GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

FOR IDENT

Optimierung von Datenumfang und -qualität von STOFF-IDENT

Marion Letzel, Veronika Gronau, Manfred Sengl,
Bayerisches Landesamt für Umwelt

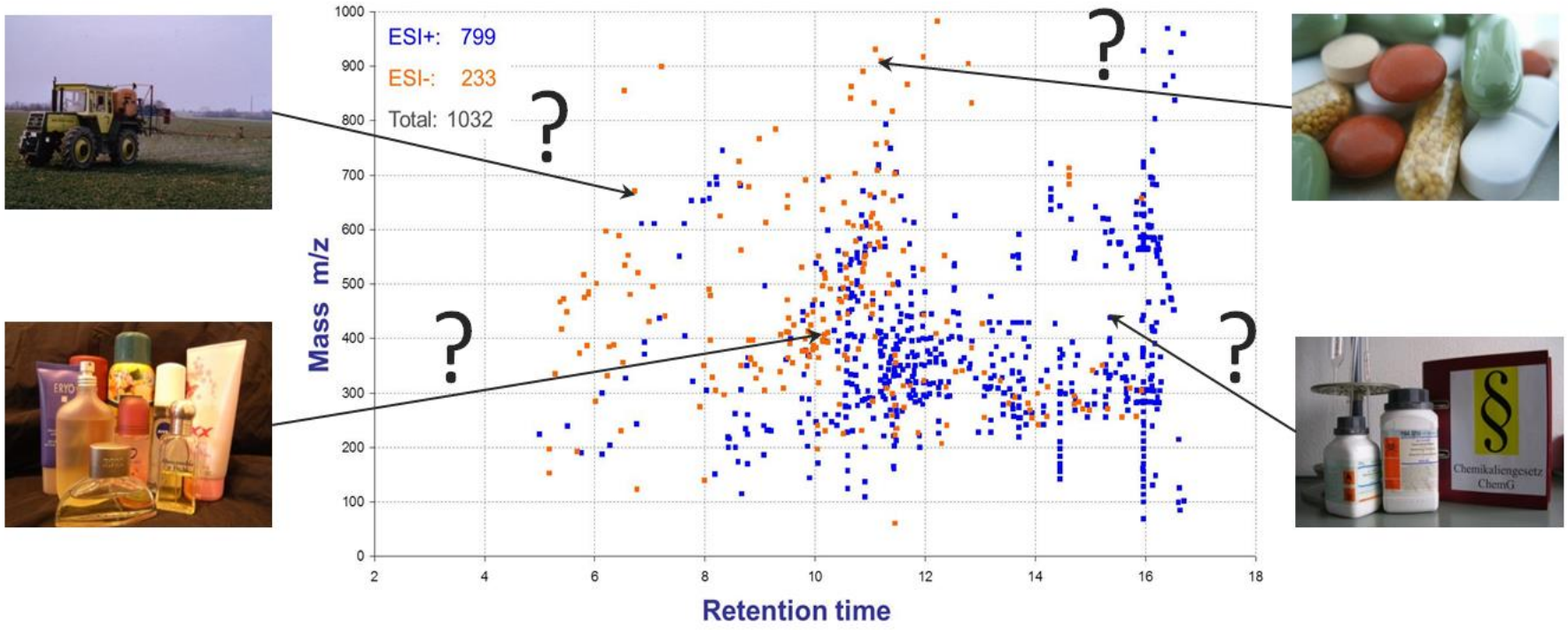
Bayerisches Landesamt für Umwelt

HOCHSCHULE WEIHENSTEPHAN-TRIESDORF UNIVERSITY OF APPLIED SCIENCES

Technische Universität München

Zweckverband Landeswasserversorgung

Berliner Wasserbetriebe



- **Fokussierung auf gewässerrelevante Stoffe**
- **Nutzung der Retentionszeit zur Priorisierung der Stoffvorschläge**
- **Frei verfügbar**

Datenattribute

- Name
- CAS-Nr.
- EC-Nr.
- SMILES
- InChI Code; InChI Key
- Chem. Name (IUPAC)
- Summenformel
- Exakte, monoisotopische Masse
- log P_{OW}
- log D_{OW} bei pH 3, 5, 7, 9
- Mengenband
- Kategorie (Stoffgruppe, Anwendung)

STOFF-IDENT: Datenbank gewässerrelevanter Stoffe

- Industriechemikalien (v.a. REACH-registrierte Stoffe)
- Humane Arzneimittelwirkstoffe und bekannte Metabolite
- Pflanzenschutzmittel und -Metabolite
- Biozide
- Weitere (bisher nachgewiesene) Stoffe
- Transformationsprodukte

Derzeit sind ca. 9.700 Stoffe in der Datenbank





Fotos: LfU

REACH

Registrierungsfristen:

- 1.12.2010: Stoffe >1000 t/a
  + R50/53 >100 t/a
  + CMR (Category 1+2) ≥1 t/a
- 1.06.2013: Stoffe >100 t/a
- 1.06.2018: Stoffe 1-100 t/a

- 7.12.2010: 2.992 Registrierungen
- 1.06.2013: 6.600 Registrierungen
- 3.12.2013: 11.766 Registrierungen
- Manuelle Entfernung von anorganischen Stoffe, Reaktions- und Stoffgemischen sowie Komplexen
- Bestand in STOFF-IDENT: ca. 4.350 Stoffe

Humane Arzneimittelwirkstoffe

- Quelle IMS MIDAS ® 2009/Umweltbundesamt
- 1517 small molecule-Wirkstoffe, aussortiert wurden u.a. Pflanzen, Tiere, Mikroorganismen, Metalle, Vitamine, Nukleine, Mineralien, Proteine/Peptide/AS, Diagnostika außer Röntgenkontrastmittel
- Manuelles Einfügen von fehlenden CAS-Nummern
- **2015 neu**: Aktualisierung um 210 neu hinzugekommene Humanarzneimittel

Pflanzenschutzmittel und ihre Metabolite

- Liste der 2014 in D zugelassenen PSM-Wirkstoffe (239 Stoffe, Quelle: BVL)
- Liste PSM-Metabolite über 1µg/L in Lysimeterversuchen (50 Stoffe, Quelle: BVL)
- Reemtsma et al. 2013: Metabolite aus Grund- und Oberflächenwasser: 105 Stoffe, teilweise mit CAS, Summenformel
- Manuelles Einfügen der restlichen CAS-Nummern

Biozide

- zugelassene Biozide (EU), Altwirkstoffe: Commission Regulation (EC) Nr. 1451/2007 (Anhang I): 583 Stoffe mit CAS
- zu prüfende Wirkstoffe (Anhang II): 234 Stoffe mit CAS, Überschneidung mit Anhang I
- Aufnahme in Anhang I der Biozid-Richtlinie 98/8/EG: 63 Stoffe mit CAS
- Alle Listen bearbeitet: Gemische, Anorganik, Mikroorganismen, natürliche Öle & Extrakte entfernt

Weitere Stoffe / Stofflisten (Auswahl)

Stoffsammlung aus Analytik-Laboren und NORMAN

- LW: 2665 Stoffe aus GC- & LC-MS-Messungen
- Forensik Innsbruck: 1207 Stoffe aus LC-MS-Messungen
- NORMAN-list of emerging substances (694); NORMAN-list of candidate emerging substances (180)
- LfU: 94 Stoffe
- **2015 neu:** Stoffliste BfG (Dank an Hr. Schlüsener)

Stofflisten Literatur

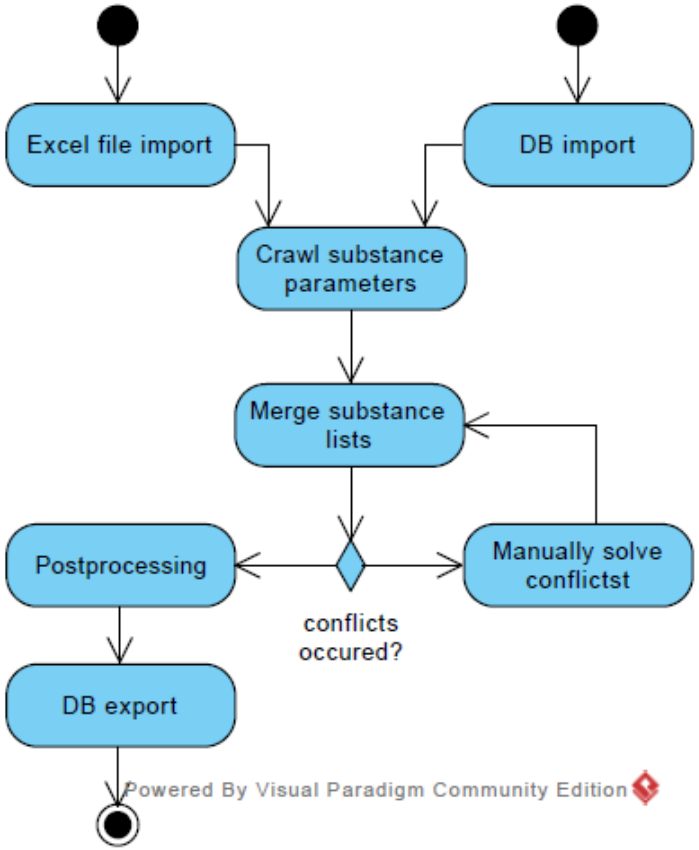
- Howard & Muir 2010 + 2013: Verunreinigungen, Byprodukte, 930 Stoffe
- Weitere, z.B. Field et al. 1994, Drewes et al. 2008, Li et al. 2014, Kern et al. 2010, RISK-IDENT, Frost & Griffiths 2008
- **2015 neu** z.B. Drewes et al. 2009, High Production Chemical Database/Colorado School of Mines/Southern Nevada Water Authority, TPs (z.B Fenner et al. 2011)

▸ *Quellen für Eintragungen in STOFF-IDENT sind in der Datenbank hinterlegt*

Sicherung Datenqualität beim Einlesen

Workflow des SI-Crawlers:

1. Prüfregeln
2. Auffüllen der Datenlücken mit chemicalize oder pubchem
3. Datenabgleich, automatisches Erstellen von „suspicious“-Listen
4. Manuelle Korrektur
5. Einlesen



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

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
		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=CO[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	

Automatisierte Fehlererkennung im **SI-Crawler**

Sicherung der Datenqualität beim Einlesen

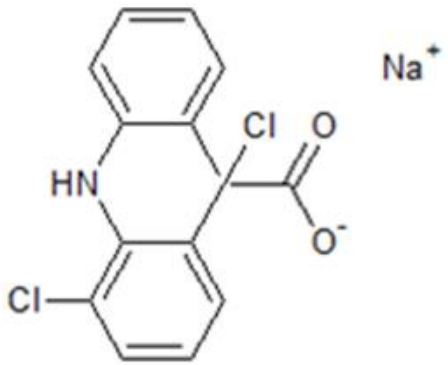
- Manuelle Korrektur der durch die Prüfregeln angezeigten Fehler
- Manuelle Korrektur der Unstimmigkeiten beim Datenabgleich (CAS, SMILES und IUPAC-Name), z.B.
  - Klärung der CAS-Nummer durch Recherchen über CAS Registry
  - Entfernung von Punkten und * in SMILES oder Summenformel
- Screening der neuen Gesamtliste auf Unstimmigkeiten
- Nachvollziehbare Dokumentation aller Bearbeitungsschritte und Datenquellen
- Schaffung konsistenter Datensätze

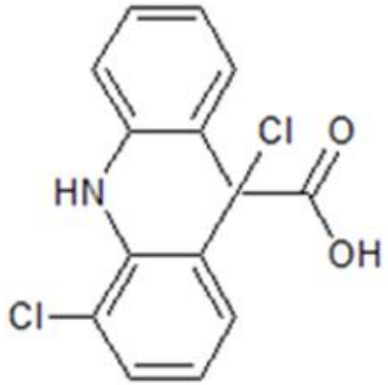
Typische Korrekturen

- 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 Einträgen

Beispiel: Salz in Säure

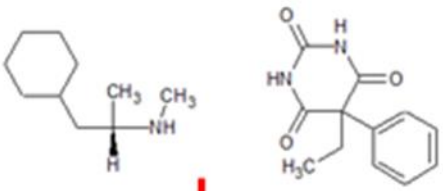




Name: Diclofenac Natrium
CAS-Nummer: 15307-79-6
SMILES: [Na+].[O-]C(=O)CC1=CC=CC=C1NC1=C(Cl)C=CC=C1Cl

Name: Diclofenac
CAS-Nummer: 15307-86-5
SMILES: [O-]C(=O)CC1=CC=CC=C1NC1=C(Cl)C=CC=C1Cl

Beispiel: Punkt im SMILES



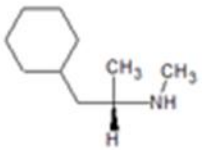
Name: barbexaclone (als Feststoffgemisch)

CAS-Nummer: 4388-82-3

CN[C@@H]©CC1CCCCC1.CCC1(C(=O)NC(=O)NC1=O)c1ccccc1

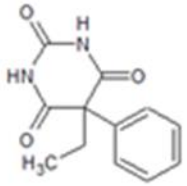
CN[C@@H]©CC1CCCCC1

CCC1(C(=O)NC(=O)NC1=O)c1ccccc1



Name: phenobaritol

CAS-Nummer: 11097-06-6



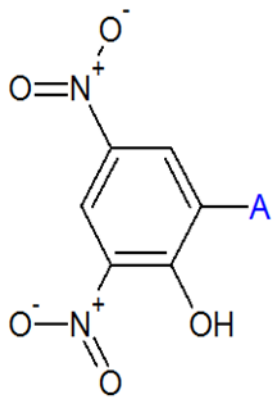
Name: levopropylhexedrine

CAS-Nummer: 6192-97-8

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: mehrere CAS-Nummern

(12262-73-6/RN)

registrated CAS-number

```
=> 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

- Laufende Aktualisierung der Datenbank
  (z.B. neu unter REACH registrierte Chemikalien)
- Eingabe weiterer Stofflisten (z.B. Norman Suspects Lists, Stoffliste NIVA (Norwegen) zusammen mit Fa. /Waters
- Stoffe aus Literatur: Fokus auf Transformationsprodukten
- Suche in einzelnen Listen (Datengrundlage auswählen)
- Optimierung der Kategorisierung
- Bearbeitung der letzten Fehlerlisten

FOR IDENT

Vielen Dank für Ihre Aufmerk-samkeit

Bayerisches Landesamt für Umwelt

HOCHSCHULE WEIHENSTEPHAN-TRIESDORF UNIVERSITY OF APPLIED SCIENCES

Technische Universität München

Zweckverband Landeswasserversorgung

Berliner Wasserbetriebe

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Optimierung und Sicherung der Datenqualität von STOFF-IDENT

FOR IDENT

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. 9700	

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Optimierung und Sicherung der Datenqualität von STOFF-IDENT

FOR IDENT

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	