

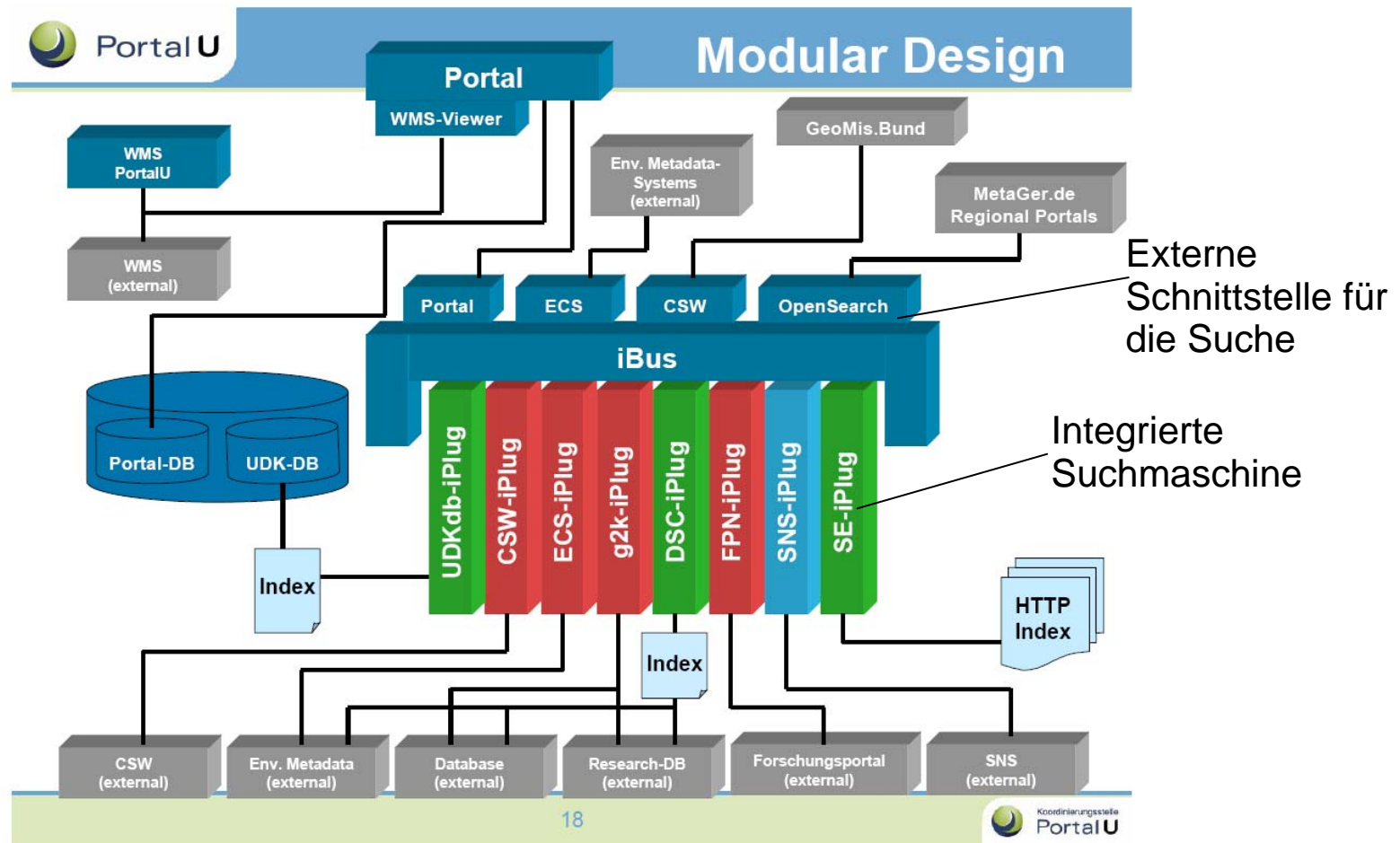
Umweltportal Deutschland PortalU Suchmaschine

Internetgruppentreffen 23.10.2007



Dr. Maria Rüther, I 1.5
maria.ruether@uba.de

Gesamtarchitektur PortalU



PortalU Suchmaschine

Nutch <http://lucene.apache.org/nutch/>

- OpenSource Software für webbasierte Suche
- Nutch setzt auf Apache Lucene Volltext Suchmaschine auf und ergänzt z.B. crawler und Parser für verschiedene Dokumentenformate

<http://lucene.apache.org/java/docs/index.html>

Nutch ist in die Gesamtarchitektur von PortalU integriert

Im PortalU Index berücksichtigte Formate

- Alle Formate, die html ausliefern (html, htm, php, jsp, ...)
- PDF
- Word
- PowerPoint
- RSS

Es werden von Nutch weitere Formate unterstützt, (z.B. MP3, zip, ...)

<http://wiki.apache.org/nutch/Features>

Durchsuchte Domainen

Webseiten-Angebot in PortalU

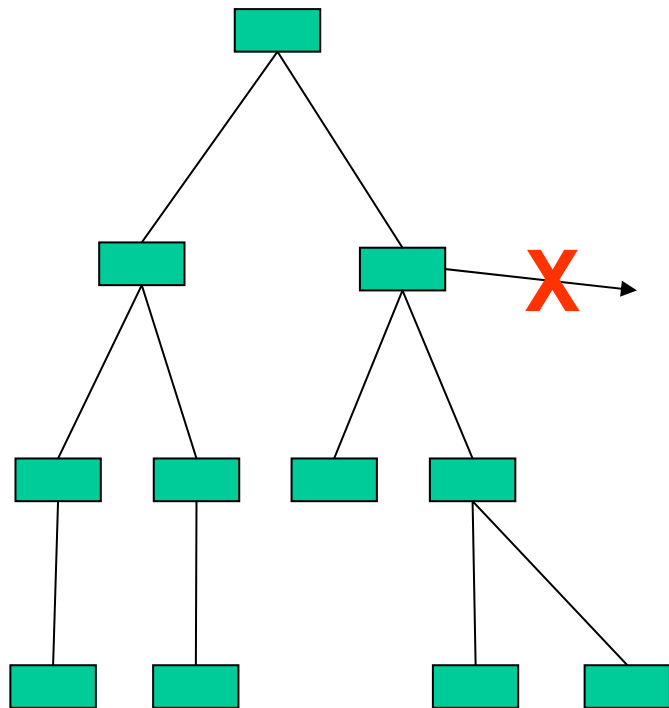
- Ca. 180 Partner in PortalU
- ca. 1.000.000 Webseiten aus verschiedenen Domainen

Pflege des Webindex

- Partner pflegen über eine Administrationsoberfläche die zu durchsuchenden Domainen
- Festzulegen sind StartURLs, LimitURLs, ExcludeURLs

Begriffsbestimmung: StartURL und LimitURL

Website: www.behoeerde.de



StartURL= Ausgangspunkt für das „Abcrawlen“ eines Webauftritts;

Gute StartURLs: Startseite, SiteMap

Schlechte StartURLs: isolierte Seiten weit unten in der Hierarchie

LimitURL= Beschränkt das „Abcrawlen“ auf den angegebenen Webauftritt (Domain), oder Teile desselben

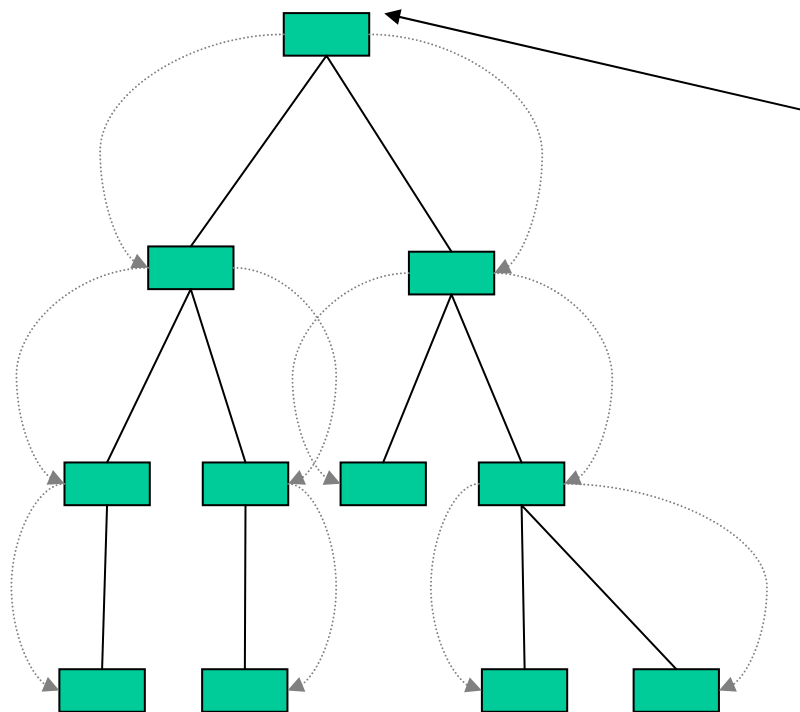
Links, die aus der Domäne herauszeigen, werden nicht verfolgt.

Typische LimitURL: Hauptdomäne

Quelle: KSt. PortalU

Indizieren eines kompletten Webauftritts

Website: www.behoerde.de



StartURL: www.behoerde.de/index.htm

LimitURL: www.behoerde.de/

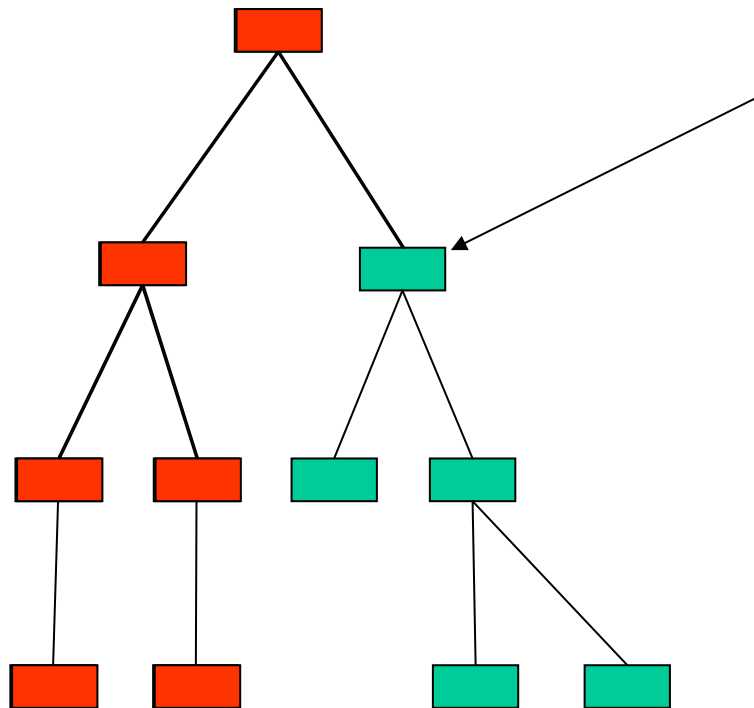
→ Gesamte Site (Baum) wird indiziert

Der Webauftritt wird Ebene für Ebene abgescrawled
(Breitensuche)

Quelle: KSt. PortalU

Indizieren eines Teil-Baums

Website: www.behoerde.de



StartURL: www.behoerde.de/umwelt/index.htm

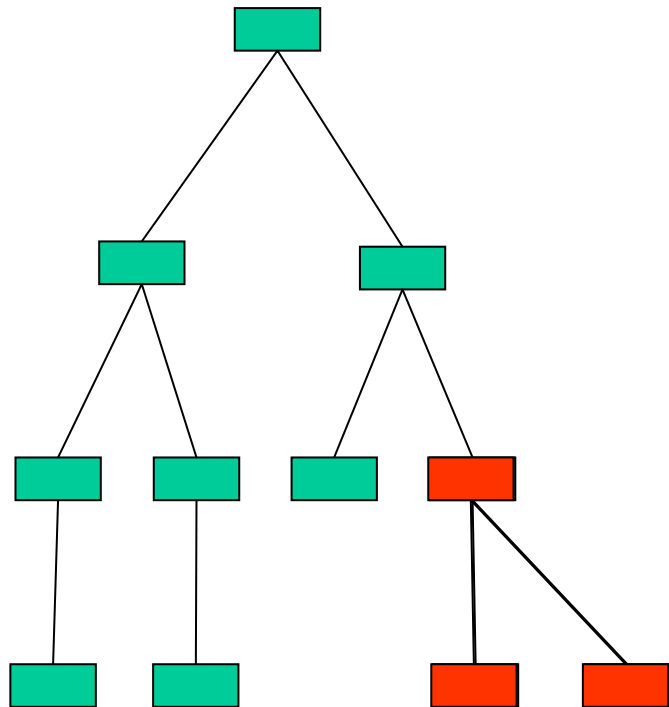
LimitURL: www.behoerde.de/umwelt/

→ Nur der Teil-Baum bzw. Zweig „/umwelt/“ wird „abgecrawled“

Quelle: KSt. PortalU

Die ExcludeURL

Website: www.behoerde.de



ExcludeURL: Legt fest, welche Seiten NICHT indiziert werden sollen.

Kann sowohl auf eine einzelne Seite als auch auf ein „Unterverzeichnis“ verweisen

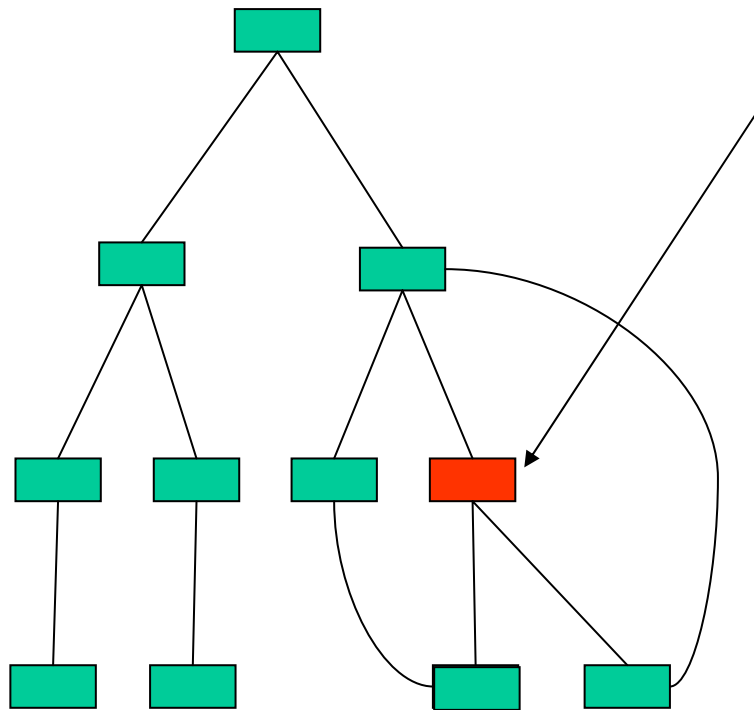
ExcludeURL: www.behoerde.de/umwelt/finnisch/

→ schneidet gesamten Zweig ab

Quelle: KSt. PortalU

Anwenden von ExcludeURLs (1/2)

Website: www.behoerde.de



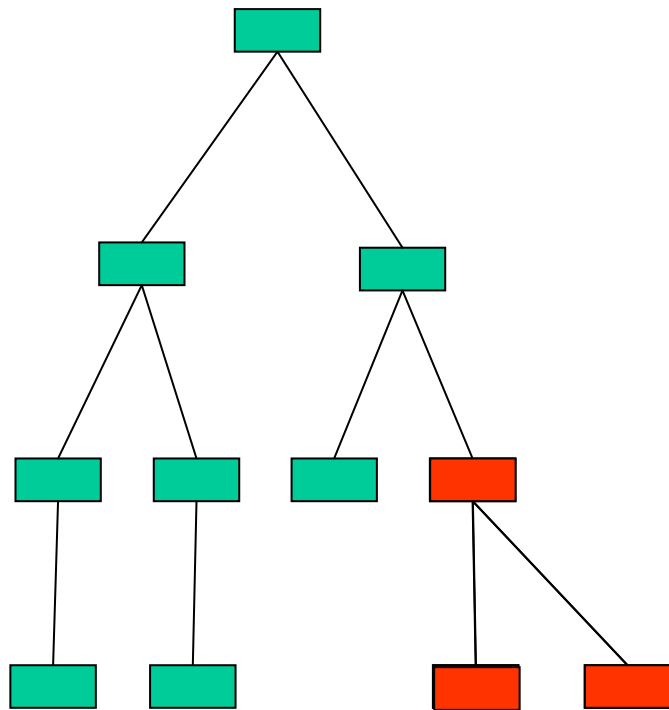
ExcludeURL: www.behoerde.de/umwelt/finnisch/dok.html

→ schließt einzelne Seite aus

Quelle: KSt. PortalU

Anwenden von ExcludeURLs (2/2)

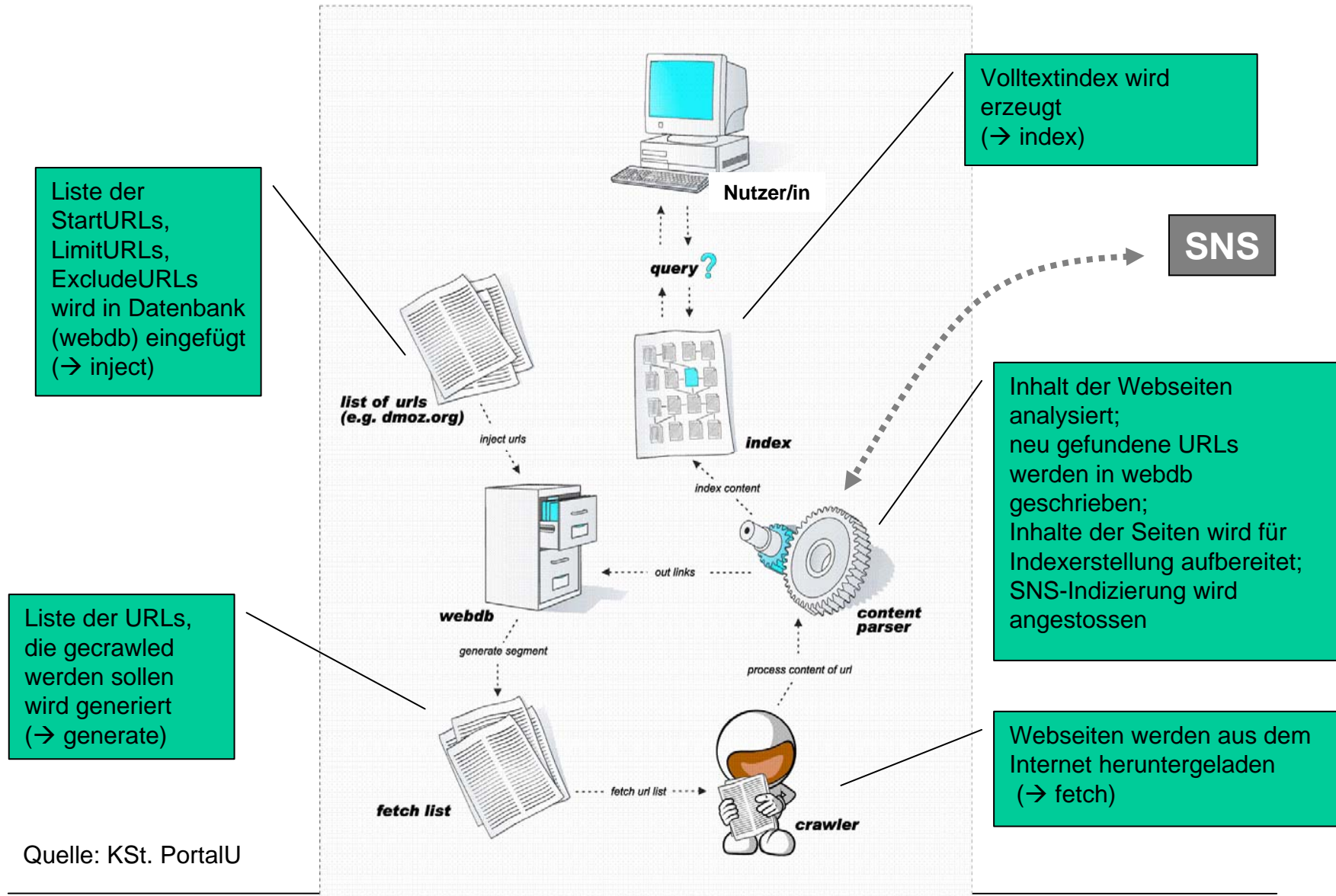
Website: www.behoerde.de



Aber: Entfernt ganzen Baum, wenn keine Verlinkung

Quelle: KSt. PortalU

Ablaufschema der Indexerstellung



Quelle: KSt. PortalU

Nutch Administration - Microsoft Internet Explorer

Adresse <http://harrison.its-technidata.de/plug-admin-se/general/pages/index.jsp> Wechseln zu

PortalU Administration Suchmaschine

Url-Pflege Index-Pflege

<< Zurück zu Partner/Anbieter Auswahl | Katalog-URLs bearbeiten | Neue Katalogseite | Web-URLs bearbeiten | **Neue Webseite** | Die letzten Test-Resultate anzeigen |

Neue Webseite: 1. Schritt - Start-URL definieren

URL:

Limit-URLs erfassen

Hilfe

Geben Sie die **Start-URL** ein.

PortalU | Copyright © Koordinierungsstelle PortalU im Nds. Umweltministerium
Alle Rechte vorbehalten

Nutch Administration - Microsoft Internet Explorer

Adresse <http://harrison.its-technidata.de/iplug-admin-se/general/pages/index.jsp> Wechseln zu

Portal U Administration Suchmaschine

Url-Pflege Index-Pflege

<< Zurück zu Partner/Anbieter Auswahl | Katalog-URLs bearbeiten | Neue Katalogseite | Web-URLs bearbeiten | **Neue Webseite** | Die letzten Test-Resultate anzeigen

Neue Webseite: 2. Schritt - Limit-URLs definieren

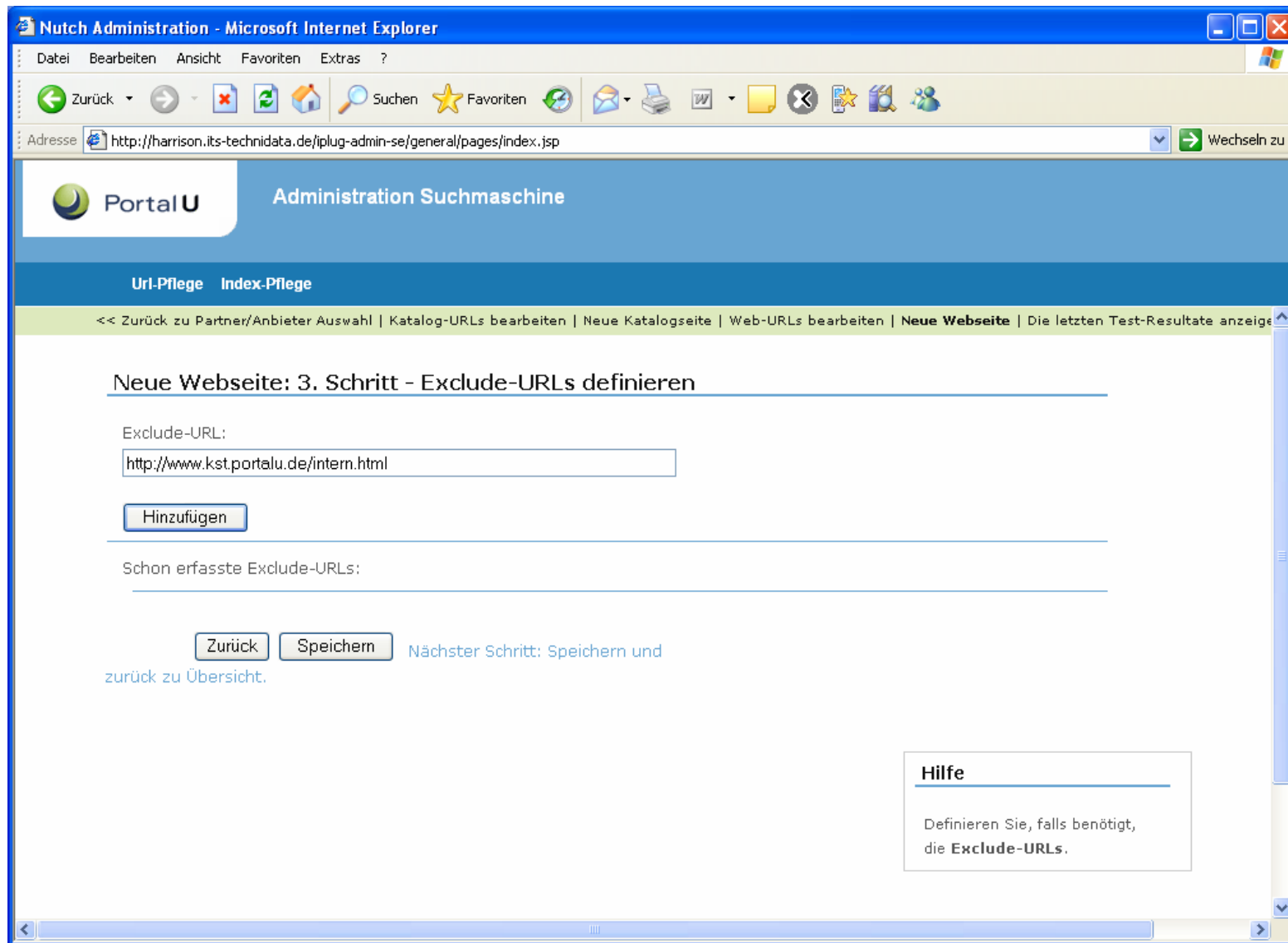
Limit-URL:

Sprache:

Eigenschaften:
 Ist eine Forschungsseite.

Schon erfasste Limit URLs:

Nächster Schritt: Exclude-URLs erfassen



Nutch Administration - Microsoft Internet Explorer

Adresse <http://harrison.its-technidata.de/plugin-admin-se/general/pages/index.jsp>

Portal U Administration Suchmaschine

Url-Pflege Index-Pflege

<< Zurück zu Partner/Anbieter Auswahl | Katalog-URLs bearbeiten | Neue Katalogseite | **Web-URLs bearbeiten** | Neue Webseite | Die letzten Test-Resultate anzeigen

Web-URLs bearbeiten

Start-URL	Aktionen
http://www.kst.portalu.de/index.html	DEL EDIT TEST
Limit-URLs	
http://www.kst.portalu.de/lang:de	
Exclude-URLs	
http://www.kst.portalu.de/intern.html	
Start-URL	Aktionen
http://www.badegewaesser.nlga.niedersachsen.de/master/C11523231_L20_D0.html	DEL EDIT TEST
Limit-URLs	
http://www.badegewaesser.nlga.niedersachsen.de/datatype:default www	
Start-URL	Aktionen
http://www.umwelt.niedersachsen.de/master/C4108003_N4107975_L20_D0_I598.html	DEL EDIT TEST
Limit-URLs	
http://www.umwelt.niedersachsen.de/	

http://harrison.its-technidata.de - PortalU - Administration Suchmaschine - Mozilla Firefox

Status	Beschreibung
0. Test run...	Der Test beginnt...
1. crawldb.inject	Die Urls werden vorbereitet...
2. bwdb.inject	Weitere Urls (Begrenzungsurls) werden vorbereitet...
3. generate	Die Urls werden in ein Segment (0)tes generiert...
4. fetch	Die Urls werden heruntergeladen und geparkt...
5. update	Die geladenen und extrahierten Urls werden in die Crawldb geschrieben...
6. linkdb	Der Linkgraph wird erstellt...
7. index	Die geladenen Urls werden indexiert
8. dedup	Es werden Urls mit doppelten Content entfernt...
9. crawl.done	Der Crawl ist beendet.

Url	Status
http://www.kst.portalu.de/archiv/index.html	Die Url wurde noch nicht gefetcht. (UNFETCHED)
http://www.kst.portalu.de/impress.html	Die Url wurde noch nicht gefetcht. (UNFETCHED)
http://www.kst.portalu.de/index.html	Die Url wurde erfolgreich gefetcht. (FETCHED)
http://www.kst.portalu.de/kontakt.html	Die Url wurde noch nicht gefetcht. (UNFETCHED)
http://www.kst.portalu.de/portalu/download.html	Die Url wurde noch nicht gefetcht. (UNFETCHED)
http://www.kst.portalu.de/portalu/index.html	Die Url wurde noch nicht gefetcht. (UNFETCHED)
http://www.kst.portalu.de/pub/index.html	Die Url wurde noch nicht gefetcht. (UNFETCHED)
http://www.kst.portalu.de/sitemap.html	Die Url wurde noch nicht gefetcht. (UNFETCHED)
http://www.kst.portalu.de/udk/index.html	Die Url wurde noch nicht gefetcht. (UNFETCHED)
http://www.kst.portalu.de/ueberuns/index.html	Die Url wurde noch nicht gefetcht. (UNFETCHED)

Fertig

Informationsanbieter

Partner

- 16 Länder
- Bund

Provider, z.B.

- Umweltbundesamt: provider:bu_uba
- RAL/Blauer Engel: provider:bu_blauerengel
- SRU: provider:bu_sru

UBA Informationsangebot

Online Administration, User/PW erforderlich:

<http://plant.its-technidata.de/iplug-admin-se-index/general/pages/index.jsp?tab=admin-url-maintenance>)

StartURLs:

- <http://www.apug.de/>
- <http://www.bvt.umweltbundesamt.de/>
- <http://www.cleaner-production.de/wwwcpg/content-frameset.php?lang=de>
- http://www.dehst.de/cln_027/nn_76410/DE/Home/homepage_node.html_nnn=true
- <http://www.diffuse-quellen.prtr.de>
- <http://www.env-it.de/luftdaten/start.fwd>
- <http://www.env-it.de/umweltdaten/jsp/documentList.do?event=show&catalogueId=0>
- <http://www.klimaschuetzen.de/>
- <http://www.ltws.de/>
- <http://www.POP-DioxinDB.de>
- <http://www.probas.umweltbundesamt.de/php/sitemap.php?>
- <http://www.reach-info.de/>
- <http://www.reach-konferenz.de/>
- <http://www.stoffdaten-stars.de/>
- <http://www.umweltbundesamt.de/>
- <http://www.umweltdaten.de/umweltbeobachtung/konkret/start.htm>
- <http://www.umweltprobenbank.de>
- <http://www.wasser-agenda.de/>

UBA Informationsangebot

Online Administration, User/PW erforderlich:

<http://plant.its-technidata.de/iplug-admin-se-index/general/pages/index.jsp?tab=admin-url-maintenance>)

StartURLs (Forts.)

- <http://osiris.uba.de/gisdienste/Herata/hmetal/>
- <http://osiris.uba.de/gisdienste/Herata/npbilanz/>
- <http://osiris.uba.de/gisdienste/Kompass/index.htm>
- <http://www.eper.de/>
- <http://www.forum.prtr.de>
- <http://www.home.eper.de>
- <http://www.home.prtr.de>
- <http://www.prtr.de/>

Parametrisierte Suche – Beispiele (1)

- PortalU-Suche über Suchfenster aus den eigenen Webseiten heraus
- Parameter gemäß PortalU-Query Syntax
Betriebshandbuch, Query-Syntax für die Suche in PortalU (InGrid-Portal)
[http://www.kst.portalu.de/wiki/index.php/PortalU® Querysyntax](http://www.kst.portalu.de/wiki/index.php/PortalU%20Querysyntax) (User/PW erforderlich)
- Ergebnis-Rückgabe im PortalU-Layout
 - <http://www.kst.portalu.de/>
 - <http://www.kst.portalu.de/udk/udklinks.html>
- Erläuterung



Microsoft
Word-Dokument

OpenSearch-Schnittstelle

- Externe Nutzung der Suche
- Basiert auf der OpenSearch Spezifikation
<http://opensearch.a9.com>
- Anfrage über HTTP/GET mit 2 Methoden:
 - Absetzen einer Suche „query“
 - Anfordern von Detailedaten „detail“ (z.B. für ein Metadatenobjekt)
- Ergebnis-Rückgabe im RSS2.0 kompatiblen Format

Quellen:

Betriebshandbuch, OpenSearch-Schnittstelle

- <http://www.kst.portalu.de/wiki/index.php/OpenSearch-Schnittstelle> (User/PW erforderlich)

Betriebshandbuch, Query-Syntax für die Suche in PortalU (InGrid-Portal)

- http://www.kst.portalu.de/wiki/index.php/PortalU®_Querysyntax (User/PW erforderlich)

OpenSearch Schnittstelle - Beispiele (2)

GET-Request und XML-Rückgabe

- Suche im Webindex = linke Ergebnisspalte in PortalU
 - http://www.portalu.de/opensearch/query?q=wasser+datatype:default+provider:bu_uba+ranking:score

Suche *nur* in Metadaten (hier: UDK-UBA)

- http://www.portalu.de/opensearch/query?q=Wasser+datatype:metadata+provider:bu_u ba+ranking:score
- Suche *nur* in Datenbanken „g2k“-Schnittstelle
 - <http://www.portalu.de/opensearch/query?q=Wasser+datatype:g2k+ranking:any>

➤ XML-Rückgabe kann über Stylesheets an das eigene Layout angepasst werden

Ende ...