

Portal-U InGrid 1.0

DV-technisches Feinkonzept Portal-U / InGrid 1.0
Migrationskonzept

Version 1.0

Stand 01.07.2005

Erstellt durch das Konsortium

 GIS-tec

GISTec GmbH

Rundeturmstraße 12
64283 Darmstadt
Tel (+49)6151 / 155-250
Fax (+49)6151 / 155-259
Mail info@gistec-online.de
<http://www.gistec-online.de>



Fraunhofer Institut für Graphische Datenverarbeitung

Fraunhoferstraße 5
64283 Darmstadt
Tel (+49)6151 / 155-0
Fax (+49)6151 / 155-199
Mail info@igd.fhg.de
<http://www.igd.fhg.de>



media style GmbH

Mansfelder Straße 13
06108 Halle
Tel (+49)345 / 5222702
Fax (+49)345 / 5222719
Mail info@media-style.com
<http://www.media-style.com>

 wemove

wemove digital solutions GmbH

Eschersheimer Landstraße 5-7
60322 Frankfurt
Tel (+49)69 / 759003-0
Fax (+49)69 / 759003-22
Mail info@wemove.com
<http://www.wemove.com>

Im Auftrag der Koordinierungsstelle UDK/GEIN



**Niedersächsisches Umweltministerium –
Koordinierungsstelle UDK/GEIN**

Archivstrasse 2
30167 Hannover
Tel (+49)511 / 120-3480
Fax (+49)511 / 120-3697
Mail kug@numis.niedersachsen.de
<http://www.udk-gein.de>

Dokumentenhistorie

Version	Datum	Bearbeiter	Anlass	Status
1.0	20.06.2005	GIStec / media style / wemove	Phase 1 Portal-U	FINAL

Inhaltsverzeichnis

1	Einleitung	4
2	Zielsetzung	5
3	Migrationskonzept	6
3.1	Vorhanden Daten: gein® und UDK-Metadaten	6
3.1.1	Portal	6
3.1.2	Suchmaschine	6
3.1.3	Windows-UDK	7
3.2	Migration	7
3.2.1	Portal	7
3.2.2	Suchmaschinen	8
3.2.3	UDK	8

1 Einleitung

Das Umweltinformationsnetz Deutschland gein® (German Environmental Information Network) ist zurzeit "das Portal für Umweltfragen" in Deutschland, welches im Internet verteilte Informationen zahlreicher öffentlicher Einrichtungen erschließt. Es wurde vom Umweltbundesamt (UBA) in den Jahren 1998 bis 2000 aufgebaut und bis Ende 2002 weiterentwickelt. Beim Umweltdatenkatalog (UDK) handelt es sich um das Informationssystem zum Nachweis von Umweltinformationen der öffentlichen Verwaltung in Deutschland. Seine Grundkonzeption entstand in den Jahren 1991 bis 1995 im Niedersächsischen Umweltministerium.

Im Jahre 2003 wurden gein® und UDK organisatorisch zusammengeführt. Beide Systeme sollen nun konzeptionell und technologisch in der Anwendung InGrid 1.0 (Information Grid, ehemaliger Arbeitstitel: gein® 2.0) integriert werden. Mit Hilfe von InGrid 1.0 soll das zukünftige "Portal für Umweltfragen" Portal-U aufgebaut werden.

Das Migrationskonzept des DV-technischen Feinkonzepts „Portal-U / InGrid 1.0“ erläutert die Vorgehensweise zur Migration der bestehenden Daten aus gein® und dem WWW-UDK 5. Der Schwerpunkt liegt hierbei auf den vorliegenden Portaldaten, den URLs für die Suchmaschine und die vorhandenen UDK-Metadaten.

In dem vorliegenden Migrationskonzept wird zunächst die Zielsetzung der Migration erläutert um dann in Kapitel 3 die zu migrierenden Daten aufzuführen. Im anschließenden Kapitel wird auf die eigentliche Migration eingegangen und die Vorgehensweise beschrieben.

2 Zielsetzung

Das Modul Migrationskonzept soll aufzeigen, wie die vorhandenen Informationen von gein® sowie die UDK-Datenbestände der Länder und des Bundes in Portal-U übernommen werden können. Hierbei wird besonders auf folgende Punkte eingegangen:

- Übernahme der administrativen Angaben
- Übernahme der UDK Metadaten

3 Migrationskonzept

3.1 Vorhanden Daten: gein® und UDK-Metadaten

Im Folgenden werden die Informationen aufgeführt, die in gein® vorhanden sind und nach Portal-U migriert werden sollen. Die Informationen sind in die drei Kategorien „Portal“, „Suchmaschine“ und „UDK-Metadaten“ untergliedert.

3.1.1 Portal

Anbieter-Informationen

- Name und Bezeichnung des Anbieters
- URL der Homepage,
- Ggf. Logo bzw. Pfad zum Logo (ist in gein® als Bilddatei abgelegt)

Informationen über Master/Submaster der jeweiligen Anbieter

- Name, Vorname des Masters/Submasters
- Adresse (Strasse, PLZ, Ort)
- Tel.
- E-mail
- Benutzername und Passwort

3.1.2 Suchmaschine

Informationen zu den gemeldeten Internetseiten der Anbieter:

- Start-URL
- Limit-URL
- Exclude-URLs
- Anbieterkennung

Informationen zu den von den Anbietern gemeldeten News- und Messwertseiten

- Die URL der News / Messwertseiten
- Titel und Beschreibung der Seiten
- Der Typ der Seiten (News oder Messwert)
- Anbieterkennung

Informationen zu den von einem Anbieter gemeldeten Themenseiten

- Die URL der Themenseite
- Untertitel und, falls kein automatisch erzeugtes Abstract angezeigt werden soll, Beschreibung der Seite
- Die Klassifizierung (UBA Umweltklassen) der Themenseite. Um eine möglichst automatische Übernahme der Themenseiten von gein® in Portal-U zu ermöglichen, wird vom AG eine Mapping-Tabelle erstellt. Diese bildet die bisher verwendeten UBA Umweltklassen auf das neue System (Thematische Kategorien, Funktionale Kategorien) ab. Es soll möglich sein, Themenseiten, die nicht automatisch gemappt werden können (für die z.B. keine funktionale Kategorie automatisch ermittelt werden kann), sondern per Hand nachbearbeitet werden müssen, zu kennzeichnen bzw. gesondert aufzulisten.
- Anbieterkennung

3.1.3 Windows-UDK

Die erfassten Metadaten des Windows-UDK 5.0 liegen in diversen Datenbanksystemen vor. Diese können mit Hilfe der bestehenden WinUDK 5.0 Exportschnittstelle in semantisches XML exportiert werden.

3.2 Migration

Die Migration erfolgt aus den in Kapitel 3.1 beschriebenen vorhandenen Daten. Diese Daten werden in Phase 2 dem Auftragnehmer nach Absprache mit dem AG in geeigneter Form zur Verfügung gestellt, damit sie gemäß den folgenden Kapiteln in Portal-U integriert werden können.

3.2.1 Portal

Die aus gein® übernommenen Daten der Anbieter werden in die Datenstrukturen von Portal-U übernommen. Die Master/Submaster-Systematik aus gein® wird in Absprache mit dem Auftraggeber auf die Partner/Infoanbieter-Datenstruktur von Portal-U abgebildet.

Die zu übernehmenden Daten werden aus gein® in CSV-Dateien überführt und in die Benutzer-Datenbank importiert. Ziel ist es, dass alle Benutzer mit einem administrativen Account unter gein® sich am Portal-U einloggen, ihre Daten vorfinden und mit diesen arbeiten können. Eine nachträgliche Importierung ausgewählter Benutzer ist über die CSV-Dateien ebenfalls möglich.

Die Master/Submaster Systematik wird in Absprache mit dem Auftraggeber auf die im Portal-U verwendete Rechte-Struktur abgebildet.

Die Angaben zu den Postschnittstellen können manuell eingepflegt werden, da es sich nur um eine relativ geringe Anzahl von DBs handelt.

3.2.2 Suchmaschinen

Um die URLs der einzelnen Anbieter inkl. Ihrer Metadaten zu übernehmen, wird für den Export ein User mit Leserechten für die derzeitige verwendete Datenbank benötigt. Für die Weiterverarbeitung der Daten wird ein Import Tool realisiert.

Mittels SQL Queries können die jeweilig gewünschten Daten in einer benötigten Struktur extrahiert und als CSV ausgegeben werden. So können zu den Anbietern die URLs und abhängig vom Indexaufbau beliebige Metadaten wie Titel und Beschreibung, Thema, Typ Aktuelles/Messwert, UBA Klasse als Text extrahiert werden.

Importer

Mit Ausnahme von Limit- und Exclude URLs müssen die Daten in die Nutch WebDB übernommen werden. Hierfür wird ein Importwerkzeug realisiert, das die CSV-Dateien parst und in die WebDB injizieren kann. Über eine Konfigurationsdatei ist einstellbar, welche „Spalte“ der CSV welchem Metadatenfeld entspricht.

Das Importer-Tool kann für die Migration CommandLine-basiert genutzt werden. Weiterhin wird es in die Administrationsoberfläche des Datasource Clients eingebunden werden, um auch später mittels eines CSV-Uploads mehrere Datensätze in einem Schritt anzulegen.

Limit- und Exclude URLs sind pro Suchmaschinen Datasource Client (also pro Anbieter) global gültig. Die o.g. CSV Dateien können im DSC weiterverwendet werden.

3.2.3 UDK

Die Windows-UDK 5.0 Metadaten, die als semantische XML-Dateien vorliegen, können mit dem InGrid XML-Importer in die entsprechende Metadatenbank importiert werden. Das Importwerkzeug ist in dem Feinkonzept iBus beschrieben.

Da das physikalische Datenbankmodell zwischen UDK 5.0 und InGrid 1.0 weitgehend identisch ist, ist es weiterhin möglich das iPlug UdkDB parallel zu einem bereits installierten WWW-UDK 5.0 zu installieren. Damit können einzelne Partner für eine Übergangszeit ihre bereits installierten WWW-UDK 5.0 weiter betreiben und ihre Daten parallel Portal-U zur Verfügung stellen. Eine Migration der Partner während Jahresfrist auf InGrid 1.0 wird dennoch empfohlen.