Arbeitskreis Umweltinformationssysteme

Workshop 03./04 Mai 2012 Dresden

Aufbau und Nutzung von Fachterminologie

Joachim Fock Umweltbundesamt

Thomas Bandholtz innoQ Deutschland GmbH

Umwelt Bundes Amt
Für Mensch und Umwelt

innoQ

Agenda

- Vorbemerkung
- Aufbau und Pflege von Fachterminologie
  - Sichtung vorliegender Fragmente
  - Analyse des Corpus
  - Begriffe, Benennungen und ihre Schreibweisen
  - Hierarchie und andere Verwandtschaften
  - Definitionen
  - Formalisieren
  - verfügbar machen
  - Persistenz und Evolution
- Nutzung von Fachterminologie
  - in der Kommunikation
  - Wissensmanagement
  - Automatisches Indexieren
  - Dokumente vergleichen
  - verlinkte Daten

Umwelt Bundes Amt
Für Mensch und Umwelt

innoQ

Vorbemerkung

- basiert auf Projekterfahrungen der Autoren und Anderen seit den 1990er Jahren
- geht teilweise darüber hinaus und entwirft ein Szenario für die nahe Zukunft
- das beschriebene Vorgehen basiert auf intellektuellen Leistungen eines Teams *und* auf automatischen Prozeduren
- will keine komplexe Ontologie, sondern eine praxis-orientierte „leichte“ Terminologie für nahe liegende Anwendungen
- kann auch ohne große Ressourcen langsam aufgebaut werden

Aufbau und Pflege

von Fachterminologie

Umwelt Bundes Amt
Für Mensch und Umwelt

innoQ

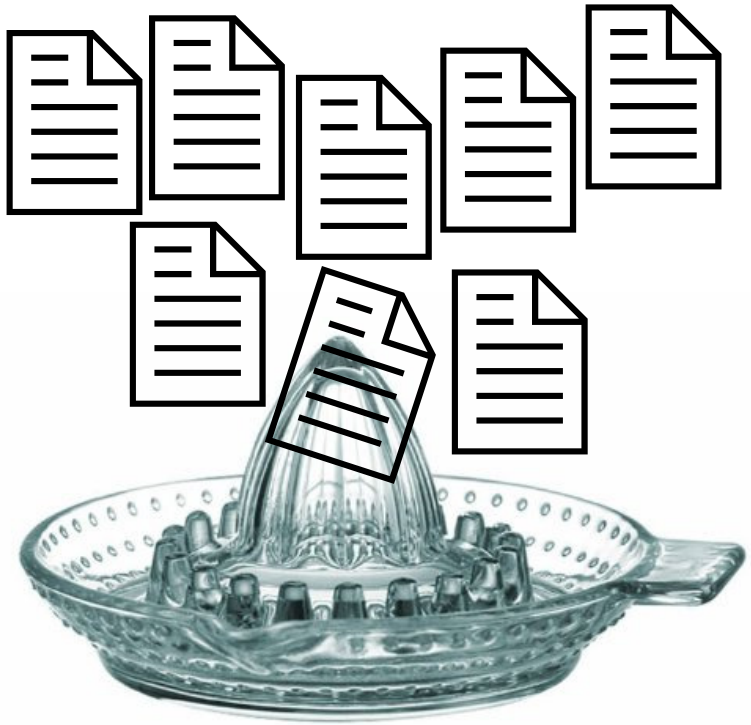
Sichtung vorliegender Fragmente

- Glossare
- Code-Listen in Datenbanken
- Sitemaps und Menüs
- Branchenstandards (GEMET, Financial Business Ontology, …)

- **TOP 100** = Liste der 100 „beliebtesten“ Begriffe – Ergebnis eines Rankings im Team

Analyse des Corpus (1)

ca. 100 repräsentative Dokumente



Zeichenkette	Wort	Dokument	Absatz	Satz	Position

(Stoppwörter werden ignoriert)

Analyse des Corpus (2)

Häufigkeit der Wörter im Corpus insgesamt

Zeichenkette	Wort	Dokument	Absatz	Satz	Position

Wort	Häufigkeit insgesamt

= das Gegenstück zu den TOP 100 des Teams

Ein Indiz für semantische Nähe:
Welche Wörter kommen wie oft gemeinsam im selben Dokument/Absatz/Satz vor?

Wort1	Wort2	Dokument	Absatz	Satz

Begriffe, Bezeichner und ihre Schreibformen

im Corpus finden wir Zeichenketten, die Schreibformen von Wörtern sind, welche Begriffe bezeichnen.

Flexionsformen der 90.000 gängigsten Wörter liegen im Deutschen Morphologie-Lexikon als Creative Commons vor.

Zusammensetzungen, die auch getrennt geschrieben werden können.

Homographe und Eigennamen, die schreibgleich mit anderen Wörtern sind.

Umwelt Bundes Amt
Für Mensch und Umwelt

innoQ

Mehrfache Bezeichner von Begriffen

“Abfall” und “Müll”

„Waldschaden“ und „Waldsterben“

sind das jeweils eigenständige Begriffe?

Vorzugs-Bezeichner und alternativer Bezeichner desselben Begriffs?

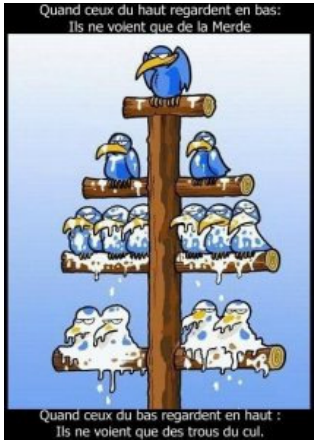
Wie genau wollen Sie unterscheiden?

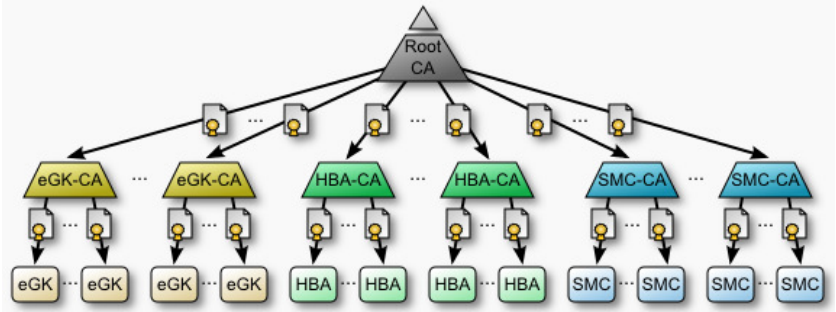
Lieber nicht zu genau, aber auch nicht zu wenig.

Aber machen Sie es überall gleich.

Hierarchie und andere Verwandtschaften (1)

- eventuell vorliegende Hierarchien bewerten
- Code-Listen: Attribute werden Oberbegriffe für ihre Wertevorräte
- Wichtig: die Begriffe der obersten Hierarchie-Ebene
  - wie viele soll es geben?
  - sind sie semantisch klar gegeneinander abgegrenzt?
  - ist ihre semantische Nähe gleichmäßig verteilt?
  - sind ihre jeweiligen Teilbäume in etwa gleich stark?





Umwelt Bundes Amt
Für Mensch und Umwelt

innoQ

Hierarchie und andere Verwandtschaften (2)

- *allgemeine* nicht-hierarchische Verwandtschaft "*related*"
- Anhaltspunkte liefert die in der Corpus-Analyse ermittelte semantische Nähe.

- *spezielle* nicht-hierarchische Verwandtschaften:
  - Rauchen *verursacht* Krebs
  - Katze *frisst* Maus
  - ….
- Verursacht Rauchen wirklich Krebs oder erhöht es nur das Risiko? Und um wie viel Prozent?
- **Nicht verzetteln – keep it simple!**

Definitionen

vorliegende Definitionen zitieren,

eigene Definitionen erst mal meiden

Definitionen sind für die meisten nahe liegenden Anwendungen gar nicht erforderlich

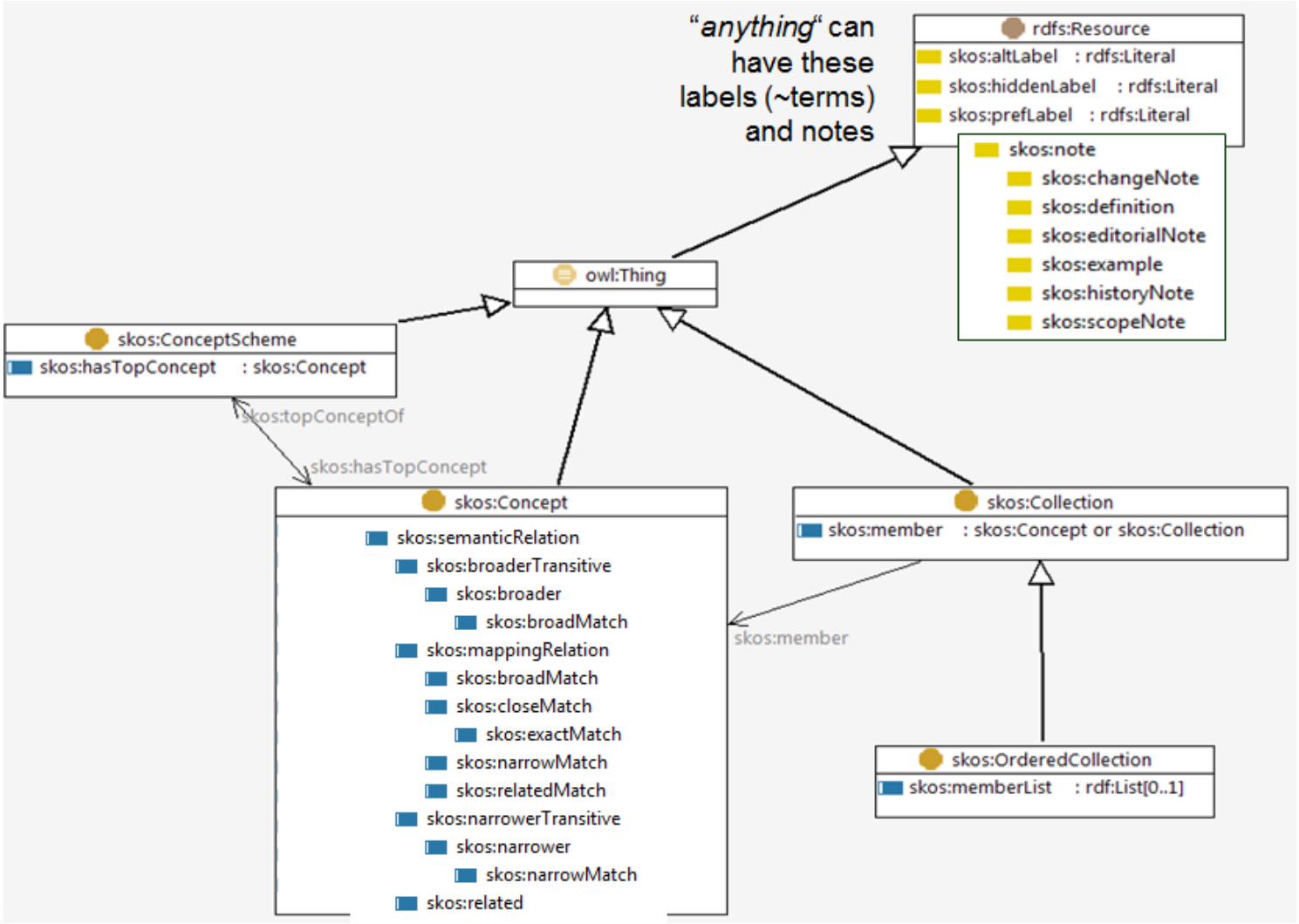
ein Begriff ist bereits durch seine Beziehungen zu anderen Begriffen und durch seine Bezeichner praktikabel identifiziert

eigene Definitionen entwickeln sich nach Bedarf, wenn die Terminologie erst einmal etabliert ist …

Formalisieren

- die Terminologie muss in eine interoperable, maschinenlesbare Form gebracht werden.
- Simple Knowledge Organisation System (SKOS), eine verbreitete W3C Recommendation auf Basis von RDF
- dies ist kein einmaliger Vorgang, sondern muss bei der Pflege laufend erfolgen
- Werkzeuge wie PoolParty (kommerziell) oder iQvoc (open source) bieten einen Editor für SKOS-kompatible Terminologie

Simple Knowledge Organisation System



Umwelt Bundes Amt
Für Mensch und Umwelt

Verfügbar machen

- Die Terminologie muss zu jeder Zeit einfach und schnell im Web (oder IntraNet) erreichbar sein
- Jeder Begriff wird durch eine eindeutige HTTP URI repräsentiert und kann mit dieser referenziert werden
- Diese URI liefert eine maschinen- oder menschenlesbare Darstellung, abhängig vom Accept Header des Aufrufs
- Dazu gehört auch Werbung – die beste Werbung sind sinnvolle, sichtbare Anwendungen
- aber auch: (Un-) Begriff des Monats, Begriffsranking nach Häufigkeit in neuen Dokumenten, jeder kann Begriffe kommentieren etc.

Persistenz und Evolution

„Cool URIs don‘t change!“ (Tim Berners-Lee 1998)

Wir wollen, dass Begriffe referenziert werden, daher dürfen wir keine toten Links generieren

Begriffe werden nicht gelöscht, sondern bei Bedarf deutlich als verfallen markiert

keine starre Versions-Strategie, sondern laufende Erweiterung um einzelne Begriffe oder Themengruppen

Es existiert immer nur die neueste Version, abwärts kompatibel zu allen früheren Versionen

Nutzung

von Fachterminologie

Umwelt Bundes Amt
Für Mensch und Umwelt

innoQ

Nutzung in der täglichen Kommunikation

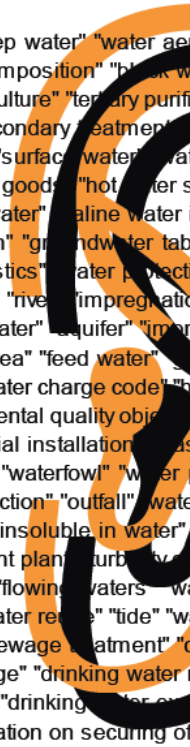
klar abgegrenzte Begriffe verringern Missverständnisse

Auswahl aus „ähnlichen Begriffen“ beim Editieren hilft beim präzisen Formulieren

auch unerwünschte Begriffe (Beispiel „Waldsterben“) können automatisch erkannt werden

Subskribieren von Schlüsselbegriffen als Filter für News-Feeds und neue Dokumente

The noise of water (2001)



"daphnia" "direct discharger" "permit for the use of water" "deep water" "water aeration" "snow water" "terms of waste water" "agricultural effluent" "boiling water reactor" "water pollution load" "combined waste water" "waste water composition" "black water" "grey water" "river works" "water board law" "water supervision" "protection area for water regulation" "groundwater resources" "salamander" "agriculture" "tertiary purification of sewage" "sewage purification close to nature" "persistent chemical substance" "primary treatment" "flood" "chemical sewage purification" "secondary treatment" "sewage" "industrial effluent" "lowering of groundwater level" "waste water purification" "bathing waters" "water protection" "waste water reclamation" "surface water" "water privilege" "underground disposal of waste water" "water rate statute" "water levy act" "hydro-isobath" "groundwater flow" "groundwater storey" "public goods" "hot water storage" "hot water heating system" "groundwater" "waste water legislation" "water quality management" "water heating" "seepage water disposal" "hot water" "saline water intrusion" "industrial water" "mineral water" "stocktaking" "municipal water management" "hydraulic and sanitary engineering" "waste water examination" "groundwater table" "void water" "seepage water treatment" "percolating water" "waste water reduction" "sewage flow" "deep sea" "sewage lagoon" "waste water statistics" "water protection policy" "well" "water quality directive" "salts" "surface water" "river water" "wastewater load" "indirect discharger" "back water" "waste water register" "river" "impregnation (materials)" "municipal sewage" "waste water sludge" "ordinance on parameters of noxiousness of waste water" "sludge" "harmfulness of wastewater" "aquifer" "impregnating agent" "sewer" "desalination of brackish water" "waste water decontamination" "brackish water" "Waste Water Origins Ordinance" "intertidal area" "feed water" "groundwater contour line" "ground water conservation" "soil moisture regime" "soil water" "small body of water" "waste water charge legislation" "waste water charge code" "human settlement" "stagnant water" "waste water charge fixation" "state water law" "waters (geographic)" "water sciences" "waste water charge" "environmental quality objective" "turbine" "residue" "water works" "waste water disposal" "proprietary right" "water course regulation" "sewage decontamination" "liquid manure" "industrial installation" "waste water disposal" "embargo" "wastewater discharge" "rinse water" "EU Water Protection Directive" "industry" "Framework Water Directive" "river filtrate" "waterfowl" "water pollution" "rhithrosphere" "dump impounded water" "turbomachine" "water supply" "water pollution prevention" "raw water" "deep water" "sea water protection" "outfall" "water evaporation" "water consumption" "water board decree" "biological water testing" "sea water fish" "water analysis" "sea water desalination" "material insoluble in water" "shore bird" "waste water disposal embargo" "waste water disposal scheme" "sea water" "drinking water preparation regulation" "sewage treatment plant" "turbine" "water" "drinking water" "water management plan" "general planning on water resources development" "sewage disposal" "residual amount of water" "flowing waters" "water management" "wastewater quality" "tail water" "condensate" "under water coating" "planning permission" "aquatic animal" "water temperature" "water regime" "tide" "waste water treatment plant" "physical sewage treatment" "mechanical sewage treatment" "water resources" "international convention" "electrochemical sewage treatment" "chemical sewage treatment" "rural area" "anaerobic sewage treatment" "aerobic sewage treatment" "water mite" "permit" "drinking water supply" "drainage" "drinking water regulation" "water quantity management" "water volume" "water market" "water statistics" "water level" "water sports" "reservoir" "waste water treatment" "drinking water ordinance" "drinking water protection area" "water shortage" "solubility in water" "drinking water quality" "water line" "water conductivity" "low water" "regulation on securing of enough water" "law on the securing of enough water" "securing enough water" "shallow water" "protected water catchment area" "Act Pertaining to Charges Levied for Discharging Waste Water into Waters" "wastewater levy" "water cooling system" "discharged water" "water cycle" "hydroelectric power plant" "water power" "water partition" "waterborne sound" "water pollutant" "rainwater" "flood runoff" "drinking water treatment plant" "drinking water treatment" "pollution of waters" "sewerage system" "water contents" "vapour" "Water Hygiene Act" "water purification" "water act" "water association" "vadose water" "inland water way" "water hygiene" "water resources policy act" "biological water balance" "water hardness" "water pollution monitoring" "water sample" "water price" "water flow" "aquatic plant" "increasing water hardness" "water penny" "water surface" "surface runoff" "water utilization" "PWR-type reactor" "deep sea fishing" "inland waters" "algae bloom" "hydrologic balance" "water movement" "water bottom" "water quality model" "water deposit" "water protection directive" "water protection legislation" "water pollution control measure" "water pollution control act" "water pollution control deputy" "toilet" "pelagial" "water demand" "hydraulic construction" "water treatment" "water quality" "long distance water supply" "water catchment" "restoration of waters" "impounded water" "water act" "water company" "regulation of waters" "water content" "bilge water" "water ouzel" "water runoff" "water endangering" "regulation concerning water endangering matter" "wet-type cooling tower" "permit to exploit water" "reject water" "fresh

Umwelt Bundes Amt
Für Mensch und Umwelt

Nutzung im Wissensmanagement

- klassische Volltextsuche vergleicht nur Zeichenketten und nicht Begriffe
- mit „Waldsterben“ wird man nicht auch „Waldschaden“ finden
- die Auswahl „ähnlicher Begriffe“ in der Suchbedingung hilft
- mit „Wald“ wird man nicht „Wälder“ finden
- ein Begriffs-basierter Index kennt bereits alle Bezeichner und Schreibweisen
- Begriffs-basierte Suche kann auch ganze Begriffsnetze als Suchbedingung verwenden

Automatische Indexierung

- Einen Begriffs-basierten Index sollte man nicht manuell aufbauen – es sei denn man hat nur wenig Dokumente und viel Personal

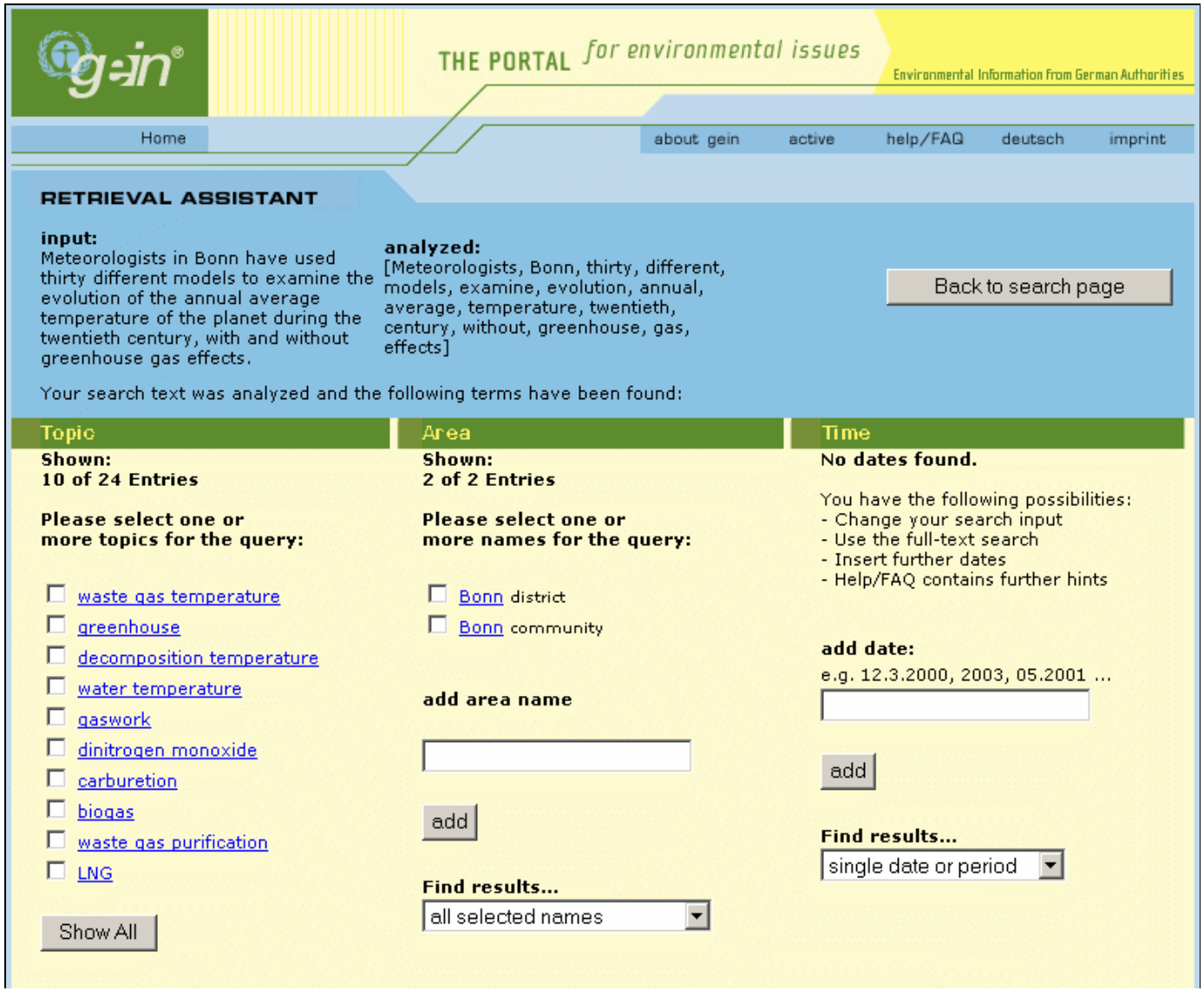
- automatische Verfahren (z.B. iQvoc) erreichen eine hinreichende Genauigkeit, abhängig von der Qualität und „Dichte“ der Terminologie.
  Insbesondere gilt dies für zusammengesetzte Begriffe.

- Die Verbesserung der Terminologie ist vielversprechender als die manuelle Verschlagwortung

Dokumentenvergleich

- aufgrund der Indexeinträge können Dokumente verglichen werden, auch wenn sie sich in der Wortwahl unterscheiden.
- Dokumente (oder Absätze daraus) können als Suchbedingung verwendet werden
- Zu jedem neuen Dokument kann automatisch eine Liste ähnlicher Dokumente erzeugt werden

Anwendung in *gein*® 2003



Umwelt Bundes Amt
Für Mensch und Umwelt

innoQ

Nutzung mit verlinkten Daten

Vier Grundprinzipen

- Use URIs as names for things
- Use HTTP URIs so that people can look up those names.
- When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
- Include links to other URIs so that they can discover more things.

Tim Berners-Lee, 2006-07-27
http://www.w3.org/DesignIssues/LinkedData.html

Nutzen von verlinkten Daten

- Datensätze referenzieren die Terminologie
- Diese Verweise können sofort verfolgt werden
- Wertevorräte für einzelne Attribute können in der Terminologie als Kollektionen (skos:Collection) abgelegt und über eine REST-Schnittstelle kontrolliert werden.
- Wenn die Datensätze selbst als Linked Data verfügbar sind, kann übergreifend nach Verweisen auf denselben Begriff gesucht werden.
- Durch weitere OWL/RDF Statements oder Regeln wird Datenintegration stark vereinfacht

Zusammenfassung

- der Aufbau von Fachterminologie ist auch ohne groß aufgehängtes Projekt machbar
- man braucht ein motiviertes Team und ein paar Werkzeuge (auch als open source verfügbar)
- Im Fokus sollten nahe liegende Anwendungen stehen, nicht die perfekte Ontologie
- Die Terminologie muss im Web verfügbar sein und bekannt gemacht werden
- Der potentielle Nutzen liegt in der täglichen Kommunikation, im Wissensmanagement und in verlinkten Daten

The End

Joachim Fock
Umweltbundesamt (de)
Wörlitzer Platz 1, 06813 Dessau-Roßlau, Germany
joachim.fock@uba.de

Thomas Bandholtz
innoQ Deutschland GmbH, Halskestr. 17, 40880 Ratingen, Germany
thomas.bandholtz@innoq.com

http://www.innoq.com/de/themen/linked-data/iQvoc

Umwelt Bundes Amt
Für Mensch und Umwelt