

# Identifying Information with Environmental Relevance in Web Sites of the Public Administration Using Automatic Classification

Franz Schenk,

*Coordination Center PortalU at the Lower Saxony Ministry of Environment and Climate Protection*

Daniel Meyerholt

*Department of Business Informatics at the University of Oldenburg*

SEARCH [Environmental Information](#) | [Laws](#) | [Research Projects](#) | [Addresses](#)PortalU Search [Advanced Search](#) [History](#) [Options](#) [Tips](#)TOPICS [Agriculture](#) · [Air and Climate](#) · [Animal Protection](#) · [Chemicals](#) ·  
[Construction](#) · [Energy](#) · [Environmental Economy](#) ·  
[Environmental Information](#) · [Forestry](#) · [Gene-Technology](#) ·  
[Geology](#) · [Health](#) · [Nature and Landscape](#) · [Noise and](#)[Agitation](#) · [Radiation](#) · [Soil](#) · [Soil Pollution](#) · [Sustainable Development](#) · [Traffic](#) ·  
[Waste](#) · [Water](#)NEWS [Glücksspiel kann süchtig machen! Aktionstag zeigt Gefahren und Hilfen](#)

28.9.2011 9:40

In der Zeit von 14 bis 18 Uhr wird in Potsdam im Bereich der Brandenburger Straße und am Brandenburger Tor mit verschiedenen Aktionen auf das Thema Glücksspielsucht und die damit verbundenen Gefahren aufmerksam gemacht sowie über Hilfsangebote für spielsüchtige Menschen und deren Angehörige im Land...

Information Provider: Ministerium für Umwelt, Gesundheit und Verbraucherschutz des Landes Brandenburg

PortalU INFO 

10-11.11. 2011: [11. Biomass-Symposia](#) at Umweltcampus Birkenfeld / FH Trier!

19.-27.11.11: [The European Week for Waste Reduction](#). Activities coordinated by [BMU](#) and [NABU](#)

Till august 2012: [New Photocompetition at the Ministry for Climate Protection](#) „New energy – climate protection of North Rhine-Westphalia“

[International year of forests 2011](#)

Our [newsletter](#) provides the latest information about PortalU!

[Subscribe newsletter](#)

Get the latest information about PortalU with our [Newsletter!](#)

ALMANAC 

*1 year ago:*

**Greenpeace activists protested against the nuclear policy of the German government**

- Information brokering system
- Environmental data, literature, research projects, metadata, ...
- Information from public authorities (all federal states and institutions of the federal government)
- Web Index, ~5mio Web pages  
all content with environmental relevance  
collaboratively compiled by the project partners

## Pattern-based definitions of relevant web pages

- Entry point (= start URL)

e.g. *sachsen.de/umwelt.html*

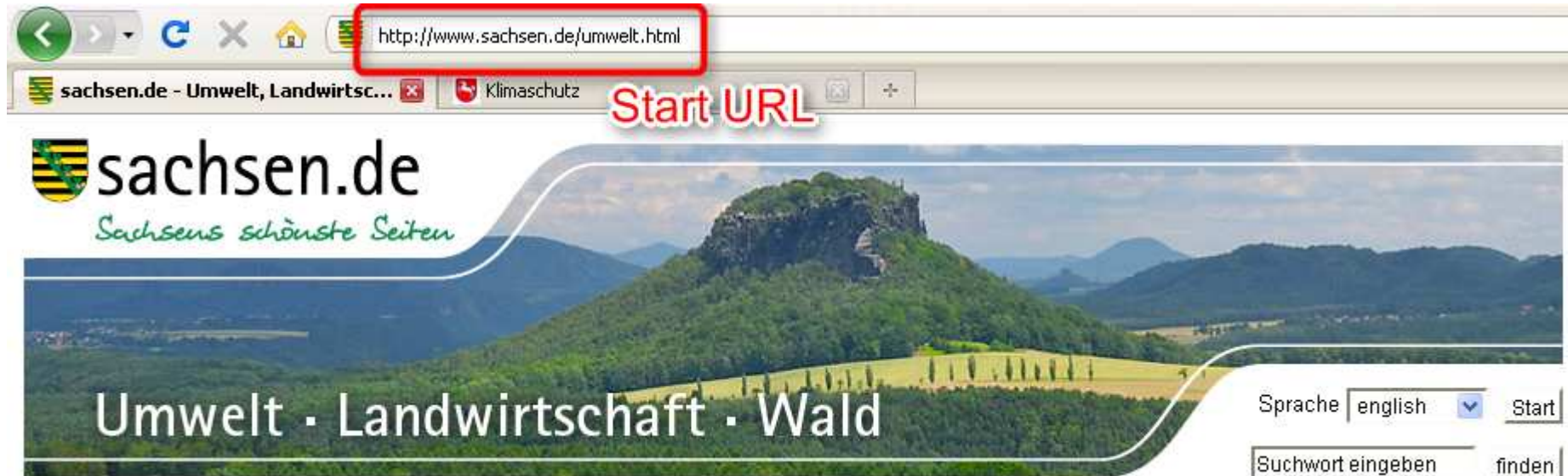
- Boundaries of the site (= limit URL )

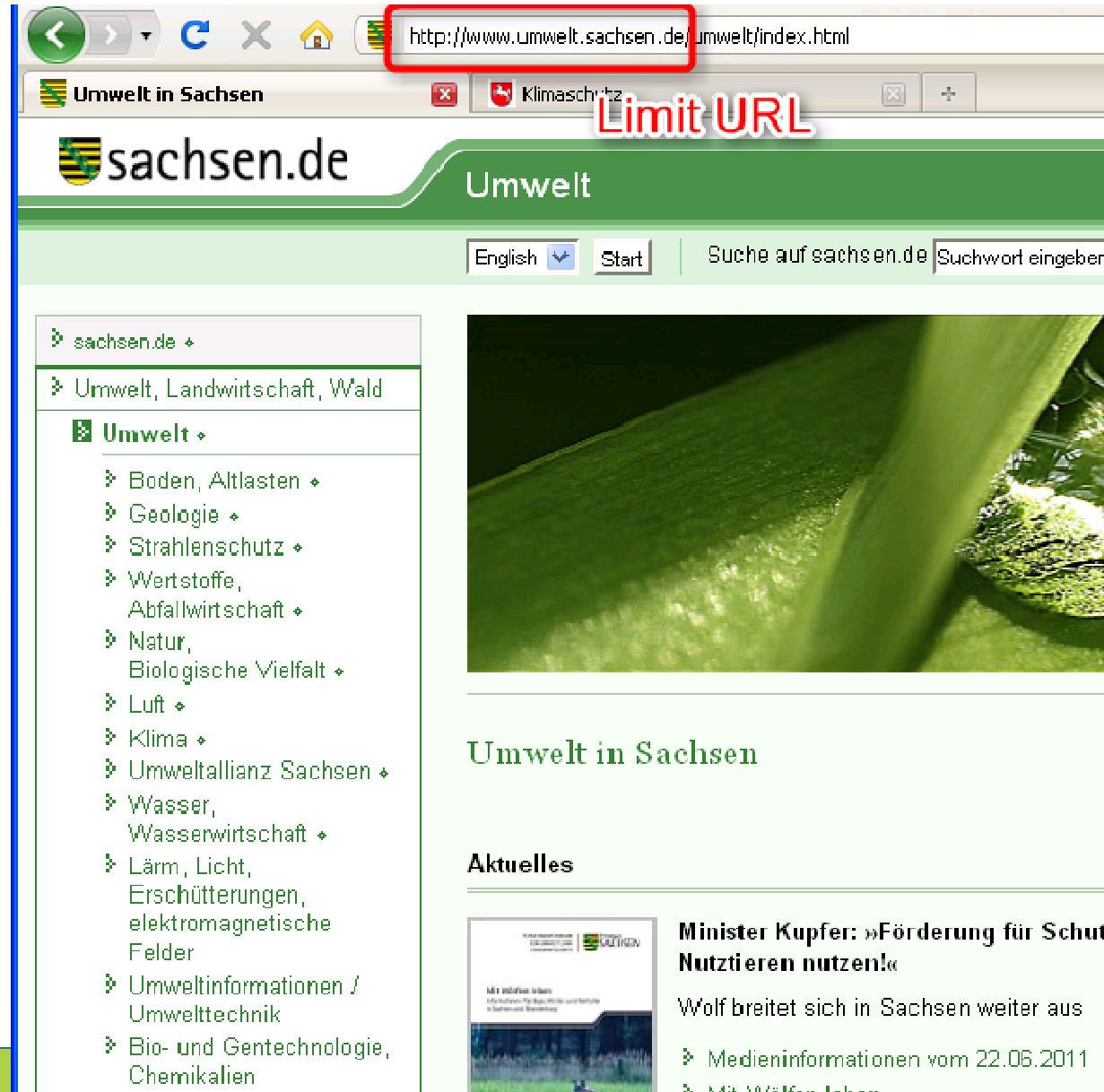
e.g. *umwelt.sachsen.de*

- Exclusion patterns (= exclude URL)

e.g. *umwelt.sachsen.de/umwelt/2420.htm*

# Web Crawler





The screenshot shows a web browser window with the address bar containing the URL `http://www.umwelt.sachsen.de/umwelt/index.html`, which is highlighted with a red rectangle. A red watermark reading "Limit URL" is overlaid on the browser window. The website header features the "sachsen.de" logo and the word "Umwelt". Below the header is a search bar with the text "Suche auf sachsen.de" and a placeholder "Suchwort eingeben". A left sidebar contains a navigation menu with the following items:

- sachsen.de +
- Umwelt, Landwirtschaft, Wald
- Umwelt +**
  - Boden, Altlasten +
  - Geologie +
  - Strahlenschutz +
  - Wertstoffe, Abfallwirtschaft +
  - Natur, Biologische Vielfalt +
  - Luft +
  - Klima +
  - Umweltallianz Sachsen +
  - Wasser, Wasserwirtschaft +
  - Lärm, Licht, Erschütterungen, elektromagnetische Felder
  - Umweltinformationen / Umwelttechnik
  - Bio- und Gentechnologie, Chemikalien

The main content area features a large green image of a plant stem. Below the image is the heading "Umwelt in Sachsen" and a section titled "Aktuelles". The "Aktuelles" section contains a news item with a thumbnail image and the following text:

**Minister Kupfer: »Förderung für Schutz Nutztieren nutzen!«**  
Wolf breitet sich in Sachsen weiter aus

- Medieninformationen vom 22.06.2011
- Mit Wölfen leben



The screenshot shows a web browser window with the address bar containing the URL `http://www.umwelt.sachsen.de/umwelt/2420.htm`, which is highlighted with a red box and labeled "Exclude URL". The browser tabs include "Portalfunktionen" and "Klimaschutz". The website header features the "sachsen.de" logo and the word "Umwelt". Below the header is a search bar with the text "Suche auf sachsen.de" and a "Suchwort eingeben" input field. A left sidebar contains a navigation menu with categories like "Umwelt", "Boden, Altlasten", "Geologie", "Strahlenschutz", "Wertstoffe, Abfallwirtschaft", "Natur, Biologische Vielfalt", "Luft", "Klima", "Umweltallianz Sachsen", "Wasser, Wasserwirtschaft", "Lärm, Licht, Erschütterungen, elektromagnetische Felder", "Umweltinformationen / Umwelttechnik", "Bio- und Gentechnologie, Chemikalien", and "Portalfunktionen". The "Portalfunktionen" section is expanded, showing links for "Übersicht", "Impressum", "Rechtliche Hinweise", and "Kontakt". The main content area is titled "Portalfunktionen" and contains text explaining that each online application under sachsen.de includes content and functions for user navigation, such as the Impressum, legal notices, overview (Sitemap), and contact information. A link "zurück zum Seitenanfang" is also present.

**Exclude URL**

http://www.umwelt.sachsen.de/umwelt/2420.htm

Portalfunktionen

sachsen.de

Umwelt

English Start Suche auf sachsen.de Suchwort eingeben

Umwelt

- Boden, Altlasten
- Geologie
- Strahlenschutz
- Wertstoffe, Abfallwirtschaft
- Natur, Biologische Vielfalt
- Luft
- Klima
- Umweltallianz Sachsen
- Wasser, Wasserwirtschaft
- Lärm, Licht, Erschütterungen, elektromagnetische Felder
- Umweltinformationen / Umwelttechnik
- Bio- und Gentechnologie, Chemikalien
- Portalfunktionen**
  - Übersicht
  - Impressum
  - Rechtliche Hinweise
  - Kontakt

### Portalfunktionen

Zu jeder Online-Anwendung unter sachsen.de gehören neben Inhaltsseiten zum jeweiligen Thema auch solche, die unterstützende Informationen und Funktionen zur Nutzung der Online-Anwendung beinhalten.

Solche Inhalte werden hier als »Portalfunktionen« bezeichnet.

Dazu gehören das Impressum, Rechtliche Hinweise, Übersicht (Sitemap) und der Kontakt.

[zurück zum Seitenanfang](#)

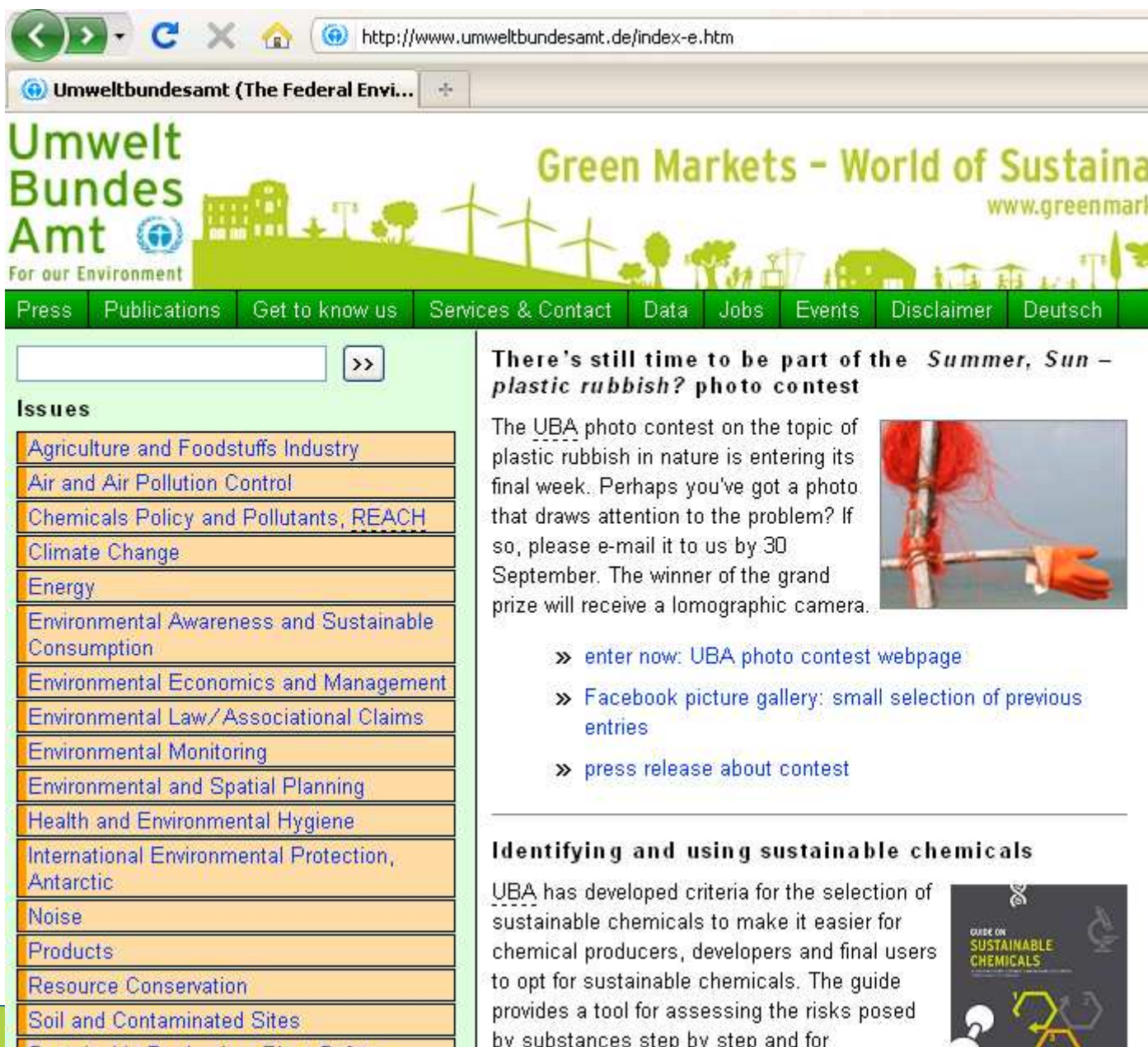
# Structure of Web Sites

The image displays two overlapping web browser windows. The top window is titled 'Umwelt in Sachsen' and shows the URL 'http://www.umwelt.sachsen.de/umwelt/index.html'. The bottom window is titled 'Saarland' and shows the URL 'http://saarland.de/79685.htm'. The 'Saarland' page has a blue header with the word 'Saarland' and a navigation menu with items: 'Saarland-Start', 'Mein Saarland', 'Themenportale', 'Politik & Verwaltung', and 'Land & Tourismus'. A sidebar on the left of the 'Saarland' page lists categories under 'Umwelt': 'Boden, Altlasten', 'Geologie', 'Strahlenschutz', 'Wertstoffe, Abfallwirtschaft', and 'Natur, Biologische Vielfalt'. The main content area of the 'Saarland' page has a 'Wasser' section with a sub-header 'Aktuelle Meldungen - Wasser'. Below this is a list of links: 'Hochwassermeldedienst', 'Hochwasserschutz im Saarland', 'Wasserhaushalt', 'Oberflächengewässer', 'Grundwasser', 'Umgang mit wassergefährdenden Stoffen (VAwS)', 'Abwasser', 'Wasserrahmenrichtlinie (WRRL)', 'Genehmigungen nach dem Wasserrecht', 'Rechtsvorschriften', and 'Nationale und Internationale'. The 'Wasser' section features an article titled 'Ist die Prims für Fische durchgängig?' dated '16.05.2011'. The article includes a photograph of a stream with a wooden structure and a speaker icon. Below the photo, the text reads: 'Seit Oktober 2010 wird im Auftrag des Ministeriums für Umwelt, Energie, Populationsstruktur und –dynamik von Leitfischarten im Primseinzl'.

- Typical application: spam filters
- Objective in the context of PortalU:
  - Integration into *nutch* Web-crawler as a plugin
  - Automatic detection of environmental relevance of single web pages
  - > only the entry points to Web sites have to be provided

- Training Sets
- Creation of dictionaries for each training set
- Training of algorithm
- Testing of algorithm (on part of the training set)
- Evaluation on a test domain

- Two sets of Web Pages.  
both public authorities  
disjunct content with respect to environmental  
relevance
- uba.de (Federal Environment Agency)
- sachsen.de (Saxony) and rlp.de (Rhineland-Palatine)



Umwelt Bundes Amt  
For our Environment

Green Markets - World of Sustainable  
www.greenmark


Press Publications Get to know us Services & Contact Data Jobs Events Disclaimer Deutsch

Issues

- Agriculture and Foodstuffs Industry
- Air and Air Pollution Control
- Chemicals Policy and Pollutants, REACH
- Climate Change
- Energy
- Environmental Awareness and Sustainable Consumption
- Environmental Economics and Management
- Environmental Law/Associational Claims
- Environmental Monitoring
- Environmental and Spatial Planning
- Health and Environmental Hygiene
- International Environmental Protection, Antarctic
- Noise
- Products
- Resource Conservation
- Soil and Contaminated Sites

**There's still time to be part of the *Summer, Sun - plastic rubbish?* photo contest**


The UBA photo contest on the topic of plastic rubbish in nature is entering its final week. Perhaps you've got a photo that draws attention to the problem? If so, please e-mail it to us by 30 September. The winner of the grand prize will receive a lomographic camera.



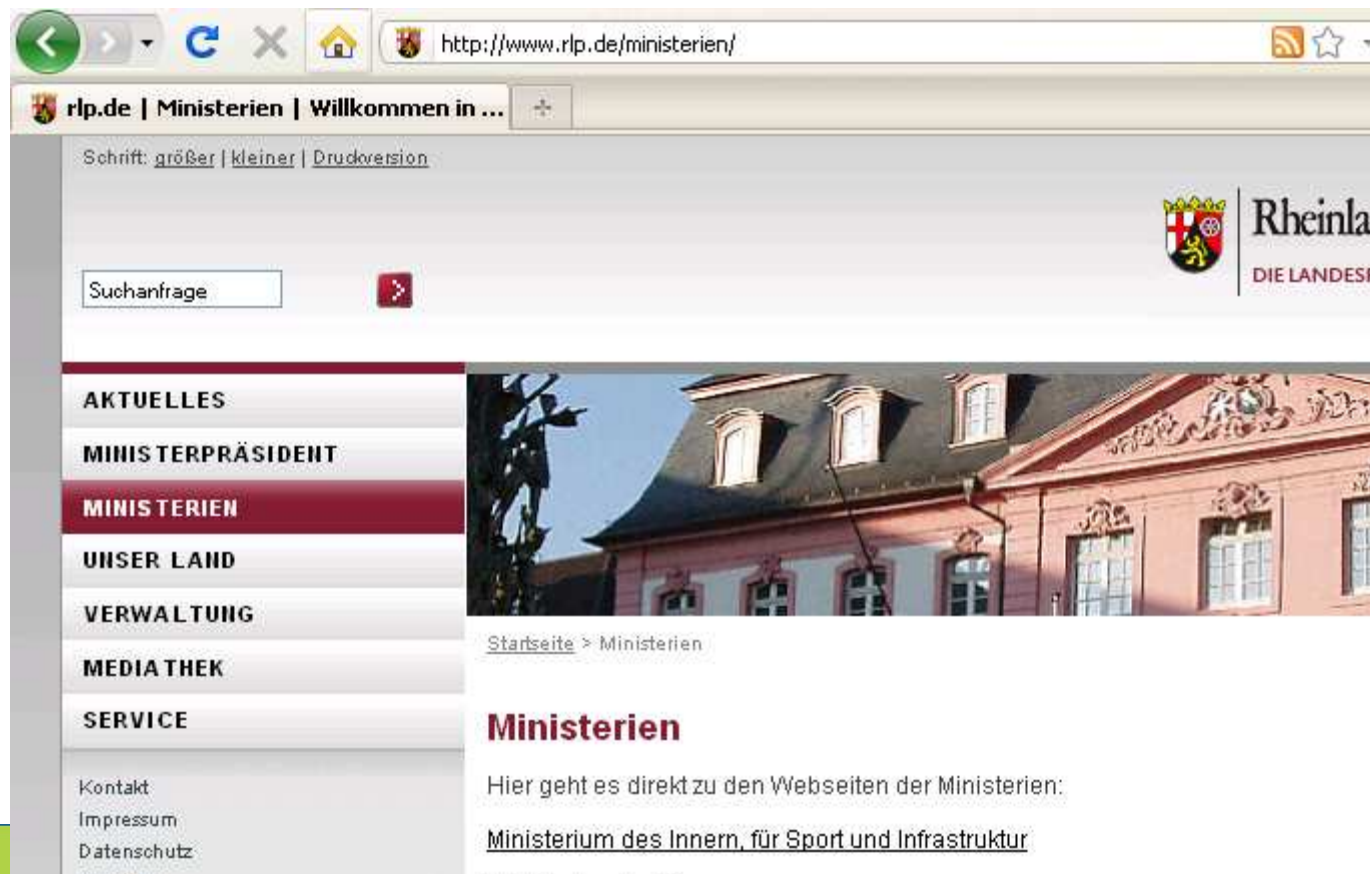
- » enter now: [UBA photo contest webpage](#)
- » [Facebook picture gallery: small selection of previous entries](#)
- » [press release about contest](#)

**Identifying and using sustainable chemicals**

UBA has developed criteria for the selection of sustainable chemicals to make it easier for chemical producers, developers and final users to opt for sustainable chemicals. The guide provides a tool for assessing the risks posed by substances step by step and for



# Training Sets



# Training Sets & Dictionaries

	uba.de (A)	sachsen.de (B1)	uba.de (A)	sachsen.de + rlp.de (B2)
Number of Web pages	5131	11527	5131	21368
Size of dictionary	176	85	153	69

available , insbesondere , texte , nachhaltiger , policy , adresse , development , pollutant , fachbibliothek , toxic , german , stellenangebote , luft , protection , resource , gesamtkatalog , rahmen , luftreinhaltung , bereich , system , consumption , conservation , antarktis , kostenlos , spatial , water , materialien , altlasten , stoffe , plant , verwandte , partnerverlage , publikation , report , chemikalienpolitik , release , mediendatenbank , internationaler , raumbezogene , energie , noise , umweltplanung , umwelt , landwirtschaft , änderung , boden , entwicklung , umweltbundesamt , erhalten , climate , planning , control , nachhaltige , health , safety , anlagensicherheit , produktion , ziel , deutschland , lärm , anforderungen , associational , technische , grundwasser , sites , internationale , rechtliche , related , order , energy , hintergrundpapiere , reach , informationen , trinkwasser , antarctic , schadstoffe , foodstuffs , gesundheit , verkehr , umweltbeobachtung , maßnahmen , umweltökonomie , umweltmanagement , change , jugendliche , sustainable , pollution , reihe , umweltrecht , products , kennnummer , kinder , waste , pollutants , cyanocenter , ressourcenschonung , environment , schutz , nahrungsmittelindustrie , services , umweltbewusstsein , awareness , environmental , papers , federal , richtlinie , economics , hygiene , transport , contaminated , wasser , teil , bewertung , free , background , produkte , konsum , gewässerschutz , umwelthygiene , bestellen , energieeffizienz , abfallwirtschaft , daten , agriculture , umsetzung , verbandsklage , wissen , sachgebiete , beiträge , langfassung , fragen , management , international , aktuelle , chemicals , germany , ergebnisse , neuerscheinungen , know , industry , klimaschutz , data , grundlagen , umweltschutz , monitoring , agency , gesamtlste , umweltinformation , umweltbundesamtes , claims , soil , transfer , ratgeber

- mixture of specialized and general terms (environment is a cross-sectoral topic)
- Mixture of english and german terms
- Manual refinement necessary: Many irrelevant terms that only concern the structure of the Web site have to be removed (e.g. site navigation, dates, Names, ... )

## Procedure

- randomly choose 2/3 of Web pages for training
- use the other 1/3 of Web pages for evaluation

## Typical measures

- Precision:  $tp/(tp + fp)$
- Recall:  $tp/(tp + fn)$
- *F1* (weighted harmonic mean, or F measure)  
trades off precision versus recall
- MicroF1 (micro averaged harmonic mean, well suited  
average value when the document classes differ greatly  
in size)

## Naïve Bayes

	A + B <sub>1</sub>		A + B <sub>2</sub>	
	A	B <sub>1</sub>	A	B <sub>2</sub>
Precision	0.969	0.871	0.918	0.927
Recall	0.656	0.991	0.726	0.982
F1	0.783	0.927	0.811	0.954
<b>MicroF1</b>	<b>0.891</b>		<b>0.925</b>	

## Support Vector Machines

	A + B <sub>1</sub>		A + B <sub>2</sub>	
	A	B <sub>1</sub>	A	B <sub>2</sub>
Precision	0.798	0.761	0.855	0.934
Recall	0.321	0.964	0.713	0.978
F1	0.458	0.851	0.789	0.955
<b>MicroF1</b>	<b>0.766</b>		<b>0.926</b>	

- Test Domain: niedersachsen.de

niedersachsen.de	Total: 141813
pages from relevant sub-domains:	Total: 35570

(pseudo a priori knowledge)

Relevant subdomains:

umwelt, numis, mu, mu1, lbeg, nlwkn, naturparke, weltnaturerbe, wattenmeerstiftung, nlga, nna-en, nna, elbtalaue, life-duemmer, naturschutzgebiete, natur-erleben, umweltbericht, umweltkarten

# Evaluation on Test Domain



The screenshot shows a web browser window with the address bar containing the URL `http://www.umwelt.niedersachsen.de/portal/live.php?navigation_id=2138&psmand=10`. A red rectangular box highlights the domain `www.umwelt.niedersachsen.de`. Below the browser window, the website header is visible, featuring the logo of the Niedersächsisches Ministerium für Umwelt und Klimaschutz (Lower Saxony Ministry for Environment and Climate Protection) and a navigation menu with items like 'Startseite', 'Kontakt', 'RSS', 'Aktuelles', 'Themen', 'Umweltbericht 2010', 'Der Minister', 'Wir über uns', and 'Service'. The main content area displays the text: 'Willkommen im Niedersächsischen Ministerium für Umwelt und Klimaschutz'.

	Naïve Bayes		Support Vector Machines	
	$C(A+B_1)$	$C(A+B_2)$	$C(A+B_1)$	$C(A+B_2)$
Precision	0.674	0.812	0.272	0.628
Recall	0.588	0.608	0.165	0.521
F1	0.628	0.696	0.206	0.569
TP	20917	18666	5883	18516
FP	10098	4315	15715	10953
FN	14653	16904	29687	17054
TN	96145	101929	90528	95290

Evaluation of results is difficult, correctness of results is hard to be proven

e.g.

[http://www.ms.niedersachsen.de/live/live.php?navigation\\_id=5097&article\\_id=12913  
&\\_psmand=17](http://www.ms.niedersachsen.de/live/live.php?navigation_id=5097&article_id=12913&_psmand=17)

Marked as *unrelevant* by Naïve Bayes

Identified as *relevant* by SVM

Not all false positives are incorrect!



http://www.ms.niedersachsen.de/portal/live.php?navigation\_id=5097&article\_id=12913&\_psmand=17

Hintergrundinformation zum Reaktorunglück in Tschernobyl

**Niedersächsisches Ministerium für Soziales, Frauen, Familie, Gesundheit und Integration**

[Startseite](#) | [Inhaltsverzeichnis](#) | [Kontakt](#) | 

[Aktuelles](#) | [Themen](#) | [Die Ministerin](#) | [Wir über uns](#) | [Service](#)

▸ [Navigation](#) ▸ [Themen](#) ▸ [Unsere Stiftungen](#) ▸ [Kinder von Tschernobyl](#)

Themen

- [Soziales](#)
- [Gleichberechtigung / Frauen](#)
- [Familie](#)
- [Bürgerschaftliches Engagement](#)
- [Kinder & Jugendliche](#)
- [Senioren / Generationen](#)
- [Bauen & Wohnen](#)

## Hintergrundinformation zum Reaktorunglück in Tschernobyl

### Der 26. April 1986 und seine Folgen

Als am 26. April 1986 der Block IV des Kernkraftwerkes Tschernobyl im Norden der Ukraine - nahe der weißrussischen Grenze - explodierte, versuchten die Verantwortlichen der damaligen Sowjetunion zunächst, die Folgen herunterzuspielen oder zu vertuschen. Selbst die in unmittelbarer Nähe wohnende Bevölkerung wurde nicht informiert. Mit der Explosion wurden große Mengen

# Again: Results

	Naïve Bayes		Support Vector Machines	
	$C(A+B_1)$	$C(A+B_2)$	$C(A+B_1)$	$C(A+B_2)$
Precision	0.674	0.812	0.272	0.628
Recall	0.588	0.608	0.165	0.521
F1	0.628	0.696	0.206	0.569
TP	20917	18666	5883	18516
FP	10098	4315	15715	10953
FN	14653	16904	29687	17054
TN	96145	101929	90528	95290

- Precise a priori knowledge on large sets of documents is needed (like reuter text corpus)
- Tests with other / larger training sets
- Classifying Web pages with respect to environmental themes