TEXTE
21/2025

Abschlussbericht

Online-Portal „Non-Target Screening für die Umweltüberwachung der Zukunft“

Ein digitales Archiv für das Aufzeichnen von stofflichen Belastungen in Gewässern

von:

Kevin Jewell, Franziska Thron, Björn Ehlig, Jonas Skottnik, Thomas Scharrenbach, Thomas Ternes, Arne Wick
Bundesanstalt für Gewässerkunde, Koblenz

Jan Koschorreck, Anna Lena Kronsbein,
Umweltbundesamt, Berlin

Herausgeber:
Umweltbundesamt

Für Mensch & Umwelt

Umwelt Bundesamt

TEXTE 21/2025

REFOPLAN des Bundesministeriums Umwelt,
Naturschutz, nukleare Sicherheit und Verbraucherschutz

Forschungskennzahl 3720 22 201 0
FB001571

Abschlussbericht

Online-Portal „Non-Target Screening für die Umweltüberwachung der Zukunft“

Ein digitales Archiv für das Aufzeichnen von stofflichen Belastungen in Gewässern

von

Kevin Jewell, Franziska Thron, Björn Ehlig, Jonas Skottnik,
Thomas Scharrenbach, Thomas Ternes, Arne Wick
Bundesanstalt für Gewässerkunde, Koblenz

Jan Koschorreck, Anna Lena Kronsbein,
Umweltbundesamt, Berlin

Im Auftrag des Umweltbundesamtes

Impressum

Herausgeber
Umweltbundesamt
Wörlitzer Platz 1
06844 Dessau-Roßlau
Tel: +49 340-2103-0
Fax: +49 340-2103-2285
buergerservice@uba.de
Internet: www.umweltbundesamt.de

Durchführung der Studie:
Bundesanstalt für Gewässerkunde
Am Mainzer Tor 1
56068 Koblenz

Abschlussdatum:
Januar 2024

Redaktion:
Fachgebiet II 2.5 Labor für Wasseranalytik
Jan Koschorreck

Publikationen als pdf:
http://www.umweltbundesamt.de/publikationen

ISSN 1862-4804

Dessau-Roßlau, Februar 2025

Kurzbeschreibung: Online-Portal „Non-Target Screening für die Umweltüberwachung der Zukunft“

Es wurde ein System zur Verwaltung großer Datenmengen aus dem Non-Target Screening (NTS) entwickelt, einer Technik zur Identifizierung von organischen Spurenstoffen im Wasser. Das System (genannt NTSPortal) ist von entscheidender Bedeutung für die Anwendung von NTS im Gewässerschutz, da NTS Daten aus tausenden von Parametern bestehen, aus vielen Proben und durch unterschiedliche Labore generiert werden. Um Vergleiche über Raum und Zeit mit diesen großen Datenmengen durchzuführen, ist ein digitales Archiv erforderlich.

Das System besteht aus Skripten für die Prozessierung von Messdaten, der Sicherung der Daten in einer Datenbank und der Visualisierung der Daten auf einem interaktiven Web-Dashboard. Dies ermöglicht die schnelle Erstellung räumlicher Übersichten, in denen Schadstoff-Hotspots hervorgehoben werden, sowie langfristige Trendanalysen (z. B. von neuen Arzneimitteln in Flüssen). Darüber hinaus erleichtert das System die Identifizierung bisher unbekannter Spurenstoffe, z.B. durch die Analyse täglicher Proben aus Oberflächengewässer.

Es bestehen jedoch weiterhin Herausforderungen. Die Vergleichbarkeit der Daten über mehrere Jahre hinweg und zwischen Laboren benötigt weitere Optimierung und die Entwicklung von Standards für die Messung und Datenprozessierung. Die Benutzeroberfläche und die Qualitätssicherung müssen weiterentwickelt werden, um die Benutzerfreundlichkeit und die Interpretation der Ergebnisse zu verbessern. Trotz dieser Herausforderungen bietet das NTSPortal einen vielversprechenden Ansatz für die Integration von NTS im Gewässerschutz.

Abstract: Online portal “Non-target screening for the environmental monitoring of the future”

A system was developed to manage large amounts of data from Non-Target Screening (NTS), a technique used to identify organic trace contaminants in surface waters. This system (called NTSPortal) is crucial because NTS generates data with hundreds of parameters from many samples, and the required comparisons across space and time can only be achieved with a digital archive.

The system consists of scripts for processing raw measurement data, storing results in a database, and data visualization with an interactive web dashboard. This allows for quick generation of regional overviews highlighting contaminant hotspots and long-term trend analysis (e.g., tracking new pharmaceuticals in rivers). Additionally, the system facilitates the identification of previously unknown contaminants, for example by analyzing daily river samples and characteristic trend data.

However, challenges remain. Data compatibility across labs and over time needs further optimization and the introduction of new standards for analysis and data processing. The user interface and data quality assurance require further development to improve usability and interpretation of results. Despite these challenges, NTSPortal offers a promising approach for integrating NTS into water quality assessment.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis und Definitionen

Abkürzung	Erläuterung
API	Application programming interface
AUE Basel	Amt für Umwelt und Energie, Basel-Stadt
BAFU	Bundesamt für Umwelt, Bern
BAM	Bundesanstalt für Materialforschung
BfG	Bundesanstalt für Gewässerkunde
BWB	Berliner Wasser Betriebe
CAS	Chemical Abstracts Service
CAS-RN	CAS registry number (Chem. Identifizierung)
cps	Counts per second
Dashboard	Web-Applikation für die dynamische Darstellung von Daten in einer Datenbank
DDA	Data-Dependent Acquisition
DIA	Data-Independent Acquisition
EAWAG	Eidgenössische Anstalt für Wasserversorgung, Abwasserreinigung und Gewässerschutz
EIC	Extrahiertes Ionenchromatogramm
ESI	Elektrospray Ionisierung
f_{FN}	Anteil falsch negativer
f_{FP}	Anteil falsch positiver
(HP)LC	Flüssigkeitschromatografie
(HR)MS	(Hochauflösender) Massenspektrometer/-metrie
ICWRGC	International Centre for Water Resources and Global Change
IKSR	Internationale Kommission zum Schutz des Rheins

Abkürzung	Erläuterung
InChI	International chemical identifier
InChI-key	Hash-Codierung der InChI
IS	Interner Standard
LANUV	Landesamt für Natur, Umwelt und Verbraucherschutz
LAWA	Bund/Länder-Arbeitsgemeinschaft Wasser
LfU-Bayern	Bayerisches Landesamt für Umwelt
LfU Brandenburg	Brandenburgisches Landesamt für Umwelt
LUBW	Landesanstalt für Umwelt, Messungen und Naturschutz Baden-Württemberg
ML	Maschinelles Lernen
MS	Massenspektrometer/-metrie
m/z	Masse-Ladungs-Verhältnis
neg	Negative Ionisierung (ESI)
NLWKN	Niedersächsischer Landesbetrieb für Wasserwirtschaft, Küsten- und Naturschutz
noSQL	Not only SQL (structured query language)
NTS	Non-Target-Screening, Non-Target-Analyse
pos	Positive Ionisierung (ESI)
QToF-MS	Quadrupol Time-of-Flight mass spectrometer (Hybrid Quadrupol-Flugzeitmassenspektrometer)
SRO	Slowakische GmbH
t_R	Retentionszeit (Chromatografie)
UBA	Umweltbundesamt, Dessau
UFZ	Umweltforschungszentrum, Leipzig

Definitionen verwendeter Begriffe

Ein zentrales Konzept bei NTS ist das **Feature** (Müller 2023). Ein Feature ist der grundlegende Datenpunkt im NTS, es besteht aus drei Kenngrößen (m/z, Retentionszeit und Signalstärke [Intensität oder Peakfläche]) und stellt einen chromatographischen Peak dar. Die Features einer Probe werden tabellarisch in der sogenannten Featureliste aufgeführt. Eine gemessene Substanz kann ein oder mehrere Features hervorrufen. Ein Feature wird gegebenenfalls mit einem extrahierten Ionenchromatogramm (EIC) und Massenspektren, ggfs. auch Fragmentspektren (MS^2) verknüpft. Im optimalen Fall wird ein Feature mit weiteren Informationen wie beispielsweise einem Substanznamen annotiert (Xiao et al. 2012).

Die Aufnahme von MS^2-Spektren kann über unterschiedliche Scanarten im Massenspektrometer durchgeführt werden. Die prominentesten Beispiele sind **Data-Dependent-Acquisition** (DDA) und **Data-Independent-Acquisition** (DIA). Bei DDA werden Molekülionen für die Fragmentiertierung isoliert, aber das kann nur für eine begrenzte Anzahl an Ionen gemacht werden. Bei DIA werden alle Ionen als Paket fragmentiert und die Relationen zwischen Fragmente und Vorläuferionen wird nachträglich (rechnerisch) ermittelt (ein Prozess namens

Deconvolution). Deconvolution ist nicht immer perfekt und DIA Spektren haben mehr Rauschen als DDA Spektren, aber es wird von nahezu jedem Molekülion ein MS^2-Spektrum gemessen.

Non-Target-Screening wird hier in drei Stufen unterteilt: das **Bibliothek-gestütztes Screening** (library screening), das **Suspect-Screening** und das **Screening nach Unbekannten**.

Beim Bibliothek-Screening wird eine Referenz-Bibliothek von Massen, Standardspektren und Retentionszeiten bekannter Substanzen (die als analytische Standards vorliegen) verwendet, um Features mit Substanznamen zu annotieren. Hier ist die Sicherheit der Annotation hoch, da ein eigens gemessener Standard als Vergleich herangezogen wird, und der Prozess wird durch Skripte automatisiert. Die Sicherheit entspricht einer Konfidenz von Level 1 in der Einstufung nach Schymanski et al. 2014a (Schymanski et al. 2014) oder Kategorie 1 in der Einstufung nach Schulz und Lucke 2019 (Schulz und Lucke 2019).

Beim Suspect-Screening wird mit der theoretischen m/z einer vermuteten Substanz nach passenden Features gesucht. Die resultierenden Fragmentspektren werden untersucht (mit der Literatur verglichen), und/oder mit nachträglich gemessenen Standards abgeglichen, um die Identität zu bestätigen. Hier kann nur der erste Teil des Prozesses automatisiert werden und die Sicherheit der Annotation hängt von der Art der Bestätigung ab (über einen Standard, hoch (Level 1), über einen Literaturvergleich, weniger hoch (Level 2 bis 5)).

Beim Screening nach Unbekannten sucht man nicht nach einer bestimmten Substanz, sondern es werden Features priorisiert, die für die jeweilige Fragestellung relevant sind. Dies geschieht meist über statistische Analysen von, beispielsweise, Zeitreihen, räumlichen Verteilungsmustern oder Stoffeigenschaften. Die priorisierten Features werden anschließend über Literaturvergleiche und chemische Strukturaufklärung identifiziert. Auch hier ist die Sicherheit der Annotation vom Grad der Bestätigung abhängig. Die Prozesse (Priorisierung, Strukturaufklärung) werden mit Hilfe von Software unterstützt, werden aber nicht durchgehend automatisiert.

Zusammenfassung

Der schnelle Datenzugriff und die skalierbare Verwaltung von Daten sind in vielen Branchen zunehmend wichtige Teile des Arbeitsalltags, und immer mehr Werkzeuge für wissenschaftliche Datenbearbeitung und das sog. Informationsmanagement werden entwickelt. Ihr Nutzen für eine schnelle Interpretation der Ergebnisse und eine datenbasierte Planung liegt auf der Hand. Für das Non-Target-Screening (NTS) und die Verbesserung der Wasserqualität der Oberflächengewässer ist die umfassende und schnelle Datenverfügbarkeit von entscheidender Bedeutung, da jede Messung eine große Anzahl von Parametern enthält. Zudem wird das NTS als Screening-Technik eingesetzt, um einen ersten Überblick über das Vorkommen organischer Spurenstoffe zu erhalten. Es müssen viele Proben - räumlich oder zeitlich - verglichen werden. Daher ist ein System, das Tausende von Messungen mit jeweils Hunderten (oder sogar Tausenden) von Parametern gleichzeitig abrufen und vergleichen kann, unerlässlich.

Der in diesem Projekt entwickelte Prototyp der Datenverarbeitung, der Datenbank und des Web-Dashboard-Systems zeigt, dass aktuelle Frameworks und Open-Source-Bibliotheken in Kombination bereits die nötige Infrastruktur für einen aggregierten, interaktiven Online-Zugang zu Daten bereitstellen können.

Die Datenbank wurde in zwei Ebenen für unterschiedliche Anwendungsfälle unterteilt: Eine Ebene umfasst Daten, die durch Bibliothek-Screening verarbeitet wurden und zeigt nur identifizierte Verbindungen mit einem möglichst geringen Anteil falsch-positiver Befunde. Die zweite Ebene erweitert diese Daten um unbekannte Signale, die nur durch Masse und Retentionszeit zugeordnet werden können. Solche Daten sind für das Suspect-Screening oder das Screening nach Unbekannten geeignet.

Dashboards wurden für die Benutzeroberfläche eingesetzt. Dies sind browserfähige Applikationen, die über eine Internet-Verbindung zur Datenbank verfügen. Mit der Online-Bereitstellung kann ein räumlicher Überblick, der Hotspots und potenzielle Schadstoffquellen zeigt und Vergleiche zwischen Standorten ermöglicht, schnell erstellt werden. Beispielsweise ließ sich anhand von Kartenüberlagerungen erkennen, dass einige Derivate von quartären Ammoniumverbindungen (die als Biozide verwendet werden) spezifisch für bestimmte Wassereinzugsgebiete sind.

Anhand von Messungen von Schwebstoffproben aus der Umweltprobenbank, die bis ins Jahr 2005 zurückreichen, wurden Zeitreihen über zwei Jahrzehnte hinweg erstellt, um langfristige zeitliche Trends von Spurenstoffen zu erkennen. Beispielsweise konnte durch diese Daten nachgewiesen werden, wie die Einführung eines neuen Arzneimittels (Sitagliptin) schnell zu einer flächenübergreifenden Verbreitung in Flüssen führte.

Die Untersuchung täglich entnommener Proben eines Flusses ermöglichen zudem die Analyse von Zeitreihen, um typische Muster anthropogener Emissionen zu finden. Dadurch können beispielsweise auch ohne vorherige Kenntnis der Produktionsstandorte im Einzugsgebiet unbekannte Verbindungen priorisiert werden, die charakteristische produktionsbedingte Intensitätsverläufe aufweisen und vermutlich aus industriellen Quellen stammen. Daten von Derivaten oder verwandten Verbindungen können dann leicht miteinander und mit bekannten Verbindungen verknüpft werden (entweder durch Korrelation von Zeit-Trends oder durch Suche nach ähnlichen Massen oder MS^2-Fragmentierungsmustern), so dass Chemiker*innen zeitnah Hinweise für die Strukturaufklärung und schließlich die Identifizierung der Substanz sammeln können. In diesem Bericht wird ein Beispiel dargestellt, in dem durch die Zeitreihenanalyse ein Feature priorisiert wurde, das einen diskontinuierlichen Eintrag aufwies.

Die Strukturaufklärung ergab mehrere potenzielle Kandidaten mit einer Tetramethylpiperidine Teilstruktur.

Die retrospektive Suche in Tausenden von historischen Messungen nach neuen interessanten Verbindungen (Bibliothek-Screening) konnte erfolgreich automatisiert werden. Monatliche Aktualisierungen der Spektrenbibliothek können nun regelmäßig und automatisch mit dem gesamten Messarchiv verknüpft werden.

Herausforderungen und weiterer Entwicklungsbedarf werden in den folgenden Bereichen gesehen:

1. Die erforderliche Datenstabilität (innerhalb eines Labors) und die Einheitlichkeit der Datenfelder für Metadaten zwischen den Laboren muss weiter verbessert werden. Im Moment schränkt das die Interpretation der Daten ein. So wurden beispielsweise im Laufe von vier Jahren unvermeidbare methodische Änderungen vorgenommen (z. B. Verbesserungen der Analysemethode), die zu Veränderungen in Intensität-Zeitreihen führten. In der Kommunikation zu den Daten (Dokumentation) wird darauf hingewiesen, dass es sich bei NTS um eine Screening-Technik handelt, die nicht als “Target-Analytik“ behandelt werden kann (Konzentrationsbestimmung mit Hilfe von Surrogatstandards (Isotopenverdünnungs-Massenspektrometrie)). Es ist auch notwendig, dass Daten mit Metainformationen (bspw. Austausch eines Analysegeräts) annotiert werden können, die an prominenten Stellen gezeigt werden.
2. Basierend auf den Nutzererfahrungen sollten zukünftig die Dashboards optimiert und ergänzt werden. Dies sind einfache Dingen wie die Sichtbarkeit von Achsen und Punkte, erstreckt sich aber auch auf komplexere Themen wie die Annotation von Zeitreihen (z.B. mit Informationen zum Status des Messgeräts zum Zeitpunkt der Messung), die Bereitstellung leistungsfähigerer, in die Website integrierter statistischer Analysewerkzeuge und das Einpflegen oder die Verknüpfung mit weiteren Daten, wie z.B. Konzentrationen oder Abflussmengen, zur Unterstützung der Interpretation.
3. Die Qualitätssicherung bei der Datenverarbeitung befindet sich noch in der Entwicklungsphase. Beispielsweise ist die automatische Erkennung von falsch-positiven Ergebnissen ein wichtiger Baustein für eine bessere Nutzbarkeit des Systems. Ein Fall wäre die falsche Zuordnung einer Substanz durch das Screening mit der Spektrenbibliothek. Eine objektive Quantifizierung der Anzahl von Falsch-Positiven ist in diesem Fall schwer. Als möglicher Lösungsansatz bietet sich die Anwendung von weiteren (unabhängigen) Algorithmen, die als nachträglicher „Prüfungsschritt“ fungieren können. Zudem bedarf es generell neuer, aussagekräftiger Kenngrößen für die schnelle und zuverlässige Erkenntnis von „Außer-Kontroll-Situationen“, z.B., die Anzahl detektierter Features in einer Referenzprobe. Nicht zuletzt ist hier auch eine klare Kommunikation der Datenqualität an die Nutzer*innen unerlässlich, um die Interpretation der Screening-Ergebnisse einzuordnen.

Das Projekt hat demonstriert, dass die automatisierte Prozessierung eine Voraussetzung für das retrospektive Bibliothek-Screening ist. Häufige Ergänzungen zur Spektrenbibliothek machten eine vollständig automatisierte Verarbeitung der Messdateien erforderlich. Andernfalls müsste die Datenverarbeitung ständig manuell wiederholt werden. Um dies zu vermeiden, wurden Skripte entwickelt, um den gesamten Bestand an Messdateien in der Datenbank regelmäßig und ohne menschliches Eingreifen termingemäß zu prozessieren. Die automatisierte Datenprozessierung in NTS ist ein wichtiges Forschungsthema und erfordert weitere Entwicklungen in der automatisierten Qualitätsprüfung.

Investitionen in weitere Software-Programmierung während des Projekts waren entweder sehr zeitaufwendig oder teuer, wenn sie ausgelagert wurden. Ein Großteil des geschriebenen Codes

wird noch weiterentwickelt und fortlaufend dokumentiert werden müssen. Die Nutzung bestehender Code-Frameworks erwies sich als der beste Weg, um Fortschritte zu erzielen. Der Nachteil war, dass innerhalb der Grenzen dieses bestehenden Codes gearbeitet werden musste, d. h. die Ziele mussten auf der Grundlage dessen, was diese bereits bestehenden Frameworks leisten konnten, angepasst werden.

Es gibt große methodische Unterschiede, wie verschiedene Labore das Non-Target-Screening durchführen, und die Datensätze können sich in ihrer Struktur und ihrem Inhalt erheblich unterscheiden. Die größte Herausforderung bestand darin, dass viele Datensätze nicht über eine vergleichbare Anzahl von MS^2-Spektren verfügen, was die Annotation und das Alignment ohne extrem genaue Retentionszeiten erschwerte. Ein stärkerer Fokus sollte daher auf MS^2-Spektren gelegt werden.

Eine der größten analytischen Herausforderungen war die Aufrechterhaltung der Stabilität von Intensitäten und Retentionszeiten über fast zehn Jahre der Analyse. Die Verwendung von Shewhart-Kontrollkarten war entscheidend, um Abweichungen zu minimieren und Probleme am Messgerät frühzeitig zu erkennen. Die Einführung der Kontrollkarten ging mit einer erkennbaren Verbesserung der Stabilität einher. Manchmal war eine Verschiebung jedoch unvermeidlich. So hat beispielsweise eine Änderung in der Produktion der chromatographischen Säule eine Verschiebung der Retentionszeiten verursacht, die eine Erhöhung der Toleranzen für den Retentionszeit-Abgleich mit der Bibliothek erforderlich gemacht hat.

Summary

Fast data accessibility and scalable management of data is advancing in many branches and software tools for information management and scientific data processing are becoming available. Their usefulness for fast interpretation of results and evidence-based decision making is clear. For non-target-screening (NTS) in surface water quality assessment, this access to data is crucial because of the large numbers of parameters in each measurement and, since NTS is used as a screening technique to obtain an initial overview of the contamination by trace organic compounds, it often requires a large number of samples to be compared, either over time or space. So, a system to be able to access and compare thousands of measurements each with hundreds (or even thousands) of parameters is essential.

The prototype of the data processing, the database and the web dashboard system developed in this project shows that current frameworks and open source libraries can, in combination, deliver the necessary infrastructure to allow for aggregated, online and interactive data access to NTS data.

The database was organized into two levels for different use-cases: One level is for data processed by library screening showing only identified compounds with a low fraction of false positives. The second level extends this data with unknown detections identifiable only by mass and retention time. This type of data is useful for suspect screening or identifying unknown contaminants.

With interactive database access through an online dashboard it could be shown that a regional overview (i.e. map overlays) can be quickly generated, showing hotspots and potential sources of contaminants and allowing for comparisons between sites. With map overlays it can be seen for example that some derivatives of quaternary ammonium compounds (used as biocides) are specific to particular watersheds.

With measurements of samples from the Environmental Specimen Bank going back to 2005, time-series over two decades could be made to view long-term trends of contaminants, even seeing how the introduction of a new pharmaceutical (e.g. Sitagliptin) quickly resulted in its dispersion in rivers.

The measurement of samples from daily composite sampling of a river has allowed the analysis of time series to find patterns typical to anthropogenic emissions, allowing for the identification of previously unknown contaminants without prior knowledge of industries in the watershed. Data from derivatives or related compounds could then be easily linked together (either by correlating time trends, or looking for similar masses or MS^2 fragmentation patterns) allowing the chemist to quickly build a “body of evidence” for structural elucidation and ultimately identifying the chemicals. In this report, an example is shown in which time series analysis led to the prioritization of a feature as likely originating from an anthropogenic emission. The structural elucidation provided several candidate structures with a tetramethylpiperidine functional group.

It was shown that retrospective searching of thousands of historical measurements for new compounds of interest can be automated. Monthly updates to the spectral library can now be regularly checked with the entire archive of measurements with an automated routine.

Challenges and further development needs are seen in the following areas:

1. The necessary data stability (within a lab) and uniformity of data fields across labs is still not where it needs to be. This limits the interpretation of data. For example, over the course of 4 years of measurements there were unavoidable changes made (e.g. improvements to the

analysis method), these led to shifts in the response (peak area) trend in time-series which resulted in artifacts. It is important to communicate that NTS is a screening technique and should not be treated as “target analysis”, i.e. concentrations determined using surrogate standards (isotope dilution mass spectrometry). Furthermore, the annotation of data with metainformation (e.g. replacement of an instrument), which can then be clearly visualized alongside the data would improve communication.

2. Based on user feedback, continued work still needs to be done to improve the experience when using the dashboards. This starts with simple things like the visibility of axis and points but extends to more complex topics like annotating trends, providing more powerful statistical analysis tools built into the website and linking to more (external) data to assist the interpretation of results.
3. The quality assurance in the data processing is still in development. For example, the automated determination of false positives in library screening (incorrect compound annotations) followed by data correction needs to be further developed for improved usability of the system. In this and other cases quantifying the number of false positives is challenging to do objectively. One possibility to explore further is the use of alternate algorithms, which are used for checking the results after processing. Additionally, the development of easily interpretable parameters for quickly determining if control limits have been exceeded are needed. A simple example is the number of detected compounds in a reference sample. Finally, clear communication to the user of data quality is again essential to improve the interpretation of screening results.

The project demonstrated that fully automated processing of measurement files is a prerequisite for retrospective analysis, which was one of the initial goals of the project. Frequent additions to the spectral library made this necessary, otherwise, data processing work, a tedious process at best, would need to be constantly repeated. To avoid this, the entire collection of measurement files in a laboratory can now be re-processed on a regular basis without any human interaction. Automated processing was not widely available during the project and so a prototype system was developed. Further improvements are necessary in the future, for example implementing further automated quality checks.

Investment in further software programming during the project was either very time-consuming or prohibitively expensive when outsourced. Much of the code written will need to be further developed and documented in the future. The use of existing code frameworks proved to be the best way to make progress. The challenge is to operate within the limits of this existing code, i.e. by creative use of existing code and adapting the work to what these pre-existing frameworks can do.

There are large differences in how different labs perform non-target-screening and datasets can vary substantially in their structure and content. The biggest challenge was that many datasets did not have a comparable number of MS^2 spectra, making annotation and alignment difficult without extremely accurate retention times. Looking ahead, we believe there should be a stronger focus on MS^2 spectra.

A further challenge was maintaining stability of intensities and retention times over almost ten years of analysis. The use of Shewhart control charts was crucial to minimize drifts and quickly recognize problems with the instrument. Sometimes a shift was unavoidable. In one case, production changes in the chromatographic column meant that retention times tolerances for the library screening and alignment needed to be increased.

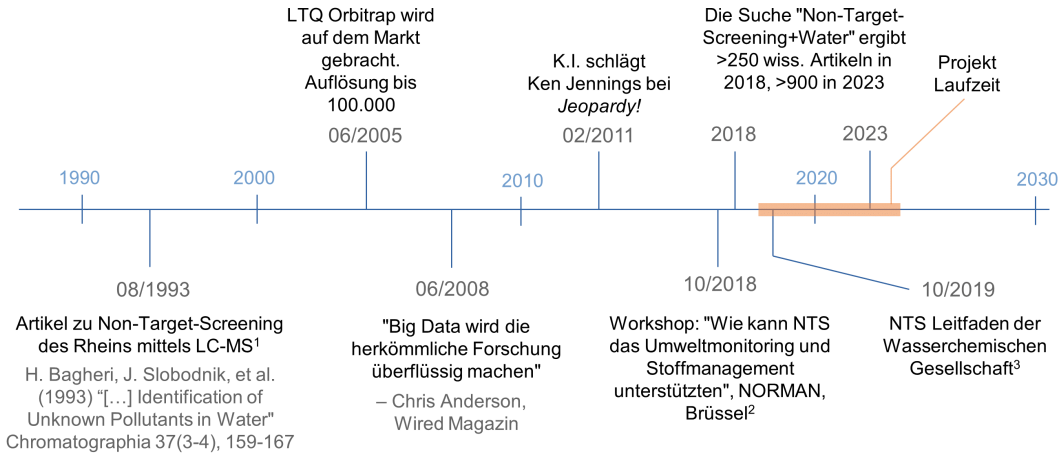
1 Einleitung

1.1 Entwicklung von Non-Target-Screening

Non-Target-Screening (NTS) ist eine massenspektrometrische Methode, in der, im Gegensatz zur klassischen „Target“-Analytik, keine Ionenvorauswahl stattfindet, sondern eine sogenannte Full-Scan Messung durchgeführt wird. Die Rohdaten (Massenspektren) werden anschließend algorithmisch nach Analyten durchsucht.

Abbildung 1: Zeitachse der Entwicklung von Non-Target-Screening

[1] (Bagheri et al. 1993) [2] (Hollender et al. 2019) [3] (Schulz und Lucke 2019)



Quelle: eigene Abbildung BfG

Die Entwicklung der Non-Target Screening Verfahren (NTS) begann in den 1990er Jahren. Allerdings haben erst Fortschritte in der hochauflösenden Massenspektrometrie, in den Datenwissenschaften und in der Informatik in den 2000er und 2010er Jahren die Technik vorangetrieben und in den Blickpunkt der Umweltanalytik gebracht (Abbildung 1). Heutzutage gibt es jährlich rund 900 Forschungsartikel zu dem Thema NTS und Wasseranalytik.

Non-Target-Screening in der Gewässerüberwachung

Die bisher mit hochsensitiven analytischen Methoden untersuchten „Targetstoffe“ in der Gewässerüberwachung – z.B. die prioritären und flussgebietsspezifischen Stoffe der Oberflächengewässerverordnung (OGewV) - stellen nur einen kleinen Teil der Stoffe anthropogenen Ursprungs dar, die tatsächlich in deutschen Fließgewässern vorkommen. Der Einsatz des NTS mit Hilfe der hochauflösenden Massenspektrometrie (in Kombination mit chromatographischer Trennung) kann einen Teil dieser Lücke schließen.

Das NTS bietet daher große Chancen für die chemische Gewässerüberwachung, vor allem zur Priorisierung von umwelt- und gesundheitsrelevanten Chemikalien, Entdeckung bisher nicht bekannter Umweltkontaminanten und der Identifizierung von Emissionsquellen (Hollender et al. 2019).

NTS Daten eröffnen Möglichkeiten für eine Reihe von Anwendungen im Gewässerschutz und in der Chemikalienbewertung:

- Frühwarnsystem zum Erkennen bislang nicht erfasster Schadstoffbelastungen in Gewässerproben

- Identifizierung der Quellen der Schadstoffbelastung
- Unterstützung bei der Überwachung der Umweltqualitätsnormen der prioritären sowie der flussgebietsspezifischen Stoffe, z.B. in Form eines Pre-Screenings
- Retrospektive Auswertungen in den digitalisierten Messungen, um beispielsweise die Verteilung eines aktuellen (bisher unbekannten) oder in der Zukunft relevanten Schadstoffs zu untersuchen
- Die Durchführung von NTS-Analysen wird bereits in Laboren angewendet, muss aber z.B. hinsichtlich der Vergleichbarkeit der Analysen, des Datenmanagements und der Datenauswertung noch weiter optimiert werden. Für einen übergreifenden Gewässerschutz ist es zu vermeiden, dass die Datensätze verschiedener Messprogramme voneinander isoliert bleiben. Die Informationen sollten stattdessen gemeinsam genutzt werden, um Ressourcen zu bündeln, Synergien zu schaffen und eine bessere und größere Datengrundlage zu bekommen. Erfahrungen, wie NTS Daten z.B. für eine gewässerübergreifende Auswertung kombiniert werden können, fehlen bislang weitestgehend.

1.2 Big-Data Ansätze und Digitalisierung

Big Data und die Forschungsgebiete des maschinellen Lernens (ML) als ein Teil der künstlichen Intelligenz (KI) wachsen seit den 2010er Jahren kontinuierlich und haben in vielen, wenn nicht allen, Industriebranchen und Forschungsbereichen, einschließlich der Umweltchemie, Anwendungen gefunden. Diese neuen Entwicklungen werden oft unter dem Begriff "Digitalisierung" zusammengefasst. NTS trägt zu diesem Wandel bei, da es sich um eine Analysemethode handelt, die stark auf automatisierte Datenverarbeitung und die Bearbeitung großer Datensätze angewiesen ist.

ML-Werkzeuge zur Analyse von Zeitreihen oder zur Mustererkennung sind als Open Source oder kommerzielle Produkte verfügbar. Diese könnten bei der Analyse und Bewertung von NTS-Daten sehr vorteilhaft eingesetzt werden. Beispielsweise kann ein ML Algorithmus mit der Hilfe eines annotierten Trainingsdatensatz Millionen von Intensitäts-Zeit-Verläufe analysieren und erlernen, welche Muster eher anthropogenen und welche eher natürlichen Ursprungs sind. Ein derart trainierter Algorithmus kann schnell neu gewonnene NTS-Daten mit diesen Informationen annotieren. Dies ist ein wichtiges Werkzeug, um die umfangreichen NTS-Ergebnisse aus Fließgewässern schneller und besser zu interpretieren und zu verwenden.

Die Herausforderung dabei besteht nicht in der Programmierung spezifischer ML-Algorithmen selbst, da diese sehr universell sind. Es sind vielmehr zwei Voraussetzungen erforderlich:

Erstens wird eine große annotierte Datenmenge benötigt, um den *supervised ML* Algorithmus zu trainieren. Bei ML Methoden ist die reine Anzahl von Bits nicht entscheidend, sondern die Anzahl der Fälle. Im oben genannten Beispiel sind die Anzahl und Länge der Zeitreihen (gruppierte Features) ausschlaggebend und nicht die Anzahl einzelner Features. NTS-Daten sind meistens sehr umfangreich, weil sie viele Spektren und ein sehr großes Signalrauschen beinhalten können. Die Anzahl der Zeitreihen aus zusammengefassten Messungen (nach Ausschluss von Signalrauschen und Hintergrund) fällt dabei oft überraschend gering aus. Es werden daher sehr viele Messungen benötigt, um seltene Signale sicher zu erfassen.

Zweitens ist die Qualität der Daten in Bezug auf die Annotationen im Trainingsdatensatz entscheidend. Je akkurater und umfangreicher die Informationen für jede Zeitreihe angegeben sind, desto besser kann das ML Verfahren trainiert werden.

Um diese Art von ML aufzubauen, sind daher große Datensammlungen hoher Qualität erforderlich, d.h. sie müssen nicht nur genau, sondern auch sehr gut annotiert sein.

1.3 Hintergrund

1.3.1 Ausgangslage und Bedarf

Die bisherige Erfahrung mit der Auswertung sehr großer NTS Datensätze (von mehreren tausend Proben) ist zum Zeitpunkt des Berichts (2024) begrenzt. Eine überregionale Auswertung über viele Datensätze und die Auswertung von Langzeittrends erfordert die Sammlung und Harmonisierung einer Vielzahl von Proben bzw. einzelner Messdateien, was gegenwärtig ein überwiegend manueller, langsamer Prozess ist. Die Rohdaten werden in der Auswertesoftware geöffnet und für die Fragestellung neu prozessiert, ein rechenintensives Vorgehen. Die Software, die dabei angewendet wird, ist meistens nicht für die Durchsuchung tausender Messdateien ausgelegt, sodass gruppenweise gearbeitet werden muss und die Auswertung dadurch weiter verlangsamt wird.

Softwaretools für die Prozessierung wurden in den letzten Jahren rasch weiterentwickelt. Die Entwicklung wurde aber nicht mit einem ebenso schnellen Aufbau von Software für die Bereitstellung und Reporting der Ergebnisse begleitet. In der bisherigen Arbeitsweise ist die Prozessierung (peak-picking, Datenbereinigung) nicht strikt von der Auswertung der prozessierten und gereinigten Daten getrennt. Diese Trennung ist für die Bereitstellung der NTS Ergebnisse für Externe jedoch essentiell, um schnell und fokussiert an der eigenen Fragestellung arbeiten zu können.

Durch die Größe der NTS Daten hat die Anwendung von Suchmaschinen und interaktiven Visualisierungen besonders große Vorteile. Fortschritte in anderen Bereichen, wie z.B. das Data Warehousing in der Wirtschaftsinformatik, können für die Analyse von NTS-Daten angewendet werden. Die Bereitstellung solcher Datenzentren über HTTP (Firmen-intern oder im Internet) werden häufiger verwendet und dementsprechend werden auch vielseitig einsetzbare Softwarelösungen angeboten.

Die Heterogenität von NTS Daten aus unterschiedlichen Laboren und auch die spezifischen Schwierigkeiten NTS Daten miteinander zu vergleichen (keine Konzentration, unklare Identität), erfordert neue Lösungen für die Harmonisierung. Hier werden Algorithmen entwickelt werden müssen, da die Datenmenge für die menschliche Bearbeitung zu groß ist.

1.3.2 Vorhergehende Entwicklungen

Die EAWAG und ein Schweizer Unternehmen hat die Software enviMass (Schymanski et al. 2014) für die Untersuchung von Langzeit-Trends durch eine zeitlich geregelte Probenahme an einer Messstelle entwickelt. Die Betrachtung von regional verteilten Proben und die Online Bereitstellung (web-hosting) war zu Beginn des Projektes noch in der Entwicklung. Sie wurde dann in dem BMBF geförderten Projekt K2I (Müller 2023) und durch das internationale “Rhein Projekt Non-Target-Screening” (Ondruch und Heinz 2024) weiterentwickelt. Die Software fokussiert sich allerdings stark auf die Priorisierung von Unbekannten und ist nicht vorrangig für die schnelle Durchforstung von Ergebnissen in der Stoffregulierung oder der Wasserbewirtschaftung entwickelt worden.

Andere Auswertetools wie z.B. mzMine (Schmid et al. 2023), PatRoon (Helmus et al. 2022) oder herstellerspezifische Software bilden eine solide Auswahl an Tools für die Auswertung an einem Desktop-PC oder in einigen Fällen auch auf einem Server. Auf diese Entwicklungen kann perspektivisch aufgebaut werden.

Für die in diesem Projekt benötigten Probenmengen sind Serverlösungen unausweichlich. Investitionen in Rechenzentren und Linux-basierten Servern für das wissenschaftliche Rechnen wurden in den letzten Jahren an einigen Ressortforschungszentren, darunter auch an der BfG, vorangetrieben. Die errungenen Kapazitäten sind eine Voraussetzung für die automatisierte Prozessierung und Harmonisierung von NTS Datensätzen im überregionalen Kontext.

Die Digital Sample Freezing Plattform (DSFP) (Alygizakis et al. 2019) und das Global Natural Products Social Molecular Networking (GNPS) (Nothias et al. 2020) sind wichtige Entwicklungen von Online-Diensten für das NTS. DSFP ist eine Entwicklung der Universität Athen und des Environmental Institute SRO im Auftrag des NORMAN Netzwerks und wird vorwiegend für die Bereitstellung von Forschungsdaten verwendet. Mit dieser Plattform können Messdaten mit entsprechenden Metadaten über eine Upload-Seite hochgeladen werden und nach einer internen Prozessierung über ein Online-Portal (Dashboards) durchsucht werden. GNPS, von der Universität California–San Diego, ist eine offene Online-Plattform für die Auswertung und Annotierung von LC-HRMS Daten für Metabolomics und den Austausch von Spektren natürlicher Substanzen. Die direkte Anwendung dieser akademischen Dienste ist für jedes Labor möglich. Sie bieten jedoch nicht die notwendige Datensouveränität und strukturelle Unabhängigkeit, die für eine nationale Datenbank mit Wasserqualitätsdaten notwendig ist.

Die Entwicklung der online recherchierbaren Datenprodukte im Umweltmanagement unterliegt einer langen Entwicklungsgeschichte. Zwei Beispiele in diesem Kontext sind die internationale Wasserqualitätsdatenbank GEMStat, die seit 2014 durch die BfG und das ICWRGC weiterentwickelt und gehostet wird (ICWRGC 2024) und das Chemikalien Infoportal Cheminfo (Cheminfo 2024), welches durch das UBA entwickelt wurde und umweltrelevante Eigenschaften von Substanzen und Mischungen für den Umweltschutz und die Gefahrenabwehr bereitstellt.

1.3.3 Aktuelle Entwicklungen

Während der Laufzeit dieses Projekts wurden mehrere nationale und internationale Projekte zum Aufbau von Datenbanken gestartet. Dazu gehören nationale Projekte in mehreren europäischen Ländern wie NTSuisse (Singer 2024) in der Schweiz und Aquaplexus in Dänemark (Kjøller 2024). Zu den internationalen Projekten gehören das NTS Projekt für den Rhein, das von der IKSR (Ondruch und Heinz 2024) koordiniert wird, und Initiativen innerhalb des PARC-Konsortiums (Europäische Partnerschaft für die Bewertung chemischer Risiken) zum Aufbau von Frühwarnsystemen auf der Grundlage von NTS (Niarchos et al. 2023).

1.4 Strategie und Zielsetzung

Basierend auf den oben genannten Bedürfnissen für die Anwendung von NTS im Kontext des Gewässerschutzes, war das übergeordnete Projektziel ein Konzept für die überregionale Sammlung von NTS Gewässerdaten und die Datenbereitstellung über das Internet. Das Konzept wurde während des Projektes prototypisch implementiert und im Testbetrieb angepasst und weiterentwickelt.

Weitere wichtige Ziele wurden am Projektanfang definiert:
Die NTS-Daten sollen Qualitätsstandards unterliegen und gesichert gespeichert werden. Das User-Interface für das Web-Portal soll es ermöglichen, die Daten zu durchsuchen und zu visualisieren. Vereinfachte, grundlegende Auswertungen sollen direkt in der Webanwendung integriert sein, zum Beispiel die Sortierung der Befunde nach Kategorien wie Intensitätsänderungen. Die Daten sollen exportierbar sein, damit Auswertungen mit geeigneter Software durchgeführt werden können. Letztlich sollen die Daten für die spätere Anwendung von KI vorbereitet und erste Ansätze getestet werden.

2 Konzeptentwicklung und operative Ziele

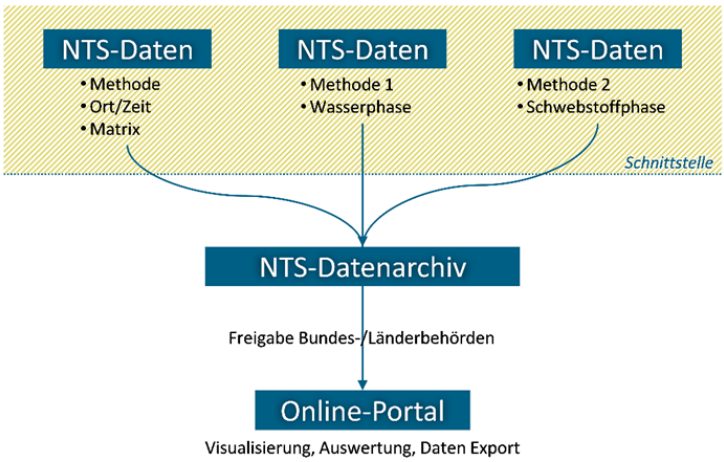
Das übergeordnete Ziel dieses Projekts ist die Entwicklung einer neuartigen und dauerhaften Datenbank- und Analyseplattform für NTS-Daten von Fließgewässern. Daten aus verschiedenen Laboren sollen kombinierbar und eine aggregierte Auswertung der Daten möglich sein. Dieses Vorhaben lässt sich in drei operative Ziele unterteilen: 1) Strategien zur Harmonisierung von NTS-Daten, 2) Implementierung und Aufbau einer zentralen Datenablage, 3) Bereitstellung von prototypischen Datenrecherche- und Visualisierungstools.

2.1 Strategien zur Harmonisierung NTS-Daten

Das erste operative Ziel dieses Projektes ist die Entwicklung von Strategien und die Implementierung von Harmonisierungsprotokollen für NTS-Daten aus verschiedenen Quellen, einschließlich Messungen in diversen Umweltmatrizes, Daten aus verschiedenen Regionen und von verschiedenen Laboren. Heterogene NTS-Datensätze müssen vergleichbar gemacht werden, sodass ein Feature aus einem Datensatz mit einem Feature eines anderen Datensatzes verglichen werden kann.

Die Harmonisierung von NTS-Datensätzen ist für die gemeinsame, interregionale Datenauswertung unerlässlich. „Harmonisierung“ beschreibt in diesem Kontext sowohl die Sicherstellung einer minimalen Vergleichbarkeit der Daten (Analysemethoden und Software für die Prozessierung), als auch die Umstrukturierung der gesammelten Datensätze in ein einheitliches Format. Ein Kriterium kann beispielsweise sein, dass alle verwendete Methoden bzw. Workflows eine Blindwertkorrektur durch Methodenblanks durchführen, um eine vergleichbare Qualität der aggregierten Daten zu haben. Die exportierten Datensätze werden anschließend durch ein Skript in ein einheitliches Datenmodell umformatiert.

Abbildung 2: Harmonisierungskonzept für das NTSPortal



Quelle: eigene Abbildung BfG

Um dieses Ziel zu erreichen, muss die Heterogenität der anzunehmenden Daten bekannt und die minimal benötigte Vergleichbarkeit der Daten erfüllt sein. Ein Konsens über die Datenqualität und Vergleichbarkeit unter den Nutzern des Datenarchivs ist dabei eine wesentliche Voraussetzung.

2.2 Implementierung einer zentralen Datenbank

Nach der Aggregation und Zusammenführung besteht das nächste operative Ziel darin, die Daten in einer geeigneten zentralen Datenbank zu sichern und den Zugriff darauf zu gewährleisten. Die Merkmale dieser Datenbank sind folgende: i) ein sicheres und leicht zugängliches Datenbanksystem, das die zeitgemäßen IT-Standards zur Datensicherheit und Kommunikation implementiert, und ii) die spezifischen Bedürfnisse der Institutionen des Bundes und der Länder zu erfüllen. Zu diesen Anforderungen gehören:

a) **Horizontale Skalierbarkeit**. Derzeit ist nicht bekannt, wie viele Daten in der Datenbank gespeichert werden müssen. Dies hängt von der Anzahl potenzieller Nutzer*innen ab, sowie vom Umfang ihrer Messdaten. Daher ist ein System erforderlich, das leicht erweitert werden kann und bei wachsenden Datenmengen performant bleibt.

b) **Ausführlich dokumentierte APIs**. Da die Datenbank die Grundlage für die Analyseplattform bildet, die höchstwahrscheinlich in R oder Python programmiert werden wird (den gebräuchlichsten Computersprachen in den Datenwissenschaften), müssen die gespeicherten Daten über APIs typ-offen für unterschiedliche Sprachen leicht zugänglich sein.

c) **Flexibilität**. NTS-Daten aus verschiedenen Laboren können sehr heterogen sein. Die flexible Anpassung von Datentypen und das Hinzufügen, Ändern und Löschen von Datenfeldern ist erforderlich. Diese Änderungen können durch Fortschritte in der Forschung oder durch Angleichung an Standards erforderlich sein.

2.3 Bereitstellung von prototypischen Datenrecherche- und Visualisierungstools

Das dritte operative Ziel des Projektes ist die Implementierung prototypischer Datenanalyse-Werkzeuge, um die in der Datenbank gespeicherten Daten auswerten und visualisieren zu können. Diese Plattform soll Nutzer*innen der Behörden und Forschung zugänglich sein. Typischerweise wird dafür ein Online-Dienst bereitgestellt, der über eine sichere (verschlüsselte) Verbindung verfügt. Innerhalb der Projektlaufzeit werden die Grundlagen für eine voll funktionsfähige Analyseplattform und prototypische Analysewerkzeuge entwickelt. Dazu gehören unter anderem die interaktive Suche nach Features und die Visualisierung von Zeitreihen oder räumlichen Verteilungen eines Stoffes auf einer Karte in einer Dashboard-Struktur. Dies wird über HTTPS zugänglich gemacht

Wie oben erwähnt, ist die Datenbank über APIs für Datenanalyse-Workflows in anderen Programmiersprachen zugänglich. Dies ermöglicht die einfache Einbindung von Rohdaten in Skripte für die statistische Auswertung oder die Entwicklung neuer, interaktiver Analysewerkzeuge.

2.4 Projektphasen

Die Arbeiten zu den drei Hauptzielen werden in zwei Phasen unterteilt. In Phase 1 werden zunächst nur Features behandelt, die bereits bekannt sind (annotiert, AP3). In Phase 2 wird die Datenbank für alle Features (bekannt und unbekannt, AP4) ausgeweitet (Tabelle 1). Die schrittweise Steigerung der Komplexität, von Phase 1 zu Phase 2, ermöglicht eine höhere Qualität bei der Entwicklung, denn Ergebnisse aus “einfacheren” Fällen können als Grundlage für Qualitätssicherung von komplexeren Fällen angewendet werden. Zum Beispiel können Ergebnisse mit annotierten Daten aus Phase 1 in Phase 2 umfassender geprüft bzw. plausibilisiert werden.

Tabelle 1: Entwicklungsphasen (AP3 und AP4)

AP	Phase	Anzahl Geräte/ Methoden	Anzahl Matrices	Welche Features
3	1a	1	1	Annotierte Features
	1b	1	≥ 2	Annotierte Features
	1c	≥ 2	≥ 2	Annotierte Features
4	2a	1	1	Alle Features
	2b	1	≥ 2	Alle Features
	2c	≥ 2	≥ 2	Alle Features

3 Umsetzung

3.1 Begleitkreise

Der Begleitkreis besteht aus Expert*innen aus dem Non-Target-Screening Bereich, sowohl aus der akademischen als auch aus der behördlichen deutschsprachigen Gewässerbeobachtung (LfU Bayern, UFZ, LUBW, LANUV, NLWKN, LfU Brandenburg, Landeslabor Berlin-Brandenburg, BWB, BAM, BAFU, Eawag, AUE Basel, UBA Wien und Uni-Luxemburg). Das Ziel der Gruppe war es, den aktuellen Stand des Projektes nach außen zu kommunizieren und gleichzeitig Anregungen und Hinweise zu erhalten.

Die offizielle Kick-Off Veranstaltung für den Begleitkreis fand am 03.12.2020 als Online-Veranstaltung mit 16 externen Expert*innen statt (siehe B.3). Es fanden zwei weitere Treffen des Begleitkreises am 31.01.2022 und 23.01.2023 statt. Das letzte Treffen wurde mit dem Projektabschlusstreffen (14.12.2023) kombiniert und als Präsenzveranstaltung in Berlin mit mehr als 80 Teilnehmern abgehalten. Der Begleitkreis bleibt in den nachfolgenden Vorhaben (Gewässerbeobachtung der Zukunft und NTSPortal) bestehen.

Eine Unterarbeitsgruppe des Begleitkreises mit einem Fokus auf Qualitätssicherung wurde im ersten Jahr des Projektes gegründet. Die Gruppe bestand aus neun Expert*innen aus sechs Instituten (UFZ, LfU Bayern, Eawag, AUE Basel, UBA und BfG). Es fanden insgesamt sechs Online-Meetings statt. Diese dienten dem Austausch zum Thema Datenqualität und Harmonisierung der Daten.

Einmalig fand ein Online Treffen für Stakeholder der Bundesländer am 12.04.2021 statt. Das Treffen wurde über die Sitzung des Ausschusses oberirdischer Gewässer und Küstengewässer (AO) der Bund/Länder-Arbeitsgemeinschaft Wasser (LAWA) initiiert. Es nahmen Vertreter*innen aus fünf Bundesländern sowie schweizerischer und österreichischer Umweltbehörden teil.

3.2 Aufbau und Weiterentwicklung der Spektrenbibliothek

Die MS^2-Spektren und Retentionszeitbibiothek (kurz: Spektrenbibliothek) beinhaltet Referenzspektren und Retentionszeiten von analytischen Standards, die für den Abgleich mit Messdaten verwendet werden (Bibliothek-gestütztes Screening) (Jewell et al. 2019). Mit Hilfe der Spektrenbibliothek werden bei der Prozessierung NTS-Daten mit Substanznamen annotiert (siehe 3.3.5.1). Die Bibliothek wird daher stets erweitert und mit priorisierten und/oder neu identifizierten Substanzen aktualisiert. Die Klassifizierung der Stoffe wird derzeit vom UBA aktualisiert, um unter anderem die einzelnen Verbindungen den Stoffgesetzen zuzuordnen.

Derzeit enthält die Bibliothek 1431 Verbindungen und 27428 Spektren von drei Behörden (BfG, LfU Bayern und UBA). Die Spektren, die durch die BfG aufgenommen wurden (19817), sind auf MassBank Europe veröffentlicht (massbank.eu). Aktualisierungen können über Skripte (Abbildung 3) mit geringem Aufwand in das MassBank-Record-Format exportiert und an MassBank weitergegeben werden.

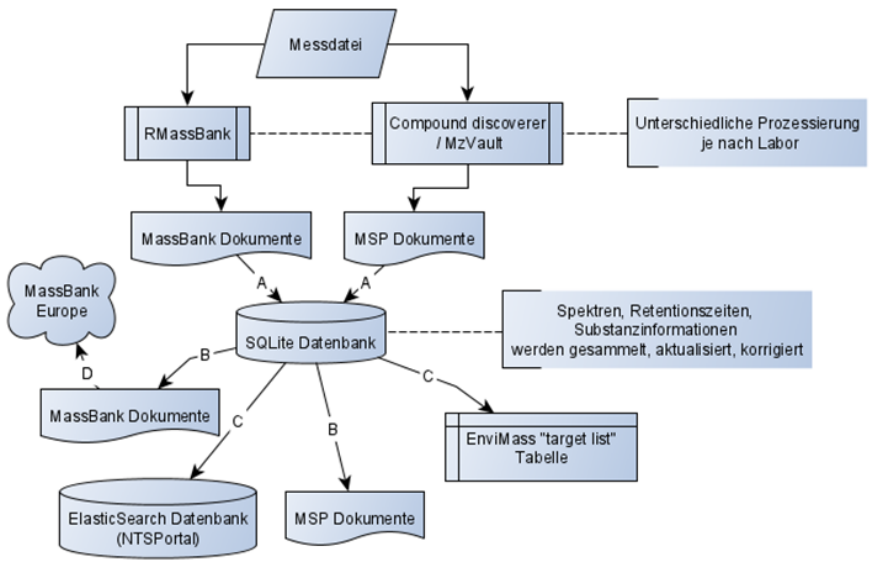
3.2.1 Zusammenarbeit zwischen Laboren über den Austausch von Spektrenbibliotheken

Beim Aufbau der Datenbank für annotierte Daten in Phase 1 des Projekts (siehe 2.4) war erwartungsgemäß festzustellen, dass die Vergleichbarkeit der angewendeten Screening-Bibliotheken einen starken Einfluss auf die Vergleichbarkeit der Ergebnisse hat. Der Austausch von Spektren und die Bildung einer „kollektiven“ Spektrenbibliothek (*Engl.:* CSL) zwischen Partnern des NTSPortals wurde daher als wichtiges Standbein der Harmonisierung gesehen.

Bei der Zusammenarbeit mit externen Laboren gab es eine hohe Bereitschaft, Standard-Spektren auszutauschen und Spektrenbibliotheken anzugleichen. Ein Konzept zur Bildung eines „kollektiven“ Datenbanksystems, das in diverse, bereits existierende Auswerteworkflows integriert werden kann, wurde deshalb verfolgt (Abbildung 3). Die bestehende SQLite Datenbank, die bereits im Vorläuferprojekt begonnen wurde (Jewell et al. 2021), bildet den zentralen Datensatz für die Harmonisierung (Import), Kuratierung und Export der Bibliothek in andere Formate. Verschiedene selbst geschriebene Python-Skripte erlauben den Import von sowohl MassBank als auch MSP (NIST) formatierten Text-Dateien. Aus der zentralen SQLite-Datenbank, die laufend auf Fehler geprüft wird (Kuratierung), können Bibliotheken in anderen Datenbank-Formaten gebildet werden (Export). MSP Dokumente sind sowohl das Export- und Importformat von der Thermo *MzVault* Software, die von Orbitrap Nutzer*innen verwendet wird, als auch ein Importformat für die Open-Source Software *PatRoon* (Helmus et al. 2022). Das MassBank-Format dient zur Veröffentlichung der Spektren auf MassBank Europe und als alternatives Import-Format für *MzVault*. Die geräteunabhängige Auswertesoftware *enviMass* (Loos 2024) kann die Datenbank in Form einer „Target List“ (CSV-Datei) anwenden. Die Spektrenbibliothek wird als JSON-Datei in *ElasticSearch* importiert, damit sie zusammen mit den NTS Ergebnissen im NTSPortal dargestellt werden kann. Das selbst geschriebene R-Paket *ntsworkflow* nutzt die Datenbank direkt als SQLite-Datei.

Abbildung 3: Das Konzept der „Kollektiven Spektrenbibliothek“ für die Angleichung von Spektrenbibliotheken

A: Python-Skripte mit Hilfe von Modulen (u.a.) sqlalchemy; B: R-Skripte mit Hilfe von Modulen (u.a.) Spectra; C: R-Skripte (hauptsächlich Eigenentwicklungen); D: GIT Workflow zum Hochladen in das MassBank-Data Repository



Quelle: eigene Abbildung BfG

Die Spektrenbibliothek wurde mit der Hilfe von externen Spektren des LfU Bayerns und des UBA erweitert. Vom LfU wurden 7522 Spektren von 702 Verbindungen beigesteuert, davon 292 neue Verbindungen. Die Daten wurden als MSP-Dateien bereitgestellt. Das UBA lieferte 6 Verbindungen (UV-Filter), diese wurden als Rohdaten (Messdateien) geliefert und mit dem RMassBank Workflow (Stravs et al. 2013), der auch intern verwendet wird, prozessiert.

3.2.1.1 Angleichung der Retentionszeiten

Retentionszeiten sind methodenspezifisch und müssen in einer Standardbibliothek für jede Methode einzeln bestimmt werden. Selbst wenn zwei Labore eine harmonisierte chromatographische Methode verwenden, können Unterschiede in den Retentionszeiten entstehen. Dies kann beispielsweise an einem unterschiedlichen instrumentellen Aufbau liegen.

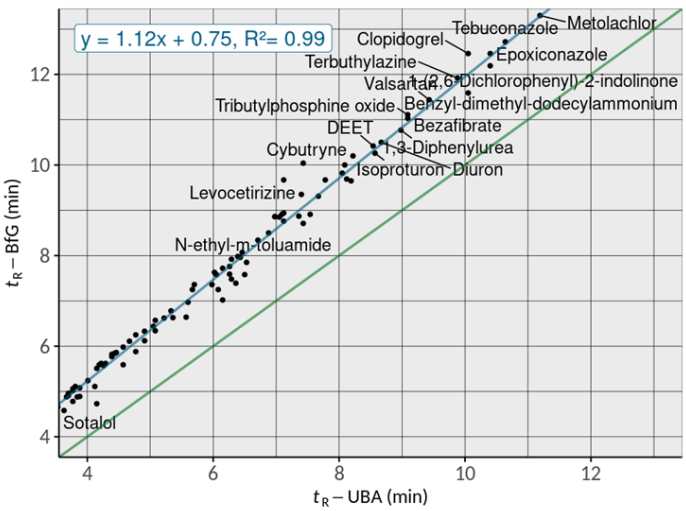
Damit für jeden Bibliothekseintrag Retentionszeiten für alle angewendeten Methoden zur Verfügung stehen, wurden die Retentionszeiten mit Regressionsmodellen umgerechnet.

Ein Beispiel sieht folgendermaßen aus: Im ersten Schritt werden Retentionszeiten von bekannten Substanzen in der BfG-Methode und in einer neuen Methode für die Erstellung eines Modells ausgewählt. Um die Retentionszeiten in der neuen Methode zu finden, werden Rohdaten ohne den Abgleich von Retentionszeiten, sondern nur über einen Abgleich der MS^2 Spektren, ausgewertet. Als Ergebnis des Modells werden die Einträge der Spektrenbibliothek um Retentionszeiten für die neue Methode ergänzt. Sodass die Rohdaten, die mit der neuen Methode gemessen worden sind, mit den berechneten Retentionszeiten (für alle Substanzen) ausgewertet werden können. Dieses Vorgehen muss für weitere, neu hinzugekommene Methoden wiederholt werden.

Die Retentionszeiten des UBA und der BfG wurden über ein lineares Modell umgerechnet (Abbildung 4). Die Retentionszeiten des LfU waren ausreichend vergleichbar, sodass diese ohne Änderung übernommen wurden (y = 1.00x + 0.30). Für die Retentionszeiten des LANUV wurde ein Generalized Additive Model angewendet (GAM) (Stanstrup et al. 2015). Bei der Auswertung (Bib.-Screening, (Jewell et al. 2019)) werden labor- und methodenspezifische Retentionszeiten passend zur Quelle der Messdaten genutzt.

Abbildung 4: Umrechnung von Retentionszeiten für Spektren des UBA

Messung von einem Standard-Mix (10 µg/L) mit 95 Verbindungen. Grüne Linie stellt die Identitätsgleichung dar.



Quelle: eigene Abbildung BfG

3.3 Phase 1: Aufbau einer Datenbank für annotierte Daten

3.3.1 Auswahl des Datenbanksystems

Unter Berücksichtigung der Bedürfnisse der Datenbank und nach Absprache mit den analytischen und IT-Fachreferaten der BfG wurde das Datenbanksystem *ElasticSearch* als Grundlage für die Implementierung gewählt. *ElasticSearch* ist eine „NoSQL-Datenbank, die Daten und Dokumente flexibel in einem sogenannten „Datensee“ speichert. Sämtliche Kommunikation mit der Datenbank (auch Datenimport und -export) kann über APIs (RESTful) durchgeführt werden. Die einzelnen Dokumente werden in sogenannten Indices gehalten. Durch die Indizierung bietet *ElasticSearch* performante Suchfunktionen und kann auf verteilten Serversystemen betrieben werden. Sie ist somit flexibel und skalierbar.

ElasticSearch ist die weltweit am häufigsten verwendete Enterprise-Suchmaschine und kann alle Anforderungen des geplanten NTS-Datenspeichers und der Analyseplattform erfüllen. Diese umfassen unter anderem:

a) Als verteilter Dokumentenspeicher werden die Daten auf getrennten Knoten eines Rechenclusters gespeichert. Damit können Daten für zusätzliche Sicherheit repliziert werden und die Datenbank kann sehr einfach skaliert werden. Um die Datenbank zu expandieren, müssen dem Cluster nur weitere Knoten hinzugefügt werden.

b) *ElasticSearch* hat eine umfangreich dokumentierte RESTful-API, eine standardisierte Schnittstelle für webbasierte Anwendungen. Diese kann verwendet werden, um die Datenbank mit maßgeschneiderter Software, die in einer beliebigen Programmiersprache erstellt wird, zu verbinden.

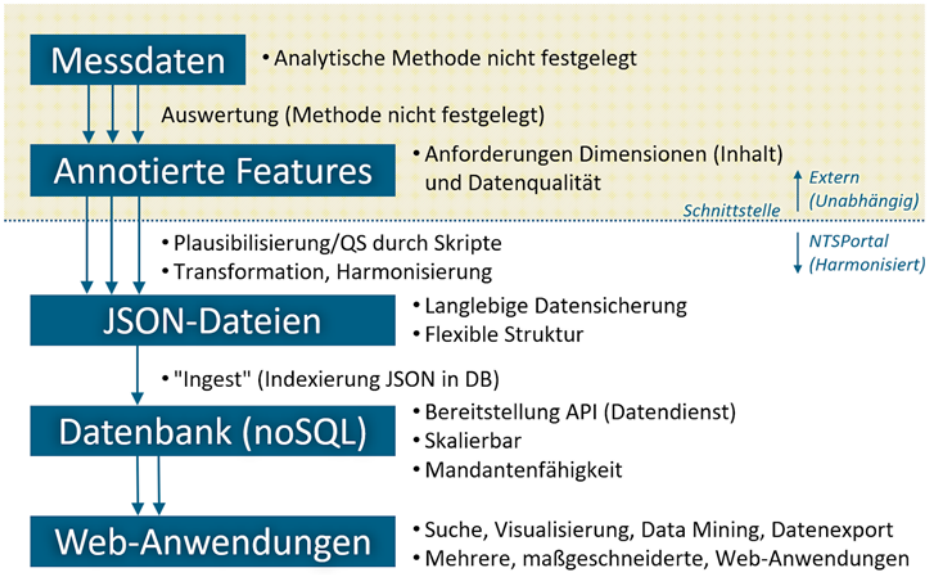
c) Als Dokumentenspeicher vom Typ noSQL ist es sehr flexibel, kann heterogene Daten speichern und um neue Felder erweitert werden,

d) Die Volltext-Suchmaschine in *ElasticSearch* kann zur Suche in unstrukturierten Textfeldern verwendet werden, z.B. in Kommentarfeldern, die Informationen zu bekannten Quellen, Fortschritt der Identifizierung oder Stoffverwendungskategorien enthalten können,

e) *ElasticSearch* wurde mit quelloffenen Bibliotheken aufgebaut und die Basis-Version kann kostenlos auf Computern jeder Größe installiert werden, so dass die in diesem Projekt entwickelten Strategien und Werkzeuge leicht von externen Institutionen, die eine ähnliche Struktur implementieren wollen, repliziert werden können.

ElasticSearch ist Teil eines größeren Softwarekonglomerats, das als *ELK Stack* bekannt ist und weitere Werkzeuge beinhaltet, beispielsweise Software für die Bearbeitung und Import von Logdaten (*Logstash*) und eine Datenanalyse- und Visualisierungsplattform, genannt *Kibana*. Letztere ist speziell konzipiert für die Analyse von Zeitreihen oder räumlich verteilten Daten. Sie bietet vorgefertigte Bausteine für den Aufbau von Dashboards, welche für eine generelle Suche und Darstellung genutzt werden können. Zudem bietet sie die Möglichkeit, weitere statistische Analysen durchzuführen, z.B. eine Sortierung anhand einer im Vorfeld (extern) berechneten Regressionsanalyse. Es ist daher mit begrenzten Entwicklungskosten möglich, die vorgefertigten Bausteine in *Kibana* für die Erstellung von Prototyp Dashboard-Ansichten zu verwenden. Dies schließt die Verwendung weiterer maßgeschneiderter Software, für anwendungsspezifische Analyse, nicht aus. Im späteren Verlauf des Projektes wurde begonnen, eine zweite Webanwendung mit Hilfe des *Shiny* Frameworks zu entwickeln (siehe 3.3.9). Beide Dashboards (*Shiny* und *Kibana*) dienen unterschiedlichen Zwecken und sollen parallel gehostet werden.

Abbildung 5: Feinkonzept NTS-Datenbank



Quelle: eigene Abbildung BfG

3.3.2 Datenbankschema für annotierte Daten

Die NTSPortal Datenbank wird unterteilt in sogenannte Indizes (separate Ablagen für JSON-Dokumente). Ein Non-Target Feature entspricht einem JSON-Dokument und beinhaltet sowohl analytische Daten aus der LC-MS Messung, als auch Informationen zur Probenahme, Analysemethode, Datenquellen und Lizenzbestimmungen (Abbildung 6).

Zum Aufbau der NTSPortal Datenbank wurden nur die Features angewendet, die durch einen Treffer in der Spektrenbibliothek mit einem Substanznamen annotiert wurden (Abbildung 6), diese können mit unterschiedlichen Analyse- und Auswertemethoden aufgenommen worden sein (dargestellt durch mehrere Pfeile). Während der Projektlaufzeit wurden verschiedene Datenquellen exemplarisch getestet. Für einen qualitätsgesicherten Import von externen Daten in das NTSPortal, wurden Skripte zur Harmonisierung der Daten angewendet. Eine Beschreibung der bisher angewendeten Methoden findet sich im Teil Harmonisierung (3.3.6).

Insgesamt befinden sich am Projektende mehr als 700.000 Features, unterteilt in mehr als 12 verschiedene Indizes im NTSPortal. Dazu gehören Features von täglichen, 24 h Oberflächenwasser Mischproben aus der Rhein-Messstation der BfG in Koblenz und der Elbe Messstation in Tangermünde für das Jahr 2019, sowie jährliche Schwebstoffmischproben von 15 Standorten aus den Jahren 2005 bis 2020. Diese Proben wurden von der Umweltprobenbank bereitgestellt und von der BfG in einem weiteren Forschungsvorhaben analysiert. Seit 2021 werden die Proben im Routinemessprogramm der Umweltprobenbank durch NTS untersucht. Die Unterteilung in Indizes erlaubt die rollenspezifische Kontrolle der Zugriffsrechte auf bestimmte Datensätze (Abschn. 3.6.2).

3.3.3 Sicherung der Daten

Die aktuelle Strategie für die langfristige Datenspeicherung besteht darin, alle importierten Dokumente in einem textbasierten Format (JSON) als sogenannte "flat-files" zu sichern (Abbildung 6). JSON ist das native Format für den Import von Daten in *ElasticSearch*, so dass diese Dateien jederzeit zum einfachen Neuaufbau der *ElasticSearch*-Datenbank verwendet werden können, sollte dies notwendig sein. Alternativ können sie auch als Datenquelle für andere Datenbanken dienen. Zusätzlich dazu werden regelmäßig sogenannte „Snapshots“ der

ElasticSearch Datenbank erstellt, um den aktuellen Zustand zu sichern. Die Dateien (Snapshots und JSONs) werden auf dem zentralen Datenspeicher der BfG abgelegt, der über mehrere Sicherungskopien und eine Versionierung verfügt.

Die *ElasticSearch* Datenbank wurde am Ende des zweiten Projektjahrs auf eine neue Infrastruktur umgesiedelt (siehe Abschn. 3.6). Diese verfügt über ein automatisches Sicherungssystem, das die komplette virtuelle Maschine (VM), inklusive aller Rohdaten (Datenbankinhalte) und Software, nach einem Ausfall selbständig neu aufsetzen kann.

Abbildung 6: Beispiel eines Dokuments (JSON) mit ausgewählten Feldern

mz: Masse-Ladungs-Verhältnis (Da), **rt**: Chrom. Retentionszeit in Min, **start**: Probenahmedatum, **duration**: Probenahmedauer (Tage), **station**: Codierung Messstelle, **pol**: Polarität, **licence**: Nutzungslizenz, **name**: Substanzname, **area**: Chromatografische Peakfläche, **area_is**: Chromatografische Peakfläche des internen Standards, **tag**: Kennzeichen für Filterung, **gkz**: Gewässerkennzahl, **km**: Fluss-km, **eic**: Extrahiertes Chromatogramm, **ms1**: Massenspektrum, **ms2**: Fragmentspektrum


```
{
  "mz": 237.1025,
  "rt": 9.0,
  "area": 321.5,
  "area_is": 0.1404,
  "name": "Carbamazepine",
  "start": "2019-03-01",
  "duration": 1,
  "station": "elbe_tan_l",
  "matrix": "water",
  "pol": "pos",
  "licence": "dl-de/by-2-0",
  "date_import": 1604071478,
  "data_source": "BfG",
  "gkz": 5,
  "km": 388.2,
  "loc": {"lat": 52.549876, "lon": 11.983321},
  "eic": [
    {"time": 301, "int": 10},
    {"time": 302, "int": 11},
    {"time": 303, "int": 8}
  ],
  "ms1": [
    {"mz": 237.1025, "int": 1.0},
    {"mz": 238.1025, "int": 0.1}
  ],
  "ms2": [
    {"mz": 194.1212, "int": 1.0},
    {"mz": 192.1345, "int": 0.8}
  ],
  …
}
```


Quelle: eigene Abbildung BfG

3.3.4 Metadaten und Dokumentation

Metadaten und Dokumentation werden auf der Wiki-Seite der NTSPortal Repository (github.com/bafg-bund/ntsportal/wiki) veröffentlicht[1]. Damit können Änderungen und Erweiterungen direkt an Anwender*innen und Interessierte weitergegeben werden. Es enthält u.a. auch interne Regeln zur einheitlichen Dateibezeichnungen. In Abschn. 3.3.4.1 ist das Inhaltsverzeichnis zum Zeitpunkt des Berichts (2024) des Wikis dargestellt. Funktionen des

[1] Die Repository ist zurzeit nur intern sichtbar und wird zum Projektabschluss veröffentlicht.

ntsportal R-Pakets werden zusätzlich, wie üblich, im Quelltext dokumentiert (github.com/bafg-bund/ntsportal).

3.3.4.1 Übersicht über die Inhalte des NTSPortal Wikis zum Zeitpunkt des Berichts

- Connecting to ElasticSearch API
- Using Query DSL for Elastic
  - Searching with ElasticSearch
  - Aggregations with ElasticSearch Query DSL
- Using R Package 'elastic' for retrieving data
- Using the Python ElasticSearch client
- Database structure and index management
  - Structure of NTSPortal
  - Delete index
  - Change alias or index name
  - Add field to existing index
  - Create index backup step-by-step
  - Index mappings (DB schema)
    - Index mapping for dbas indices
    - Index mapping for dbas_analysis indices
    - Index mapping for msrawfiles indices
    - Index mapping for nts indices
    - Index mapping for spectral_library indices
    - Index mapping of ufid indices
- Adding files to msrawfiles
  - Datafile naming conventions
  - Add measurement file to msrawfiles step by step
- Make changes to existing documents
  - Change path msrawfiles
- NTSPortal server infrastructure
- Naming conventions for NTSPortal
  - Station naming
  - Chromatographic methods
  - Sample and data sources
- Using ntsportal.bafg.de
- Anomaly Detection in Kibana
- Library screening processing algorithm

3.3.5 Automatisierte Messdatenprozessierung

Die Prozessierung von Messdaten (Rohdateien) wurde mit Skripten durchgeführt, die auf dem sogenannten Bibliothek-gestützten Screening Verfahren (Jewell et al. 2019) basieren. Während des Projekts wurde festgestellt, dass eine autark laufende Prozessierung unabdingbar ist.

Um eine autarke Prozessierung zu erreichen, wurden die Skripte schrittweise weiter optimiert, sodass ein höherer Automatisierungsgrad erreicht werden konnte. Zum Projektende können alle vorhandenen Rohdaten prozessiert und in die *ElasticSearch* Datenbank geladen werden. Neue Daten werden erkannt und ebenfalls prozessiert. Die Berechnungszeit für 8000 Rohdaten betrug ca. 70 h. Wichtige Schritte für die zukünftige Weiterentwicklung werden in Abschnitt 4 (Ausblick) erläutert.

3.3.5.1 Funktionsweise der automatisierten Datenprozessierung

Alle Messdateien werden über den *ElasticSearch* Index *msrawfiles* verwaltet. Hier werden alle notwendigen Informationen für die Auswertung der jeweiligen Messdatei gehalten: sämtliche

Metadaten zur Probe und Messung, alle notwendigen Parameter für die Prozessierung und letztendlich auch der Speicherort der Messdateien[2]. Die Messdateien selbst werden in *ElasticSearch* nicht gespeichert.

Ein Prozessierungsskript durchläuft den *msrawfiles*-Index, führt das Bibliothek-Screening für jede Datei durch und speichert alle Ergebnisse in einen Index für Ergebnisse. Das R-Paket *ntsworkflow* und die oben genannte Spektrenbibliothek werden als Grundlage für diese Prozessierung angewendet. Weitere Dokumentation zur Funktionsweise der Skripte befindet sich in den Repositories der R-Pakete *ntsworkflow* und *ntsportal.*

3.3.6 Datenimport

Um die Flexibilität des Konzepts (Abbildung 5) darzustellen, wurden Daten aus unterschiedlichen Quellen in das NTSPortal importiert.

- Massenspektrometrische Messdateien (Sciex ‘wiff’, Thermo ‘raw’). Über diesen Weg wurden auch die meisten internen (in der BfG aufgenommen) Daten importiert.
  - Konvertierung in das offene Format mzXML und Prozessierung mit *ntsworkflow*
  - Umformatierung der Ergebnisse in JSON und anschließender Import in das NTSPortal
- Tabellarische Daten von externen Partnern (csv, xlsx)
  - Ergebnisse von NTS Messungen in Tabellen wurden in JSON umformatiert und anschließend importiert
  - Exemplarisch für Daten von BWB (Thermo *Compound Discoverer*) und Landeswasserversorgung (Sciex *PeakView*) durchgeführt.
- JSON Dokumente direkt aus der Prozessierungssoftware
  - Export von annotierten NTS-Daten („Targets“) als JSON (im NTSPortal Format) aus der Prozessierungssoftware *enviMass*. Daten werden geprüft und anschließend importiert (siehe Abschn. 3.3.6.1)

3.3.6.1 Daten von *enviMass*

EnviBee GmbH, das Schweizer Unternehmen hinter der NTS-Software *enviMass*, wurde im Rahmen dieses Projektes beauftragt, eine Export-Funktion in *enviMass* einzubauen, die den Export von Ergebnissen im JSON-Format ermöglicht. Hiermit ist der Export von annotierten Features (in *enviMass* als „Targets“ bezeichnet), zusammen mit Probenahme- und Methodeninformation in der *enviMass* Software per Knopfdruck möglich. Es bedarf lediglich einer nachträglichen Überprüfung der JSON-Datei (R Skript) vor der Eintragung der Daten ins NTSPortal. Für die Anwendung von *enviMass* wurde eine ausführliche Hilfe in einem modalen Dialog-Fenster zur Funktion geschrieben.

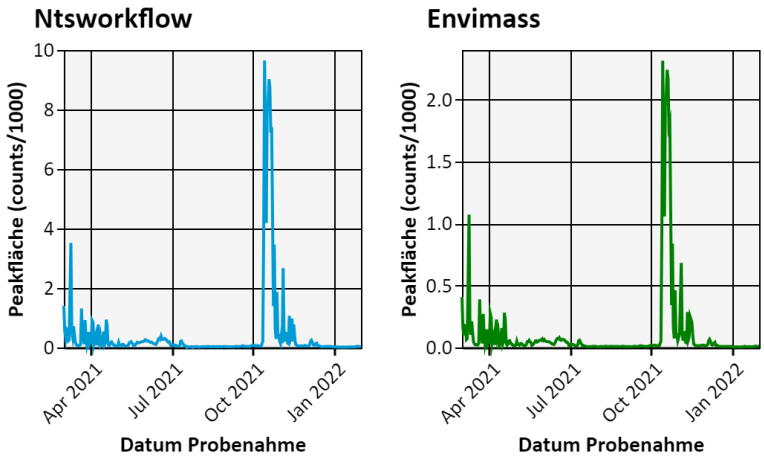
Aufgrund der unterschiedlichen Funktionsweisen von *enviMass* im Vergleich zu den Prozessierungsskripten der BfG (basierend auf *ntsworkflow*), sind die Ergebnisdaten der beiden Workflows nicht identisch. Beispielsweise beinhalten MS^2 Spektren exportiert aus *enviMass* nur diejenigen Fragmentmassen, die in der Target-Liste explizit eingetragen wurden. Andererseits kann *enviMass* mehrere Addukte für eine Substanz exportieren, was bei *ntsworkflow* derzeit nicht möglich ist.

[2] Siehe Feldbennenung im NTSPortal Wiki (3.3.4)

Der Import von *enviMass* Ergebnissen wurde testweise durchgeführt und die Vergleichbarkeit der Daten überprüft. In diesem Versuch wurden Messungen von Proben des Rheins bei Koblenz (Tagesmischproben) sowohl mit *enviMass* als auch mit *ntsworkflow* ausgewertet. Beide Datensätze wurden in das NTSPortal importiert und mit Hilfe der Spektrenbibliothek gescreent (*enviMass*: „Target Screening“, *ntsworkflow*: „Annotation“). Die Ergebnisse für den Zeitraum 2021-03-01 bis 2022-01-31 zeigen eine Überschneidung von 26% der Befunde. Die meisten nicht übereinstimmenden Ergebnisse waren zusätzliche Befunde in *enviMass* (631 Befunde mit *enviMass*, 259 mit *ntsworkflow*). Dies liegt daran, dass *enviMass* nur m/z und t_R als Kriterien für die Suche heranzieht, während *ntsworkflow* auch einen MS^2 Abgleich durchführt. Die Nachweisgrenzen sind für *enviMass* deshalb geringer (die Aufnahme von MS^2 Spektren benötigt eine höhere Intensität), allerdings mit dem Nachteil, dass die Zuordnung weniger sicher ist. So enthalten die Ergebnisse möglicherweise einen höheren Falsch-Positiv-Anteil. Weitere Unterschiede finden sich bei den berechneten Peakflächen aufgrund unterschiedlicher Integrationsalgorithmen (Abbildung 7). Weiterführende Studien über längere Zeiträume werden benötigt, um die Vergleichbarkeit weiter zu prüfen und gegebenenfalls durch Anpassungen in der Prozessierung und Harmonisierung zu optimieren.

Abbildung 7: Vergleich von Zeitreihen ausgewertet mit enviMass und mit ntsworkflow

Tetrabutylammonium im Rhein bei Koblenz



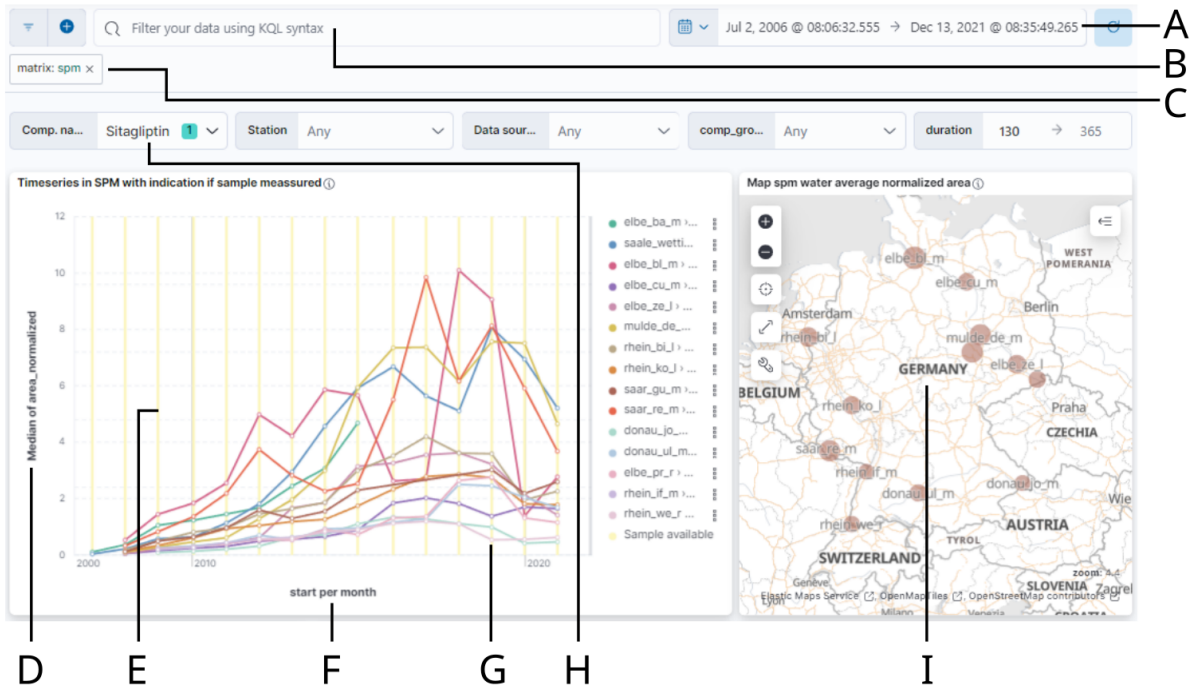
Quelle: eigene Abbildung BfG

3.3.7 Web-basierte Anwendung für das NTSPortal (Phase 1)

Über eine Web-Anwendung, ein sogenanntes *Front-End*, können Anwender mit Hilfe einer grafischen Benutzeroberfläche die Daten im NTSPortal durchsuchen, visualisieren lassen und exportieren. Zusätzlich wurden Funktionen für die Sortierung (Priorisierung) von Befunden, basierend auf ihrer räumlichen Verteilung oder Änderung in der Intensität, über die Zeit entwickelt.

Abbildung 8: Screenshot des prototypischen Dashboards (erstellt mit Kibana) zur Visualisierung von Peakflächen-Zeitreihen

Suche nach einer Substanz in den Schwebstoffzeitreihen (Sitagliptin). A: Eingabe Zeitraum, B: Suchbox für komplexere Suchen (Kibana Query Language), C: Gesetzte Filter (Matrix: SPM) D: Y-Achse (Peakfläche, relativ zum IS), E: Gelbe Linien: Probe von diesem Zeitpunkt ist vorhanden, F: Zeitachse (Zeitpunkt der Probenahme, 2000 bis 2025), G: Zeitreihen von Sitagliptin für jede Messstelle, H: Drop-Down-Menüs (Substanzname, Messstelle, Datenquelle, Substanzgruppe, Mischprobendauer), I: Karte



Quelle: eigene Abbildung BfG

Diese sogenannten *Dashboard*-Anwendungen wurden mit Hilfe der Software *Kibana* entwickelt (Abbildung 8 zeigt das Beispiel einer Zeitreihe). Die Entwicklungen von neuen *Dashboards* mit eingebauten Recherchefunktionen wurden im Laufe des Projektes auf der *Kibana* Plattform stetig fortgesetzt und Änderungen wurden im laufenden Betrieb vorgenommen. Das Portal beinhaltet derzeit 24 Seiten mit verschiedenen Such- und Darstellungsfunktionen und kann auf https://ntsportal.bafg.de abgerufen werden (nach vorheriger IP-Zulassung, siehe 3.6.1). Die Anwendung der Seite wird in einem Tutorial-Video exemplarisch erklärt. Das Video wird bei Anmeldung zur Verfügung gestellt und ist auch auf der FTP-Seite der BfG verfügbar[3]. Weitere Dokumentation befindet sich auf der Seite selber und im NTSPortal-Wiki[4]. In der Sonderpublikation zum Langenauer Wasserforum 2021 wurde ein Artikel zum NTSPortal veröffentlicht und zwei räumliche Analysefunktionen näher beschrieben (siehe Anhang D)

3.3.7.1 Interaktive Priorisierungstools auf der NTSPortal Dashboard-Oberfläche

Veränderung des Signals über die Zeit

Die Schwebstoffprobendaten können nach Veränderung der IS-relativen Peakfläche über die Zeit sortiert werden und Substanzen, die zum Beispiel ein steigendes Signal (analytisches Signal, z.B. IS-normalisierte Peakfläche) aufweisen, priorisiert werden.

Hierzu wird eine lineare Regression für jede Substanz an jeder Station extern in R berechnet. Die resultierenden Steigungswerte werden als neue Datenbank in das NTSPortal gespeichert. Eine

[3] ftp.bafg.de/pub/REFERATE/g2/ntsportal/NTSPortal_tutorial_basic_v4.mp4
[4] https://github.com/bafg-bund/ntsportal/wiki/Using-ntsportal.bafg.de

Tabelle mit Substanz, Messstelle und Steigung wird auf dem „Time Series SPM“ Dashboard angezeigt und kann für die Sortierung angewendet werden. Die Berechnung findet somit nicht “interaktiv” statt, sondern im Vorfeld. Bei der Eintragung von neuen Daten muss diese Berechnung neu durchgeführt werden.

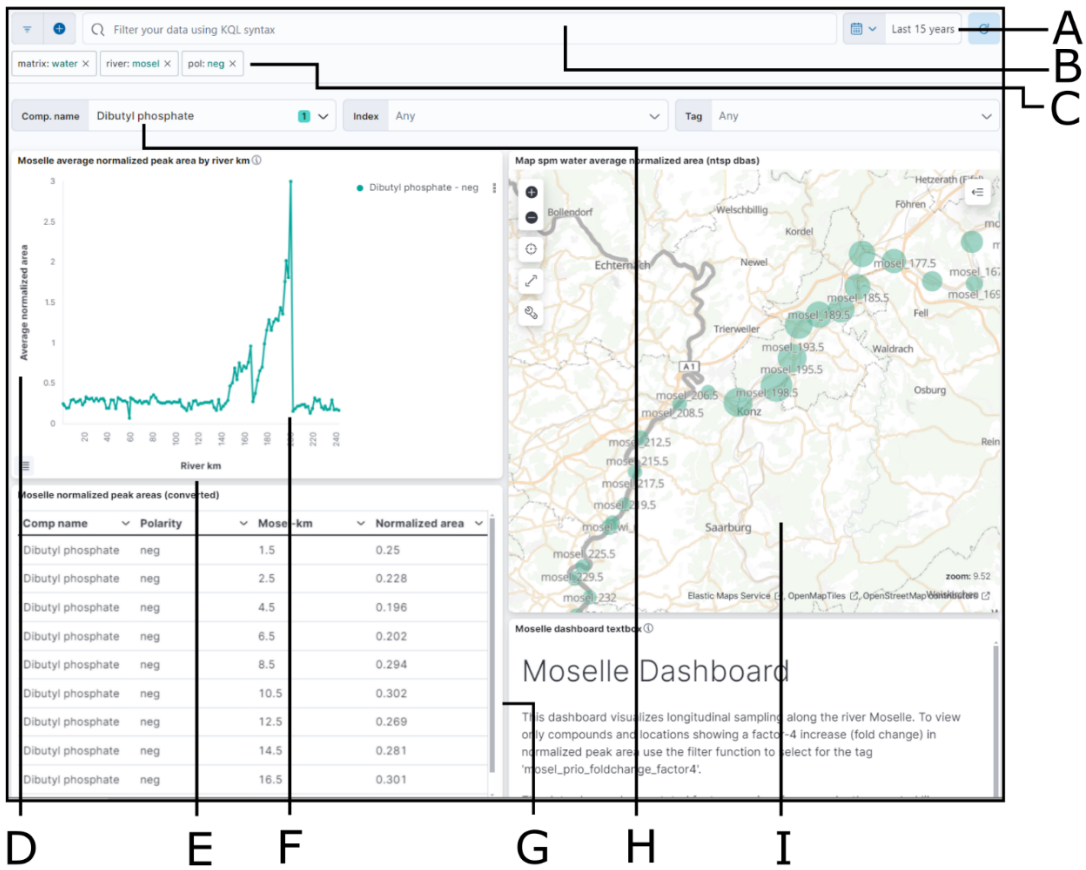
Zwei Beispiele dieser Priorisierung, die aus dem Datensatz erhoben wurden, sind das Desinfektionsmittel Chlorhexidin in Weil am Rhein (Verdoppelung der instrumentellen Response von 2016 bis 2021) und der Lebensmittelfarbstoff Brilliant Blue FCF an der Mulde bei Dessau (fast Verzehnfachung der Response von 2013 bis 2020).

Fold-Change zur Analyse von räumlich und zeitlich zusammenhängenden Proben

Beim Vergleich von räumlichen Daten (z.B. eine Fließzeit-korrespondierende Probenahme entlang eines Flussverlaufs), bedeutet der sogenannte Fold-Change die relative Veränderung der Response zwischen zwei benachbarten Messstellen. Als Priorisierungsmethode können Substanzen hervorgehoben werden, die einen Fold-Change größer eines bestimmten Wertes aufweisen. Das Vorgehen kann an der Längsbeprobung der Mosel (Abbildung 9) getestet werden. Im Dashboard kann über das Suchfeld das Label „mosel_foldchange_factor4“ ausgewählt werden. Als Ergebnis zeigt eine Tabelle Substanzen und den jeweiligen Fluss-km an, an dem ein Fold-Change zum vorherigen Messpunkt von >4 (>400%) vorliegt. Die Tabelle kann nach Response oder Fluss-km sortiert werden und ausgewählte Substanzen können näher untersucht werden. Somit kann der Ort der Response-Erhöhung über die Kartendarstellung visualisiert werden. In dem genannten Beispiel zeigte die Substanz Dibutylphosphat einen Fold-Change >4 von Fluss-km 202 bis 200.5 (die Mosel-Kilometrierung beginnt in Koblenz). Bei km 200 mündet die Saar in die Mosel, daher liegt die Quelle der Emission vermutlich im Saar-Einzugsgebiet.

Abbildung 9: Screenshot eines Dashboards für Probenahmen entlang eines Flussverlaufs

Darstellung einer Substanz in der Mosel (Dibutylphosphat). A: Eingabe Zeitraum, B: Suchbox für komplexere Suchen (Kibana Query Language), C: Gesetzte Filter (Matrix: Wasser, Fluss: Mosel, Polarität: Negativ), D: Y-Achse (Peakfläche, relativ zum IS), E: X-Achse Fluss-km (Fließrichtung Rechts nach Links), F: Anstieg bei km 200,5 (Mündung der Saar), G: Tabelle mit Werten H: Drop-Down-Menüs (Substanzname, Index-Name und Label), I: Kartendarstellung, Punktgröße zeigt Peakfläche, relativ zum IS.



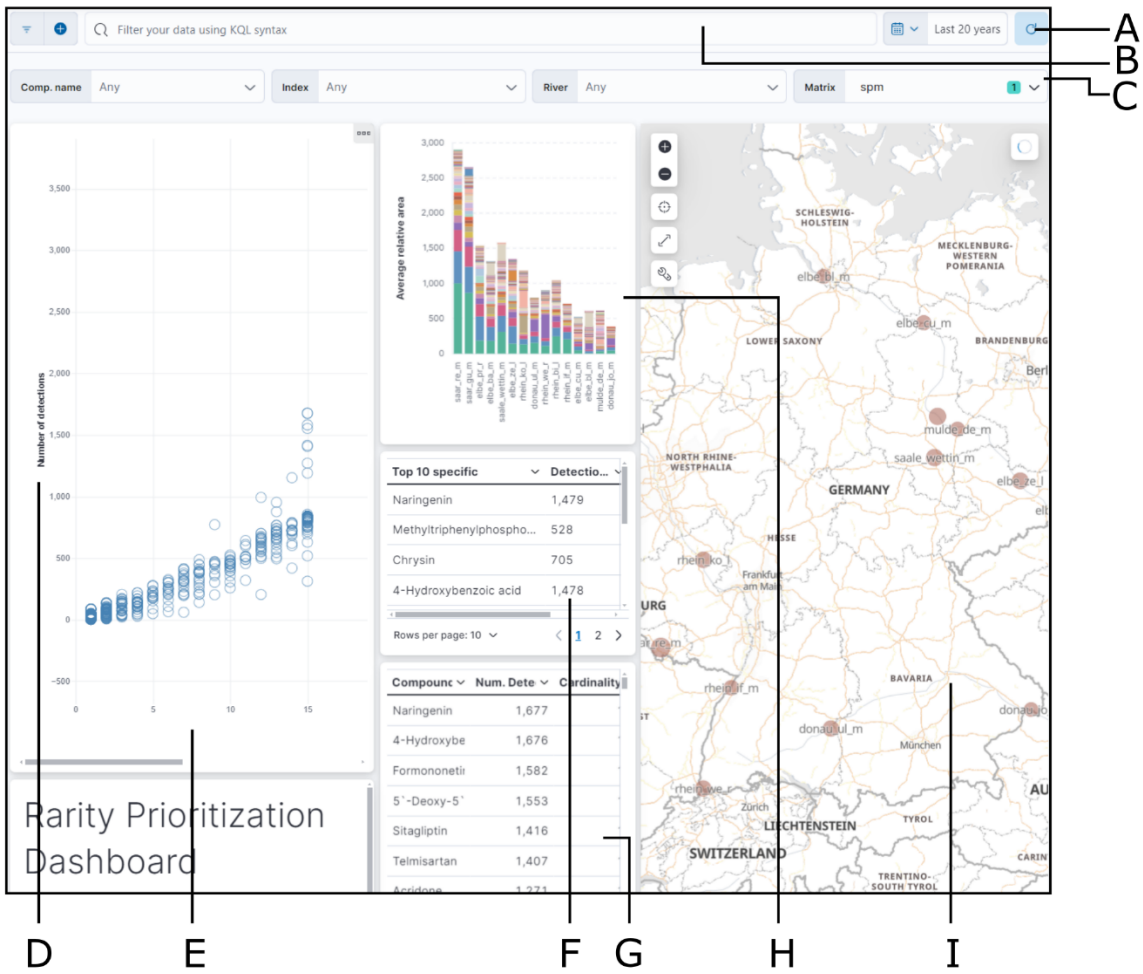
Quelle: eigene Abbildung BfG

Rarity Dashboard

Dieses Dashboard zeigt auf der linken Seite ein Streudiagramm von jeder detektierten Substanz und auf der rechten Seite eine Karte mit den Messstellen. Auf der horizontalen Achse ist die Anzahl der Messstellen, an denen die jeweilige Substanz detektiert wurde, auf der vertikalen Achse die Anzahl der Detektionen der Substanz dargestellt. Alle Substanzen der Datenbank sind als blaue Punkte aufgetragen. Auf der rechten Seite wird eine Karte angezeigt, in der die räumliche Verteilung aller Befunde dargestellt wird. Das Streudiagramm kann als einfaches Priorisierungstool fungieren. Punkte, die oben rechts im Diagramm plotten sind Substanzen, die überall und sehr häufig detektiert werden, während sich oben links Substanzen befinden, die häufig detektiert werden, aber an einer begrenzten Anzahl von Messstellen. Der Vergleich von ubiquitär auftretenden Verbindungen mit Substanzen, die eine räumlich eingeschränkte Verteilung aufweisen, sind für Untersuchungen von Eintragswegen von besonderem Interesse. Nach der Auswahl einer Substanz zeigt die Kartendarstellung die Verteilungsmuster dieser Substanz und gibt somit Hinweise auf mögliche Eintragswege (Abbildung 10).

Abbildung 10: Screenshot des “Rarity Dashboards”

Darstellung aller Substanzen in Messproben von Schwebstoffen (Umweltprobenbank). A: Eingabe Zeitraum, B: Suchbox für komplexere Suchen (Kibana Query Language), C: Gesetzte Filter (Matrix: Schwebstoff), D: Scatterplot Substanzen: Y-Achse (Anzahl Detektionen), E: Scatterplot: X-Achse (Anzahl Messstellen), F: Tabelle für die Auflistung häufig detektierter Verbindungen, G: Tabelle mit allen Verbindungen H: Gestapelte Säulendiagramm: gemittelte relative Peakfläche, relativ zum IS, nach Messstelle. I: Kartendarstellung, Punktgröße zeigt Peakfläche, relativ zum IS.



Quelle: eigene Abbildung BfG

3.3.8 Anwendungsbeispiele des NTSPortal

Diese Beispiele sollen zeigen, wie das NTSPortal einen Überblick über die Verteilung einer Substanz geben kann. Dies kann beispielsweise der Regulatorik in der Expositionsabschätzung von Einzelstoffen oder Mischungen sowie der Vorbereitung einer Analyse- oder Messkampagne dienen.

3.3.8.1 Anwendungsbeispiel: Informationen zu einer Substanz: 6-PPD-Chinon

6-PPD-Chinon wird durch Reifenabrieb in die Gewässer eingetragen. Es wirkt toxisch auf ausgewählte Salmonidenarten mit 24 h LC50-Werten von 0.95 µg/L bei juvenilen Coho-Lachsen (Tian et al. 2020), 0.50 µg/L bei juvenilen Bachsaiblingen und 1 µg/L bei juvenilen Regenbogenforellen (Brinkmann et al. 2022). Aufgrund der großen Anzahl von Publikationen, wurde ein Referenzspektrum von 6-PPD-Chinon für die Spektrenbibliothek im April 2021 aufgenommen. Das hat einen retrospektiven Nachweis ermöglicht. Mit Hilfe des Dashboards

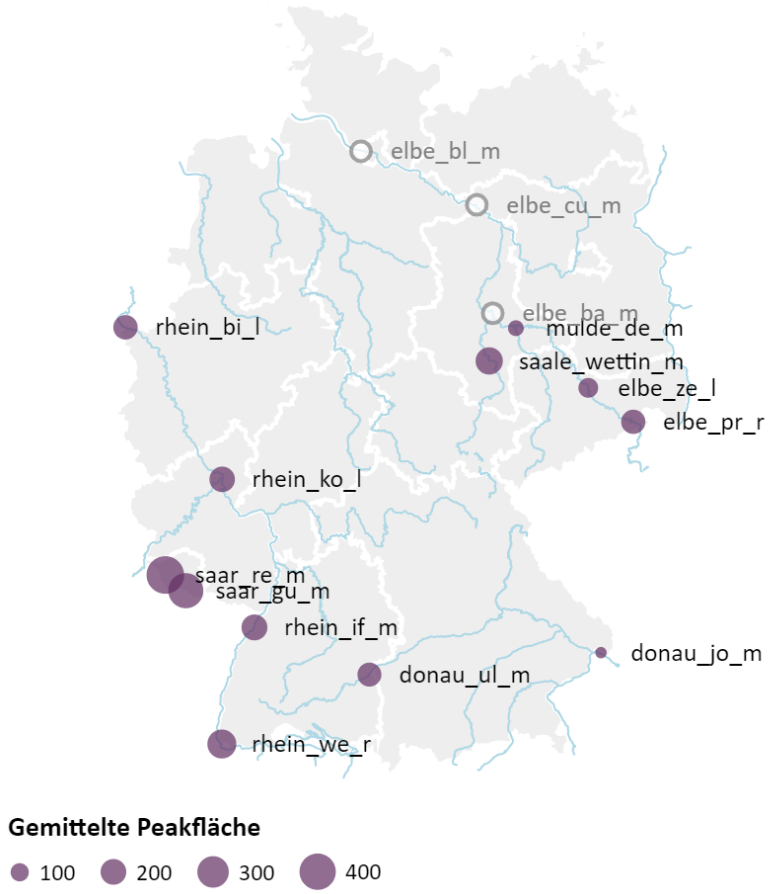
“Search for a substance” werden tabellarisch und kartographisch Informationen zu den Befunden von 6-PPD-Chinon dargestellt. Bei Bedarf kann im Dashboard "Spectra and Chromatograms" die Vertrauenswürdigkeit der Zuordnung geprüft werden, indem die Fragmentspektren der Proben mit der Spektrenbibliothek nebeneinander dargestellt werden.

Die Informationen aus der Datenbank zeigen eine weitreichende Verteilung der Substanz in den Schwebstoffproben im gesamten Beobachtungszeitraum und sporadische Befunde in den Wasserproben. In den Wasserproben ist das Signal nah an der Bestimmungsgrenze. Fast alle Messstationen für Schwebstoffe (12 von 15) weisen Befunde für 6-PPD-Chinon auf, aber die höchsten Peakflächen wurden an den Messstellen Rehlingen und Güdingen an der Saar gefunden. Die Response an allen Messstellen hat sich in Laufe der letzten 20 Jahren in den Schwebstoffproben nicht wesentlich verändert (Abbildung 11 und Abbildung 12).

Die Informationen aus dem NTSPortal geben eine Orientierung für detaillierte Untersuchungen (Bestimmung der Konzentration, höhere Probenahmefrequenz für ausgewählte Messstellen an der Saar). In einem verwandten Projekt wurde eine retrospektive Quantifizierung der Substanz in den Schwebstoffproben der Saar durchgeführt, hierdurch wurden Konzentrationen zwischen 10 und 100 ng/g gemessen (Dierkes et al. 2024).

Abbildung 11: Regionale Verteilung von 6PPD-Chinon in Schwebstoffproben der Umweltprobenbank

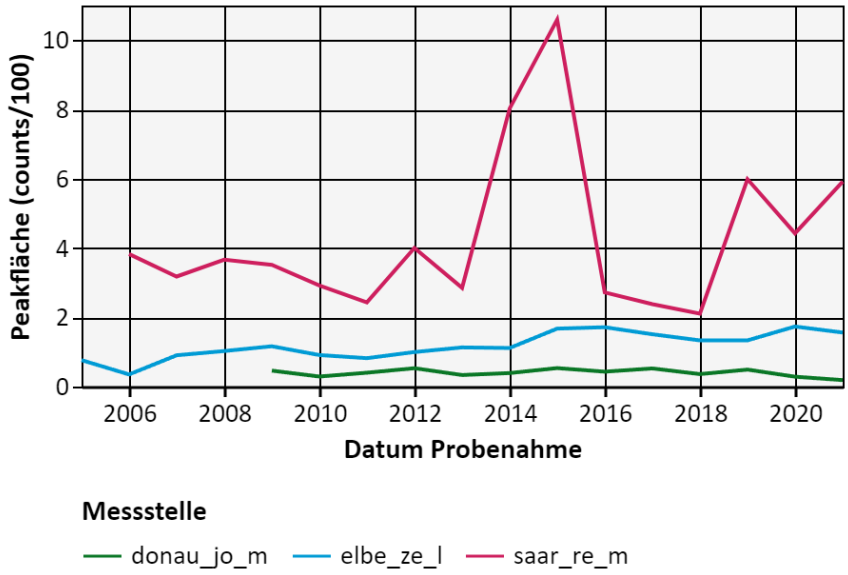
Gemittelte Peakfläche an Messstellen von Schwebstoffen der Umweltprobenbank. An grauen Punkten wurde 6PPD-Chinon nicht detektiert (unter Nachweisgrenze). Messstellencodierung: siehe Anhang E.



Quelle: eigene Abbildung BfG

Abbildung 12: Zeitverlauf von 6PPD-Chinon in Schwebstoffproben der Umweltprobenbank

Zeitreihen der Peakfläche an drei ausgewählten Messstellen: donau_jo_m: Jochenstein, Messungen ab 2008; elbe_ze_l: Zehren, Messungen ab 2005; saar_re_m: Rehlingen, Messungen ab 2006.



Quelle: eigene Abbildung BfG

3.3.8.2 Anwendungsbeispiel: Informationen zu einer Stoffgruppe

Das NTSPortal kann auch für die Suche von Substanzgruppen angewendet werden. Ein Beispiel hierfür ist die Gruppe der Quartären Ammoniumverbindungen (QAV), die unter anderem als Biozide in Desinfektionsmitteln oder Katalysatoren Einsatz finden. Um diese Gruppe mit dem NTSPortal zu untersuchen, kann in dem Dashboard “Search for a Substance” im KQL Suchfeld der Suchbegriff “name: *ammonium” eingegeben werden. Mit diesem sog. Wildcard Suchbegriff werden alle detektierten QAV Verbindungen dargestellt (41 Verbindungen von 47 in der Spektrenbibliothek, siehe Tabelle 2). Diese Verbindungen werden in allen bisher untersuchten Regionen detektiert.

Tabelle 2: Liste der detektierten quartären Ammonium Verbindungen

Substanzname	Anzahl Befunde Wasser	Anzahl Befunde SPM	Anzahl Messstellen
Tetrabutylammonium	3,509	836	235
Tetrapropylammonium	2,619	393	30
Benzyl-triethylammonium	2,105	509	192
Tributylmethylammonium	1,279	193	6
Dibenzyldimethyl ammonium	823	771	16
Hexadecyl-trimethylammonium	800	788	194
Ethyl-hexadecyl-dimethyl ammonium	538	648	66
Octyl-decyl-dimethylammonium	382	738	169
Octadecyl-trimethylammonium	330	799	141
Dodecyltrimethylammonium	330	723	15

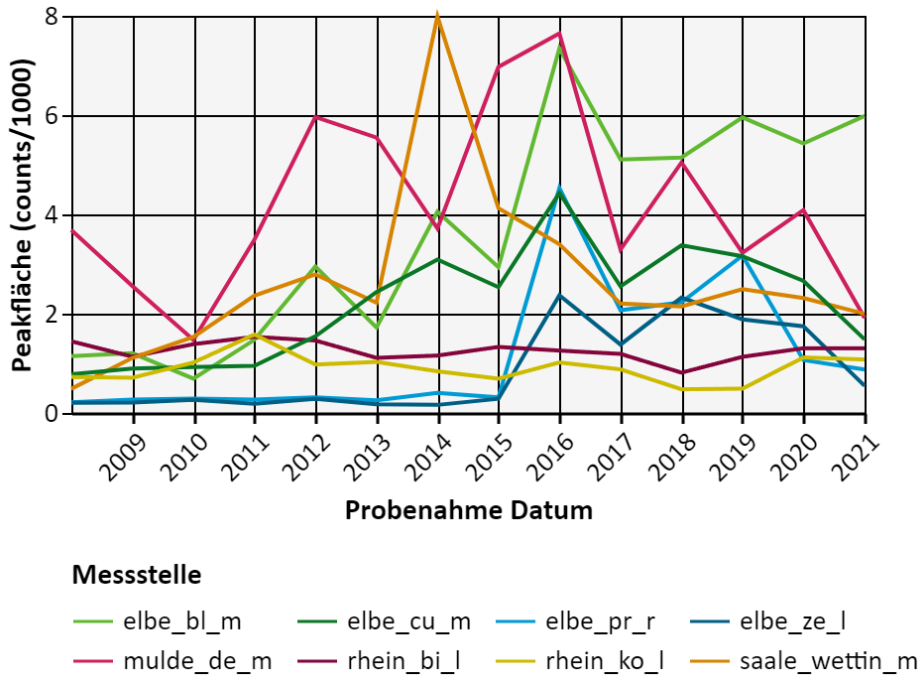
Substanzname	Anzahl Befunde Wasser	Anzahl Befunde SPM	Anzahl Messstellen
Tetraethylammonium	320	3	14
Benzyl-dimethyl-hexadecylammonium	250	821	16
Benzyl-dimethyl-dodecylammonium	237	839	31
Benzyl-tributylammonium	186	698	15
Dimethyldioctylammonium	144	839	15
Eicosyltrimethylammonium	141	709	13
Trimethylphenylammonium	128	0	1
Benzyl-dimethyl-tetradecylammonium	123	839	53
Didecyl-dimethylammonium	121	814	59
N-Benzyldimethylstearyl ammonium	114	808	35

Durch die Anwendung des “Rarity Prioritization” Dashboards (Abbildung 10) kann die Häufigkeit und Verbreitung einzelner QAVs verglichen werden. Es wurde festgestellt, dass Tetrabutylammonium fast überall vorkommt (>200 Messstellen) und sehr häufig detektiert wurde (>4000 Befunde), während Tetrapropylammonium und Tributylmethylammonium eher vereinzelt vorkommen (<30 Messstellen), aber dennoch zu den am häufigsten nachgewiesenen Verbindungen gehören (>1400 Befunde).

Tetrabutylammonium (TBA) wurde in allen Flusssystemen sowohl in Wasser- als auch in Schwebstoffproben gefunden. Für die Schwebstoffproben wurden die höchsten Peakflächen im Elbe-Einzugsgebiet gefunden, während im Wasser die höchsten Werte in der Region um das Hessische Ried und den Urselbach zu verzeichnen waren (für diese Regionen waren allerdings keine Schwebstoffproben zum Vergleich verfügbar). Eine zukünftige Quantifizierung des TBA in Wasser-, Schwebstoff- und Sedimentproben in den Gebieten Hessisches Ried und Urselbach könnte die Belastung dort im Vergleich zu anderen Gebieten einordnen. Betrachtet man alle Probenahmestellen für die Schwebstoffe, ist seit 2010 ein allgemeiner Anstieg in der Konzentration zu verzeichnen (Abbildung 13). Die Konzentrationsentwicklung in aktuellen Proben sollte daher weiterhin beobachtet werden.

Abbildung 13: Zeitreihe von Tetrabutylammonium an Schwebstoffmessstellen der Umweltprobenbank 2009 bis 2021

Messstellencodierung: Siehe Anhang E

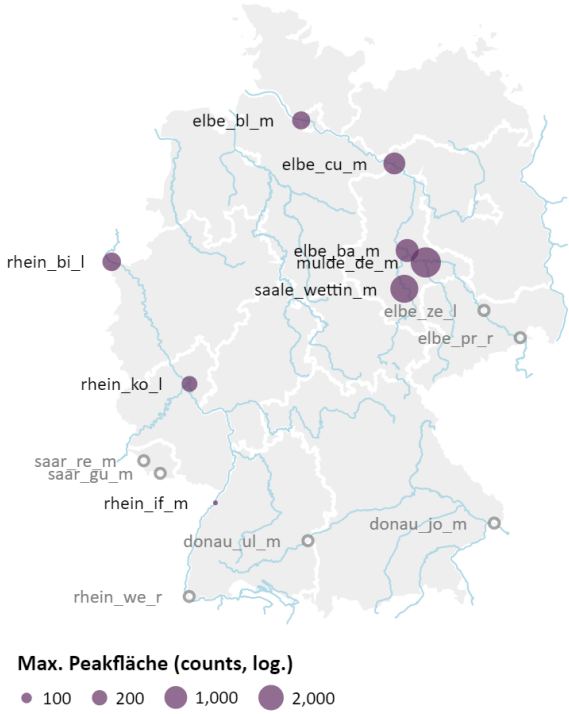


Quelle: eigene Abbildung BfG

Tetrapropylammonium wurde sowohl im Rhein als auch in der Elbe detektiert, aber das Signal in den Schwebstoffproben war im Elbe-Einzugsgebiet höher. In der Mulde (Nebenfluss der Elbe) war das Signal 15-mal so hoch wie im Rhein, aber die Werte haben sich bis zur Elbmündung die im Rhein wieder angeglichen (wahrscheinlich aufgrund der Verdünnung) (Abbildung 14). Die Quelle der Verbindung scheint in der Region um Leipzig und im Einzugsgebiet der Münzbach zu liegen (Abbildung 15), diese Information könnte für die Planung weiterer Untersuchungen hilfreich sein. Die Schwebstoffdaten der Mulde zeigen eine starke Fluktuation des Signals seit 2018, und bei den monatlichen Probenahmen (nur das Jahr 2021 vorhanden) wurde ebenfalls eine starke Fluktuation des Signals innerhalb eines Jahres beobachtet. Dies ist ein Hinweis auf die Art der Einleitung (industriell anstatt kommunal). Für den Rhein konnte nur beobachtet werden, dass die Quelle flussaufwärts von Koblenz liegt, so dass der Datensatz um weitere Messstationen ergänzt werden muss, um die Quelle zu bestimmen.

Abbildung 14: Regionale Verteilung von Tetrapropylammonium in Schwebstoff- und Wasserproben (Deutschlandweit)

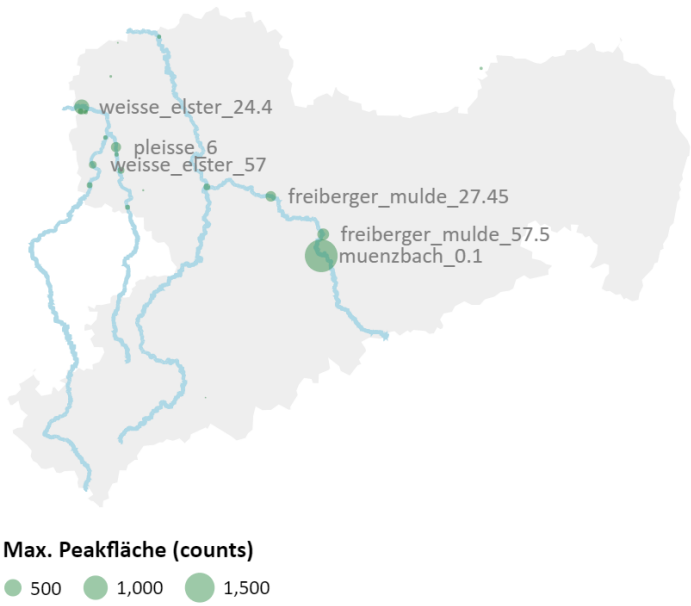
Fuchsia Punkte zeigen Messstellen von Schwebstoffen der Umweltprobenbank. An grauen Punkten wurde Tetrapropylammonium nicht detektiert (unter Nachweisgrenze).



Quelle: eigene Abbildung BfG

Abbildung 15: Regionale Verteilung von Tetrapropylammonium in Schwebstoff- und Wasserproben (Sachsen)

Ausschnitt Land Sachsen. Grüne Punkte: Oberflächenwasserproben aus einem Kooperationsprojekt mit dem Land Sachsen (Köppe et al. 2024). Messstellencodierung: Siehe Anhang E



Quelle: eigene Abbildung BfG

Tributylmethylammonium wird in Koblenz (Rhein), Iffezheim (Rhein), Tangermünde (Elbe) und Dessau (Mulde) gefunden, ein Hinweis auf möglichen Quellen flussaufwärts von Iffezheim und Dessau. In der Vergangenheit war das Signal in Blankenese ca. 10x so hoch wie in Dessau, wahrscheinlich aufgrund weiterer Quellen in der Elbe. Seit 2016 ist das Signal in Blankenese jedoch rückläufig. Eine genauere Untersuchung des Einzugsgebiets der Mulde könnte weitere Informationen zu der noch aktiven Quelle liefern. Am Rhein ist das Signal in Bimmen durchweg höher als in Koblenz und Iffezheim, was bedeutet, dass es in diesem Einzugsgebiet möglicherweise mehrere Emissionsquellen gibt. Weitere Probenahmen im Flussverlauf zwischen Koblenz und Bimmen werden benötigt, um die Quellen einzugrenzen.

3.3.9 Entwicklung eines weiteren Front-Ends für ergänzende Funktionalitäten

Die Anwendung von *Kibana* zur Bildung des Front-Ends war für grundlegende Funktionen wie die schnelle Suche und Darstellung von Zeitreihen und Karten sehr geeignet. Die Plattform ist für große Datensätze konzipiert und integriert wichtige Mechanismen, um die Benutzererfahrung zu optimieren (schnelle Reaktionszeiten, minimale Last für den Client-Rechner). Es bietet grafische Werkzeuge für die schnelle Entwicklung von Visualisierungen, ohne programmieren zu müssen. Zudem können viele Datenbank-Management Aufgaben, wie z.B. das Nutzerkonto-Management, grafisch mit Kibana durchgeführt werden, ohne Kenntnisse der ElasticSearch API haben zu müssen.

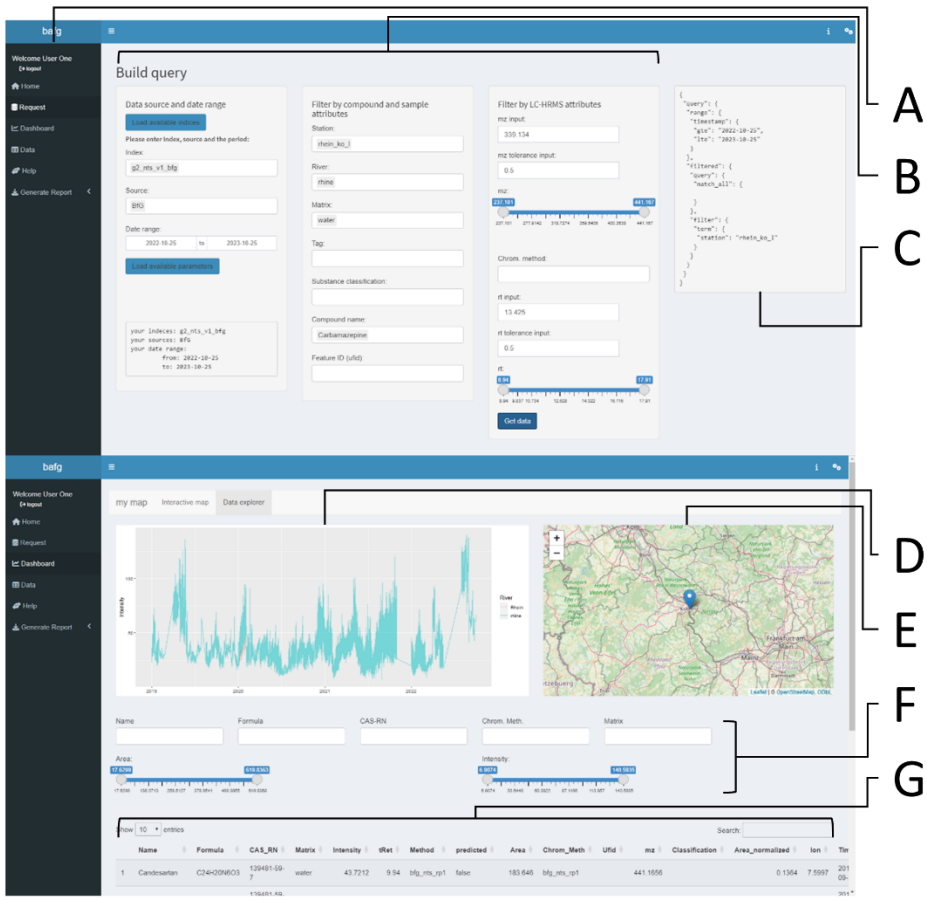
Kibana hat allerdings, trotz seiner Vorteile, eine begrenzte Flexibilität für den Aufbau spezialisierter Dashboards und Webanwendungen. Daher wurde, im Hinblick auf die geplante Fortsetzung des NTSPortals über die Projektlaufzeit hinaus, die Entwicklung eines zweiten Front-Ends zur Ergänzung der bestehenden Dashboards im letzten Projektjahr gestartet.

Für diese Entwicklung wurde das *Shiny* Framework von Posit (Kasprzak et al. 2021) verwendet. Diese JavaScript-Bibliothek bietet eine Einbindung von R, sodass es für statistische Berechnungen benutzt werden kann. Des Weiteren gibt es bereits Erfahrungen in der Anwendung und der passenden Infrastruktur für das Hosting von *Shiny* Webanwendungen an der BfG.

Die Entwicklung wurde als Auftrag an die SVA GmbH (Rahmenvertragspartner) vergeben. Die Webanwendung wurde modular aufgebaut, damit neue Funktionen schrittweise eingebaut werden können. Die ersten Funktionen, die entwickelt wurden, sind die Vereinfachung von Suchanfragen in den komplexen Datenstrukturen von NTS Daten und der Export von vollständigen Datensätzen. Mit *Kibana* konnten bisher zusammengefasste Datensätze als CSV-Dateien exportiert werden, Rohdaten jedoch sind bisher als JSON nur über die Konsole erhältlich (durch Anwendung der ElasticSearch-Querysprache). Gute Exportmöglichkeiten sind allerdings eine wichtige Voraussetzung für weiterführende Auswertungen der Daten aus dem NTSPortal und die Verknüpfung mit weiteren Daten. Die Nutzung der Konsole ist für die meisten Anwender*innen jedoch nicht praktikabel. Neben der Exportfunktion werden Zeitreihen und Karten zur Visualisierung der zum Export ausgewählten Daten und eine Hilfe-Seite aufgebaut (Abbildung 16). Zum Projektende wurde eine erste Version des Front-Ends erstellt, allerdings sind weitere Entwicklungsarbeiten notwendig, bevor die Web-App für externe Nutzer*innen publiziert werden kann. Die weitere Entwicklung wird während des fortschreitenden Ausbaus der Datenbank erfolgen.

Abbildung 16: Screenshots des zweiten, mit Shiny gebauten Front-Ends im derzeitigen Entwicklungsstand (Entwurf)

Die Screenshots zeigen eine vereinfachte Methode für die Filterung von Daten (Build Query) und die Zeitreihen- und Kartendarstellung. A: Navigationsleiste. B: Eingabefelder zum Aufbau eines Queries, C: API Code des erstellen Queries (zur Speicherung), D: Darstellung Zeitreihen. E: Darstellung Karte. F: Interaktive Filter. G: Ergebnistabelle.



Quelle: eigene Abbildung BfG

3.4 Phase 2: Aufbau einer Datenbank für bekannte und unbekannte Features

Die in Phase 1 entwickelte Datenbank enthält nur annotierte Features. Das sind Features, die durch den Abgleich mit der Spektrenbibliothek mit einem Substanznamen annotiert wurden (Bibliothek-Screening). In Phase 2 wurde eine Datenbank für alle Features des Datensatzes entwickelt (auch die, die nicht annotiert wurden) (Tabelle 1). Mit Hilfe dieser Daten können zwei andere Ausrichtungen des NTS (Suspect Screening und Suche nach Unbekannten, siehe Abkürzungsverzeichnis) durchgeführt werden.

Sowohl die Struktur der Datenbank in Phase 2 (DB-Schema), als auch das Design und die Funktionalität der Dashboards wurden von Phase 1 übernommen. Der alleinige Unterschied liegt in der Verlinkung oder Gruppierung von Features einer Substanz (Alignment). Bei Phase 1 wird das Alignment über den Substanznamen gemacht. In Phase 2 wurden nur ein Bruchteil der Features annotiert (ca. 5%). Alle Features einer, in ca. 95% der Fälle, unbekannten Substanz werden deshalb über eine ID verlinkt (sogenannte Universal Feature ID, oder *Ufid*). Mit Hilfe eines neu entwickelten Algorithmus werden Features derselben Substanz in einer *Ufid* gruppiert (siehe 3.4.1). Der entwickelte Algorithmus wurde mit Hilfe eines Testdatensatzes von ca. 2 Millionen Features validiert. Dabei wurden Substanznamen (Annotation durch Bibliothek-Screening wie in Phase 1) zur Kontrolle der Ufid-Zuordnung herangezogen, um die Rate der falschen Zuordnungen einschätzen zu können, siehe hierfür Abschn. 3.4.2.

3.4.1 Beschreibung des Algorithmuses zur Ufid-Zuordnung

Der Algorithmus nutzt ein Clustering-Verfahren, um Features anhand ihrer Ähnlichkeit in den Variablen 1) m/z, 2) Retentionszeit und 3) MS^2 zu gruppieren. Die drei Dimensionen werden durch ein additives Scoring-System kombiniert, dabei werden Grenzwerte für jede Dimension individuell festgelegt.

Die NTS Daten werden mit einer Data-Dependent MS^2-Methode aufgenommen. Nur ca. ⅔ der Features haben deshalb ein assoziiertes MS^2-Spektrum. Features ohne MS^2 werden durch einen nachträglichen “Gap-Filling” Schritt zu bestehenden Gruppen zugeordnet.

Der oben beschriebene Algorithmus wurde “Level 1 Ufid-Zuordnung” genannt. Um die Vor- und Nachteile der Anwendung von MS^2 zu testen, wurde ein zweiter Algorithmus entwickelt. Dieser nutzt dasselbe Clustering- und Scoring-System, allerdings werden zum Vergleich nur die Dimensionen 1) m/z und 2) Retentionszeit herangezogen. Der zweite Algorithmus wurde “Level 2 Ufid-Zuordnung” genannt. Für die Level 2 Ufid-Zuordnung ist kein Gap-Filling-Schritt notwendig, denn MS^2-Spektren spielen hierbei keine Rolle.

3.4.1.1 Level 1 Ufid-Zuordnung

Im ersten Schritt wird eine Distanzmatrix aufgebaut, indem die Ähnlichkeiten zwischen allen Features paarweise berechnet werden. Die Ähnlichkeit von zwei Features wird über einen Score bestimmt, der aus der Summe der Scores der drei genannten Kriterien (m/z, Retentionszeit und MS^2) berechnet wird. Ein Score von 0 bedeutet, dass die verglichenen Features in allen drei Kriterien ähnlich sind. Ein Score größer Null bedeutet, dass ein oder mehrere Kriterien nicht übereingestimmt haben. Die Berechnung ist in der leistungsstarken C++ Sprache programmiert und läuft parallelisiert auf einem Hochleistungsrechner (32 Kerne) mit Hilfe der „parallelDist“ R-Bibliothek (Eckert 2018). Die resultierende Matrix wäre zu groß, um alle Features der gesamten Datenbank in einem Schritt miteinander zu vergleichen, deshalb wird die Arbeit in Pakete von je 10.000 Features aufgeteilt. Dabei entsteht für jedes Paket eine Distanzmatrix von 50 Millionen Ähnlichkeitswerten (Scores).

Im zweiten Schritt wird mit Hilfe der Distanzmatrix ein hierarchisches Clustering durchgeführt (Complete-Linkage-Verfahren) (Han et al. 2011). Das resultierende Dendrogramm wird auf der ersten Ebene geschnitten und die so gebildeten Gruppen bekommen jeweils eine Ufid. Das Schneiden des Dendrogramms auf der ersten Stufe bedeutet, dass alle Features in einer Gruppe einen Vergleichs-Score von 0 zueinander haben (innerhalb der Toleranzen in allen Kriterien). Es werden mindestens 5 Features benötigt, um eine Gruppe zu bilden.

Im nächsten Schritt wird das sogenannte Gap-Filling durchgeführt. Dabei werden Features, die kein MS2-Spektrum haben, zu den bestehenden Ufid-Gruppen zugewiesen. Dies geschieht über

den Vergleich von m/z, Retentionszeit und chromatographischer Peakform. Verglichen wird das neue Feature mit jeden bereits existierenden Gruppe, über den Mittelwerten von m/z und t_R. Die Peakform wird mit dem intensivsten Feature der jeweiligen Gruppe verglichen.

Im letzten Schritt werden die Ufids zu den Features in der Datenbank zugeschrieben (im Feld *ufid*).

Beim Eintragen neuer Ergebnisse (Features) in die Datenbank wäre im Prinzip ein erneutes Clustering der gesamten Datenbank nötig. Dabei werden die Ufids neu zugewiesen und somit die Verfolgung von Unbekannten erschwert. Deshalb werden neue Features erst zu bestehenden Ufids zugeordnet und nur Features, die zu keiner Gruppe passen, werden wie oben beschrieben geclustert. Um diese Zuordnung von bereits bekannten Ufids durchzuführen, wird eine zusätzliche Datenbank (Ufid-Bibliothek) für existierende Ufid-Gruppen angewendet. Diese Bibliothek wurde mit Hilfe von SQLite aufgebaut. Die Zuordnung von Ufids (mit der Ufid-Bib.) erfolgt analog zur Annotation mit Substanznamen (mit der Spektrenbibliothek, siehe 3.3).

3.4.1.2 Level 2 Ufid-Zuordnung

Eine zweite Vorgehensweise für das gruppieren von Features wurde getestet, die nur m/z und t_R-Informationen für das Clustering heranzieht, und keinen MS^2 Abgleich durchführt. Der Vorgang läuft unabhängig von der Level 1 Zuordnung und wird in einem separaten Feld hinterlegt (ufid2). Somit ist es möglich, beide Stufen der Ufid-Zuordnung parallel zu implementieren. Für ufid2 werden keine MS^2-Spektren benötigt und es kann somit eine deutlich höhere Anzahl an Dokumenten (Features) gruppiert werden. Unter ufid2 werden mehr Gruppen gebildet, auch von Features mit geringeren Intensitäten. Die Anwendung einer Ufid-Bibliothek ist auch nicht notwendig, da keine MS^2-Spektren gespeichert werden müssen. Die Vertrauenswürdigkeit oder Qualität der Zuordnung (im Sinne von falsch Positiven) ist vermutlich schlechter, wegen der höheren Wahrscheinlichkeit, dass isobare, ko-eluierende Substanzen (zwei Substanzen mit ähnlichen m/z und t_R) zusammen gruppiert werden.

3.4.2 Erprobung der Ufid-Zuordnung

Der Algorithmus wurde auf eine Testdatenbank mit über 2 Millionen Features angewendet. Die Datenbank teilt sich in drei Datensätzen: 1) Messungen täglicher Mischproben des Rheins bei Koblenz (Oberflächenwasserproben, 2128 Messungen), 2) Schwebstoffmessungen von Proben der Umweltprobenbank (7 Messstellen, 172 Messungen) und 3) 24-Std.-Mischproben der Ruhr bei Mülheim (Oberflächenwasserproben, 63 Messungen). Die Proben der Ruhr wurden durch das LANUV mit einer eigenen NTS-Methode gemessen. Dieser Datensatz wurde hinzugefügt, um die Harmonisierung externer Daten zu demonstrieren.

3.4.2.1 Harmonisierung Ruhr-Daten

Die Daten der Ruhr wurden mit einer anderen chromatographischen Methode aufgenommen (LANUV Methode, (Brüggen und Schmitz 2018)). Um die Retentionszeiten vergleichen zu können, wurde zunächst ein Umrechnungsmodell angewandt, das eine Angleichung der BfG- und LANUV-Retentionszeiten ermöglicht (Stanstrup et al. 2015). Das Modell wurde mit Hilfe von annotierten Verbindungen (Bibliothek Screening), die in beiden Datensätzen gefunden wurden, aufgebaut. Da beide Datensätze genug intrinsische Marker (6171 Datenpunkte von 196 Verbindungen) hatten, um das Modell aufzubauen, konnte die Anwendung eines gemeinsamen Standard-Mix umgangen werden. Die Retentionszeiten aller Features des LANUV-Datensatzes (auch aller Unbekannten) wurden mit diesem Modell in BfG-Retentionszeiten umgerechnet (siehe auch Abbildung 4: Umrechnung von Retentionszeiten für Spektren des UBA).

3.4.2.2 Erprobung der Level 1 Ufid-Zuordnung

Von insgesamt 2 Millionen Features wurden 69% der Features in 26,600 Gruppen (Ufids) eingeteilt. Die Parameter für die Ufid-Zuordnung sind in Tabelle 3 aufgelistet. Die übrigen 31% nicht gruppierten Features (im folgenden Waisen genannt) waren entweder Features ohne MS^2, die auch durch das Gap-Filling nicht zugeordnet werden konnten oder es waren vereinzelte Befunde, die nicht die Mindestmenge von 5 Features je Gruppe erreichen konnten. Der Anteil von Waisen hing hauptsächlich von der Häufigkeit von MS^2-Spektren ab. Daher ist eine optimierte Data-Dependent MS^2 Methode, mit möglichst vielen MS^2-Aufnahmen, für die Level 1 Ufid-Zuordnung unabdingbar.

Die Richtigkeit der Ufid-Zuordnung wurde durch die Übereinstimmung der Substanz-Annotationen (Bibliothek Screening) innerhalb einer Ufid-Gruppe überprüft. Idealerweise entspricht eine Ufid-Gruppe einer Substanz. Wenn eine Ufid fälschlicherweise mehreren Substanzen entspricht, wurden alle Features einer der Annotationen (der am häufigsten auftretende Name) als “richtige” Ufid-Zuordnungen gezählt und alle anderen Features als “falsche” Ufid-Zuordnungen gezählt. Diese Berechnung beschränkt sich auf annotierte Features (72000, ca. 4% aller Features) und gilt unter der Annahme, dass die Annotation über das Bibliothek Screening richtig ist, was mit 95%iger Sicherheit gegeben ist (Jewell et al. 2019). Letztendlich hatten 7% der Features eine falsche Ufid-Zuordnung. Darunter fallen Substanzen, die selbst bei händischer Auswertung schwer zu unterscheiden sind (z.B. Regioisomere). Die Anwendung der hierdurch gebildeten Zeitreihen zum “Screening nach Unbekannten” werden in Abschn. 3.4.3.2 demonstriert.

3.4.2.3 Erprobung der Level 2 Ufid-Zuordnung

Die Anzahl an Waisen (siehe Abschn. 3.4.2.2) ist mit 3% bei der Level 2 Ufid-Zuordnung wesentlich geringer. Auch Features mit geringeren Intensitäten, wofür keine MS^2-Spektren aufgenommen wurden, wurden gruppiert. Allerdings bestehen diese zusätzlichen Gruppen aus Features, die sich nicht oder nur mit erheblich mehr Aufwand identifizieren lassen, da sie die zur Strukturaufklärung erforderlichen MS^2-Spektren nicht besitzen.

Die Anzahl an falschen Zuordnungen (basierend auf dem Bibliothek Screening, siehe Abschn. 3.4.2.2), durch isobare Substanzen, die über das Clustering anhand von m/z und der Retentionszeit nicht unterschieden werden konnten, betrug 11%. Sie ist nur geringfügig schlechter als bei der Level 1 Ufid-Zuordnung.

Die hierdurch gebildeten Zeitreihen von Rhein Oberflächenwasser bei Koblenz wurden zur Entwicklung und Erprobung weiterer Analysewerkzeuge angewendet (siehe 3.5).

Tabelle 3: Parameter für das Clustering und Ufid-Zuordnung

Parameter	Wert
m/z Toleranz	7 mDa
RT Toleranz	1 Min.
MS2 Dot-Product Schwellenwert	50 (von 1000)
Minimum Anzahl Features für ein Cluster (minPts)	5

3.4.2.4 Zukünftige Schritte in der Arbeit mit Unbekannten

Die prototypischen Algorithmen und abgeschlossenen Versuche bilden eine Basis für die weitere Entwicklung des Alignments nach dem Projektende. Ziel ist es, einen automatisierten Workflow zu haben, der ohne menschliches Eingreifen eingetragene Features weitgehend fehlerfrei in bestehende oder neue Gruppen (Ufids) einsortiert.

Um die Anzahl der sog. Waisen und falsche Zuordnungen zu reduzieren, sowie den Code robuster gegen Ausreißer zu machen, wird noch weitere Entwicklungsarbeit benötigt (siehe Abschnitt 4). Anschließend kann die Auswertung weiterer Messdaten, über den oben genannten Testdatensatz hinaus, erfolgen.

3.4.3 Erweiterung des Front-Ends für nicht annotierte Features

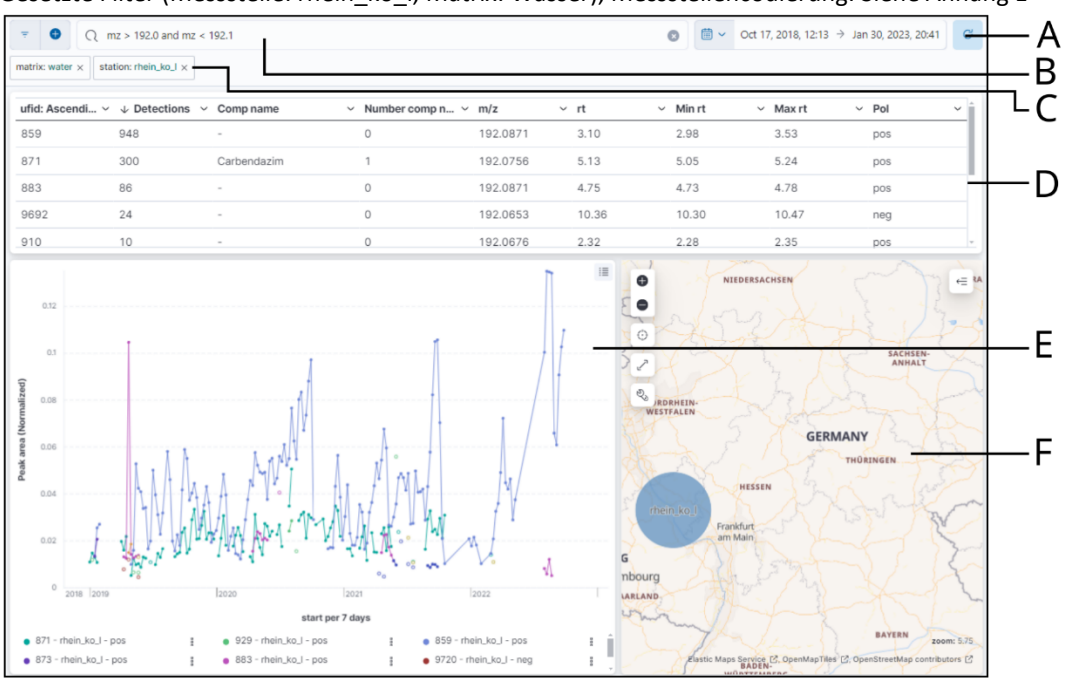
Die vorgestellten Dashboards in Phase 1 des Projektes können nur für die Suche und Darstellung der annotierten Features angewendet werden. Für die Suche von nicht annotierten Features über m/z, t_R (z.B. für das Suspect-Screening) oder für die Suche nach unbekannten Features und die anschließende Identifizierung, wurden neue Dashboards benötigt. Die Funktionsweise und das Aussehen der Dashboards ist vergleichbar mit den Dashboards aus Phase 1, allerdings ist der Entwicklungsstand in Phase 2 weniger fortgeschritten. Beispielsweise wurden die interaktiven Priorisierungstools (siehe 3.3.7.1) noch nicht für unbekannte Features implementiert. Die Anzahl an Proben, die für Phase 2 ausgewertet wurden (1277 Proben), ist auch geringer als für Phase 1 (5920 Proben).

3.4.3.1 Suspect Screening

Mit Hilfe des Dashboards “Time Series Unknown Compounds”, können die Anwender*innen einen Massenbereich für eine Substanz in der Suchbox eingeben (unter Anwendung einer Schiebereglers oder eines Eingabefeldes) und so die Treffer als Zeitreihen ansehen. So kann im Suspect Screening nach einer Substanz mit bekannter m/z (für Substanz und Addukt) gesucht werden. Treffer werden tabellarisch, nach Ufid sortiert, gelistet. Weitere, verknüpfte Dashboards bieten zum Beispiel die Überprüfung der MS^2-Spektren an (Abbildung 17).

Abbildung 17: Bildung von Zeitreihen für Suspect-Screening in Wasserproben

Screenshot eines Dashboards für Suspect Screening A: Eingabe Zeitraum, B: Suchbox mittels “Kibana Query Language”, in diesem Fall wurde “mz > 192.0 and mz < 192.1” eingegeben. Siehe Datenbankschema im Wiki, C: Gesetzte Filter (Messstelle: rhein_ko_l, Matrix: Wasser), Messstellencodierung: Siehe Anhang E



Quelle: eigene Abbildung BfG

3.4.3.2 Suche nach Unbekannten

Für diese Variante des NTS wird nach bisher unbekannten Substanzen, die eine Umweltrelevanz haben, gesucht. Eine Strukturaufklärung kann je nach Relevanz folgen.

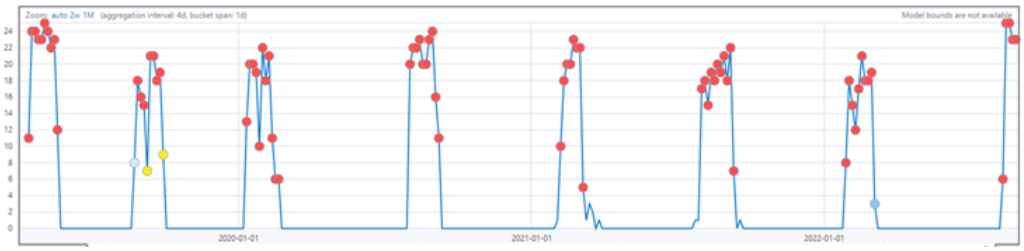
Ein typisches Beispiel ist die Zeitreihenanalyse. Industriell eingetragene Substanzen weisen oft eine diskontinuierliche zeitliche Emission auf und können somit vom natürlichen Hintergrund unterschieden werden (oft “Priorisierung” genannt). Es gibt eine Vielzahl verschiedener Software für die Zeitreihenanalyse.

Beispielhafte Zeitreihenanalyse

Das Anomaly Detection Modul in Kibana wurde angewendet, um die Zeitreihen der Substanzen der Tagesmischproben aus Koblenz (Rhein, Oberflächenwasser, 2019 bis 2022) zu sortieren (26.606 Zeitreihen). Die Anzahl der Detektionen, die maximale Peakfläche und die Veränderung der Peakfläche wurden als Faktoren bei der Berechnung des “Anomaly Score” angewendet. Diese Kennzahl erlaubt eine Sortierung der Zeitreihen, die für die manuelle Plausibilisierung im Detail visualisiert werden kann. Abbildung 18 zeigt eine der Top 50 Zeitreihen.

Abbildung 18: Screenshot aus dem „Anomaly Detection Module“ in Kibana und Visualisierung einer priorisierten Zeitreihe

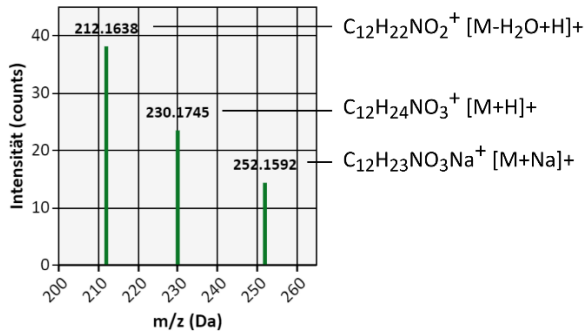
Y-Achse: Peakfläche. X-Achse: Jahresverlauf von 2019 bis 2022. Farblich markierte Punkte zeigen Messpunkte mit einem “anomaly score” (rot: hoch, gelb: niedrig).



Quelle: eigene Abbildung BfG

Das Feature aus Abbildung 18 wurde näher untersucht. Im ersten Schritt wurden die Addukte und Isotopologen dieses Features als Gruppe betrachtet, dadurch konnte festgestellt werden, dass es sich um ein ESI-Quellen-Fragment (Verlust von H_2O) handelte. Zudem konnten das Natrium-Addukt und das Wasserstoff-Addukt detektiert werden, allerdings mit niedrigeren Peakflächen. Durch die Anwendung von Genform (Meringer et al. 2011) konnten mögliche Summenformeln generiert werden (Abbildung 19).

Abbildung 19: Massenspektrum des Unbekannten Features bei m/z 230.1738 und Retentionszeit 9,4 Minuten

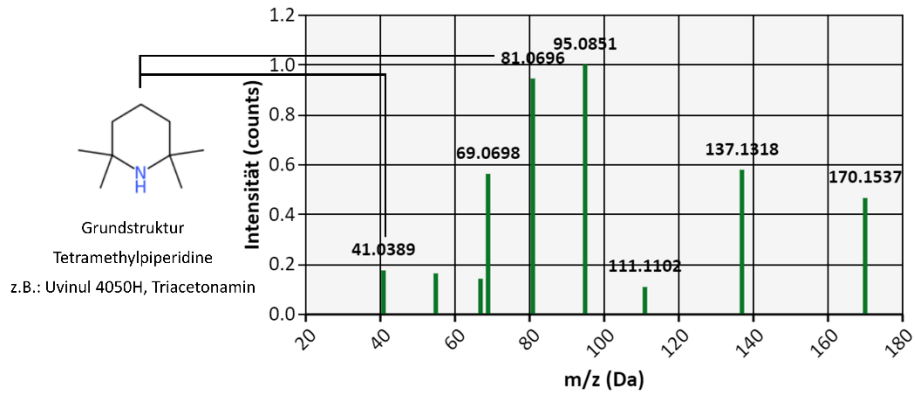


Quelle: eigene Abbildung BfG

Eine der vorgeschlagenen Summenformeln hat 26 Substanzvorschläge in der *PubChemLite* Liste ergeben (Schymanski et al. 2021). Diese Substanzen wurden zusammen mit dem MS^2 Spektrum betrachtet. Mögliche Fragmentierungswege, die zur Struktur passen, wurden postuliert. Eine mögliche Grundstruktur, die in einigen Kandidaten vorhanden war, ist das Tetramethylpiperidine. Das MS^2-Spektrum zeigt zwei Fragmente, die häufig durch diese Struktur verursacht werden (Abbildung 20).

Die Strukturaufklärung konnte während des Projektes nicht abgeschlossen werden. Jedoch hat das Beispiel gezeigt, dass die Prozessierung und Qualitätssicherung der Messdaten mit dem NTSPortal automatisiert und im Voraus durchgeführt werden kann. Die großen Datensätze, in diesem Fall Zeitreihen über 3 Jahre mit einer hohen Auflösung, waren schnell aufrufbar und konnten in ihrer Gesamtheit in wenigen Schritten durchsucht werden. Die Identifizierung von Unbekannten, ein arbeitsintensiver Schritt, kann somit fokussierter erfolgen, gegebenenfalls durch spezialisiertes Personal.

Abbildung 20: MS²-Spekrum des unbekannten Features bei m/z 230.1738 und Retentionszeit 9,4 Minuten



Quelle: eigene Abbildung BfG

3.5 Versuche zum maschinellen Lernen

3.5.1 Zusammenarbeit mit Horváth & Partners

Im Januar 2022 begann die Firma Horváth & Partners mit einem vom UBA beauftragten Projekt zur Entwicklung von KI-unterstützten Werkzeugen für das NTSPortal. Diese wurden in einer grafischen Applikation verfügbar gemacht und sollen bei der Auswertung von NTS-Datensätzen eingesetzt werden. Der Hintergrund dieses Vorhabens war eine Initiative des UBA, mit dem Ziel, KI-Projekte in den eigenen Resorts zu fördern. Das NTSPortal wurde als Fallstudie ausgewählt, da ein relativ großer Datensatz zur Verfügung stand und eine konkrete Fragestellung bereits vorlag.

Die Anforderung an die Applikation war die Möglichkeit der Zeitreihenanalyse zur Mustererkennung und Priorisierung von Features. Das Ziel war es, Features zu finden, die wahrscheinlich einen anthropogenen Ursprung haben.

Die Datensätze für diese Anwendung wurden vorbereitet und an Horváth in Form eines *Elastic Dumps* übergeben. Eine ElasticSearch Datenbank als Spiegelung der BfG Datenbank wurde bei Horváth & Partners aufgebaut. Neben der Vorstellung und Beratung zu den Ergebnissen (*Sprint Review*), fanden eine Reihe von Beratungsgesprächen zwischen BfG und Horváth & Partners zu der Theorie des Non-Target-Screenings statt. Verschiedene Ansätze zum Clustering der Daten wurden in Python ausprobiert. Es wurde eine Kombination aus Werkzeugen entwickelt, die als *Recommender* für Zeitreihen beschrieben wurden. Der *Recommender* soll es ermöglichen, Zeitreihen zu priorisieren. Hierfür schlägt der User eine oder mehrere Zeitreihen vor und die KI zeigt ähnlichen Zeitreihen im Datensatz. Beispielsweise gibt der User eine Zeitreihe von einer bekannten Industriechemikalie vor (z.B. Tetrabutylammonium) und es werden Zeitreihen, die auch ein diskontinuierliches Eintragungsmuster aufweisen, ausgewählt und vorgeschlagen.

Tests zu Anomalie-Detektion und Vorhersage in einer Web-Anwendung

Kibana wurde als Plattform für die Erstellung des Web-Dashboards zur Recherche und Visualisierung verwendet. Wie in Abschn. 3.3.7 beschrieben, umfasst *Kibana* anpassbare Analyse-Werkzeuge zur Erkennung von Anomalien und zur Trendvorhersage (sogenannte *Machine Learning Module*). Diese Werkzeuge sind ideal, um mit der Implementierung von ML

mit der NTS-Datenbank zu beginnen. Die statistischen Analysemethoden, die in diesem Modul angewendet werden, sind auf der Internetseite von Elastic dokumentiert.

Die Anomalie-Detektion wurde exemplarisch angewendet, um Befunde mit außergewöhnlich hohen oder niedrigen Peakflächen in den Zeitreihen von bekannten Substanzen im Rhein bei Koblenz zu priorisieren und die Anwendung von Algorithmen des maschinellen Lernens in der Praxis zu testen. Hierfür wurde die medial-normierte Peakfläche [A/med(A)] als Kenngröße herangezogen. Anomalien bekommen eine Bewertung zwischen 1 und 100, wonach der Anwender die Ergebnisse (nach Feature oder nach Substanz) sortieren kann (Abbildung 21A). Gefundene Anomalien können in einigen Fällen diskontinuierlichen Einträgen, typischerweise durch industrielle Emissionen, zugeordnet werden (siehe Beispiel Tetrabutylammonium, Abbildung 21B). Die Anomalie-Detektion kann somit als Priorisierungs-Tool dienen. In vereinzelten Fällen waren die Anomalien offensichtlich Ausreißer (Artefakte, siehe Beispiel Abbildung 21C und Abbildung 21D), die nachträglich korrigiert werden können (z.B. durch Entfernung des Eintrags). Die Anomalie-Detektion kann somit auch zur Qualitätssicherung beitragen.

Abbildung 21: Analyse von Zeitreihendaten mit Hilfe der „Anomaly Detection Module“ in Kibana

Datenquelle: normierte Peakfläche, tägliche Wasserproben aus Rhein/Koblenz.
A: Auflistung der gefundenen Anomalien und der Substanzen, die dazu beigetragen haben. Rote Markierungen in der Heatmap zeigen Anomalien mit einem Score > 75.
B: Visualisierung der Zeitreihen mit Annotationen. C: Zeitreihe von Valsartan mit einem unerwartet hohen Anomalie-Score im Sommer 2020 (Score: 99). D: Zoomansicht der Anomalie von Valsartan (Ausreißer)

Quelle: eigene Abbildung BfG

3.6 IT Infrastruktur und Online Bereitstellung

Zu Beginn des Projekts wurde die vorhandene Server-Infrastruktur der BfG genutzt, da der Speicherbedarf relativ gering war und mit vorhandener Hardware gedeckt werden konnte. Im

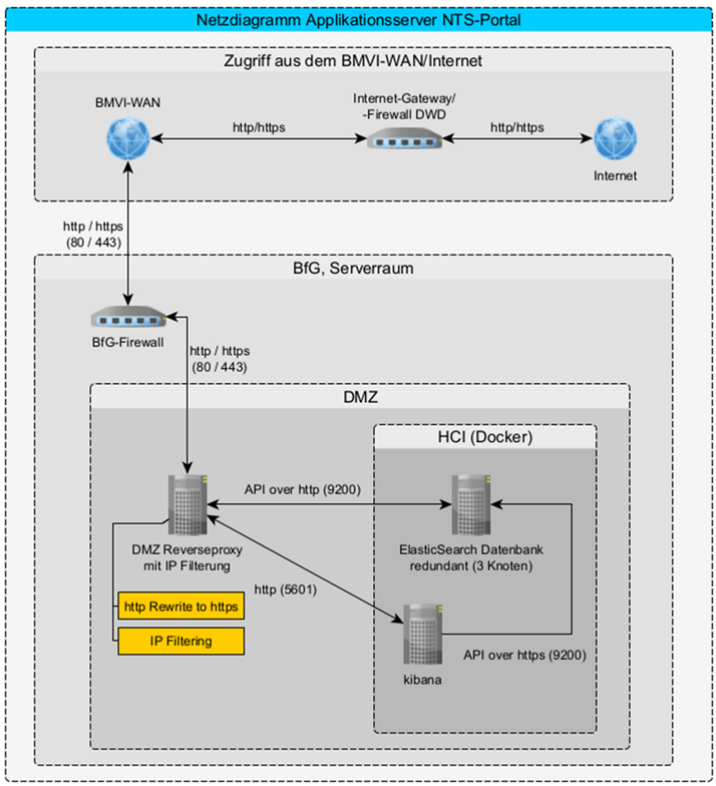
Laufe des Projektes wurde das System in eine Hyperconverged Infrastructure (HCI) Umgebung übertragen. HCI ist ein IT-Architekturansatz, bei dem die Komponenten der IT-Infrastruktur in einem einzigen Software-Defined-System zusammengefasst sind und über diese administriert werden. Dazu zählen Server (Prozessoren, Arbeitsspeicher, Storage), virtuelle Maschinen und auch Netzwerkkomponenten. Dieser Umzug ins HCI System brachte höhere Verfügbarkeiten und bessere automatisierte Sicherungsmaßnahmen.

Durch die eingesetzte Virtualisierungstechnik *Docker* konnte die Sicherheit erhöht werden und ein direktes Zugreifen auf einzelne Server zielgerichtet ermöglicht werden. Durch Docker ist es ebenfalls möglich, die Container (Virtualisierungseinheit) dynamisch zu skalieren, falls es zukünftig benötigt wird.

Es wurde folgende (virtualisierte) Hardware im HCI System benutzt (Abbildung 22):

- Eine Virtuelle Maschine mit Rocky Linux in der Version 8, auf dem HCI System, das die Grundlage für die Virtualisierung mit Docker bereitstellt.
  - Für die Datenbank: *ElasticSearch* wurden drei virtualisierte Knoten mit Hilfe von Docker erzeugt und
  - für das Web-Portal: *Kibana* ein Knoten mit Docker.
- Es wurden jeweils die offiziell von Elastic zur Verfügung gestellten Docker Images genutzt. Für ElasticSearch ist dies zu finden unter: https://www.docker.elastic.co/r/elasticsearch und für Kibana unter: https://www.docker.elastic.co/r/kibana

Abbildung 22: Übersicht der Server-Infrastruktur



Quelle: eigene Abbildung BfG

3.6.1 Erreichbarkeit außerhalb des BfG Intranets

Von außen ist *Kibana* bzw. das NTSPortal erreichbar über https://ntsportal.bafg.de, abgesichert mit einem SSL Zertifikat von *Let's Encrypt* für den verschlüsselten Übertragungsweg. Eine Freigabe der ElasticSearch Rest API (nur lesender Zugriff) wurde auf ähnliche Weise eingerichtet und ist über https://elastic.bafg.de zu erreichen.

Zudem erfolgt die gezielte Freischaltung der mitarbeitenden Behörden anhand derer IP-Adressen / IP-Adressbereiche über einen Apache Proxy Server (IP-Whitelisting). Es besteht dadurch kein globaler Zugang aus dem Internet. Aus Sicherheitsgründen ist dies für die Entwicklungsphase nötig.

Kibana wurde zudem durch Benutzername und Passwort abgesichert, sodass jeder Nutzer für den Login einen zuvor angelegten Benutzernamen mit Passwort benötigt, der jederzeit durch Admins deaktiviert oder gelöscht werden kann. Die API-Schnittstelle wurde zudem mit Benutzernamen und Passwort und einen API Key abgesichert, sodass der Zugriff für Unbefugte verhindert wird. Für die API Schnittstelle gelten dieselben Rechte wie für Benutzer von *Kibana*. Für genauere Informationen bezüglich des Rechtekonzepts siehe Abschn. 3.6.2.

Für den Aufbau des NTSPortals wurde BfG-intern ein IT-Sicherheitskonzept erstellt.

3.6.2 Rechte- / Rollenkonzept

Daten werden in unterschiedliche Indices (Datenbanken) derselben Struktur geteilt, je nach Institution oder Projekt. Diese Aufteilung ist für den Anwender nicht sichtbar, aber es erlaubt die Zuordnung von Rollen und Datensätzen. Jedes Nutzerkonto ist mit einer oder mehreren Rollen bestückt. Eine Rolle hat gewisse Rechte, beispielsweise das Lesen eines bestimmten Indizes.

Das Rollenkonzept funktioniert wie folgt: Nach dem Einloggen öffnet sich ntsportal.bafg.de mit einem Dashboard, das auf Indizes mit Ergebnissen des Bibliothek Screenings zugreift. Entsprechend der zugewiesenen Rolle erhalten die Nutzer*innen Zugriff auf bestimmte Indizes und nur diese werden sichtbar.

Bei der Abgabe von Daten (z.B. durch ein Landesumweltamt als “Datenlieferant”) wird besprochen, wer die Daten im NTSPortal sehen darf und was damit gemacht werden darf. Der einfachste und transparenteste Weg dafür ist, eine Open-Data-Lizenz anzuwenden. Als Beispiel werden alle Daten der BfG, UBA und Sachsen mit der Deutschland-Lizenz-Namensnennung-2 (dl-de/by-2-0) versehen und können somit nach den Lizenzbestimmungen dargestellt und weiterverwendet werden.

4 Ausblick

Das NTSPortal in seiner jetzigen Form wird über die Projektlaufzeit hinaus weiterhin aufrufbar bleiben und weiterentwickelt. Es ist geplant, den Zugang schrittweise für weitere Institutionen und zukünftig für alle auszuweiten. Das System wird weiterhin von der BfG verwaltet und fest in die bestehende Datendienste der BfG und UBA integriert.

Ein Anschlussvorhaben, finanziert durch und in Kooperation mit dem UBA, wird die Weiterentwicklung des NTSPortals voranbringen (Lessmann et al. 2024). Die Weiterentwicklung bezieht sich auf die derzeitigen Herausforderungen, die zum Teil schon angesprochen wurden:

1. Weiterentwicklung der automatisierten Prozessierung, um die Qualität und Zuverlässigkeit der Routinen zu verbessern
2. NTS-Daten mit weiteren Informationsquellen besser zu integrieren, oder mit diesen zu verknüpfen, z.B. Abflussdaten, Konzentrationen (Target-Daten) oder (öko)toxikologische und gesetzliche Informationen zu einzelnen Verbindungen.
3. Entwicklung interaktiver Anwendungen, bevorzugt als Onlinedienst, für die statistische oder vergleichende Analyse der Daten (z.B. für die Priorisierung von Features).
4. die Integration von geschätzten Konzentrationen für annotierte Verbindungen als Unterstützung für die Risikobewertung. Diese können retrospektiv berechnet werden, beispielsweise durch externe Kalibration. Alternative Ansätze, z.B. die Modellierung der ESI Ionisierungseffizienz (Malm et al. 2021) werden aktuell erforscht.

Die Kooperationen mit anderen nationalen und internationalen NTS Vorhaben, die in diesem Projekt initiiert wurden, bleiben durch die regelmäßige Veranstaltung von Begleitkreistreffen in den REFOPLAN Projekten Gewässerbeobachtung der Zukunft I und II bestehen. Das NTSPortal ist eine wichtige Ansprechstelle für NTS Strategien der Bundesländer. Vernetzungen und Austausch fanden bereits auf der Abschlussveranstaltung im Dezember 2023 und während der Begleitkreistreffen statt. Die BfG ist auch ein Partner in dem Rhein-Projekt-NTS (Ondruch und Heinz 2024), koordiniert durch die IKSR, in dem eine Datenbank für den Austausch von NTS Daten zwischen den Messstationen am Rhein aufgebaut wurde. Ein Austausch von Spektrenbibliotheken hat bereits zwischen einigen Partner stattgefunden und eine Harmonisierung von Bibliotheken über das komplette Konsortium wird angestrebt. Zukünftig ist auch ein Austausch oder eine Verknüpfung der NTS-Daten verschiedener Datenbanken geplant und die Entwicklung zur Harmonisierung von NTS Daten, die in diesem Projekt gemacht worden sind, kann dafür als Grundlage dienen. Dem europäischen PARC Projekt (Partnership for the Assessment of Risks from Chemicals) wurden NTS Daten der Schwebstoffzeitreihen zu Verfügung gestellt, die als Beispieldatensatz für die Entwicklung von sog. “Early Warning”-Systemen dienen kann. Dem NORMAN Konsortium wurden ebenfalls Daten für die Entwicklung der Digital Sample Freezing Plattform (DSFP) zur Verfügung gestellt. Ein weiterer wichtiger Zwischenschritt zur Verknüpfung der Datenbanken ist die Entwicklung einer einheitlichen Ontologie für Non-Target-Screening-Datenbanken (harmonisierte Begriffe sowie Einigung auf Dateiformate für den Austausch). Dies wird auch in Begleitkreistreffen oder im bilateralen Austausch diskutiert, aber die Bildung eines Konsortiums mit allen beteiligten Institutionen und Stakeholder steht noch aus. Die Veröffentlichung (sowohl des dokumentierten Codes als auch wissenschaftlichen Publikationen zum Grundprinzip und Anwendung der Datenbank) für das NTSPortal wird als erste wichtige Voraussetzung für die Harmonisierung und Kooperation gesehen.

5 Quellen

Monografien:

Dierkes, G., L. Boulard und A. Wick (2024). Non-Target Screening in Schwebstoff- und Biotaproben. Texte. Dessau-Roßlau, Umweltbundesamt. **77**.

Han, J., M. Kamber und J. Pei (2011). Data Mining : concepts and techniques. Burlington, Elsevier Science, 10.1016/C2009-0-61819-5.

Jewell, K. S., N. Hermes, B. Ehlig, F. Thron, T. Köppe, U. Thorenz, M. P. Schlüsener, C. Dietrich, A. Wick und T. A. Ternes (2021). Methodik zur Anwendung von Non-Target Screening (NTS) mittels LC-MS/MS in der Gewässerüberwachung. Texte. Dessau-Roßlau, Umweltbundesamt. **144**.

Köppe, T., N. Hermes, S. Rohde und A. Wick (2024). Studie zur Ermittlung der stofflichen Belastung, Belastungsschwerpunkte und Eintragspfade in ausgewählten sächsischen Gewässern durch ein Non-Target-Screening. Dresden, Sächsisches Landesamt für Umwelt Landwirtschaft und Geologie.

Lessmann, O., E. Rosenheinrich, K. Jewell, A. Wick, A. L. Kronsbein und J. Koschorreck (2024). Weiterentwicklung des Online-Portals für die Gewässerbeobachtung der Zukunft. Berlin, UBA.

Niarchos, G., M. Engwall, L. I. Bengtström, M. Larsson, J. Lundqvist und L. Ahrens (2023). Concept and plan of effect-based monitoring and effect directed analysis (EDA) of chemicals towards EWS. Lund, Swedish University of Agricultural Sciences (SLU).

Schulz, W. und T. Lucke (2019). Non-Target Screening in der Wasseranalytik - Ein Leitfaden zur Anwendung der LC-ESI-HRMS für Screening-Untersuchungen. Langenau, Wasserchemische Gesellschaft (GDCh).

Schulze, T. und M. Ricking (2005). Entwicklung einer Verfahrensrichtlinie „Sedimente und Schwebstoffe“.

Zeitschriftenaufsätze:

Alygizakis, N. A., P. Oswald, N. S. Thomaidis, E. L. Schymanski, R. Aalizadeh, T. Schulze, M. Oswaldova und J. Slobodnik (2019). "NORMAN digital sample freezing platform: A European virtual platform to exchange liquid chromatography high resolution-mass spectrometry data and screen suspects in “digitally frozen” environmental samples." In: TrAC Trends in Analytical Chemistry, 115, 129-137, 10.1016/j.trac.2019.04.008.

Bagheri, H., J. Slobodnik, R. M. M. Recasens, R. T. Ghijsen und U. A. T. Brinkman (1993). "Liquid chromatography — Particle beam mass spectrometry for identification of unknown pollutants in water." In: Chromatographia, 37, 3, 159-167, 10.1007/BF02275854.

Brinkmann, M., D. Montgomery, S. Selinger, J. G. Miller, E. Stock, A. J. Alcaraz, J. K. Challis, L. Weber, D. Janz und M. Hecker (2022). "Acute toxicity of the tire rubber-derived chemical 6PPD-quinone to four fishes of commercial, cultural, and ecological importance." In: Environmental Science & Technology Letters, 9, 4, 333-338, 10.1021/acs.estlett.2c00050.

Brüggen, S. und O. J. Schmitz (2018). "A New Concept for Regulatory Water Monitoring Via High-Performance Liquid Chromatography Coupled to High-Resolution Mass Spectrometry." In: Journal of Analysis and Testing, 2, 4, 342-351, 10.1007/s41664-018-0081-5.

Helmus, R., B. van de Velde, A. M. Brunner, T. L. ter Laak, A. P. van Wezel und E. Schymanski (2022). "patRoon 2.0: Improved non-target analysis workflows including automated transformation product screening." In: Journal of Open Source Software, 7, 71, 10.21105/joss.04029.

Hollender, J., B. van Bavel, V. Dulio, E. Farmen, K. Furtmann, J. Koschorreck, U. Kunkel, M. Krauss, J. Munthe und M. Schlabach (2019). "High resolution mass spectrometry-based non-target screening can support

regulatory environmental monitoring and chemicals management." In: Environmental Sciences Europe, 31, 1, 42, 10.1186/s12302-019-0225-x.

Jewell, K. S., U. Kunkel, B. Ehlig, F. Thron, M. Schlüsener, C. Dietrich, A. Wick und T. A. Ternes (2019). "Comparing mass, retention time and MS2 spectra as criteria for the automated screening of small molecules in aqueous environmental samples analyzed by LC-QToF-MS/MS." In: Rapid Communications in Mass Spectrometry, 34, e8541, 10.1002/rcm.8541.

Jewell, K. S., F. Thron, B. Ehlig, N. Hermes, S. Quanz, M. P. Schlüsener, I. Fettig, J. Koschorreck, K. Kramer, T. Scharrenbach, T. A. Ternes und A. Wick (2021). Non-Target-Screening, Identifikation und räumliche Eingrenzung von Stoffeinträgen in Gewässern. Wasserforum. Langenau, Dr. Margareta Dellert-Ritter. **14:** 48-50, https://www.umweltbundesamt.de/publikationen/methodik-zur-anwendung-von-non-target-screening-nts.

Kasprzak, P., L. Mitchell, O. Kravchuk und A. Timmins (2021). "Six Years of Shiny in Research--Collaborative Development of Web Tools in R." In: arXiv preprint arXiv:2101.10948, 10.48550/arXiv.2101.10948.

Malm, L., E. Palm, A. Souihi, M. Plassmann, J. Liigand und A. Kruve (2021). "Guide to Semi-Quantitative Non-Targeted Screening Using LC/ESI/HRMS." In: Molecules, 26, 12, 10.3390/molecules26123524.

Meringer, M., S. Reinker, J. Zhang und A. J. M. C. M. C. C. Muller (2011). "MS/MS data improves automated determination of molecular formulas by mass spectrometry." In, 65, 2, 259-290.

Nothias, L.-F., D. Petras, R. Schmid, K. Dührkop, J. Rainer, A. Sarvepalli, I. Protsyuk, M. Ernst, H. Tsugawa und M. Fleischauer (2020). "Feature-based molecular networking in the GNPS analysis environment." In: Nature methods, 17, 9, 905-908, 10.1038/s41592-020-0933-6.

Schmid, R., S. Heuckeroth, A. Korf, A. Smirnov, O. Myers, T. S. Dyrlund, R. Bushuiev, K. J. Murray, N. Hoffmann, M. Lu, A. Sarvepalli, Z. Zhang, M. Fleischauer, K. Dührkop, M. Wesner, S. J. Hoogstra, E. Rudt, O. Mokshyna, C. Brungs, K. Ponomarov, L. Mutabdžija, T. Damiani, C. J. Pudney, M. Earll, P. O. Helmer, T. R. Fallon, T. Schulze, A. Rivas-Ubach, A. Bilbao, H. Richter, L.-F. Nothias, M. Wang, M. Orešič, J.-K. Weng, S. Böcker, A. Jeibmann, H. Hayen, U. Karst, P. C. Dorrestein, D. Petras, X. Du und T. Pluskal (2023). "Integrative analysis of multimodal mass spectrometry data in MZmine 3." In: Nature Biotechnology, 41, 4, 447-449, 10.1038/s41587-023-01690-2.

Schymanski, E. L., J. Jeon, R. Gulde, K. Fenner, M. Ruff, H. P. Singer und J. Hollender (2014). "Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence." In: Environmental Science & Technology, 48, 4, 2097-2098, 10.1021/es5002105.

Schymanski, E. L., T. Kondić, S. Neumann, P. A. Thiessen, J. Zhang und E. E. Bolton (2021). "Empowering large chemical knowledge bases for exposomics: PubChemLite meets MetFrag." In: Journal of Cheminformatics, 13, 1, 19, 10.1186/s13321-021-00489-0.

Schymanski, E. L., H. P. Singer, P. Longree, M. Loos, M. Ruff, M. A. Stravs, C. R. Vidal und J. Hollender (2014). "Strategies to Characterize Polar Organic Contamination in Wastewater: Exploring the Capability of High Resolution Mass Spectrometry." In: Environmental Science & Technology, 48, 3, 1811-1818, 10.1021/es4044374.

Stanstrup, J., S. Neumann und U. Vrhovšek (2015). "PredRet: Prediction of Retention Time by Direct Mapping between Multiple Chromatographic Systems." In: Analytical Chemistry, 87, 18, 9421-9428, 10.1021/acs.analchem.5b02287.

Stravs, M. A., E. L. Schymanski, H. P. Singer und J. Hollender (2013). "Automatic recalibration and processing of tandem mass spectra using formula annotation." In: Journal of Mass Spectrometry, 48, 1, 89-99, 10.1002/jms.3131.

Tian, Z., H. Zhao, K. T. Peter, M. Gonzalez, J. Wetzel, C. Wu, X. Hu, J. Prat, E. Mudrock, R. Hettinger, A. E. Cortina, R. G. Biswas, F. V. C. Kock, R. Soong, A. Jenne, B. Du, F. Hou, H. He, R. Lundeen, A. Gilbreath, R. Sutton, N. L. Scholz, J. W. Davis, M. C. Dodd, A. Simpson, J. K. McIntyre und E. P. Kolodziej (2020). "A ubiquitous tire

rubber–derived chemical induces acute mortality in coho salmon." In: Science, eabd6951, 10.1126/science.abd6951.

Xiao, J. F., B. Zhou und H. W. Ressom (2012). "Metabolite identification and quantitation in LC-MS/MS-based metabolomics." In: Trends in analytical chemistry : TRAC, 32, 1-14, 10.1016/j.trac.2011.08.009.

Internetadressen:

Cheminfo. (2024). Cheminfo. 2024, https://chemikalieninfo.de/ (02.06.2024).

Eckert, A. (2018). Parallel distance matrix computation using multiple threads [R package paralleldist version 0.2.4]. https://cran.r-project.org/web/packages/parallelDist/index.html (01.11.21), 10.32614/CRAN.package.parallelDist.

ICWRGC. (2024). GEMStat. 2024, https://gemstat.org/ (02.06.2024).

Kjøller, C. (2024). AQUAPLEXUS. 2024, https://aquaplexus.dk/ (04.06.2024).

Loos, M. (2024). enviMass - mass spec analysis workflow - version 4.4. 2024, www.envibee.ch (06.06.2024).

Müller, U. (2023). K2I Spurenstoff-Tracker. https://www.k2i-tracker.de/ (02.06.2024).

Ondruch, P. und M. D. Heinz. (2024). Pilot project on non-target screening. 2024, https://www.iksr.org/en/iksr/rhein-2040/rhine-project-non-target-screening (02.06.2024).

Singer, H. (2024). NTSuisse. 2024, https://www.ntsuisse.ch/ (02.06.2024).

A Stand Veröffentlichungen und Kostübersicht NTSPortal

A.1 Vorträge und Veröffentlichungen

Tabelle 4: Vorträge und Veröffentlichungen

Zitat	Typ
K. Jewell, F. Thron, M. Schlüsener, K. Kramer, T. Scharrenbach, I. Fettig, J. Koschorreck, C. Schulte, T. Ternes, and A. Wick, A Database Model for Aggregating Non-Target-Screening Data. Symposium „Monitoring Station of the Future“ Bundesanstalt für Gewässerkunde, 12.04.2021, Online/Koblenz	Vortrag
K. Jewell, F. Thron, M. Schlüsener, K. Kramer, T. Scharrenbach, I. Fettig, J. Koschorreck, C. Schulte, T. Ternes, and A. Wick, Ein Datenbankmodell für aggregierte Non-Target-Screening-Ergebnisse. Wasser 2021 - Jahrestagung der Wasserchemischen Gesellschaft 10.-12.05.2021, Online	Poster
K. Jewell, F. Thron, M. Schlüsener, K. Kramer, T. Scharrenbach, I. Fettig, J. Koschorreck, C. Schulte, T. Ternes, and A. Wick, Ein Datenbankmodell für aggregierte Non-Target-Screening-Ergebnisse. *Vom Wasser*, **2021**, 119(3), 94-96.	Kurzbeitrag
I. Fettig, J. Koschorreck, K. Jewell, A. Wick. Online database and analysis platform for NTS data for environmental monitoring. 6th International Conference on Environmental Specimen Banks, 29.09.21, Online/Incheon	Vortrag
I. Fettig, J. Koschorreck, K. Jewell, F. Thron, B. Ehlig, G. Dierkes, K. Kramer, J. Skottnik, T. Scharrenbach, T. Ternes, A. Wick. Online database and analysis platform for NTS data for chemical water monitoring. International Conference for Non-Target-Screening (ICNTS), 05.10.21, Online/Erding	Vortrag
K. Jewell, F. Thron, B. Ehlig, S. Quanz, M. Schlüsener, K. Kramer, T. Scharrenbach, I. Fettig, J. Koschorreck, T. Ternes, and A. Wick, Identifikation und räumliche Eingrenzung von Stoffeinträgen in Gewässern. *Sonderpublikation zum Langenauer Wasserforum.* **2021.** Zweckverband Landeswasserversorgung, Langenau	Kurzbeitrag und Vortrag
Jewell, K. S., Bader, T., Ondruch, P., Fettig, I., Koschorreck, J., Wick, A., Heinz, M.-D., Müller, U., Winzenbarcher, R., Ternes, T. A. **2022**. Die drei ??? der Non-Target-Analytik - ein Fall für die kollaborative Detektivarbeit Wasser 2022, Wasserchemische Gesellschaft der GDCh, Online.	Kurzbeitrag und Vortrag
Kevin Jewell*, Jan Koschorreck*, Franziska Thron, Björn Ehlig, Jonas Skottnik, Alexander Badry, Anna Lena Kronsbein, Thomas Scharrenbach, Arne Wick, Thomas Ternes **2023** Harmonizing Non-Target-Screening Results for Aggregated Analysis SETAC Europe 2023, SETAC, Dublin	Poster

A.2 Abschätzungen zu den laufenden Betriebskosten für das NTSPortal

NTSPortal wurde in die bereits bestehende IT-Infrastruktur der BfG aufgebaut. Diese beinhaltet Netzlaufwerke, Server für die Datenbearbeitung und vieles mehr. Die gesamte Infrastruktur kann hier nicht abgebildet werden. Die Kosten (Abbildung 23) sind deshalb nur eine Einschätzung und beziehen sich nur auf den Server für die ElasticSearch Datenbank. Die Lizenz für das ElasticSearch Platinum Modul wurde über einen Rahmenvertragspartner gekauft und die Kosten sind nicht öffentlich. Dieses Modul bietet Zusatzfunktionen und ist für den grundlegenden Betrieb von NTSPortal nicht erforderlich.

Abbildung 23: Kostenabschätzungen für den Aufbau einer ElasticSearch Datenbank

Kostenabschätzungen für den Aufbau einer ElasticSearch Datenbank

Posten	€/Jahr
Servercluster, 3 Knoten	4 000€ (Hardware und Support)
IT Administration (50% E7 TVöD)	43 000€

Quelle: eigene Abbildung BfG

B Agenda und Protokoll des Begleitkreis Kick-Off Treffens und Abschlussworkshops

B.1 Agenda des Begleitkreis Kick-Off Treffens am 03.12.2020

Abbildung 24: Agenda des Begleitkreis Kick-Off Treffens am 03.12.2020 - Seite1

Agenda des Begleitkreis Treffens am 03.12.2020 - Seite 1 (mit alten Logos)

Für Mensch und Umwelt

Stand: 1. Dezember 2020

Begleitkreis des REFOPLAN Forschungsvorhabens “Online-Portal Non-Target Screening für die Umweltüberwachung der Zukunft”

Kick-Off Veranstaltung

Ort: Online (WebEx, Siehe Einladungslink per E-Mail)

Zeit: 03.12.2020, 13:00-15:30 (12:30-13:00 Technik-/Verbindungstest)

Agenda

Beginn: 13:00	
Begrüßung, kurze Vorstellungsrunde und Einleitung – (UBA, Alle)	15 Min
Strukturiertes zusammenbringen von NTS-Daten: Zwei Herausforderungen – (BfG)	30 Min
NFDI4Chem - Digitalisierung der chemischen Forschung – Tobias Schulze, UFZ	15 Min + 10 Min Diskussion
Demonstration von ElasticSearch und Kibana – (BfG)	15 Min
Diskussion/Fragerunde (Unterpunkte auf Seite 2) – (Alle)	60 Min
Nächster Termin – (Alle)	5 Min
Voraussichtliches Ende: 15:30	

Moderation: Ina Fettig, UBA

Bundesanstalt für Gewässerkunde (BfG)
Referat G2 (Gewässerchemie)
Referat Z2 (Informationstechnik und -management)
Am Mainzer Tor 1
56068 Koblenz
www.bafg.de

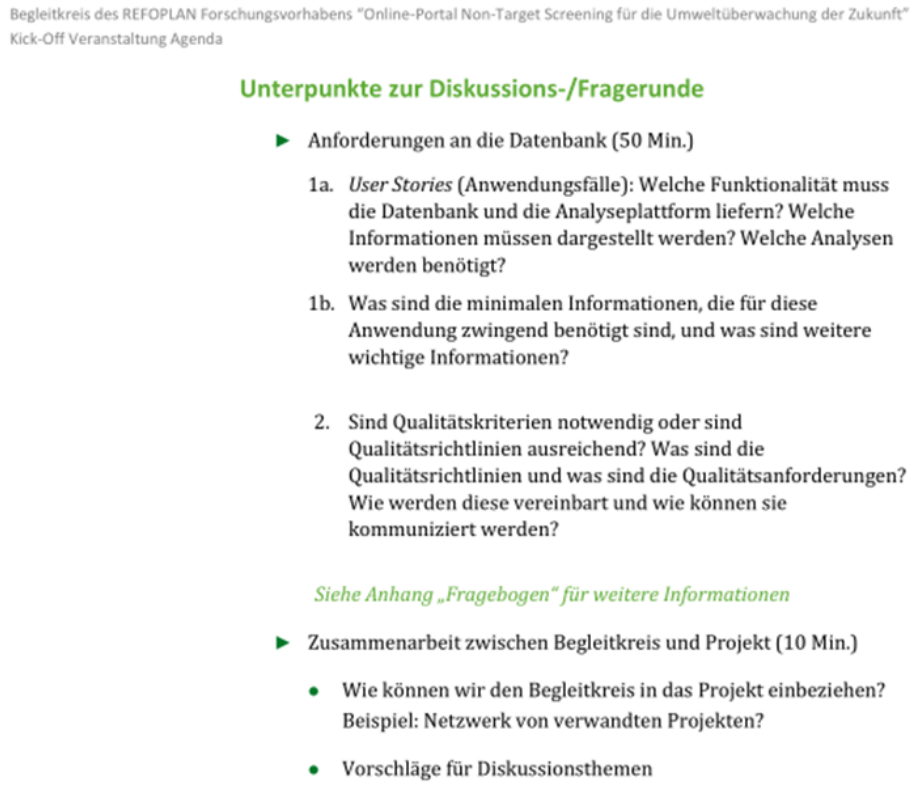
Umweltbundesamt
Fachgebiet II 2.4
Wörlitzer Platz 1
06844 Dessau-Roßlau
www.umweltbundesamt.de

Auftraggeber: UBA
Auftragnehmer: BfG
FKZ: 3720 22 0201 0
Laufzeit: 01.05.2020 für 3 Jahre

Quelle: eigene Abbildung UBA/BfG

Abbildung 25: Agenda des Begleitkreis Kick-Off Treffens am 03.12.2020 - Seite2

Agenda des Begleitkreis Treffens am 03.12.2020 - Seite 2 (mit alten Logos)


Begleitkreis des REFOPLAN Forschungsvorhabens "Online-Portal Non-Target Screening für die Umweltüberwachung der Zukunft"
Kick-Off Veranstaltung Agenda

Unterpunkte zur Diskussions-/Fragerunde

- Anforderungen an die Datenbank (50 Min.)

  1a. *User Stories* (Anwendungsfälle): Welche Funktionalität muss die Datenbank und die Analyseplattform liefern? Welche Informationen müssen dargestellt werden? Welche Analysen werden benötigt?

  1b. Was sind die minimalen Informationen, die für diese Anwendung zwingend benötigt sind, und was sind weitere wichtige Informationen?

  2. Sind Qualitätskriterien notwendig oder sind Qualitätsrichtlinien ausreichend? Was sind die Qualitätsrichtlinien und was sind die Qualitätsanforderungen? Wie werden diese vereinbart und wie können sie kommuniziert werden?

  Siehe Anhang „Fragebogen“ für weitere Informationen

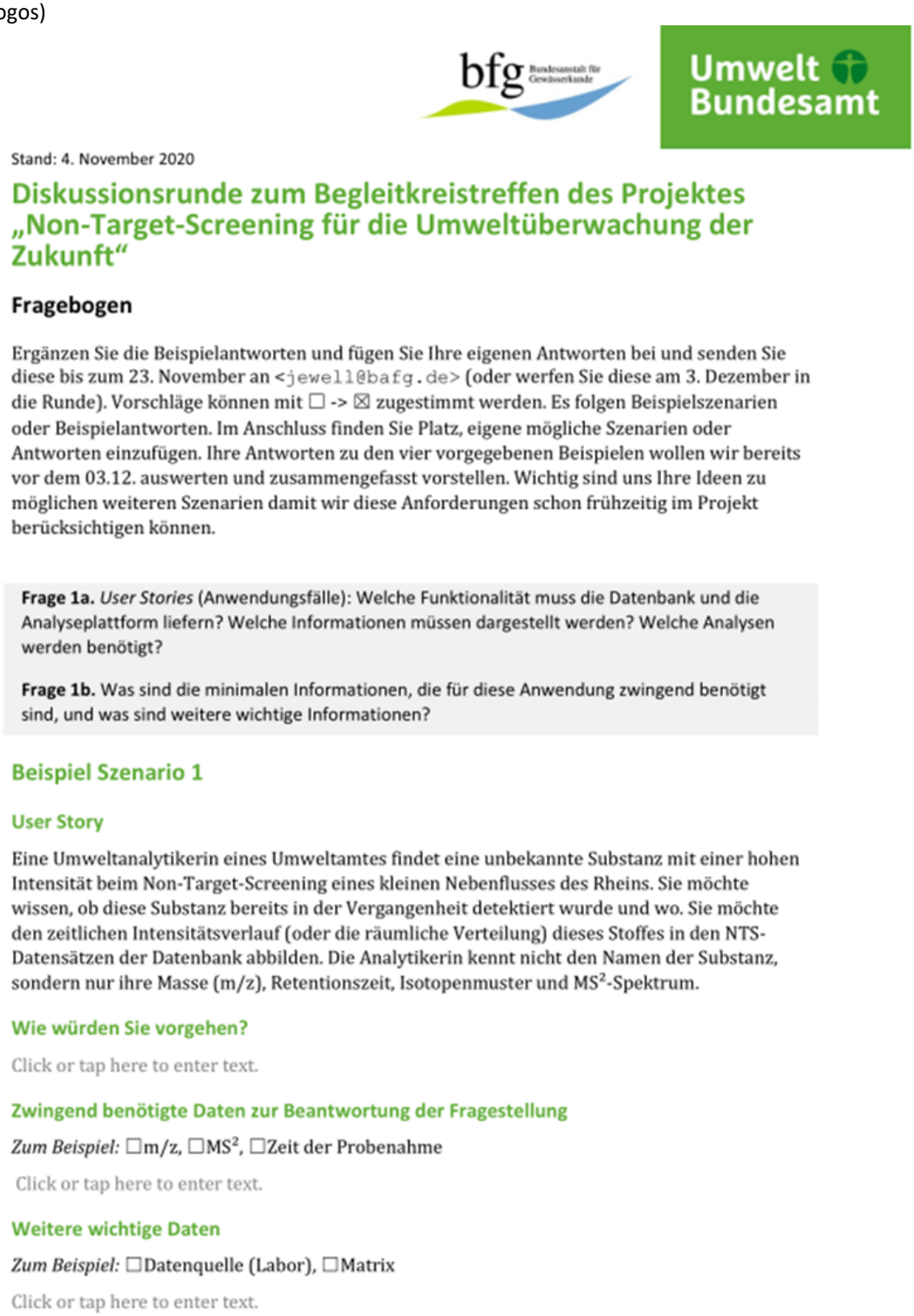
- Zusammenarbeit zwischen Begleitkreis und Projekt (10 Min.)
  - Wie können wir den Begleitkreis in das Projekt einbeziehen? Beispiel: Netzwerk von verwandten Projekten?
  - Vorschläge für Diskussionsthemen

Quelle: eigene Abbildung UBA/BfG

B.2 Fragebogen zur Vorbereitung des 1. Treffens

Abbildung 26: Fragebogen zur Vorbereitung des 1. Treffens – Seite 1

Zur Vorbereitung des 1. Treffens wurde vorab ein Fragebogen an die Teilnehmer versendet, welche beantwortet werden sollten. Die Antworten sollten an jewell@bafg.de gesendet werden. - Seite 1 (mit alten Logos)



Stand: 4. November 2020

Diskussionsrunde zum Begleitkreistreffen des Projektes „Non-Target-Screening für die Umweltüberwachung der Zukunft“

Fragebogen

Ergänzen Sie die Beispielantworten und fügen Sie Ihre eigenen Antworten bei und senden Sie diese bis zum 23. November an <jewell@bafg.de> (oder werfen Sie diese am 3. Dezember in die Runde). Vorschläge können mit ☐ -> ☒ zugestimmt werden. Es folgen Beispielszenarien oder Beispielantworten. Im Anschluss finden Sie Platz, eigene mögliche Szenarien oder Antworten einzufügen. Ihre Antworten zu den vier vorgegebenen Beispielen wollen wir bereits vor dem 03.12. auswerten und zusammengefasst vorstellen. Wichtig sind uns Ihre Ideen zu möglichen weiteren Szenarien damit wir diese Anforderungen schon frühzeitig im Projekt berücksichtigen können.

Frage 1a. *User Stories* (Anwendungsfälle): Welche Funktionalität muss die Datenbank und die Analyseplattform liefern? Welche Informationen müssen dargestellt werden? Welche Analysen werden benötigt?

Frage 1b. Was sind die minimalen Informationen, die für diese Anwendung zwingend benötigt sind, und was sind weitere wichtige Informationen?

Beispiel Szenario 1

User Story

Eine Umweltanalytikerin eines Umweltamtes findet eine unbekannte Substanz mit einer hohen Intensität beim Non-Target-Screening eines kleinen Nebenflusses des Rheins. Sie möchte wissen, ob diese Substanz bereits in der Vergangenheit detektiert wurde und wo. Sie möchte den zeitlichen Intensitätsverlauf (oder die räumliche Verteilung) dieses Stoffes in den NTS-Datensätzen der Datenbank abbilden. Die Analytikerin kennt nicht den Namen der Substanz, sondern nur ihre Masse (m/z), Retentionszeit, Isotopenmuster und MS^2-Spektrum.

Wie würden Sie vorgehen?

Click or tap here to enter text.

Zwingend benötigte Daten zur Beantwortung der Fragestellung

Zum Beispiel: ☐m/z, ☐MS^2, ☐Zeit der Probenahme

Click or tap here to enter text.

Weitere wichtige Daten

Zum Beispiel: ☐Datenquelle (Labor), ☐Matrix

Click or tap here to enter text.

Quelle: eigene Abbildung UBA/BfG

Abbildung 27: Fragebogen zur Vorbereitung des 1. Treffens – Seite 2

Zur Vorbereitung des 1. Treffens wurde vorab ein Fragebogen an die Teilnehmer versendet, welche beantwortet werden sollten. Die Antworten sollten an jewell@bafg.de gesendet werden. - Seite 2

Beispiel Szenario 2

User Story

Ein Mitarbeiter im Regierungspräsidium wurde über einen lokalen chemischen Industrieunfall unterrichtet. Er möchte wissen, ob durch diesen Unfall „neue“ Features (bisher unbekannte Substanzen) mittels NTS detektiert worden sind. Er möchte Features sehen, die spezifisch für eine Region und für ein bestimmtes Zeitintervall sind.

Wie würden Sie vorgehen?

Click or tap here to enter text.

Zwingend benötigte Daten zur Beantwortung der Fragestellung

Zum Beispiel: ☐Zeit der Probenahme, ☐Ort der Probenahme, ☐Intensität

Click or tap here to enter text.

Weitere wichtige Daten

Zum Beispiel: ☐m/z, ☐Retentionszeit, ☐EIC

Click or tap here to enter text.

Ihre Vorschläge zu weiteren möglichen Szenarien

User Story

Click or tap here to enter text.

Zwingend benötigte Daten zur Beantwortung der Fragestellung

Click or tap here to enter text.

Weitere wichtige Daten

Click or tap here to enter text.

Quelle: eigene Abbildung UBA/BfG

Abbildung 28: Fragebogen zur Vorbereitung des 1. Treffens – Seite 3

Zur Vorbereitung des 1. Treffens wurde vorab ein Fragebogen an die Teilnehmer versendet, welche beantwortet werden sollten. Die Antworten sollten an jewell@bafg.de gesendet werden. - Seite 3

Frage 2. Sind Qualitätskriterien notwendig oder sind Qualitätsrichtlinien ausreichend? Was sind die Qualitätsrichtlinien und was sind die Qualitätsanforderungen? Wie werden diese vereinbart und wie können sie kommuniziert werden?

Beispiel 1

Qualitätsgrenzen sind notwendig und sollten so streng wie möglich sein, ohne eine Mehrheit der Nutzer auszuschließen, d.h. wir müssen eine gemeinsame Vereinbarung über a) die Bewertung der Qualität und b) die Qualitätsgrenzen finden. *Zum Beispiel:* m/z-Fehler unter XX mDa, Intensitätsabweichung eines Standards kleiner als Faktor XX.

☐ Zustimmung

☐ Ablehnung –

Begründung: Click or tap here to enter text.

Beispiel 2

Qualitätsrichtlinien sind ausreichend. Die Daten werden unabhängig von der Qualität akzeptiert, und die Algorithmen können, dank der großen Datenmenge, mit Ausreißern umgehen. Wenn einige Daten von schlechter Qualität sind, wird dies bei der Bildung von Durchschnittswerten nicht sichtbar sein. Auf diese Weise bildet man die größtmögliche Datenbasis.

☐ Zustimmung

☐ Ablehnung –

Begründung: Click or tap here to enter text.

Ihre Vorschläge

Click or tap here to enter text.

Quelle: eigene Abbildung UBA/BfG

B.3 Protokoll des Belgleitkreis Kick-Off Treffens

Abbildung 29: Ergebnisprotokoll Begleitkreis Kick-Off Treffen – Seite 1

Protokoll des Belgleitkreis Kick-Off Treffens

Allgemeine Anmerkungen

- Im Rahmen der Diskussion wurde die Notwendigkeit einer detaillierten Qualitätssicherung vor der Eintragung der Daten in die Datenbank betont.
- Es werden im neuen Jahr Ansätze für die QS in einem kleineren Kreis erarbeitet und beim nächsten Treffen des Begleitkreises vorgestellt.

Ergebnisprotokoll

Vortrag: Einleitungsvortrag (Ina Fettig)

Erläuterungen zur Bewilligung des Projektes und zu der Trägerschaft auf der Ebene des UBA und des BMU.

Das Projekt ermöglicht die gemeinsame Nutzung von NTS Gewässerdaten für die Unterstützung der Umwelt- und Stoffgesetze und ein gewässerübergreifendes Frühwarnsystem und ist ein Vorzeigebeispiel für die digitale Transformation der Umweltbeobachtung.

Identifizierung von Schnittstellen z.B. zu Flussgebietsgemeinschaften und -kommissionen, Chemikaliengesetze, Digitalisierungsstrategie des Bundes, Green Deal etc.

Tabelle: Inhalt der Diskussion um „Einleitungsvortrag“

Das Projekt initiiert einen Begleitkreis (wovon hier berichtet wird) und eine Stakeholder-Gruppe, die Anfang 2021 Tagen wird	Stakeholder-Gruppe: Themen wie Implementierung und Datenrechte werden diskutiert, Fachwissen von Non-Target-Screening ist nicht vorausgesetzt. Organisiert über die LAWA mit VertreterInnen der Bundesländer sowie BAFU, UBA Wien Begleitkreis: Experten-Ebene, hier wird die fachliche Umsetzung diskutiert

Quelle: eigene Abbildung BfG

Abbildung 30: Ergebnisprotokoll Begleitkreis Kick-Off Treffen – Seite 2

Vortrag: Strukturiertes Zusammenbringen von NTS-Daten: Zwei Herausforderungen (Kevin Jewell)

Kurze Beschreibung der Projektziele, Schilderung der größten Herausforderungen und ein kurzer Bericht zum aktuellen Stand der Projektarbeiten

Tabelle: Inhalt der Diskussion um Vortrag „Strukturiertes Zusammenbringen von NTS-Daten: Zwei Herausforderungen“

Allgemein	Details
Alignment über Substanzname oder CAS	Substanz oder CAS nicht eindeutig genug - Teilweise Synonyme - Teilweise keine CAS-RN vorhanden Eigener Identifier somit zwingend notwendig - Es sollte direkt mit einer Universal Alignment ID (UAID) gestartet werden - Alle Infos (Synonyme etc.) unter der UAID zusammenfassen - Eventuell Erfahrungen aus Norman abfragen **Weiteres Vorgehen** Es wird noch in Phase 1 (annotierte Features) mit einem UAID System begonnen
Alignment der RT/RTI	Eine gemeinsame Liste an IS ist notwendig, um RTIs zu bestimmen. **Weiteres Vorgehen** Eine Abfrage innerhalb des BK zu möglichen gemeinsamen IS Kandidaten wird in 2021 durchgeführt
Vergleichbarkeit Peakflächen/Intensitäten	Software geben oft nur Fläche oder Intensität aus, nicht beides.

Quelle: eigene Abbildung BfG

Abbildung 31: Ergebnisprotokoll Begleitkreis Kick-Off Treffen – Seite 3

Vortrag: NFDI4Chem - Digitalisierung der chemischen Forschung (Tobias Schulze)

Vorstellung des Nationalen Forschungsprogramms NFDI4Chem

Tabelle: Inhalt der Diskussion um Vortrag „NFDI4Chem - Digitalisierung der chemischen Forschung“

Entwicklung des Electronic Labbooks	Es wird auf die Entwicklung vom KIT zurückgegriffen und diese wird weiterentwickelt

Demonstration: prototypisches Dashboard (Kevin Jewell)

Vorstellung der Suche und Visulisierung von Daten in der Datenbank über ein Dashboard (Erstellt mit Hilfe von Kibana)

Tabelle: Inhalt der Diskussion um Demonstration „prototypisches Dashboard“

Allgemein	Details
Bedienung und Funktionen der Oberfläche	Die Filterung nach Masse ist noch sehr umständlich. Die Annotation von Features nach der Eintragung in der Datenbank wird als weitere Funktion empfohlen. Daten können als csv exportiert werden. Ein entsprechendes Export-Tool ist schon in Kibana integriert. Anwendung von Boolean Operatoren: Mit KQL (Kibana Query Language) kann mit den Operatoren „and“ und „or“ Suchen kombiniert werden. In Bezug auf die Diskussion über die Qualitätskennzeichnung (Diskussionsrunde), wurde ein einfaches Datenfilterungssystem („Preset Filters“) empfohlen. **Weiteres Vorgehen** Es wird weiter an der Dashboard Oberfläche im Laufe des Projektes gearbeitet.
Datenformat für eingehende Daten	Die Daten sind text-basierte Dokumente in JSON Format mit Schlüssel-Wert Paaren und Tabellen für Chromatogramm, MS- (Isotopenmuster) und MS^2-Spektrum. Das Format kann jederzeit um gewünschte Parameter erweitert werden. Die JSON-Dateien werden an der BfG durch Harmonisierungsskripte erzeugt. Geliefert werden die Messdaten in irgendein offenes, maschinenlesbares Format – csv, txt, xlsx, mzXML etc. JSON-Dateien können selbstverständlich auf Nachfrage zurück an den Datenlieferanten gesendet werden.
Datenrechte	Freigaben und Nutzungsrechte müssen vorher gekennzeichnet werden. Hierzu wurde eine CC-BY

Quelle: eigene Abbildung BfG

Abbildung 32: Ergebnisprotokoll Begleitkreis Kick-Off Treffen – Seite 4

Allgemein	Details
	Lizenz vorgeschlagen. CC-BY-SA-NC wurde explizit nicht empfohlen. Dies ist auch ein Thema bei NFDI. **Weiteres Vorgehen** Angaben zu Datenrechten werden als weitere Parameter zu den Daten hinzugefügt (wie genau ist noch nicht klar) und werden individuell angepasst. Abstimmungen zu den Datenrechten werden im Laufe des Projektes und in Zusammenhang mit der Stakeholder-Gruppe stattfinden.
Qualitätssicherungsmerkmale in die Datenbank	QS-Parameter, wie z.B. Massengenauigkeit, werden noch nicht in der Datenbank mitgeführt. **Weiteres Vorgehen** Die Entwicklung der QS wird auch die Mitführung dieser Daten berücksichtigen.
Metainformationen	Jedes Dokument (Feature) enthält auch Metadaten. ElasticSearch optimiert die Ablage dieser Daten im Hintergrund. Die Kodierung der Messstellen (bspw. „rhein_ko_l“) wird in einer separaten Tabelle geführt.

Quelle: eigene Abbildung BfG

Abbildung 33: Ergebnisprotokoll Begleitkreis Kick-Off Treffen – Seite 5

Diskussionsrunde: User Stories und Qualitätssicherung

Auswertung des Fragebogens und Diskussion von weiteren Themen

Tabelle: Inhalt der Diskussion um Diskussionsrunde „User Stories und Qualitätssicherung“

Allgemein	Details
Plausibilisierung und die Minimierung von False Positives	Bei allen User Stories wurden die QS und Plausibilisierung der Ergebnisse als sehr wichtig erachtet. Vorschläge für die Plausibilisierung: - Erwartete Ionisierbarkeit/ chromatographische Trennung - Qualitätsdaten überprüfen, flaggen von nicht eindeutigen Ergebnissen (siehe oben) - Online Datenbanken für Isomere überprüfen - Voreingestellte Filter für sichere Daten (z.B. durch Qualitätsangaben, siehe unten) **Weiteres Vorgehen** Die Vorschläge werden bei der Entwicklung der QS-Strategie berücksichtigt.
Bestimmung „Kritischen Fold Change“ für User Story 2	Der Wert muss statistisch abgeleitet werden. Beispielsweise über die Standardabweichung des Vorlaufs (6-10 Fach). Der Begriff „Fold Change“ findet nicht überall Verwendung.
User Story 3 und 4: Keywords für Annotationen	Die Annotation von Features mit Hinweisen für die Identifizierung (bspw. Herkunft, bekannte funktionelle Gruppen usw.) wurde als sehr wichtig erachtet. Standardisierte Bezeichnungen (kontrolliertes Vokabular) wären jedoch dafür notwendig.
User Story 5: „Screening“ für mehrere Substanzen	Vorgeschlagen wurde, dass eine Abfrage für mehrere Substanzen mit einer Liste oder Tabelle (bspw. csv) erfolgen kann.
User Story 6: Matrizes-Übergreifende Vergleiche	Erfahrungen in NORMAN (bspw. NormaNEWS2) könnten hierfür hilfreich sein
Qualitätssicherung	Die Anwendung von Qualitätskennzeichnung mit strikten Kriterien wurde vorgeschlagen. Eine Filterfunktion im Dashboard könnte mit Hilfe dieser Kennzeichnung implementiert werden. Am besten als max. zwei-stufiges System. **Weiteres Vorgehen** Eine QS-Strategie wird in einer kleineren Gruppe erarbeitet und dem BK zur Abstimmung (auch in

Quelle: eigene Abbildung BfG

Abbildung 34: Ergebnisprotokoll Begleitkreis Kick-Off Treffen – Seite 6

Allgemein	Details
	Bezug auf die noch festzulegenden Toleranzen) vorgestellt.
Vergleiche über mehrere Methoden	Ein Vergleich oder Alignment über mehrere Methoden wird als sehr schwierig erachtet.
	Weiteres Vorgehen
	Methodenvergleiche werden im Verlauf des Projektes durchgeführt, um die Grenzen der Vergleichbarkeit zwischen Methoden zu bestimmen.

Weiteres Vorgehen

- QS Diskussionen in kleinerem Kreis (u.a. IKSR-Projekt, TrinkIDENT, Fachausschuss GDCh), Erarbeitung von Vorschlägen für nächstes Begleitkreistreffen.
- Stakeholder-Treffen Frühjahr 2021 (Ankündigung und Terminfindung Ende Januar 2021)
- 2. Begleitkreistreffen Mitte 2021

Protokoll: Nina Hermes, Kevin Jewell, Ina Fettig

Quelle: eigene Abbildung BfG

C Zusammenfassung des Abschlussworkshops, 14.-15.12.2023, Berlin

Abbildung 35: Zusammenfassung des Abschlussworkshops 14.-15.12.2023, Berlin – Seite 1

Zusammenfassung des Abschlussworkshops 14.-15.12.2023, Berlin (Seite 1)

Umwelt Bundesamt

Für Mensch und Umwelt

Stand: 1. März 2023

Workshop „Non-Target Screening im Gewässerschutz“

Zusammenfassung

14.-15.12.2023, Landesvertretung Sachsen-Anhalt, Berlin

50 Jahre Umweltbundesamt 1974–2024

1 Kleingruppen

1.1 Daten sammeln

- Standardisierung/Harmonisierung
  - Aktuelle Leitfäden und Harmonisierungsvorhaben (GDCh, DIN, NORMAN) sind bereits gute Ansätze für harmonisierte NTS Daten
  - Hinweise für Anforderungen und Maßnahmen der Qualitätssicherung für Vergaben an Dritte erforderlich
  - Ein einheitlicher interner Standardmix (aktuell in Arbeit beim GDCh Fachausschuss NTS), den die Labore möglichst übergreifend einsetzen, ist ein wichtiger Baustein zur Harmonisierung des NTS
  - Metadatenerfassung/-speicherung ist für die langfristige Nutzung und den Austausch der NTS Daten wichtig
  - Einheitliche Messmethode in großen Projekten (bspw. IKSR) ist ein pragmatischer Ansatz, um eine gute Vergleichbarkeit der Daten zu ermöglichen
  - Gesetzliche Grundlage für das Teilen von Mindestinformationen neuer Substanzen (Summenformel, MS^2 Spektren, Bereitstellung von Standards) bei bspw. REACH-Registrierung und Zulassungsverfahren für Pflanzenschutzmittel, Biozide und Arzneimitteln ist erforderlich
  - Zukunftsaufgabe: Datenbanken intelligent bauen und nach FAIR Prinzipien zusammenführen und sichtbar machen
- Daten teilen
  - Rechteverwaltung, Lizensierung, Nachverfolgung bei Datennutzung – reichen die FAIR Prinzipien für diesen Fall aus?
  - Automatisierung durch künstliche Intelligenz (KI)?

1

Quelle: eigene Abbildung UBA/BfG

Abbildung 36: Zusammenfassung des Abschlussworkshops 14.-15.12.2023, Berlin – Seite 2

Zusammenfassung des Abschlussworkshops 14.-15.12.2023, Berlin (Seite 2)

1.2 Daten verknüpfen

- Big Data/KI
  - Große Hoffnung für NTS, aufgrund des großen Datenvolumens und Komplexität der Auswertung
  - Zwar gibt es im NTS eine große Datenmenge (Terabytebereich), aber die Datenstruktur (Umweltmonitoring entspricht keinem konventionellem Experimentdesign (Treatment/Kontrolle)) und die Anzahl von Datensätzen ist eine Herausforderung für KI basierte Anwendungen, was bspw. (unüberwachtes) maschinelles Lernen erschwert
  - Vision: automatisierte Datenbankabfragen (bspw. Toxizitätsdaten → Mischungsbewertung), ‚Bewertung auf Knopfdruck‘
  - Kann in Kombination mit NTS ein Echtzeit-Uumweltmonitoring ermöglichen, bspw. durch automatisierte/zeitnahe Mustererkennung anthropogener Einträge
- Daten verknüpfen
  - Aktuell ist es ist es noch einfacher Target Daten (Konzentrationen) mit bestehenden Non-Target Datensätzen zu kombinieren, andersherum ist es schwieriger
    - NTS Datenschema ist allgemeingültiger und kann als Standard für sowohl Target als auch NTS dienen
    - Durch eine Kalibrierung und Spiken der Targetstandards können aus HRMS NTS Messungen Targetdaten generiert werden
    - Vorteile: Datensätze können direkt verknüpft werden (NTS + Target), retrospektive Auswertung der NTS Datensätze möglich
    - Nachteile: Höherer Aufwand als „nur“ eine Targetmessung durchzuführen → geringerer Durchsatz, derzeit noch weniger sensibel
- NTS bietet die Chance Forschungsfelder zu verknüpfen
  - Bspw. Biologie und Chemie: Mit derselben Probe NTS plus wirkungsbezogene Analytik (WBA), eDNA, Metabol-/Transkriptomik etc. durchführen → umfassendere Qualitätsaussage zur Probe möglich
  - voneinander lernen: Metabolomik und NTS sehr ähnliche Herausforderungen

Quelle: eigene Abbildung UBA/BfG

Abbildung 37: Zusammenfassung des Abschlussworkshops 14.-15.12.2023, Berlin – Seite 3

Zusammenfassung des Abschlussworkshops 14.-15.12.2023, Berlin (Seite 3)

1.3 Daten bewerten

- Aktuelle Anwendungsfelder des NTS in der Umweltbeobachtung
  - Zustandsbeschreibung und -verbesserung
    - Prozessüberwachung/Erfolgskontrolle, bspw. Trinkwasseraufbereitung, 4. Reinigungsstufe
    - Einleitungsquellen bestimmen
    - Substitutionseffekte erfassen
    - Räumlich/zeitliche Trends ableiten
    - Fingerprinting typischer Stoffmuster von Proben/Standorten
    - Komplexität der Belastung (Mischungen) erfassen
  - NTS Daten als Entscheidungskriterium für weitere Maßnahmen
    - Stoffpriorisierung: Veranlassung Target Screening, wenn Konzentrationen erforderlich
    - Priorisierung der Messstellen (bspw. Flussabschnitt) für Umweltbeobachtungsprogramme
    - Early warning Systeme (EWS) für Umwelt-, Emissions- und Stoffgesetze
    - Erfassung unbekannter Stoffe/Transformationsprodukte
  - Rechtlicher Kontext:
    - Nicht immer sind quantitative (Target) Daten notwendig
      - Genaue Abwägung, wann qualitativer Nachweis (Suspect Screening) ausreichend ist
    - REACH:
      - Verwendung von NTS Daten für Umweltmedien und Biota bei der Bewertung von Persistenz und Bioakkumulation
      - NTS-Daten können als Unterstützung in einem Weight-of-Evidence-Ansatz verwendet werden
      - Erkennen von Trends
      - Informationen zur Exposition
    - Wasserrahmenrichtlinie: Vision, möglichst alle und nicht nur die prioritären und flussgebietsspezifischen Substanzen für den chemischen Zustand eines Gewässers erfassen→ NTS kann einen Teil der Lücke schließen

Quelle: eigene Abbildung UBA/BfG

Abbildung 38: Zusammenfassung des Abschlussworkshops 14.-15.12.2023, Berlin – Seite 4

Zusammenfassung des Abschlussworkshops 14.-15.12.2023, Berlin (Seite 4)

- Mischungscharakterisierung
  - MS^2 Fragmente mit Toxikophoren verknüpfen (bspw. funktionelle Gruppen mit bekannter Wirkung)
- Semi-quantitative Daten
  - Quantifizierungsmethoden für NTS Daten, die für risikobasierte Bewertungen erforderlich sind, werden aktuell entwickelt
  - Aktuell: Target Screening zwingend für Risikobewertung nötig
  - Können Suspect Screening Daten in der Umweltbewertung als Nachweis gelten?
  - Rechtssicherheit semi-quantitativer NTS Daten unklar

2 Podiumsdiskussionen

2.1 Standardisieren

- Entscheidend ist die Vergleichbarkeit der Bewertungsergebnisse, nicht alle Einzelheiten der Methode
  - Konventionelle Target Analytik als Basis verwenden
  - Eine Stabilität und Vergleichbarkeit basierend auf nur wenigen Parametern ist wünschenswert
- Ausschreibungen: es werden Kriterien benötigt, um sinnvoll ausschreiben zu können, Angebote zu unterscheiden bzw. eine Auswahl zu begründen
  - BP4NTA Study Reporting Tool kann ein Anhaltspunkt sein
- Entwicklung einer DIN/ISO Norm
  - Es wird keine strenge IS Liste in der Norm geben, die mitgemessen werden müssen, sondern Kriterien für die Substanzauswahl
  - Vergleichsuntersuchungen werden durchgeführt (in Planung)

2.2 Zusammenarbeit

- Nationale Datenbanken sind wichtig, langfristig benötigen wir eine europäische Lösung
- Herausforderungen:
  - Daten liegen oft dort, wo sie erhoben wurden
  - Bei Kooperationsprojekten bestimmt der konservativste Partner die Daten-Policy für das Projekt

Quelle: eigene Abbildung UBA/BfG

Abbildung 39: Zusammenfassung des Abschlussworkshops 14.-15.12.2023, Berlin – Seite 5

Zusammenfassung des Abschlussworkshops 14.-15.12.2023, Berlin (Seite 5)

- LAWA
  - NTS auf Landesebene bisher kaum thematisiert
  - Zentrale Datenhaltung auch bei übrigen Themen entscheidend und viel diskutiert
- EU/DG Environment
  - Bisher noch keine EU Initiative
  - Prozess beschleunigen und Impulse geben, bspw. durch einen EU NTS Workshop in Brüssel

Impressum

Herausgeber

Umweltbundesamt
Wörlitzer Platz 1
06844 Dessau-Roßlau
Tel: +49 340-2103-0
buergerservice@uba.de
Internet:
www.umweltbundesamt.de
/umweltbundesamt.de
/umweltbundesamt

Stand: März/2024

Autorenschaft, Institution

Anna Lena Kronsbein, Umweltbundesamt

Kevin Jewell, Bundesanstalt für Gewässerkunde

Jan Koschorreck, Umweltbundesamt

Quelle: eigene Abbildung UBA/BfG

D Beiträge für Konferenzen

D.1 Beitrag für „Wasser 2021“ eingereicht am 01.12.2020

Abbildung 40: Beitrag für „Wasser 2021“ eingereicht am 01.12.2020 – Seite 1

Beitrag für „Wasser 2021“ eingereicht am 01.12.2020 (Seite 1)

Ein Datenbankmodell für aggregierte Non-Target-Screening Ergebnisse

K. Jewell, Koblenz/D, F. Thron, Koblenz/D, M. Schlüsener, Koblenz/D, K. Kramer, Koblenz/D, T. Scharrenbach, Koblenz/D, I. Fettig, Berlin/D, Jan Koschorreck, Berlin/D, C. Schulte, Dessau-Roßlau/D, T. Ternes, Koblenz/D, A. Wick, Koblenz/D

Dr. Kevin S. Jewell, Bundesanstalt für Gewässerkunde, Am Mainzer Tor 1, Koblenz/D

Einleitung

Das Non-Target-Screening (NTS) bietet große Chancen für die chemische Gewässerüberwachung, vor allem zum Screening oder zur Priorisierung von umwelt- und gesundheitsrelevanten Chemikalien, Auffindung bisher nicht bekannter Umweltkontaminanten und der Zuordnung von Eintragsquellen [1]. Des Weiteren dient es in der gewässerchemischen Forschung zur Untersuchung von Transformations- oder Verwitterungsprozessen [2] oder auch zur Untersuchung von saisonalen Veränderungen [3].

Die Durchführung von NTS-Analysen wird bereits in einigen Forschungseinrichtungen, Instituten und Landeslaboren zu diesen Zwecken angewendet. Für einen übergreifenden Gewässerschutz ist es aber zu vermeiden, dass die Datensätze verschiedener Messprogramme voneinander isoliert bleiben. Bei den Messungen werden jeweils große Datenmengen produziert, oft wird aber nur „die Spitze des Eisbergs“ untersucht, und die Datenmengen erreichen schnell unhandliche Größen. Zudem kommt hinzu, dass NTS-Daten, dank der unspezifischen Messung, oft für mehrere Untersuchungsziele verwendet werden können. Es besteht der Bedarf an einem digitalen Portal um NTS-Gewässerdaten zu aggregieren, damit diese für weitere Untersuchungen im selben Labor und auch für analytische und behördliche Forschungspartner bereitstehen. Die Informationen können somit gemeinsam genutzt werden, um Ressourcen zu bündeln, Synergien zu schaffen und eine bessere und größere Datengrundlage für die Forschung und für die Chemikalienbewertung zu generieren. Nur so wird es möglich sein, stoffliche Belastungen gewässerübergreifend zu identifizieren und den Gewässerschutz in Deutschland nachhaltig zu unterstützen. Erfahrungen, wie NTS Daten von verschiedenen Laboren für eine gemeinsame Auswertung aggregiert werden können, fehlen jedoch weitestgehend.

Das UBA-REFOPLAN-Projekt „Online-Portal: Non-Target Screening für die Umweltüberwachung der Zukunft“ wurde im Mai 2020, mit dem Ziel eine neuartige Bund-Länder-Datenbank- und Analyseplattform für NTS-Daten von Flüssen zu entwickeln, begonnen. Eine Datenbank, die in der Lage ist, NTS Ergebnisse aus verschiedenen Laboren zu kombinieren und eine aggregierte Re-Analyse der Daten zu ermöglichen. Dieses Vorhaben lässt sich in drei Zwischenziele unterteilen (Abb. 1):

- Strategien zur Harmonisierung von NTS Ergebnissen
- Implementierung einer zentralen Datenablage
- Bereitstellung von prototypischen Datenrecherche- und Analysetools

Quelle: eigene Abbildung BfG

Abbildung 41: Beitrag für „Wasser 2021“ eingereicht am 01.12.2020 – Seite 2

Beitrag für „Wasser 2021“ eingereicht am 01.12.2020 (Seite 2)

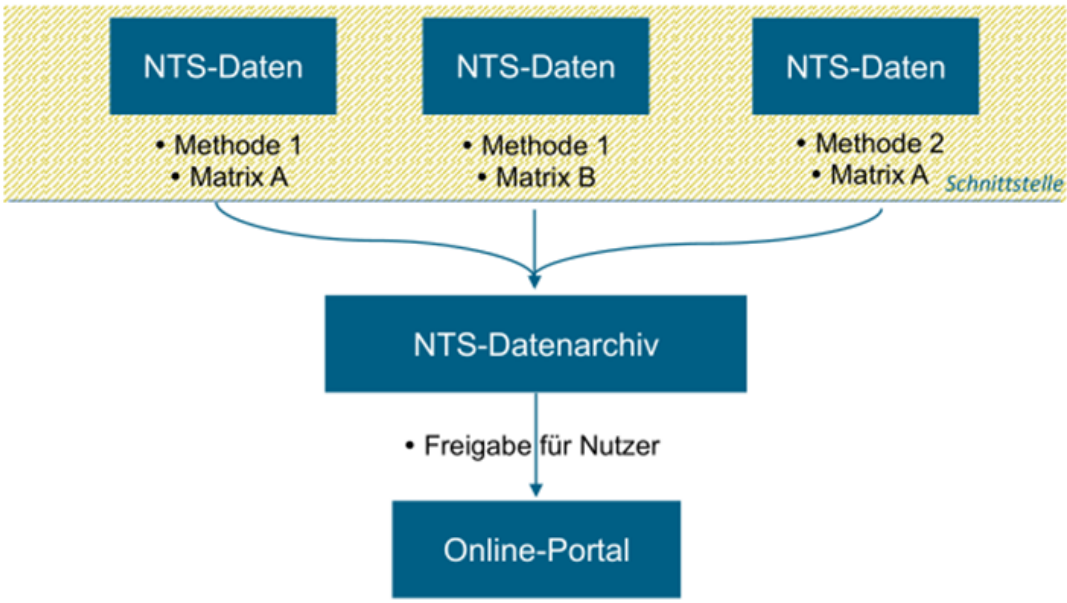


Abbildung 1 Konzeptdarstellung des NTS-Datenarchivs

Die erste Aufgabe zur Erreichung dieser Ziele besteht darin, ein einheitliches Datenmodell als Grundlage für die zukünftigen Harmonisierungsstrategien zu entwickeln. Hier wird die Entwicklung dieses Modells und die Implementierung eines Testdatensatzes vorgestellt.

Analytische Methoden

Proben von täglichen 24 h Mischproben (Wasser) und jährlichen Mischproben von Schwebstoffen (Umweltprobenbank [4]) von der Messstelle Koblenz/Rhein (BfG) wurden mit einer LC-QToF-MS/MS Methode gemessen [5] und mit einem Workflow in R ausgewertet [6]. Dabei wurde eine Spektrendatenbank mit 987 für Gewässer typische organische Spurenstoffe für die automatisierte Annotierung der Daten angewendet. Die Ergebnisse wurden anschließend in das Datenarchiv geladen.

Datenbankmodell

Unter Berücksichtigung der Bedürfnisse der Datenbank wurde ElasticSearch als Grundlage für die Implementierung gewählt. ElasticSearch ist eine verteilte Open-Source-Dokumentspeicher mit integrierter Suchmaschine und fällt unter die Kategorie „NoSQL“-Datenbanken. Vorteile von ElasticSearch sind unter anderem: a) Als verteilte Datenbank werden die Daten auf getrennten Knoten eines Rechenclusters gespeichert. Damit kann sie große Datenmengen ohne Einbußen in der Funktionalität speichern und aufrufen. b) ElasticSearch hat eine umfangreich dokumentierte RESTful-API, eine

Quelle: eigene Abbildung BfG

Abbildung 42: Beitrag für „Wasser 2021“ eingereicht am 01.12.2020 – Seite 3

Beitrag für „Wasser 2021“ eingereicht am 01.12.2020 (Seite 3)

standardisierte Schnittstelle für webbasierte Anwendungen. Diese kann verwendet werden, um die Datenbank mit maßgeschneiderter Analyse- oder Visualisierungssoftware zu verbinden. c) Die Basisversion von ElasticSearch ist quelloffen und kann kostenlos auch auf Desktop-PCs installiert werden.

Die Einträge in ElasticSearch basieren auf Dokumenten im JSON-Standardformat, wobei ein Dokument ein Non-Target Feature mit folgender Struktur ist:

```
POST /nts_archiv/_doc
{
  "mz": 237.1025,                          # Masse-Ladungs-Verhältnis vom Ion (mDa)
  "rt": 9.0,                               # Chromat. Retentionszeit in Minuten
  "date": "2019-03-01",                    # Probenahmedatum
  "duration": 1,                           # Probenahmedauer für Mischproben (Tage)
  "station": "elbe_tan_l",                 # Codierung Messstelle
  "matrix": "Wasser",                      # Matrix
  "hplc_method": "10.1016/j.chroma.2015.11.014",    # Beschreibung Methode (DOI)
  "pol": "pos",                            # Polarität
  "date_import": 1604071478,               # Zeit Datenimport (epoch s)
  "data_source": "BfG",                    # Daten-Quelle, Institut, Labor
  "name": "Carbamazepine",                 # Substanzname (falls bekannt)
  "cas": "298-46-4",                       # CAS-RN
  "area": 321.5,                           # Chromat. Peakfläche (Gerätabhängig)
  "area_is": 2290,                         # Chromat. Peakfläche des intern. Standards
  "norm_a": 0.1404,                        # Normalisierte Peakfläche (area / area_is)
  "loc": {"lat": 52.549876,                # Koordinaten Probenahmestelle
          "lon": 11.983321},
  "comment": "isomers found",              # Kommentarfeld (freier Text)
  "eic": [                                 # Chromatogramm (Zeit in s, Intensität)
    {"time": 301, "int": 10},
    {"time": 302, "int": 11},
    {"time": 303, "int": 8}
  ],
  "ms1": [                                 # Massenspektrum
    {"mz": 237.1025, "int": 1.0},
    {"mz": 238.1025, "int": 0.1}
  ],
  "ms2": [                                 # Fragmentspektrum
    {"mz": 194.1212, "int": 1.0},
    {"mz": 192.1345, "int": 0.8}
  ]
}
```

Es wurde ein pre-deployment ElasticSearch *Index* (Datenbank) erstellt, der ca. 150.000 Dokumente mit annotierten (mit chemischer Bezeichnung) Non-Target Features enthält. Dazu gehören Features von täglichen, 24 h Mischwasserproben aus der Rhein-Messstation der BfG in Koblenz für das Jahr 2017 bis 2020 und jährliche Schwebstoffmischproben vom gleichen Standort aus den Jahren 2006 bis 2018.

Ein Entwurf eines Visualisierungs-Dashboards wurde mit Hilfe von Kibana als Analyseplattform entwickelt, um Zeitreihen der mit IS normierten Peakflächen für ausgewählte Verbindungen darzustellen (Abb. 2). Die Suche kann unter Verwendung der Kibana Query Language (KQL) erfolgen. Beispielsweise wird die Suche nach einem

Quelle: eigene Abbildung BfG

Abbildung 43: Beitrag für „Wasser 2021“ eingereicht am 01.12.2020 – Seite 4

Beitrag für „Wasser 2021“ eingereicht am 01.12.2020 (Seite 4)

Feature anhand des m/z des Molekül-Ions und des m/z eines Fragment-Ions folgendermaßen eingegeben:

```
KQL: mz > 242.2 and mz < 242.3 and ms2: { mz > 142.1 and mz < 142.2 }
```

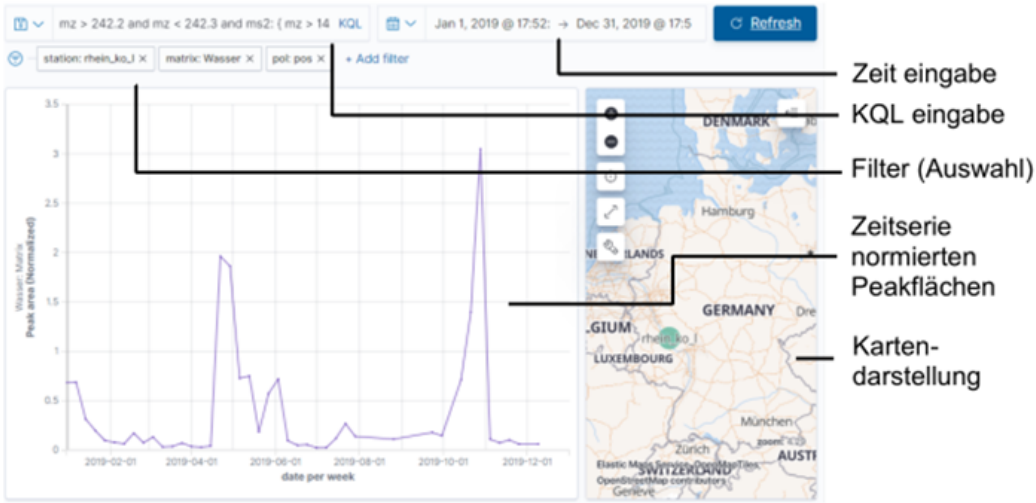


Abbildung 2 Screenshot des Dashboards (erstellt mit Kibana) zur Visualisierung von Peakflächen-Zeitserien. Suche nach einer Substanz mit der Masse 242.2842 und MS² Fragmentmasse 142.1597 (Tetrabutylammonium).

Fazit

Durch die Anwendung der vorgestellten Datenstruktur und mit Hilfe der Analyseplattform Kibana können NTS-Daten aggregiert und anhand von Substanznamen oder auch m/z, Retentionszeit und MS²-m/z durchsucht und analysiert werden. Dies bietet eine mögliche Grundlage für die zukünftigen Harmonisierungsstrategien, womit Daten aus unterschiedlichen Quellen integrativ ausgewertet werden können.

Ausblick

Im nächsten Schritt wird die Harmonisierung von NTS-Daten aus weiteren Auswertemethoden und in Zusammenarbeit mit anderen Umweltlaboren entwickelt. Das Projekt sieht vor, diese Ergebnisse in die Datenbank zu integrieren und für beteiligte Institute freizuschalten.

Danksagung

Finanzierung: UBA-REFOPLAN Projekt „Online-Portal: Non-Target Screening für die Umweltüberwachung der Zukunft“ FKZ: 3720 22 201 0.

Literatur

[1] J. Hollender, B. van Bavel, V. Dulio, E. Farmen, K. Furtmann, J. Koschorreck, U. Kunkel, M. Krauss, J. Munthe, and M. Schlabach, *Environ. Sci. Eur.*, **2019**, 31(1), 42.

Quelle: eigene Abbildung BfG

Abbildung 44: Beitrag für „Wasser 2021“ eingereicht am 01.12.2020 – Seite 5

Beitrag für „Wasser 2021“ eingereicht am 01.12.2020 (Seite 5)

[2] S. Brand, L. Veith, R. Baier, C. Dietrich, M. J. Schmid, and T. A. Ternes, *J. Hazard. Mater.*, **2020**, 395, 122289.
[3] C. M. G. Carpenter, L. Y. J. Wong, C. A. Johnson, and D. E. Helbling, *Environ. Sci. Techn.*, **2019**, 53(1), 77-87.
[4] Umweltbundesamt. *Umweltprobenbank des Bundes*. **2020** [Abgerufen: 09.09.2020]; von: www.umweltprobenbank.de.
[5] G. Nürenberg, M. Schulz, U. Kunkel, and T. A. Ternes, *J. Chromatogr. A.*, **2015**, 1426, 77-90.
[6] K. S. Jewell, U. Kunkel, B. Ehlig, F. Thron, M. Schlüsener, C. Dietrich, A. Wick, and T. A. Ternes, *Rapid Commun. Mass Sp.*, **2019**, 34, e8541.

Quelle: eigene Abbildung BfG

D.2 Beitrag für „Wasser 2022“ eingereicht am 27.04.2022

Abbildung 45: Beitrag für „Wasser 2022“ eingereicht am 27.04.2022 – Seite 1

Beitrag für „Wasser 2022“ eingereicht am 27.04.2022 (Seite 1)

Die drei ??? der Non-Target-Analytik - ein Fall für die kollaborative Detektivarbeit

K. Jewell, Koblenz/D, T. Bader, Langenau/D, P. Ondruch, Koblenz/D, I. Fettig, Berlin/D, J. Koschorreck, Berlin/D, A. Wick, Koblenz/D, M. D. Heinz, Koblenz/D, U. Müller, Karlsruhe/D, R. Winzenbacher, Langenau/D, T. Ternes, Koblenz/D.

Einleitung

Anthropogene Spurenstoffe werden über Punktquellen oder diffuse Einleitungen in Oberflächengewässer eingetragen. Bisherige Überwachungskonzepte basieren auf Analysen (sog. Target-Analysen) von bestimmten bekannten Substanzen. Diese erfassen jedoch nur einen Bruchteil der vorkommenden Spurenstoffe. Selbst bei diesen bekannten Stoffen lassen sich aber Eintragsorte häufig nicht oder nur mit großem Zeitversatz ermitteln, weshalb Emittenten oft unentdeckt bleiben. Unzulänglichkeiten in den Meldeketten forcieren die zeitliche Verzögerung zusätzlich.

Die genannten Nachteile können durch ein laborübergreifendes Non-Target-Screening (NTS) ausgeglichen werden. NTS bietet große Chancen für die Gewässerüberwachung [1], vor allem zur Priorisierung von umwelt- und gesundheitsrelevanten Chemikalien. Bei der Auffindung bisher nicht bekannter Kontaminanten im Spurenbereich ist NTS als alternativlos zu sehen.

NTS wird von einigen Laboren in der Gewässer- und Trinkwasserüberwachung bereits erfolgreich eingesetzt. Trotz umfangreicher Stoff- und Spektrenbibliotheken bleiben viele individuelle Identifizierungsversuche dabei aber noch erfolglos. Durch das Zusammenführen der NTS-Daten verschiedener Labore – also durch „kollaborative Detektivarbeit“ – können Emittenten eingegrenzt oder sogar ermittelt werden. Dies erhöht die Wahrscheinlichkeit einer Identifizierung enorm. Und selbst ohne chemische Identifizierung ließen sich die Emissionsquellen unbekannter Stoffe so schneller ermitteln und womit Maßnahmen zur Reduzierung der Emission möglich werden. Von der EU werden Maßnahmen an der Eintragsquelle (Verursacherprinzip) als nachhaltigste, wirksamste und kostengünstigste Variante zur Vermeidung chemischer Verschmutzungen beschrieben [2].

Das Zusammenführen und Auswerten der zeitlichen und räumlichen Daten von mehreren Laboren, welche mit unterschiedlichen NTS-Geräten und -Methoden analysieren, stellt gegenwärtig eine große Herausforderung dar. Folgende Punkte sind hierbei primär zur berücksichtigen:

- Entwicklung und Anwendung von Ansätzen zur Erkennung relevanter Befunde (z.B. Anomalie-Erkennung, Differenzierung anthropogener Stoffe von natürlichem Hintergrund)
- Bedarf an großen Datenmengen zum Trainieren von Algorithmen (maschinelles Lernen)
- Entwicklung leistungsstarker Datenbanksysteme, welche für das Management und schnelle Durchsuchen großer Datenmengen geeignet sind
- Entwicklung von Algorithmen zum „Verschneiden“ der Datensätze verschiedener Labore
- Hoher Entwicklungsaufwand für die Programmierung von APIs bzw. graphischer Oberflächen zur Interaktion und zur Visualisierung der Daten und Ergebnissen

Um die genannten Herausforderungen anzugehen, laufen aktuell drei Forschungsprojekte:

1. EU-Rhine Project Non-Target-Screening (IKSR, AUE, LUBW, BfG, LANUV, RWS)
2. UBA-REFOPLAN Projekt „NTSPortal“ (BfG, UBA)
3. BMBF-Projekt K2I (TZW, LW, LRZ, TUM)

Im Folgenden werden die drei Forschungsprojekte kurz vorgestellt sowie mögliche zukünftige Kollaborationen beschrieben.

Quelle: eigene Abbildung BfG

Abbildung 46: Beitrag für „Wasser 2022“ eingereicht am 27.04.2022 – Seite 2

Beitrag für „Wasser 2022“ eingereicht am 27.04.2022 (Seite 2)

1. EU-Rhine Project Non-Target-Screening

   Das Projekt „Monitoring the pollution of water through non-target screening on the Rhine“ ist der Verbesserung der Analyse und Erhöhung der Vergleichbarkeit von NTS Daten gewidmet, die an internationalen Messstationen entlang des Rheins erhoben werden.

   Die anthropogenen Einflüsse sowie die nötigen ökologischen Funktionen des Rheins und die Trinkwasserbereitstellung stellen ein Spannungsfeld dar. Deswegen ist der Schutz des Rheins im Fokus der nationalen Institutionen, des internationalen Übereinkommens [3] und der IKSR. Vor diesem Hintergrund ist das Ziel des Projektes die Entwicklung und Anwendung eines Tools für a) eine automatisierte Erkennung von Emissionen, b) eine schnelle Klassifizierung und Identifizierung von Schadstoffquellen und c) eine effizientere Identifizierung bisher unerkannter Schadstoffe. Dieses Tool basiert auf einem täglichen Wassermonitoring mittels Flüssigkeitschromatographie gekoppelt mit hochauflösender Massenspektrometrie (LC-HRMS) und der nachfolgenden Auswertung („Data-Mining“) der NTS Daten. Eine Harmonisierung der LC-HRMS-Methoden ermöglicht eine hohe Vergleichbarkeit der Daten, die von mehreren Stationen und unterschiedlichen Instrumenten (Orbitrap- und ToF-HRMS) generiert werden. Das dazu benötigte Software-Tool für die Datenprozessierung wird zusammen mit Umweltschutzbehörden der Rheinanliegerstaaten, der IKSR und dem Privatsektor entwickelt und vom EU-Programm “LIFE“ gefördert. Eine erfolgreiche Anwendung des Tools wird zu einer enormen Verbesserung in der Überwachung, ergänzend zum Internationalen Warn- und Alarmplan Rhein – IWAP, führen. Die Vision des Projektes ist die Durchführung einer „Echtzeit“-Flussüberwachung, transnational über verschiedene Länder in Europa hinweg.

1. UBA-REFOPLAN Projekt „NTSPortal“

   Das Forschungsvorhaben ‘Online-Portal „Non-Target Screening für die Umweltüberwachung der Zukunft“‘ soll Grundlagen schaffen, um NTS-Daten zusammenzuführen und Anwendungen für die Recherche und das Management von NTS-Daten im Gewässerschutz bereitzustellen. NTS-Daten können dann aus unterschiedlichen Regionen vergleichend und rückblickend betrachtet und für ein laborübergreifendes Data-Mining genutzt werden. Im Vordergrund stehen die überregionalen und retrospektiven Analysen sowie die Verschneidung von Messdaten aus unterschiedlichen Matrizes. Für die NTS-Datensammlung wurde eine Datenbank mit Hilfe der ElasticSearch-Plattform aufgebaut [4]. Die Datenbank enthält bereits mehr als 1,8 Millionen Einträge von über 100 Messstellen verteilt über Deutschlands Haupt- und Nebengewässern, und zwar sowohl für Oberflächenwasser- als auch für Schwebstoffe. Die Schwebstoffmessdaten stammen von Proben der Umweltprobenbank des Bundes für 13 Messstellen an Bundeswasserstraßen wie Rhein und Elbe. Sie reichen bis ins Jahr 2005 zurück [5].

   Eine web-basierte Nutzeroberfläche erlaubt die interaktive Echtzeitsuche und die graphische Darstellung von Ergebnissen. Zudem enthält sie bereits einige Funktionen der statistischen Analyse, z.B. der Trendanalyse. Die Funktionsweise der Plattform ist unabhängig von der angewendeten NTS-Auswertesoftware (Vorprozessierung). Bisher wurden drei Workflows für die Vorprozessierung als Datenquelle getestet. Zurzeit wird an der Harmonisierung und Umformatierung von weiteren Formaten der Ergebnisdaten und am Ausbau der web-basierten Benutzeroberfläche zur Recherche, Auswertung und zum Export der Daten gearbeitet.

2. BMBF-Projekt K2I

   Im K2I-Verbundprojekt „Künstliche und Kollektive Intelligenz zum Spurenstoff-Tracking in Oberflächenwasser für eine nachhaltige Trinkwassergewinnung“ soll ein Demonstrator eines cloudbasierten Systems für NTS-Analysen konzipiert und implementiert werden (k2i-tracker.de). Im Zusammenwirken der Cloudlösung mit örtlich verteilen LC-HRMS-Laboren sollen die Quellen bekannter und unbekannter Stoffe mittels künstlicher Intelligenz (KI) rasch eingegrenzt werden („Spurenstoff-Tracker“). Durch die vernetzte Auswertung der Labordaten in Kombination mit KI-

Quelle: eigene Abbildung BfG

Abbildung 47: Beitrag für „Wasser 2022“ eingereicht am 27.04.2022 – Seite 3

Beitrag für „Wasser 2022“ eingereicht am 27.04.2022 (Seite 3)

Algorithmen dürfte ein echter Mehrwert für die Quellenzuordnung und damit auch für die Identifizierung von neuen Spurenstoffen entstehen.

Im Rahmen einer Proof-of-Concept-Studie werden in einer Modellregion („Donau und Zuflüsse im Großraum Ulm“) Proben entnommen und von den Projektpartnern sowie den assoziierten Partner (Bodensee-Wasserversorgung, Hamburg Wasser, Hessenwasser und Westfälische Wasser- und Umweltanalytik) analysiert. Die Rohdaten werden im Anschluss – über ein eigenes entwickeltes Userinterface – auf den Server zur Auswertung geladen.

In der ersten Projektphase wurde der Schwerpunkt auf die Backend-Programmierung gelegt. Für die Datenprozessierung wurde die *enviMass*-Software (enviBee GmbH) an die Cloud-Umgebung eingebunden. Damit wird ein umfassendes Preprocessing der NTS-Daten für verschiedene LC-HRMS Geräte bereitstellt. Danach werden die Ergebnisse in eine *ElasticSearch*-Datenbank überführt. Machine- oder Deep-Learning-Algorithmen greifen wiederum auf diese Datenbank zu. Als IT-Partner entwickelt das Leibniz Rechenzentrum (LRZ) die Cloud-Infrastruktur und prüft diverse Methoden zur Datenauswertung.

In der zweiten Projektphase werden auch Schnittstellen und User Interfaces weiter ausgebaut. Durch das enge Zusammenspiel von IT- und Analytik-Spezialisten soll schon bei dem Demonstrator eine hohe Praxistauglichkeit erreicht werden.

Ausblick Zusammenarbeit

Trotz der unterschiedlichen Nutzergruppen (behördliche Gewässerüberwachung, Trinkwasserversorger, Forschungslabore) können Synergieeffekte unter den drei Forschungsprojekten genutzt werden. Bereits im Dezember 2021 wurde über gemeinsame Fragestellungen und Probleme diskutiert.

Machine- oder Deep-Learning-Ansätze benötigen riesige Datenmengen zum Training und zur Validierung der Methoden. Zwar sind NTS-Datensätze meist sehr groß, allerdings fast immer überbestimmt (Anzahl Features >> Anzahl Proben). Je mehr Daten zur Verfügung stehen, desto leistungsfähigere Methoden können genutzt werden und desto belastbarer sind die Ergebnisse. Beispielsweise können „sequence transduction models“ mit extrem großen Datenmengen umgehen, ohne numerisch instabil zu werden. Dem Datenaustausch von ausgewählten Datensätzen innerhalb der Projekte wurde daher unisono zugestimmt.

Neben dem Austausch von analytischen Messdaten wurde die Datenbank *ElasticSearch* für alle drei Projekte ausgewählt sowie gemeinsame Feldnamen und -typen vereinbart. Dies bietet den großen Vorteil, dass die aufwändige Programmierung von APIs bzw. graphischer Oberflächen unter den Projekten ausgetauscht werden kann bzw. zukünftig eine gemeinsame Datenbank verwaltet werden könnte.

Zur Früherkennung von Umweltrisiken und zur effizienten und umfassenden Gewässer- und Trinkwasserüberwachung stellt NTS eine vielversprechende Lösung dar. Eine Zusammenarbeit trägt dazu bei, künftige Herausforderungen effizient zu lösen und damit die NTS-Analytik für breite Anwendungen nutzbar zu machen. Die drei Forschungsprojekte leisten hierfür einen Auftakt. Mittel- bis langfristig müssen jedoch nachhaltige Lösungen unabhängig von Projektlaufzeiten entwickelt werden.

Danksagung

Neben den hier genannten Autoren, tragen in den einzelnen Projekten noch viele weitere Partner, assoziierte Partner oder Begleitgruppen zum Gelingen bei.

Finanzierung

► UBA-REFOPLAN Projekt „Online-Portal: Non-Target Screening für die Umweltüberwachung der Zukunft“ FKZ: 3720 22 201 0.

Quelle: eigene Abbildung BfG

Abbildung 48: Beitrag für „Wasser 2022“ eingereicht am 27.04.2022 – Seite 4

Beitrag für „Wasser 2022“ eingereicht am 27.04.2022 (Seite 4)

- Grant agreement for an action under life sub-programme for the environment: “Monitoring the pollution of water through non-target screening on the Rhine” 07.0201/2020/837943/SUB/ENV.C1

- BMBF Verbundprojekt „Künstliche und kollektive Intelligenz zum Spurenstoff-Tracking in Oberflächengewässern für eine nachhaltige Trinkwassergewinnung (K2I)“, Fördermaßnahme "Digital Green Tech - Umwelttechnik trifft Digitalisierung" (digitalgreentech.de) FKZ: 02WDG1593A-D.

Literatur

[1] Hollender, J., van Bavel, B., Dulio, V., Farmen, E., Furtmann, K., Koschorreck, J., Kunkel, U., Krauss, M., Munthe, J., Tornero, V., et al. Environ. Sci. Eur., 2019, 31(1), 42.

[2] Amtsblatt der Europäische Union C 445/126 vom 29.10.2021 „Umsetzung der Wassergesetzgebung der EU“ [eur-lex.europa.eu] Abgerufen am 20.04.2022

[3] Internationale Kommission zum Schutz des Rheins „Übereinkommen zum Schutz des Rheins“,12.04.1999, Bern. [www.iksr.org] Abgerufen am 25.04.2022

[4] Jewell, K., Thron, F., Schlüsener, M., Kramer, K., Scharrenbach, T., Fettig, I., Wick, A. et al. 2021, Ein Datenbankmodell für aggregierte Non-Target-Screening-Ergebnisse. Vom Wasser, 119(3), 94-96. [doi.org/10.1002/vomw.202100020]

[5] Umweltbundesamt 2020, Schwebstoffe - Suspendierte Stoffe. Umweltprobenbank des Bundes. [www.umweltprobenbank.de] Abgerufen am 20.04.2022.

Quelle: eigene Abbildung BfG

D.3 Kurzbeitrag für die „Sonderpublikation zum Langenauer Wasserforum“

Abbildung 49: Kurzbeitrag für die „Sonderpublikation zum Langenauer Wasserforum“ – Seite 1

Kurzbeitrag für die „Sonderpublikation zum Langenauer Wasserforum“ (Seite 1)

Dr. Kevin Jewell | Franziska Thron | Björn Ehlig | Dr. Nina Hermes | Sabrina Quanz | Dr. Michael Schlüsener | Ina Fettig | Jan Koschorreck | Kasjen Kramer
Dr. Thomas Scharrenbach | Prof. Dr. Thomas Ternes | Dr. Arne Wick

Aufbau einer Datenbank und Entwicklung eines Web-basierten Recherche- und Analysetools als Grundlage für ein überregionales Non-Target-Screening in der Umweltüberwachung der Zukunft.

Bisherige Erfahrungen in der Analyse von NTS-Daten zur Lokalisierung von Schadstoff-Einträgen basieren oft auf maßgeschneiderten, studien-spezifischen Analyseroutinen, die mit Skriptsprachen wie R, Python oder Matlab geschrieben wurden (Köppe et al. 2020, Krauss et al. 2019). Wir entwickeln eine Datenbank mit einem Web-basierten Datenrecherche- und Analysetool (Online-Portal: Non-Target Screening für die Umweltüberwachung der Zukunft, „NTS-Portal“, Finanzierung: REFOPLAN), um solche NTS-Analysen auch Studien- und Labor-übergreifend durchzuführen und damit auch überregional Aussagen über Vorkommen und Einleitungen der detektierten Schadstoffen treffen zu können. Mit den Analysefunktionen soll es unter anderem möglich sein, die über NTS detektierten Schadstoffe nicht nur überregional zu vergleichen, sondern auch die Quellen spezifischer Einleitungen einzugrenzen. Auf Basis dieser Informationen könnten zudem weitergehende Probenahmen und Target-Analysen engmaschiger, gezielter und effektiver aufgestellt werden. Die Herausforderungen für das NTS-Portal bestehen darin, die Analyseroutinen für die NTS-Daten zu generalisieren und mit einem intuitiven Interface auszustatten.
Im Folgenden werden zwei Beispiele von maßgeschneiderten Analyseskripten vorgestellt, die als Inspiration für Funktionen auf der Online-Plattform dienen.

Beispiele von Studien-spezifischen, räumlichen Untersuchungen mittels Non-Target-Screening

Im Rahmen des vom BMBF geförderten Verbundvorhabens NiddaMan wurde eine NTS-Studie an der Nidda und ihrer Nebenflüsse durchgeführt. Zur Identifizierung und Lokalisierung der Einleitungen wurde der sogenannte „Fold-Change“-Ansatz gewählt. Die Priorisierung von Features erfolgte durch einen Vergleich der Intensitäten von Features an der jeweiligen Messstelle mit allen Messstellen flussaufwärts (sowohl im Nebenfluss als auch im Hauptstrom der Nidda). Features mit einer mehr als 4-fach höheren Intensität (Fold-Change) als im Oberlauf wurden priorisiert (Abb. 1). Der Fold-Change von vier wurde empirisch ermittelt. Zum Beispiel zeigten 4% der Features an der Messstelle H4 mindestens 4-fach höhere Intensitäten im Vergleich zum Maximum der Intensitäten an den flussaufwärts gelegenen Messstellen der Horloff und der Nidda. Mit dieser Strategie konnten einige bisher unbekannte Substanzen in der Nidda identifiziert und der Ort des Eintrags lokalisiert oder zumindest eingegrenzt werden (Köppe et al. 2020).
Das zweite Beispiel befasst sich mit einem Sondermessprogramm am Rhein von 2017, das durch die IKSR Gruppe „SANA“ organisiert und durch viele Partnerinstitutionen entlang des Rheins durchgeführt wurde. Bei einer Auswertung des NTS-Datensatzes, die sich auf einen Teil der Proben konzentriert hat, wurde ebenfalls ein Fold-Change-Ansatz angewendet (Abb. 2). Features an den Messstellen, die eine 4-fach höhere Intensität als die maximale Intensität der überliegenden Messstellen aufweisen, wurden für die Identifizierung von Unbekannten priorisiert (Jewell et al. in prep).

Implementierung im NTS-Portal

Um ein Angebot für Behörden zu schaffen, damit NTS-basierte Gewässerdaten aus Forschungs- und Messprogrammen des Bundes und der Länder zusammengefasst und in einem gemeinsamen Portal genutzt werden können, sollen die Analysestrategien mit einer grafischen Webapplikation implementiert werden. Bei den Entwicklungen werden zwei gegensätzliche Ziele abgewogen: Die Vereinfachung der Analysestrategien (damit sie allgemein anwendbar sind) und deren Nutzbarkeit für spezifische Fragestellungen. Für das NTS-Portal wurde eine nicht relationale Datenbank (ElasticSearch) für prozessierte NTS-Daten aufgebaut. In den folgenden Beispielen

Quelle: (Jewell et al. 2021)

Abbildung 50: Kurzbeitrag für die „Sonderpublikation zum Langenauer Wasserforum“ – Seite 2

Kurzbeitrag für die „Sonderpublikation zum Langenauer Wasserforum“ (Seite 2)

werden Daten herangezogen, die mit einer für NTS optimierten LC-HRMS/MS-Methode gemessen wurden (Nürenberg et al. 2015). In dieser ersten Phase der Entwicklung des NTS-Portals sind nur annotierte Non-Target-Features in der Datenbank, d.h. Features, die über eine Spektrenbibliothek mit einem Substanznamen annotiert wurden (Jewell et al. 2019). Der Aufbau der Webapplikation wurde mit Hilfe der browserbasierten frei verfügbaren Analyseplattform Kibana durchgeführt.

Die Seite des NTS-Portals verfügt über unterschiedliche grafische Benutzeroberflächen zur Visualisierung von Daten, sogenannte „Dashboards“. Um sich beispielsweise den Intensitätsverlauf von NTS-Features in der Mosel anzeigen zu lassen (127 Stichproben des Oberflächenwassers im Längsverlauf), muss eine Suchanfrage (Query) an ElasticSearch mit den Parametern „Fluss: Mosel“ und „Matrix: Wasser“ gestellt werden. Die grafische Benutzeroberfläche nutzt dafür visuelle Kästchen als „Filter“ und übersetzt dabei die Anfrage für ElasticSearch.

Die Daten werden im Voraus (nicht Echtzeit) in R analysiert und Features, die einen Fold-Change von vier aufweisen, werden mit einem Label versehen („mosel_prio_foldchange_factor4“). Über dieses Label kann der Datensatz weiter eingegrenzt werden (Abb. 3a). Mit Hilfe der Tabelle können Substanzen ausgewählt und der komplette Datensatz für eine ausgewählte Substanz angeschaut werden (Abb. 3b). Dibutyl-Phosphat bspw. weist einen hohen Anstieg in der Intensität nach der Saarmündung auf. Die starke Abnahme der Intensität im Flussverlauf könnte z.B. mit den zeitlichen Abständen der Probenahme und einer nur kurzzeitigen Einleitung zusammenhängen.

Durch die Analyse der Daten im Voraus ist die Funktionalität des Dashboards schnell und stabil, allerdings kann der Fold-Change-Faktor nicht interaktiv justiert werden. Es können jedoch weitere Labels eingefügt werden, wie zum Beispiel „mosel_prio_foldchange_factor10“ usw.

Der hierfür benötigte Intensitätsvergleich ist nur bei zeitlich zusammenhängenden Proben möglich. Für die Analyse von überregional aggregierten NTS-Daten ohne zeitlichen Zusammenhang oder NTS-Daten von unterschiedlichen Matrices wurde eine alternative Analysestrategie implementiert, die nur die Anzahl der Befunde in die Statistik heranzieht (keine Vergleiche der Intensitäten). In diesem Fall werden Messstellen ausgewählt und dann Substanzen priorisiert, die besonders häufig an diesen Messstellen auftreten. Abbildung 4 zeigt ein Beispiel von Daten der unteren Elbe, hierbei wurden Messungen von Schwebstoffen (2007-2018, Umweltprobenbank des Bundes) und Oberflächenwasserproben von der Messstelle Tangermünde (2019, tägliche Mischproben) gemeinsam analysiert. Trotz der zeitlichen und methodischen Unterschiede konnten einige Substanzen priorisiert werden, die vermehrt in der unteren Elbe auftreten. Ein bekanntes Beispiel hier ist das Tetrabutylphosphonium (Brand et al. 2018). Die Aussagekraft steigt mit der Datengrundlage. Vor allem eine gute regionale Verteilung der Messungen wird für die Statistik benötigt.

Fazit und Ausblick

Die labor- und flussgebietsübergreifende Datenhaltung von Non-Target-Messungen sowie die Implementierung einfach nutzbarer und gleichzeitig generell einsetzbarer Funktionen für die Recherche und Priorisierung von Substanzen ist eine große Herausforderung – aber auch eine große Chance – für das Chemikalienmanagement. Die ersten Ansätze zeigen Potenzial und werden weiter ausgebaut. Es ist von Vorteil, möglichst viele qualitätsgesicherte NTS-Daten unterschiedlicher Programme in das Datenportal zu integrieren, denn die räumliche Verteilung der Proben hat eine entscheidende Rolle bei der räumlichen Eingrenzung der Belastung, unabhängig von der Größe des Non-Target-Datensatzes oder der Menge gefundener Features.

Neben der Entwicklung von weiteren grafischen Analysetools ist auch die Erweiterung der Plattform auf unbekannte (nicht annotierte) Features geplant. Die vorgestellten Funktionen können dann ebenso für die Priorisierung von Non-Targets und ihre anschließende Identifizierung genutzt werden.

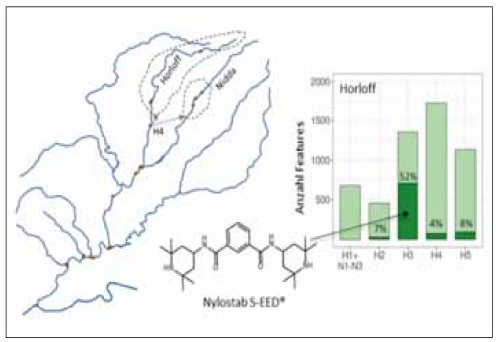


Abb. 1: Priorisierung von Non-Target-Features anhand der Fold-Change-Methodik in Köppe et al. (2020), positive Ionisierung. Hell grüne Balken zeigen die gesamte Anzahl von Features an den Messstellen, dunkelgrüne Anteile sind die priorisierten Features. Eine durch die Priorisierung identifizierte Substanz an Messstelle H3 ist Nylostab S-EED, das in der Textilherstellung Verwendung findet.

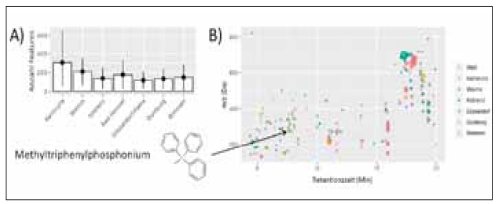


Abb. 2: a) Anzahl der priorisierten Features durch den Fold-Change-Ansatz am Rhein (positive Ionisierung). b) Scatterplot der ersten 10 priorisierten Features an jeder Messstelle (sortiert nach absteigender Intensität). Punktgröße: gemittelte Intensität.

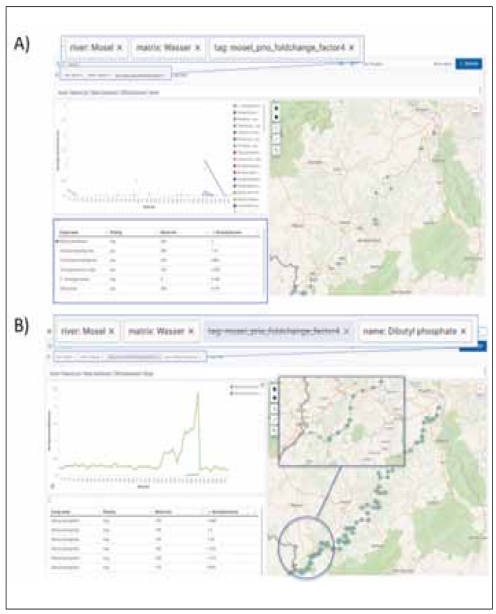


Abb. 3: Implementierung eines Fold-Change-Filters auf einer grafischen Oberfläche. A) Nach der Auswahl des Filters „mosel_prio_foldchange_factor4“ werden die priorisierten Substanzen in einer Tabelle dargestellt (mit absteigender Intensität). B) Nach Auswahl der ersten Substanz (Dibutyl-Phosphat) und Ausschaltung des Fold-Change-Filters werden alle Daten dieser Substanz gezeigt. Zahlen an den Messstellen sind die in Flussrichtung abnehmenden Moselkilometer. Punktgröße: gemittelte Intensität.

49

Quelle: (Jewell et al. 2021)

Abbildung 51: Kurzbeitrag für die „Sonderpublikation zum Langenauer Wasserforum“ – Seite 3

Kurzbeitrag für die „Sonderpublikation zum Langenauer Wasserforum“ (Seite 3)

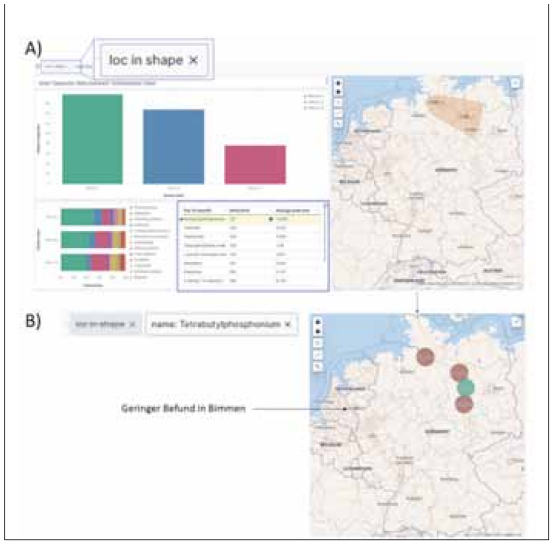


Abb. 4: Priorisierung von Substanzen, die vermehrt an ausgewählten Messstellen vorkommen (nach Anzahl Befunde) A) Nach der Auswahl der Messstellen auf einer Karte (mit der Maus, in Orange), werden die besonders häufig detektierten Substanzen in einer Tabelle gezeigt („loc“: location). B) Nach der Auswahl einer Substanz und Ausschaltung des Messstellen-Filters („loc in shape“) werden alle Daten der Substanz gezeigt, um die regionale Verteilung anzuschauen und somit möglicherweise die Quelle der Substanz einzugrenzen.

Brand, S., Schlüsener, M.P., Albrecht, D., Kunkel, U., Strobel, C., Grummt, T. and Ternes, T.A. (2018) Quaternary (triphenyl-) phosphonium compounds: Environmental behavior and toxicity. Water Research 136, 207–219.
Jewell, K.S., Kunkel, U., Ehlig, B., Thron, F., Schlüsener, M., Dietrich, C., Wick, A. and Ternes, T.A. (2019) Comparing mass, retention time and MS2 spectra as criteria for the automated screening of small molecules in aqueous environmental samples analyzed by LC-QToF-MS/MS. Rapid Communications in Mass Spectrometry 34, e8541.
Jewell, K.S., Schlüsener, M.P., Holz, J., Kunkel, U., Brüggen, S., Albrecht, D., Ternes, T.T. and Wick, A. (in prep).
Köppe, T., Jewell, K.S., Dietrich, C., Wick, A. and Ternes, T.A. (2020) Application of a non-target workflow for the identification of specific contaminants using the example of the Nidda river basin. Water Research 178, 115703.
Krauss, M., Hug, C., Bloch, R., Schulze, T. and Brack, W. (2019) Prioritising site-specific micropollutants in surface water from LC-HRMS non-target screening data using a rarity score. J Environmental Sciences Europe 31(1), 45.
Nürenberg, G., Schulz, M., Kunkel, U. and Ternes, T.A. (2015) Development and validation of a generic nontarget method based on liquid chromatography – high resolution mass spectrometry analysis for the evaluation of different wastewater treatment options. Journal of Chromatography A 1426, 77–90.

Autoren

Dr. Kevin Jewell[1], Franziska Thron[1], Björn Ehlig[1], Dr. Nina Hermes[1], Sabrina Quanz[1], Dr. Michael Schlüsener[1], Ina Fettig[2], Jan Koschorreck[2], Kasjen Kramer[1], Dr. Thomas Scharrenbach[1], Prof. Dr. Thomas Ternes[1], Dr. Arne Wick[1]
[1] Bundesanstalt für Gewässerkunde (BfG) | Am Mainzer Tor 1 | 56068 Koblenz
[2] Umweltbundesamt (UBA) | Bismarckplatz 1 | 14193 Berlin

Kontakt

Dr. Kevin S. Jewell, Wissenschaftlicher Mitarbeiter
Dr. Arne Wick, Referatsleiter Gewässerchemie
Bundesanstalt für Gewässerkunde (Federal Institute of Hydrology)
Referat G2 – Gewässerchemie –
Am Mainzer Tor 1 | 56068 Koblenz | www.bafg.de

Quelle: (Jewell et al. 2021)

E Codierung der Messstellen

Tabelle 5: Codierung der Messstellen

Beschreibungen: (Schulze und Ricking 2005)

Codierung Messstelle	Beschreibung
donau_jo_m	Donau, Stauhaltung Jochenstein, Donau km 2210, Flussmitte, Breitengrad: 48.567305, Längengrad: 13.604114
donau_ul_m	Donau, oberhalb Illermündung Ulm, Donau km 2593, Flussmitte, Breitengrad: 48.336936, Längengrad: 9.933709
elbe_ba_m	Elbe, Unterhalb Saalemündung (Barby), Elbe km 296, Flussmitte, Breitengrad: 51.996668, Längengrad: 11.888593
elbe_bl_m	Elbe, Elbmündung in Deutsche Bucht (Blankenese), Elbe km 634, Flussmitte, Breitengrad: 53.547521, Längengrad: 9.800194
elbe_cu_m	Elbe, Landschaftsschutzgebiet Elbe-Aland-Niederung (Cumlosen), Elbe km 470, Flussmitte, Breitengrad: 53.040792, Längengrad: 11.636399
elbe_tan_l	Elbe, Tangermünde (WSA Magdeburg), Elbe km 389.0, linkes Ufer, Breitengrad: 52.549876, Längengrad: 11.983321
elbe_pr_r	Elbe, Hafenbecken Prossen, Elbe km 13, rechtes Ufer, Breitengrad: 50.927016, Längengrad: 14.116322
elbe_ze_l	Elbe, Landschaftsschutzgebiet Zehren, Elbe km 96, linkes Ufer, Breitengrad: 51.264465, Längengrad: 13.402703
rhein_bi_l	Rhein, Messplattform der Internationalen Messstation Bimmen-Lobith, Rhein km 865, linkes Ufer, Breitengrad: 51.859995, Längengrad: 6.066983
rhein_if_m	Rhein, Schleuse Iffezheim (WSA Freiburg), Rhein km 333.2, Flussmitte, Breitengrad: 48.830796, Längengrad: 8.110804
rhein_ko_l	Rhein, Koblenz, Messstation der BfG, Rhein km 590.4, linkes Ufer, Breitengrad: 50.349601, Längengrad: 7.599691
rhein_we_r	Rhein, YC Weil-Märkt, Rhein km 172.2, rechtes Ufer, Breitengrad: 47.601589, Längengrad: 7.594340
saar_gu_m	Saar, Schleuse Güdingen, Saar km 92,9, Flussmitte, Breitengrad: 49.212001, Längengrad: 7.023477
saar_re_m	Saar, Schleuse Rehlingen, Saar km 54.2, Flussmitte, Breitengrad: 49.374082, Längengrad: 6.698360