Trainer/ Administrator) mitgeführt werden, sodass damit leicht ein Berechtigungskonzept umgesetzt werden kann.

## Ergebnisse und Ausblick

Die mobile Applikation wurde für iOS und Android in den jeweiligen Distributionsplattformen (App Stores) mit Unterstützung für Smartphones und Tablets veröffentlicht<sup>8,9</sup>. Die Registrierung der Nutzer per E-Mail-Adresse wurde individuell koordiniert, um eine schwer zu kontrollierende und potenziell risikobehaftete freie Registrierung zu vermeiden. Eine externe Bewerbung der App erfolgte nicht, um die Sichtbarkeit auf den für das Projekt notwendigen Personenkreis zu beschränken.

Es wurden über zwei Jahre hinweg vor allem die *Anwesenheit von Spielern und die inhaltlichen Schwerpunkte von Trainings* erfasst. Daneben wurden auch - obgleich

---

<sup>7</sup> <https://tools.ietf.org/html/rfc7519>

<sup>8</sup> Android: <https://play.google.com/store/apps/details?id=de.thbrandenburg.prelitycaspi>

<sup>9</sup> iOS: <https://itunes.apple.com/us/app/sport-performance-insight/id1233629508?ls=1&mt=8>

deutlich weniger - individuelle Trainings und Leistungskennzahlen aufgezeichnet. Hier bieten sich noch zahlreiche Gelegenheiten, vor allem die zum Teil noch sehr jungen Nutzer im Rahmen spielerischer Anreizgestaltung zur Erfassung von Daten zu motivieren (vergl. „Gamification“). Weder auf Seiten der Trainer noch der Spieler sind maßgebliche Probleme in der Bedienbarkeit der App aufgetreten.

Als erster wesentlicher Motivator für die Erfassung von Daten aus Nutzersicht, hat sich die *Visualisierung der Datenqualität* gezeigt, die über einen Fortschrittsbalken die Vollständigkeit der erfassten Trainings in der aktuellen Saison zeigte („Pflegezustand“).

Aus den erfassten Daten ließen sich bisher die Trainingsintensität und auch das *Engagement von einzelnen Spielern und Gruppen* zielgerichtet zusammenfassen und bewerten. Auch erste tiefergehende Analysen (Korrelation von Alter, Schlagweite, Wetter, gespielten Turnieren etc.) wurden mit diesem Datenbestand möglich. Die Daten stehen darüber hinaus für weiterführende Auswertungen in der App und im Web-Frontend zur Verfügung. Durch den direkten Zugriff auf die Datenbank ist auch eine explorative Analyse mit externen Werkzeugen und eine Anreicherung mit anderen Datenquellen möglich.

Die App kann über die Pflege entsprechender Datenbankfelder dynamisch um *weitere Leistungsparameter* erweitert werden, sodass auch zukünftige Kennzahlen ohne ein neues Release auf den Distributionsplattformen integriert werden können. Zudem lässt sich das gesamte Informationssystem dadurch in nahezu beliebigen Anwendungsgebieten einsetzen. Dazu zählen nicht nur andere Sportarten, sondern auch betriebswirtschaftliche Kennzahlen.

Mittlerweile unterstützt das gewählte technologische Framework auch *Progressive Web Apps* (PWAs). Die App lässt sich durch die zugrundeliegenden Web-Technologien leicht modernisieren und auf weitere Plattformen oder neuere Bibliotheken portieren.

Die nächsten Entwicklungsschritte nach der Erfassung eigener Daten und der Integration zusätzlicher Datenquellen ist nun die *Entwicklung erster Präskriptionen* für Trainer und Spieler. Abbildung 3 zeigt beispielhaft, wie ein erster - zunächst rein deskriptiver Schritt - in diese Richtung aussehen kann. Zu sehen sind einfache Beobachtungen, die sich aus den bereits vorhandenen Daten problemlos für alle Spie-

ler berechnen lassen. Diese Beobachtungen können zu konkreten Handlungen anregen, z. B., die eigene Verbesserung in der Spielerakte anzusehen ("Zum Profil") oder zusätzliche Leistungskennzahlen zu erfassen. Im nächsten Schritt müssen Wege gefunden werden, die Unterschiede in den erfassten Kennzahlen der Spieler darzustellen und zu einer eindeutigen Handlungsempfehlung für Spieler, Trainer oder beide zu überführen. Das könnte zunächst nur eine allgemeine Empfehlung für einen Spieler zu mehr Fitness-Training sein, wenn große Abweichungen in einer durch Fitness-Training beeinflussbaren Kennzahl, z. B. Schlagweite, innerhalb der Altersklasse auftreten. Weiterhin könnten dem Trainer kaum spürbare Entwicklungen wie der stetige Rückgang der Trainingsintensität einzelner Spieler hervorgehoben werden, um rechtzeitig positiv einzuwirken. Auch das ist bereits mit den vorhandenen Daten möglich. Diese allgemeinen Empfehlungen müssen dann mit zunehmender Datenbasis stückweise spezifischer gestaltet werden, bis eine gewisse Menge an relevanten, belastbaren und automatisierten Präskriptionen tatsächlich das individuelle Training und das Training in Gruppen unterstützen kann.

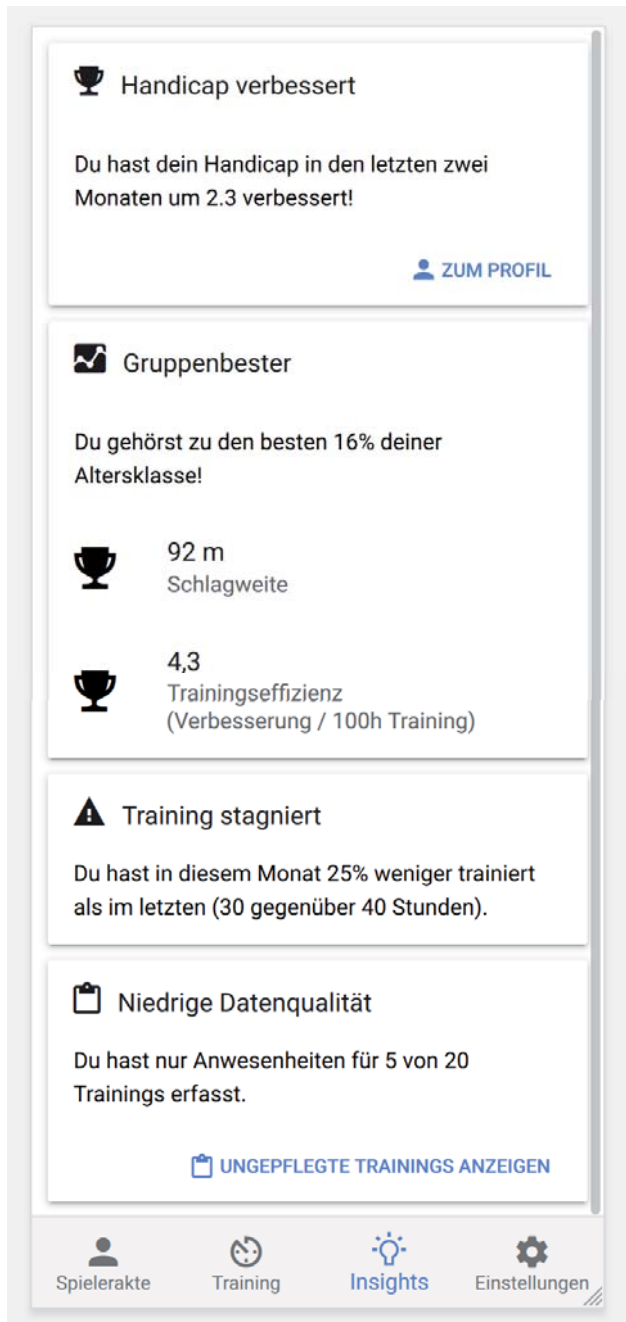


Abbildung 3: Entwurf für präskriptive Funktionen in der Entwicklungsumgebung der mobilen App

## Quellen

- [1] NITZE, ANDRÉ : The Case for Web Technologies in Mobile Business Apps. In: *Berliner Schriften zu modernen Integrationsarchitekturen, Bewertungsaspekte service- und cloudbasierter Architekturen*. BSOA/Bcloud 2014, November 2014. — ISBN 978-3-8440-2940-6.
- [2] KIERTSCHER, TOBIAS ; NITZE, ANDRÉ : Contract-first Development - Vom OpenAPI-basierten Vertrag zur Implementierung. In: SCHMIETENDORF, A. (Hrsg.): *ESAPI 2017 - 1. Workshop Evaluation of Service-APIs*. Berlin, 2. November 2017. Berliner Schriften zu modernen Integrationsarchitekturen, Shaker-Verlag, Aachen, November 2017. — ISBN 978-3-8440-5559-7.

# Leistungsfähigkeit von Solarstrom- anlagen beeinflussende klimatische Faktoren

Stefan Rieder

## Inhalt

|                                |     |
|--------------------------------|-----|
| Einleitung .....               | 194 |
| Einflussfaktoren .....         | 194 |
| Modultemperatur.....           | 195 |
| Ablagerungen.....              | 196 |
| Mechanische Beanspruchung..... | 198 |
| Einstrahlung.....              | 199 |
| Abbildung in Klimadaten .....  | 200 |
| Fazit.....                     | 202 |
| Quellen.....                   | 203 |



## Einleitung

Im Zuge des Projektes Präskriptive Analytik für KMU in Brandenburg (PreLytica) sollen im Teilprojekt Solar Potential Analytics (SPA) auf Basis heterogener Datenquellen Handlungsempfehlungen für kleine mittelständische Unternehmen gegeben werden. Es wird untersucht, in welchen Fällen sich eine Solarstromanlage rentiert und entsprechende Simulationen werden transparent dargestellt. Da ein Bau einer Photovoltaikanlage im gewerblichen Bereich mit Investitionen in andere unternehmerische Bereiche Vergleichen Stand halten muss, ist es wichtig sämtliche Unsicherheiten zu berücksichtigen um eine exakte Planung zu ermöglichen. Mit einer geplanten Laufzeit von über 20 Jahren addieren sich sämtliche Faktoren, die die Anlageneffizienz beeinflussen. Meist wird dabei an technische Parameter wie Anlagenleistung, Moduleffizienz und Anlagentechnik gedacht. Einige Hersteller werben mit einer garantierten maximalen technischen Degradation von 20 % nach 25 Jahren Betrieb. Jedoch haben auch umweltbezogene variable Faktoren einen enormen Einfluss auf die Effizienz und am Ende auch auf die Wirtschaftlichkeit der Solarstromanlage. Daher werden bei den Simulationen des Teilprojektes **SPA** auch Einflussparameter berücksichtigt, die vorwiegend vom Standort abhängen und beim Bau der Anlage nur bedingt änderbar sind. Es gilt, einflussnehmende Faktoren zu benennen und diese am Ende anhand empirischer Daten auch in die Simulationen mit auf zu nehmen. Über eine langjährige Laufzeit ist weiterhin zu untersuchen, ob sich aufgrund des Klimawandels Veränderungen in absehbarer Zukunft ergeben und wie sich diese auswirken könnten.

## Einflussfaktoren

Unter dem Begriff der Degradation versteht man die Leistungsreduzierung von Modulen bzw. Solarstromanlagen über die Zeit. Auf globaler Ebene lässt sich die Verschlechterung bzw. Variabilität der Leistungsfähigkeit von Photovoltaikanlagen nicht einfach beziffern. Regionale klimatische Unterschiede lassen eine generelle Aussage schlichtweg nicht zu. Während in tropischen Regionen eine hohe Luftfeuchte herrscht, die beispielsweise Korrosion beschleunigen könnte, stehen dem eine trockene staubige Luft in ariden Regionen gegenüber. Auch hohe

Temperaturen beeinflussen die Leistungsfähigkeit von Solarstromanlagen merklich. Mit Hinblick auf potentielle Standorte von PV-Anlagen in Brandenburg gilt es herauszufinden, welche Faktoren einen Effekt auf den dort erzielbaren Ertrag haben könnten und wie diese in Simulationen berücksichtigt werden können. Einerseits können Parameter, die in anderen Regionen der Erde einen erheblichen Effekt haben, im kühlgemäßtem Klima Deutschlands eine untergeordnete Rolle spielen. Auf der anderen Seite sind in dem erwähnten Projektgebiet im Sommer auch ausgedehnte Trockenperioden (wie zuletzt 2018) möglich, die Anlehnungen an aride Gebiete zulassen. Im Folgenden soll eine Übersicht von Einflussfaktoren gegeben werden, die im Laufe des Projektes empirisch evaluiert und in die Potentialberechnung einfließen werden.

## Modultemperatur

Die Standard-Testbedingungen (STC für *Standard Test Conditions*) für Photovoltaikmodule sehen eine Modultemperatur von 25°C vor. Obwohl dies als Optimum angesehen werden kann, wird dieser Wert in der Realität in Zeiten ohne Bewölkung und hoher Sonneneinstrahlung (im Frühling & Sommer) nur selten erreicht. Abhängig von Umgebungstemperatur und Einstrahlung kann die Modultemperatur nahezu doppelt so hoch wie die der Umgebungsluft sein (vergl. Abbildung).

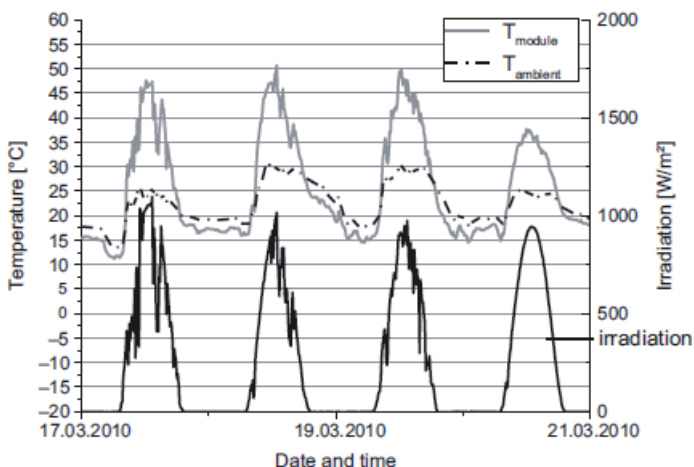


Abb. 1: Monitoring der Einstrahlung sowie der Modul- und Umgebungstemperatur an vier aufeinanderfolgenden Tagen in Freiburg ([1], S. 94).

Auch wenn erhöhte Temperaturen keinen direkten Stress auf PV-Module ausüben, so beschleunigen sie den Degradationsprozess (vergl. [1], S. 97) und senken die aktuelle Leistungsfähigkeit des Moduls. Der Temperaturkoeffizient eines Moduls, der in jedem Datenblatt zu finden ist, beschreibt den relativen Wirkungsgradverlust pro Grad Abweichung von dem Optimum. Dieser beträgt bei kristallinen Modulen meist 0,4 %. Demzufolge müsste allein aufgrund der um 25° erhöhten Modultemperaturen (vergl. Abbildung 1) ein Abschlag von 10 % auf den Wirkungsgrad der Anlagen berechnet werden.

Wind oder leichte Konvektionsströme sorgen hingegen für eine Kühlung der Module. Bei identischen Rahmenbedingungen hat also auch die Art der Anlagenerrichtung einen Einfluss auf die Modultemperatur. Während auf Flachdächern aufgeständerte Anlagen viel Platz zum Belüften aufweisen, sind auf geneigten Dächern angebrachte Module aufgrund der geringen Hinterlüftung temperaturkritischer. Besonders empfindlich reagieren gebäudeintegrierte Photovoltaikanlagen, die in das Dach eingelassen werden und keinerlei Raum für kühlende Luftströme zulassen. Eine Untersuchung von Koehl (vergl. [2], S. 84) zeigt, dass Module, die in Gebäude integriert wurden, bei gleichen Bedingungen 17,5° wärmer waren als aufgeständerte Module. Dies würde einen weiteren zusätzlichen Verlust von 7 % bedeuten. Für den weiteren Projektverlauf ist wichtig, dass sich die Moduloberflächentemperatur auf Basis der klimatologischen Parameter Wind, Lufttemperatur und Einstrahlung sehr gut simulieren lässt (vergl. [3] & [2], S. 82).

## Ablagerungen

Ablagerungen kleiner in der Luft enthaltenen Partikel können einen starken Einfluss auf die Leistung von Photovoltaikanlagen haben. Neben Staub aus der Atmosphäre zählen auch Ablagerungen von pflanzlichen Produkten (z. B. Pollen oder Blätter), Ruß und Salz. Dieser als „soiling“ beschriebene Effekt wird als signifikant bezeichnet (vergl. [4], S. 2). Untersuchungen in ariden Regionen der Erde zeigen, dass der Energieertrag unter realen Bedingungen um bis zu 1,3 % pro Jahr gemindert werden kann (vergl. [4], S. 2). Sowohl am Standort Köln als auch auf der Zugspitze konnte bereits nach einem Jahr ein Leistungsverlust von jeweils 0,5 % festgestellt werden (vergl. [2], S. 145).

In unseren Simulationen wird der Fokus auf Beeinträchtigungen durch Staub gesetzt, da dieser von gut messbaren empirischen Parametern abhängig ist. Dazu zählen einerseits die Lage und Umgebung des Anlagenstandortes, andererseits hat das Klima, bzw. die vorherrschende Witterung, einen großen Effekt auf solche Ablagerungen. Je nach Niederschlagshäufigkeit, Staub- und Pollenstürmen sowie Windgeschwindigkeit nimmt die Ablagerung beschriebener Partikel zu oder es kommt zu einer Selbstreinigung der Moduloberflächen. Um mögliche Effekte genau beschreiben zu können, ist es jedoch notwendig neben den Umweltfaktoren auch die Anlagenparameter zu berücksichtigen. Vor Allem die Neigung der Solarmodule hat einen Einfluss auf etwaige Ablagerungen.

Die Sedimentation von Partikeln kann im Grundlegenden in zwei Prozesse aufgeteilt werden: Auch in der Luft enthaltene Partikel folgen den Gesetzen der Schwerkraft und somit einem gravitativ bedingtem Absinken. Je größer (bzw. schwerer) ein Partikel, desto vertikaler ist dessen Ablagerungsrichtung. Daneben gibt es eine turbulente Ablagerung, die je nach Wind vor allem kleinere Partikel betrifft. Diese lagern sich nicht nur auf nach oben richtenden Flächen ab, sondern können auch vertikale Ebenen belagern. Betrachtet man zwischen  $0^\circ$  und  $90^\circ$  geneigte Solarmodule, so wird klar, dass bei horizontal liegenden Kollektoren mehr Fläche mit Staubpartikeln exponiert ist, als dies bei vertikal geneigten Kollektoren der Fall ist.

Folgende Abbildung zeigt die Ergebnisse eines Experiments, bei dem Glastafeln mit unterschiedlicher Neigung über mehrere Monate den alltäglichen Bedingungen in einer ariden Region Indiens ausgesetzt und im Beobachtungszeitraum nicht manuell gereinigt wurden (vergl. [4], S. 4).

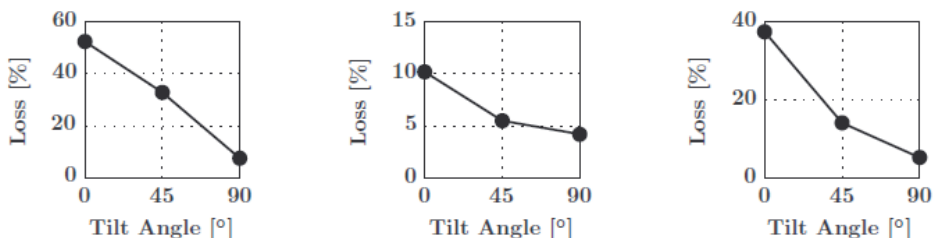


Abb.2: Transmissionsverluste geneigter Glasflächen von Mai 1986 bis Dezember 1987.  
 Links: 11. Juni 1986 bis 10. Dezember 1986 (182 Tage); Mitte: 20. Januar 1987 bis 18. Juli 1987 (180 Tage); Rechts: 21. Juli 1987 bis 15. September 1987 (56 Tage).

Auch wenn die dort vorherrschende Staubbelastung nicht direkt auf das Projektgebiet in Deutschland übertragbar ist, lassen sich die aus den Daten hervorgehenden Aussagen zum relativem Leistungsverlust auf Basis verschiedener Neigungswinkel jedoch durchaus auf mitteleuropäische Situationen anwenden. Die Erkenntnisse geben die zwei grundlegenden Prozesse der Staubablagerung (gravitativ und turbulent) gut wieder. Auffallend ist, dass der 180-tägige Beobachtungszeitraum einen deutlich geringeren Verlust mit sich bringt, als der deutlich kürzere 56 Tage andauernde Zeitraum. Grund hierfür sind starke Regenfälle in den 180-Tages-Linien, die in den anderen Beobachtungszeiträumen nicht auftraten (vergl. [4], S. 4).

Nicht nur die Ablagerung, sondern auch die durch Niederschlag und Wind mögliche Selbstreinigung der Kollektorflächen wird von Partikelgröße und Modulneigung beeinflusst. Eine Reinigung durch Wind ist nur bei großen Partikeln möglich (vergl. [4], S. 3), da Adhäsionskräfte im Verhältnis zur möglichen Angriffsfläche zu stark sind. Für eine Beseitigung sämtlicher Partikel reichen Windstärken von 180 km/h nicht aus (vergl. [4], S. 3). Den besten Selbstreinigungseffekt erzielt Regen, sofern er in ausreichender Intensität auftritt. Fällt nur wenig Niederschlag in staubiger Atmosphäre oder auf eine verschmutzte Oberfläche, so binden die einzelnen Tropfen die Partikel und bilden eine harte Schicht, die sich nur schwer von alleine löst. Ein ähnlicher Effekt kann bei hoher Luftfeuchtigkeit auftreten. In solchen Fällen verschlechtert der Niederschlag die Effizienz von Solarmodulen sogar deutlich. Kommt es jedoch zu ausreichend intensivem Niederschlag, so setzt ein Selbstreinigungsprozess ein, der bei zunehmender Modulneigung aufgrund des besseren Abflusses effizienter wird. Eine geneigte Modulfläche ist somit nicht nur weniger anfällig für Staubpartikel, sondern weist auch eine bessere Selbstreinigung bei Regen auf. Während Regen bei auf 15° geneigten Modulen nahezu keinen Effekt hatte (vergl. [2], S. 145), wurden steiler aufgeständerte Module von schweren Verschmutzungen, die 80 % Verlust mit sich brachten, nahezu komplett gesäubert (vergl. [2], S. 147).

## Mechanische Beanspruchung

Neben kontinuierlich Einfluss nehmenden Parametern wie der oben beschriebenen Temperatur, dem Wind oder der Ablagerungen gibt es auch Ereignisse, die nur sporadisch auftreten. Diese können verheerende Folgen für den

Betrieb einer Solarstromanlage haben. Extremereignisse führen selten zu einer steigenden Degradation der Anlage, sondern beschädigen diese nachhaltig. Es wurden drei Ereignisse erfasst, die eine Beschädigung nach sich ziehen können.

Starker Wind oder Orkanböen können aufgrund der mit ihnen auftretenden Kräfte PV-Anlagen aus Ihren Verankerungen reißen oder Module beschädigen. Besonders aufgeständerte Anlagen auf Flachdächern sind für Schäden durch Windsog anfällig. Eine entsprechend dimensionierte Ballastierung und Arretierung ist bei diesem Anlagentyp unabdingbar. Direkt auf dem Dach platzierte oder gar integrierte Anlagen betrifft dies kaum, da deutlich weniger Angriffsfläche geboten wird. Gleiches gilt für hohe Lasten, die durch Schnee verursacht werden können und ebenfalls zu Schäden führen können. Moderne Module verfügen über eine überaus stabile Oberfläche, die sie sehr belastbar macht, punktuelle Einschläge durch Hagelkörner bilden jedoch das dritte Extremereignis, das zu Schäden führen kann.

Da durch die beschriebenen Parameter auftretende Schäden meist zügig durch Austausch oder Reparatur behoben werden, ist aktuell keine Simulation dieser Werte geplant. Des Weiteren sind Extremwetterereignisse räumlich und zeitlich sehr begrenzt und nicht nur deshalb sehr schwer zu simulieren.

## Einstrahlung

Der Vollständigkeit halber sei an dieser Stelle noch die solare Einstrahlung erwähnt, die als dominierender Faktor für die Leistungsfähigkeit von Solarstromanlagen gilt. Bei den Simulationen werden, den Dachteilflächen aus der 3D – Potentialanalyse entsprechend, Einstrahlungswerte verwendet (vergl. Artikel *Nutzen, Vorteile und Optimierung des Energieertrags*, S. 205). Die mittlere Einstrahlung in Deutschland liegt (auf einer horizontalen Fläche) pro Quadratmeter zwischen 900 und 1200 kWh/Jahr (vergl. [5]). Mit einem langjährigen Mittel von 1100 kWh/Jahr ist der Standort Brandenburg generell gut für Photovoltaik geeignet. Saisonale Schwankungen fallen in Deutschland sehr hoch aus, in den Monaten April bis August werden am Standort Brandenburg 150 kWh/m<sup>2</sup> erreicht, in den Winter- und Übergangsmonaten teilweise nur 20 kWh/m<sup>2</sup> (vergl. [5]). Dieser Fakt wird sich auch in Zukunft aufgrund des Klimawandels nicht ändern. Jedoch ist es möglich, dass aufgrund unterschiedlicher Niederschlagsmengen und damit verbundener Bewölkung die Übergangszeiten

sonniger oder die Sommer noch trockener werden. Diese Art der Veränderungen können auf Basis der Klimadaten simuliert werden.

## Abbildung in Klimadaten

Sieht man von chemischen Prozessen ab, die für die Degradation von Photovoltaikmodulen mitverantwortlich sind, lassen sich die relevanten Einflussfaktoren auf Basis von Klimadaten abbilden. Zuerst wäre dies die Modultemperatur, die maßgeblich für den Wirkungsgrad an sonnigen Tagen ist. Dem gegenüber kann ebenfalls der Kühleffekt durch Wind zuverlässig simuliert werden. Betrachtet man die Beeinflussung der Moduleffizienz aufgrund von Ablagerungen auf der Kollektorfläche, so ist es mit weiteren Überlegungen auch möglich diesen Verlust über die Zeit zu simulieren. Auf Basis der Parameter Temperatur und Niederschlag kann auf länger andauernde Trocken- und Hitzeperioden geschlossen werden, die in einer erhöhten Staubsedimentation resultieren. Diese würden je nach Neigung der Module auch durch Regen wieder gereinigt. In Kombination mit Daten zur solaren Einstrahlung entsteht somit ein komplexes klimatologisch-technisches Modell mit entsprechenden Wechselwirkungen, das auf Basis der Anlagenparameter etwaige Verluste oder Gewinne simuliert. Für den Projektkontext sind eben diese Simulationen von besonderem Interesse, da sich Rückschlüsse auf sich verändernde Rahmenbedingungen beim langjährigen Betrieb einer Solarstromanlage im unternehmerischen Kontext ziehen lassen.

Sämtliche zur Simulation notwendigen Basisdaten stehen bereits zur Verfügung. Wie im Artikel *Nutzen, Vorteile und Optimierung des Energieertrags (S. 250)* erwähnt, besteht die Möglichkeit das Wetter (bzw. dessen Parameter Temperatur, Niederschlag, Wind und Einstrahlung) der letzten Jahre stundengenau abzufragen und somit auch die vorherrschende Witterungslage zu erfassen. Auf Basis dieser Werte ist eine Modellierung der letzten Jahre für Kalibrierungszwecke möglich.

Da potenzielle Photovoltaikanlagen jedoch heute geplant und für mindestens die nächsten 20 Jahre oder mehr geplant, ausgelegt und betrieben werden, ist es von essentieller Bedeutung auch die zukünftigen Entwicklungen zu beachten. Beeinflussen Auswirkungen des vielschichtigen Klimawandels die Effizienz von Solarstromanlagen? Entstehen im Sommer mehr Trockenperioden die zu erhöhter

Staubbelastung und somit zu einem Minderertrag führen? Oder führen bewölkungsärmere Übergangsjahreszeiten zu erhöhter Sonneneinstrahlung und einem Mehrertrag? Diese und weitere Fragen lassen sich auf Basis regionaler Klimamodelle, die bis zum Jahre 2100 Daten liefern, und den herausgearbeiteten Prozessen beantworten. Im Rahmen des Projektes wird auf Ergebnisse des CORDEX Projekts zurückgegriffen. Hierbei wurden die Ergebnisse der CMIP5 – Daten, auf denen der 5. Sachstandsbericht des IPCC (*Intergovernmental Panel on Climate Change*) basiert, regionalisiert. Globale Klimamodelldaten lassen nur in sehr geringem Maße Aussagen zu regionalen Entwicklungen zu, was eine solche Skalierung notwendig macht. Erstellt und simuliert wurden verschiedene Szenarien (*Representative Concentration Pathways „rcp“*), die die Variabilität der Bandbreite der globalen Emissionsverläufe abbilden (vergl. [6]).

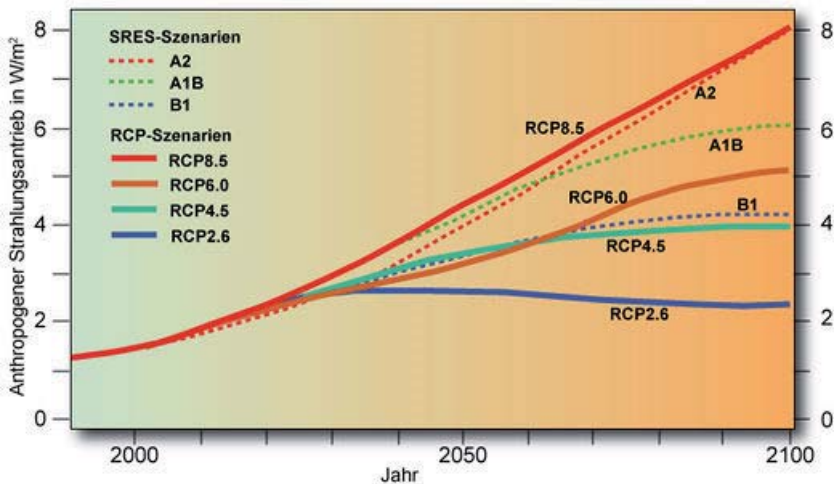


Abb.3: Strahlungsantrieb bis 2100; Vergleich SRES und RCP Szenarien (vergl. [7])

Bei den für die bisherigen Modelle entwickelten SRES Szenarien, wurden die Treibhausgasemissionen und damit die Klimaentwicklung auf Basis der Entwicklung von Energienutzung, Bevölkerung und Landwirtschaft abgeleitet. Die RCP Szenarien beschreiben in der Literatur veröffentlichte mögliche Entwicklungen der Treibhausgasemissionen bzw. des Strahlungsantriebs, wodurch Rückkopplungseffekte besser berücksichtigt werden können. Setzt man die neuen Szenarien in einen sozioökonomischen Kontext, so wäre rcp85 bei einer Zunahme der Weltbevölkerung auf 12 Mrd. Menschen bis 2100, einer Verdreifachung des Primärenergieverbrauchs und eines Öl- und Kohleanteils von



50 % realistisch. rcp26 bildet das andere Extrem, hier wird eine Weltbevölkerung von 9 Mrd. Menschen abgebildet, die mit sehr wenig Öl auskommt. Ein weiterer Unterschied ist der Ausstoß von CO<sub>2</sub>, der von heute fast 10 Gt/Jahr bei rcp85 auf 30 Gt/Jahr steigt und bei rcp26 im Jahr 2080 auf null fällt (vergl. [7]). Im Rahmen des Projekts **SPA** und der Simulation der Auswirkungen auf den Ertrag von Solarstromanlagen werden Klimamodelldaten aus dem Szenario rcp45 berücksichtigt, das unter heutigen politischen und sozioökonomischen Aspekten eine sehr realistische Entwicklung prognostiziert. Weiterhin wird jedoch auch der „worst-case“ simuliert, indem Daten aus dem rcp85 Szenario berücksichtigt werden.

## Fazit

Bei der Betrachtung der Leistungsfähigkeit von Solarstromanlagen dürfen nicht nur technische und chemische Prozesse berücksichtigt werden. Sicherlich hat die altersbedingte Degradation von Solarmodulen einen großen Effekt, jedoch gibt es auch weitere Variablen, die zu einer signifikanten Ertragsänderung führen können. Die hier herausgearbeiteten klimatologischen Parameter sind nicht nur heute von großer Bedeutung für die Abschätzung der Risiken bei der Planung einer Photovoltaikanlage, sondern können über die nächsten Jahrzehnte sehr variabel sein. Vor allem Temperaturveränderungen und das Auftreten langanhaltender Trockenperioden im Sommer werden die möglichen Erträge beeinflussen. Diese potentiellen Veränderungen werden im Rahmen des Projekts anhand verschiedener Szenarien simuliert. Dabei fließen aktuelle Erkenntnisse und etablierte Datensätze aus der Wissenschaft ein, die mögliche Entwicklungen abdecken. Am Ende stehen neben den konkreten Handlungsempfehlungen zu einer Solarstromanlage auch gut verständliche und wissenschaftlich fundierte Tipps und Hinweise zu weiterführenden Überlegungen, die bei der Auslegung und dem Betrieb einer Solarstromanlage in den nächsten Jahrzehnten berücksichtigt werden sollen. Nur so ist eine vollständige Abschätzung sämtlicher Risiken möglich, die eine Investition in saubere Energie mit sich bringt.

## Quellen

- [1] KOEHL, M. ; HECK, M. ; WIESMEIER, S. : *Categorization of weathering stresses for photovoltaic modules*. In: Energy Science & Engineering 2018 6(2), S. 93-111
- [2] KOEHL, M. : *Grundlegende Untersuchungen zur Gebrauchsdaueranalyse von Photovoltaik-Modulen*. 2011, Hagen, Universität, Fakultät für Mathematik und Informatik Elektrotechnik
- [3] EUROPEAN COMMISSION, JOINT RESEARCH CENTRE : *Overview of PVGIS data sources and calculation methods*. 2018. — URL [http://re.jrc.ec.europa.eu ... /pvg\\_static/methods.html](http://re.jrc.ec.europa.eu/pvg_static/methods.html)
- [4] SAYYAH, A. ; HORENSTEIN, M. N. ; MAZUMDER, M. K. : *Energy yield loss caused by dust deposition on photovoltaic panels*. In: Solar Energy. 2014 107, S. 576-604
- [5] European Union: *Photovoltaic Geographical Information System*.  
— URL [http://re.jrc.ec.europa.eu/pvg\\_tools/en/tools.html#PVP](http://re.jrc.ec.europa.eu/pvg_tools/en/tools.html#PVP)
- [6] MAX-PLANCK-GESELLSCHAFT ZUR FÖRDERUNG DER WISSENSCHAFTEN E.V. : *CMIP5 – Klimasimulationen und Datenbasis für den nächsten Weltklimabericht*.  
— URL [https://www.mpg.de/6880685/IB\\_20131](https://www.mpg.de/6880685/IB_20131)
- [7] *RCP-Szenarien*. — URL [http://wiki.bildungsserver.de/klimawandel ... /index.php/RCP-Szenarien](http://wiki.bildungsserver.de/klimawandel/index.php/RCP-Szenarien)



# Nutzen, Vorteile und Optimierung des Energieertrags gewerblicher PV-Anlagen auf Basis heterogener Datenquellen

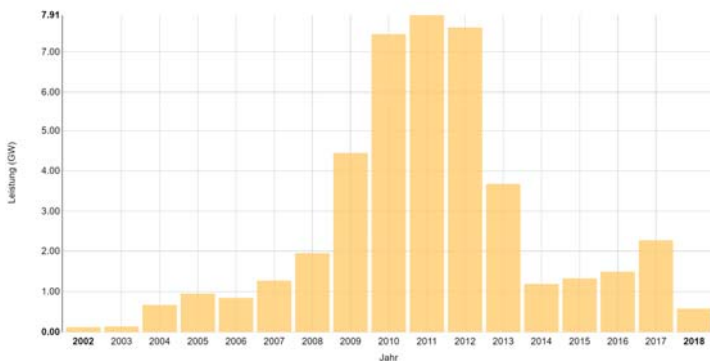
Stefan Rieder

## Inhalt

|                                               |     |
|-----------------------------------------------|-----|
| Einleitung .....                              | 206 |
| Ziel .....                                    | 207 |
| Datengrundlage .....                          | 208 |
| 3D – Potentialdaten.....                      | 208 |
| Klimadaten.....                               | 209 |
| Wetterdaten.....                              | 212 |
| Technische Parameter .....                    | 213 |
| Lastprofile .....                             | 214 |
| Solarpotential Brandenburg an der Havel ..... | 216 |
| Methodik.....                                 | 218 |
| Anwendung.....                                | 222 |
| Fazit & Ausblick.....                         | 223 |
| Quellen.....                                  | 225 |

## Einleitung

Den erneuerbaren Energien kommt bei der Erfüllung der durch die Bundesregierung gesetzten Klimaschutzziele eine besondere Rolle zu. Sie sollen bis 2040 mindestens 65 % (vergl. [1]) des Bruttostromverbrauchs decken. Im Jahr 2017 betrug dieser Anteil 33,1 %, die Photovoltaik sorgte für eine Erzeugung von 7,0 % (vergl. [2]). Der stattfindende Wandel im Mobilitäts- und Wärmesektor wird durch eingesetzte Energieeffizienzmaßnahmen (z. B. in den Bereichen Licht, Dämmung und Geräteeffizienz) nicht kompensiert werden können, was in einem steigenden Energiebedarf innerhalb der nächsten Jahrzehnte resultieren wird. Die Förderung und gleichzeitige Aufklärung des Nutzens von Solarenergie mit dem Ziel eines wieder zunehmenden Ausbaus ist daher nicht nur aus heutigen ökologischen und ökonomischen Aspekten notwendig, sondern auch nötig um die gesteckten politischen Ziele zu erfüllen und kommenden Entwicklungen entgegenzuwirken. Auf den Dächern des Stadtgebiets Brandenburg an der Havel werden (Stand 2017) nur 2,4 % des möglichen Energiepotentials genutzt. Dies zeigt, dass ein weiterer Ausbau nötig aber auch möglich ist. Die folgende Abbildung zeigt den Ausbau der Photovoltaikanlagen in Deutschland in den vergangenen Jahren. Dieser gibt die Änderungen im Erneuerbaren Energien Gesetz (EEG) gut wieder und macht den starken Rückgang des Zubaus in den letzten Jahren sichtbar.



**Abb. 1: Jährlicher Zu- und Rückbau an installierter Netto-PV-Leistung in Deutschland  
(vergl. [2], eigene Erstellung)**

Vor allem auf großen gewerblichen Bauten steckt viel Energiepotential, das bisher nicht beworben wird und daher oft ungenutzt bleibt. Im gewerblichen Bereich stellt

die Energiewende hohe Anforderungen an Wirtschaftlichkeit und Transparenz. Die Investition in eine Solarstromanlage steht im gewerblichen Umfeld meist in direkter Konkurrenz zu anderen Objekten, die Erwartungen an eine hohe Rendite erfüllen sollen. Ökologische Aspekte stehen bei vielen Unternehmen bis heute leider nicht im Vordergrund. Aufgrund stetig sinkender Modul- und Komponentenpreise werden Solarstromanlagen immer günstiger und können bei richtiger Planung und Nutzung durchaus mit hohen Renditen betrieben werden. Bedingt durch die sinkende Einspeisevergütung und steigende Strombezugspreise muss die Ausrichtung der Anlage auf Autarkie und Eigenverbrauch im Vordergrund stehen. Durch die teilweise geringe Akzeptanz ist es weiterhin wichtig, transparente Berechnungen und realistische Vorteile zu errechnen. Aktuelle Verfahren werden diesen Anforderungen nur selten gerecht. Im Teilprojekt **Solar Potential Analytics** (SPA) des Projekts **PreLytica** steht daher die analytische Simulation und Optimierung möglicher Solarstromanlagen in Bezug auf den individuellen Stromverbrauch eines Unternehmens im Vordergrund.

## Ziel

Das Kernziel des Teilprojektes **Solar Potential Analytics** ist es, auf Basis heterogener Datenquellen Handlungsempfehlungen für kleine mittelständische Unternehmen ableiten zu können. Auf dem Markt stehen viele Webseiten und Anwendungen bereit, die die etwaigen Vorteile einer Solarstromanlage auch im gewerblichen Bereich ausgeben. Da diese jedoch oft von Fachfirmen oder Energieversorgern entwickelt und angeboten werden, stehen meist kommerzielle Interessen im Vordergrund. Selten sind die im Hintergrund arbeitenden Algorithmen nicht transparent und die zu Grunde liegenden Berechnungsparameter und Annahmen nicht einsehbar. In dem Projekt SPA sollen genau diese Punkte aufgegriffen und in Form eines neuen Solarrechners implementiert werden. Mit nur wenig Vorwissen und minimalen notwendigen Dateneingaben soll so ein aussagekräftiges Ergebnis erzeugt werden, das dem Nutzer bei seinen weiteren Überlegungen begleitet und sein Vorhaben stützt. Durch den wissenschaftlichen Ansatz sind die ausgegebenen Daten vertrauensbildend und unabhängig von kommerziellen Interessen. Die voll transparente Berechnung und optimierte Ergebnisdarstellung soll neben wissenschaftlichen Erkenntnissen Entscheidungshilfen für Unternehmer liefern und somit einen Beitrag zur Energiewende leisten.

## Datengrundlage

Beschriebenes Ziel des Projektes ist es, Handlungsempfehlungen auf Basis verschiedenster Datengrundlagen abzuleiten. Im Zuge der Aktualität rund um Open Data und INSPIRE sollen neben Daten kommerzieller Anbieter auch Informationen aus offenen Quellen in das Projekt einfließen. Um das gesteckte Ziel zu erreichen, ist es nötig verschiedene technische, klimatologische und meteorologische Datenquellen mit Potentialdaten von Gebäuden zu kombinieren.

### 3D – Potentialdaten

Eine exklusiv für dieses Projekt von dem Unternehmen tetraeder.solar gmbh durchgeführte 3D-Solarpotentialanalyse sämtlicher Bestandsgebäude im Untersuchungsgebiet der Städte Berlin und Brandenburg an der Havel liefert die Basis für alle folgenden Untersuchungen.

Die Erzeugung der Potentialdaten basiert auf Gebäudeumringen und Laser-scandaten, die in der Regel aus Befliegungen erzeugt werden. Im ersten Schritt wird ein mit etwa 20 cm aufgelöstes Oberflächenmodell erstellt, das die nähere Umgebung (500 m) des zu untersuchenden Gebäudes abbildet. Darin erkennbar sind sowohl Dachstrukturen wie Dachformen und -winkel, Dachaufbauten (Schornsteine, Klimaanlage, Gauben) als auch Elemente der näheren Umgebung, die aufgrund ihrer Position und Höhe Schatten auf das zu untersuchende Dach werfen könnten. In einem zweiten Oberflächenmodell, das gröber aufgelöst ist, wird die Topografie des Umlands abgebildet um beispielsweise Berge oder Täler, die für einen verzögerten Sonnenauf- oder untergang verantwortlich sind, berücksichtigen zu können. Mit Hilfe eines Raytracing Algorithmus werden daraufhin für jeden Messpunkt eines Gebäudes die Sonnenstrahlen in stündlichen Intervallen simuliert. Trifft dieser Strahl auf den Punkt, so steigt der potentielle Ertrag, wird der Strahl auf seinem Vektor unterbrochen, so gibt es ein verschattendes Element. Im letzten Schritt folgt die Zusammenlegung homogener, d. h. gleich ausgerichteter, geneigter und ähnlich eingestrahelter Messpunkte zu einer Dachteilfläche (vergl. Abbildung 2). Je nach Jahreseinstrahlung werden diese in Eignungsklassen eingeordnet. Ein Gebäude besteht daher in der Regel aus mehreren Dachteilflächen. Ein typisches Satteldach, meist anzutreffen bei Wohnbebauung, weist somit zwei verschiedene

Flächen auf. Große Industriehallen können je nach Dachstruktur in beliebig viele Dachflächen aufgeteilt werden.

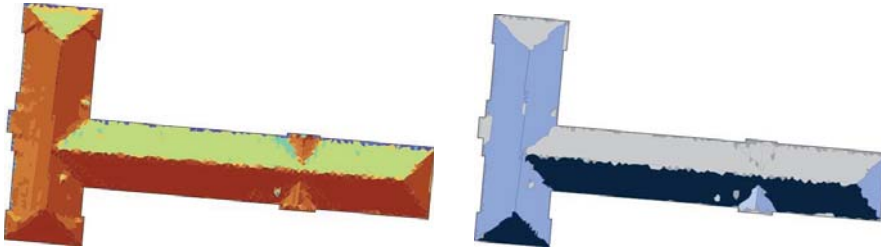


Abb. 2: Solare Einstrahlung auf einem Gebäude (links) und daraus resultierende geeignete Dachteilflächen (rechts)

Der so erzeugte Datensatz steht als SQL-Datenbank mit mehreren Tabellen zur Verfügung, weiterhin können Gebäudeinformationen und Bilder über die interaktive REST-API für einzelne Gebäude bezogen werden. Diese liefert weiterführende Funktionen, weshalb entschieden wurde mit dieser Schnittstelle zu arbeiten. Von vorrangigem Interesse sind die einzelnen Flächen der zu untersuchenden Gebäude sowie deren Dachneigung, Dachausrichtung und Einstrahlung.

Von Seiten des Dienstleisters wird nur die jährliche Strahlung einer Dachteilfläche bereitgestellt, weshalb es für die Fragestellung zwingend nötig ist, diesen präzisen Wert, der die reale Verschattungssituation berücksichtigt, in eine höhere zeitliche Auflösung zu interpolieren.

## Klimadaten

Aufgrund der genannten Notwendigkeit die Jahresstrahlung zu interpolieren, ist eine valide Datenbasis für zeitlich hoch aufgelöste Einstrahlungswerte nötig. Weltweit stehen verschiedene Modelle zur Verfügung, die Einstrahlung auf stündlicher bis viertelstündlicher Basis liefern. Die unten stehende Tabelle gibt einen Überblick über die verschiedenen Varianten. Alle hier aufgelisteten Projekte unterliegen strengen wissenschaftlichen Anforderungen und werden in der Forschung vielfach verwendet. Urheber dieser Daten sind einerseits Hochschulen



(z. B. HelioClim von MINES ParisTech oder Satel-Light der Universität Oldenburg), staatliche Institutionen (NASA SSE) oder wissenschaftliche Einrichtungen von Staatenverbänden (PVGIS der European Commission, Joint Research Centre). Mit Meteororm wird auch ein kommerzieller Anbieter in die Auflistung übernommen.

#### Übersicht verschiedener Datensätze zur solaren Einstrahlung

| Datenbasis  | Abdeckung               | Daten | Datengrundlage       | Zeitl. Aufl. | Kosten | Ind. Ausrichtg. | Zeitreihe    |
|-------------|-------------------------|-------|----------------------|--------------|--------|-----------------|--------------|
| Meteororm   | Weltweit                | API   | Messwerte & Satellit | 60 min       | ja     | Nein            | 1998 - heute |
| Satel-Light | Europa                  | Bild  | Satellit             | 30 min       | nein   | Nein            | 1996 - 2000  |
| HelioClim-1 | Afrika, Europa          | txt   | Satellit             | 24 h         | nein   | Nein            | 1985 - 2005  |
| HelioClim-3 | Afrika, Europa          | txt   | Satellit             | 15 min       | nein   | Ja              | 2004 - 2006  |
| PVGIS 4     | Afrika, Europa          | txt   | Messwerte & Satellit | 15 min*      | nein   | Ja              | 1981 - 1990  |
| PVGIS 5     | Afrika, Europa, Amerika | txt   | Messwerte & Satellit | 60 min       | nein   | Ja              | 2005 - 2015  |
| NASA SSE    | Weltweit                | API   | Satellit             | 180 min      | nein   | Nein            | 2003 - heute |

\* Durchschnittswerte (TMY) für den gesamten Zeitraum

Bei der Auswahl der am besten geeigneten Datenquelle ist die Möglichkeit einer Eingabe eigener Neigungs- und Ausrichtungswinkel Voraussetzung, damit jedes Gebäude mit den korrekten Werten simuliert werden kann. Ein weiterer Vorteil wäre ein offener Zugang zu den Daten, um das eigens gesteckte Ziel, möglichst viele OpenData – Daten zu nutzen, erfüllen zu können.

Den genannten Anforderungen entsprechen mit HelioClim in Version 3 und PVGIS lediglich zwei verfügbare Datensätze. Das in Europa weit verbreitete PVGIS erfüllt mit aktuell (Stand 10/2018) beiden verfügbaren Versionen die Anforderungen. Zum Projektstart (März 2017) gab es leider noch keinerlei Anzeichen für Version 5 von PVGIS, die Veröffentlichung der neuen Datensätze wurde im September 2017 abgeschlossen. Demzufolge fiel die Wahl auf PVGIS 4, da die in HelioClim-3 erzeugten Daten lediglich auf Grundlage von drei Beobachtungsjahren basieren. Aus klimatologischer Sicht ist dieser Zeitraum zu kurz, um für die genannte Fragestellung des Projekts valide und langfristig gültige Aussagen treffen zu können. Im Vergleich dazu bietet PVGIS 4 mit den Grunddaten aus 1981-1990 einen zehn Jahre andauernden Beobachtungszeitraum, was aufgrund der Ausgleiche jährlicher Schwankungen die Aussagekraft deutlich erhöht. Des Weiteren gibt PVGIS 4 keine Zeitreihe aus, sondern übermittelt Daten eines „Standardjahres“ (TMY). Diese Werte bilden somit keinen spezifischen Zeitraum ab, ermöglichen dadurch aber eine Aussage über die langfristigen potentiellen Erträge einer Solarstromanlage. Aus diesen Gründen wurde entschieden PVGIS 4 als Datenquelle zur Interpolation der solaren Einstrahlung zu nutzen. Es steht keine Schnittstelle bereit, über die die Daten bezogen werden können. Stattdessen ist es notwendig die Informationen mittels Screen Scraping auszulesen. Mehr Informationen zu diesem Verfahren sind in dem Artikel *Screen Scraping* von Tobias Kiertscher auf Seite 17 zu finden.

Unter dem Aspekt, dass die Herausgeber von PVGIS im Laufe des Projektzeitraums eine neue Version veröffentlicht haben wird überprüft, wie ein Update auf PVGIS 5 implementiert werden kann. Dies wird vor allem interessant, da die Version 4 Mitte Oktober 2018 offline genommen wird. Da Daten gecached werden, ist das weitere Nutzen des Programms für die festgelegten Standorte Brandenburg an der Havel und Berlin allerdings auch mit den bestehenden Daten möglich. Ein wesentlicher Vorteil der neuen Version sind aktuellere Bezugsjahre, die im Vergleich zu den 30 Jahre alten Daten aus PVGIS 4 etwaige Veränderungen des Klimas abbilden. Dies geht jedoch mit einer Verschlechterung der zeitlichen Auflösung auf Stundenwerte einher. Da die Daten mit zeitlich hoch aufgelösten Lastprofilen simuliert werden, würde dies zu einer deutlichen Reduzierung der erwarteten Genauigkeit führen.

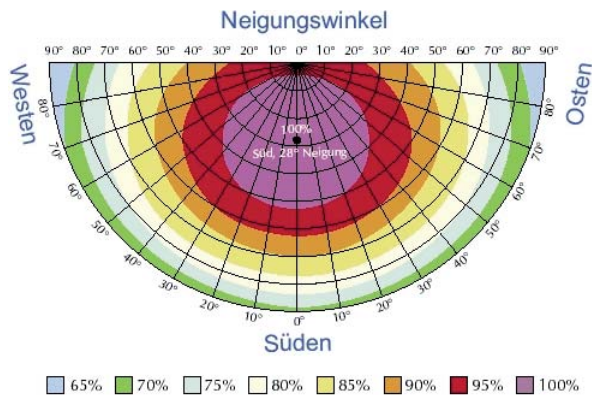
## Wetterdaten

Neben der Interpolation der Ergebnisse der Verschattungsanalyse mit Daten aus Klimanormalperioden ist auch eine genaue Betrachtung der vergangenen Jahre interessant. Durch dieses Vorgehen ist es möglich, „was wäre wenn“-Szenarien zu simulieren und deren Ergebnisse unter wirtschaftlichen Aspekten zu bewerten. So kann beispielsweise der realistische aber nicht realisierte Ertrag einer Solarstromanlage der letzten Jahre auf einem spezifischen Dach exakt bestimmt werden. Dabei würde man keine gemittelten Werte der letzten Jahre und auch keine grob aufgelösten Satellitendaten verwenden, sondern auf stündlich aufgelöste Einstrahlungs- und Wetterdaten der letzten Jahre zurückgreifen. Unter den deutschen öffentlichen Datenanbietern gab es leider keinerlei Datensatz, der diesen Anforderungen gerecht wird. Auch der Deutsche Wetterdienst veröffentlicht lediglich Daten zu seinen Messstationen, die jedoch teilweise lückenhaft und räumlich nicht interpoliert sind. Daher konnte hier leider nicht auf öffentliche bzw. offene Daten zurück gegriffen werden. Kommerzielle Datenlieferanten bieten benötigte Werte jedoch bis zurück in die 1980er Jahre an. Für dieses Projekt relevante Daten der jüngeren Vergangenheit (genutzt ab 2013) werden auf Basis von Mess-, Reanalyse- und Satellitendaten erstellt. Von besonderem Interesse ist allen voran die Einstrahlung, welche maßgeblich von der Bewölkung beeinflusst wird. Diese ist als stündliche Globalstrahlung in Watt pro  $\text{m}^2$  auf horizontaler Ebene verfügbar. Individuell geneigte und ausgerichtete Flächen sind aufgrund der erfassten Basisdaten nicht abrufbar und müssen interpoliert werden. Die Datensätze bilden aufgrund der hohen räumlichen und zeitlichen Auflösung nicht nur klimatologische Eigenschaften oder die über mehrere Tage vorherrschende Witterung ab, sondern beschreiben das sich täglich wechselnde Wetter. Aufgrund des so eingebrachten Realismus werden die Ertrags- und Lastkurven nachvollziehbar, bekannte tages- oder jahreszeitliche Schwankungen zeigen mögliche Probleme im potentiellen Betrieb einer Solarstromanlage auf. Bei der Wahl des Datenlieferanten fiel die Wahl auf die Meteomatics GmbH mit Sitz in Berlin, die ebenfalls das Teilprojekt SPI mit Wetterdaten versorgt. Von Seiten des Dienstleisters wird eine REST-API bereitgestellt.

## Technische Parameter

Nachdem sämtliche Datenquellen und -sätze zur globalen Einstrahlung und Dachflächen im Untersuchungsgebiet definiert wurden, müssen technische Parameter beschrieben werden, die die Umwandlung der solaren Strahlung in nutzbaren Strom ermöglichen. Seit Beginn des weitläufigen Solaranlagenausbaus 2009 haben sich auch die technischen Möglichkeiten enorm weiterentwickelt. Neben Senkungen der Anlagenverluste kam es in den letzten Jahren zu einer deutlichen Steigerung der Modulleistungen bei sinkenden Anlagenkosten. Dementsprechend werden in der Anwendung aktuelle Module und Anlagenwerte in auswählbaren Voreinstellungen definiert, die auf Wunsch jedoch individuell geändert werden können. Das Standardmodul beschreibt ein 270 Wp Modul mit den Abmessungen von 0,99 m \* 1,65 m, was einem Wirkungsgrad von 16,53 % entspricht. Der Verlust durch Wechselrichter und Kabel beträgt voreingestellt 20 %. Somit wird mit einem Gesamtwirkungsgrad der Solarstromanlage von 13,22 % gerechnet.

Neben den rein technischen Parametern spielt auch die Ausrichtung der Solarstromanlage eine entscheidende Rolle. Während diese bei Sattel- oder Walmdächern aufgrund ihrer Neigung fix ist, bieten Flachdächer aufgrund der Aufständermöglichkeit viel Flexibilität. Als flach gilt ein Dach mit einer Neigung von weniger als 10°. Die früher allseits beliebte Ausrichtung der Anlagen gen Süden ist heute nicht in allen Fällen sinnvoll, da dabei zur Mittagszeit sehr viel Strom erzeugt wird, während die Morgen- und Abendstunden jedoch nicht ausgenutzt werden. In Zeiten hoher Einspeisevergütung war dies jedoch höchst effektiv, da durch dieses Vorgehen der maximale Ertrag erzeugt und voll ins Netz eingespeist werden konnte. Unter Berücksichtigung dieser Tatsache wird die Solarstromanlage in der Voreinstellung mit einer Ost-West – Aufständerung simuliert. Bei dieser Art der Auslegung sind die Module gegenläufig angelehnt und zeigen in 90° und 270° Richtung. Die Neigung beträgt dabei 15°. Als zweites steht die klassische Art der Aufständerung zur Auswahl, bei der die Module mit 30° Neigung gen Süden aufgeständert werden.



**Abb.3: Zusammenhang Ausrichtung, Neigung und Ertrag einer Solarstromanlage  
(vergl. [3])**

Die oben stehende Abbildung zeigt, dass in unseren geographischen Breiten eine Aufständerung in Ost-West lediglich einen Minderertrag von 5 % mit sich bringt. Gleichzeitig führt dies jedoch zu einem höheren Eigenverbrauchsanteil, da auch die Morgen- und Abendstunden genutzt werden. Dies gilt unter den wirtschaftlichen Aspekten und dem heute vorherrschenden Förderregime als sehr sinnvoll.

## Lastprofile

Um auf Basis der oben genannten Datensätze und Parameter erzeugten Ertragskurven potentieller Solarstromanlagen die ökonomischen und ökologischen Vorteile errechnen zu können, werden Informationen zu dem Stromverbrauch eines Gebäudes benötigt. Dabei ist es nicht nur wichtig zu wissen wie hoch der Jahresstromverbrauch ist, sondern vor allem die Verteilung dessen auf das Jahr bzw. den Tag. Je nach zeitlicher Variabilität des Verbrauchs kann ein und dieselbe Solarstromanlage wirtschaftlich sinnvoll oder unrentabel sein. Nachdem die Einspeisevergütung in den letzten Jahren an Attraktivität verloren hat, ist es bei steigenden Strombezugspreisen umso sinnvoller die Anlagen auf den Eigenverbrauch zu optimieren. Je mehr auf dem Dach erzeugter Strom selbst verbraucht werden kann, desto wirtschaftlicher ist eine Solarstromanlage. Vor allem bei Flachdächern ist es bei der Anlagenplanung aufgrund der Möglichkeit, Anlagen aufständern zu können sinnvoll das vorherrschende Lastprofil zu berücksichtigen.

Der Bundesverband der Energie- und Wasserwirtschaft e.V. (BDEW) ist zuständig für die Bereitstellung von Standardlastprofilen. Notwendig sind diese für Verbrauchergruppen, die nicht leistungsgemessen sind, sprich weniger als 100'000 kWh Strom pro Jahr verbrauchen. Dazu zählen Privathaushalte, kleine Unternehmen, Gewerbe und Landwirtschaftsbetriebe. Aufgrund fehlender Leistungsmessung ist bei diesen Abnehmern keine genaue Messung des tatsächlichen Verbrauchs möglich, diese lassen sich auf Basis von Prognosen und Stichproben jedoch hinreichend genau errechnen. Die für dieses Projekt relevanten repräsentativen Profile für den Strombezug sind in verschiedenen Ausprägungen je nach Kundengruppe verfügbar. Folgende Lastprofile werden vom BDEW zur freien Verfügung veröffentlicht (vergl. [4], S.5 ff):

- G0: Gewerbe allgemein
- G1: Gewerbe werktags 8 – 18 Uhr
- G2: Gewerbe mit starkem bis überwiegendem Verbrauch in den Abendstunden
- G3: Gewerbe durchlaufend
- G4: Laden / Friseur
- G5: Bäckerei mit Backstube
- G6: Wochenendbetrieb
- L0: Landwirtschaftsbetriebe
- L1: Landwirtschaftsbetriebe mit Milchwirtschaft / Nebenerwerbs-Tierzucht
- L2: Übrige Landwirtschaftsbetriebe
- H0: Haushalt

Sämtliche Profile sind auf 1000 kWh / Jahr normiert und liegen in 15-minütiger Auflösung vor. Getrennt sind die Profile jeweils in verschiedene Jahreszeiten (Winter, Sommer, Übergangszeit), darunter in Tagesausprägungen (Samstag, Sonntag, Werktag). Auf dieser Basis lässt sich ein statistischer Jahresverlauf erzeugen, der ein ganzes Jahr in 15-minütiger Auflösung abbildet. Auch wenn die jeweiligen Profile kein spezifisches Unternehmen widerspiegeln, so lassen sich auf Grundlage dieser Datensätze ausreichend genaue Aussagen über die Lastverteilung

über den Tagesverlauf bzw. die Jahreszeiten treffen. Um die Lastprofile in einen Jahresgang zu bringen wurde ein eigenes Skript geschrieben, dass die verschiedenen Ausprägungen entsprechend prozessiert. Aufgrund der Normierung ist dieses durch Verrechnung mit dem Jahresstromverbrauch variabel nutzbar.

Sollte ein an dem Projektergebnis interessiertes Unternehmen leistungsgemessen sein und Zugriff auf das eigene Lastprofil haben, so ließe sich auch dieses in die Berechnung integrieren.

In der Anfangsphase des Projektes wurde versucht mit regionalen Unternehmen und Stromversorgern einen Datenaustausch zu vereinbaren, um so gemessene Lastprofile zu erhalten. Diese hätten dann dem jeweiligen Gebäude zugewiesen werden können und somit eine hundertprozentige Genauigkeit erzielt. Nicht zuletzt aufgrund hoher Hürden im Bereich des Datenschutzes war dieses Vorhaben jedoch leider erfolglos. Weiterhin lagen nicht so viele Daten im Untersuchungsgebiet vor wie erhofft, auch der Rollout der Smart-Meter war noch nicht ausreichend fortgeschritten bzw. vor zu kurzer Zeit erfolgt um ausreichend Daten zu generieren.

## Solarpotential Brandenburg an der Havel

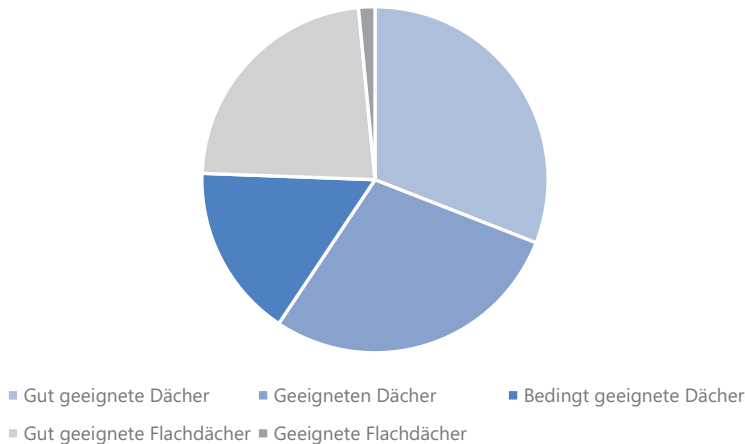
Die bereits erwähnte Potentialanalyse aller Bestandsgebäude im Stadtgebiet Brandenburg an der Havel wurde bereits abgeschlossen und sämtliche Ergebnisse stehen zur Verfügung. Neben den benötigten Daten zu individuellen Gebäuden, die in der Entwicklung des Solarrechners genutzt werden, liefert die 3D-Analyse auch interessante Potentialstatistiken des Stadtgebiets:

Von den insgesamt rund 47'000 untersuchten Gebäuden sind 20'000 Gebäude, was 42,6 % entspricht, für die Erzeugung von Solarstrom geeignet. Auf den ersten Blick erscheint dies nicht viel. Berücksichtigt man jedoch die Tatsache, dass die meisten Garagen und eingemessenen Gartenhütten schon aufgrund ihrer fehlenden Größe ungeeignet sind, relativiert sich diese Ansicht. Auf den Dächern Brandenburgs sind 3.304.000 m<sup>2</sup> für die Erzeugung von Solarenergie geeignet. Geneigte Dächer (steiler als 10°) stellen 51 % dieser Fläche (1.675.500 m<sup>2</sup>), Flachdächer kommen auf 1.628.500 m<sup>2</sup> (49 %).

Betrachtet man die maximal installierbare Leistung (in kWp) auf diesen Flächen wird ersichtlich, dass sich 345.359 kWp auf den Dächern der Stadt Brandenburg an

der Havel installieren ließen. Die Aufteilung in Dachneigung und Eignungsklasse ist der folgenden Abbildung zu entnehmen. Auffällig ist, dass nur knapp 25 % der möglichen Leistung auf die Flachdachflächen entfallen, die rund 50 % der Fläche bilden. Dies resultiert aus der notwendigen Aufständigung der Solarstromanlagen auf diesen Flächen. Durch Wartungsgänge und Abstände, die die Selbstverschattung der Anlage minimieren sollen, ist bei einer Aufständigung gen Süden nur rund ein Drittel der Fläche nutzbar. Höher wäre der Flachdachanteil der installierbaren Leistung, würde man eine Ost-West – Konfiguration annehmen.

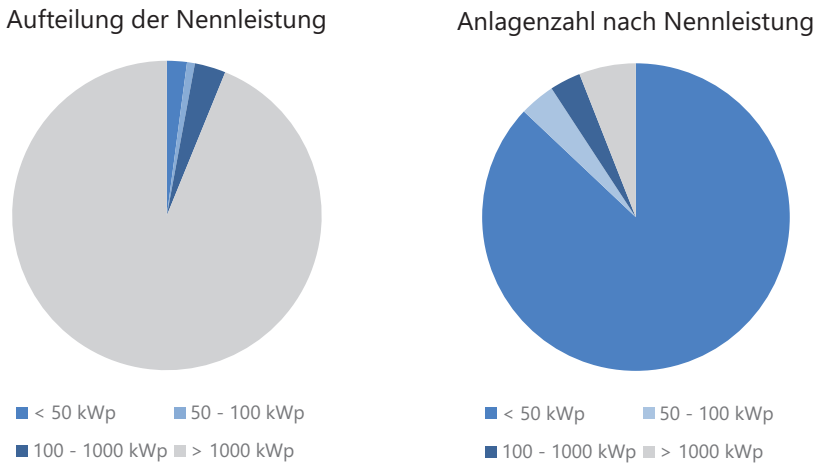
Maximal installierbare Leistung (kWp)



**Abb. 4: Maximal installierbare Leistung (in kWp) in Brandenburg an der Havel, unterteilt nach Eignung und Dachwinkel**

Mit der angenommenen Aufständigung bei Flachdächern gen Süden liegt die Summe des möglichen jährlichen Stromertrags aller Dachflächen bei 289.000 MW. Dies entspricht einer bilanziellen Versorgung von 192.667 Bürgern (bei Annahme von 1500 kWh / Jahr und Bürger). Nimmt man diesen Vollausbau des Potentials an, so ließen sich pro Jahr etwa 133.000 Tonnen CO<sub>2</sub> einsparen. Ein Blick in das Anlagenregister der Bundesnetzagentur (vergl. [5], eigene Bearbeitung) zeigt, dass in den Postleitzahlgebieten der Stadt Brandenburg an der Havel (14770, 14772, 14774, 14776) 399 Solarstromanlagen mit einer Gesamtleistung von 139.891 kWp Nennleistung installiert sind. Die folgende Abbildung zeigt, dass auf 24 Anlagen mit einer Nennleistung über 1000 kWp ganze 131.477 kWp entfallen, während die verbleibenden 375 Anlagen 8.415 kWp stellen.





**Abb. 5: Verteilung der PV-Bestandanlagen in Brandenburg an der Havel nach Anlagengröße. Links: Leistung nach Nennleistung der Anlage, rechts: Anzahl der Anlagen nach Nennleistung.**

Die Grenze von 1000 kWp wird in dieser Betrachtung als Unterscheidungsmerkmal zwischen Dach- und Freiflächenanlagen angenommen. Das beschriebene Verhältnis wird aufgrund der großen Solarparks geprägt, die in das Stadtgebiet fallen, z.B. am Flugplatz Briest. Dass von dem oben genannten Gesamtdachpotential lediglich 2,4 % genutzt werden zeigt, dass hier dringender Handlungsbedarf besteht, um einen weiteren Ausbau zu forcieren.

## Methodik

Nachdem alle für das Projektvorhaben relevanten Datenquellen zur Verfügung stehen, wurde mit der Entwicklung der nötigen Rechenlogik begonnen. Den Anfang machte das Backend, in dem alle Simulationen und Analysen laufen. Neben mathematischen Operationen wurden auch Lösungen für Probleme aufgrund klimatischer Zusammenhänge gesucht. Erst nach Vollendung dieses Bausteins wird mit der Erstellung des nutzerfreundlichen Solarrechners gestartet.

Sämtliche Algorithmen sollen zum Start und automatisierten Ablauf der kompletten Analyse lediglich den Jahresstromverbrauch und der Adresse des Gebäudes erwarten. Der Stromverbrauch ist in der Regel bekannt oder kann aus

Abrechnungen abgelesen werden. Nach Übergabe der Adresse wird diese geokodiert, um geographische Koordinaten (im Referenzsystem WGS84, EPSG 4326) zu erhalten. Mit diesen können alle nachfolgenden Prozesse gestartet werden.

Zuerst wird die API der Solarpotentialdaten angesprochen um zu überprüfen, ob das gewählte Gebäude überhaupt für die Erzeugung von Solarstrom geeignet ist. Ist dies der Fall, so werden Informationen über die nutzbaren Teildachflächen (Größe, Neigung, Ausrichtung, Einstrahlung und Verschattung) abgefragt. Wie bereits erwähnt, wird in diesem Datensatz lediglich ein Wert für die jährliche Einstrahlung ausgeliefert, der jedoch die Verschattungssituation berücksichtigt. Handelt es sich bei einer Dachteilfläche um ein Flachdach (Neigung  $< 10^\circ$ ), so wird die voreingestellte bzw. ausgewählte Aufständerungsart simuliert. Dabei werden Neigung und Ausrichtung des Dachs mit den vordefinierten Werten überschrieben und die Einstrahlung wird dementsprechend an die neuen Werte angepasst. Dies geschieht unter Zuhilfenahme des Datensatzes aus PVGIS und einem Korrekturfaktor, um die Verschattungssituation auf den neuen Wert zu übertragen.

Nach Sortierung der Dachteilflächen nach ihrer Eignung (Kombination aus Größe und Einstrahlung) wird berechnet, wie groß eine Solarstromanlage sein muss, um die Anforderungen des Nutzers zu erfüllen. Zu diesem Zeitpunkt wird ermittelt, wie viele Module nötig sind. Als notwendiger Stromertrag gilt dabei, dass dieser dem Wert des Jahresstromverbrauchs entsprechen soll. In der Regel führt dies zu Autarkie- und Eigenverbrauchswerten um die 30 %. Dies stellt eine solide Ausgangsbasis für spätere Optimierungen dar. Im Anschluss wird geprüft, ob es eine gut geeignete Fläche gibt, auf die die ermittelte Anlage aufgrund des Flächenbedarfs platzierbar ist. Ist dies aufgrund hohen Platzbedarfs oder aufgrund einer hohen Fragmentierung des Gebäudedachs nicht möglich, so muss die Solarstromanlage auf mehrere Dachteilflächen verteilt werden. Alle so aktiven Dachteilflächen werden für die nächsten Schritte zwischengespeichert.

Für jede Dachteilfläche, die eine Anlage oder einen Teil der Anlage beinhaltet, muss die erfasste Jahreseinstrahlung nun auf Werte im Viertelstundentakt verrechnet werden. Die für diese Interpolation notwendigen Daten stammen aus PVGIS. Zur Verfügung stehen für jeden Standort in Europa Einstrahlungsdaten für Monate und für jede Neigungs- und Ausrichtungskombination zusätzlich Viertelstundenwerte (in  $\text{W/m}^2$ ). Letztere müssen für jeden Monat einzeln abgefragt werden und bilden

einen typischen mittleren Tag im Monat ab. Um eine einheitliche Einheitenbasis zu erhalten, werden die Daten in kWh umgerechnet. Aus diesen Werten werden die PVGIS – Einstrahlungswerte in prozentualen Anteilen auf die Monate und die Viertelstunden verteilt. Berücksichtigung findet dabei die Tatsache, dass Monate unterschiedlich lang sind. Somit entsteht ein Jahresverlauf mit einem Stichtag pro Monat, der für eine Standort-, Neigung- und Ausrichtungskombination gültig und normiert ist. Es ist somit bekannt, wie viel Prozent der Jahreseinstrahlung in welcher Viertelstunde eines Monats zur Verfügung steht. Nach Multiplikation dieser Faktoren mit den Jahreswerten aus der Verschattungsanalyse entsteht ein genauer Verlauf, der alle notwendigen und verfügbaren Einflussparameter berücksichtigt. Da auch die Verschattungsanalyse, die eine komplette Simulation des Sonnenverlaufs über ein Jahr beinhaltet, anhand von PVGIS kalibriert wurde, liegt dem gesamten Vorgehen eine gemeinsame Datenbasis zugrunde.

Die Wetterdaten liegen als Ergebnis der API-Anfrage in stündlicher Auflösung in kWh/m<sup>2</sup> vor. Dabei gilt dieser Wert für eine horizontale Fläche, sämtliche Winkelkombinationen müssen darauf basierend interpoliert werden. Dazu werden anhand der TMY-Daten („*typical meteorological year*“) monatliche und stündliche Korrekturfaktoren zwischen der Einstrahlung auf eine horizontale und auf eine entsprechend der zu planenden Solaranlage geneigten Fläche gebildet. Mit Hilfe derer lassen sich die gemessenen Daten auf die Modulflächen übertragen. Im Gegensatz zu dem „gemittelten“ Strahlungsverlauf der TMY-Daten stehen nun reale Wetterdaten in stündlicher Auflösung verschiedener Jahre zur Verfügung. In dem Projekt werden die Jahre 2013 bis 2017 verwendet, eine Erweiterung bis 2018 wird im nächsten Projektjahr stattfinden. Dadurch lassen sich konkrete Fragestellungen wie „Wie viel Euro hätten wir mit Hilfe einer Solarstromanlage im Jahr 2017 einsparen können?“ beantworten.

Um von dem Klimaparameter, der Einstrahlung, auf den möglichen Stromertrag zu kommen, werden die bereits beschriebenen definierten technischen Rahmenparameter genutzt. Da die für die erste Anlagenkonfiguration benötigte Modulanzahl feststeht, wird auf Basis dessen für jede Dachteilfläche die Kollektorflächen berechnet. Mit Multiplikation des Gesamtwirkungsgrad der Anlage von 13,22 % und der Erzeugerfläche ergibt sich so der potentielle Stromertrag für jede Stunde.

Um letztendlich an den Rechner gestellte Fragen und den allgemeinen Nutzen einer Solarstromanlage berechnen zu können, wird im weiteren Verlauf das im Vorfeld aufbereitete Lastprofil eingelesen. Dieses liegt ebenfalls in viertelstündiger Auflösung in kWh vor und berücksichtigt die verschiedenen Tagesausprägungen (Werktag, Samstag, Sonntag). Bei der Verrechnung des Lastprofils mit den TMY-Daten, die mit Hilfe von PVGIS erzeugt wurden, ist Vorsicht geboten, da nicht einfach der 15. eines jeden Monats genommen werden kann. Stattdessen muss ein gewichtetes Mittel der verschiedenen Tage erstellt werden, das repräsentativ gültig ist. Bei dem Abgleich mit den stundengenauen Wetterdaten ist dies hingegen nicht notwendig.

Nun liegen Zeitreihen vor, die Auskunft über den Netzbezug oder die Netzeinspeisung geben. Über die Jahressumme dieser Daten lässt sich daher ermitteln, wie viel Kilowattstunden direkt aus der Solarstromanlage verbraucht und wie viel aufgrund des zeitlichen Versatzes von Ertrag und Verbrauch in das Netz eingespeist werden. Eigenverbrauchs- und Autarkiewerte lassen sich daraus ableiten. Da das ursprüngliche Lastprofil bekannt ist, lässt sich die Differenz bilden, die in einem wirtschaftlichen Vorteil enden kann. Eingespartes Geld wird im aktuellen Stand der Entwicklung bewusst nicht berücksichtigt, da die Anlagenpreise je nach Anbieter sehr variabel und von weiteren, aus der Ferne nicht abschätzbaren, Parametern abhängen. Eingesparte und eingespeiste Strommengen können jedoch in finanziellen Vorteilen klar ausgegeben werden. Der ökologische Vorteil in Form von CO<sub>2</sub> – Einsparung kann auf Basis der Jahresstromproduktion der Anlage errechnet werden. Laut Umweltbundesamt (vergl. [6]) erzeugt eine Kilowattstunde im Strommix 2017 ohne Berücksichtigung des Stromhandelssaldos 489 g/kWh.

## Anwendung

Um Akzeptanz im Bereich der Solarenergie zu schaffen und Interesse an der Thematik zu wecken, ist eine nutzerfreundliche Anwendung gefragt, die keine weiteren Hemmnisse aufbaut. Ein rein technisches Skript würde dieser Anforderung nicht gerecht werden. Aus diesen Gründen wurde entschieden, eine den gängigen Standards der Webentwicklung entsprechende Anwendung zu entwickeln, in der Nutzer geführt und informiert werden.

Da im Falle einer Webanwendung auch das Thema Datenschutz relevant ist, wird Wert darauf gelegt so wenig persönliche Daten wie möglich zu erfassen. Um jedoch sinnvolle Simulationen erstellen und somit hilfreiche Handlungsempfehlungen ableiten zu können, ist es nötig die Adresse des Gebäudes sowie den jährlichen Stromverbrauch zu kennen. Informationen zum Lastprofil und Präferenzen der Anlage sind optional, können das Ergebnis jedoch verbessern. Auf Tracking Tools oder Kontaktdaten wird bewusst verzichtet, eingegebene personenbezogene Daten werden nicht gespeichert.

Die Webanwendung ist noch nicht fertig entwickelt, dennoch kann das Grobkonzept kurz skizziert werden. Nach dem initialen Aufruf des Solarrechners wird der Nutzer nach der Adresse des Gebäudes gefragt, für das eine Solarstromanlage simuliert werden soll. Nach Eingabe und Bestätigung dieser auf einer Karte wird das allgemeine Solarpotential des Gebäudes aufgezeigt. Neben Werten zur maximal möglichen Stromerzeugung, der geeigneten Fläche und der maximal möglichen Anlage wird das Einstrahlungsbild des Daches aus der 3D – Analyse angezeigt. Dies suggeriert bereits, dass in dieser Anwendung keine simplen Annahmen getroffen werden, sondern Berechnungen anhand stichfester Daten durchgeführt werden. Auf dem Bild lassen sich auch für Laien gut eingestrahlte und verschattete Bereiche erkennen, je nach Dachbeschaffenheit sind auch Dachaufbauten (z. B. Klimaanlage) erkennbar. Nach Eingabe des Jahresstromverbrauchs und Bestätigung (oder Anpassung) der voreingestellten Parameter (Lastprofil, Aufständigung, Modul) wird die Analyse und Anlagensimulation gestartet. Ist diese nach wenigen Sekunden abgeschlossen, wird der Nutzer auf die Ergebnisseite weitergeleitet. Hier können bereits Werte zur Wirtschaftlichkeit der Anlage abgelesen werden. Weiterhin besteht die Möglichkeit, den zeitlichen Verlauf von Lastprofil, Solarertrag sowie deren Zusammenspiel graphisch visualisieren. So ist nachvollziehbar, zu welchen

Zeiten (morgens und abends, aber auch auf jahreszeitlicher Ebene) der Stromertrag der Anlage genutzt werden kann und wann der benötigte Strom aus dem Netz bezogen werden muss. Neben den rein visuellen Elementen können auf dieser Seite die Berechnungsparameter (Anlagengröße, Richtung der Aufständigung, Module) interaktiv angepasst und mit den vorherigen Simulationsläufen verglichen werden. Dies erzeugt Verständnis in diesem technischen Bereich und ein Gefühl für eine sinnvolle Anlagenkonfiguration. Die berechneten Daten und Ergebnisse können am Ende je nach Vorliebe gespeichert oder ausgedruckt werden.

## Fazit & Ausblick

Nach der genauen Projektkonzeption sowie der Organisation sämtlicher benötigter Daten hat das Projekt nun seine nächste Phase, in der die beschriebenen Datensätze kombiniert und zu allgemein verständlichen Informationen prozessiert werden, weitestgehend abgeschlossen. Während die meisten Datenquellen konsistent sind oder extra für dieses Forschungsprojekt angefertigt wurden, stellen offene Daten ein Problem dar. Im Bereich der Wetterdaten war es leider nicht möglich an entsprechende Datensätze zu gelangen, da diese benötigten Daten in Deutschland von offizieller Seite nicht verfügbar sind. Eine valide interpolierte Datenbasis stellen hier nur kommerzielle Anbieter bereit. Bei den Einstrahlungsdaten, die für das Projektvorhaben fundamental sind, steht im Jahr 2018 eine Umstellung an. Das vorgestellte und bisher genutzte PVGIS4 wird in naher Zukunft abgestellt und auf Version 5 gehoben. Damit geht leider auch eine Änderung des Funktionsumfangs einher. Darum muss im Rahmen des Projektes dieses Update noch auf Anwendbarkeit geprüft werden, da diese Umstellung mit Vor- aber auch Nachteilen verbunden ist.

Neben Anpassungen an sich ändernde Webdienste soll eine eigenständige Optimierung der Solarstromanlage auf die vom Nutzer eingegebenen Daten erfolgen. Ziel ist eine möglichst hohe Stromeinsparung auf Basis der TMY – Daten, was in einer guten Wirtschaftlichkeit resultiert. Solarstromanlagen werden heute mit einer realen Laufzeit von über 20 Jahren ausgelegt und geplant. In Zeiten, in denen das Thema Klimawandel stets präsent ist, wäre es natürlich interessant, ob und in welchem Umfang der Ertrag von Solarstromanlagen vom Klimawandel in Brandenburg betroffen ist. Daten eines europäischen Klimamodells werden

dahingehend auf Veränderungen untersucht, die für die regenerative Stromerzeugung durch Photovoltaik relevant sind. Dadurch kann beispielsweise abgeleitet werden, ob es in Zukunft vermehrt Bewölkung im Sommer gibt oder ob langwierige Trockenphasen dominieren. Somit stehen auch Informationen über die Laufzeit zur Verfügung, die eine Risikoabschätzung der Investition ermöglichen. Mehr Informationen zu diesem Aspekt finden sich im Artikel *Leistungsfähigkeit von Solarstromanlagen* (S. 193).

Die beschriebenen Daten zu der aktuellen Situation des Ausbaus der Solarenergie auf den Dächern der Stadt Brandenburg zeigt, dass in diesem Bereich der Entwicklung noch Handlungsbedarf besteht. Mit dem vorgestellten Projekt soll ein Beitrag geleistet werden, der eine aktuelle Problematik aufgreift und zu mehr Aufklärung und transparenteren Informationen sowie Daten führt. Damit können mittelständische Unternehmen eigene Vorteile für sich generieren, aber auch einen ökologischen Beitrag zur Energiewende leisten.

## Quellen

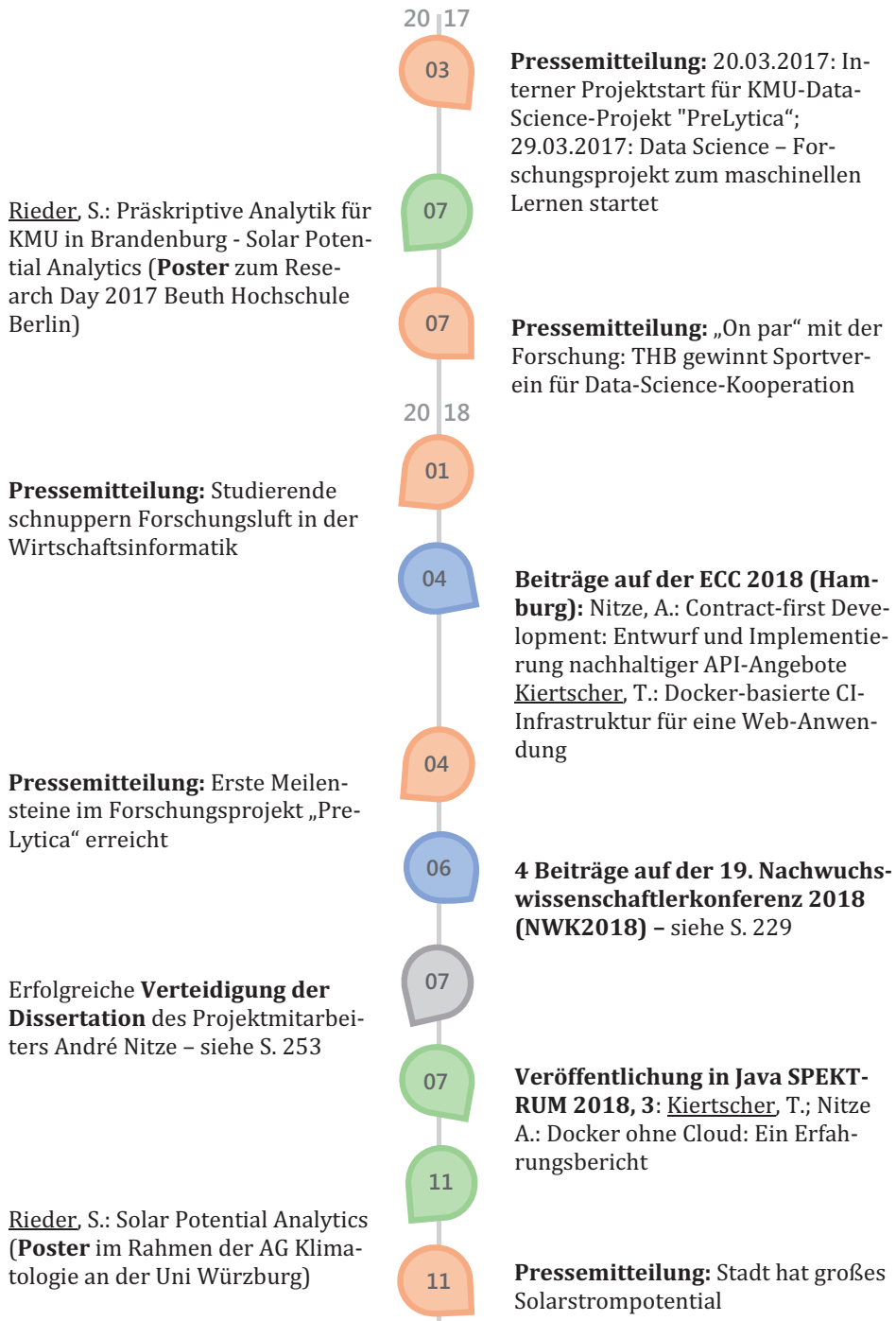
- [1] BMWI : *Das Erneuerbare-Energien-Gesetz*. 2018. — URL [https://www ...  
.erneuerbare-energien.de/EE/Redaktion/DE/Dossier/eeg.html ...  
?cms\\_docId=71802](https://www.erneuerbare-energien.de/EE/Redaktion/DE/Dossier/eeg.html?cms_docId=71802)
- [2] FRAUNHOFER-ISE : *Stromerzeugung in Deutschland in 2017*. 03/2018.  
— URL [https://www.energy-charts.de/energy\\_pie\\_de.htm?year=2017](https://www.energy-charts.de/energy_pie_de.htm?year=2017)
- [3] TST ARGE SOLARSTROM : *Neigung und Ausrichtung*. — URL [http://www.pv- ...  
ertrag.com/neigung-und-ausrichtung/](http://www.pv-ertrag.com/neigung-und-ausrichtung/)
- [4] VDEW : *Anwendung der Repräsentativen VDEW-Lastprofile step-by-step*. 2000.  
— URL [https://www.bdew.de/media/documents/2000131\\_Anwendung- ...  
repraesentativen\\_Lastprofile-Step-by-step.pdf](https://www.bdew.de/media/documents/2000131_Anwendung-repraesentativen_Lastprofile-Step-by-step.pdf)
- [5] BUNDESNETZAGENTUR FÜR ELEKTRIZITÄT, GAS, TELEKOMMUNIKATION, POST UND EISENBAHNEN : *Archivierte EEG-Vergütungssätze und Datenmeldungen*. 2018.  
— URL [https://www.bundesnetzagentur.de/DE/Sachgebiete/Elektri ...  
zitaetundGas/Unternehmen\\_Institutionen/ErneuerbareEnergien ...  
/ZahlenDatenInformationen/EEG\\_Registerdaten/ArchivDatenMeldgn ...  
/ArchivDatenMeldgn\\_node.html](https://www.bundesnetzagentur.de/DE/Sachgebiete/ElektrizitaetundGas/Unternehmen_Institutionen/ErneuerbareEnergien/ZahlenDatenInformationen/EEG_Registerdaten/ArchivDatenMeldgn/ArchivDatenMeldgn_node.html)
- [6] UMWELTBUNDESAMT : *Strom- und Wärmeversorgung in Zahlen*. 2018.  
— URL [https://www.umweltbundesamt.de/themen/klima-energie ...  
/energieversorgung/strom-waermeversorgung-in-zahlen ...  
?sprungmarke=Strommix#Kraftwerke](https://www.umweltbundesamt.de/themen/klima-energie/energieversorgung/strom-waermeversorgung-in-zahlen?sprungmarke=Strommix#Kraftwerke)





## Timeline von Publikationen und Pressemitteilungen

Alle Pressemitteilungen und weitere Informationen zu den Veröffentlichungen sind abrufbar unter <https://prelytica.th-brandenburg.de/> unter Neuigkeiten und Publikationen.





# Beiträge zur Nachwuchswissenschaftlerkonferenz 2018

## 19. NWK

### Inhalt

|                                                                                                                                       |     |
|---------------------------------------------------------------------------------------------------------------------------------------|-----|
| Automated User Interface Testing in Low-Scaled Web Development Projects ....                                                          | 230 |
| PreLytics SPI – Sport Performance Insight .....                                                                                       | 232 |
| Entwurf und prototypische Implementierung eines Präskriptionssystems<br>für den Leistungssport .....                                  | 234 |
| Methodische Beschreibung einer Optimierung des Energieertrags von<br>gewerblichen PV-Anlagen auf Basis heterogener Datenquellen ..... | 236 |

# Automated User Interface Testing in Low-Scaled Web Development Projects

Kilian Friedrich, André Nitze

In diesem Beitrag soll dargestellt werden, wie Software-Qualitätssicherung in kleinen Entwicklungsprojekten realisiert werden kann. Die grundlegende Forschungsfrage dabei ist, wie Software-Tests eingesetzt werden können, um Risiken wie Nichtverfügbarkeit oder funktionale Einschränkungen zu minimieren (vgl. [1]). In einer Fallstudie wurden dafür die bisher manuell durchgeführten Tests automatisiert. Die Ausgangssituation bestand also darin, dass vorher keine Implementierung von Nutzerakzeptanztests vorhanden war. Dazu wurden zunächst verschiedene Test-Frameworks evaluiert. Die Evaluation wurde anhand der Anforderungen des konkreten Projekts und den darin verfügbaren Ressourcen durchgeführt. Es wurden kommerzielle Produkte und Open-Source-Lösungen untersucht. Eine zentrale Anforderung war dabei, dass eine End-to-End-Test-Lösung für die vorliegende Client-Server-Architektur in einer heterogenen Technologi Landschaft aus mobilen und Desktop-basierten Endgeräten gefunden werden musste. Die Vorgehensweise wurde dokumentiert, Herausforderungen werden diskutiert und auf Replizierbarkeit hin bewertet. Weiterhin wurde geprüft, wie die erstellten Test-Suiten in eine Werkzeugkette für kontinuierliche Auslieferung integriert werden können. Die Erkenntnisse aus der Fallstudie deuten darauf hin, dass geschäftliche Risiken mit Testautomatisierung wesentlich verringert werden können. Mit der Umsetzung im Test-Framework "Cucumber", in dem ein verhaltensgetriebener Entwicklungsansatz (vgl. [2]) verfolgt wird, ergaben sich mehrere positive Effekte, darunter die Erhöhung der Systemverfügbarkeit und eine gesteigerte Entwickler-Awareness für Komponenten mit vielen Abhängigkeiten. Weiterhin ist anzunehmen, dass die höhere Verfügbarkeit des Systems zu einer Steigerung des Vertrauens beim Endnutzer führt. Die gesammelten Erfahrungen wurden strukturiert für ähnliche Projekte aufbereitet. Abschließend wird diskutiert, welche praktische Relevanz theoretische Testansätze grundsätzlich in einem beschränkten Projektrahmen besitzen.

Das vorliegende Projekt wird finanziert durch den Europäischen Fonds für regionale Entwicklung (EFRE) und die Technische Hochschule Brandenburg.

```
...F...[]  
..  
  
7 passing (21.80s)  
1 skipped  
1 failing  
  
1) Click on the "Nutzungsvereinbarung"9 I click on "Nutzungsvereinbarung":  
An element could not be located on the page using the given search parameters.  
running chrome
```

Abb.1: Ergebnis einer Test-Prozedur mit Cucumber

## Quellen

[1] Peter Liggesmeyer. Software-Qualität: Testen, Analysieren und Verifizieren von Software. 2. Aufl. Heidelberg: Spektrum Akademischer Verlag, 2009.

[2] Dan North. Behavior Modification: The evolution of behavior-driven development. 2006 <http://www.stickyminds.com/better-software-magazine/behavior-modification> 17.11.2015.

# PreLytica SPI – Sport Performance Insight

Sophia Schubert, André Nitze

SPI (Sport Performance Insight) ist Teilprojekt des Projekts PreLytica - Prescriptive Analytics. Kardinale Frage des Forschungsprojekts: Die Entwicklung von Analysesystemen und Prognosemodellen, welche der gesundheitlichen Vorsorge und dem sportlichen Erfolg dienen. Durch die systematische Anwendung statistischer und präskriptiver Verfahren werden Trends erkannt, Querverbindungen aufgezeigt (Mustererkennung) und Handlungsempfehlungen gegeben, um Entscheidungen im Training zu unterstützen. Genau hier zeigt sich auch der Unterschied zu anderen gängigen Verfahren wie deskriptiven, diagnostischen und prädiktiven Verfahren. Es werden zusätzlich Handlungsempfehlungen gegeben und es besteht die Möglichkeit der Trendbeeinflussung. Bei dem Projektpartner, einem Golfclub, startet die Empirie. Es werden eigene und öffentlich verfügbare Daten (z. B., Turnierergebnisse) erhoben. Dazu zählen personenbezogene Daten wie Alter, Geschlecht, Größe und Gewicht aber auch die Trainingsaktivität, gespielte Distanzen sowie Schlagzahlen (siehe Abbildung 1). Eine erste Analyse könnte sein: Lassen sich Sportler in Cluster einteilen, bspw. durch die Variablen "Häufigkeit des Fitnesstrainings", "Körpergewicht" und "Trainingsaktivität"? Weiterhin werden Abhängigkeiten zwischen Leistungskennzahlen und Alter und Geschlecht und deren Wirkrichtung untersucht. Mittels diverser Instrumente der Statistik wie Data Mining und Clusterwise Regression Models soll ein Prognosemodell formuliert werden (vgl. [1], S.176ff). Anders als bei Maennig und Wellbrock (Vorhersage olympischer Medaillen für teilnehmende Länder; vgl. [2], S.131ff) und den Methoden bei diversen Sportwetten steht hier nicht eine Erfolgsvorhersage der Sportler im Vordergrund, sondern deren Leistungsverbesserung durch Trainingsoptimierung. Weiterhin sollen Ansätze zum maschinellen Lernen, Monte Carlo Simulationen und Szenariotechniken Anwendung finden. Die Erkenntnisse sollen sich auf weitere Sportarten und Datenquellen (z. B. aus Fitnesstrackern, anderen Wearables und Health-Apps) übertragen lassen. Das Projekt wird finanziert durch den Europäischen Fonds für regionale Entwicklung (EFRE).

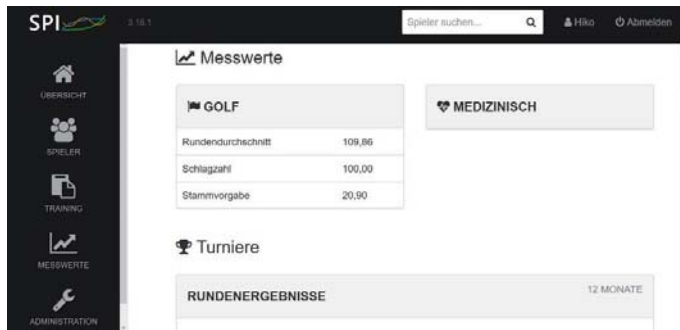


Abb.1: SPI-Webversion, Ansicht als Trainer

## Quellen

- [1] Konradt, U.; Albers, S.; Wolf, J.(Hrsg.) (2009): Methodik der Empirischen Forschung. 3.Aufl. Wiesbaden: Gabler.
- [2] Maennig, W.; Wellbrock, C. (2008): Sozioökonomische Schätzungen olympischer Medaillengewinne. Sportwissenschaft, Vol.38, S.131–148.



# Entwurf und prototypische Implementierung eines Präskriptionssystems für den Leistungssport

Daniel Kiertscher

Im Leistungssport sind Folgen von Entscheidungen über die Auswahl von Sportlern und Trainingsinhalten sehr unsicher und lassen sich nicht revidieren. Softwaregestützte Darstellungen der Entscheidungsgrundlagen können helfen, einen Teil dieser Unsicherheiten zu mindern. Jedoch sind hierfür die Möglichkeiten sehr vielfältig. Zudem sorgen variierende Herausforderungen bei der Datenerfassung und Aufarbeitung und eine wachsende Anzahl an Fragestellungen und Analyseansätzen für eine häufige Aktualisierung des Analyseprozesses. Dies erschwert es, Spiel- und Trainingsverläufe zur Entscheidungsfindung effektiv und effizient zu betrachten. Für die Nachwuchsförderung im Golfsport werden Analysen dieser Art aufgrund der stark begrenzten Anwenderzahl in manueller und schwer reproduzierbarer Weise erstellt. In dieser Arbeit wurde deshalb der Erkenntnisgewinnungsprozess für Golftrainer in mehrere Phasen aufgeteilt, welche es erleichtern, das breite Spektrum der erforderlichen Kompetenzen und Ressourcen zu integrieren. Hierbei haben die Kommunikation zwischen Analysten und Trainern und die Operationen auf den Daten einen wesentlichen Einfluss auf die Qualität der Analysen. Deshalb müssen hier trotz der notwendigerweise automatisierten und minimal invasiven Datenerfassung im Trainingsbetrieb Möglichkeiten zur interaktiven Erkenntnissuche unterstützt werden. Zur Konzeption und Visualisierung einer Präskription für Golftrainer wurden in einem Prototyp Trainings- und Wettkampfdaten zusammengeführt und analysiert. Dabei wurde festgestellt, dass Transformieren und Visualisieren mit hoher Ausdrucksstärke es ermöglichen, wiederholt zügig Erkenntnisse aus den lückenhaften Datenbeständen zu gewinnen. Zentraler Bestandteil der Vorgehensweise ist der interaktive Einsatz einer Programmiersprache zur Visualisierung und Gestaltung des Domänenmodells vor der Analyse der

statistischen Eigenschaften der Daten. Abbildung 1 zeigt mögliche visuelle Ausgaben für die Rolle des Programmierers, des Analysten und des Trainers. Die aus dem Projekt hervorgegangenen Erkenntnisse zeigen, dass sich mit dieser Methode Handlungsempfehlungen ermitteln und kommunizieren lassen, welche im Leistungssport helfen, tragfähige Entscheidungen zu treffen. Das vorliegende Projekt wird finanziert durch den Europäischen Fonds für regionale Entwicklung (EFRE) und die Technische Hochschule Brandenburg.

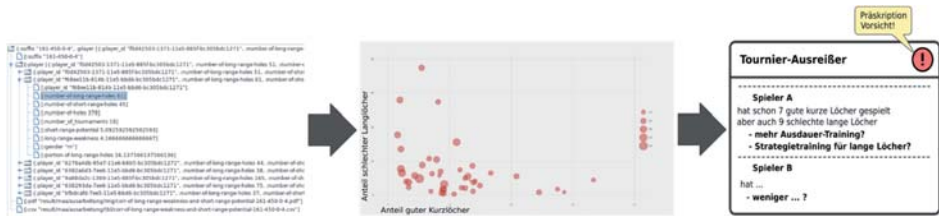


Abb. 1: Informationsfluss vom Programmierer über den Analysten zum Trainer

# Methodische Beschreibung einer Optimierung des Energieertrags von gewerblichen PV-Anlagen auf Basis heterogener Datenquellen

Stefan Rieder, André Nitze

Den erneuerbaren Energien kommt bei der Erfüllung der durch die Bundesregierung gesetzten Klimaschutzziele eine besondere Rolle zu. Sie sollen bis 2040 mindestens 65 % der Bruttostromerzeugung erzeugen. Im Jahr 2017 betrug dieser Anteil 33,1 %, die Photovoltaik sorgte für eine Erzeugung von 6,1 %. Vor allem auf großen gewerblichen Bauten steckt jedoch viel Potential, das bisher nicht beworben wird und daher oft ungenutzt bleibt. Die Energiewende im gewerblichen Bereich stellt hohe Anforderungen an Wirtschaftlichkeit und Transparenz. Aktuelle Verfahren werden diesen Anforderungen nicht gerecht. Im Vordergrund des Projekts steht die analytische Simulation und Optimierung möglicher PV-Anlagen in Bezug auf den individuellen Stromverbrauch eines Unternehmens. Diese Analyse soll in hoher räumlicher und zeitlicher Auflösung durchgeführt und die erhaltenen Ergebnisse nachvollziehbar präsentiert werden. Um dies zu erreichen, gilt es verschiedene themenspezifische Daten aus unterschiedlichen Quellen miteinander zu verknüpfen. Im ersten Schritt werden benötigte Daten erfasst, dazu zählen neben Lastprofilen und gemessenen sowie interpolierten Wetter- und Klimadaten eine exklusiv für dieses Projekt durchgeführte 3D-Solarpotentialanalyse (vgl. Abb.

1) sämtlicher Bestandsgebäude im Untersuchungsgebiet der Städte Berlin und Brandenburg an der Havel. In der so erzeugten, 8 GB großen Datenbank, finden sich detaillierte Informationen zu solarer Einstrahlung (vgl. Abb. 2) und Dachteilflächen (vgl. Abb. 3) und solarer Einstrahlung. Nach Validierung der Datensätze werden diese auf Basis physikalischer sowie klimatologischer Gegebenheiten und unter Zuhilfenahme statistischer Verfahren zusammengeführt. Auf Grundlage des so erstellten Datenpools werden weiterführende Algorithmen entwickelt, die

spezifische optimierende Analysen für PV-Anlagen ermöglichen und individuelle Prognosen liefern, um potentielle wirtschaftliche Vorteile aufzuzeigen. Zur transparenten Darstellung der Ergebnisse wird im letzten Schritt ein interaktives Webportal erstellt. Durch die voll transparente Berechnung und optimierte Ergebnisdarstellung soll das Projekt neben wissenschaftlichen Erkenntnissen Entscheidungshilfen für Unternehmer liefern und somit einen Beitrag zur Energiewende leisten. Das vorliegende Projekt wird finanziert durch den Europäischen Fonds für regionale Entwicklung (EFRE) und die Technische Hochschule Brandenburg.



Abb. 1: Ausschnitt aus der durchgeführten Solarpotentialanalyse

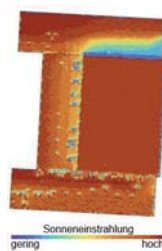


Abb. 2: Solare Einstrahlung



Abb. 3: Geeignete Dachteilflächen



# Pressemitteilungen

## Inhalt

|                                                                                       |     |
|---------------------------------------------------------------------------------------|-----|
| Interner Projektstart für KMU-Data-Science-Projekt "PreLytica" .....                  | 240 |
| Data Science – Forschungsprojekt zum maschinellen Lernen startet .....                | 241 |
| „On par“ mit der Forschung: THB gewinnt Sportverein für Data-Science-Kooperation..... | 243 |
| Studierende schnuppern Forschungsluft in der Wirtschaftsinformatik .....              | 245 |
| Erste Meilensteine im Forschungsprojekt „PreLytica“ erreicht .....                    | 246 |
| Dächer in Brandenburg an der Havel mit Potential .....                                | 248 |

## Interner Projektstart für KMU-Data-Science-Projekt "PreLytica"

20.03.2017

Am 16.03.2017 wurde das EU-geförderte Projekt "Präskriptive Analytik im Land Brandenburg (PreLytica)" mit einem internen Team-Kickoff an der Technischen Hochschule Brandenburg (THB) begonnen.

Die Arbeit am vom Europäischen Fond für regionale Entwicklung (EFRE) geförderten Projekt wurde offiziell bereits am 01.03.2017 aufgenommen. Das Projekt untergliedert sich in zwei Teilprojekte, in denen Methoden des maschinellen Lernens auf große Datenmengen angewandt werden sollen. Ziel ist jeweils die automatisierte Ableitung relevanter Handlungsvorschläge für die Anwender.

Das Teilprojekt **Sport Performance Insight (SPI)** beschäftigt sich dabei mit der Analyse und Auswertung von medizinischen Daten im Leistungssport zur Verbesserung individueller Trainingsleistungen. Im Teilprojekt **Solar Potential Analytics (SPA)** geht es um die Potenzialanalyse von Dachflächen aus Solarkatasterdaten. Beide Projekte werden durch ihre konzeptionellen Grundlagen und Zielstellungen vereint.

Unter Mitwirkung von acht Mitarbeitern soll innerhalb von drei Jahren für jedes Teilprojekt ein anwendungsnaher Prototyp entstehen, in dem neue Erkenntnisse, Technologien und Methoden beispielhaft demonstriert werden können.

Aktuelle Informationen zum Projekt, Publikationen und Zwischenergebnisse werden auf <https://prelytica.th-brandenburg.de> veröffentlicht.

# Data Science – Forschungsprojekt zum maschinellen Lernen startet

29.03.2017

An der Technischen Hochschule Brandenburg (THB) beginnt ein Forschungsprojekt im Bereich „Data Science“. Die Wissenschaftler mit dem „Sexiest Job des 21. Jahrhunderts“ suchen nach Mustern in unstrukturierten Datenmengen, um unternehmerische Entscheidungen zu erleichtern.

## Datenwissenschaft im Land Brandenburg

Die Arbeit an dem vom Europäischen Fond für regionale Entwicklung (EFRE) geförderten Projekt "Präskriptive Analytik im Land Brandenburg" (PreLytica) wurde bereits am 01.03.2017 aufgenommen. In zwei Teilprojekten sollen Methoden des maschinellen Lernens erforscht und angewandt werden. Aus unstrukturierten Datenmengen sollen dabei automatisiert relevante Handlungsvorschläge abgeleitet werden, um gezielt unternehmerische Entscheidungen in kleinen und mittelständischen Unternehmen (KMU) im Land Brandenburg zu erleichtern.

## Kickoff-Event zum Projektstart

Am 28.03.2017 wurde das Projekt mit einer Kickoff-Veranstaltung offiziell an der THB vorgestellt. Interessierte konnten sich über Inhalte und Ziele des Projekts informieren, die neu bezogenen Projekträume besichtigen und mit Mitarbeitern über das Trendthema „Data Science“ diskutieren.

Im Teilprojekt **Sport Performance Insight (SPI)** befassen sich die Forscher mit der Analyse und Auswertung medizinischer Daten im Leistungssport. Im Teilprojekt **Solar Potential Analytics (SPA)** geht es um die Potenzialanalyse von Dachflächen aus Solarkatasterdaten. Beide Teilprojekte werden durch ihre Grundlagen im Bereich „Data Science“ (= Datenwissenschaft) vereint, der vor allem Methoden aus Statistik, Informatik und Mathematik beinhaltet.

Die Kombination dieser Disziplinen macht das Projekt zu einer echten Herausforderung, der sich insgesamt acht Mitarbeiter der THB stellen. Innerhalb von drei Jahren



soll für jedes Teilprojekt ein anwendungsnaher Prototyp entstehen, in dem die neuen Erkenntnisse, Methoden und Technologien demonstriert werden.

## Verbindung von Forschung und Lehre

Das Projekt mit sechsstelliger EU-Finanzierung ist ein Pfeiler der strategischen Ausrichtung der Hochschule auf das allgegenwärtige Thema „Digitalisierung“. Fähige Studierende können in dem Projekt zudem als Nachwuchswissenschaftler ein Praktikum absolvieren und ihre Abschlussarbeit (Bachelor- oder Masterthesis) verfassen. Dadurch wird ein Austausch zwischen Forschung und Lehre ermöglicht und Studierende werden direkt in die Spitzenforschung eingebunden.

Aktuelle Informationen und Zwischenergebnisse zum Projekt werden auf der Projektwebseite (<https://prelytica.th-brandenburg.de>) veröffentlicht.

# „On par“ mit der Forschung: THB gewinnt Sportverein für Data-Science-Kooperation

14.07.2017

Im Forschungsprojekt „PreLytica“ der Technischen Hochschule Brandenburg (THB) konnte ein erster Projektpartner gewonnen werden. Ziel der Kooperation ist die Anwendung maschineller Lernverfahren auf Daten aus dem Leistungssport.

## Kooperation mit regionalem Sportverein

Im Rahmen des Projekts „PreLytica“ soll erforscht werden, wie maschinelle Lernverfahren auch in kleinen und mittelständischen Unternehmen eingesetzt werden können. Dabei umfasst das Teilprojekt „Sport Performance Insight“ die Analyse von Daten mit Methoden aus dem Bereich „Data Science“.

Über 300 Kinder und Jugendliche trainieren regelmäßig im Golf- und Land-Club Berlin Wannsee e.V. (GLCBW). Die zahlreichen Vorurteile gegenüber der Sportart sind längst überholt. Golf ist ein Breitensport und die Nachwuchsförderung ist wie beim Fußball oder Handball in Vereinen organisiert. Im Bereich des Leistungssports muss auch im Golf regelmäßig und hart trainiert werden, um in Turnieren bestehen zu können.

Der Zweck dieser Kooperation ist, die verfügbaren Daten im Leistungssportbereich zu analysieren und damit die bereits sehr starke Nachwuchsförderung des GLCBW weiter zu verbessern. Das vorbildliche Jugend-Engagement bildete in der Vergangenheit die Grundlage etlicher Mannschafts- und Einzelerfolge, darunter deutsche Meistertitel und Berufungen in den deutschen Nationalkader. Diese Ergebnisse beruhen bisher vor allem auf den Erfahrungen der Trainer und dem Engagement der einzelnen Spieler. „Die Kooperation mit der THB sehen wir als wichtige strategische Säule unserer Jugendarbeit. Wir erhoffen uns zahlreiche neue Einblicke in die Faktoren, die zum Erfolg der Spieler und des Vereins beitragen.“, kommentiert Leistungssportkoordinatorin Miriam Hiller vom GLCBW.

## Mit maschinellem Lernen zum Erfolg

Auf Basis der gewonnenen Erkenntnisse sollen die individuellen Leistungen von Spielern und Trainingsgruppen verbessert werden. Dabei wird die individuelle Entwicklung jedes einzelnen Spielers einbezogen. – Das wiederum ist nur mit Hilfe von Algorithmen möglich, denen beigebracht werden soll, wie eine optimale Leistungsentwicklung aussieht. Dafür bedarf es jedoch zahlreicher Daten, anhand derer das System die Muster lernt, die ein erfahrener Trainer als Bauchgefühl mitbringt.

Das Forschungsteam rund um Projektleiter Prof. Robert Franz will anhand der Daten Zusammenhänge einzelner Einflussfaktoren, wie der Trainingsintensität und der Turnierergebnisse, finden. Mit den gewonnenen Erkenntnissen sollen mittelfristig die Leistungen der Spieler und Trainingsgruppen verbessert werden. Dafür werden die individuellen Entwicklungen der einzelnen Spieler einbezogen.

## Analyse unstrukturierter Daten zur Entscheidungsunterstützung

Maschinelle Lernverfahren funktionieren besonders gut mit sehr großen und vielfältigen Datenmengen. Für die Analysen sollen daher öffentliche verfügbare Daten, aber auch die internen Datenbestände des Sportvereins eingesetzt werden. „Dafür müssen wir viele unterschiedliche Datenquellen und Datenformate ineinander integrieren.“, weist Projektleiter Prof. Dr. Robert Franz auf eine der zentralen Herausforderungen des Projekts hin.

Das System, das in den kommenden zwei Jahren entwickelt werden soll, soll am Ende also konkrete Handlungsvorschläge zum optimalen Training der Nachwuchsspieler geben. Die eingesetzten Methoden und Algorithmen lassen sich dann prinzipiell auf andere Sportarten, aber auch auf völlig andere Bereiche wie die Unternehmensführung übertragen.

# Studierende schnuppern Forschungsluft in der Wirtschaftsinformatik

04.01.2018

Studierende der Technischen Hochschule Brandenburg (THB) machen wissenschaftliche Gehversuche im Forschungsprojekt "PreLytica".

## Projektstudium

Zu Beginn ihres Studiums müssen alle Studierenden der Wirtschaftsinformatik an der THB im Rahmen von Projekten erste Erfahrungen mit wissenschaftlichen Arbeitsweisen sammeln. Innerhalb der ersten acht Wochen heißt es dann: Literatur zusammentragen, Konzepte bewerten und Prototypen entwickeln. Am Ende wurden die Ergebnisse vor ca. 200 Schülerinnen und Schülern aus Brandenburg im Audimax der THB präsentiert.

## Forschungsprojekt "PreLytica"

Im EU-geförderten Forschungsprojekt "PreLytica" geht es um die Nutzung verschiedener Datenquellen zur Generierung von neuem Wissen. Damit soll die Entscheidungsfindung unterstützt werden. Im Teilprojekt "Solar Potential Analytics" (SPA) geht es um die Identifizierung geeigneter Dachflächen zur Nutzung für Photovoltaik-Anlagen.

Im Rahmen des Projektstudiums sollten die Studierenden mit den konkreten Anforderungen aus diesem Forschungsprojekt verschiedene Software-Produkte für ein Auftragsvergabeportal bewerten. Bereits im ersten Semester haben die Studierenden damit einen Einblick in aktuelle Forschungsthemen erhalten und konnten gleichzeitig einen aktiven Beitrag für das Projekt leisten.

# Erste Meilensteine im Forschungsprojekt „PreLytica“ erreicht

**09.04.2018**

Vor einem Jahr wurde das Forschungsprojekt „PreLytica“ mit einer Kickoff-Veranstaltung von Herrn Prof. Robert Franz und seinen Mitarbeitern offiziell an der Technischen Hochschule Brandenburg vorgestellt.

Am 04.04.2018 öffnete das Projektteam erneut seine Türen und stellte seine ersten Ergebnisse und den weiteren Projektverlauf der beiden Teilprojekte zur Diskussion.

## Entwicklung mobiler und Web-basierter Prototypen

Im Teilprojekt „Sport Performance Insight“ (SPI) befasst sich das Projektteam mit der Analyse und Auswertung medizinischer Daten im Leistungssport. Im Teilprojekt „Solar Potential Analytics“ (SPA) geht es um die Potenzialanalyse von Dachflächen aus Solarkatasterdaten. Bereits jetzt konnten wesentliche Meilensteine erreicht werden. Dazu gehören der Aufbau einer Datenbank, die Bereinigung und Vorverarbeitung von Daten sowie die Entwicklung mobiler und Web-basierter Prototypen zur Datenerfassung und -analyse. Weiterhin erfolgte durch das Projektteam eine erste Generierung und praktische Validierung von Handlungsempfehlungen auf Basis der vorhandenen Daten.

Die Ziele des weiteren Projektverlaufs wurden mit den Gästen konstruktiv diskutiert. In naher Zukunft sollen weitere Kooperationspartner akquiriert werden, um einen verlässlichen Datenbestand aufzubauen. Der neue Fokus liegt dann auf der präskriptiven Analyse der Daten. Durch die gewonnenen Erkenntnisse sollen die individuellen Leistungen von Spielern und Trainingsgruppen im Leistungssport verbessert werden.

## Datenschutz ist ein wichtiger Aspekt

Der Austausch mit Experten zum Thema Datenschutz und Datensicherheit wird ein weiterer Meilenstein im Forschungsprojekt PreLytica sein. Neben dem Aufbau von Kompetenzen werden vom Projektteam IT-Werkzeuge entwickelt, die den Prozess

der Integration und Vorverarbeitung von Daten für Data-Science-Anwendungen auch für KMU realisierbar machen.

Weitere Informationen zu den beiden Teilprojekten finden Sie auf der Projektwebseite „PreLytica“.

# Dächer in Brandenburg an der Havel mit Potential

23.11.2018

Im Rahmen des Projekts PreLytica am Fachbereich Wirtschaft der TH Brandenburg wurde eine 3D-Solarpotentialanalyse sämtlicher Bestandsgebäude im Untersuchungsgebiet der Stadt Brandenburg an der Havel durchgeführt.

Stefan Rieder, Mitarbeiter des Projekts am Lehrstuhl von Prof. Dr. Robert Franz, liefert mit seinen Ergebnissen interessante Einblicke in den möglichen Ausbau von Solarstromanlagen auf den Dächern der Stadt.

Auf Basis von 3D-Daten wurde jedes eingemessene Gebäude im Stadtgebiet auf seine Eignung für Solarstromanlagen untersucht. Dabei berücksichtigt wurden nicht nur die individuellen Dachformen, sondern auch Verschattungen aufgrund von Vegetation, benachbarten Objekten oder Dachaufbauten.

## 3.304.000 m<sup>2</sup> für die Erzeugung von Solarenergie geeignet

Neben den benötigten Daten zu individuellen Gebäuden, die in dem Teilprojekt Solar Potentials Analytics (SPA) genutzt werden, liefert die 3D-Analyse auch interessante Potentialstatistiken des Stadtgebiets:

Von den insgesamt rund 47.000 untersuchten Gebäuden sind 20.000 Gebäude (ca. 42%) für die Erzeugung von Solarstrom geeignet. Berücksichtigt werden muss die Tatsache, dass die meisten Garagen und eingemessenen Gartenhütten schon aufgrund ihrer fehlenden Größe ungeeignet sind. Dennoch sind auf den Dächern Brandenburgs 3.304.000 m<sup>2</sup> für die Erzeugung von Solarenergie geeignet (51 % Geneigte Dächer; 49 % Flachdächer).

Die maximal installierbare Leistung (in kWp - Kilowatt peak) auf den Dächern der Stadt Brandenburg an der Havel wäre 345.359 kWp. Die Aufteilung in Dachneigung und Eignungsklasse ist der Abbildung zu entnehmen. Nur knapp 25% der möglichen Leistung entfallen auf die Flachdachflächen, bilden aber rund 50% (genau 49 %) der Fläche der geeigneten Dächer. Dies resultiert aus der notwendigen Aufständigung der Solarstromanlagen auf diesen Flächen. Durch Wartungsgänge und Abstände,

welche die Selbstverschattung der Anlage minimieren sollen, ist bei einer Aufständigung gen Süden nur rund ein Drittel der Fläche nutzbar. Höher wäre der Flachdachanteil der installierbaren Leistung, würde eine Ost-West – Konfiguration angenommen werden.

## Ein Stromertrag von 289.000 MW sind möglich

Mit der angenommenen Aufständigung bei Flachdächern gen Süden liegt die Summe des möglichen jährlichen Stromertrags aller Dachflächen bei 289.000 MW. Dies entspricht einer bilanziellen Versorgung von 192.667 Bürgern (bei Annahme eines Verbrauchs von 1500 kWh/Jahr und Bürger). Nimmt man diesen Vollausbau des Potentials an, so ließen sich pro Jahr etwa 133.000 Tonnen CO<sub>2</sub> einsparen. Ein Blick in das Anlagenregister der Bundesnetzagentur (BUNDESNETZAGENTUR 2018, Stand 07/2017) zeigt, dass in den Postleitzahlgebieten der Stadt Brandenburg an der Havel (14770, 14772, 14774, 14776) 399 Solarstromanlagen mit einer Gesamtleistung von 139.891 kWp Nennleistung installiert sind. Dass von dem oben genannten Gesamtdachpotential lediglich 2,4% genutzt werden zeigt, dass hier dringender Handlungsbedarf besteht, um einen weiteren Ausbau zu forcieren.

Zukünftig wären Daten für das ganze Bundesland Brandenburg wünschenswert. In anderen Bundesländern (Berlin, Nordrhein-Westfalen und Thüringen) werden sämtliche Geodaten bereits seit längerer Zeit öffentlich und kostenfrei als Open Data bereitgestellt.

Weitere Informationen zu den beiden Teilprojekten finden Sie auf der Projektwebseite „PreLytics“.





## Transfer in die Lehre

Die Hochschule ist stark auf das allgegenwärtige Thema „Digitalisierung“ ausgerichtet. Für fähige Studierende gibt es zahlreiche Möglichkeiten im Projekt mitzuwirken. Sie können bspw. in dem Projekt als Nachwuchswissenschaftler ein Praktikum absolvieren und ihre Abschlussarbeit (Bachelor- oder Masterthesis) verfassen. Studierende werden so direkt in die Forschung mit eingebunden. Es erfolgt so ein Austausch zwischen Forschung und Lehre. In der bisherigen Projektlaufzeit haben bereits verschiedene Studierende mit verschiedenen Aktivitäten erfolgreich am Projekt mitgewirkt:

- Projektstudium (Im Rahmen des Studiengangs Wirtschaftsinformatik B.Sc.)
- Leonid Drost - Ergebnisse des Praktikums waren eine konzeptionelle Skizze und ein Tätigkeitsbericht: Entwicklung eines Synchronisations- und Caching-Konzepts für den Offline-Betrieb mobiler Apps
- Kilian Friedrich - ehemaliger Mitarbeiter: Konzeption und Umsetzung eines Stundenplan-Administrationswerkzeugs auf Basis von Angularjs
- Markus Lampert – Praktikum incl. erfolgreich abgeschlossener Bachelorthesis: Konzeption und prototypische Implementierung einer Web-API für ausgewählte Wetterdaten des DWD
- Daniel Kiertscher – Projektmitarbeiter und Masterand: KDD zur Entscheidungsfindung im Leistungssport: Ein explorativer Ansatz
- Dr.-Ing. André Nitze – ehemaliger Projektkoordinator, erfolgreich abgeschlossene Promotion: Entwicklung eines prozessorientierten Modells zur konstruktiven Qualitätssicherung mobiler Unternehmens-Applikationen

Bei der Planung und dem Aufbau des Projekts wirkten weitere Praktikanten/Praktikantinnen erfolgreich mit:

- Elvira Schäfer – erfolgreich abgeschlossene Bachelorarbeit: Kriterienbasierte Evaluation von Frameworks für die Dokumentation von Programmierschnittstellen
- Enrico Bunde- erfolgreich abgeschlossene Bachelorarbeit: Softwaretests und Behaviour-Driven Development im Kontext hybrider App Entwicklung



# Entwicklung eines prozessorientierten Modells zur konstruktiven Qualitätssicherung mobiler Unternehmens-Applikationen

André Nitze

## Zusammenfassung Dissertation

Im Vorlauf und innerhalb des EU-geförderten Forschungsprojekts "PreLytics" an der Technischen Hochschule Brandenburg wurde ein Promotionsprojekt durchgeführt, das mit der Verteidigung der Arbeit am 28.06.2018 an der Otto-von-Guericke-Universität Magdeburg erfolgreichen abgeschlossen wurde. Die Arbeit beschäftigt sich mit der Fragestellung, wie fachliche Anforderungen möglichst effizient formalisiert und in einem strukturierten Entwicklungsprozess in ein werthaltiges digitales Produkt überführt werden können. Dafür wurden eine systematische Literaturrecherche, eine empirische Analyse und zahlreiche Experteninterviews durchgeführt und daraus schließlich ein konstruktives Prozessmodell zur qualitätsgesicherten Entwicklung mobiler Applikationen entwickelt und experimentell validiert. Die im November 2018 veröffentlichte Dissertationsschrift ist inhaltlich mit dem Projekt verwoben, was sich nachträglich über die Bezüge zum organisatorischen Entwicklungsprozess, der technischen Realisierung des Informationssystems und den thematisch verwandten studentischen Abschlussarbeiten zeigt.

## Dissertation

*Entwicklung eines prozessorientierten Modells zur konstruktiven Qualitätssicherung mobiler Unternehmens-Applikationen.* Dissertationsschrift. November 2018, Shaker-Verlag, Aachen. — ISBN 978-3-8440-6275-5.

## Bachelor-Arbeiten

BUNDE, E. : *Softwaretests und Behaviour-Driven Development im Kontext hybrider App-Entwicklung*. Bachelor-Thesis. 2016.

FRIEDRICH, K. : *Konzeption und Umsetzung eines Stundenplan-Administrationswerkzeugs auf Basis von AngularJS*. Bachelor-Thesis. 2016.

SCHÄFER, E. : *Kriterienbasierte Evaluation von Frameworks für die Dokumentation von Programmierschnittstellen*. Bachelor-Thesis. 2016.

## Zusammenfassung und Ausblick

In diesem Band wurden die ersten Projektergebnisse des Forschungsvorhabens PreLytica vorgestellt. Alle Kapitel bezogen sich auf Ergebnisse, die in der ersten Projekthälfte im Zeitraum Dezember 2016 bis etwa September 2018 erzielt wurden. Hauptthema war in der Startphase des Projekts die Entwicklung von Szenarios, die sich mit dem geplanten prototypischen Ansatz umsetzen lassen. Dazu mussten Datenquellen gesichtet und erschlossen werden, Projektpartner wurden gesucht und gefunden, Webportale aufgebaut, Rollen in den Prozessen definiert, mobile Apps zur Datenerfassung entwickelt und vorhandene Datenquellen aus dem Internet automatisiert in die Projektdaten integriert.

Es wurden viele heterogene Datenquellen identifiziert. Darunter befanden sich offene und proprietäre Daten. Daten wurden auf mehreren Wegen beschafft. Für Teilprojekt SPA wurden Solarpotenzialdaten von einem externen Dienstleister lizenziert. Für Teilprojekt SPI wurden die eigenen Daten eines Projektpartners erfasst und mit Daten aus offenen Quellen kombiniert.

Die geplanten Arbeitspakete des Projekts wurden begonnen und sind zum Teil bereits abgearbeitet. Dazu zählen:

- Ist-Analyse inklusive Identifizierung und Akquise relevanter Kooperationspartner
- Konzeption, Entwicklung und Betrieb einer technologischen Infrastruktur für die sichere Datenverarbeitung
- Aufbau einer Datenbank mit proprietären Datenquellen und selbst erfassten Daten
- Konzeption eines Schemas zur Vereinheitlichung der heterogenen Datenquellen
- Bereinigung und Vorverarbeitung von Daten
- Entwicklung mobiler und Web-basierter Prototypen zur Datenerfassung und -analyse

- Erste Generierung und praktische Validierung von Handlungsempfehlungen auf Basis der vorhandenen Daten

Damit konnten wesentliche geplante Meilensteine erreicht werden. Offene Forschungsfragen bestehen in Hinblick auf die Datenqualität der erfassten Datenquellen, insbesondere in den Aspekten Vollständigkeit, Konsistenz und historische Spannweite. Dieser Umstand beschränkt die Aussagekraft der identifizierten maschinellen Lernverfahren. Im weiteren Verlauf des Projekts soll dieser Herausforderung unter Hinzuziehung weiterer Datenquellen und der Validierung zusätzlicher Lernverfahren begegnet werden.

Die grundlegenden Vorarbeiten sind aber umgesetzt, nun folgt als nächster Projektschritt die Visualisierung von Daten sowie das Ableiten von Entscheidungen. Ein typischer Informationsfluss ist in Abbildung 1 dargestellt.

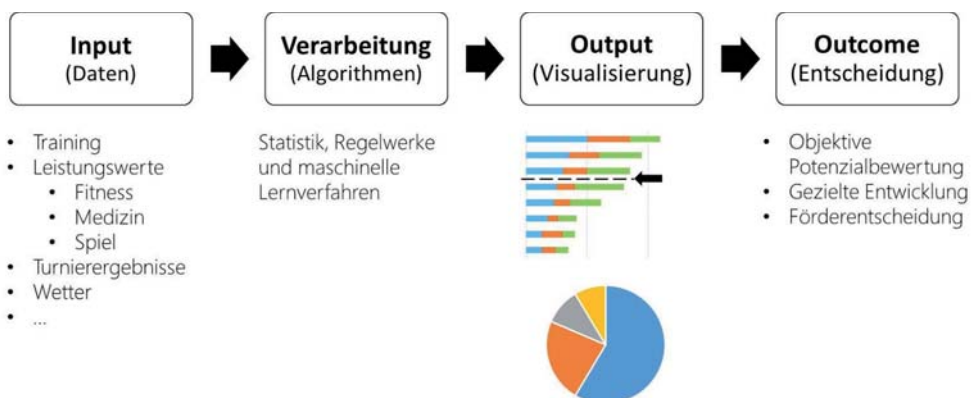


Abb. 1: Aus Daten werden Entscheidungen

Da das Projektziel ist, aus vorhandenen und selbst erzeugten Daten neues Wissen zu generieren und daraus Entscheidungen abzuleiten, idealerweise sogar direkt das System Entscheidungsvorschläge automatisiert erarbeiten zu lassen, wird sich der weitere Projektverlauf nun mit der Datenauswertung befassen. Die Infrastrukturen für die Prototypen sind fertig gestellt, so dass der Projektfokus auf die maschinellen Lernverfahren wechseln kann.