## Datensatz\_1: Projects | P3\_GrantExport\_with\_abstracts.csv

Das ist der Hauptdatensatz und die meisten Variablen, die beim Sampling oder in dem weiteren Projekt benutzt werden können befinden sich hier.

#### Project Number

Eindeutiges ID, die zwischen den unterschiedlichen Datensätze Projekte verbindet.

#### Project Number String

Sieht eher aus wie dieses ID nur intern eine Bedeutung hat.

#### Project Title

Alle Projekte haben Überschriften aber hier befinden sich unterschiedliche Sprachen.

#### Project Title English

Nur 40% von den Projekte haben eine englische Überschrift.

#### Reponsible Applicant

Der Name vom der verantwortlichen Antragstellerin. Falls wir nach der Analyse die beteiligten Personen in den einzelnen Projekten zugreifen wollen, können wir sie mit dem *Project Number* in dem *P3\_PersonExport.csv* finden (der Datensatz zeigt nicht nur *Responsible Applicant* sondern auch die anderen).

#### Funding Instrument & Funding Instrument Hierarchy

Diese sind Förderungsinstitutionen und ihre übergeordnete Kategorien. Sie könnten eher bei der Konklusionen der Recherche hilfreich sein. Z.B.: *SI aspect has a higher visibility in funding category …*

#### Institution & University

Ähnlich wie Förderungskategorien könnte die Information über Institutionen am Ende hilfreich sein, um über SI in bestimmten Institutionen zu sprechen.

#### Institution Country | **Sampling**

Vor dem Sampling können wir bestimmen was für eine Länderverteilung wir haben wollen. Diese Variable macht uns auch möglich eine regionale Eingrenzung zu machen

#### Discipline Number & Discipline Name & Discipline Name Hierarchy & Top level of discipline | **Sampling**

Die Fachrichtungen, ihre übergeordneten Disziplinen und als ein weiteres Level Forschungsgebiete. Diese Variablen können uns helfen einen Rahmen von den wiss. Disziplinen herzustellen und das als eine Eingrenzung zu benutzen. Weiters können diese wieder über die unterschiedlichen Aspekte in unterschiedlichen Felder zu argumentieren.

Es gibt 159 unterschiedliche Fachrichtungen, 24 Disziplinen und 3 Forschungsgebiete.

#### Start Date & End Date | **Sampling**

Diese sind wieder direkt zum Sampling relevant, weil wir bereits 2015-2018 als die Zeitspanne ausgewählt haben. Da bestimmte Projekte nicht beendet sind, *End Date* zu beachten liefert bessere Ergebnisse.

* Zwischen 2015-2018 gibt es ~11.000 Projekte.
* Zwischen 2015-2019 ~13.500 Projekte.

#### Approved Amount

Könnte wieder die Aussagen am Ende unterstützen aber fürs Sampling hat es keine Anwendungsmöglichkeit.

#### Keywords | **?**

Bestimmte Aspekte können durch die Keywords bestimmten werden. Potentiell könnten wir auch beim Sampling eine Sammlung von den Keywords beachten. Jedoch brauchen wir weitere Daten-Exploration.

#### Abstract (Project)

Nur ~40% von den Abstracts sind vorhanden und die Sprachen hier sind wieder gemischt. Wenn wir keine kontextuelle Analyse durchführen werden sind sie mindestens für jetzt nicht relevant.

#### Lay Summary - Lead & Lay Summary

Zusammenfassungen für nicht wiss. Menschen. Nur seltene Projekte beinhalten so eine Zusammenfassung und es gibt unterschiedliche Sprachen. Nicht relevant momentan.

## Datensatz\_2: People | P3\_PersonExport.csv

In den Projekten beteiligte Personen.

#### Last Name & First Name & Institute Name

Diese Informationen können durch *Project Number* zugegriffen werden.

#### Gender | **Sampling(?)**

Beim Sampling könnte das Gender relevant sein, wenn wir die Verteilung der Geschlechter beachten wollen.

#### Institute Place

Die Städter von den Institutionen. Diese Granulation brauchen wir höchstwahrscheinlich nicht aber prinzipiell könnten wir hierdurch *Country*-Variable in *Country,City* umwandeln.

#### Person ID SNSF

IDs für die Personen aber diese Information ist nicht im dem Datensatz von den Projekte vorhanden. Also, hilft uns nicht.

#### ORCID

Nur wenige haben eine ORCID-Nummer.

#### Projects as responsible Applicant & Projects as Partner & Projects as Practice Partner & Projects as Employee & Projects as Contact Person

In diesen Spalten stehen die zu dem Person assoziierten *Project Numbers*. Durch diese IDs können wir relevante Personen identifizieren.

## Datensatz\_3: Publications | P3\_PublicationExport.csv

Der Datensatz ist uns gerade nicht relevant. Wir haben die Informationen über Projekte bereits in dem Datensatz\_1.

## Datensatz\_4: Output Data | P3\_GrantOutputDataExport.csv

In diesem Datensatz sehen wir was für *Outputs* von unterschiedlichen Projekten produziert wurde.

#### Output Type

Unterschiedliche *Output*-Typen wie print media, online media, books, brochures, leaflets etc. unter dieser Variable vorhanden. Könnte eher am Ende die Aussagen unterstützen. Z.B.: *Die Projekte, in denen SI-Aspekt stärker war haben eher mehr gedruckte Medien produziert.*

#### Year

Natürlich nicht alle Projekte haben solche *Outputs* aber es sieht aus wie seit ~2014 immer mehrere Projekte sind dabei.

## Datensatz\_5: Collaboration | P3\_CollaborationExport.csv

Unterschiedliche Arten der Kollaboration in den Projekte

#### Group/Person

Beteiligte und ihre Organisationen sind hier wieder zu sehen.

#### Types of collaboration | **?**

Die Arten der Kollaboration. Leider sind sie nur organisatorisch gedacht und die Beschreibung enthält nicht so viele Details. Die Kategorien sind folgend:

* in-depth/constructive exchanges on approaches, methods or results
* Publication
* Research Infrastructure
* Industry/business/other use-inspired collaboration
* Exchange of personnel

Wir könnten auf bestimmte Kollaborationsarten fokussieren aber diese Kategorien sagen zu wenig.

#### Country

Hier haben wir die kollaborierende Länder. Über den SI-Aspekt in unterschiedlichen Kombinationen der Länder könnte am Ende in die Analyse gebracht werden.