GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Optimierung und Sicherung der Datenqualität von STOFF-IDENT

V. Gronau, M. Sengl, M. Letzel
Bayerisches Landesamt für Umwelt

Bayerisches Landesamt für Umwelt, Bürgermeister-Ulrich-Str. 160, 86179 Augsburg, veronika.gronau@lfu.bayern.de

STOFF-IDENT- Datenbank

- Hilfsmittel zur Identifizierung bislang unbekannter gewässerrelevanter Stoffe mithilfe der Non- und Suspected-Target Analytik
- Einbindung von STOFF-IDENT in eine Arbeitsplattform zur Verknüpfung mit anderen Recherche-Tools (DAIOS, MassBank etc.)
- Integration der Stoffdaten aus Stoffvollzügen (REACH, Biozid-Richtlinie…)

AP3 – Datenqualität und Datenumfang

- Integration weiterer gewässerrelevanter Stoffe in die Datenbank
- Aktualisierung vorhandener Stoffgruppen
- Prüfung des Datenumfangs bei der Anwendung von STOFF-IDENT
- **Kontinuierliche Prüfung und Optimierung der Datenqualität**
- **Fehlersuche und -korrektur**

Bisher in STOFF-IDENT enthaltene Stoffe

Stoff	Quelle	Anzahl	Jahr
REACH	ECHA	ca. 4350	bis Dez 2013
Arzneimittel	UBA	ca. 1500	2009
PSM & PSM-TP	Bundesamt für Verbraucherschutz und Lebensmittelsicherheit, Reemtsma et al. 2013	ca. 400	2014 in D zugelassen
Biozide	Commission Regulation (EC) Anhang I & II	ca. 300	2007
Weitere schon nach-gewiesene Stoffe & TP	Zweckverband Landeswasserversorgung (LW)	2665	Feb 2014
	Fachpublikationen	ca. 1000	
	NORMAN: Candidates, List of Emerging Pollutants	ca. 870	2013
	Forensik Innsbruck (Oberacher 2011)	1207	2011
Gesamt alt (abzüglich der Überschneidungen)		ca. 8500	

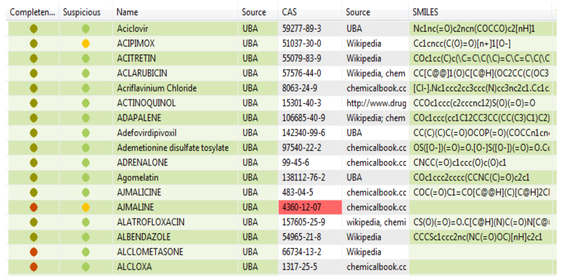
Neue in STOFF-IDENT enthaltene Stoffe

Stoff	Quelle	Anzahl	Jahr
Chemikalien	BfG, Schlüsener et al.	930	2015
Arzneimittel	UBA	210	
Haushalts-chemikalien	Drewes et al.	30	2009
Haushalts-chemikalien	High Production Chemical Database/Colorado School of Mines/Southern Nevada Water Authority	630	
Haushalts-chemikalien	Huntscha et al.	110	2012
TP‘s	GWA (Fenner et al. 2011)	75	2011
Gesamt neu (abzüglich der Überschneidungen)		ca. 9100	

Prüfregeln zum Einlesen in die Datenbank

- Substanz muss CAS-Nummer oder SMILES-Code haben
- Richtigkeit der CAS-Nummer wird über deren Prüfziffer überprüft
- SMILES und Summenformel dürfen keinen Punkt enthalten
- SMILES darf kein * enthalten
- -20 < logP < 20

Automatisierte Fehlererkennung im **SI-Crawler**, dem Daten-Einlesetool



Completen...	Suspicious	Name	Source	CAS	Source	SMILES
		Aciclovir	UBA	59277-89-3	UBA	Nc1nc(=O)c2ncn(COCCO)c2[nH]1
		ACIPIMOX	UBA	51037-30-0	Wikipedia	Cc1cncc(C(O)=O)[n+]1[O-]
		ACITRETIN	UBA	55079-83-9	Wikipedia	COc1cc(C)c(\C=C\C(\C)=C\C=C\C=C\C(\C
		ACLARUBICIN	UBA	57576-44-0	Wikipedia, chem	CC[C@@]1(O)C[C@H](OC2CC(C(OC3
		Acriflavinium Chloride	UBA	8063-24-9	chemicalbook.cc	[Cl-].Nc1ccc2cc3ccc[N]cc3nc2c1.Cc1c
		ACTINOQUINOL	UBA	15301-40-3	http://www.drug	CCOc1ccc(c2cccnc12)S(O)(=O)=O
		ADAPALENE	UBA	106685-40-9	Wikipedia; chem	COc1ccc(cc1C12CC3CC(CC(C3)C1)C2)
		Adefovirdipivoxil	UBA	142340-99-6	UBA	CC(C)(C)C(=O)OCOP(=O)(COCCn1cn
		Ademetionine disulfate tosylate	UBA	97540-22-2	chemicalbook.cc	OS([O-])(=O)=O.[O-]S([O-])(=O)=O.Cc
		ADRENALONE	UBA	99-45-6	chemicalbook.cc	CNCC(=O)c1ccc(O)c(O)c1
		Agomelatin	UBA	138112-76-2	UBA	COc1ccc2cccc(CCNC(C)=O)c2c1
		AJMALICINE	UBA	483-04-5	chemicalbook.cc	COC(=O)C1=C(O[C@@H](C)[C@H]2Cl
		AJMALINE	UBA	4360-12-07	chemicalbook.cc	
		ALATROFLOXACIN	UBA	157605-25-9	wikipedia, chemi	CS(O)(=O)=O.C[C@H](N(C(=O)N[C@
		ALBENDAZOLE	UBA	54965-21-8	Wikipedia	CCCSc1ccc2nc(NC(=O)OC)[nH]c2c1
		ALCLOMETASONE	UBA	66734-13-2	Wikipedia	
		ALCLOXA	UBA	1317-25-5	chemicalbook.cc	

Sicherung der Datenqualität bei neuen Daten

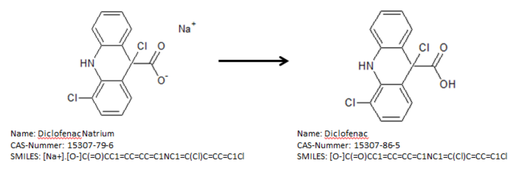
- Manuelle Korrektur der durch die Prüfregeln angezeigten Fehler
- Nichtübereinstimmung von CAS, SMILES und IUPAC bei nochmals eingelesenen Datensätzen ⟶ Korrektur
- Manuelle Fehlerkorrekturen in neuen Datensätzen, z.B. Klärung der CAS-Nummer durch Recherchen über CAS Registry
- Nachvollziehbare Dokumentation aller Bearbeitungsschritte und Datenquellen
- Schaffung konsistenter Datensätze: Salze in Säuren umwandeln d.h. Korrektur von Masse, SMILES und CAS-Nummer

Fehlersuche

- Salze, Komplexe und Gemische
- Punkte in über 120 SMILES-Codes
- Punkte in der Summenformel
- Mehrere CAS-Nummern bei über 60 Substanzen (nur eine der CAS-Nummern ist tatsächlich registriert)
- Substanzen ohne SMILES-CODE in STOFF-IDENT
- Falsche Übersetzung von Sonderzeichen in **?**

⟶ manuelle Korrektur von ca. 900 „Fehlern“

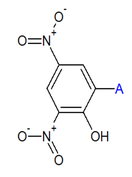
Beispiel: Salze



Beispiel: * im SMILES

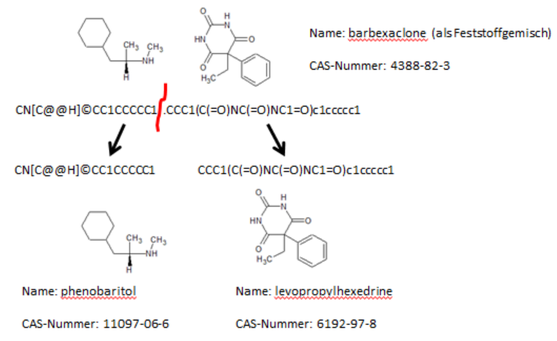
Beispiel: 2,4-dinitro-6-(1-alkyl)phenol

SMILES: Oc1c(*)cc(cc1[N+]([O-])=O)[N+]([O-])=O



* entspricht der Alkylgruppe

Beispiel: Punkt im SMILES



Beispiel: mehrere CAS-Nummern

registrated CAS-number

```
                (12262-73-6/RN)

=> d L3 REG
1          RN                77-92-9  REGISTRY
DR   12262-73-6, 43136-35-2, 136108-93-5, 245654-34-6, 623158-96-3,
     856568-15-5, 878903-72-1, 890704-54-8, 896506-46-0, 906507-37-7,
     1192555-95-5
```

deleted CAS-numbers

Ausblick

- Fokus auf Eingabe von Transformationsprodukten
- Optimierung der Kategorisierung
- Laufende Aktualisierung der Datenbank (z.B. neu unter REACH registrierte Chemikalien)

Bayerisches Landesamt für Umwelt · HOCHSCHULE WEIHENSTEPHAN-TRIESDORF UNIVERSITY OF APPLIED SCIENCES · Technische Universität München · Zweckverband Landeswasserversorgung · Berliner Wasserbetriebe