



# **Datenintegration durch semantische Normalisierung**

**UFOPLAN 3712 12 100 „Linked Environment Data“**

**Joachim Fock, Maria Rüther (Umweltbundesamt)**

**Thomas Bandholtz (innoQ Deutschland GmbH)**

# Agenda

- Worum geht es?
- Das Linked-Data 5-Star Ranking
- seine Erweiterung durch semantische Normalisierung
- SDMX
- Data Cubes Vocabulary (qb)
- Simple Knowledge Organisation System (skos)
- Semantische Normalisierung am Beispiel von
  - Umweltprobenbank
  - Internationale Kommission zum Schutz des Rheins
  - EEA Waterbase Rivers
- Was haben wir davon?

# Worum geht es?

# Datenintegration – same old story again

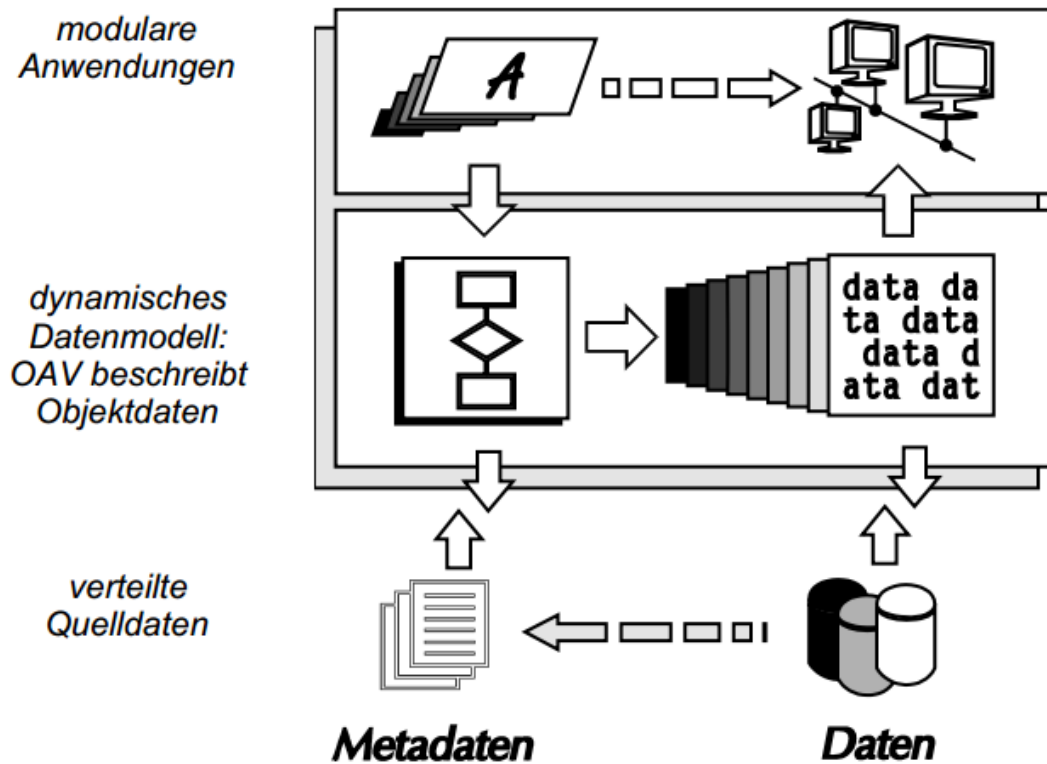





Abb. 5 Systemmodell eines integrativen IS

Fachübergreifende Integration von Umweltdaten. In: Integration von Umweltdaten, 2. Workshop, 2.-4. Februar 1994, Schloss Dagstuhl.

# Drei Ebenen der Datenintegration

- organisatorisch 
- methodisch 
- technisch 

Gerlinde Knetsch: Medienübergreifendes Monitoring. In: RUNDBRIEF DES FACHAUSSCHUSSES FÜR UMWELTINFORMATIK. FEB 2013.

# Das Linked Data 5 Star Ranking und seine Erweiterung durch semantische Normalisierung

# Linked Data 5-Star Ranking



Available on the web (whatever format) *but with an open licence, to be Open Data*



Available as machine-readable structured data (e.g. excel instead of image scan of a table)



as (2) plus non-proprietary format (e.g. CSV instead of excel)



All the above plus, Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff



All the above, plus: Link your data to other people's data to provide context

<http://www.w3.org/DesignIssues/LinkedData.html>



# Semantische Normalisierung

\*\*\*\*\* 5 Sterne plus ...

- gleich strukturierte Daten nutzen dasselbe RDF Schema, z.B. Data Cubes Vocabulary (qb)
- gleich bedeutende Datendimensionen nutzen dieselben Konzepte, z.B. Simple Knowledge Organisation System (skos)

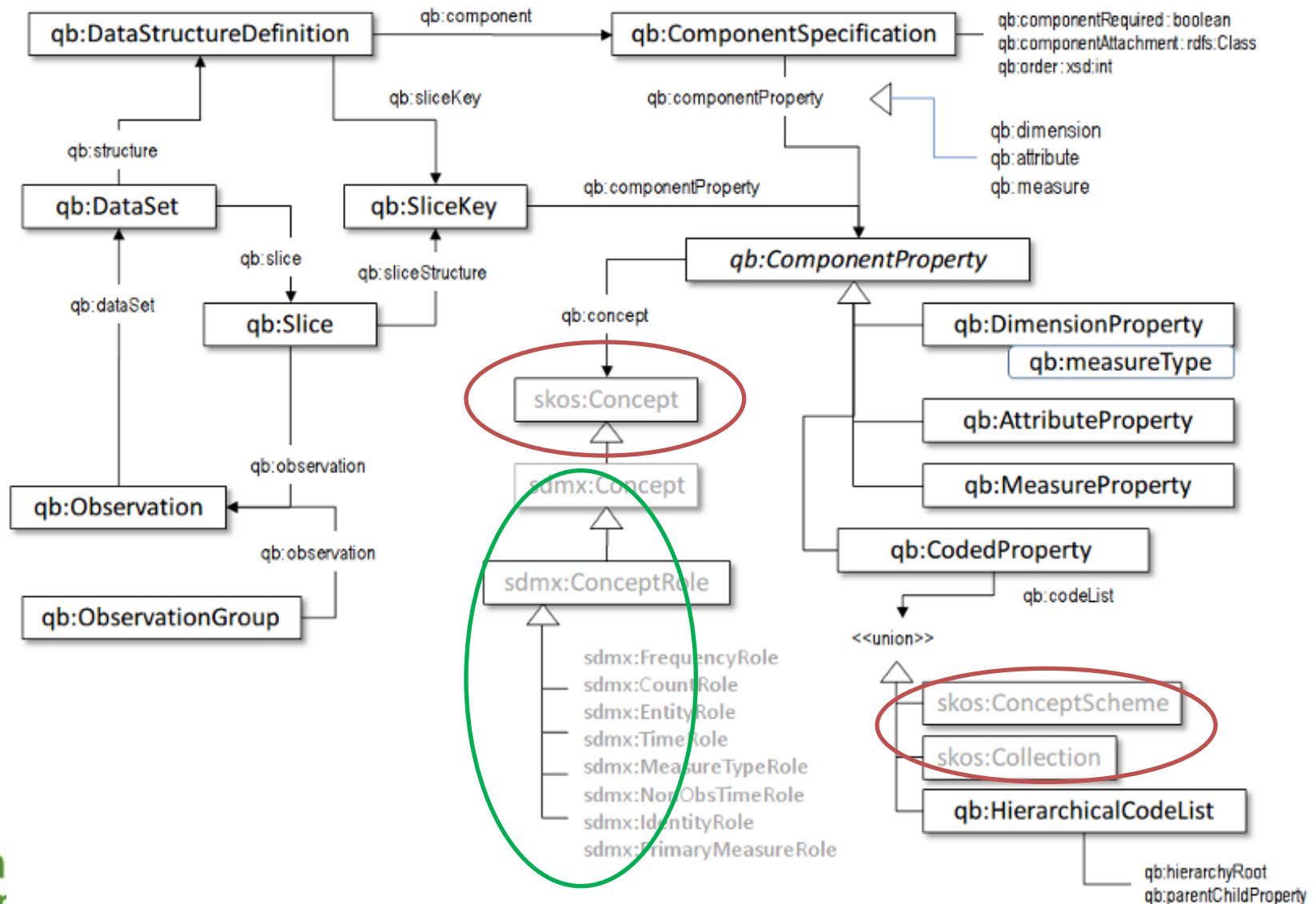


# SDMX, Data Cubes Vocabulary (qb) und Simple Knowledge Organisation System (skos)

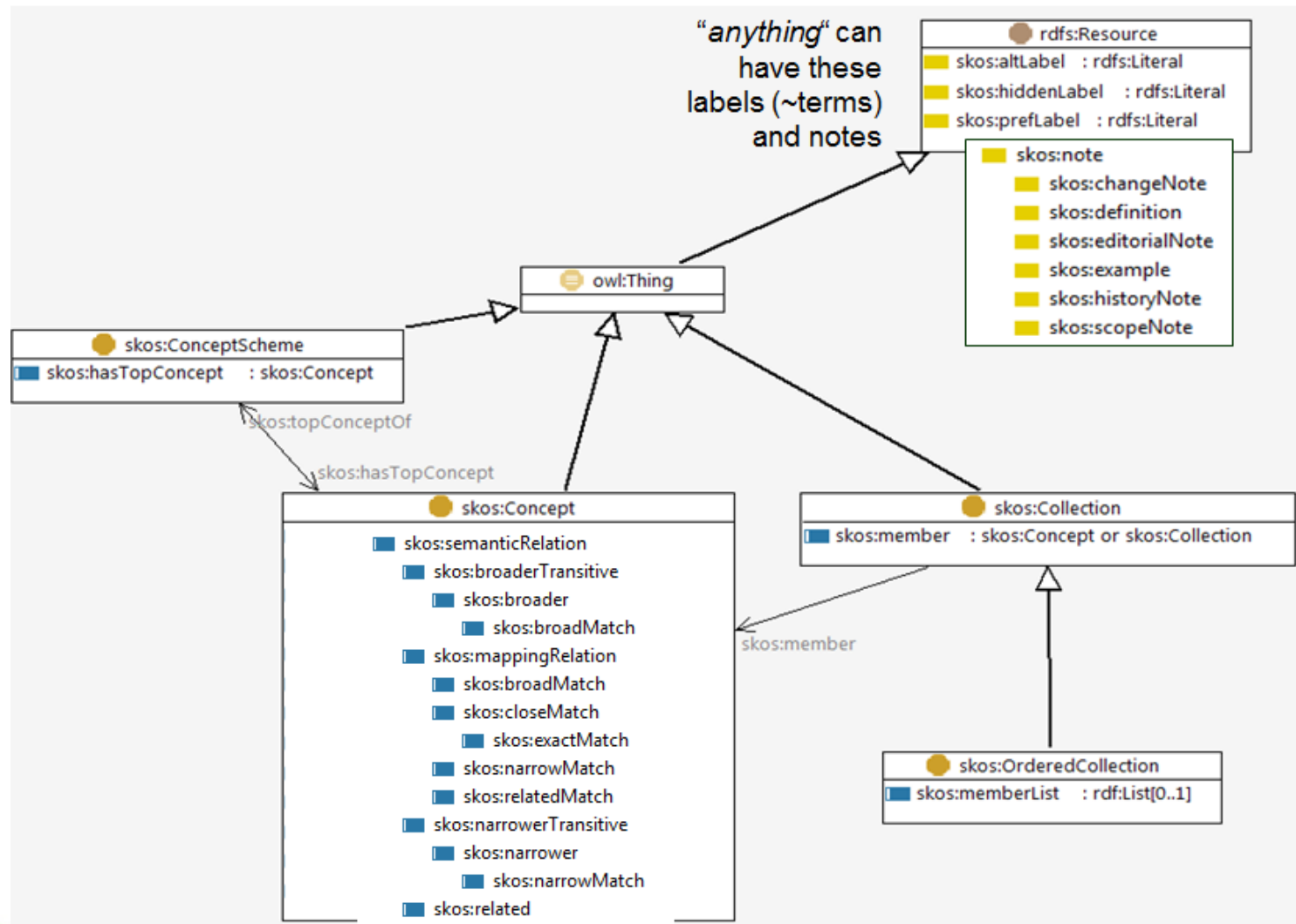
# Statistical Data and Metadata Exchange

- „SDMX is an initiative to foster standards for the exchange of statistical information.”
- European regulation (EU408, 2011) “concerning statistics on pesticides, as regards transmission format”:
- “Member States shall transmit the statistical data on the placing on the market of pesticides as described in Annex I to Regulation (EC) No 1185/2009 using the statistical data and metadata exchange (SDMX) format. The data shall be transmitted or uploaded by electronic means to the single entry point for data at Eurostat.”
- UBA Fachbereich IV 1 Internationales und Pestizide

# Data Cubes Vocabulary (qb)



# Simple Knowledge Organisation System



# Semantische Normalisierung am Beispiel von

- Umweltprobenbank (UPB)
- Internationale Kommission zum Schutz des Rheins (IKSR)
- EEA Waterbase Rivers (WbR)

# Gemeinsamkeiten

- alle beschreiben langfristig das Vorkommen von potentiellen Schadstoffen an gleich bleibenden Orten
- IKSR und WbR beschreiben Gewässerproben
- UPB entnimmt Proben aus unterschiedlichen Medien, darunter auch von Fischen aus denselben Gewässern
- Es gibt Überschneidungen hinsichtlich der Orte und der Analyte

# Unterschiede

- konkrete Datenmodelle (hier als CSV Export)
  - ▀ Vereinheitlichung auf das Data Cubes Format
- Kodierungen / Benennungen
  - ▀ URI-Verweise auf vereinheitliche, gemeinsame Konzepte

<https://internal.innoq.com/exchange/users/fnd/led/eea.ttl>

# Diversität der Tabellenstrukturen

IKSR	UPB	WbR	LED
Probenahmestelle	Probenahmegebiet	NationalStationID	dim. location
(implizit)	Probenart	(implizit)	dim. observedMedia
Parameter	Analyt	Determinand	dim. analyte
-	Messeinheit	Unit	att. unit
-	Extraktionsmethode	-	att. method
Maßeinheit	Statistischer Parameter	spezifische Spalten	spezifische measureProperties
Jahr	(alle Jahre in einer Zeile)	Year	dim. time
Turnus	-	AggregationPeriod	att.cycle
Sonderzeichen	-	-	att.signed
Wert	(betitelt mit dem jeweiligen Jahr)	-	spezifische measureProperties



# normalisierte Beobachtungs-Datensätze

## UPB

**ubp:4711** a qb:observation ;  
qb:dataset ledds:**ubp-exposure** ;  
ledds:observedMedia ledcs:**\_10221** ;  
ledds:analyte ledcs:\_10082 ;  
ledds:location ledcs:\_10122 ;  
ledds:time 1995 ;  
ledds:mean **2.1** ;  
ledds:unit „mg/kg TG“ ;  
(...) .

## IKSR

**iksr:0815** a qb:observation ;  
qb:dataset ledds:**iksr-exposure** ;  
ledds:observedMedia ledcs:**water** ;  
ledds:analyte ledcs:\_10082 ;  
ledds:location ledcs:\_10122 ;  
ledds:time 1995 ;  
ledds:mean **7.4** ;  
ledds:unit „mg/l“ ;  
(...) .

# normalisierte Konzepte

ledcs:[analytes](#) a skos:ConceptScheme .

**ledcs:\_10082 a skos:Concept ;**

skos:prefLabel "Cobalt"@de ;

skos:altLabel "Cobalt (Co)"@de ;

skos:inScheme ledcs:[analytes](#) .

ledds:analyte a qb:DimensionProperty ;

skos:prefLabel "Analyt"@de ;

skos:altLabel "Parameter"@de ;

skos:altLabel „Determinand"@en ;

qb:codeList ledcs:[analytes](#) .

# Was haben wir davon?

# Übergreifende Recherche-Dimensionen

- jedes Datenattribut, das mit einem skos:conceptScheme hinterlegt ist, entspricht einer Dimension
- Substanzen
- Raumbezug
- Spezies
- allgemeines Schlagwort
- Untersuchungsmedium
- ...
- Zeit (ISO 8601 statt skos:)

# Schritte der übergreifenden Recherche

1. Auswahl einer Dimension
2. Auswahl eines Konzepts dieser Dimension
3. -> Datenbanken, die dieses Konzept verwenden
4. Auswahl weiterer Dimensionen und Konzepte
5. -> Datenbanken, die alle gewählten Konzepte verwenden
6. Recherche abschicken

# Anzeige der Ergebnisse, z.B.

- gemeinsame tabellarische Darstellung von gefundenen Messdaten mehrerer Quellen (für aller normalisierten Attribute)
- gemeinsame geographische Darstellung der Orte
- graphische Trendvergleiche

# Zusammenfassung

- Mit SDMX, Data Cubes und SKOS können wir verteilte Beobachtungsdaten medienübergreifend strukturieren
- Linked Data Technologie ermöglicht die integrierte Navigation und Recherche
- Europäische Regulierungen verlangen SDMX, auch im Umweltbereich
- Offene Daten vernetzen!

# - Ende der Präsentation -

## **Joachim Fock**

Umweltbundesamt (de)  
Wörzlitzer Platz 1, 06844 Dessau-Roßlau  
[joachim.fock@uba.de](mailto:joachim.fock@uba.de)

## **Maria Rüther**

Umweltbundesamt (de)  
Corrensplatz 1, 14195 Berlin  
[maria.ruether@uba.de](mailto:maria.ruether@uba.de)

## **Thomas Bandholtz**

innoQ Deutschland GmbH, Krischerstr. 100, 40789 Monheim am Rhein  
[thomas.bandholtz@innoq.com](mailto:thomas.bandholtz@innoq.com)

## **UFOPLAN 3712 12 100 „Linked Environment Data“**