# DQ-Bericht

## Erarbeitung und Evaluierung von Methoden der Qualitätssicherung von Daten und Metadaten für die POP-Dioxin-Datenbank des Bundes und der Länder

condat®

# DQ-Bericht

# Zusammenfassung des Umweltbundesamt

Auftraggeber
Umweltbundesamt
Wörlitzer Platz 1
06844 Dessau-Roßlau

Ansprechpartner:
Umweltbundesamt FG IV 2.1 / Herr Gärtner, philipp.gaertner@uba.de

Auftragnehmer
Condat AG
Alt-Moabit 91d
10559 Berlin

Laufzeit: Januar bis September 2013

# 1. Aufgaben- und Zielstellung

Das Umweltbundesamt (UBA) beabsichtigte den zuletzt in 2006 qualitätsgeprüften Datenbestand der Dioxindatenbank des Bundes und der Länder (kurz Dioxindatenbank) in 2013 wiederholt einer qualitätssichernden Prüfung zu unterziehen, um den Datenbestand an eine Datenqualität heranzuführen, die einer Open Data Governance-Strategie entspricht.

Ziel ist die Herstellung einer Datenqualität, die den Anforderungen an eine Open Data –UBA-Strategie entspricht. Das Projekt setzt in einem ersten Schritt das Regierungsprogramm „Vernetzte und transparente Verwaltung“ um mit dem Ziel, eine bundesweite Plattform für öffentlich zugängliche Daten zu schaffen. Zum einen sollen qualitätssicherende Metriken auf die Datenbestände angewendet werden. Zum anderen sollen Ergebnisse aus 2006 mit neuen Ergebnissen verglichen werden. Basis für den Vergleich bilden die in 2006 eingeführten automatisierten qualitätssichernden Verfahren auf vorhandene Datenbestände (in der Dioxindatenbank als Bewertungskriterien hinterlegt). Neben speziellen Fragenstellungen, wie beispielsweise der Untersuchung von sogenannten Hot Spots, sind im Rahmen dieses Projektes ebenfalls Überlegungen zur Etablierung einer DQ-Strategie zu tätigen.

Die Ziele des Projektes sind:

- die Analyse des Datenniveaus der Dioxindatenbank nach vorgegebenen Qualitätskriterien,
- die fachlich-technische Bewertung der Datenqualität und der Vergleich mit Ergebnissen aus 2006.

Die Durchführung von qualitätssichernden Maßnahmen in der Datenbank selbst (wie es in 2006 der Fall war) ist nicht Gegenstand dieses Projektes, sondern lediglich die Analyse und Bewertung der Datenqualität der Dioxindatenbank.

Das Projekt ist untergliedert in folgende Arbeitspakete:

1. **Qualitätskontrolle** des Datenbestandes der Dioxindatenbank,
2. **Quantitätskontrolle** in Bezug auf die Messwerte der Analysen-Ergebnisse, die TEQ-Berechnungen und die Ermittlung von Ausreißern,
3. Vorschlag zur Etablierung einer **DQ-Strategie** für die Dioxindatenbank,
4. Prüfung von **speziellen Fragestellungen**, wie beispielsweise sogenannter Hot Spots,
5. Zusammenführung aller Ergebnisse in einem **DQ-Bericht**.

Leistungsgegenstand ist die Durchführung, Dokumentation und Bewertung analytischer Qualitätsuntersuchungen für die **Umweltkompartimente.** Die Kompartimente Lebensmittel, Futtermittel und Humanmatrizes sind nicht Gegenstand des Vorhabens.

Im Rahmen des Projektes werden festgestellte Abweichungen und Defizite dokumentiert, jedoch keine Korrekturen von Daten im Quellsystem vorgenommen. Soweit für die Fortsetzung von Arbeiten erforderlich, werden Anpassungen ausschließlich in einer simulierten Arbeitsumgebung ausgeführt.

Grundanliegen des Vorhabens ist nicht nur die Feststellung von Zuständen sondern ergänzend auch deren fachlich-technische Bewertung. In erster Instanz wird dabei geprüft, ob eine bestimmte Sollvorgabe bzw. Erwartung erfüllt ist oder nicht. Diese Zielgrößen werden im Rahmen des Projektes mit dem Auftraggeber abgestimmt und in Form von Business Rules bereitgestellt. In einem zweiten Schritt kann dann bestimmt werden, ob die festgestellten Abweichungen in ihrer Qualität geschäftskritisch sind und welche Verbesserungsmaßnahmen in Betracht kommen.

## 1.1 Untersuchungsgegenstand

Für die durchzuführenden Untersuchungen wurde seitens des Auftraggebers UBA der gesamte Datenbestand der Dioxindatenbank auf QS-relevante Aspekte fokussiert. Demnach ergeben sich folgende fachliche Abgrenzungen.

Tabelle 1: Untersuchungsgegenstand

| Aspekt | Fokussierung |
|---|---|
| **Haupttabellen** | T_TITEL, T_STANDORT, T_PROBENAHME, T_PROBE, T_ANALYSEN_ERGEBNISSE |
| **Umwelt-kompartimente** | Abwasser, Stäube, Immissionen, Deposition, Emissionen, Innenraumluft, Sediment, Produkte, Boden terr., Boden subh., Biota, Wasser, Abfall |
| **Stoffarten** | Dioxine (PCDD), Furanen (PCDF) und Polychlorierte Biphenyle (PCB) |

Bei den Haupttabellen ist eine strenge Unterscheidung zwischen Daten und Metadaten nicht notwendig, da es in der Natur der in der Dioxindatenbank implementierten logischen Sicht ist, dass viele der gespeicherten Informationen eine Doppelrolle einnehmen und eine Unterscheidung für die Analysen nicht relevant ist.

Spezielle Fragenstellungen betreffen auch Tabellen wie beispielsweise den Gemeindeschlüsselkatalog.

## 1.2 Business Rules

Im Rahmen der Konkretisierung der Aufgabenstellung wurden zwischen AG und AN eine Reihe von fachlichen und technischen Business Rules abgestimmt, die in diesem Projekt zu überprüfen sind. Die Business Rules lassen sich den folgenden Bezugsebenen zuordnen:

- **Attribut**
  Hierbei handelt es sich um Analysen bezogen auf genau ein Attribut in einer Tabelle. Die zu prüfenden Business Rules sind überwiegend technischer Natur und behandeln beispielsweise Themen wie Vollständigkeit, Fehlerfreiheit, Eindeutigkeit und Widerspruchsfreiheit. Ergebnisse zu diesen Business Rules werden zu einem großen Teil durch Methoden des Data Profiling geliefert, aber auch durch konkrete Überprüfung fachlicher Regeln. Ein Beispiel für eine Business Rule auf Attribut-Ebene ist die Überprüfung von Pflichtfeldern.
- **Tabelle**
  Analysen bezogen auf mehr als ein Attribut (auf sogenannte Tupel) einer Tabelle überprüfen regelbasierte Zusammenhänge zwischen Attributen. Die zugrundeliegenden Business Rules sind generell fachliche Regeln.
  Ein Beispiel für eine Business Rule auf Tabellen-Ebene ist die Überprüfung des Zusammenhangs zwischen den Attributen M01 (bezieht sich auf den Messwert der Spalte in der Tabelle T_ANALYSEN_ERGEBNIS der Dioxindatenbank) und B01 (bezieht sich auf den Wert für Bestimmungsgrenze der Spalte in der Tabelle T_ANALYSEN_ERGEBNIS der Dioxindatenbank) der Analysenergebnisse.
- **Datenbank**
  Bei Analysen, die sich auf ein oder mehrere Attribute in mehr als einer Tabelle beziehen, ist potentiell die gesamte Datenbank betroffen. Welche Tabellen und Attribute in welchem Zusammenhang stehen, werden durch fachliche Business Rules bestimmt. Eine besondere Stellung hierbei nimmt das Thema der referentiellen Integrität (also die Sicherstellung der Existenz von Datensätzen zweier relationaler Tabellen durch Fremdschlüssel auf Basis von Integritätsbedingungen oder sogenannten Constraints) ein, da es sowohl technisch als auch fachlich sein kann. Da die technische referentielle Integrität bei der Verwendung von Datenbanken im Allgemeinen durch die Datenbank selbst mittels Constraints sichergestellt werden kann, spielt sie bei der DQ-Untersuchung eine untergeordnete Rolle. Die Überprüfung von fachlichen (regelbasierten) Zusammenhängen ist auf dieser Bezugsebene in der Regel sehr individuell und nur durch explizite Definition durchführbar.
  Ein Beispiel für eine Business Rule auf Datenbank-Ebene ist die Überprüfung des Zusammenhangs zwischen den Attributen STOFFSPEKTRUM_ID der Tabelle T_ANALYSEN_ERGEBNIS sowie KOMPARTIMENT_ID der Tabelle T_STANDORT und KOMPARTIMENT_ID und STOFFSPEKTRUM_ID der Tabelle T_STOFFSPEKTRUM_PROFILE.

Eine Zuordnung zu den genannten Bezugsebenen ist nicht immer leicht möglich, da jede Business Rule immer durch verschiedene Aspekte bestimmt ist. Beispielsweise ist in diesem Projekt der Untersuchungsgegenstand auf ausgewählte Umweltkompartimente abgegrenzt. Da nicht in jeder Tabelle sämtliche Informationen zu Kompartimente verfügbar sind, ist in der Regel diese Information aus anderen Tabellen zu ermitteln. Hinsichtlich der Überprüfung von Wertebereichen sind die vom Auftraggeber vorgegebenen Business Rules ebenfalls recht unterschiedlich. In vielen Fällen sind die Wertebereiche direkt angegeben (beispielsweise kann die Überprüfung des Wertebereichs des Merkmals „Art der Probe“ (Datenbankattribut PROBENART in der Tabelle T_PROBENAHME) anhand der vorgegebenen Werte 1 = Einzelprobe, 2 = Referenzprobe und 3 = Mischprobe erfolgen) und in anderen Fällen hingegen werden Referenztabellen zur Ermittlung der betreffenden Wertebereiche benannt (wie beispielsweise im Falle des Merkmals „Feuchte“ (Datenbankattribut DURCHFEUCHTUNG_ID in der Tabelle T_PROBE), für das die Bedingung „Select [Id],[Kurzbezeichnung],[Name] From [T_DURCHFEUCHTUNG];“ zu prüfen ist). Bei der Ergebnispräsentation wird versucht, diesem Umstand Rechnung zu tragen.

## 1.3 BiPRO-Bericht (2006)

Im Rahmen des Forschungsprojektes FKZ 204 62 251 „Qualitätssicherung und Erweiterung des Datenbestandes der Dioxindatenbank des Bundes und der Länder, einschließlich der Auswertung und Bewertung der Daten“ führte die BiPRO GmbH im Jahr 2006 Qualitätssicherungsroutinen durch. Im Ergebnis des Projektes fand eine Qualitätskennzeichnung der Probendatensätze mittels eines Bewertungskriteriums statt.

Die im BiPRO-Bericht (2006) dargestellten Untersuchungen des Bewertungskriteriums sind mit dem aktuellen Datenbestand zu vergleichen. Die Qualitätskriterien, welche dem BiPRO-Bericht zugrundeliegen sind[1]:

**Wertebereiche**
Für die Qualitätssicherung der Daten ist die Angabe des Wertebereiches der einzelnen Spalten wesentlich. Die Datensätze der Dioxindatenbank wurden auf die Einhaltung der Wertebereiche überprüft.

**Einheiten**
Falsche Einheiten verursachen Fehler von drei Größenordnungen oder mehr (nach oben oder unten). Abweichende Einheiten lassen sich daher bei der Kontrolle von Ausreißern erkennen.

**Falsche Werte**
Zur Identifikation von unglaubwürdigen Werten wurden statistische Ausreißertests (Nalimov-Test mit P=99%) und zusätzlich ein visuelles Screening auf Wertesprünge durchgeführt. Werte die außerhalb der Spannweite liegen (statistische Ausreißer) können sowohl außergewöhnlich hoch belastete Proben (echte Ausreißer) als auch fehlerhaft übertragene Daten repräsentieren.

Negative Analyseergebnisse können in der Realität nicht vorkommen.

**Bestimmungsgrenze**
Für jeden Analysewert, der „0“ ist, muss eine Bestimmungsgrenze angegeben sein. Die Bestimmungsgrenzen wurden mit statistischen Tests auf Plausibilität und Qualität hinsichtlich des zu diesem Zeitpunkt bestehenden Qualitätsstandards geprüft. Hierzu werden übliche Spannen für Qualitätsstandards für verschieden Probenahmezeiträume bestimmt. Bestimmungsgrenzen, die außerhalb der Spannen lagen, wurden auf Plausibilität geprüft.

**Ortsangaben, Datumsangaben, Quellenangaben**
Ortsangaben, Datumsangaben und Quellenangaben beinhalten unerlässliche Informationen für die Auswertung der Daten. Zur Abfrage räumlicher oder zeitlicher Trends wie z.B. saisonale Schwankungen muss die Datumsangabe auch den Monat der Probenahme enthalten. Für erforderliche Nachfragen muss jeder Datensatz auch Informationen zu Ansprechpartner oder Quellliteratur enthalten. Die Datensätze wurden diesbezüglich auf ihre Vollständigkeit geprüft.

[1] Textpassagen als Auszüge wiedergegeben. Die für das aktuelle Projekt nicht relevanten Aspekte wurden weggelassen.

**Kompartimente**

Für die Auswertung der Datensätze nach Trends ist die Angabe des jeweiligen Kompartiments erforderlich. Daher wurden die Probendatensätze auf ihre Vollständigkeit bezüglich der Kompartimentangabe untersucht. Darüber hinaus wurde die Plausibilität der Kompartimentangabe überprüft.

# 2. Methodik und Vorgehen

Dieses Kapitel beschreibt die methodische Vorgehensweise zur Vorbereitung, Durchführung und Bewertung der Untersuchungsergebnisse.

## 2.1 Allgemeiner Überblick

Das gesamte Projektvorhaben gliedert sich in Arbeitspakete, die sich direkt aus der Aufgaben- und Zielstellung ergeben. Das methodische Vorgehen ist für die Arbeitspakete „Qualitätskontrolle“, „Quantitätskontrolle“ und „spezielle Fragestellungen“ in den nachfolgenden Abschnitten beschrieben. Für die Arbeitspakete „DQ-Strategie“ und „DQ-Bericht“ gibt es keine besondere Methode. Die DQ-Strategie leitet sich aus den Ergebnissen ab und der DQ-Bericht fasst die Ergebnisse zusammen. In der Auswertung der dokumentierten Ergebnisse sind Empfehlungen für das zukünftige QS-Management abgeleitet und dokumentiert. Alle von den Arbeitspaketen gelieferten dokumentierten Ergebnisse, werden in eine zusammenfassende und kontinuierlich fortgeschriebene Ergebnisdokumentation.

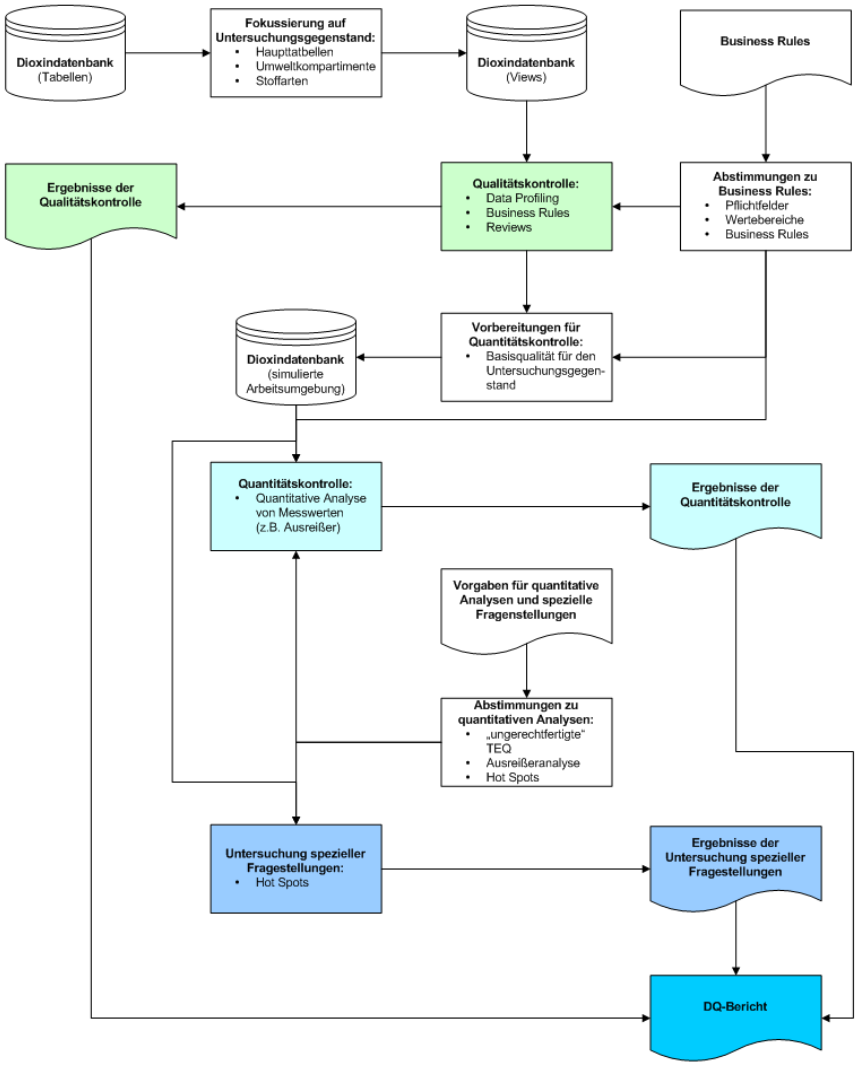



Abbildung 1: Überblick Projektablauf

Im Rahmen der Vorbereitung wird zunächst der Untersuchungsgegenstand datenbankseitig definiert. Dazu werden entsprechende Views angelegt, die sowohl hinsichtlich der benötigten Informationen als auch im Hinblick auf die Fokussierung auf ausgewählte Umweltkompartimente Festlegungen beinhalten. Parallel dazu wird begonnen, gemeinsam mit dem Auftraggeber die bereitgestellten Business Rules zu prüfen und abzustimmen. Die Überprüfung der Business Rules hinsichtlich Inhalt und Erwartungshaltung zu den formulierten Regeln ist notwendig, da sich die einzelnen Business Rules stark unterscheiden sowohl in Bezug auf Interpretation ihrer Fragestellung als auch in Bezug auf ihre Korrektheit. Da die Überprüfung der Business Rules arbeitspaketspezifisch erfolgt, ist es nicht notwendig bei der Bearbeitung der Qualitätskontrolle bereits Sachverhalte für andere Arbeitspakete zu berücksichtigen. Je nach Bearbeitungskontext werden die notwendigen Sachverhalte geklärt. Anschließend werden die abgestimmten Business Rules der Qualitätskontrolle technisch umgesetzt. Dazu wird mit Hilfe von Datenverarbeitungs-Werkzeugen[2] ein technisches Instrumentarium geschaffen, welches eine weitgehende Wiederholbarkeit ermöglichen soll. Da dies aufgrund der Individualität der einzelnen Fragestellungen nicht immer gewährleistet ist, kann (insbesondere bei sich ändernden Voraussetzungen) keine hundertprozentige wiederholbare automatische Ausführung erreicht werden.

Die Business Rules, welche für die Qualitätskontrolle untersucht werden, müssen beispielsweise durch spezielle aus der jeweiligen Regel abgeleitete Datenbankabfragen definiert werden, Untersuchungen bzgl. der Pflichtfelder oder der Wertebereiche lassen sich sehr gut anhand der beigestellten Informationen generieren. Sämtliche Ergebnisse aus allen Untersuchungen müssen für die Übernahme in den DQ-Bericht überarbeitet werden, was ebenfalls nicht automatisiert abläuft. Die anschließende Begutachtung und Bewertung der Ergebnisse ist wesentlicher Aspekt der Bearbeitung und im vorliegenden Projekt generell nicht als automatisierbarer Prozess vorgesehen. Wie dem dargestellten Workflow entnommen werden kann, bauen die weiteren Arbeitspakete auf der Qualitätskontrolle auf (vgl. Abbildung 1).

Das Datenmodell der Dioxindatenbank ist im nachfolgenden Schaubild (vgl. Abbildung 2) vereinfacht abgebildet und zeigt nur die im Rahmen der Untersuchungen betrachteten Tabellen. Dabei sind die Haupttabellen blau umrandet und die Referenztabellen grün umrandet dargestellt. Aus Gründen der Übersichtlichkeit wird darauf verzichtet, die Kreuztabelle (T_KREUZTABELLE) im Schaubild ebenfalls zu zeigen, sie steht aus fachlicher Sicht mit allen Haupttabellen (und weiteren Tabellen) in Beziehung.

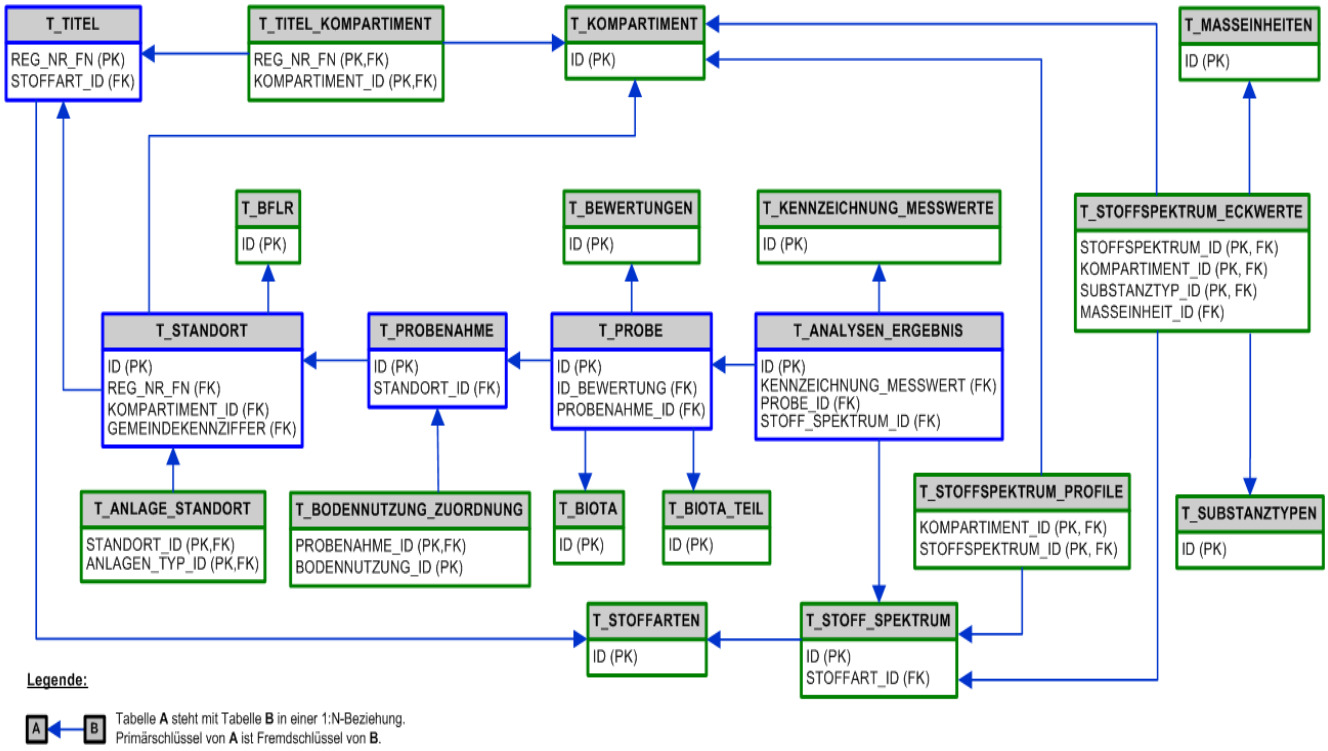



Abbildung 2: Datenmodell der Dioxindatenbank (vereinfacht dargestellt im Kontext Qualitätskontrolle)

[2] Werkzeuge bestehen bei den Business Rules zu etwa 50% aus SQL-Skripten, beim Data Profiling wird kein SQL eingesetzt.

## 2.2 Qualitätskontrolle

Im Fokus stehen hierbei für die benannten Haupttabellen die Qualitätskriterien „Relevanz“, „Vollständigkeit“, „Fehlerfreiheit“, „Eindeutigkeit“ und „Widerspruchsfreiheit“, soweit diese nicht bereits durch implementierte Datenbankmechanismen sichergestellt sind.

Ziel der Qualitätskontrolle ist zum einen, den Untersuchungsgegenstand anhand von ausgewählten DQ-Metriken geeignet darzustellen, um so schnell einen Überblick über die Datenlage zu bekommen und daraus erste Schlüsse für potentielle Probleme in den Daten erkennen zu können. Zum anderen ist es Ziel, Zusammenhänge innerhalb des Untersuchungsgegenstandes anhand bestimmter vorgegebener Regeln zu prüfen, in denen fachliche Erwartungen an die Daten formuliert werden. Als Techniken kommen hier sowohl das

- **Data Profiling** als auch
- die Überprüfung von **Business Rules**

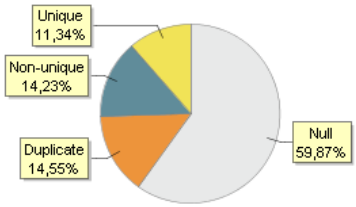
zur Anwendung.

**Data Profiling** ist eine Methode zur Analyse verfügbarer Daten in einer Datenbank, die darauf ausgerichtet ist, mittels verschiedener statistischer und weiterer Informationen einzelne Attribute einer Tabelle zu charakterisieren. Data Profiling kann dabei helfen, den Zustand von Daten in einer Datenbank anhand bestimmter DQ-Metriken zu ermitteln. Dabei hilft Data Profiling in erster Linie ein Bild von den Daten zu gewinnen, um Aussagen über die Beschaffenheit der Datenqualität machen zu können. Im Allgemeinen wird Data Profiling eingesetzt, um die Qualität von Daten zu bewerten, und nach einer entsprechenden Verbesserung der Datensituation eine wiederholte Analyse durchzuführen und mit der vorhergehenden Analyse zu vergleichen, um so Aussagen über eine Verbesserung zu bekommen (Monitoring).

Beim Data Profiling werden für bestimmte Tabellen und Attribute diverse Metriken berechnet, die in ihrer Auswertung als Qualitätssensoren genutzt werden können. Solche Metriken bzw. Auswertungen beinhalten unter anderem:

- Spaltenanalyse
- Häufigkeitsanalyse
- Domainanalyse
- Patternanalyse

**Spaltenanalyse**
Die Spaltenanalyse liefert statistische Aussagen über Attribute einer Tabelle mit den Informationen zum Datentyp, zur Wert-Anzahl, zur Anzahl der Datensätze, Anzahl der Datensätze mit „*NULL*“ bzw. „not *NULL*“, Anzahl der unterschiedlichen (distinct) Werte eines Attributs etc. Die Abbildung 3 zeigt ein Beispiel für ein (fiktives) Ergebnis einer durchgeführten Spaltenanalyse aus dem dann entsprechende Schlüsse gezogen werden können.




| Type | Count | % |
|---|---|---|
| Null | 83 413 | 59,87% |
| Non-null | 55 904 | 40,13% |
| Duplicate | 20 274 | 14,55% |
| Distinct | 35 630 | 25,57% |
| Non-unique | 19 830 | 14,23% |
| Unique | 15 800 | 11,34% |

Abbildung 3: Beispiel für eine Spaltenanalyse (ohne Bezug zur Dioxin-Datenbank)

Die Anzahlen in der gezeigten Abbildung haben dabei folgende Bedeutungen:

***NULL***: Anzahl aller Datensätze, bei denen die betreffende Spalte entweder leer ist oder den Wert *NULL* enthält.
**Non-*NULL***: Anzahl aller Datensätze, bei denen die betreffende Spalte entweder nicht leer ist und nicht den Wert *NULL* enthält (ist stets die Summe aus Duplicate + Distinct).

**Duplicate**: Anzahl der Datensätze, bei denen die betreffende Spalte mehrfache Werte enthält
**Distinct**: Anzahl der Datensätze, bei denen die betreffende Spalte nicht leere unterscheidbare Werte enthält (non-unique + unique)

**Non-unique**: Anzahl von Werten, die mindestens einen doppelten Wert aufweisen
**Unique**: Anzahl von Werten, die keinen doppelten Wert aufweisen

**Häufigkeitsanalyse**
Die Häufigkeitsanalyse (z.B. Werteverteilung und Häufigkeiten bei diskreten Werten) zeigt wie oft jeder Wert in den Daten enthalten ist (absolut und relativ zur Gesamtanzahl in Prozent).

**Domainanalyse**
In der Domainanalyse werden für alle nominal- und ordinalskalierten Datentypen die möglichen Ausprägungen und ihre Häufigkeit dargestellt, wobei unterscheidbare Werte extra angezeigt werden.

**Patternanalyse**
Die Pattern- oder Mask-Analyse wird für alle Spalten durchgeführt und zeigt die Wert-Struktur an. Durch eine starke Clusterbildung der Werte ist schnell erkennbar, ob die Struktur der Werte gleich ist. „W“ ist Platzhalter für ein Wort, „L“ für einen Buchstaben und „D“ für eine Ziffer. Nicht erwartete Werte und Wertemuster können mittels der Pattern- oder Mask-Analyse schnell erkannt werden.

**Business Rules** beschreiben gegenüber dem eher technisch gestalteten Data Profiling fachliche Erwartungshaltungen an die Daten. Wie bereits im Abschnitt Business Rules (vgl. Abschnitt 1.2, Seite 4) dargestellt, adressieren Business Rules unterschiedliche Bezugsebenen. Je nach Komplexität der betreffenden Regel ist auch ein entsprechend komplexes Vorgehen je Regel-Überprüfung notwendig. Diese zumeist speziellen Untersuchungen messen den Erfüllungsgrad einer Regel und stellen diesen der Erwartungshaltung gegenüber. Daraus ergeben sich dann Schlussfolgerungen, die als Ergebnisse in den DQ-Bericht gelangen.

Sowohl das Data Profiling als auch die Auswertung von Business Rules erfolgt im vorliegenden Projekt je nach Kontext mit oder ohne Berücksichtigung des Kompartiments. Werden Auswertungen nach (mehreren) Kompartimenten durchgeführt, so

werden diesen Auswertungen stets auch solche für alle betrachteten Kompartimente vorangestellt. Werden Auswertungen ausschließlich für ein Kompartiment durchgeführt, so bleiben sowohl die nicht betroffenen Kompartimente als auch alle anderen in ihrer Gesamtheit unberücksichtigt.

In diesem Zusammenhang sei angemerkt, dass bezüglich der Analysedaten im Rahmen der Qualitätskontrolle keine statistische Ausreißerbetrachtung erfolgt ,(sondern beispielsweise die Validität von Angaben (Messwerten) dahingehend geprüft wird, ob sie versehentlich mit Faktoren auf Basis abweichender Maßeinheiten versehen sind oder außerhalb erwarteter Genauigkeitsbereiche liegen.

Bei der Überprüfung von Business Rules werden Erwartungen an die Daten formuliert, deren Erfüllung oder Nicht-Erfüllung anschließend überprüft werden kann. Diese Regeln können einzelne Attribute betreffen, wie beispielsweise bei der Prüfung von Wertebereichen, oder beziehen sich auf Beziehungen zwischen Datensätzen oder Attributen. Im letzteren Fall sind insbesondere auch Integritätsbetrachtungen zwischen den Tabellen von Interesse, um Aussagen zur Vollständigkeit und Korrektheit der gegenseitigen Referenzierungen zu erhalten.

Die aus den Regelüberprüfungen (Business Rules) gewonnenen Ergebnisse werden anschließend den Sollvorgaben gegenübergestellt. Es wird festgestellt, ob die an dieser Stelle formulierte Erwartung bzw. Regelkonformität erfüllt ist. Im Falle einer Abweichung kann diese quantifiziert werden (in x % aller Fälle). Es liegt dann in der Verantwortung des Auftraggebers eine anzustrebende Erfüllungsrate festzulegen. Folgende Reaktionen sind möglich:

- Die Abweichung wird in Art und Umfang zur Kenntnis genommen, ohne dass ein begründetes Erfordernis vorliegt, diesen Zustand zu ändern.
- Es ist eine Verbesserung anzustreben, wobei Ziel die 100%ige Erreichung eines definierten Sollzustandes oder auch nur eine Anhebung des Qualitätsniveaus sein können.

Im Falle einer angestrebten Verbesserung sind Maßnahmen zu entwickeln und zu bewerten, die eine Zustandsbesserung durch Datenkorrektur zum Ziel haben. Des Weiteren sind Präventivmaßnahmen (i.d.R. Prozessverbesserungen) anzuleiten, die ein höheres DQ-Niveau nachhaltig und dauerhaft sicher stellen.

## 2.3 Quantitätskontrolle

Im Rahmen der Quantitätskontrolle werden insbesondere die Analysedaten untersucht. Die Untersuchung der Analysedaten (Messergebnisse) erfolgt, sofern notwendig, auf Basis von Daten, die hinsichtlich ihrer qualitätsbezogenen Eigenschaften einer mit dem Auftraggeber abgestimmten Basisqualität genügen. Erforderliche Datentransformationen, um diese Basisqualität zu simulieren, werden in einem temporären Arbeitsbereich in Vorbereitung der Untersuchungen ausgeführt.

In der nachfolgenden Tabelle sind die vom Auftraggeber festgelegten generellen Qualitätskriterien für die Quantitätskontrolle beschrieben.

Tabelle 2: Generelle Ausschlusskriterien

| Kombination | Anzahl Analysendatensätze insgesamt | Anzahl Analysendatensätze bereinigt[3] | Bemerkung |
|---|---|---|---|
| Für M01 und B01 ist mindestens ein Wert nicht mit 0 oder *NULL* belegt, d.h. ((M01 gleich 0 oder M01 gleich *NULL*) und B01 ungleich 0) oder (M01 ungleich 0 und (B01 gleich 0 oder B01 gleich *NULL*)). | 466.703 | 453.649 | Diese Kombination bildet den Normalfall der Analyseergebnisse der Dioxindatenbank. Da es vereinzelt vorkommen kann, dass M01/B01 anstelle eines Eintrags „0“ leer ist, ist die Berücksichtigung dieser „Qualitätsfehler“ in der Quantitätsanalyse dennoch begründet. |

Desweiteren wurde seitens des Auftraggebers der Untersuchungsgegenstand hinsichtlich der Stoffarten und Einzelkongenere festgelegt (vgl. auch Abschnitt 0, Seite 4). Die folgende Tabelle zeigt die festgelegten Einzelkongenere in den jeweiligen Stoffarten.

Tabelle 3: Beschränkung des Untersuchungsgegenstands auf Stoffarten und Einzelkongenere bzw. Homologene

| Stoffarten | Einzelkongenere und Homologene (STOFF_SPEKTRUM_ID) |
|---|---|
| dl-PCB | PCB 77 (3001),PCB 81 (3002), PCB 105 (3013), PCB 114 (3014), PCB 118 (3015), PCB 123 (3016), PCB 126 (3003), PCB 156 (3017), PCB 157 (3018), PCB 167 (3019), PCB 169 (3004), PCB 189 (3021) |
| Indikator PCB | PCB 28 (3005), PCB 52 (3006), PCB 101 (3007), PCB 138 (3008), PCB 153 (3009), PCB 180 (3010) |
| PCDD/PCDF | 1,2,3,4,7,8-HxCDF (1015), 1,2,3,6,7,8-HxCDF (1016), 2,3,4,6,7,8-HxCDF (1017), 1,2,3,7,8,9-HxCDF (1018), 1,2,3,4,6,7,8-HpCDF (1019), 1,2,3,4,7,8,9-HpCDF (1020), OCDF (1021), TCDF (1022), PeCDF (1023), HxCDF (1024), HpCDF (1025), 2,3,4,7,8-PeCDF (1014), 1,2,3,7,8-PeCDF (1013), 2,3,7,8-TCDF (1012), HpCDD (1011), HxCDD (1010), PeCDD (1009), TCDD (1008), OCDD (1007), 1,2,3,4,6,7,8-HpCDD (1006), 1,2,3,6,7,8-HxCDD (1005), 1,2,3,7,8,9-HxCDD (1004), 1,2,3,4,7,8-HxCDD (1003), 1,2,3,7,8-PeCDD (1002), 2,3,7,8-TCDD (1001), 1,2,3,4,6,7,9-HpCDD (1030), 1,2,3,7,8-/1,3,4,6,9-PeCDF (1028), 1,2,3,7,8-/1,2,3,4,8-PeCDF (1026), 1,2,3,4,7,8-/1,2,3,4,7,9-HxCDF (1027) |
| PCB-Homologene und weitere Einzelkongenere | TriCB (3022), TetraCB (3023), PentaCB (3024), HexaCB (3025), HeptaCB (3026), OctaCB (3027), NonaCB (3028), PCB 110 (3029), PCB 149 (3030), PCB 60 (3011), PCB 74 (3012), PCB 170 (3020) |

Auf Basis der sich ergebenen Daten werden die Quantitätskontrollen durchgeführt und die Ergebnisse mit entsprechenden Soll-Erwartungen abgeglichen und Ausreißer ermittelt. Auffälligkeiten und schwerpunktmäßig gewünschte Überprüfungen werden mittels weiterführender Analysen untersucht.

Die Untersuchungen werden, soweit dies sinnvoll ist, Kompartiment stratifiziert durchgeführt. Neben der Basisqualität bezogen auf die Analysewerte selbst, ist hier ein weiterer Aspekt der Basisqualität insbesondere die Eindeutigkeit bzw. Widerspruchsfreiheit der multiple referenzierten Kompartimente relevant.

Das Vorgehen und die eingesetzten Techniken orientieren sich zum einen an denen aus dem Arbeitspaket Qualitätskontrolle, wobei hier zu erwarten ist, dass vorrangig die Formulierung und Prüfung von Regeln zur Anwendung kommen. Die Folgeschritte sind analog dem Arbeitspaket Qualitätskontrolle. Ergänzend werden auch Zählungen bzgl. Einzelkongeneren innerhalb der Kompartimente durchgeführt.

[3] „bereinigt“ bedeutet in diesem Zusammenhang, die Anzahl sich ergebener Analysedatensätze unter Anwendung des in Spalte „Kombination“ genannten Einschlusskriteriums. Es wurden folglich nur Datensätze gezählt, die das genannte Einschlusskriterium erfüllen.

Im Rahm der Quantitätskontrolle stehen die Analysewerte im Fokus der Betrachtung. Eine Beeinträchtigung der Datenqualität ergibt sich insbesondere durch

- Ausreißer und
- fehlende Werte.

Auch bei der Qualitätskontrolle werden Untersuchungen hinsichtlich der Analysewerte durchgeführt. Beispielsweise wird die Zuordnung von Maßeinheiten zu Messwerten untersucht, weil falsche Maßeinheiten und darauf basierende Umrechnungen ebenfalls zu fehlerhaften Werten führen können.

**Ausreißer**

Ist ein vorhandener Wert im Verhältnis zu den meisten anderen Werten ungewöhnlich hoch bzw. tief, so wird dieser Wert als Ausreißer bezeichnet. Bzgl. Ausreißer können typischerweise folgende Feststellungen getroffen werden:

- der Wert basiert auf falschen Beobachtungen, falschen Messungen oder falschen Datenerfassungen,
- die Beobachtungen oder Messungen stammen aus einer anderen Population, als die der meisten anderen Werte,
- der Wert ist korrekt, stellt aber ein ungewöhnliches Ergebnis dar.

In den ersten beiden Fällen, würden die betroffenen Werte als „falsch“ interpretiert werden, im letzten Fall als „richtig“. Beispielsweise können diese „richtigen“ Werte im Zusammenhang mit besonderen Belastungssituationen (sogenannten „Hot Spots“gemessen werden. Als eine Konsequenz daraus ist der Prozess der Ermittlung von Ausreißern stets durch die beiden Phasen (1) Ausreißer-Ermittlung und (2) Ausreißer-Entscheidung charakterisiert. Zur Unterstützung dieses Prozesses werden statistische Tests bzw. Clustering-Methoden und Verteilungsdiagramme verwendet (sogenannte Scatterplots). Die Identifizierung eines tatsächlichen Ausreißers zu automatisieren birgt daher gewisse Schwierigkeiten, die letztlich nur durch eine sachkundige Bewertung von Menschen anhand von geeigneten Visualisierungshilfen getroffen werden kann, beispielsweise durch Scatterplots. Die folgende Abbildung zeigt einen Ausschnitt aus einem Scatterplot-Diagramm in Matrixform.

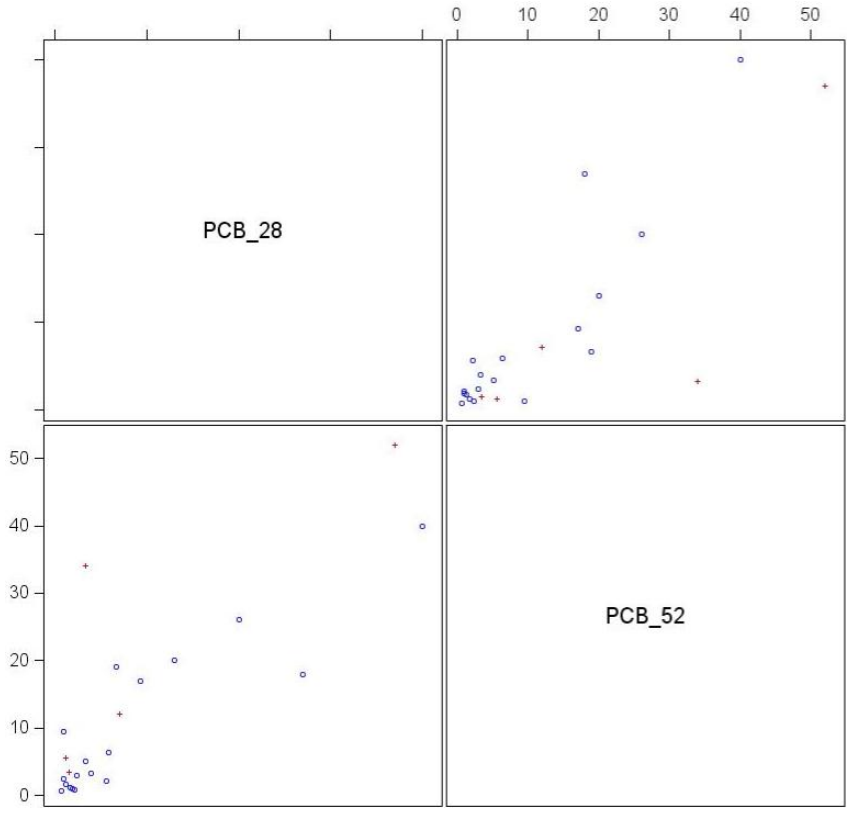



Abbildung 4: Beispiel für einen Scatterplot (Ausschnitt)

Ausreißer sind im gezeigten Scatterplot durch rote +-Zeichen dargestellt, andere Werte durch blaue o-Zeichen.

Die folgende Abbildung zeigt eine schematische Darstellung eines Boxplots. Boxplots werden häufig in Zusammenhang mit der Darstellung von Perzentilen verwendet, da sie eine sehr kompakte Form bieten viele Perzentile in einem Diagramm darstellen zu können.

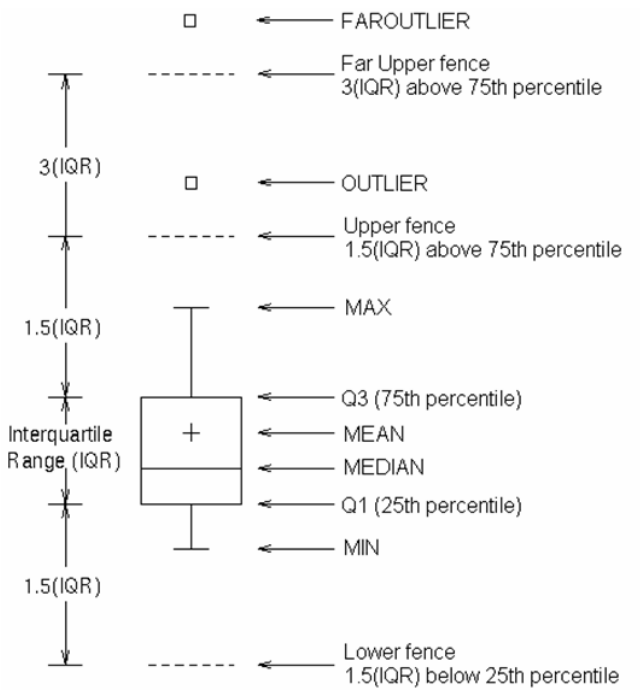



Abbildung 5: Beispiel für einen Boxplot (Schemaskizze)

Der hier gezeigte Mittelwert (engl. Mean) ist in der Abbildung durch ein +-Zeichen gekennzeichnet, die Ausreißer sind durch ein □-Zeichen gekennzeichnet. Oft werden für beide Zeichen alternativ auch o-Zeichen im Boxplot verwendet.

**Fehlende Werte**

Fehlende Werte lassen sich gegenüber den Ausreißern leichter ermitteln, jedoch ist bei vorliegenden fehlenden Werten zu entscheiden, ob der fehlende Wert durch einen anderen Wert ersetzt werden muss und kann oder ob der Wert fehlend bleibt und entsprechend in Analysen ausgeschlossen werden muss. Zur Ersetzung fehlender Werte gibt es zwei Varianten:

- der fehlende Wert wird auf Basis der Quelle „nachgetragen“,
- der fehlende Wert wird mittels mathematischer Methoden ergänzt (z.B. auf Basis eines interpolierten Wertes).

Im Rahmen des vorliegenden Projektes werden bei der Feststellung fehlender Werte keine Ersetzungen durchgeführt, sondern stattdessen die betreffenden Werte von den Analysen ausgeschlossen (eine detaillierte Beschreibung des durchgeführten Verfahrens befindet sich in Abschnitt 2.3.2, Seite 16).

In den nachfolgenden Abschnitten werden die Verfahren zur Ermittlung von „ungerechtfertigt“ berechneten TEQs und zur Ermittlung von Ausreißern detailliert beschrieben.

### 2.3.1 Ermittlung von „ungerechtfertigt“ berechneten TEQs

Ein TEQ-Wert ist dann als "ungerechtfertigt" anzusehen, wenn er existiert und die Voraussetzungen für eine Berechnung nicht vorliegen. Voraussetzungen für eine "gerechtfertigte" Berechnung eines TEQ-Wertes ergeben sich aus dem BiPRO-Bericht:

Im Einzelnen wurde ein Ausschluss von Proben vorgenommen wenn:

- für mehr als 30% der Einzelkongenere keine Werte vorlagen, d.h. die Eingabezellen leer waren (insbesondere für solche mit hohem TEF),
- entweder nur Daten für octachlorierte Kongenere bzw. nur Homologensummen in der Datenbank vorliegen und daher die Bestimmung der TEQ-Werte nicht vorgenommen werden konnte.

Anzumerken ist hierbei, dass "hohe TEFs" für die PCDD/F und PCBs nicht definiert wurden. Das Kriterium "octachlorierte Kongenere" ist nicht relevant, weil "octachlorierte Kongenere" bereits im 30%-Kriterium enthalten sind. Die abgestimmte Begründung hierfür lautet: Das genannte Ausschlusskriterium bzgl. "octachlorierte Kongenere", d.h. Einzelkongenere mit der STOFF_SPEKTRUM_ID "1007" ("OCDF") oder "1021" ("OCDD") scheint generell irrelevant zu sein, da das 30%-Kriterium generell verletzt sein würde, falls nur "octachlorierte Kongenere" verfügbar wären, d.h. max. 2 Einzelkongenere von insgesamt 17 Kongeneren, macht eine Quote von 88% fehlende Werte.). Das Kriterium "Homologensummen" (d.h. Datensätze mit T_STOFF_SPEKTRUM.SYMBOL = "S") ist nicht relevant, weil durch expliziter Angabe der zu berücksichtigenden Einzelkongeneren anhand des „Hintergrundpapiers zur TEQ-Berechnung“[4] keine Schnittmenge existiert (Homologensummen sind Stoffarten mit STOFF_SPEKTRUM_NAME = (PeCDD, HpCDD, TCDD, HxCDF, TriCB, OctaCB, HexaCB, NonaCB, HxCDD, PentaCB, TCDF, PeCDF, HpCDF, HeptaCB)).

Somit gilt einzig und allein die 30%-Regel.

Die Zuordnung von TEQ-Werten der Kreuztabelle (vgl. entsprechende Spalten in T_KREUZTABELLE) zu Einzelkongeneren erfolgt anhand des Dokuments „Hintergrundpapiers zur TEQ-Berechnung“ wie folgt:

Tabelle 1: Zuordnung von TEQ-Werten der Kreuztabelle zu Einzelkongeneren und Homologenen

| TEQ-Werten der Kreuztabelle | Einzelkongenere und Homologene |
|---|---|
| "I-TEQ";<br>"WHO-PCDD/F-TEQ (2005)";<br>"WHO-PCDD/F-TEQ" | "2,3,7,8-TCDD"; "1,2,3,7,8-PeCDD"; "1,2,3,4,7,8-HxCDD"; "1,2,3,6,7,8-HxCDD"; "1,2,3,7,8,9-HxCDD"; "1,2,3,4,6,7,8-HpCDD"; "OCDD"; "2,3,7,8-TCDF"; "1,2,3,7,8-PeCDF"; "2,3,4,7,8-PeCDF"; "1,2,3,4,7,8-HxCDF"; "1,2,3,6,7,8-HxCDF"; "1,2,3,7,8,9-HxCDF"; "2,3,4,6,7,8-HxCDF"; "1,2,3,4,6,7,8-HpCDF"; "1,2,3,4,7,8,9-HpCDF"; "OCDF" |
| "WHO-PCB-TEQ (2005)";<br>"WHO-PCB-TEQ" | "PCB 77"; "PCB 81"; "PCB 126"; "PCB 169"; "PCB 105"; "PCB 114"; "PCB 118"; "PCB 123"; "PCB 156"; "PCB 157"; "PCB 167"; "PCB 189" |
| "WHO-PCDD/F-PCB-TEQ (2005)";<br>"WHO-PCDD/F-PCB-TEQ" | "2,3,7,8-TCDD"; "1,2,3,7,8-PeCDD"; "1,2,3,4,7,8-HxCDD"; "1,2,3,6,7,8-HxCDD"; "1,2,3,7,8,9-HxCDD"; "1,2,3,4,6,7,8-HpCDD"; "OCDD"; "2,3,7,8-TCDF"; "1,2,3,7,8-PeCDF"; "2,3,4,7,8-PeCDF"; "1,2,3,4,7,8-HxCDF"; "1,2,3,6,7,8-HxCDF"; "1,2,3,7,8,9-HxCDF"; "2,3,4,6,7,8-HxCDF"; "1,2,3,4,6,7,8-HpCDF"; "1,2,3,4,7,8,9-HpCDF"; "OCDF"<br>"PCB 77"; "PCB 81"; "PCB 126"; "PCB 169"; "PCB 105"; "PCB 114"; "PCB 118"; "PCB 123"; "PCB 156"; "PCB 157"; "PCB 167"; "PCB 189" |

In der Kreuztabelle gibt es stets Werte für **drei berechnete TEQs** mit den Faktoren der I-TEFs, der WHO-TEFs$_{1998}$ und der WHO-TEFs$_{2005}$. Jedes der drei mit unterschiedlichen Faktoren berechneten Toxizitätsäquivalente enthält wiederum drei TEQs, die auf Basis von Berechnungsalgorithmen mit den unterschiedlichen Bestimmungsgrenzen durchgeführt werden. Liegt kein Messwert für ein Einzelkongener vor, aber eine Bestimmungsgrenze, so wird der TEQ unter Einbeziehung der vollen, halbe und gleich Null - Bestimmungsgrenze berechnet. In der Datenbank sind das die Kennzeichen für ein TEQ wie folgt:

"lb" (lower bound), "mb" (middle bound) und "ub" (upper bound).

Es gibt unterschiedliche Fälle, die sich bei der Überprüfung als Resultat ergeben können:

(1) der TEQ-Wert in der Kreuztabelle existiert und die Voraussetzungen sind ebenfalls erfüllt, d.h. die Berechnung ist "gerechtfertigt",

[4] Vgl. auch „Hintergrundpapier_TEQ_Berechnung.docx“ als Beistellung des Auftraggebers.

(2) der TEQ-Wert in der Kreuztabelle existiert nicht und die Voraussetzungen sind nicht erfüllt, d.h. die nicht erfolgte Berechnung ist ebenfalls "gerechtfertigt",

(3) der TEQ-Wert in der Kreuztabelle existiert aber die Voraussetzungen sind nicht erfüllt, d.h. die Berechnung ist somit "ungerechtfertigt",

(4) der TEQ-Wert in der Kreuztabelle existiert nicht obwohl die Voraussetzungen erfüllt sind, d.h. nicht erfolgte Berechnung ist "ungerechtfertigt".

Von den genannten Fällen sind (3) und (4) interessant und als Fehler einzustufen.

### 2.3.2 Ermittlung von Ausreißern

In diesem Abschnitt werden die Kriterien zum generellen Ausschluss von Datensätzen und zur Stratifizierung beschrieben sowie das Verfahren zur Ermittlung von Ausreißern skizziert.

Für eine Ausreißeranalyse müssen mindestens doppelt so viele Datensätze wie Dimensionen vorhanden sein. Bei niedrigerer Probenanzahl ist keine Ausreißeranalyse durchzuführen.

Tabelle 2: Minimal benötigte Anzahl an Analysedatensätzen pro Stoffart[5]

| Stoffart | Dimensionen | Minimal benötigte Anzahl Analysedatensätze |
|---|---|---|
| dl-PCB | 12 | 24 |
| Indikator PCB | 6 | 12 |
| PCDD/PCDF | 29 | 58 |
| PCB | 12 | 24 |

Die Ausreißeranalysen erfolgen stets getrennt nach:

- Kompartiment,
- Stoffart,
- und je nach Kompartiment nach weiteren Auswahlkriterien.

Da es sich bei den Kongeneren um Kompositionsdaten handelt, müssen die Daten zunächst transformiert werden. Je nach verwendeter Einheit ist entweder eine ilr- oder eine log-Transformation durchzuführen. Bei der log-Transformation werden alle Werte logarithmiert zur Basis e verwendet. Die ilr-Transformation erfolgt anhand der folgenden Formel:

*ilr*(x) = ($z_1$, ..., $z_{D-1}$) mit

$$z_i = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j}}{x_i} \quad \text{for } i = 1, \ldots, D-1$$

wobei D die Anzahl der Dimensionen ist.

Zur Durchführung der ilr-Transformation wird daher immer ein Restwert benötigt, der sich je nach Einheit anhand der folgenden Basiswerte zur Restwertbestimmung ergeben:

µg/kg → 10e+9
µg/g → 10e+6
ng/kg → 10e+12
ng/g → 10e+9

[5] Die Werte ergaben sich aus den Anzahlen der in Tabelle 1 genannten Einzelkongeneren und Homologenen.

| | | |
|---|---|---|
| pg/kg | → | 10e+15 |
| pg/g | → | 10e+12 |
| fg/kg | → | 10e+18 |
| fg/g | → | 10e+15 |

Der Ausreißertest wird mittels eines robusten Verfahrens nach folgender Methode berechnet[6]:

METHOD=M(SCALE=MED) This option obtains $\hat{\sigma}$ by the iteration

$$\hat{\sigma}^{(m+1)} = \text{med}_{i=1}^{n} |y_i - x_i^T \hat{\theta}^{(m)}| / \beta_0$$

where $\beta_0 = \Phi^{-1}(.75)$ is the constant such that the solution $\hat{\sigma}$ is asymptotically consistent when $L(\cdot/\sigma) = \Phi(\cdot)$ (refer to Hampel et al. 1986, p. 312).

[6] Eine detaillierte Darstellung der Methode und des verwendeten robusten Verfahrens kann in Colin Chen: Robust Regression and Outlier Detection with the ROBUSTREG Procedure SUGI 27 Paper 265-27 nachgelesen werden.

# 3. Evaluierung und Ergebnisse

## 3.1 Zusammenfassung

Die im durchgeführten DQ-Projekt ermittelten Ergebnisse zeigen kein vollständiges Bild der Dioxindatenbank. Durch Eingrenzung des Untersuchungsgegenstands auf die für die Auftraggeber relevanten Aspekte und Festlegung der anzuwendenden analytischen Schwerpunkte ist es aber gut möglich, eine Bewertung der Datenqualität der Dioxindatenbank für den Untersuchungsgegenstand vorzunehmen und daraus Empfehlungen für ein weiteres Vorgehen abzuleiten.

In den nachfolgenden Abschnitten werden die Ergebnisse der Untersuchungen dargestellt. Wegen der sehr umfangreichen Ergebnisausgaben für einzelne Analysen, werden in diesem Dokument nicht alle Ergebnisausgaben im Detail abgebildet, sondern die wichtigsten Ergebnisse summarisch zusammengefasst.

Der letzte Abschnitt befasst sich mit der Bewertung der Datenqualität der Dioxindatenbank für den Untersuchungsgegenstand und daraus ableitbare Empfehlungen für ein weiteres Vorgehen. In Erweiterung der Bewertung werden für das Thema Data Governance als strategische Ausrichtung Feststellungen getroffen, die im Rahmen der Umsetzung einer entsprechenden Initiative hilfreiche Aussagen liefern können.

## 3.2 Übersicht über den Datenbestand

Dieser Abschnitt gibt zunächst eine Übersicht über den Datenbestand eingeschränkt auf den Untersuchungsgegenstand und mit den zur Beantwortung relevanter Fragestellungen referenzierten weiteren Tabellen. Die Tabelle 7 zeigt die in die Untersuchung einbezogenen Datensätze der analysierten Haupttabellen.

Tabelle 7: Datenbestand der Haupttabellen für ausgewählte Umweltkompartimente

| Datenbestand | Tabelle | Anzahl Datensätze |
|---|---|---|
| Titel/Messprogramm | T_TITEL[7] | 217 (160) |
| Standorte | T_STANDORT | 9.323 |
| Probenahmen | T_PROBENAHME | 18.252 |
| Proben | T_PROBE | 21.775 |
| Analysenergebnisse | T_ANALYSEN_ERGEBNISSE | 466.703 |

Die Dioxindatenbank teilt die Bodenproben in die Subkompartimente „Boden terrestrisch“ und „Boden subhydrisch“ ein. Im vorliegenden DQ-Bericht erfolgt die Darstellung generell getrennt für die beiden Subkompartimente. Das Gleiche gilt für das Kompartiment Luft, aus dem nur die Subkompartimente „Stäube“, „Immissionen“, „Deposition“, „Emissionen“ und „Innenraumluft“ im Bericht unabhängig voneinander dargestellt sind.

Eine Übersicht über die Anzahl der Probendatensätze des Datenbestands der Dioxindatenbank (Stand Januar 2013) zu den betrachteten Umweltkompartimenten ist Abbildung 6 zu entnehmen.

[7] Die erste Anzahl Datensätze der Tabelle T_TITEL bezieht sich auf alle sich ergebene Datensätze der Relation zwischen T_TITEL und T_STANDORT, der Wert in Klammern bezieht sich auf eindeutige Werte des Primary Keys REG_NR_FN.

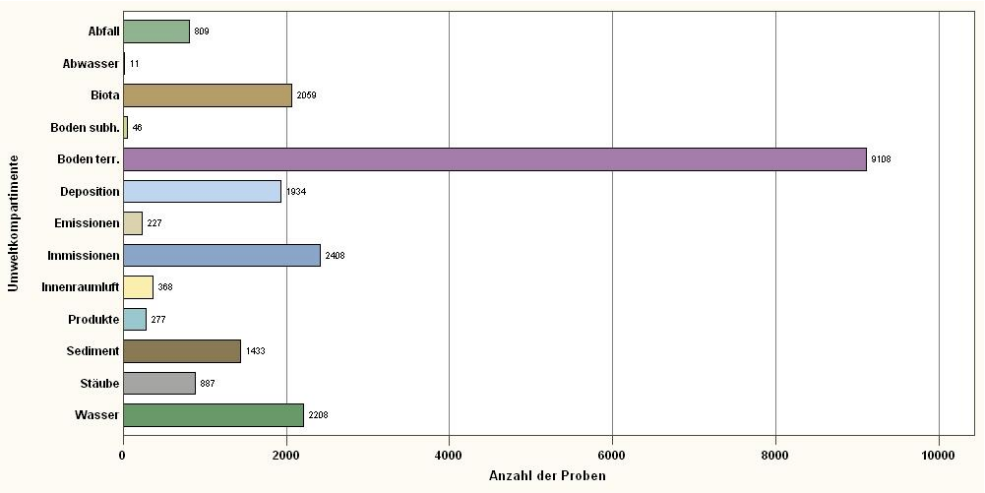



Abbildung 6: Anzahl der Proben in der Dioxindatenbank nach Umweltkompartimenten (Stand Januar 2013)

Wie der Abbildung 6 zu entnehmen ist, sind in der Dioxindatenbank gespeicherte Probendatensätze für das Kompartiment **Boden terrestrisch** (über 9.000 Probendatensätze) am besten repräsentiert. Anschließend folgen Datensätze für das Kompartiment **Immissionen** (über 2.400 Probendatensätze), **Wasser** (über 2.200 Probendatensätze) und **Biota** (über 2.000 Probendatensätze). Danach folgen Datensätze für das Kompartiment **Deposition** (knapp unter 2.000 Probendatensätze), **Sediment** (über 1.400 Probendatensätze), **Stäube** (ca. 890 Probendatensätze) und **Abfall** (ca. 800 Probendatensätze). Alle weiteren Kompartimente sind nur recht wenig vertreten.

## 3.3 Qualitätskontrolle

In diesem Abschnitt sind die Untersuchungsergebnisse der Qualitätskontrolle für den gesamten Untersuchungsgegenstand und alle ausgewählten Umweltkompartimente dargestellt. Wo eine detaillierte Darstellung je Umweltkompartiment sinnvoll und möglich ist, werden im Allgemeinen zunächst eine Zusammenfassung für alle Umweltkompartimente vorgenommen und anschließend das Ergebnis je Umweltkompartiment dargestellt. Die Reihenfolge der Darstellung folgt hierbei den Themenschwerpunkten der Aufgabenstellung. Sofern einzelne Kompartimente getrennt dargestellt werden, ist die Reihenfolge der Kompartimente alphabetisch.

Eine Darstellung aller Untersuchungsergebnisse pro Umweltkompartiment, wie es beispielsweise im BiPRO-Bericht der Fall ist, hat sich im vorliegenden Fall nicht als nützlich erwiesen, weil durch die vielzähligen verschiedenen Untersuchungen die Vergleichbarkeit in Bezug auf die Aufgabenschwerpunkte durch eine kompartimentsbezogene Darstellung erschwert würde. Auch ist die Datenbasis für jede Fragenstellung nicht gleich, sodass ein Vergleich über Kompartimente besser ist als pro Kompartiment ein Vergleich über alle Fragestellungen zu haben.

Die Dokumentation der Ergebnisse erfolgt in der Regel tabellen- und attributbezogen, sodass im Falle von identifizierten DQ-Problemen eine einfache Lokalisierung der Quelle in der Dioxindatenbank möglich ist. Bis auf die Ergebnisse zum Data Profiling und zum Vergleich mit dem BiPRO-Bericht wurden alle ermittelten DQ-Probleme und Auffälligkeiten durch die Angabe einer Fehlerquote quantifiziert. Somit ist auf der gewählten Darstellungsebene eine schnelle Bewertung des Sachverhalts möglich. In den meisten Fällen ist der Auftragnehmer nicht in der Lage die Kritikalität der Sachlage zu beurteilen. Somit obliegen die Finalbewertung und die Entscheidung über die Einleitung von Maßnahmen in Art und Umfang dem Auftraggeber.

### 3.3.1 Ergebnisse für Data Profiling

Zusammenfassend lässt sich feststellen, dass die meisten entdeckten Auffälligkeiten vermutlich dadurch entstehen, dass analoge Wertemengen nicht vereinheitlicht sind. Eine strengere Formalisierung bezüglich dieser Wertemengen könnte die Nutzbarkeit der Angaben jedoch verbessern. Insbesondere sind hierbei hervorzuheben, dass *NULL*-Werte oft die Rolle von Werten wie „0“ oder „unbekannt“ übernehmen. Einige Attribute scheinen gar nicht benötigt zu werden (die entsprechenden Werte sind stets *NULL*). In diesen Fällen wäre die Entfernung der Attribute überlegenswert, da die Verständlichkeit des zugrundeliegenden Datenmodells verbessert werden könnte. In anderen Fällen ist offensichtlich die Feldlänge der Attribute nicht ausreichend, um die betreffenden Informationen zu speichern. In diesen Fällen wäre neben der Möglichkeit der Feldlängenerweiterung auch eine entsprechende Formalisierung der Werte überlegenswert. In jedem Fall ist zu prüfen, ob die betreffenden Datenwerte ggf. zu korrigieren sind.

## 3.4 Quantitätskontrolle

Zunächst werden die Ergebnisse bzgl. der Häufigkeit verfügbarer Messwerte dargestellt. Hierfür wurde eine grafische Darstellung in Form von horizontalen Balkendiagrammen vom Auftraggeber gewünscht. Von den insgesamt 32 Diagrammen sind 3 im vorliegenden Bericht dargestellt.  Häufigkeit verfügbarer Messwerte

Hinsichtlich der Häufigkeit verfügbarer Messwerte kann generell festgestellt werden, dass das Bild recht unterschiedlich ist, je nach dem welches Kompartiment und welche Stoffart betrachtet wird. Die im Folgenden dargestellten Diagramme geben ein gutes Zeugnis über die Verteilung der Häufigkeiten der Messwertverfügbarkeit ab. Als Diagramme werden im Dokument gezeigt:

- Kompartiment Deposition / Stoffart dl-PCB
- Kompartiment Innenraumluft / Stoffart Indikator PCB
- Kompartiment Produkte / Stoffart PCDD/PCDF

### 3.4.1 Kompartiment Deposition / Stoffart dl-PCB

Für die Stoffart dl-PCB im Kompartiment Deposition liegt die Häufigkeit der verfügbaren Messwerte insgesamt etwa bei 50%, allerdings ist gar kein Kongener zu 100% verfügbar. Die Abbildung 7 zeigt die Verteilung der Einzelkongenere für die Stoffart dl-PCB im Kompartiment Deposition.

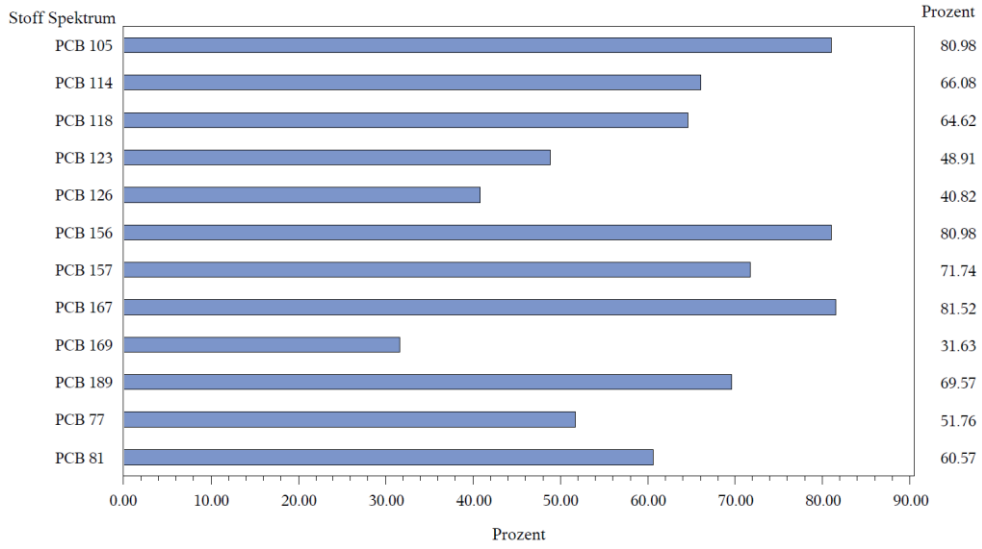



Abbildung 7: Häufigkeit verfügbarer Messwerte für das Kompartiment Deposition und der Stoffgruppe dl-PCB

### 3.4.2 Kompartiment Innenraumluft / Stoffart Indikator PCB

Für die Stoffart Indikator PCB im Kompartiment Innenraumlauf liegt die Häufigkeit der verfügbaren Messwerte insgesamt etwa bei 30%, keines der sechs Indikator-PCB ist zu 100% verfügbar. Die Abbildung 8 zeigt die Verteilung der Einzelkongenere für die Stoffart Indikator PCB im Kompartiment Innenraumlauf.

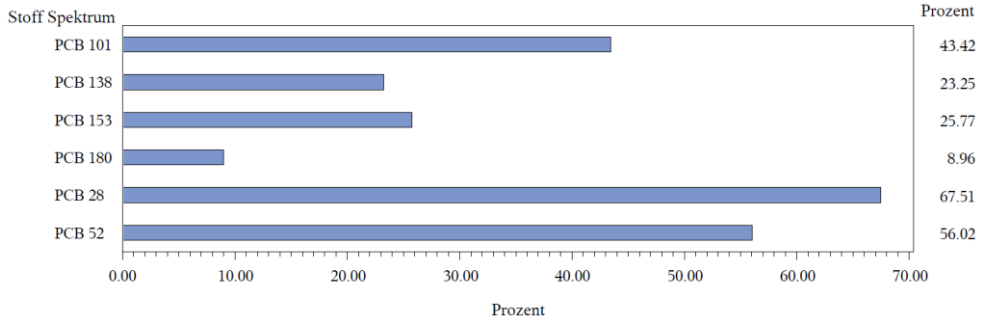



Abbildung 8: Häufigkeit verfügbarer Messwerte für das Kompartiment Innenraumluft und der Stoffgruppe Indikator PCB

### 3.4.3 Kompartiment Produkte / Stoffart PCDD/PCDF

Für die Stoffart PCDD/PCDF im Kompartiment Produkte liegt die Häufigkeit der verfügbaren Messwerte insgesamt etwa bei 65%, nur die Kongenere „1,2,3,4,7,8 HxCDF“ und „1,2,3,4,7,9 HxCDF“ sind zu 100% verfügbar. Die Abbildung 9 zeigt die Verteilung der Einzelkongenere für die Stoffart PCDD/PCDF im Kompartiment Produkte. Anzumerken ist hierbei, dass 2 Einzelkongenere fehlen (d.h. entsprechende Datensätze sind nicht verfügbar), wodurch im Diagramm nicht alle möglichen Einzelkongenere darstellt sind.

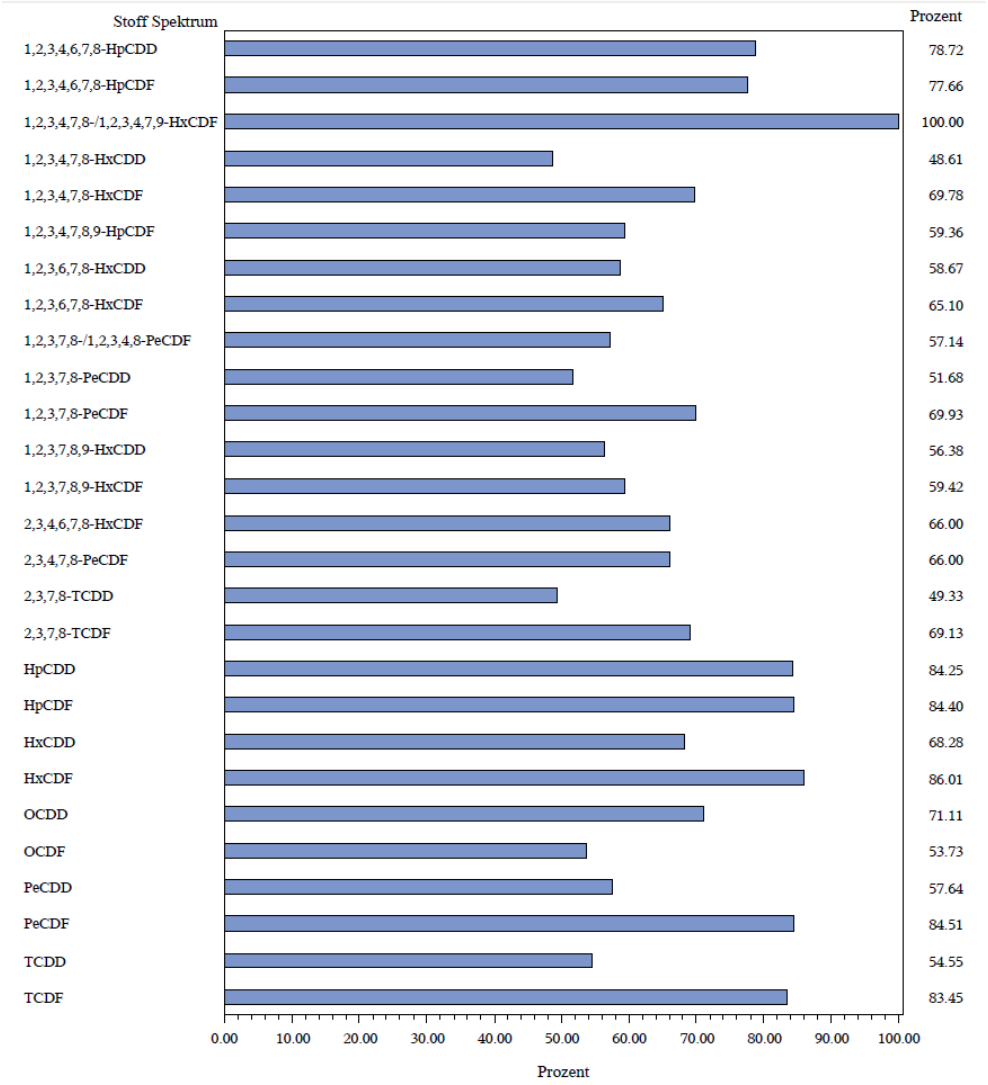



Abbildung 9: Häufigkeit verfügbarer Messwerte für das Kompartiment Produkte und der Stoffgruppe PCDD/PCDF

# 4. Empfehlungen und weiteres Vorgehen

## 4.1 Untersuchung zur Datenqualität POP-Dioxin-Datenbank und Open Data Governance als strategische Ausrichtung

Das Vorhaben basiert nicht auf einer isolierten Initiative zur Untersuchung eines bestimmten Datenbestandes im Umweltbundesamt unter dem Gesichtspunkt seiner spezifischen Datenqualität, sondern stellt bereits mit seiner Aufgabenstellung auch einen übergreifenden Zusammenhang her. Die Leistungsbeschreibung legt als mitgeltendes Ziel fest, dass „*der seit 1995 im Ausbau befindliche Datenbestand der POP-Dioxin-Datenbank des Bundes und der Länder an diejenige Datenqualität herangeführt werden (soll), die einer Open Data Governance Strategie entspricht. … Neue technische Möglichkeiten, verstärkte partizipative Bestrebungen und ökonomische Erwägungen (Mehrfachnutzung von Daten) erfordern neue und übergreifende Strategien zur Bereitstellung und Nutzbarkeit von Daten aus der öffentlichen Verwaltung. Die über Portale mit intelligenten Such- und Zugriffsmechanismen erreichbaren Daten fordern am Anfang qualitätsgesicherte und verlässliche Daten*“. Es werden Schlussfolgerungen für die Etablierung einer Data Governance Strategie gezogen und Empfehlungen für eine (schrittweise) Etablierung ausgesprochen: für die POP-Dioxindatenbank des Bundes und der Länder und für weitere gleichartige Datenbestände im Umweltbundesamt.

Über die eigenen bzw. kooperativen Verwaltungs- und Forschungsaufgaben sowie die wissenschaftliche Beratung, Unterstützung und Koordination von Maßnahmen des Bundes hinaus gewinnen folgende Aspekte zunehmend an Bedeutung:

- die Aufklärung der Öffentlichkeit in Umweltfragen,
- die Veröffentlichung von Umweltdaten.

Ein spezieller Treiber sind die Initiativen und besonderen Verpflichtungen in Politik und Verwaltung zu **Open Data**. Auf der Grundlage des Regierungsprogramms „Vernetzte und transparente Verwaltung" setzt sich die Bundesregierung das Ziel, bis 2013 eine bundesweite Plattform für Open Data zu schaffen. Die Herstellung von Datenqualität vor der öffentlichen Bereitstellung liegt im genuinen Interesse von (Umwelt-) Verwaltungen und ist Element einer aktiven Open Data Strategie, einer **Open Data Governance**.

Die Veröffentlichung von Open (Government) Data im Rahmen der Wertschöpfungskette stellt stets eine verantwortungsbewusst vorgenommene kontrollierte Bereitstellung von Daten und Informationen dar. Sicherzustellen sind unter anderem:

- die Vertrauenswürdigkeit der Informationen als glaubwürdig und belastbar,
- die Eignung für den vorgesehenen Verwendungszweck bezüglich Relevanz und Angemessenheit,
- die Integrationsfähigkeit, um Informationen untereinander in Beziehung zu setzen und somit anreichern zu können,
- die Transformierbarkeit je nach Verwendungskontext,
- die Beschreibbarkeit und Recherchierbarkeit,
- die Schutzwürdigkeit von personenbezogenen und sicherheitsrelevanten Daten.

Daten gewinnen in ihrer Rolle als Open Data an Gewichtigkeit und Relevanz und sind somit als wertvolles Gut zu behandeln. Datenschätze werden gehoben und durch die Zugänglichkeit für ein großes Publikum zwangsläufig aufgewertet.

Mit dieser Öffnung verbunden sind u.a.:

- ein weitgehend unbekannter und in seiner Zusammensetzung und Motivation inhomogener Nutzerkreis,
- die Bereitstellung flexibler Auskunfts- und Analyse-Infrastrukturen für Self-Services, eine zunehmende Sensibilisierung für die Wahrnehmung, neue Formen der Verantwortung und die Reputation / Glaubwürdigkeit als Bundesbehörde in der Öffentlichkeit.

Die Erfüllung dieser Anforderungen passiert nicht im Selbstlauf, sondern erfordert ein System von Regeln, Verantwortlichkeiten, Prozessen und Ressourcen um dieser Erwartungshaltung und Anforderung gerecht werden zu können. Sie erfordern somit die Etablierung einer Data Governance. Data Governance umfasst die Personen, die Prozesse und die Technologien, die zur Verwaltung und zum Schutz des Datenkapitals in einer Organisation benötigt werden, um allgemein verständliche, korrekte, vollständige, vertrauenswürdige, sichere und auffindbare Unternehmensdaten garantieren zu können. Data Governance definiert Regeln, Organisationsstrukturen, Prozesse, Rollen, Datenarchitekturen und Technik und schafft so die allgemein anerkannten und verbindlichen Grundlagen für Datenintegration, Datenqualität, Stammdatenmanagement, Metadatenmanagement und den Datenschutz in einer Organisation.

## 4.2 Rückschlüsse, Bewertungen und Empfehlungen

Eine übergreifende bzw. auf die Dioxin-Datenbank bezogene DQ-Strategie obliegt in erster Linie der Verantwortung des Auftraggebers. Sie muss dort durch Personen und Überzeugungen getragen und an übergeordneten Zielen ausgerichtet sein.

Im Rahmen des durchgeführten DQ-Projektes kann ein Beitrag für eine solche Strategie durch Formulierung von Maßnahmeschwerpunkten und Vorschlägen zur Etablierung einer Data Governance Initiative geleistet werden. Dies erfolgt auf zwei Wegen:

- Auswertung der in den Arbeitspaketen #1 und #2 identifizierten systematischen Defektmuster und Proklamierung als Arbeitsschwerpunkte im Rahmen einer Strategieumsetzung.
- Verallgemeinerung von Erfahrungen aus der Projektdurchführung und Übertragung als Best-Practice-Erfahrungen auf künftige Vorhaben zu Data Quality und Data Governance.

In den nachfolgenden Abschnitten werden dazu ausgewählte Aspekte in loser Folge dargestellt und diskutiert. Sie sollen als Anregungen verstanden sein. Konkret projektbezogene Feststellungen und Ableitungen sind durch einen farbigen Hintergrund hervorgehoben.

### 4.2.1 Datenqualitätskriterien

Datenqualität kann pragmatisch als die Eignung von Daten für den vorgesehenen Verwendungszweck oder als System von Datenqualitätsdimensionen wie nachfolgend dargestellt beschrieben werden.

Die formalisierte Sicht auf Datenqualität wird über entsprechende Kriterien vermittelt, die vor dem Beginn einer Datenqualitätsprüfung gleichberechtigte Kandidaten darstellen.

Ein Beispiel für ein System von Datenqualitätskriterien stellt die von der DGIQ (Deutsche Gesellschaft für Informations- und Datenqualität) entwickelte Systematik dar, die auf den Arbeiten von Wang und Strong beruht (die Erläuterungen zu den einzelnen DQ-Kriterien finden sich u.a. auf www.dgiq.de).

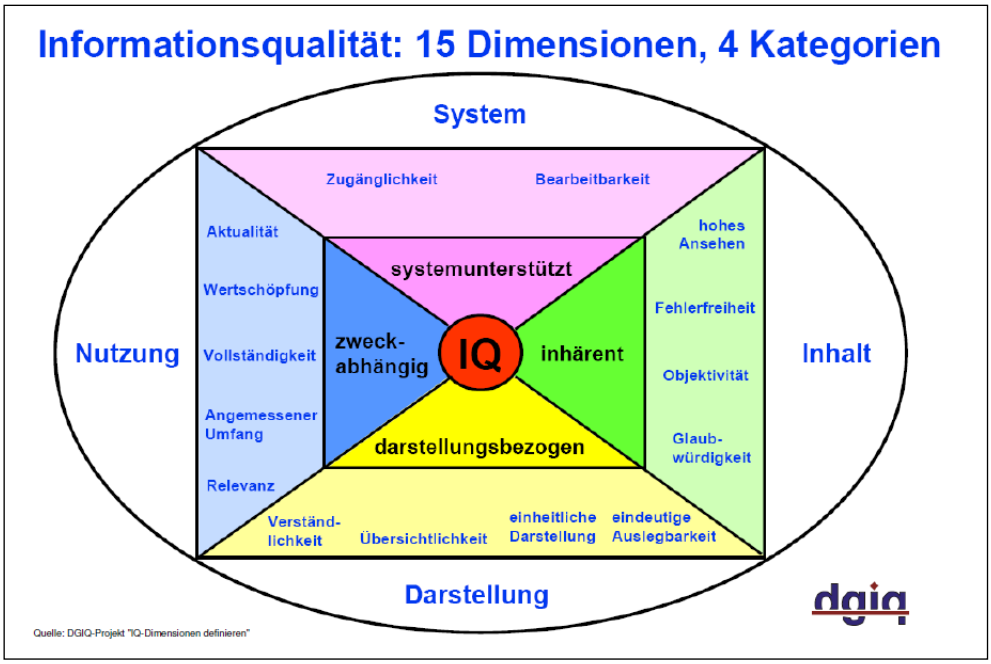



Abbildung 10: Datenqualitätsdimensionen, Quelle: DGIQ

Sind bei der Erzeugung und Verteilung von Informationen innerhalb einer Organisation oder in definierten Beziehungen zwischen verschiedenen Organisationen die Informationsempfänger in der Regel bekannt, ist in dem Umfeld von Open Data das Empfängerspektrum breit gestreut. Der vorgesehene Verwendungszweck und vorrangige Qualitätsmaßstäbe sind somit nur schwer eingrenzbar. Eine Reduktion von Datenqualität auf die Sicherstellung ausgewählter Einzelaspekte ist daher nicht zulässig. Im Kontext von Open Data ist somit der 360°-Ansatz zu verfolgen.

Was ist gute Datenqualität und was ist schlechte? Die Frage kann einerseits aus dem Gefühl und der Erfahrung heraus beantwortet werden – in Kenntnis bekannter Defizite und der mit ihnen verbundenen Nachteile sowie den resultierenden Aufwendungen zu ihrer Überwindung. Besser ist in jedem Fall ein systematischer Ansatz durch regelmäßige und wiederholte Messung des Datenqualitätsniveaus.

Im DQ-Projekt zur POP-Dioxin-Datenbank wurden rückblickend vorrangig folgenden DQ-Kriterien betrachtet:

- Fehlerfreiheit,
- Vollständigkeit,
- Einheitliche Darstellung (eingeschränkt),
- Eindeutige Auslegbarkeit (eingeschränkt),
- Wertschöpfung (eingeschränkt),
- Relevanz (eingeschränkt),
- Aktualität (eingeschränkt).

Eine bewusste Fokussierung und Priorisierung wurde vorab allerdings nicht vorgenommen, die dargestellte Auswahl hat sich letztendlich eher intuitiv ergeben.

I.d.R. sind aber die Prüfung auf Vollständigkeit und Fehlerfreiheit die Standardfragestellungen. Hierfür kamen im Projekt die für die Beantwortung dieser Fragestellungen bewährten Standardinstrumente zur Anwendung:

- Data Profiling
- Formulierung und Prüfung des Erfüllungsgrades von Business Rules

Durch das Data Profiling und die damit verbundenen qualitätsprüfenden Metriken wird eine Vielzahl von Defekten transparent gemacht und identifiziert. Ein Beispiel ist der Belegungsgrad einzelner Felder, mit der Pflichtfeldern geprüft werden können. Ein weiteres Beispiel ist die insbesondere für Schlüsselfelder relevante Uniqueness.

Im durchgeführten Projekt wurden die Profiling-Ergebnisse einer Sichtung bzgl. erkennbarer Auffälligkeiten unterzogen. Bei diesem Verfahren werden jedoch noch nicht alle Potentiale des Data Profilings ausgeschöpft. Die ermittelten Wertebereiche, Wertemuster und Werteverteilungen gewinnen insbesondere dann an Wert, wenn sie vereinbarten Sollvorgaben gegenüber gestellt und somit bewertet werden können.

Eine Diskussion zu den Business Rules findet sich im nachfolgenden Abschnitt.

Aus Blick einer Open Data Strategie müssen in zukünftigen Folgeprojekten zum POP-Dioxin-Datenbestand entsprechend den Eingangsbemerkungen verstärkt auch folgende DQ-Kriterien in den Fokus genommen werden, die bisher nicht angemessen betrachtet wurden:

- Aktualität (intensiviert),
- Wertschöpfung (intensiviert),
- angemessener Umfang,
- Relevanz (intensiviert),
- Verständlichkeit,
- Übersichtlichkeit,
- Glaubwürdigkeit,
- Zugänglichkeit.

Einzelne dieser Kriterien werden in dem Abschnitt „Fokuserweiterungen“ reflektiert.

#### 4.2.1.1 Datenqualitätsmessung und die Rolle von Business Rules

Wie kann eine Messung und Quantifizierung vollzogen werden? Letztendlich steht immer die Frage im Mittelpunkt, in welchem Umfang die Erwartungen an die Daten erfüllt sind. Um diese Frage zu beantworten, werden Erwartungsregeln gesammelt und so formuliert, dass ihre Einhaltung mit technischen Hilfsmitteln überprüft werden kann. Wichtig dabei ist, auch scheinbar triviale Regeln aufzunehmen, weil bereits bei diesen oftmals unerwartete Regelverstöße aufgedeckt werden können. Beispiele dafür sind: „Das Feld xyz einer Datenbanktabelle muss immer gefüllt sein“ oder „Der vorgefundene Wert muss in einem bestimmten Intervall liegen oder einer vorgegebenen Notation entsprechen“.

Die Erfüllung dieser einfachen Erwartungen stellt die Voraussetzung für die Sicherstellung einer Basisqualität dar, auf deren Basis komplexere Regeln aufgestellt und überhaupt erst sinnhaft überprüft werden können. Beispiele für derartige Regeln sind spalten-, zeilen- und tabellenübergreifende Wenn-Dann-Beziehungen sowie Fragen des Matchings von Daten. Matching innerhalb einer Tabelle ist oftmals die Basis für die Identifikation und Bereinigung von Dubletten, Matching-Untersuchungen zwischen Tabellen dienen zumeist der Sicherstellung der Integrationsfähigkeit von Daten.

Je Regel lässt sich ermitteln, ob oder ob sie nicht erfüllt und wie groß das Maß der Abweichung ist. Diese Metriken können zu einem Datenqualitätsindex kombiniert und in ihrer Veränderung über die Zeit beobachtet werden. Eine stetige Verbesserung

im Zusammenhang mit eingeleiteten Verbesserungsmaßnahmen manifestiert die Korrektheit des eingeschlagenen Weges und der hierbei getätigten Investitionen in Datenqualitätsinitiativen.

Was tun, wenn Regelverstöße, das heißt Abweichungen von den Erwartungen ermittelt werden? Zwischen Fachabteilung und IT ist je Regelverletzung eine Bewertung sowie – in Kenntnis geeigneter und finanzierbarer Maßnahmen – eine Zielvereinbarung vorzunehmen:

1. Die festgestellte Regelverletzung ist geschäftskritisch und nicht akzeptabel. Ziel ist die 100%ige Regelerfüllung.

2. Die Regelverletzung ist mehr oder weniger geschäftskritisch, eine 100%ige Regelerfüllung entweder nicht erforderlich oder nicht realistisch bzw. nicht wirtschaftlich erreichbar. Ziel ist ein höherer Erfüllungsgrad < 100%.

3. Die Regelverletzung ist zwar festgestellt, für das operative bzw. dispositive Geschäft ohne Belang. Die Einleitung von Verbesserungsmaßnahmen ist nicht vorzusehen. Eine Optimierung kann ggf. in der Richtung vorgesehen werden, eine zwar vorhandene, in der Praxis aber nicht verwendete Information aus den Systemen zu entfernen. Die Vermeidung von ungenutzten Datenfriedhöfen ist eine durchaus qualitätssteigernde und effizienzfördernde Verbesserung.

Auch für das betrachtete Projekt war die Arbeit mit Business Rules von großem Wert. Die dabei gemachten Erfahrungen sollten unbedingt in zukünftige vergleichbare Vorhaben einfließen:

- Bei der Formulierung der Business Rules muss darauf geachtet werden, dass sie von allen Beteiligten zweifelsfrei in gleicher Art und Weise verstanden werden.

- Business Rules sollen generell als Soll-Anforderungen definiert werden (nicht als Vorgabe des Überprüfungsweges, wie z.B. „Erzeugung eines Reports, der alle Datensätze mit … enthält“).

- Generische Formulierung sind nicht nur zulässig, sondern aus Effektivitätsgründen gezielt zu verwenden (z.B. „Für alle Datumsfelder gilt: …“).

- Als gut zu bewerten ist die im Projekt gewählte Möglichkeit, für die Präsentation der Prüfergebnisse unterschiedliche Darstellungsformen zu wählen. So wurde z.B. für die Darstellung der Ergebnisse von Häufigkeitsanalysen auf eine tabellarische Ausgabe verzichtet und eine durchgängige Visualisierung in Form von Balkendiagrammen gewählt. Andere angewandte grafische Darstellungsformen sind Scatterplots und Box-Plots.

- Für die Pflege der Business Rules und für die Ergebnisse aus deren (wiederholter) Überprüfung empfiehlt sich die Einrichtung eines zentralen Repositories, in dem die Business Rules identifiziert, kategorisiert und beschrieben sind. Damit wird ein definierter Standort zur Abstimmung und Verabschiedung von Business Rules geschaffen (mit all seinen Vorteilen gegenüber einem in Form von E-Mails, Telefonaten, Protokollen usw. geführten und im Nachgang nicht mehr nachvollziehbaren Definitionsprozess). Zu jeder Business Rule wird gleichzeitig der zu implementierende Algorithmus zu deren Überprüfung im Datenbestand mit abgelegt. Dieser kann als weiteres Element zur Qualitätssicherung der Business Rules herangezogen werden (Fehlerfreiheit, Eindeutigkeit, Verständlichkeit). Gleichzeitig wird damit die Wiederholbarkeit von Regelüberprüfungen unter Einhaltung gleicher Voraussetzungen sicher gestellt.

#### 4.2.1.2 Maßnahmen zur Verbesserung der Datenqualität

Verbesserungsmaßnahmen zielen grundsätzlich in zwei Stoßrichtungen:

1. Datenbereinigung in den vorhandenen Datenfonds durch Korrektur oder regelbasierte Transformation an der Quelle oder in abgeleiteten Datenbeständen.

2. Präventive Prozess- und Systemverbesserungen zur Vermeidung des erneuten Auftretens von Abweichungen.

Innerhalb wirtschaftlicher Einheiten wird der Umfang und die Tiefe dieser Maßnahmen oftmals durch den erwarteten Return on Invest (ROI) bestimmt. Allerdings nicht als alleiniges Kriterium – sobald das Image des Unternehmens und das Vertrauen in seine Produkte und Dienstleistungen betroffen sind, haben diese Effekte Vorrang. Das deckt sich mit dem Anliegen von Behörden, die Daten für die Öffentlichkeit bereitstellen – auch sie möchten als vertrauenswürdig und verlässlich wahrgenommen und den hoch gesetzten Ansprüchen bei der Erfüllung des politischen Auftrages gerecht werden. Herausforderung ist die Sicherstellung von

- Erwartungsgerechtigkeit,
- Vertrauenswürdigkeit
- Belastbarkeit
- Regelkonformität (Compliance).

Die Durchführung von Datenbereinigungsmaßnahmen war nicht Gegenstand und Zweck des durchgeführten Projektes. Aus den Untersuchungsergebnissen lassen sich jedoch folgende Empfehlungen ableiten, an denen sich auch die in späteren Abschnitt dokumentierten Maßnahmenvorschläge ausrichten:

- Eine weiterführende Untersuchung daraufhin, inwieweit systematische oder zufällige Fehler vorliegen, um Hinweise für Bereinigungsstrategien zu erhalten.
- Nach der Durchführung einmaliger Datenbereinigungen Erfolgsmessung durchführen und regelmäßig wiederholen.
- Alle abhängigen und abgeleiteten Größen identifizieren und neu berechnen.
- Zur Nachvollziehbarkeit von Datenkorrekturen eignet sich das Instrument des Audit Trails. Wenn die Möglichkeit besteht, sollten im Rahmen der Traceability und zur Revisionssicherheit TimeStamps und Aktiv-Flags genutzt werden, um die zeitliche Gültigkeit von Datenwerten zu kennzeichnen. Dadurch wird u.a. verhindert, dass in der Vergangenheit als sachlich richtig freigegebene Berichte nicht nachträglich als falsch eingestuft werden müssen. Die Korrektheit entsprechend Informationsstand zum Erzeugungszeitpunkt kann jederzeit belegt werden.
- Fehlerhafte Altwerte sollten niemals gelöscht, sondern lediglich durch Kennzeichnung von der weiteren Verwendung ausgeschlossen werden. Im einfachsten Fall, falls die Implementierung von Gültigkeitszeiträumen ausscheidet, durch Kopieren in zusätzliche Spalten.
- Fehlende Werte können unter Umständen auch durch Schätzwerte oder durch Defaults ersetzt werden.
- Im Rahmen der durchgeführten Untersuchungen nicht im Vordergrund, aber wichtige Verfahren zur Datenbereinigung sind die Standardisierung, die De-Duplizierung und das Parsing.
- Mit den Stammdaten beginnen, gefolgt von den Metadaten und den Bewegungsdaten.

Im Hinblick auf Präventivmaßnahmen ergeben sich folgende begründete Ansätze:

- In einem nächsten Arbeitsschritt sollte auf der Grundlage der vorliegenden Ergebnisse untersucht werden, inwieweit systematische oder zufällige Fehler vorliegen, um Hinweise für Präventionsstrategien zu erhalten. Die Feststellung systematischer Fehler ist die Grundlage für die Formulierung von Präventionsstrategien, die bei den datenliefernden Systemen ansetzen wie einheitliche Vorgaben für Datenerfassung und für Datenformate.

- In diesem Zusammenhang sollten generell für alle Seiten akzeptable Möglichkeiten der Datenqualifizierung „an den Quellen“ vereinbart werden. Aus Sicht der Qualitätssicherung begrüßenswert wären „Ausgangsprüfungen“ als der „Eingangsprüfung“ im UBA vorgelagerte Prüfungen bei den Datenlieferanten beispielsweise Checklisten und Prüfregelen zur Prüfung vor der Datenweitergabe der datenliefernden Organisationen.
- Dabei ist nicht nur an technische „Prüfregeln“ zu denken, sondern es sollten insbesondere auch prozessurale Verbesserungsmöglichkeiten – wie Rollen und Verantwortungen, Vereinbarungen zu Prozessschritten und zu Zeitplänen - diskutiert werden.
- Durchführung von Bereinigungsmaßnahmen in vorgelagerten Quellsystemen, wenn späterer erneuter Import von Daten nicht ausgeschlossen ist (ggf. auch nur formal aus Gründen der Konsistenz).
- Implementierung eines kontinuierlichen DQ-Monitorings mit DQ-Index und Dashboard. Der DQ-Index ist das Gesamtmaß aller Regelerfüllungsgrade. Er kann über die Zeit verfolgt werden und gibt Auskunft über Fortschritte in der Datenqualifizierung des beobachteten Datenbestandes.
- Datenschnittstellen absichern.

#### 4.2.1.3 Fokuserweiterungen

Bereits eingangs wurde auf die Tendenz neuer und erweiterter Nutzerkreise und der damit verbundenen vielfältigen Nutzungsformen verwiesen.

- Der Kreis der Informationsempfänger öffnet sich von dem vormaligen engen Kreis der Fachanwender zunehmend für eine in Ihren Motivationen und Informationsbedürfnissen eher inhomogene, anonyme und in ihrem Umfang unbekannte Gruppe neuen Adressatentyps.
- Interdisziplinäre Vernetzung von Informationen: Einzelne Datenbestände des Umweltbundesamtes sind mit anderen Fachdaten im Hause sowie von externer Quelle in Beziehung zu setzen – genau wie andersherum die Bedürfnisse zur integrierten Verwendung von Daten von der POP-Dioxin-Datenbank aus einem anderen Fachkontext heraus wachsen.

Dies erfordert einen erweiterten Blick auf Datenqualitätskriterien wie Zugänglichkeit, Verständlichkeit, Übersichtlichkeit und einheitliche Darstellung.

Das vorhandene komplexe Datenmodell der POP-Dioxin-Datenbank kommt den Bedürfnissen der (zukünftigen) potentiellen Empfänger (leichte und sichere Informationsgewinnung zur Beantwortung individueller Fragestellungen) nur unzureichend entgegen. Auch entsprechende Abfragetools, die einen von der technischen Abbildung abstrahierten fachlichen Zugang zu den Daten ermöglichen sollen, müssen sich den vorgefundenen Restriktionen beugen.

Beispiele für die vorgefundenen Restriktionen sind:

- Es liegt ein (teil-)normalisiertes zweidimensionales Datenmodell mit einer Vielzahl fachlich-technischer Tabellen zugrunde unter teilweiser Verfolgung eines hierarchischen Ansatzes: Selbst wenig komplexe Regelanfragen (Proben an einem Standort in einem bestimmten Zeitraum zu einem bestimmten Kompartiment) sind nur unter Kenntnis der fachlichen Zusammenhänge und des technischen Datenmodells mit Joins oder unter Verwendung von vordefinierten Views möglich.

Es wird daher angeregt, für diese Zwecke eine aus dem bisherigen POP-Dioxin-Datenbestand abgeleitete Informations- und Recherchedatenbasis zu implementieren, die mit den entsprechenden Fakten- und Dimensionstabellen dem Ansatz einer mehrdimensionalen Datenbank folgt und die dazu erforderlichen Abfragetechniken (OLAP, MDX, Pivot-Tabellen) bereits integriert.

Für die Verlinkung und Vernetzung der darin enthaltenen Informationen sind organisationsübergreifende Metadata Repositories (Data Dictionaries) vorzusehen. Die hierüber adressierten unterschiedlichen Datenbestände müssen dabei zwingend den Konventionen einer Data Governance unterliegen, da ansonsten die Voraussetzungen für eine funktionierende technische und semantische Verknüpfung nicht gegeben sind.

An dieser Stelle nicht weiter diskutiert aber dennoch hervorgehoben wird auf die entscheidende und zentrale Rolle von Metadaten und auch die von Stammdaten für den Erfolg einer Data Governance Strategie verwiesen.

#### 4.2.1.4 Projekterkenntnisse für die Strategieumsetzung

DQ-Projekte im Rahmen von Data Governance Initiativen sollen bzgl. Umfang und Dauer generell überschaubar definiert und abgewickelt werden. Gegenüber großen Projekten mit ungewissem Ergebnis und unvorhersehbarem Aufwand ist ein iteratives Vorgehen mit der Definition abgegrenzter Pilot- und Teilprojekten vorzuziehen. Sie sind Elemente und Bausteine bei Umsetzung einer Data Governance Strategie.

Das durchgeführte Projekt folgt diesem Ansatz durch Abgrenzung und Konzentration auf eine Teilmenge der Daten (benannte Haupttabellen der Datenbank und einzelne Umweltkompartimente) und wenige ausgewählte Fragestellungen mit einem entsprechend limitierten Zeit- und Budgetrahmen. Zu entscheiden ist jedoch noch der Umgang mit den in dem aktuellen Projekt nicht betrachteten verbleibenden Daten.

Solange DQ-Projekte noch keine regelmäßige und standardisierte betriebliche Übung sind, übernehmen sie die Rolle von „Testläufen“, um die besten Vorgehens- und Arbeitsweisen herauszufinden und als gesicherte Erkenntnisse Nachfolgeprojekten zur Verfügung zu stellen. DQ-Projekte sind nicht trivial, es ist daher nicht zu erwarten, dass dabei immer alles „glatt“ läuft. Dies ist kein Ausdruck des Scheiterns, sondern eine zur Erreichung der Vision erforderliche Lernstufe (Lessons Learned).

Das durchgeführte Projekt hat eine ganze Reihe an wertvollen und strategierelevanten Erkenntnissen gebracht, die im nachfolgenden Abschnitt „Herausforderungen und Arbeitsweisen“ im Einzelnen aufgeführt sind.

Die Umsetzung eines Data Governance Programms und von DQ-Initiativen ist ein ressourcenintensiver Prozess, der nicht „nebenbei“ und ohne die entsprechende Qualifikation mit erledigt werden kann. Er erfordert eine Bündelung von Maßnahmen über einen längeren Zeitraum mit entsprechendem Investitionsbedarf. Darin eingeschlossen sind:

- die Entwicklung von Kompetenzen,
- die Motivation und Befähigung von Mitarbeitern,
- die personelle Ausstattung,
- der Planung und Bewilligung angemessener Projektbudgets,
- die Beschaffung geeigneter Tools,
- die Etablierung langjähriger Partnerschaften mit verlässlichen Beratungs- und Projektpartnern.

Das durchgeführte Projekt wurde in der Definition stark auf einen relativ kleinen Untersuchungsgegenstand abgegrenzt. Dies war im Ansatz richtig, hat sich in der Durchführung bzgl. Budgetierung letztendlich aber doch als zu klein dimensioniert erwiesen. Es fehlten Reserven für die erweiterte Projektkommunikation, für explorative und iterative Vorgehensweisen (ergebnisabhängige schrittweise Nachjustierung von Prüfregeln) sowie für situationsbezogene Vertiefungen bzw. Erweiterungen von Fragestellungen. Einzuplanen sind bei diesem inhaltlich neuen Typus von Projekt ebenfalls die entsprechenden Lernkurven.

Im Rahmen einer Strategieumsetzung wird schrittweise von höheren Aufgaben- und Projektvolumina auszugehen sein. Die Erprobung valider und im Sinne der Zielerreichung begründeter Zeit- und Kostenabschätzungen im zunächst kleinen Rahmen

bewahrt daher vor unangenehmen Überraschungen nicht eingehaltener Kostenrahmen oder verfehlter Zielerreichungen in der Zukunft.

#### 4.2.1.5 Herausforderungen und Arbeitsweisen

**Das durchgeführte Vorhaben hat gezeigt, dass DQ-Projekte ein genaues und sorgfältiges Arbeiten auch im Detail erfordern.**

- Dies hat sich insbesondere bei der Definition und der Abstimmung der Business Rules für die Ermittlung von DQ-Metriken gezeigt. Uneindeutige und unvollständige Formulierungen haben trotz des vorher investierten hohen Abstimmungsaufwandes dazu geführt, dass die den Vorgaben folgenden und somit formal korrekten Prüfungen unerwartete bzw. ungewollte Ergebnisse geliefert haben.

- Im Ergebnis wurden ungeplante Aufwände durch neuerliche Abstimmungsaufwände beim Auftraggeber und vertraglich nicht abgedeckte Change Requests zur Neuimplementierung, Durchführung, Dokumentation und Bewertung von Prüfungen beim Auftragnehmer erforderlich.

- DQ-Projekte können nicht „nebenbei“ durchgeführt werden, sondern erfordern eine Konzentration auf die Aufgabe und eine weitgehende Störungsfreiheit von den sonstigen Tagesaufgaben. Es ist daher angeraten, periodisch im Projektzeitraum und in Abstimmung mit dem Projektplan die Mitarbeiter z.B. tagesweise weitestgehend von anderen Aufgaben freizustellen. Durch das Management bzw. die Projektleitung sind die erforderlichen Zeitkontingente und organisatorischen Freiräume sicherzustellen.

- Es wird empfohlen, bei der Definition von Business Rules zukünftig folgende Verbesserungsansätze zu berücksichtigen:

- Vor Freigabe einer Regel Feedback durch Dritte, wie der Inhalt der Regel verstanden wurde
  - Untersetzung und Verständnisunterstützung durch Beispiele
  - Dokumentation des vorab erwarteten Ergebnisses im Ergebnis der Regelüberprüfungen

**Eine funktionierende und effektive Kommunikation ist in DQ-Projekten unverzichtbar.**

- In einem DQ-Projekt arbeiten Experten unterschiedlicher Fachgebiete interdisziplinär zusammen. Dies muss in der Kommunikation aktiv berücksichtigt werden (welche Informationen benötigt mein Gegenüber, wie muss ich sie transportieren, so dass sie verstanden und eingeordnet werden können). Die fachlichen und die technischen Experten müssen eine gemeinsame Sprache und ein gemeinsames Verständnis finden. Konsequenzen wären ansonsten Zeitverluste, fehlerhafte Arbeitsergebnisse und Zusatzaufwände.

  Im durchgeführten Projekt hat sich gezeigt, dass die Kommunikationswege und die Verantwortlichkeiten im Kommunikationsprozess nicht immer eingehalten wurden und auch der Inhalt der transportierten Informationen unterschiedlich genau war.

  Zu überdenken sind auch die Kommunikationsinstrumente. Komplexe E-Mails mit einer Vielzahl von Fragestellungen führen auf Dauer dazu, den Überblick zu verlieren. Besser geeignet sind gemeinsame Workshops ( am besten tageweise), in denen vorher kommunizierte Fragestellungen konzentriert diskutiert und entschieden werden können.

  Für den in schriftlicher Form erfolgenden Klärungs- und Abstimmungsprozess von Einzelfragen sollte auch geprüft werden, geeignete Instrumente wie z.B. Jira oder Bugzilla zum Einsatz zu bringen.

**DQ-Projekte erfordern ein ausgeprägt methodisches und strukturiertes Vorgehen.**

- Erforderlich sind dabei klar definierte und abgegrenzte Projektrollen über die gesamte Projektzeit mit der Vermeidung des Wechsels und der Teilung von Verantwortlichkeiten. Der Projektleiter sollte wenn möglich

sowohl die fachliche und als auch die organisatorische Verantwortung aus einer Hand leisten (bei den eher kleineren Projektgrößen effektiver als die Etablierung geteilter Verantwortungen). Auch die Forderungen nach einer fachlich-organisatorisch „formalen“ und durchgängig dokumentierten Durchführung sind sicher nicht geliebt aber notwendig.

### 4.2.2 Empfehlungen I: Etablierung einer Data Governance

Ab einer bestimmten Größe der Organisation muss die Etablierung einer Data Governance über Abteilungsgrenzen hinweg auf der Ebene der Gesamtorganisation unter Einbeziehung externer Informationslieferanten und Informationsempfänger erfolgen.

Die Etablierung geht einher mit Veränderungen in Verantwortlichkeiten, Prozessen und Technologien und ist in ihren wesentlichen Herausforderungen auf folgende Kernbereiche ausgerichtet:

- Organisation
  -> Offene und veränderungsbereite Unternehmenskultur
  -> Mut und Sensibilität zur Definition und Zuweisung neuer Rollen und Verantwortlichkeiten
- Kommunikation
  -> Förderung von Transparenz und Zusammenarbeit
  -> Gemeinsame Abstimmung und fachliche Klärungen möglichst konzentriert in Workshops
- Budgets und Stakeholder
  -> „Chefthema“
  -> interne Vermarktung und Überzeugungen
  -> Angemessene Reaktion auf den Handlungsdruck durch Investition in Budgets und Ressourcen
- Standardisierung und Flexibilität
  -> Flexibilität im Denken und Handeln / im Business und in der Verwaltung
  -> Standards für einheitliche und effiziente Handlungsweisen
  -> Offenheit für Veränderung und schnelle Reaktion aus sich verändernde Anforderungen

Das Vorgehen unterscheidet sich nicht grundsätzlich von dem anderer strategischer Vorhaben:

- Ziele definieren und Nutzen benennen
- Ist-Analyse und Soll-Ist-Vergleich
- Entwicklung eines Vorgehensmodells
- Abwägung von Chancen und Risiken
- Investitionen / Budgets sicherstellen
- Data Governance Programm im Detail entwickeln und planen
- Data Governance Programm umsetzen mit regelmäßiger Erfolgskontrolle

Die Etablierung kann auch durch die Implementierung aufeinander abgestimmter Teilprogramme für einzelne Teilbereiche wie Datenqualität, Master Data Management oder Meta Data Management mit jeweils unterschiedlicher Ausgestaltung erfolgen.

### 4.2.3 Empfehlungen II: Best Practices

- Klarheit in der Motivation (Pflichterfüllung versus Bedürfnis)
- Einbettung in Strategien (Data Governance Initiative) und Sponsoren finden
- Start small and focused („Klasse statt Masse“)
- Messbarkeit sicherstellen (-> Erfüllungs- bzw. Abweichungsgrad pro Regel -> DQ-Index über die Gesamtheit)
- Immer mit dem Data Profiling beginnen (unverzichtbar!) und Erwartungen vorab formulieren
- Überprüfbare Erwartungen an die Daten formulieren (auch wenn sie noch so trivial erscheinen) => Aufstellen von einfachen bis komplexen Regeln (unverzichtbar!)
- Repository für Definition und lfd. Überwachung der Einhaltung von Business Rules implementieren
- Zunächst stets Basisqualität und Vergleichbarkeit herstellen bzw. simulieren (s. temporäre Korrektur die BR 2 verletzender fehlerhafter Hoch- und rechtswerten durch Multiplikation mit 10).
- Als ein wichtiges Kriterium von Basisqualität die Eindeutigkeit von Daten durch Standardisierung, De-Duplizierung und Parsing herstellen.
- Grenzen von Excel und SQL akzeptieren
- Tools für hochwertige unscharfe Verfahren (Parsing, Standardisierung, Matching) einsetzen
- Weiterhin auch visuelle / manuelle (Prüf-)Tätigkeiten vorsehen
- In der Lage und gewillt sein, Entscheidungen zu treffen
- Zusammenarbeit IT, Fachabteilungen, Data Stewards organisationsübergreifend organisieren
- Erfüllung von Erwartungshaltungen bestätigen / Defizite identifizieren, quantifizieren und bewerten
- Realistische sowie bezahlbare Verbesserungsziele und –maßnahmen setzen
- Im Ergebnis jeder durchgeführten Maßnahme - erneut messen

### 4.2.4 Weitere Schritte ...

#### 4.2.4.1 ... speziell zu POP-Dioxin-Datenbank

- Genauere Untersuchung der Natur und der Ursachen identifizierter Abweichungen um geeignete (wenn möglich automatisierbare) Bereinigungsstrategien abzuleiten.
- Weiterführende Ursachenermittlung.
- Definition und Umsetzung von Bereinigungsmaßnahmen zu den bereits vorliegenden Ergebnissen (ansonsten das gesamte Vorhaben ohne Effekt und die bisher bereitgestellten Mittel umsonst investiert).
- Identifikation von DQ-relevanten Fragestellungen zu der bisher nicht untersuchten POP-D-„Restdatenmenge“.

- Erhöhtes Augenmerk auf eine gesicherte Basisqualität (Beispiel: Sowohl korrekte Hoch- und Rechtwerte als auch gleichermaßen fehlerfreie Messergebnisse sind unabdingbare Voraussetzung, um eine aussagekräftige und belastbare HotSpot-Analyse durchführen zu können – im anderen Fall werden Verfälschung oder Unvollständigkeit provoziert).
- Planung, Beantragung und Bewilligung deutlich höher dimensionierter Projektbudgets, da ansonsten die Aktivitäten Stückwerk bleiben.
- Weiterführende und kontinuierliche DQ- und Prozessberatung im POPD-Umfeld.

#### 4.2.4.2 ... zur Entwicklung und Umsetzung einer Strategie

- Data Governance zur Chefsache und gleichermaßen zur gemeinschaftlichen Aufgabe machen.
- Daten als wertvollen Rohstoff ansehen, der zu veredeln ist.
- Ansprache und Sammlung von „Gleichgesinnten“:
  - ➔ Erfahrungsaustausch über verschiedene Fachdomainen hinweg,
  - ➔ die jeweils spezifischen Kompetenzen und Sichten einbringen.
  - ➔ Identifikation von Problemschnittmengen bzw. von direkten Datenausaustauschbeziehungen unter dem DQ-Aspekt.
  - ➔ Vision einer Data Governance entwickeln.
- Gremien etablieren (Arbeitsgruppen, Fachausschüsse).
- Organisationseinheiten und Datenqualitätsprozesse sowie der fachlichen Verantwortung durch Datenqualitätsbeauftragte wie Data Stewarts etablieren.
- Organisationsweites DQ-Monitoring implementieren.
- Vorhaben methodisch und steuerungstechnisch mit Instrumenten wie Balanced Scorecards begleiten.
- Langfristig ausgerichtete Partnerschaften mit externen Beratungs- und Projektpartnern zur schnellen und einfachen Mobilisierung bereits vorhandener Prozess- und DQ-Expertise etablieren.
- Interner Aufbau von Kompetenzen, Befähigung von Mitarbeitern, Sicherung der personellen und finanziellen Ausstattung.