

# Phase 1 and Phase 2 Summary SDG Project by SpringerNature, VSNU/UKB, Digital Science

London, December 2019

Juergen Wastl, Director Academic Relations & Consultancy

Mario Diwersy, Chief Technology Officer

Project Management: Timon Oefelein, Springer Nature

Part of **DIGITAL**science

# Slide deck contents

- Executive Summary
- The five SDG goals
- Phase 1: Training sets: concept, results, experience & lessons learned
- Phase 2: NLP/Machine Learning for automated classification

# Executive Summary

SpringerNature, together with Digital Science, and VSNU/UKB created a model from a selection of Sustainable Development Goals (SDG) focussing on societal aspects in the UN Sustainability Agenda.

Keyword search strings for five goals were defined, with input from the project partners, in order to produce training sets based on publications from the Dimensions platform. Using improved search strings instead of a manual build-up of respective sets of SDG related publications, the created training sets were used to apply Natural Language Processing and Machine Learning resulting in a classification scheme based on five UN SDGs.



GOAL 1: No Poverty

GOAL 2: Zero Hunger

GOAL 3: Good Health and Well-being

GOAL 4: Quality Education

GOAL 5: Gender Equality

GOAL 6: Clean Water and Sanitation

GOAL 7: Affordable and Clean Energy

GOAL 8: Decent Work and Economic Growth

GOAL 9: Industry, Innovation and Infrastructure

GOAL 10: Reduced Inequality

GOAL 11: Sustainable Cities and Communities

GOAL 12: Responsible Consumption and Production

GOAL 13: Climate Action

GOAL 14: Life Below Water

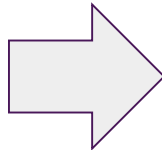
GOAL 15: Life on Land


GOAL 16: Peace and Justice Strong Institutions

GOAL 17: Partnerships to achieve the Goal


<https://www.un.org/development/desa/disabilities/envision2030.html>


# The Sustainability goals for this project



3: Good Health and Well-being 

4: Quality Education 

7: Affordable and Clean Energy 

11 Sustainable Cities and Communities 

16: Peace and Justice Strong Institutions 

These goals were chosen by the UKB-Coordination Point for Research Impact and they present an excellent fit to the SpringerNature Sustainability Development Program

# Goals focussed on the Society Aspects of SDG

- 3: Good Health and Well-being
- 4: Quality Education
- 7: Affordable and Clean Energy
- 11: Sustainable Cities and Communities
- 16: Peace and Justice Strong Institutions

Focus is on society aspects and, to a lesser extent, on biosphere and economy related goals.

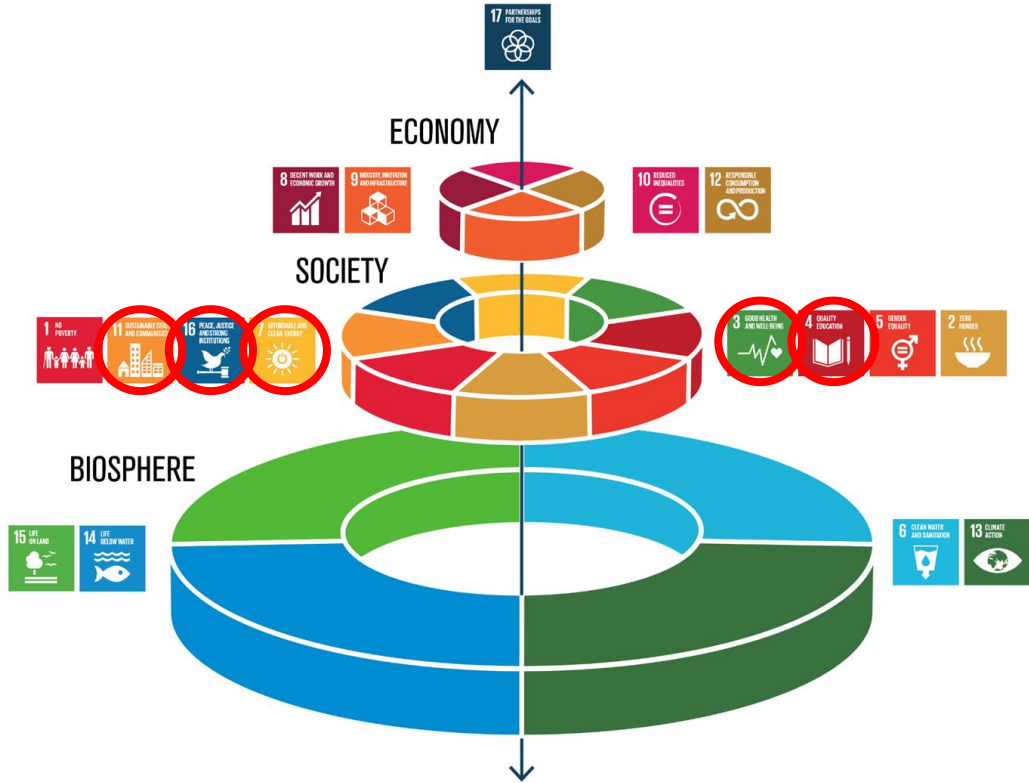


Image credit: <https://www.stockholmresilience.org/research/research-news/2016-06-14-how-food-connects-all-the-sdgs.html>

# Concept of modelling an SDG training set (Phase 1)

- 1) Analyse SDG goal description (incl indicator and targets based on the UN definitions) Extract key words and phrases
- 1) Optimise each search string for use in Dimensions (proximity search; language ambiguities)
- 1) Collect search terms in a spreadsheet and combine to retrieve a SDG specific, overarching search string per goal
- 1) Reiterative process: Check search results for false positives and improve search string and settings; no manual deletion of false positives from the resulting set of publications
- 1) Provide this scaffold search strings (spreadsheet per goal) to Partner (VSNU/UKB) for analysis, crosscheck and evaluation. Allow reiterative edits
- 1) Extend this process to allow SN editorial staff to broaden the subject expertise per goal: Evaluation of existing search strings and addition of new keywords and phrases

# Extract Search terms

Key phrases and terminology based on UN definitions of SDGs, including target and indicator definitions and narratives

NB: One training set per goal; no differentiation into targets or indicators

The image shows two screenshots of the Sustainable Development Goals Knowledge Platform website. The top screenshot displays the header with navigation links (HOME, SDGS, HLPF, STATES, SIDS, UN SYSTEM, STAKEHOLDERS, TOPICS, PARTNERSHIPS, RESOURCES, ABOUT) and a main banner for 'SUSTAINABLE DEVELOPMENT GOAL 16' with the text 'Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels'. Below the banner are tabs for 'PROGRESS & INFO (2019)', 'PROGRESS & INFO (2018)', and 'PROGRESS & INFO (2017)'. The bottom screenshot shows the 'PROGRESS OF GOAL 16 IN 2019' section, which includes a paragraph of text and a bulleted list of key findings. Below this is a table with two columns: 'TARGETS' and 'INDICATORS'. The 'TARGETS' column lists 16.1 and 16.2, and the 'INDICATORS' column lists 16.1.1, 16.1.2, 16.1.3, 16.1.4, 16.2.1, 16.2.2, and 16.2.3.

**PROGRESS OF GOAL 16 IN 2019**

Advances in ending violence, promoting the rule of law are uneven and continue to deprive millions of their access to public services and broader economic development progress. Renewed efforts are essential to move towards sustainable development.

- The number of intentional homicides per 100,000 population, which saw a slight uptick was largely the result of an increase in and some countries in sub-Saharan Africa.
- Various forms of violence against children persist. Recent data on the subject, nearly 8 in 10 children experienced psychological aggression and/or physical punishment, and half of children experienced violent disciplinary practices. At least 5 per cent of women between the ages of 15 and 49 reported experiencing at least one of their children's rights violations. Based on the limited data available, at least 5 per cent of women between the ages of 15 and 49 reported experiencing at least one of their children's rights violations. Based on the limited data available, at least 5 per cent of women between the ages of 15 and 49 reported experiencing at least one of their children's rights violations.
- There has been an overall increase in the detection of human trafficking, either a positive (enhanced efforts by authorities) or a negative (detected domestically: 58 per cent in 2016, up from 45 per cent in 2015). Detected victims of human trafficking were mostly women and children.
- The share of unsentenced detainees in the overall prison population has increased in recent years. This occurred in a context of a global increase in the number of prisoners while remaining constant as a share of the total population.
- Killings of human rights defenders, journalists and

TARGETS	INDICATORS
<b>16.1</b> Significantly reduce all forms of violence and related death rates everywhere	<b>16.1.1</b> Number of victims of intentional homicide per 100,000 population, by sex and age
	<b>16.1.2</b> Conflict-related deaths per 100,000 population, by sex, age and cause
	<b>16.1.3</b> Proportion of population subjected to physical, psychological or sexual violence in the previous 12 months
	<b>16.1.4</b> Proportion of population that feel safe walking alone around the area they live
<b>16.2</b> End abuse, exploitation, trafficking and all forms of violence against and torture of children	<b>16.2.1</b> Proportion of children aged 1-17 years who experienced any physical punishment and/or psychological aggression by caregivers in the past month
	<b>16.2.2</b> Number of victims of human trafficking per 100,000 population, by sex, age and form of exploitation
	<b>16.2.3</b> Proportion of young women and men aged 18-29 years who experienced sexual violence by age 18

<https://sustainabledevelopment.un.org/sdg16>

# QA process for generating the training sets

## Basic Process

### *Search strings*

Acknowledge **language ambiguities** (i.e. American English vs British English)

Implement **proximity searches**

Involve **subject matter experts** and add/deduct additional term/phrases

## Quality assurance

### *Minimising false positives*

Additional fine tuning, e.g. improving of proximity searches

Check against existing categorisations FoR (Fields of Research)

## *Challenge*

The creation of training sets comprising perfect search strings as a golden standard for Phase 2 (supervised Machine Learning) is knowingly not possible therefore the aim was to create the best training set possible



# Results - Search strings

Creating lists with results including identifying search string(s) per individual publication

Identifying false positives  
by checking search string  
results

Goal 3	((vaccine OR vaccination) AND (HIV OR "Human Immunodeficiency Virus"))	pub.1001133702	Analysis of Memory B Cell Responses and Isolation of Novel Monoclonal Antibodies with Neutralizing Breadth from HIV-1-Infected Individuals
Goal 3	(mortality AND premature)	pub.1001367611	Mortality benefits of population-wide adherence to national physical activity guidelines: a prospective cohort study
Goal 3	("Human African trypanosomiasis" OR "sleeping sickness")	pub.1001436503	Meiosis and Haploid Gametes in the Pathogen <i>Trypanosoma brucei</i>
Goal 3	((vaccine OR vaccination) AND (HIV OR "Human Immunodeficiency Virus"))	pub.1001465621	Genetic Imprint of Vaccination on Simian/Human Immunodeficiency Virus Type 1 Transmitted Viral Genomes in Rhesus Macaques
Goal 3	("access healthcare"~3)	pub.1001659728	Self-Screening and Non-Physician Screening for Hypertension in Communities: A Systematic Review

NB: Publications can be found by more than one search setting

# Minimising false positives by optimising proximity searches

## *Proximity Searches*

("x y"~2)... ("x y"~3)... ("x y"~5)... ("x y"~10)...

Example:

"Free primary education"~3

"Free primary education"~5

**"Free dental care during primary education" NOT SDG4**

"Free primary education"~10

**Need to strike a balance:**

**Relaxing proximity introduces false positives vs Tightening proximity limits true positives**

# Subject matter experts (SME) improve search strings and proximity searches

SME were asked to

- (i) check and evaluate existing search strings
- (ii) add and amend (edit & delete) the quantity and quality of the search strings

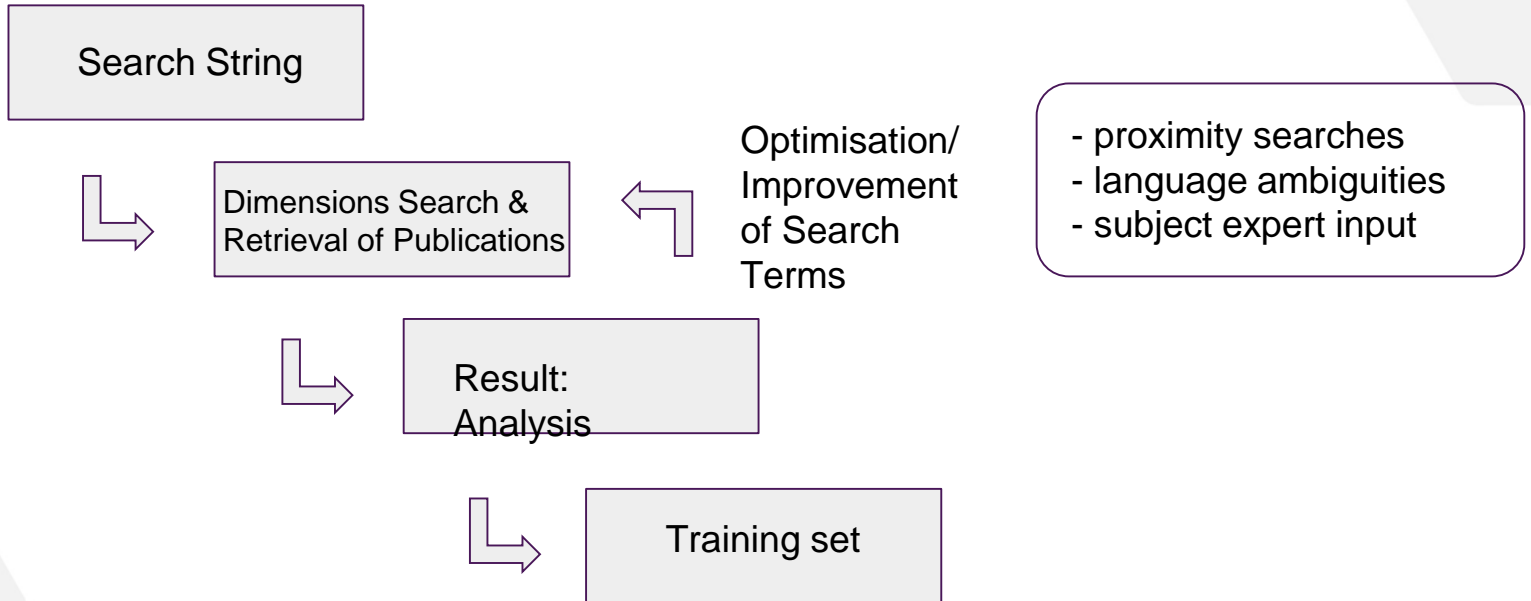
First iteration by VSNU/UKB academic input

Second iteration by SpringerNature editorial staff input

SME added terms and phrases were quality-checked by the Digital Sciece team (e.g. improving proximity searches)






# Improving the quality of the search terms through re-iteration

Bringing it all together:








# Training sets - Retrieval of results per SDG search string

Results (nr of publications) per training set:

 Goal 3	260464
 Goal 4	86557
 Goal 7	388640
 Goal 11	65027
 Goal 16	165707

Restricted to publications from 2010 onwards (to Sept 2019)

# Type of publication per goal

	Article	Chapter	Proceeding	Preprint	Mono graph	Edited book	
 goal 3	91.4%	4.9%	2.3%	1.2%	0.1%	0.1%	Goal 3 (Good Health and Well-being) overwhelmingly articles type
 goal 4	76.3%	13.7%	6.8%	1.6%	1.2%	0.4%	Goal 4 (Quality Education), 11 (sustainable cities) and 16 (Peace, strong Inst) with highest percentage on chapters
 goal 7	66.9%	7.1%	23.4%	2.2%	0.3%	0.1%	monographs and edited books
 goal 11	75.6%	13.8%	7.5%	1.8%	1.0%	0.3%	
 goal 16	70.3%	18.3%	1.5%	7.3%	2.3%	0.3%	Goal 7 (Affordable Energy) with highest share of proceedings

# Training sets - Retrieval of result per SDG search string

Results per training set:



Goal 3

260464



Goal 4

86557



Goal 7

388640



Goal 11

65027



Goal 16

165707

		citations total	cit mean		Dutch publ	share of publ	citation mean
		2.9m	10.67		6415	2.37%	24.66
		427k	4.7		1431	1.58%	12.9
		4.5m	11.03		6111	1.52%	16.72
		495k	7.21		1607	2.35%	14.24
		763k	4.19		3451	1.90%	9.34

# Results - Experience

Modelling per SDG was limited to one training set per goal - no drill-down into individual training sets per targets or indicators per goal were established; these were not in the scope of this project.

Robust, large training sets, however differences in size of training sets vary between 60k and 360k

Eliminating false positives by improving the overall search string, not by manual intervention in the result list

Repeated manual checks of resulting search strings

General search terms vs specialist search terms (e.g. “well-being”): Need for exclusion and careful intervention



# Lessons Learned

Effort to create training sets based on improved search queries is less time intense (with the available resources ) than in a bottom-up approach of combining & creating a training set based on collecting publications manually

Subject matter expertise is crucial in refining the training sets

Mobilising academic expertise and editorial staff expertise is time and resource intense

Avoid very generic search terms (well-being, sustainable, etc) to minimise false positives in the training set

Strive for the best possible search string for generating the training sets but perfection is not possible

# Phase 2: Automated classification via supervised Machine Learning (ML) - Rationale

Using supervised Machine Learning to build a **classification model** and subsequently a service that enables **classification of texts** to one of the project's five SDGs

Need for identifying and classifying text, also beyond publications

- Instant & fast

- Reliable

Basis: Publication (text) as training set

- Extensive set of relevant publications identified (Phase 1)

# Concept of establishing the classification service via ML

Training data:

Training set (positive examples per goal , Phase 1)

+

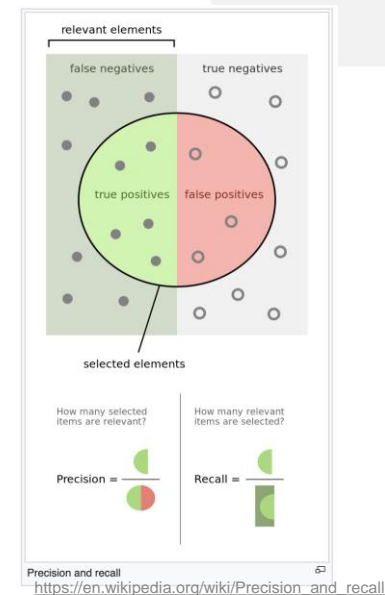
Random sample (1.3 million publications NOT part of the training set and not matching any of the five SDG)

*Inclusion of random publications to account for out-of-domain (i.e. non SDG relevant) vocabulary and improve machine learning generalizability*

Goal	Publication count
 3	260,464
 4	86,557
 7	388,640
 11	65,027
 16	165,707
Total goals (unique)	949,523
Additional random	1,290,517
Goals + random (unique)	2,240,040






# Machine Learning Methodology

- o Machine learning classifier
  - Based on the support vector machines (SVM) algorithm
  - Data processing included tokenization, stop-word removal and TF-IDF vectorization on word uni-grams with feature selection
  - All 950k publications were used as positive examples for the respective SDGs
  - The 950k documents were supplemented by about 1.3m random publications (not matching any of the goals) to account for out-of-domain vocabulary. This makes the classifier more generalizable (robust) to any (including out-of-domain) publications.
  - SVM optimizer set to favour precision (i.e how many selected items are relevant) over recall (i.e. how many relevant items are selected)



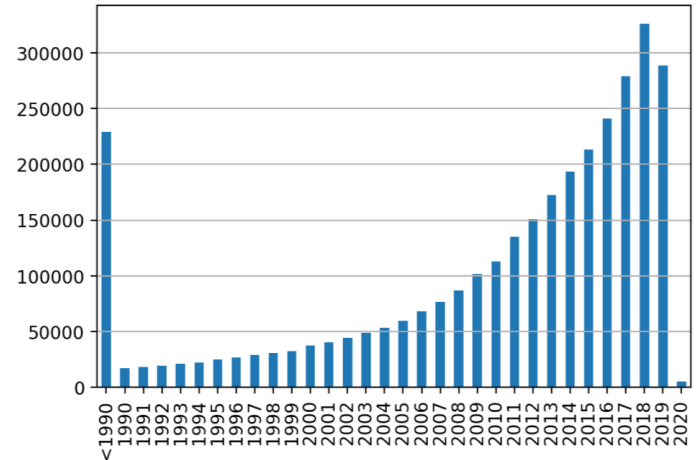
# Phase 2 Results

- o Classifier cross-validation on the training data
  - F1 of ~96% when cross-validated on the in-domain-only data (i.e. the 950k publications)
  - F1 of ~87% when cross-validated on all the training data (i.e. the 950k + 1.3Mn random publications)
  - The effectiveness values should serve only as a rough guide due to the training data not being manually annotated, i.e., the 950k + 1.3m publications have not been manually inspected for relevance - they are assumed to be correct based on the complex but limited keyword search in Phase 1

Goal	precision	recall	F1
 3	0.90	0.84	0.87
 4	0.83	0.72	0.77
 7	0.94	0.92	0.93
 11	0.79	0.61	0.69
 16	0.90	0.83	0.86
Overall	0.90	0.84	0.87

# Phase 2 Results (cont'd)

- o Application of the classifier to the entire Dimensions publication platform (>100 million publications)
  - In total 3.2 million publications were assigned to one of the five SDG
  - 2.1 million publications were assigned to one of the five SDG in the period 2010-current (October 2019)
  - Overlap of about 87% between the 3.2m publications and the 950k documents from the query result set which validates the estimate obtained on the training data



# Classification Service

- o The produced classification model is wrapped in and exposed as a web service containing
  - A REST API that abides by the industry standard openAPI 3.0.
  - Swagger UI (a third party user interface that displays our API documentation and usage information)
  - A one page UI for non-programmatic access to the classification capabilities

# Classification Service

- o The API has multiple access points that serve
  - classification system information (for both human and machine consumption)
  - goal predictions for input texts
  - relevance scores broken down to each goal for input texts



# Summary/Conclusions

- The classifier is the classic TF-IDF linear model with fairly standard parameters tipped slightly towards increased precision
- The choice of the model and parameters was dictated by the nature of the training data. This is not gold standard (that could only be achieved if each training example was inspected manually in its entirety).
- As a consequence:
  - linear models learn Phase 1 query keywords quite well (96% F1 on the 5 Goals-only data)
  - there is no reliable measure of true effectiveness which is necessary for the process of increasing classifier's performance