CEDEX eine erweiterbare Ontologie für ökologische Daten

Schentz Herbert, Mirtl Michael, Umweltbundesamt GmbH (Österreich),
herbert.schentz@umweltbundesamt.at

Abstract

Because of the success we had with the ontology of MORIS within that information system we tested the possibility to use it as an Ontology for the Semantic Web. Thus we created a basic ontology for ecological data in UML and OWL Syntax, which is completely object orientated, can be easily extent and can deal with "soft relationships". We call that ontology CEDEX and present it on http://www.umweltbundesamt.at/CEDEX

1 Zielsetzung

Das von der Umweltbundesamt GmbH für das Integrated Monitoring geschaffene Informationssystem MORIS wird erfolgreich auch für die Abbildung, Verwaltung und Auswertung von Daten diverser ökologischer Projekte eingesetzt. Das heißt, die Ontologie von MORIS ist für die Abbildung von zu mindestens vielen unterschiedlichen, wenn nicht von fast allen ökologischen Themen geeignet. (Ich möchte eine Ontologie als eine Gesamtheit von Semantiken, Strukturen Funktionen, Methoden und Modellen zur Abbildung eines Teils der Wirklichkeit verstehen, im Bewusstsein, dass dies weit weniger ist als „die Welt wie sie ist"[wörtliche Übersetzung]).

Will man Daten horizontal (themenübergreifend) und vertikal (institutionsübergreifend) im Web vernetzen, dann braucht man dazu nebst Webservices und Verzeichnissen darüber eine gemeinsame maschinenverständliche Sprache und Datenstruktur.

Wir haben uns also gefragt, ob die MORIS Ontologie auch für das semantische Web geeignet wäre. Das Ergebnis dieses Ansatzes, den wir

CEDEX (**C**lasses for **E**nvironmental **D**ata **EX**change) nennen, soll hier zur Diskussion gestellt werden.

2 Ansatz für die Erstellung von CEDEX

Folgende Grundsätze sollen von Anfang an gelten:

Die Definition hat objektorientiert zu erfolgen (was leicht ist, da MORIS bereits objektrelational aufgebaut ist)

Die Definition soll möglichst unabhängig von den Umsetzungswerkzeugen mit UML erfolgen und erst im letzten Schritt z.B. auf OWL, der von W3C empfohlenen Web Ontology Language, gemappt werden.

Die Basisontologie soll sehr einfach durch voneinander möglichst unabhängige Extensions erweiterbar sein.

Es sollen nicht nur harte Relationen, sondern auch Beziehungen wie „ist verwandt mit“, ist „ähnlich“ abgebildet werden können.

2.1 Objektorientierte Ontologie

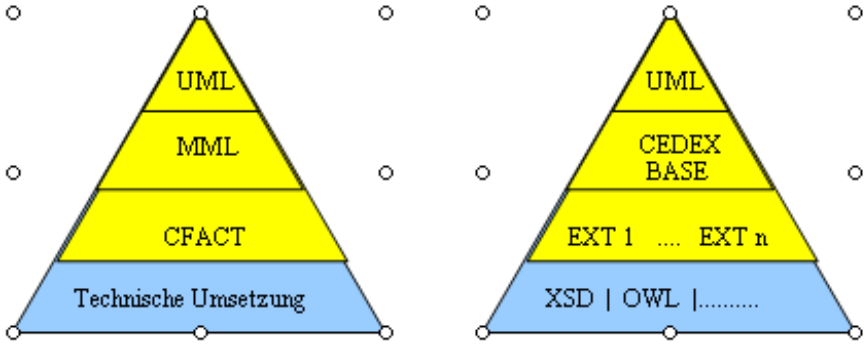
Wir gehen von der Vermutung aus, dass objektorientierte Strukturen in der nächsten Zeit in der IT-Welt state of the art sein werden, insbesondere bei Datenaustauschelementen. Da es sinnlos ist, dafür Definitionsmethoden abseits des mainstreams fest zu legen, nehmen wir komentarlos die formale Sprachen UML und für die Umwetzung OWL.

2.2 Vorgehensmodell aus CFACT (e-business)

Die UN Arbeitsgruppe geht bei der Erstellung von CFACT davon aus, dass die Definition und die Lebensdauer des Modelles länger sein wird als der Bestand einer formalen Sprache, in die es umgesetzt werden kann. Daher erfolgt die Objektdefinition auf einer abstrakteren Ebene (UML), von der man selbstverständlich auch nicht behaupten kann, dass sie technologieunabhängig wäre. Da diese Sprache dann zu allgemein wäre, wird daraus noch eine eigene e-business orientierte Metasprache abgeleitet. Erst auf dieser aufsetzend, werden die fachspezifischen Objekte definiert. Je nach Erfordernis, state of the art, ... , kann diese Ontologie auf

Schnittstellendefinitionen, Kataloge, Datenbanken, ... unterschiedlicher Technologie gemapped werden.

Abb 1 Das Vorgehensmodell von CFACT und der Vorschlag für CEDEX



2.3 Einfache Erweiterbarkeit:

Wir sind zutiefst überzeugt, dass es weder möglich ist, mit einem Schlag eine vollständige ökologische Ontologie zu entwickeln, noch, je eine auf zu bauen, die für alle Ewigkeit hält. Zuviele Themen, Sprachen, Terminologien und Institutionen sind involviert. Andererseits soll die Ontologie so aufgebaut werden, dass Zusammenhänge ihrer Objekte und anderer Elemente auffindbar, notierbar, überprüfbar und darstellbar werden, also nicht soweit auseinander liegen, dass sie nichts mehr miteinander zu tun haben.

2.4 Ähnlichkeit und nicht Gleichheit:

Wir IT – Leute wünschen uns meist möglichst klare Definitionen und eindeutige Schlüssel. Genau diese Homogenisierung sollte zwar immer angestrebt werden, ist aber aus etlichen Gründen nicht immer möglich. Einige seien hier aufgezählt:

Verwendung historischer Ergebnisse trotz Weiterentwicklung in Methodik und Technik.

Gewinnung der Ergebnisse unter vollkommen unterschiedlichen Bedingungen. (z.B. Meeresküste / Alpen).

Gewinnung der Ergebnisse unter unterschiedlichen gesetzlichen Vorraussetzungen und daher unterschiedlichen geforderten Methodiken.

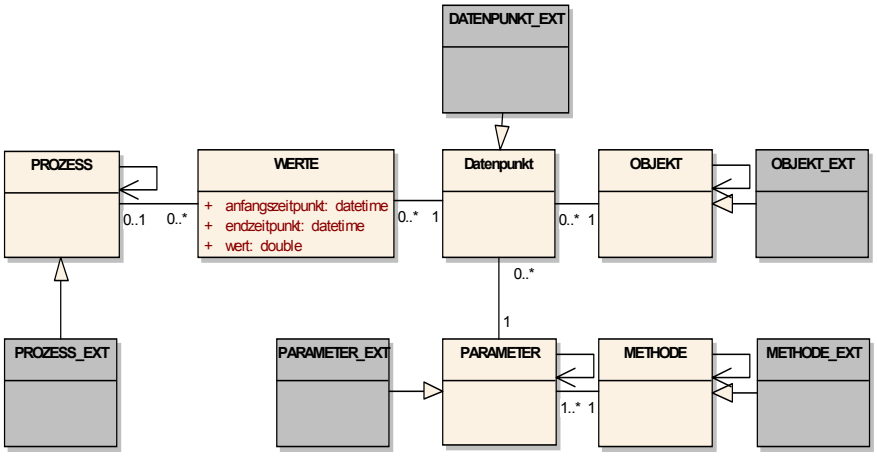
3 Umsetzung

3.1 CEDEX BASE

Es wird eine einfache Basisontologie (CEDEX – BASE) gebildet.

Das Core von CEDEX-BASE setzt die Strukturelemente **"Raumbezug, Zeitbezug, Sachbezug"** (© UDK et al.) in „OBJEKT“, „PARAMETER“, deren Assoziationsklasse „Datenpunkt“ und „WERTE“ um und erweitert diese um **Methoden** und **Prozesse** .

Abb. 2 Core von CEDEX BASE und die dazugehörigen Extension Klassen



(grau, nur der erste Vererbungsschritt)

Dieses Core wird um 4 Attachmentklassen **Akteure, Projekte, Gesetze, Dokumente** ergänzt, die nach den gleichen Regeln wie die Coreklassen erweitert werden können.

Hinzu kommen noch einige Zusatzklassen, für die keine Erweiterbarkeit vorgesehen ist, wie Dimensionen, Umrechnungsfaktoren, Skalierungen, ...

Besondere Bedeutung kommt dem Datenpunkt zu, da er die Quelle für Umweltdatenkataloge sein kann. (Was wird wo gemessen?)

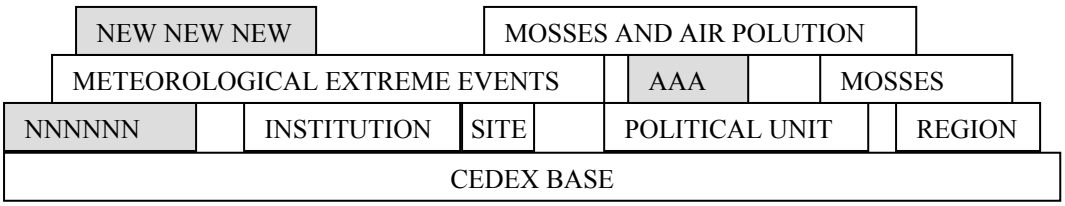
3.2 Extensions

Alle Erweiterungs - Klassen werden aus CEDEX - BASE Klassen abgeleitet

Alle vorkommenden Relationen sind in CEDEX BASE abgebildet. In den Extensions werden ausschließlich erlaubte Klassen der Basisrelationen eingeschränkt.

So können voneinander weitestgehend unabhängige Extensions für bestimmte Themen aufgesetzt werden.

Abb. 3 CEDEX BASE und einige Extensions: . Hat man sich einer Extension nicht committet, dann verliert man nur die Relationen zu diesen, die übrige Information kann man aber nützen.



Dadurch sollten beides berücksichtigt werden können:

Die Ontologie kann wachsen und muss nicht von Anfang an „allumfassend sein“. (Was Fertigstellungsdatum St. Nimmerleinstag zur Folge hätte)

Programme die die Basisontologie an der Schnittstelle einmal verwirklichen, brauchen für die Erweiterungen nicht mehr umgebaut werden, sondern müssen nur mit neuen Klassendefinitionen (etwa über OWL) versorgt werden.

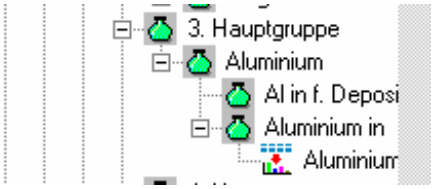
Das Frame Work ist sehr einfach, weshalb man keinen besonders grossen Aufwand braucht, um die Regeln kennen zu lernen, sei es um die Ontologie für bestimmte neue Gebiete zu erweitern, sei es um für eine Applikation eine Schnittstelle zu bauen. (Die Abbildung von Fachgebieten ist trotzdem nicht trivial!)

3.3 Weiche Relationen

"Weiche" Relationen wie zum Beispiel "ist verwandt mit" werden über Polyhierarchien innerhalb von Klassen, die aus einer Basisklasse abgeleitet wurden, umgesetzt.

Abb. 4 Ähnlichkeiten diverser Parameter "Aluminium". Die Methodik zur Ermittlung des Parameter Al in f. Deposition ist aus der (allgemeinen) Methodik zur Ermittlung des Parameters "Aluminium" abgeleitet, aber mit

dieser nicht vollkommen ident, weshalb die zugehörigen Parameter eine "parent - child" Relation besitzen.



3.4 Funktionen und Methoden

CEDEX BASE verfügt in seiner UML Definition bereits über etliche Funktionen und Methoden, die nur exemplarisch gebracht werden, um zu erläutern, dass durch das Modell eine Funktionen recht einfach werden:

Statistische Funktionen über Attribute untergeordneter Objekte (Maximum, Minimum, Mittelwert, Anzahl....). Dabei ist interessant, dass man, sobald die Funktionen als Webservices implementiert sind, sehr einfach neue Collections zusammenstellen kann, für die dann ohne irgendeinen Programmieraufwand die Funktion am Server vorhanden ist.

Aggregierfunktionen für Werte (Messreihen, Beobachtungsergebnisse) gemäß der Zeit, der untergeordneten eines Objektes (Collections s.o.) und der Untergeordneten eines Parameters, oder eines Prozesses (und dessen Untergeordneter). Auch da ist wieder interessant, dass ein Prozess (ein Event) im Nachhinein auf den Datenbestand gelegt werden kann und dafür dann aggregiert werden kann.

Funktionen über Updates. Diese Funktionen könnten besonders für ein Caching innerhalb von Rechnerverbünden interessant sein.

3.5 Modelle und Hypothesen

Viele, wenn auch nicht alle Fach- Modelle und Hypothesen können als Regeln innerhalb der „Datenpunkt" -Klassen dargestellt werden, wobei für die Abbildung vollkommen unbedeutend ist, ob Datenpunkte und Datenpunktklassen nun mit tatsächlichen Messungen verbunden sind oder hypothetisch sind:

Bsp1.: Die Nitratkonzentration im Boden zu einem bestimmten Zeitpunkt ist x * der Summe der Zeitfunktion (ausgebrachte Düngung, Datum).

Bsp2.: Die Häufigkeit von Extremniederschlägen in Westeuropa korreliert mit der mittleren Wassertemperatur des Golfstromes in den Monaten März bis Oktober.

4 Überprüfung, ob diese Ontologie hinreichend ist

Es ist nicht gut möglich, theoretisch zu überprüfen, ob ökologische Themen mit diesem Modell hinreichend abgebildet werden können. Es kann nur taxativ gecheckt werden.

Die mehrmaligen Anwendung von MORIS, aus dem CEDEX abgeleitet ist, legt die Vermutung nahe, dass dieser einfache Aufbau hinreichend ist:

4.1 Integrated Monitoring

Da CEDEX weitestgehend eine Umsetzung von MORIS ist, weiss man, dass folgende Themen erfolgreich in CEDEX abbildbar sind:

- Luftchemie
- Wasserchemie
- Bodenanalytik
- Bodenwasseranalytik
- Vegetationskunde
- Biodiversity
- Geologie
- Depositionsanalytik

4.2 Schwermetalle in Moosen

Die 5 jährig wiederkehrende Untersuchung von Schwermetallen in Moosen wurde deshalb ins MORIS abgebildet, weil nach einem bestehenden Informationssystem gesucht wurde, in welches Daten und Metadaten umgehend eingebracht werden können.

Sowohl die Beschreibung der Methodik als auch Standortdaten, Parameterdaten, Probenattribute, Werte und Wertattribute konnten ohne Probleme eingebracht verwaltet, extrahiert, einem Statistikpaket zur Verfügung gestellt, in einer Zeitreihe dargestellt und geographisch visualisiert werden.

Die CEDEX Extension für dieses Thema liegt als Beispiel vor.

4.3 Klimatische Extremereignisse

Aus Anlass des Hochwassers des Jahres 2002 entstand in Österreich die StartClim Initiative, ein Projekt, in dem Theorien, Daten und Auswertungen aus Meteorologie, geomorphologischer Beobachtung, ökonomischer Beobachtung, landwirtschaftlicher Beobachtung rund um meteorologische Extremereignisse miteinander vernetzt werden sollten. Dafür bedurfte es rasch und günstig eines Informationssystemes.

Da hier nicht nur ökologische Daten miteinander in Verbindung zu bringen sind, und anfangs auch nicht so eindeutig schien, was ein metoerologisches Extremereignis ist, war uns die Abbildbarkeit in die Struktur von MORIS nicht selbstverständlich. Umso erfreulicher war der Erfolg. Die Abbildung liegt als CEDEX Extension vor.

5 Mapping auf Technologien

CEDEX BASE und einige Extensions wurden von uns auf OWL (W3C empfohlene Web Ontology Language) und das proprietäre Format von Protégé, einem Werkzeug der Standford University gemappt. Es scheint uns sehr wahrscheinlich, dass CEDEX BASE und dessen Extensions auch auf DAML & OIL und andere formale, objektorientierte Sprachen abbildbar ist. Da es im Semantic Web die Hauptaufgabe ist und sein wird, dass sich möglichst viele Organisationen auf eine Semantik einigen, sind wir sehr an Kritiken und Diskussionen interessiert und haben diese Ontologie, z.B. auch in der Protégé Community, zur Diskussion gestellt.

5.1 Vergleich mit anderen Ontologien

Zur Zeit gibt es vielerorts Bemühungen um eiheitliche Ontologien für den ökologischen Bereich. Aus diesem Grund sollen einige rudimentäre

Querverweise und Vergleiche mit ökologischem Bezug angestellt werden: Dublin Core, ISO 19115, EML eine Markuplanguage der amerikanischen LTER Community, Importprofil von GEIN (G2K) , ...

Ziel eines solchen Vergleiches ist es vor allem, damit einen Beitrag zu einem breiten internationalen Diskussionsprozess zu leisten. Der Vergleich kann vorläufig nur in groben Zügen erfolgen. Mit Beginn einer intensiveren Diskussion würden wir uns allerdings freuen, einen intensiveren Vergleich an zu stellen und Modifikationen durch zu führen.

5.2 Vergleich mit G2K

CEDEX	G2k	entspricht
OBJEKT	Location	Gut
PROCESS	Event	Gut
PARAMETER	Descriptor	In etwa
DATENPKT	Where/what	inhaltlich

Der Unterschied in den Assoziationsklassen kommt wohl daher, dass G2K (noch) nicht als Schnittstelle für (Mess)Werte gedacht ist. Darum gibt es auch kein Pendant zum Objekt WERT von CEDEX.

Diese Gegenüberstellung berücksichtigt nur jenen kleinen Ausschnitt von G2K, der Vergleiche zuläßt. Selbstverständlich ist der polyhierarchische Tessaurus in CEDEX abbildbar

5.3 Vergleich mit EML

CEDEX	EML	entspricht
OBJEKT	Geographic Coverage	etwa
PARAMETER	Taxonomic Coverage	inhaltlich
WERT		Anders abgebildet
PROCESS	-------	
METHODE	Methode + Protocoll	inhaltlich

Diese Gegenüberstellung berücksichtigt nur jenen kleinen Ausschnitt von EML der Vergleiche zuläßt.

5.4 Vergleich mit Dublin Core

Dublin Core ist eine perfekte Struktur zur Übertragung von Metadaten und Daten, wenn man sich vorher auf einem anderen Weg über die Ontologie, also Semantik und Zusammenhänge geeinigt hat.Gerade das soll aber durch die Verwendung von CEDEX nicht notwendig sein.CEDEX soll die maschinlesbare Übertragung der gesamten Ontologie erlauben.

5.5 Vergleich mit ISO 19115

ISO 19115 definiert erfolgreich die Metadatenbeschreib-ung geographischer Daten.Sie erlaubt jedoch nicht die Abbildung von Zusammen-hängen, die über geographische Relationen hinausgehen, innerhalb der Metadaten. Diese sind aber für die Interpretation der Daten durch Programme notwendig. Es scheint aber nicht schwierig zu sein, für bestimmte Zwecke, Daten aus einem CEDEX Format in ISO 19115 zu transformieren.

5.6 Warum dann überhaupt die Mühe CEDEX zu definieren?

3 Details hat CEDEX, die uns bisher äußerst praktisch erschienen sind, und die wir sonst noch nicht in der Klarheit gefunden haben:

Datenpunktklasse

Datenpunkt

Polyhierarchie.

In der Hoffnung, eine rege Diskussion um diese Details an zu kurbeln und, ist diese Präsentation entstanden.

6 Downloads

Über die Seite http://www.umweltbundesamt.at/CEDEX der Umweltbundesamt GmbH sind CEDEX und folgende Extensions

CEDEX_INSTITUTION (zur Erläuterung der Wirkung der Polyhierarchie)

CEDEX_REGION (weil Küsten, Flusseinzugsgebiete, Gebirge, ... so oft vorkommen)

CEDEX_POLITICAL_UNITS (weil sie so einfach und Standardbedarf sind)

CEDEX_SITE (weil Messstandorte zum Standard der ökologischen Beobachtung gehören)

CEDEX_HEAVY_METAL (das Thema „Schwermetalluntersuchung in Moosen“)

CEDEX_CLIMATE (das Thema „klimatische Extremereignisse)

In folgenden Formaten downloadable :

XMI in den Styles für Enterprise Architekt und Protege

OWL

Protégé

Dazu gibt es Dokumentationen.

Es sei ausdrücklich darauf verwiesen, dass Protégé weder das einzige Tool zum Editieren von OWL Schemata ist, noch, dass die Autoren es für das beste halten. Es ist einfach das einzige Tool, auf das von der W3C Beschreibung für OWL hin verwiesen wird.

7 Literatur

World Wide Web consortium (W3C), OWL Web Ontology Language, http://www.w3.org/2001/sw/WebOnt/

Science Environment for Ecological Knowledge (SEEK): http://seek.ecoinformatics.org

Stanford University School of Medicine, Protégé, Tool für Ontologien / OWL http://protege.stanford.edu/index.html

Aifb Uni Karlsruhe, KAON (Tool für Ontologien / OWL) http://kaon.semanticweb.org/

Ontoprise GmbH, Ontoedit (Tool für Ontologien / semantisches Web)

http://www.ontoprise.de/products/ontoedit

Ecological Metadata Language (EML): http://knb.ecoinformatics.org/software/eml/

Umweltbundesamt, GEIN 2000, DV-Konzept: http://www.gein.de/2000/g2k-dvk.zip

Umweltbundesamt, G2K Profil: http://www.gein.de/2000/profile-11.htm

ISO 19115 - Geographic information - Metadata: http://metadata.dgiwg.org/standard/index.htm

ISO 19119 Open GIS definition: http://www.opengis.org/docs/02-028.pdf

UN / CEFACT's Business Colaboration Framework: http://www.unbcf.org/specials.html

Mirtl, M., Schentz, H. 2002. Strukturen und Funktionen zur Abbildung interdisziplinärer Langzeitprojekte im Bereich von Ökosystem-Monitoring und -Forschung: Der Weg zum Hauptmenü von MORIS. In: Pillmann,W., Tochtermann, K.(eds) Environmental Communication in the Information Society. International Society for Environmental Protection, Vienna, pp 106-117.

Schentz H., Zechmeister H.,Riss A., Mirtl M. 2002. Der Umgang mit nicht harmonisierten Untersuchungsergebnissen am Beispiel der Verwaltung von

Moosmonitoringdaten des UBA Wien mittels MORIS. AK Umweltdatenbanken, Illmenau http://www.umwelt.schleswig-holstein.de/?AKUmweltdatenbanken

Schentz H., Mirtl M. 2001. Vorstellung des Softwarepaketes MORIS (MOnitoring and Research Information System). AK Umweltdatenbanken, Jena http://www.umwelt.schleswig-holstein.de/?AKUmweltdatenbanken