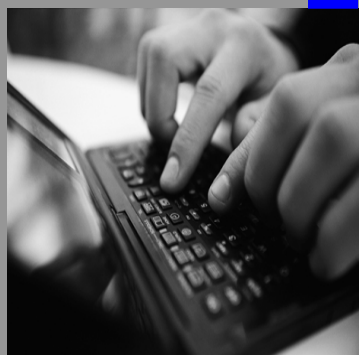


How-to Guide
SAP NetWeaver '04s

How To Extract DC Metadata from Office Documents for Indexing and Searching

Version 1.00 – May 2006

Applicable Releases:
SAP NetWeaver 2004s

THE BEST-RUN BUSINESSES RUN SAP

© Copyright 2006 SAP AG. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or for any purpose without the express permission of SAP AG. The information contained herein may be changed without prior notice.

Some software products marketed by SAP AG and its distributors contain proprietary software components of other software vendors.

Microsoft, Windows, Outlook, and PowerPoint are registered trademarks of Microsoft Corporation.

IBM, DB2, DB2 Universal Database, OS/2, Parallel Sysplex, MVS/ESA, AIX, S/390, AS/400, OS/390, OS/400, iSeries, pSeries, xSeries, zSeries, z/OS, AFP, Intelligent Miner, WebSphere, Netfinity, Tivoli, and Informix are trademarks or registered trademarks of IBM Corporation in the United States and/or other countries.

Oracle is a registered trademark of Oracle Corporation.

UNIX, X/Open, OSF/1, and Motif are registered trademarks of the Open Group.

Citrix, ICA, Program Neighborhood, MetaFrame, WinFrame, VideoFrame, and MultiWin are trademarks or registered trademarks of Citrix Systems, Inc.

HTML, XML, XHTML and W3C are trademarks or registered trademarks of W3C®, World Wide Web Consortium, Massachusetts Institute of Technology.

Java is a registered trademark of Sun Microsystems, Inc.

JavaScript is a registered trademark of Sun Microsystems, Inc., used under license for technology invented and implemented by Netscape.

MaxDB is a trademark of MySQL AB, Sweden.

SAP, R/3, mySAP, mySAP.com, xApps, xApp, and other SAP products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of SAP AG in Germany and in several other countries all over the world. All other product and service names mentioned are the trademarks of their respective companies. Data contained in this document serves informational purposes only. National product specifications may vary.

These materials are subject to change without notice. These materials are provided by SAP AG and its affiliated companies ("SAP Group") for informational purposes only, without representation or warranty of any kind, and SAP Group shall not be liable for errors or omissions with respect to the materials. The only warranties for SAP Group products and services are those that are set forth in the express warranty statements accompanying such products and services, if any. Nothing herein should be construed as constituting an additional warranty.

These materials are provided “as is” without a warranty of any kind, either express or implied, including but not limited to, the implied warranties of merchantability, fitness for a particular purpose, or non-infringement. SAP shall not be liable for damages of any kind including without limitation direct, special, indirect, or consequential damages that may result from the use of these materials.
SAP does not warrant the accuracy or completeness of the information, text, graphics, links or other items contained within these materials. SAP has no control over the information that you may access through the use of hot links contained in these materials and does not endorse your use of third party web pages nor provide any warranty whatsoever relating to third party web pages.
SAP NetWeaver “How-to” Guides are intended to simplify the product implementation. While specific product features and procedures typically are explained in a practical business context, it is not implied that those features and procedures are the only approach in solving a specific business problem using SAP NetWeaver. Should you wish to receive additional information, clarification or support, please refer to SAP Consulting.
Any software coding and/or code lines / strings (“Code”) included in this documentation are only examples and are not intended to be used in a productive system environment. The Code is only intended better explain and visualize the syntax and phrasing rules of certain coding. SAP does not warrant the correctness and completeness of the Code given herein, and SAP shall not be liable for errors or damages caused by the usage of the Code, except if such damages were caused by SAP intentionally or grossly negligent.

1 Scenario

In the process of indexing documents in KMC, attributes of documents are indexed, stored in the indexes and can then be searched. In the standard system, only KMC attributes are indexable and searchable.
Customers often want the attributes that are stored within their MS Office to be indexed as well. KMC does not support the extraction of these attributes – they are not automatically created in KMC. These attributes are also referred to as Dublin Core (DC) metadata. You can find them when you go to *File – Properties* within an MS Office document.

This paper describes how you can extract this DC metadata (attributes) using TREX functions. You can then use the KMC search interface to search for attributes. However, the attributes are not created in KMC and therefore cannot be maintained using KMC; they can only be displayed in the properties in the *Details* dialog of a resource.
TREX provides a Python extension that allows indexing of DC metadata attributes. In this scenario, we use this feature to index a document containing DC attributes. In a second step, we create a predefined property in KMC that allows users to search for these DC attributes.

2 Introduction

Before going into detail, here is an overview of the steps to be done:

- Prerequisites:
  Enterprise Portal Installation: Netweaver 2004 SPS 18
  Netweaver 2004s SPS 9
- Activate the Python extension in the TREX installation for Dublin Core metadata
- Create an index using KMC functions
- Activate the Dublin Core Python extension for this specific index in TREX
- Index the file containing the DC metadata
- In KMC, create the property to display DC metadata in the same way as KMC properties are displayed
- Create the search option set containing search fields for the extracted metadata with dropdown lists to display default values. Insert this search option set into a search iView
- Start the search using the DC metadata as the search string

Note: Default values for the DC metadata can only be provided with TREX 7.0 or higher. This feature is based on KMC guided navigation functions that do not work with TREX 6.1.

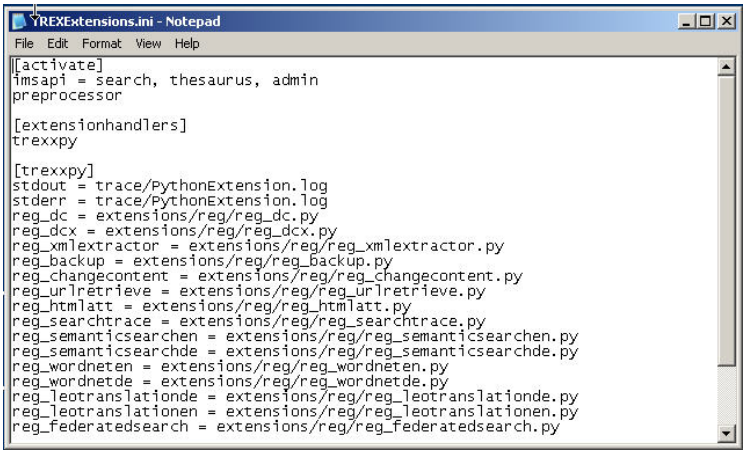
For information and terminology on Dublin Core metadata, see http://en.wikipedia.org/wiki/Dublin_Core.

3 The Step-By-Step Solution

First, you have to configure TREX to enable the extraction and indexing of the DC metadata. You have to activate the specific Python extension. If you now index a file containing DC metadata, the metadata is indexed as attributes of the documents.
You can only activate the Python extension for an existing index – global activation is not possible.
Secondly, you have to make these attributes available in KMC. You have to create predefined properties and provide them in a search option set to make them searchable.

3.1 Configure TREX

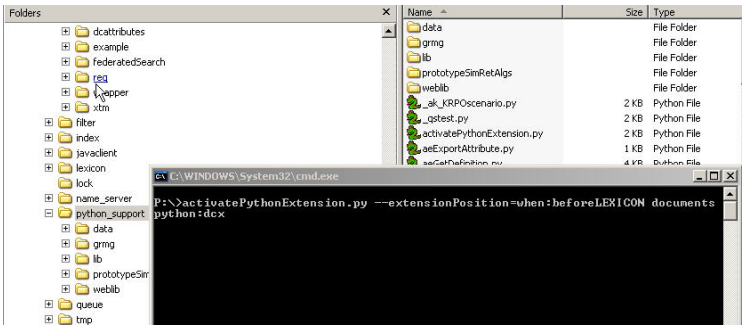
1. Edit the *TREX directory/TREXExtensions.ini* file and activate extension handling by removing the # in the [extensionhandlers] section. This causes *trexxpy.dll* to be loaded.



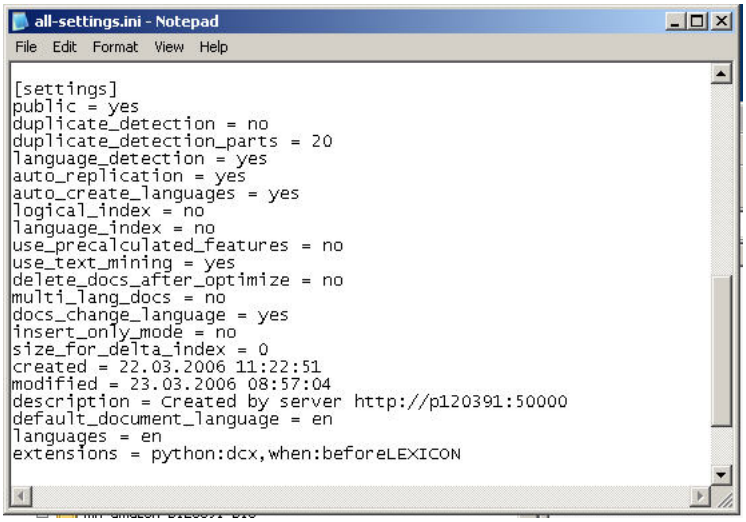
2. Create an index using KMC. Later, you assign documents containing DC metadata to this index.
3. In TREX, execute the *activatePythonExtension.py* script using the following parameter: *--extensionPosition=when:beforeLEXICON <indexId> python:dcx*

   This creates the entry *extensions=python:dcx, when:beforeLEXICON* in the *all-settings.ini* file for your index.
4. Restart the TREX preprocessor. You can do this by ending the *TREXPreprocessor.exe* process in Windows Task Manager. The TREX Daemon restarts this process.

   For more information about the Python extension that you have just

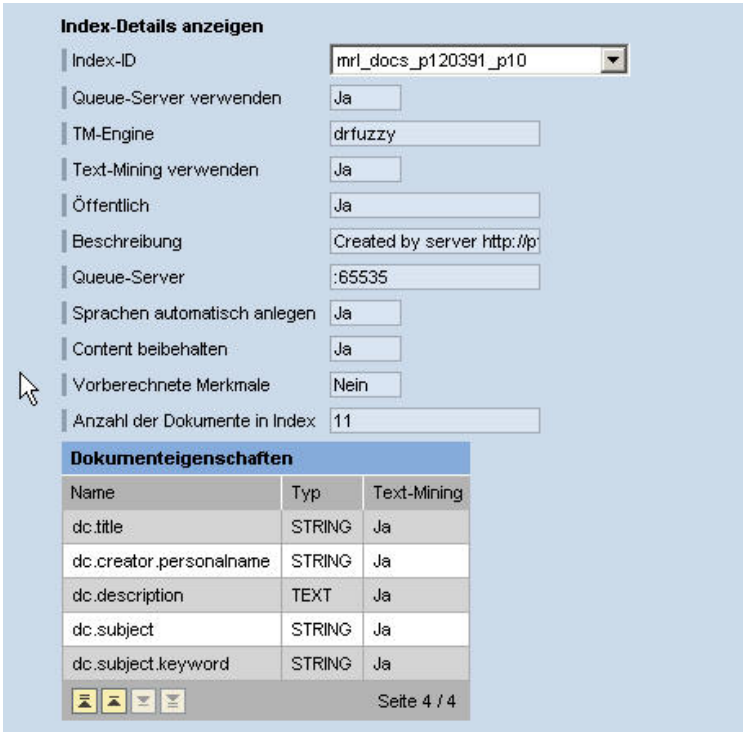


activated, see *readme.txt* in *TREX directory/extensions/dcattributes.* Here you can also look up the attribute names that TREX assigns to the DC metadata.



5. Index a document or a set of documents containing DC metadata by adding it as a data source to the index you created earlier.

   As a result, you can see in the index details view in KMC the attributes that TREX created from the DC metadata. Here the system displays the attributes that TREX extracts and stores with the index.



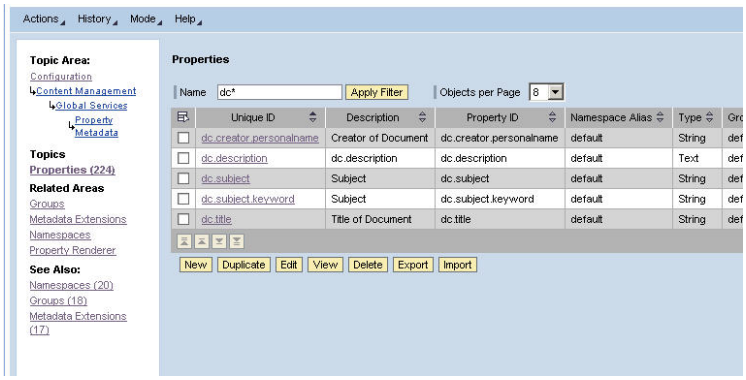
3.2 Configuration in KMC

So far you have enabled the extraction of DC metadata in TREX; this metadata is displayed as index attributes in *Index Properties*.
Now you have to make the attributes available in KMC to allow users to search for documents containing them.

6. Create a KM property to display DC metadata from TREX:

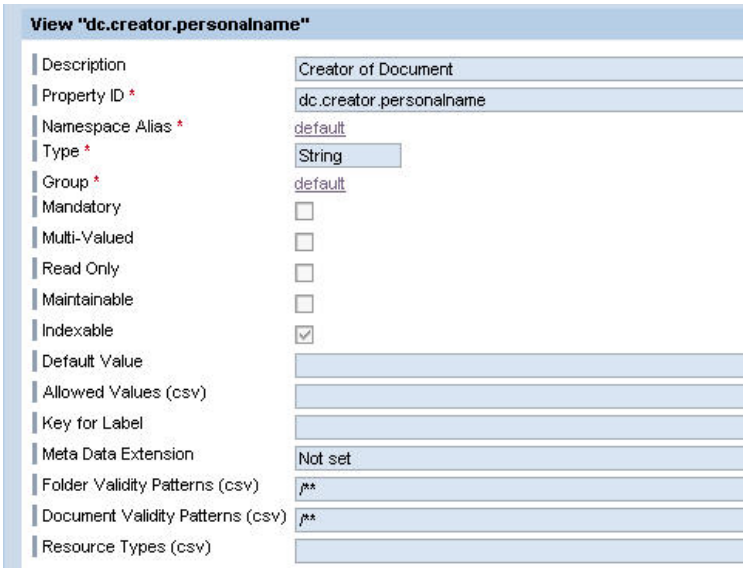
   In the KM Configuration Framework, go to *Content Management* → *Global Services* → *Property Metadata* and create a new property.

   Our example lists five KM properties that were created to reflect five elements of the DC metadata. The *DC.description* element is of type *Text.*



   The data shown in the screenshot includes:

   - Description: Creator of the document
   - Property ID: dc.creator.personalname
   - Namespace Alias: default (mandatory)
   - Enable the *Indexable* flag. (Setting the *Read Only* flag makes the property visible in the *Details* dialog. This might cause confusion because the property cannot be maintained on the KM UI)



   Optional:

   - Key for Label: *Author* (see section 3.3)
   - Metadata Extension: *MetadataExtensions* (see section 3.3)

7. Create the search option set containing DC metadata:

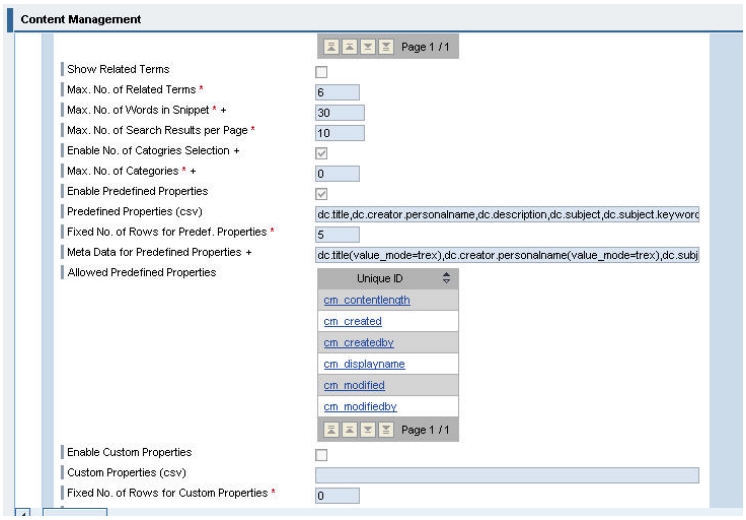
   Go to *Content Management* → *User Interface* → *Search* → *Search Option Set*. Create a *DublinCore* search option set by duplicating the *UIsearch* set using the entries indicated in the screenshot (depending on how many properties you defined in step six):

   Predefined Properties: *dc.title,dc.creator.personalname*, *dc.description*, *dc.subject*, *dc.subject.keyword*

   Fixed Number of Rows for Predef.Properties: 5

   Meta Data for Predefined Properties(advanced options): *dc.title(value_mode=trex)*,*dc.creator.personalname(value_mode=trex)*,*dc.subject(value_mode=trex)*,*dc.subject.keyword(value_mode=trex)*

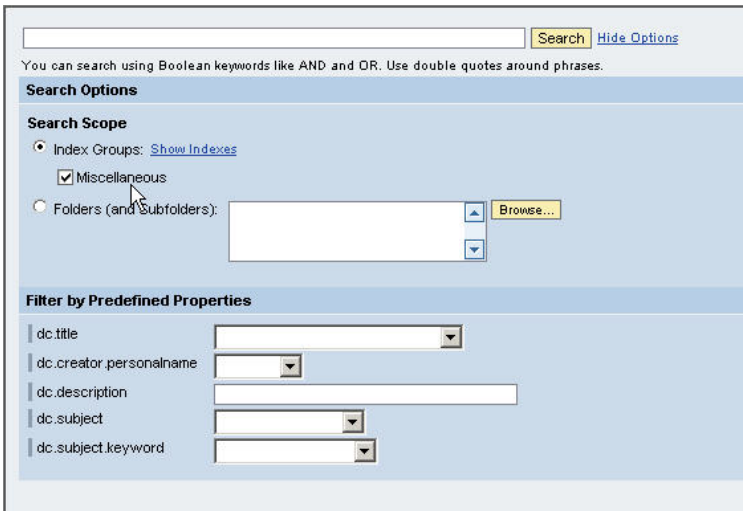
   This makes available default values for the properties from TREX. This feature is available with TREX 7.0 and higher. *Description* is a text property and hence not taken into account here.



8. Create the search UI with the search option set containing search fields for DC metadata:

   There are various ways to use the search option set in the search UI, for more information, see the standard documentation on the *Standard Search Dialog Box*.

   For example, if you change the default search dialog box by replacing the *UISearch* search option set with the one you have created, you get the search UI shown here.



9. You can now search by selecting a search term from the default values for the DC metadata (TREX 7.0) or by entering a search term (TREX 6.1). You do not need to enter anything in the general search field.

3.3 Label for Properties

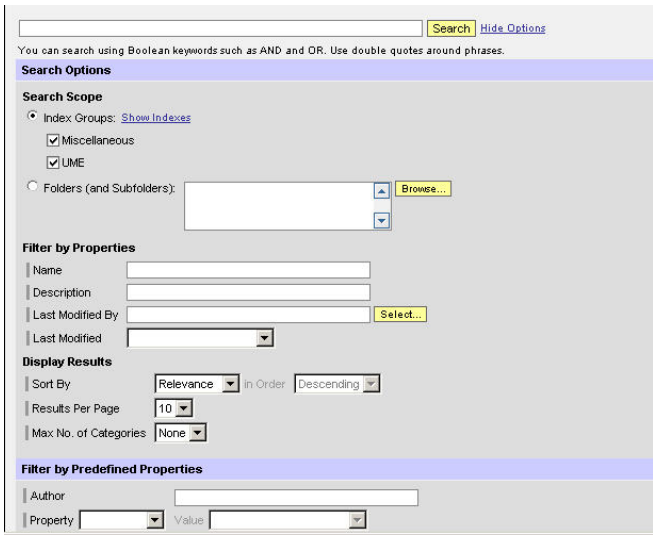
10. This section concerns the descriptive name of the DC metadata for the UI.

    You cannot choose names for the DC metadata properties in KMC, instead you have to use the attribute names that TREX assigns. You can see these names in *Index Details* in the TREX monitor.

    In this example, the DC attribute *Creator* is used. TREX generates the *dc.creator.personalname* attribute. If you now create the property in KMC, you have to use this name as the property ID to refer to this DC metadata element.

    To avoid displaying this name on the UI, you have to create a bundle file. A preconfigured example and a description of how to proceed here can be found in the documentation changing labels for properties. For more information, we recommend reading the *Translating CM Properties* how-to guide, You can download this guide from SAP Service Marketplace at https://websmp204.sap-ag.de/nw04 (choose How-To Guides → Portal, KM, and Collaboration in the navigation).

    - The metadata extension that is created here to map the information from the bundle file to the property is called *MetaExtensions*.
    - The key used in the properties files is *Author*.



http://service.sap.com/nw2004s-howtoguides

THE BEST-RUN BUSINESSES RUN SAP