PNNL-15384-1

Pacific Northwest National Laboratory
Operated by Battelle for the
U.S. Department of Energy

Universal Parsing Agent (UPA) User Guide

W. E. Cowley
N. O. Cramer
A. G. Gibson
R. T. Scott
M. A. Whiting
C. Winters

September 2005

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor Battelle Memorial Institute, nor any of their employees, makes **any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights**. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or Battelle Memorial Institute. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062;
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@adonis.osti.gov

Available to the public from the National Technical Information Service,
U.S. Department of Commerce, 5285 Port Royal Rd., Springfield, VA 22161
ph: (800) 553-6847
fax: (703) 605-6900
email: orders@ntis.fedworld.gov
online ordering: http://www.ntis.gov/ordering.htm

This document was printed on recycled paper.

PNNL- 15384-1

Universal Parsing Agent (UPA) User Guide

W. E. Cowley	R. T. Scott
N. O. Cramer	M. A. Whiting
A. G. Gibson	C. Winters

September 2005

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Table of Contents

Universal Parsing Agent (UPA)

About the UPA User Guide

The Universal Parsing Agent (UPA) processes text documents, extracts information, and stores that information in XML markup files for further use by other software products. This provides users with more time for analysis by automating document processing. UPA can process formatted and semi-formatted documents once the processing templates have been defined by the user.

Purpose of the User Guide

The Universal Parsing Agent (UPA) User Guide provides background for UPA operating procedures. This guide explains the basic procedures for creating templates and using the system.

Scope

Use this guide to learn the basic techniques for working with UPA.

Audience

The primary audience for this guide is the general user of UPA. Database and system administrators and managers are the secondary audience.

Prerequisites

Users should have a working knowledge of Windows-based personal computers and a current version of an internet browser: Internet Explorer, Mozilla, Firefox, Opera, or Netscape. Familiarity with regular PERL expressions is necessary if you want to create custom parsing logic. Those performing software maintenance and troubleshooting should also be comfortable working with computer system software and related hardware.

How to Use This Guide

Use this guide to learn how to work with UPA.

Read the Overview for an introduction to UPA and its functions.

Read Starting UPA to learn how to start the application, log in, and select a project database, and Exiting UPA to learn how to save your work and exit the application.

Read Creating Input Templates to learn how to build templates used for later document processing and transfer.

Read Assigning or Removing Templates to learn how to select and assign a template to a document database. The template will then be used to process all documents in the database.

Read Selecting Input Documents to learn how to select documents for viewing, correction, or for use in building templates for processing.

Read Working with the Wait Bin to learn how to use UPA to create XML marked-up documents.

Read the Pattern Definition Reference for a quick reference guide to PERL regular expressions.

Notation Conventions

Throughout this guide, capitalized text is used to emphasize names of software and menu items in the software. The following notation conventions mark certain important points or procedures in this Guide.

Menu Item	indicates the names of menu items, buttons, and fields.
Tips:	application usage tips.
Note:	additional helpful details about a procedure.
Caution:	indicates that data might be lost.

Results

After reading this document, you will be able to:

- Create and modify templates
- Construct document parsing logic
- Assign templates to directories
- Process documents
- Review, correct, and complete processed documents

Related Information

The following book is recommended for users needing complete information on constructing PERL regular expressions.

Programming Perl Third Edition, Larry Wall, Tom Christiansen, and Jon Orwant, 2000, O'Reilly & Associates, Inc. This book provides complete instructions on programming in Perl, including in-depth information on the use of regular expressions for pattern matching.

Acknowledgements

The following software products were used in the development of UPA.

JDOM

Copyright © 2000-2004 Jason Hunter & Brett McLaughlin. All rights reserved.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

1. Redistributions of source code must retain the above copyright notice, this list of conditions, and the following disclaimer.
2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions, and the disclaimer that follows these conditions in the documentation and/or other materials provided with the distribution.
3. The name "JDOM" must not be used to endorse or promote products derived from this software without prior written permission. For written permission, please contact <request_AT_jdom_DOT_org>.
4. Products derived from this software may not be called "JDOM", nor may "JDOM" appear in their name, without prior written permission from the JDOM Project Management <request_AT_jdom_DOT_org>.

In addition, we request (but do not require) that you include in the end-user documentation provided with the redistribution and/or in the software itself an acknowledgement equivalent to the following:

"This product includes software developed by the JDOM Project (http://www.jdom.org/)."

Alternatively, the acknowledgment may be graphical using the logos available at http://www.jdom.org/images/logos.

THIS SOFTWARE IS PROVIDED ``AS IS'' AND ANY EXPRESSED OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE JDOM AUTHORS OR THE PROJECT CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF

THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

This software consists of voluntary contributions made by many individuals on behalf of the JDOM Project and was originally created by Jason Hunter <jhunter_AT_jdom_DOT_org> and Brett McLaughlin <brett_AT_jdom_DOT_org>. For more information on the JDOM Project, please see <http://www.jdom.org/>.

Apache Software

(xalan/xerces/saxpath/xml-apis/log4j)

The Apache Software License, Version 1.1

Copyright (c) 1999-2003 The Apache Software Foundation.All rights reserved.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

3. The end-user documentation included with the redistribution, if any, must include the following acknowledgment:
"This product includes software developed by the Apache Software Foundation (http://www.apache.org/)."
Alternately, this acknowledgment may appear in the software itself, if and wherever such third-party acknowledgments normally appear.

4. The names "Xerces" and "Apache Software Foundation" must not be used to endorse or promote products derived from this software without prior written permission. For written permission, please contact apache@apache.org.

5. Products derived from this software may not be called "Apache", nor may "Apache" appear in their name, without prior written permission of the Apache Software Foundation.

THIS SOFTWARE IS PROVIDED ``AS IS'' AND ANY EXPRESSED OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE APACHE SOFTWARE FOUNDATION OR ITS CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL

DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.
==

This software consists of voluntary contributions made by many individuals on behalf of the Apache Software Foundation and was originally based on software copyright (c) 1999, International Business Machines, Inc., http://www.ibm.com. For more information on the Apache Software Foundation, please see <http://www.apache.org/>.

Jakarta ORO

Apache License
Version 2.0, January 2004
http://www.apache.org/licenses/
TERMS AND CONDITIONS FOR USE, REPRODUCTION, AND DISTRIBUTION

1. Definitions.

"License" shall mean the terms and conditions for use, reproduction, and distribution as defined by Sections 1 through 9 of this document.

"Licensor" shall mean the copyright owner or entity authorized by the copyright owner that is granting the License.

"Legal Entity" shall mean the union of the acting entity and all other entities that control, are controlled by, or are under common control with that entity. For the purposes of this definition,

"control" means (i) the power, direct or indirect, to cause the direction or management of such entity, whether by contract or otherwise, or (ii) ownership of fifty percent (50%) or more of the outstanding shares, or (iii) beneficial ownership of such entity.

"You" (or "Your") shall mean an individual or Legal Entity exercising permissions granted by this License.

"Source" form shall mean the preferred form for making modifications, including but not limited to software source code,
documentation source, and configuration files.

"Object" form shall mean any form resulting from mechanical transformation or translation of a Source form, including but not limited to compiled object code, generated documentation, and conversions to other media types.

"Work" shall mean the work of authorship, whether in Source or Object form, made available under the License, as indicated by a copyright notice that is included in or attached to the work (an example is provided in the Appendix below).

"Derivative Works" shall mean any work, whether in Source or Object form, that is based on (or derived from) the Work and for which the editorial revisions, annotations, elaborations, or other modifications represent, as a whole, an original work of authorship.

For the purposes of this License, Derivative Works shall not include works that remain separable from, or merely link (or bind by name) to the interfaces of, the Work and Derivative Works thereof.

"Contribution" shall mean any work of authorship, including the original version of the Work and any modifications or additions to that Work or Derivative Works thereof, that is intentionally submitted to Licensor for inclusion in the Work by the copyright owner or by an individual or Legal Entity authorized to submit on behalf of the copyright owner. For the purposes of this definition, "submitted" means any form of electronic, verbal, or written communication sent to the Licensor or its representatives, including but not limited to communication on electronic mailing lists, source code control systems, and issue tracking systems that are managed by, or on behalf of, the Licensor for the purpose of discussing and improving the Work, but excluding communication that is conspicuously marked or otherwise designated in writing by the copyright owner as "Not a Contribution."

"Contributor" shall mean Licensor and any individual or Legal Entity on behalf of whom a Contribution has been received by Licensor and subsequently incorporated within the Work.

2. Grant of Copyright License. Subject to the terms and conditions of this License, each Contributor hereby grants to You a perpetual, worldwide, non-exclusive, no-charge, royalty-free, irrevocable copyright license to reproduce, prepare Derivative Works of, publicly display, publicly perform, sublicense, and distribute the Work and such Derivative Works in Source or Object form.

3. Grant of Patent License. Subject to the terms and conditions of this License, each Contributor hereby grants to You a perpetual, worldwide, non-exclusive, no-charge, royalty-free, irrevocable (except as stated in this section) patent license to make, have made, use, offer to sell, sell, import, and otherwise transfer the Work, where such license applies only to those patent claims licensable by such Contributor that are necessarily infringed by their Contribution(s) alone or by combination of their Contribution(s) with the Work to which such Contribution(s) was submitted. If You institute patent litigation against any entity (including a cross-claim or counterclaim in a lawsuit) alleging that the Work or a Contribution incorporated within the Work constitutes direct

or contributory patent infringement, then any patent licenses granted to You under this License for that Work shall terminate as of the date such litigation is filed.

4. Redistribution. You may reproduce and distribute copies of the Work or Derivative Works thereof in any medium, with or without modifications, and in Source or Object form, provided that You meet the following conditions:

a. You must give any other recipients of the Work or Derivative Works a copy of this License; and
b. You must cause any modified files to carry prominent notices stating that You changed the files; and
c. You must retain, in the Source form of any Derivative Works that You distribute, all copyright, patent, trademark, and attribution notices from the Source form of the Work, excluding those notices that do not pertain to any part of the Derivative Works; and
d. If the Work includes a "NOTICE" text file as part of its distribution, then any Derivative Works that You distribute must include a readable copy of the attribution notices contained within such NOTICE file, excluding those notices that do not pertain to any part of the Derivative Works, in at least one of the following places: within a NOTICE text file distributed as part of the Derivative Works; within the Source form or documentation, if provided along with the Derivative Works; or, within a display generated by the Derivative Works, if and wherever such third-party notices normally appear. The contents of the NOTICE file are for informational purposes only and do not modify the License. You may add Your own attribution notices within Derivative Works that You distribute, alongside or as an addendum to the NOTICE text from the Work, provided that such additional attribution notices cannot be construed as modifying the License. You may add Your own copyright statement to Your modifications and may provide additional or different license terms and conditions for use, reproduction, or distribution of Your modifications, or for any such Derivative Works as a whole, provided Your use, reproduction, and distribution of the Work otherwise complies with the conditions stated in this License.

5. Submission of Contributions. Unless You explicitly state otherwise, any Contribution intentionally submitted for inclusion in the Work by You to the Licensor shall be under the terms and conditions of this License, without any additional terms or conditions.

Notwithstanding the above, nothing herein shall supersede or modify the terms of any separate license agreement you may have executed with Licensor regarding such Contributions.

6. Trademarks. This License does not grant permission to use the trade names, trademarks, service marks, or product names of the Licensor, except as required for reasonable and

customary use in describing the origin of the Work and reproducing the content of the NOTICE file.

7. Disclaimer of Warranty. Unless required by applicable law or agreed to in writing, Licensor provides the Work (and each Contributor provides its Contributions) on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied, including, without limitation, any warranties or conditions of TITLE, NON-INFRINGEMENT, MERCHANTABILITY, or FITNESS FOR A PARTICULAR PURPOSE. You are solely responsible for determining the appropriateness of using or redistributing the Work and assume any risks associated with Your exercise of permissions under this License.

8. Limitation of Liability. In no event and under no legal theory, whether in tort (including negligence), contract, or otherwise, unless required by applicable law (such as deliberate and grossly negligent acts) or agreed to in writing, shall any Contributor be liable to You for damages, including any direct, indirect, special, incidental, or consequential damages of any character arising as a result of this License or out of the use or inability to use the Work (including but not limited to damages for loss of goodwill, work stoppage, computer failure or malfunction, or any and all other commercial damages or losses), even if such Contributor has been advised of the possibility of such damages.

9. Accepting Warranty or Additional Liability. While redistributing the Work or Derivative Works thereof, You may choose to offer, and charge a fee for, acceptance of support, warranty, indemnity, or other liability obligations and/or rights consistent with this License. However, in accepting such obligations, You may act only on Your own behalf and on Your sole responsibility, not on behalf of any other Contributor, and only if You agree to indemnify, defend, and hold each Contributor harmless for any liability incurred by, or claims asserted against, such Contributor by reason of your accepting any such warranty or additional liability.

JGraph -- jgraph-4.0-java1.3

Copyright (c) 2001-2004, Gaudenz Alder

All rights reserved.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

- Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
- Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the

following disclaimer in the documentation and/or other materials provided with the distribution.

- Neither the name of JGraph nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS"AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE)ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

Jaxen

(Used from JDOM)

Redistribution and use of this software and associated documentation ("Software"), with or without modification, are permitted provided that the following conditions are met:

1. Redistributions of source code must retain copyright statements and notices. Redistributions must also contain a copy of this document.

2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

3. The name "Jaxen" must not be used to endorse or promote products derived from this Software without prior written permission of werken digital. For written permission, please contact bob@werken.com.

4. Products derived from this Software may not be called "Jaxen" nor may "Jaxen" appear in their names without prior written permission of werken digital. Jaxen is a registered trademark of werken digital

5. Due credit should be given to the Jaxen Project (http://jaxen.org/).

THIS SOFTWARE IS PROVIDED BY METASTUFF, LTD. AND ONTRIBUTORS ``AS IS'' AND ANY EXPRESSED OR IMPLIED WARRANTIES, NCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL METASTUFF, LTD. OR ITS

CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES(INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

Copyright 2001 (C) werken digital. All Rights Reserved.

JSSE

http://java.sun.com/products/jsse/LICENSE.html
Sun Microsystems, Inc.

Binary Code License Agreement

READ THE TERMS OF THIS AGREEMENT AND ANY PROVIDED SUPPLEMENTAL LICENSE TERMS (COLLECTIVELY "AGREEMENT") CAREFULLY BEFORE OPENING THE SOFTWARE MEDIA PACKAGE. BY OPENING THE SOFTWARE MEDIA PACKAGE, YOU AGREE TO THE TERMS OF THIS AGREEMENT. IF YOU ARE ACCESSING THE SOFTWARE ELECTRONICALLY, INDICATE YOUR ACCEPTANCE OF THESE TERMS BY SELECTING THE "ACCEPT" BUTTON AT THE END OF THIS AGREEMENT. IF YOU DO NOT AGREE TO ALL THESE TERMS, PROMPTLY RETURN THE UNUSED SOFTWARE TO YOUR PLACE OF PURCHASE FOR A REFUND OR, IF THE SOFTWARE IS ACCESSED ELECTRONICALLY, SELECT THE "DECLINE" BUTTON AT THE END OF THIS AGREEMENT.

1. LICENSE TO USE. Sun grants you a non-exclusive and non-transferable license for the internal use only of the accompanying software and documentation and any error corrections provided by Sun (collectively "Software"), by the number of users and the class of computer hardware for which the corresponding fee has been paid.

2. RESTRICTIONS. Software is confidential and copyrighted. Title to Software and all associated intellectual property rights is retained by Sun and/or its licensors. Except as specifically authorized in any Supplemental License Terms, you may not make copies of Software, other than a single copy of Software for archival purposes. Unless enforcement is prohibited by applicable law, you may not modify, decompile, or reverse engineer Software. Licensee acknowledges that Licensed Software is not designed or intended for use in the design, construction, operation or maintenance of any nuclear facility. Sun Microsystems, Inc. disclaims any express or implied warranty of fitness for such uses. No right, title or interest in or to any trademark, service mark, logo or trade name of Sun or its licensors is granted under this Agreement.

3. LIMITED WARRANTY. Sun warrants to you that for a period of ninety (90) days from the date of purchase, as evidenced by a copy of the receipt, the media on which Software is furnished (if any) will be free of defects in materials and workmanship

under normal use. Except for the foregoing, Software is provided "AS IS". Your exclusive remedy and Sun's entire liability under this limited warranty will be at Sun's option to replace Software media or refund the fee paid for Software.

4. DISCLAIMER OF WARRANTY. UNLESS SPECIFIED IN THIS AGREEMENT, ALL EXPRESS OR IMPLIED CONDITIONS, REPRESENTATIONS AND WARRANTIES, INCLUDING ANY IMPLIED WARRANTY OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NON-INFRINGEMENT ARE DISCLAIMED, EXCEPT TO THE EXTENT THAT THESE DISCLAIMERS ARE HELD TO BE LEGALLY INVALID.

5. LIMITATION OF LIABILITY. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL SUN OR ITS LICENSORS BE LIABLE FOR ANY LOST REVENUE, PROFIT OR DATA, OR FOR SPECIAL, INDIRECT, CONSEQUENTIAL, INCIDENTAL OR PUNITIVE DAMAGES, HOWEVER CAUSED REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF OR RELATED TO THE USE OF OR INABILITY TO USE SOFTWARE, EVEN IF SUN HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. In no event will Sun's liability to you, whether in contract, tort (including negligence), or otherwise, exceed the amount paid by you for Software under this Agreement. The foregoing limitations will apply even if the above stated warranty fails of its essential purpose.

6. Termination. This Agreement is effective until terminated. You may terminate this Agreement at any time by destroying all copies of Software. This Agreement will terminate immediately without notice from Sun if you fail to comply with any provision of this Agreement. Upon Termination, you must destroy all copies of Software.

7. Export Regulations. All Software and technical data delivered under this Agreement are subject to US export control laws and may be subject to export or import regulations in other countries. You agree to comply strictly with all such laws and regulations and acknowledge that you have the responsibility to obtain such licenses to export, re-export, or import as may be required after delivery to you.

8. U.S. Government Restricted Rights. If Software is being acquired by or on behalf of the U.S. Government or by a U.S. Government prime contractor or subcontractor (at any tier), then the Government's rights in Software and accompanying documentation will be only as set forth in this Agreement; this is in accordance with 48 CFR 227.7201 through 227.7202-4 (for Department of Defense (DOD) acquisitions) and with 48 CFR 2.101 and 12.212 (for non-DOD acquisitions).

9. Governing Law. Any action related to this Agreement will be governed by California law and controlling U.S. federal law. No choice of law rules of any jurisdiction will apply.

10. Severability. If any provision of this Agreement is held to be unenforceable, this Agreement will remain in effect with the provision omitted, unless omission would frustrate the

intent of the parties, in which case this Agreement will immediately terminate.

11. Integration. This Agreement is the entire agreement between you and Sun relating to its subject matter. It supersedes all prior or contemporaneous oral or written communications, proposals, representations and warranties and prevails over any conflicting or additional terms of any quote, order, acknowledgment, or other communication between the parties relating to its subject matter during the term of this Agreement. No modification of this Agreement will be binding, unless in writing and signed by an authorized representative of each party.

JAVA™ OPTIONAL PACKAGE

JAVA SECURE SOCKET EXTENSION, VERSION 1.0.3_XX
SUPPLEMENTAL LICENSE TERMS
These supplemental license terms ("Supplemental Terms") add to or modify the terms of the Binary Code License Agreement (collectively, the "Agreement"). Capitalized terms not defined in these Supplemental Terms shall have the same meanings ascribed to them in the Agreement. These Supplemental Terms shall supersede any inconsistent or conflicting terms in the Agreement, or in any license contained within the Software.

1. Software Internal Use and Development License Grant. Subject to the terms and conditions of this Agreement, including, but not limited to Section 3 (Java Technology Restrictions) of these Supplemental Terms, Sun grants you a non-exclusive, non-transferable, limited license to reproduce internally and use internally the binary form of the Software, complete and unmodified, for the sole purpose of designing, developing and testing your Java applets and applications ("Programs").

2. License to Distribute Software. In addition to the license granted in Section 1 (Software Internal Use and Development License Grant) of these Supplemental Terms, subject to the terms and conditions of this Agreement, including but not limited to, Section 3 (Java Technology Restrictions) of these Supplemental Terms, Sun grants you a non-exclusive, non-transferable, limited license to reproduce and distribute the Software in binary code form only, provided that you (i) distribute the Software complete and unmodified and only bundled as part of your Programs, (ii) do not distribute additional software intended to replace any component(s) of the Software, (iii) do not remove or alter any proprietary legends or notices contained in the Software, (iv) only distribute the Software subject to a license agreement that protects Sun's interests consistent with the terms contained in this Agreement, (v) agree to defend and indemnify Sun and its licensors from and against any damages, costs, liabilities, settlement

3. Java Technology Restrictions. You may not modify the Java Platform Interface ("JPI", identified as classes contained within the "java" package or any subpackages of the "java" package), by creating additional classes within the JPI or otherwise causing the addition to or modification of the classes in the JPI. In the event that you create an additional class and associated API(s) which (i) extends the functionality of the Java platform, and (ii) is exposed to third party software developers for the purpose of developing additional software which invokes such additional API, you must promptly publish broadly an accurate specification for such API for free use by all developers. You may not create, or authorize your licensees to create additional classes, interfaces, or subpackages that are in any way identified as "java", "javax", "sun" or similar convention as specified by Sun in any naming convention designation.

4. Trademarks and Logos. You acknowledge and agree as between you and Sun that Sun owns the SUN, SOLARIS, JAVA, JINI, FORTE, and iPLANET trademarks and all SUN, SOLARIS, JAVA, JINI, FORTE, and iPLANET-related trademarks, service marks, logos and other brand designations ("Sun Marks"), and you agree to comply with the Sun Trademark and Logo Usage Requirements currently located at http://www.sun.com/policies/trademarks. Any use you make of the Sun Marks inures to Sun's benefit.

5. Source Code. Software may contain source code that is provided solely for reference purposes pursuant to the terms of this Agreement. Source code may not be redistributed unless expressly provided for in this Agreement.

6. Termination for Infringement. Either party may terminate this Agreement immediately should any Software become, or in either party's opinion be likely to become, the subject of a claim of infringement of any intellectual property right.
For inquiries please contact: Sun Microsystems, Inc. 901 San Antonio Road, Palo Alto, California 94303

(LFI#138694/Form ID#011801)

Overview

UPA parses information contained within chosen text files using templates developed by analysts based on their current needs. Many files can be automatically processed at one time and the results presented to the analyst. Alternatively, single files can be processed as well on a case-by-case basis as designated by the analyst. These templates can be reused and new ones created for later or immediate use.

- UPA provides systematic document preparation for the efficient ingest of data into data analysis tools. By systematizing the process a common ground is established for the analysis of future documents.
- Other tools treat each document as a whole, UPA improves the quality of data going into an analytical system by breaking documents into meaningful components that can be dynamically defined on an as-needed basis.
- UPA can be applied to both structured and unstructured data.

A Typical Scenario

Analysts have access to a file repository and can select directories or individual files that they would like processed to see if they contain particular information. From their file resource they select the directories or individual files. These are then placed in the Input Bin for later batch mode processing. Once processed the results are placed in one of three repositories within UPA so analysts can browse the results.

Wait Bin: View files that matched required items in a template. The analyst can

- Manually revise these patterns to become part of a new template or
- Designate the file as complete.

Incomplete Bin: Lists all files where no direct matches were found in any available templates. You can view these files and modify or create new templates to reprocess any failed files. You can test the modified template by running the parser on the single document.

Complete Bin: Lists files that have been successfully parsed with the template that was used to parse them. You can view both the processed files and the related template.

Analysts can also create templates using existing templates or by starting with a blank template. A set of parser tools are provided to assist the analyst in defining patterns or expressions that will look for specific information contained in a file. The parser patterns that you create are saved as a template. It is this template that is used to parse files.

While defining patterns, analysts do so in the context of how they traditionally structure a document. Analysts typically define a header area, the main message area, and the footer area of a document. Within each of these main areas they can define searches for specific information items or content.

Analysts have the ability to create and save draft templates to work on and finalize later. They can also create statistical data about the files being processed. This information tells the analyst how many files have been parsed in the various categories. In the future, this feature could begin to aggregate information to enable analysts to view trends on user-specified data.

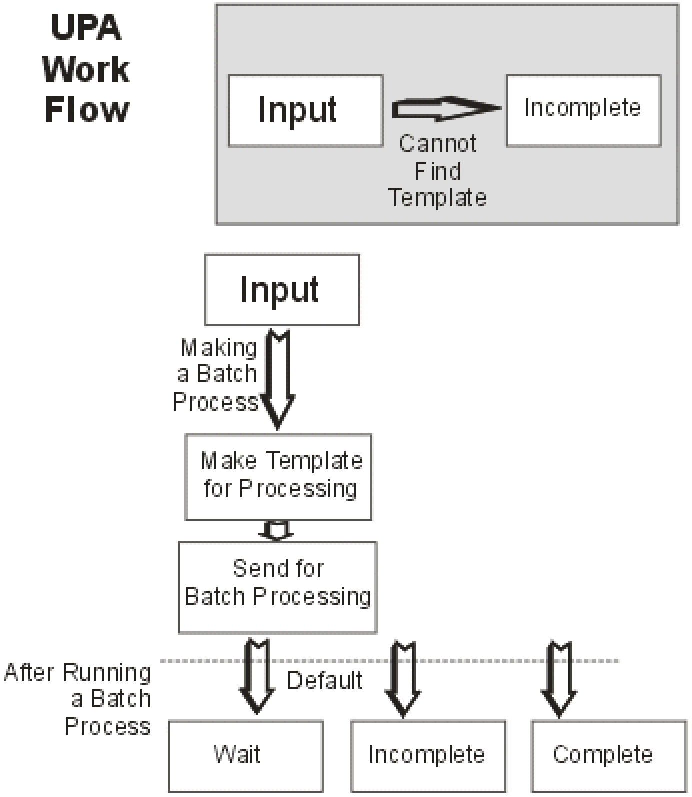


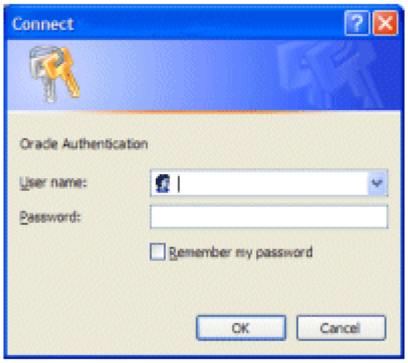
Figure 1. UPA Work Flow Diagram

Starting UPA

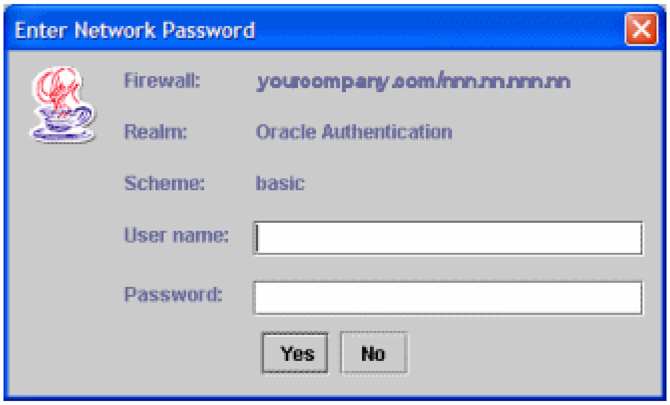
All activities are started from the UPA main window.

To start UPA, perform the following steps.

1. Start a browser such as Microsoft Internet Explorer or Mozilla.
2. If this is the first time you have used UPA, from the browser File menu select Open. Browse to the file, UPAClient.html, select the file, and click Open .
   Or if UPA has been set up for you, from your browser Favorites list, select UPA.
   The Oracle authentication window will display.



3. Enter your user name and password and click Ok. The Enter Network Password window will display.



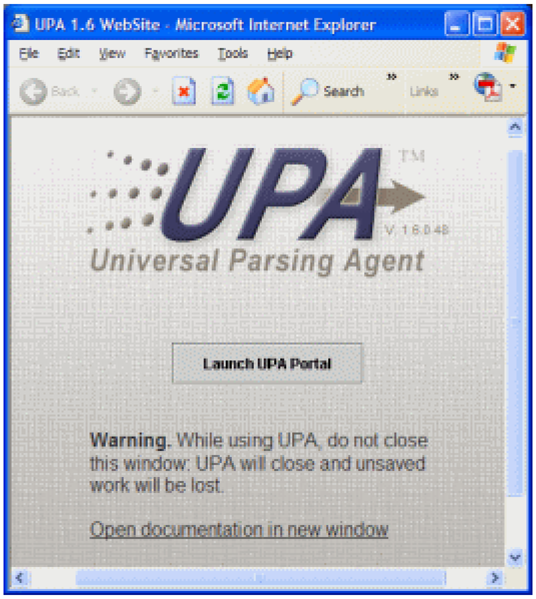
Enter Network Password

Firewall: yourcompany.com/nnn.nn.nnn.nn

Realm: Oracle Authentication

Scheme: basic

4. Enter your user name and password and click Yes.
   The UPA Website window will display in your browser.

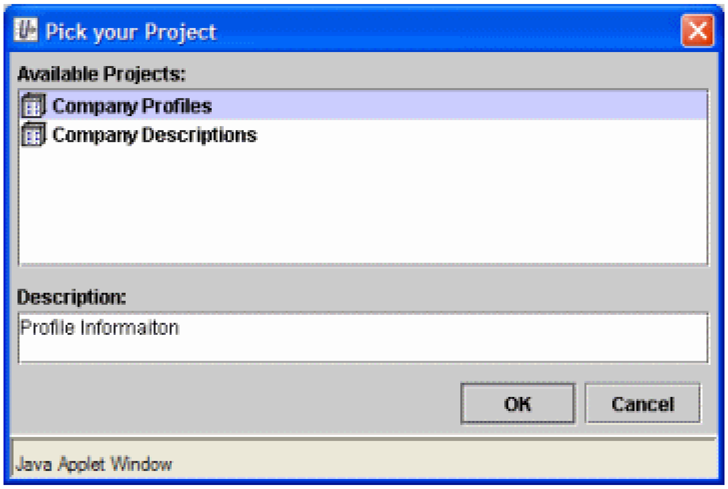

UPA 1.6 WebSite - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Links
UPA
TM
V 1.6.0.48
Universal Parsing Agent
Launch UPA Portal

Warning. While using UPA, do not close this window: UPA will close and unsaved work will be lost.

Open documentation in new window

Caution: Do not close your browser window while working with UPA. Closing your browser will close UPA as well and all your unsaved work will be lost. UPA is a web-based application and needs your browser to run.

5. From the UPA Website window, click Launch UPA Portal. The Pick Your Project window will display.



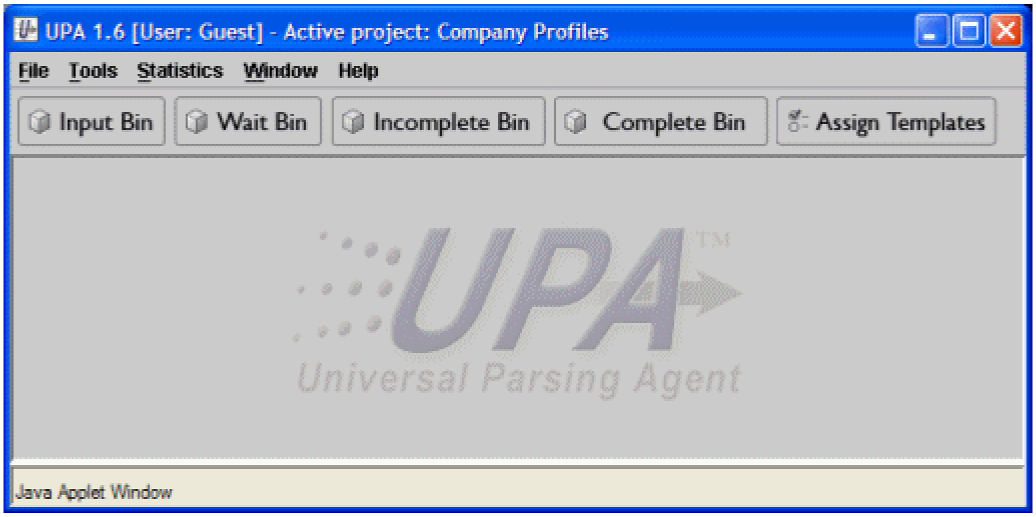
Available Projects:

- Company Profiles
- Company Descriptions

Description:

Profile Informaiton

6. Select a project from the Available Projects list and click Ok. The UPA Active Database window will display.



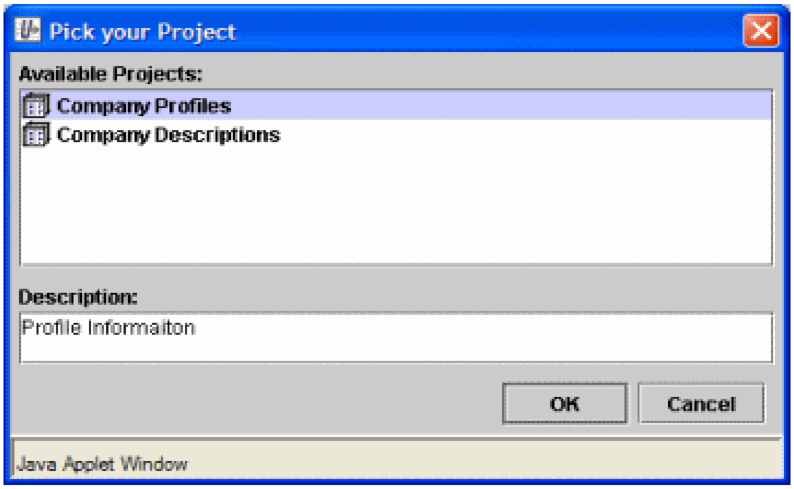
This is the UPA main window and toolbar. From the main window you can perform the following activities.

- Create Input templates
- Modify Input templates
- Check and edit documents in the Wait Bin
- View results in the Complete Bin
- Assign templates to be used in document processing

Selecting a Project

Once you have started UPA, you can select another project at any time. Projects consist of an Oracle database containing documents and the templates used to process them. To select a new project

1. From the UPA main window File menu, pick Select Project. The Pick Your Project window will display.



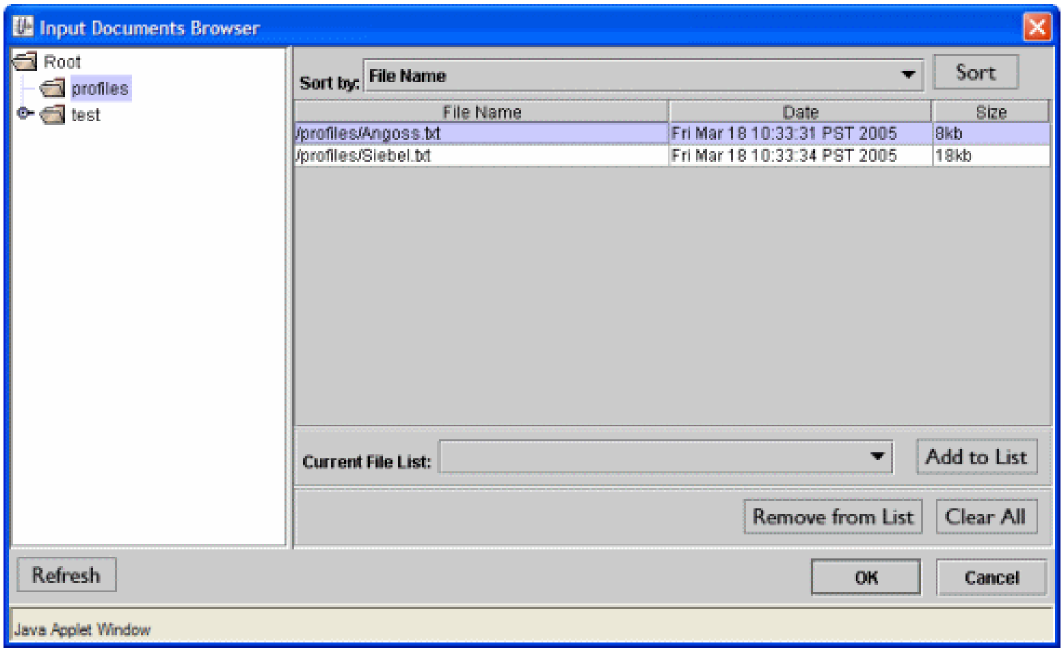
Profile Informaiton

2. From the Available Projects list, select a project. Project descriptions display in the Description area. Click Ok. The UPA main window title bar will display the name of your selected project database. You can now work with templates or documents in the various bins.

Selecting Input Documents

To select one or more documents to work with, perform the following procedures.

1. From the UPA main window, click on Input Bin Input Bin. The Input Documents Browser will display.



Sort by: File Name

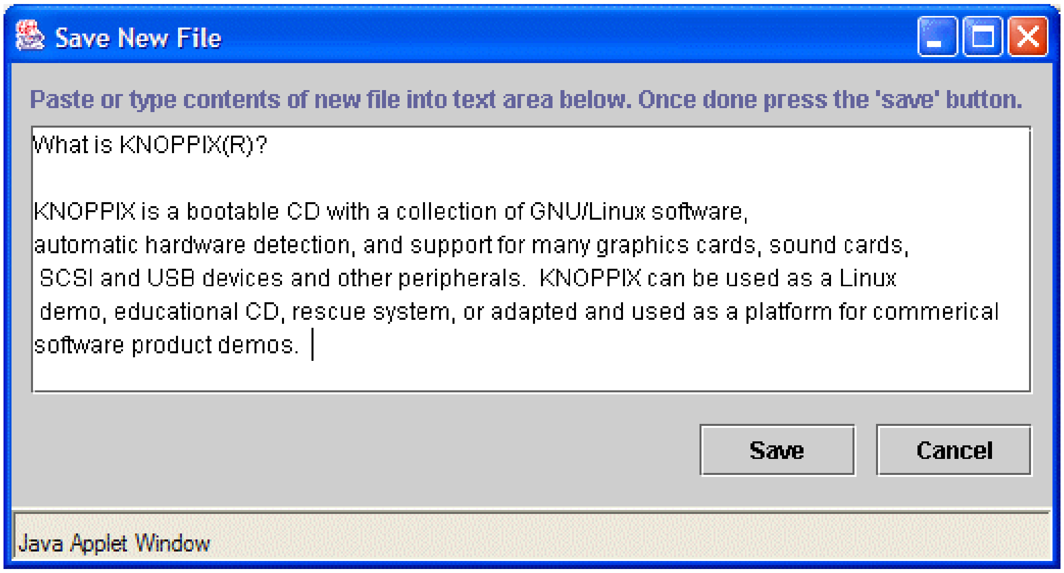
File Name	Date	Size
/profiles/Angoss.txt	Fri Mar 18 10:33:31 PST 2005	8kb
/profiles/Siebel.txt	Fri Mar 18 10:33:34 PST 2005	18kb

2. From the left hand side of the Documents Browser, open a folder and select one or more documents from the bin.
   - Tips:
   - You can select multiple documents by holding down the Ctrl key and clicking on all the documents you want in the bin.
   - You can add documents to the list by double-clicking on a file.
   - You can open documents in other directories by using the directory folders on the left side of the Document Browser window.
3. Click Add to List. The selected documents will be placed in the Current File List.
4. To see the documents on the list, click the drop-down button on the Current File List.
5. To remove a document from the list, select the file from the drop-down menu and click Remove from List.
6. To remove all the selected documents from the list, click Clear All.
7. To begin working with your selected documents, click OK.

Pasting a New Document

Sometimes you may want to add a new document for processing to those already in your database. To add a document

1. From the UPA main window File menu, select Paste New Document. The Save New File window will display.

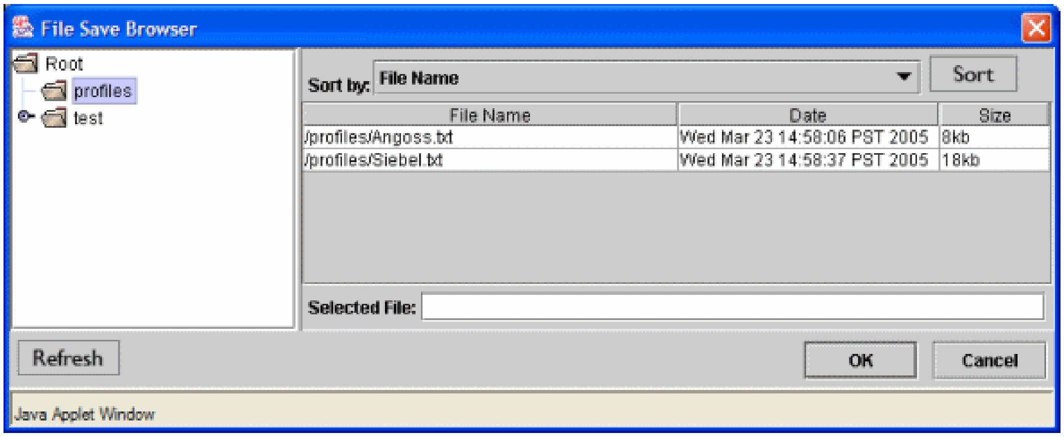


Paste or type contents of new file into text area below. Once done press the 'save' button.

What is KNOPPIX(R)?

KNOPPIX is a bootable CD with a collection of GNU/Linux software,
automatic hardware detection, and support for many graphics cards, sound cards,
SCSI and USB devices and other peripherals. KNOPPIX can be used as a Linux
demo, educational CD, rescue system, or adapted and used as a platform for commerical
software product demos.

2. Enter the contents of the new document in the window by cutting and pasting from the original document. Click Save. The File Save Browser window will display.



Sort by: File Name

File Name	Date	Size
/profiles/Angoss.txt	Wed Mar 23 14:58:06 PST 2005	8kb
/profiles/Siebel.txt	Wed Mar 23 14:58:37 PST 2005	18kb

Selected File:

3. Navigate to the directory where you want to save the document.

4. Enter the file name in the Selected File field and click Ok to save the document.

New documents will be processed with other documents in the directory, using the assigned template.

Pattern Definition Reference

Patterns are defined by "expressions;" sets of characters that enable the UPA software to seek and find similar patterns in text documents. UPA uses PERL regular expressions when you create the Pattern Definitions used by templates.

Expression Name	**Use this Code**	**ASCII Character Definition**
Whitespace	\s	[\t\n\r\f]
Non-whitespace	\S	[^\t\n\r\f]
Word character	\w	[a-zA-Z_0-9]
Non-word characters (to exclude words from your search but include such characters as # or @)	\W	[^a-zA-Z_0-9]
Digit	\d	[0-9]
Non-digits (to exclude digits from your search)	\D	[^0-9]
backspace within character sets	\b	

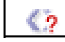
What it Does	**Use this Code**
Match at least n but not more than m times, for example: {2,5} would find matches at least 2 times but not more than 5.	{n,m}
Match at least n times, for example {3,} would find at least 3 matches and maybe more.	{n,}
Match at exactly n times, for example {2} would find exactly 2 matches.	{n}
Match 0 or more times.	*
Match 1 or more times.	+
Match 0 or 1 time.	?

What it Does	Use this Code
precise match - matches any single character from a set inside brackets	[...]
group items to find (object selection)- use the parentheses to group items in a search so they will be found and kept together	( )
wildcard - matches any single character including whitespace	.
logical or - this character OR that character	|
logical not - not this character - null token matching the beginning of a set of characters or line	^
null token matching the end of a set of characters or line	$
null token matching a word boundary (\w on one side and \W on the other)	\b

Author:\s(.*?)\n	This example pattern will search for the phrase "Author:" and will save into the database anything after the word to the end of the line. The UPA application will not save either the word 'Author' or the semicolon after the word 'Author' to the database. Only the pattern inside the parentheses will be saved.

Template and Document Icons

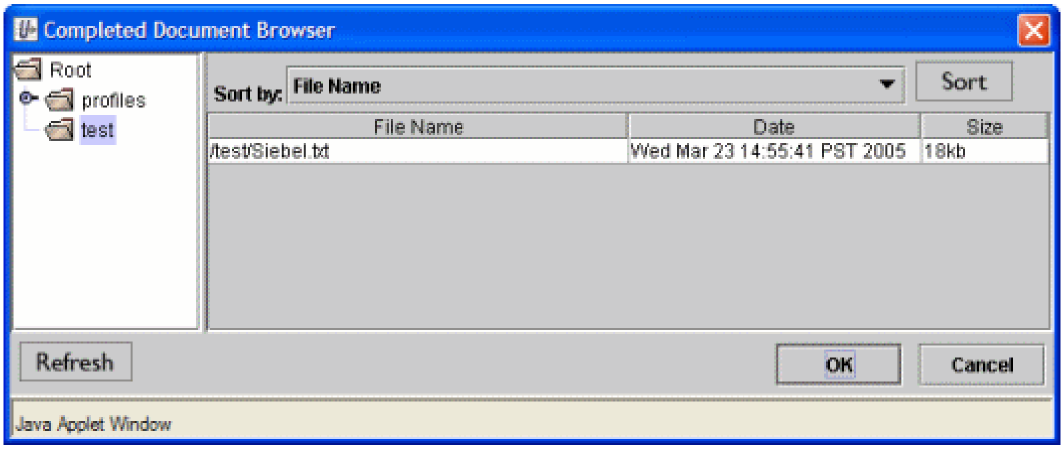
The right hand side of the UPA main window displays a listing of items contained in the template and related documents. These icons will change depending upon which Bin you are working in.

Icon	What It Means
	Parser steps have been assigned to this template item. The number of steps defined displays inside the icon.
	No parser steps have yet been assigned to this item.
	XML data has been entered by hand for this item.
	XML data was parsed and saved for this item.
	XML data was parsed and saved for this item, but it does not meet all criteria specified. Please verify and correct if necessary.
	XML data was not parsed or found for this item.

Viewing Completed Documents

Documents that have successfully completed processing will be moved to the Complete Bin. You can view completed documents, which consist of the source text, XML source (marked-up XML created by the parser), the extracted value, and the parsing diagram for each part of the template. To view a completed document

1. From the UPA main window, click on Complete Bin . The Complete Documents Browser will display.

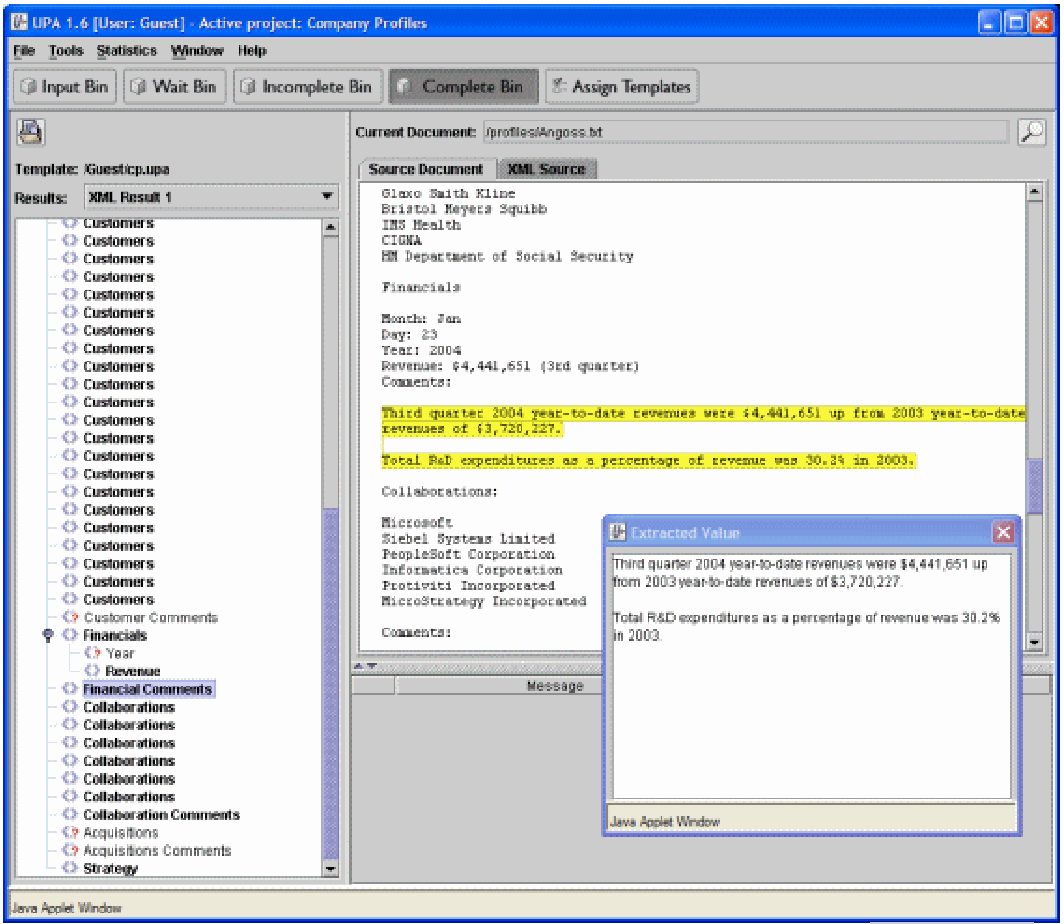


Sort by: File Name

File Name	Date	Size
/test/Siebel.txt	Wed Mar 23 14:55:41 PST 2005	18kb

2. From the left hand side of the Documents Browser, open a folder and select a document from the bin. You can open documents in other directories by using the directory folders on the left side of the Document Browser window.

3. To view your selected document, click OK. The document will display in the UPA main window. The initial view will be of the Source Document tab, which displays the original source text of the parsed document.



Template: /Guest/cp.upa

Results: XML Result 1

Current Document: /profiles/Angoss.txt

```
Glaxo Smith Kline
Bristol Meyers Squibb
IMS Health
CIGNA
HM Department of Social Security

Financials

Month: Jan
Day: 23
Year: 2004
Revenue: $4,441,651 (3rd quarter)
Comments:

Third quarter 2004 year-to-date revenues were $4,441,651 up from 2003 year-to-date
revenues of $3,720,227.

Total R&D expenditures as a percentage of revenue was 30.2% in 2003.

Collaborations:

Microsoft
Siebel Systems Limited
PeopleSoft Corporation
Informatica Corporation
Protiviti Incorporated
MicroStrategy Incorporated

Comments:
```

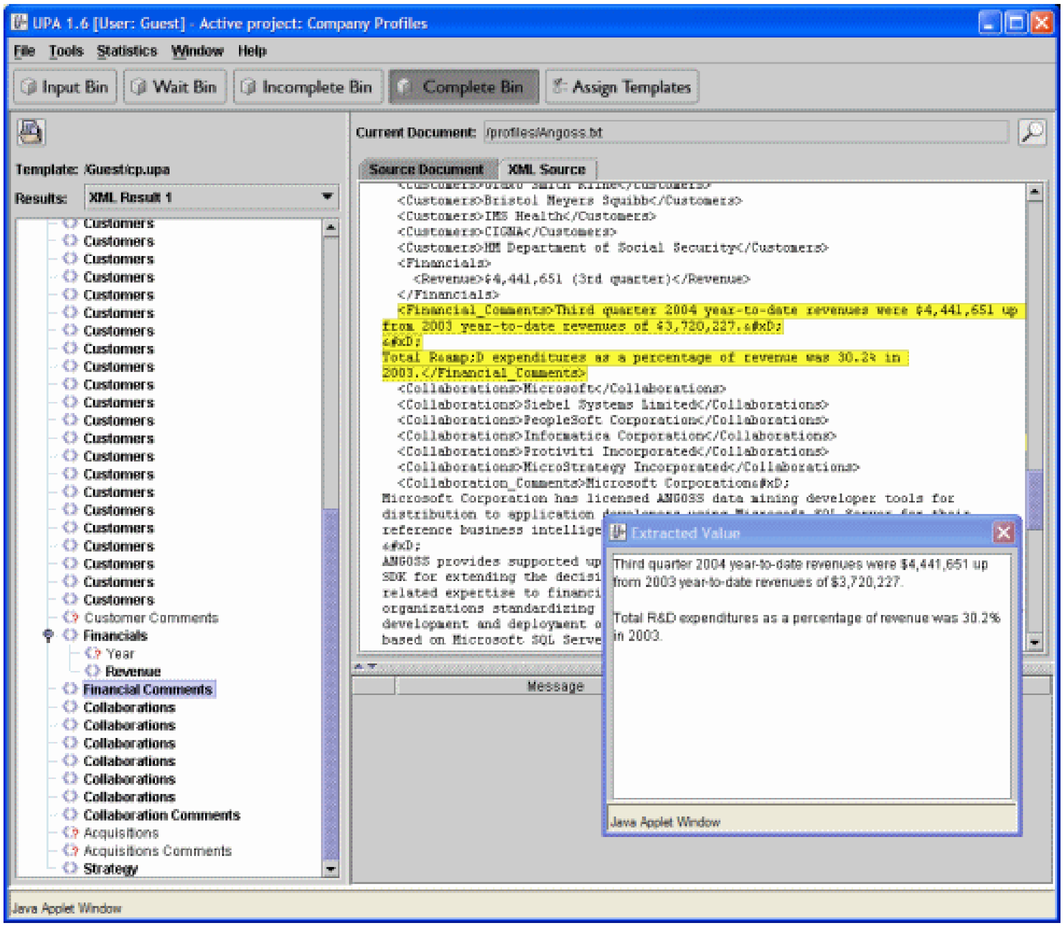
Extracted Value

Third quarter 2004 year-to-date revenues were $4,441,651 up from 2003 year-to-date revenues of $3,720,227.

Total R&D expenditures as a percentage of revenue was 30.2% in 2003.

Java Applet Window

4. To view the document XML output, click the XML Source tab XML Source. The XML file created from the source text will display.



```
  <Customers>Bristol Meyers Squibb</Customers>
  <Customers>IMS Health</Customers>
  <Customers>CIGNA</Customers>
  <Customers>HM Department of Social Security</Customers>
  <Financials>
    <Revenue>$4,441,651 (3rd quarter)</Revenue>
  </Financials>
  <Financial_Comments>Third quarter 2004 year-to-date revenues were $4,441,651 up
from 2003 year-to-date revenues of $3,720,227.&#xD;
&#xD;
Total R&amp;D expenditures as a percentage of revenue was 30.2% in
2003.</Financial_Comments>
  <Collaborations>Microsoft</Collaborations>
  <Collaborations>Siebel Systems Limited</Collaborations>
  <Collaborations>PeopleSoft Corporation</Collaborations>
  <Collaborations>Informatica Corporation</Collaborations>
  <Collaborations>Protiviti Incorporated</Collaborations>
  <Collaborations>MicroStrategy Incorporated</Collaborations>
  <Collaboration_Comments>Microsoft Corporation&#xD;
Microsoft Corporation has licensed ANGOSS data mining developer tools for
distribution to application
reference business intellige
&#xD;
ANGOSS provides supported up
SDK for extending the decisi
related expertise to financi
organizations standardizing
development and deployment o
based on Microsoft SQL Serve
```

Third quarter 2004 year-to-date revenues were $4,441,651 up from 2003 year-to-date revenues of $3,720,227.

Total R&D expenditures as a percentage of revenue was 30.2% in 2003.

5. Clicking on the fields in the Results panel will display the XML extracted data in the Extracted Value window.

Note: You can not change the source text document or the XML Source.

Exiting UPA

When you are finished working, close the UPA application. You do not need to have UPA running on your PC for batch processing to take place. To exit and close UPA, perform the following procedure.

1. Before you exit UPA, remember to save any of the work you want to keep by selecting Save from the File menu or by clicking Save 💾.
2. To exit the UPA main window, from the File menu, select Quit. The UPA main window will close. Your browser window will display the UPA WebSite.
3. You can now exit the UPA WebSite window in your browser by closing the browser window. A message similar to the following will display:

   Are you sure you want to navigate away from this page?

4. Click Ok. The UPA WebSite window will close.

Creating Input Templates

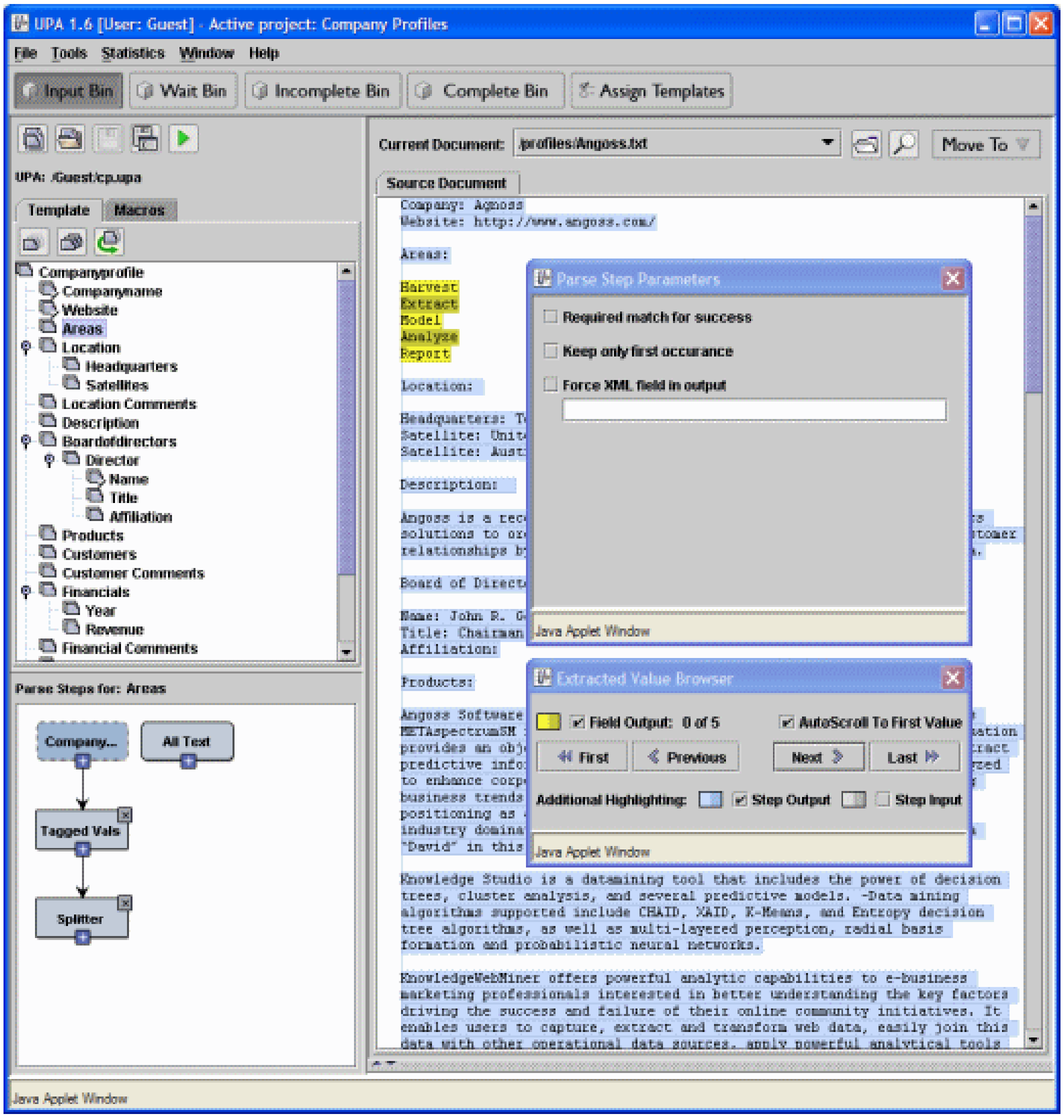
The main task you will perform from the Input Bin is to create templates used by UPA for document input processing. This task can consist of several parts:

- Creating a New Template
    - Multi-Step Parsing
    - Generating a Template Automatically
- Modifying an Existing Template

Creating a New Template

To create a new template, you will need to define the parsing patterns which will be part of the template. These patterns will be used later during batch document processing. Templates are created using typical text documents from your database as a guide to the patterns you will build. To create a template for document processing, start UPA from your internet browser and perform the following procedure.

1. Select and open a text document. The UPA Input Bin template builder will display, along with a source document. The UPA Template Builder will display a new template.



UPA: /Guest/cp.upa

Parse Steps for: Areas

Current Document: /profiles/Angoss.txt

```
Company: Angoss
Website: http://www.angoss.com/

Areas:

Harvest
Extract
Model
Analyze
Report

Location:

Headquarters: T
Satellite: Unit
Satellite: Aust

Description:

Angoss is a rec
solutions to or
relationships b

Board of Direct

Name: John R. G
Title: Chairman
Affiliation:

Products:

Angoss Software
METAspectrumSM
provides an obj
predictive info
to enhance corp
business trends
positioning as
industry domina
"David" in this

Knowledge Studio is a datamining tool that includes the power of decision
trees, cluster analysis, and several predictive models. -Data mining
algorithms supported include CHAID, XAID, K-Means, and Entropy decision
tree algorithms, as well as multi-layered perception, radial basis
formation and probabilistic neural networks.

KnowledgeWebMiner offers powerful analytic capabilities to e-business
marketing professionals interested in better understanding the key factors
driving the success and failure of their online community initiatives. It
enables users to capture, extract and transform web data, easily join this
data with other operational data sources, apply powerful analytical tools
```

Parse Step Parameters:

☐ Required match for success

☐ Keep only first occurance

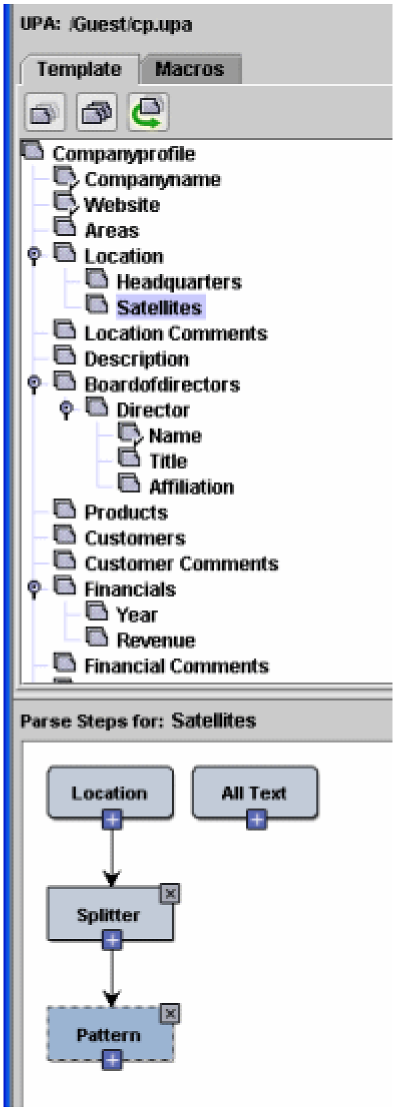
☐ Force XML field in output

Extracted Value Browser:

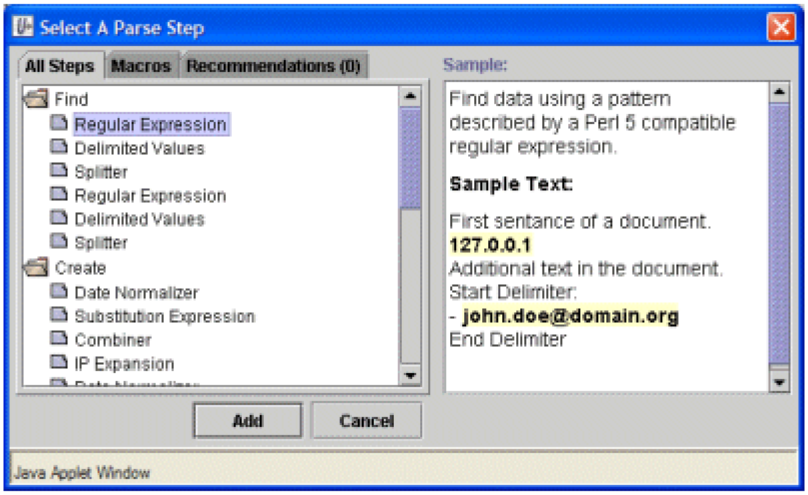
☑ Field Output: 0 of 5 ☑ AutoScroll To First Value

Additional Highlighting: ☑ Step Output ☐ Step Input

2. You can work with the default template that initially displays or you can create an entirely new template by clicking New Template 🗋. If you want to open another existing template, click the Open Template button 📂. **Note**: The status of items in the list of template parser elements is indicated by template icons beside each element.
3. Select an item from the list of template parser elements. The parse steps, if any, will display in the Parse Steps for: [parse step name] area. If no parse steps have been defined, the default parse steps will display.



4. To define a parse step, click on the plus ⊞ on the parse step box. The Select a Parse Step window will display.



- Find
  - Regular Expression
  - Delimited Values
  - Splitter
  - Regular Expression
  - Delimited Values
  - Splitter
- Create
  - Date Normalizer
  - Substitution Expression
  - Combiner
  - IP Expansion

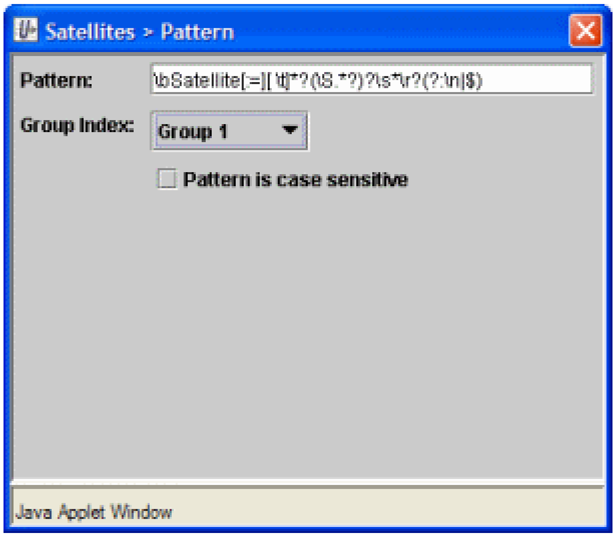
Sample:

Find data using a pattern described by a Perl 5 compatible regular expression.

Sample Text:

First sentance of a document.
127.0.0.1
Additional text in the document.
Start Delimiter:
- **john.doe@domain.org**
End Delimiter

5. Select the type of parse step you want to add from the list. As you click on a type of step, an example of the type of parsing output will display in the Sample column. Click Add. The type of pattern you selected will display in the Parse Step pattern window.



Pattern: \bSatellite[:=][\t]*?(\S.*?)?\s*\r?(?:\n|$)

Group Index: Group 1

□ Pattern is case sensitive

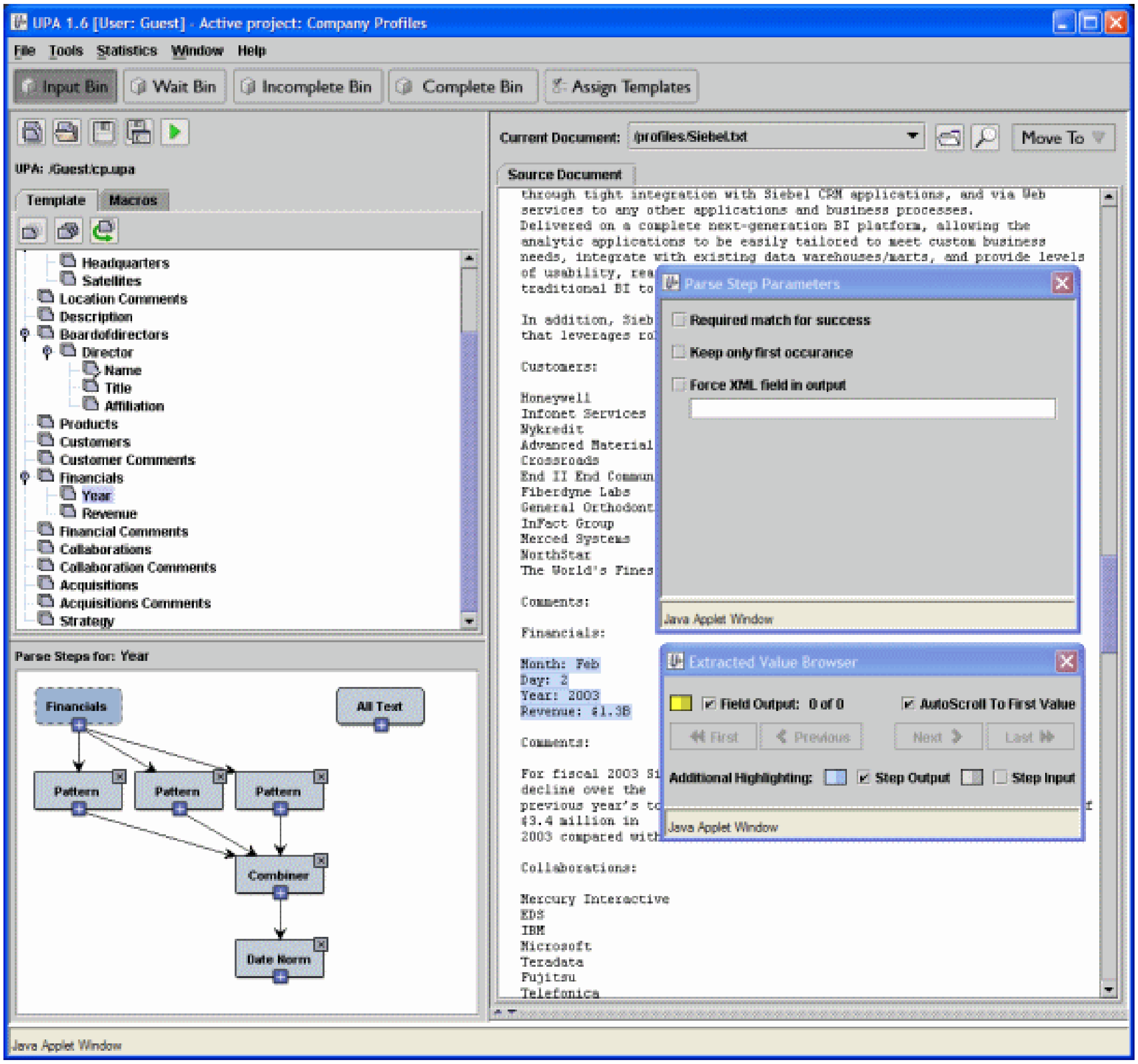
Note that these windows will display different options and parameters depending upon the type of parse step you selected earlier.

6. Add any elements to the parser pattern, customizing it to find the parts of the document you want to save after processing. You can use a combination of PERL regular expressions, words, numbers, or symbols. Once you add a pattern to the selected parser element, it will highlight in bold in the list of template parser elements.

7. Click Save Template 💾 to save your template.

Multi-Step Parsing

For some results, you will want to create a multi-step parser pattern. A multi-step pattern consists of a series of parser steps, the output of one step flowing into another step for further processing. This enables you to not only find information in a document, but to process it into a different and perhaps more useful format. For example, you can create a step to find a date and time in a document, and then create a step to 'normalize' the date/time into a format you prefer, such as day, month, year, hours, seconds, minutes.



Current Document: /profiles/Siebel.txt

```
through tight integration with Siebel CRM applications, and via Web
services to any other applications and business processes.
Delivered on a complete next-generation BI platform, allowing the
analytic applications to be easily tailored to meet custom business
needs, integrate with existing data warehouses/marts, and provide levels
of usability, rea
traditional BI to

In addition, Sieb
that leverages ro

Customers:

Honeywell
Infonet Services
Nykredit
Advanced Material
Crossroads
End II End Commun
Fiberdyne Labs
General Orthodont
InFact Group
Merced Systems
NorthStar
The World's Fines

Comments:

Financials:

Month: Feb
Day: 2
Year: 2003
Revenue: $1.3B

Comments:

For fiscal 2003 Si
decline over the
previous year's to
$3.4 million in
2003 compared with

Collaborations:

Mercury Interactive
EDS
IBM
Microsoft
Teradata
Fujitsu
Telefonica
```

Parse Step Parameters:

- ☐ Required match for success
- ☐ Keep only first occurance
- ☐ Force XML field in output

Extracted Value Browser:

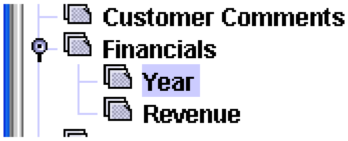
- 🟨 ☑ Field Output: 0 of 0
- ☑ AutoScroll To First Value
- Additional Highlighting: ☑ Step Output ☐ Step Input

In the example above, you can see the final derived date output and the saved multiple steps required, as well as splits, or branches in the process.

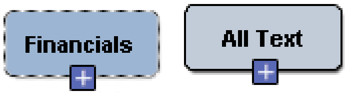
Multi-step Tutorial

To create a multi-step process

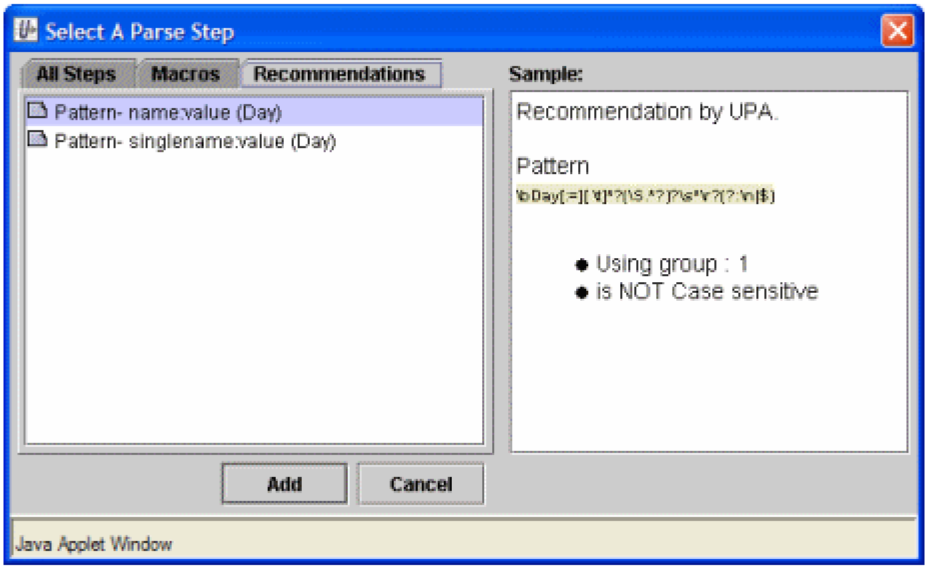
1. Select a template or create a new template.
2. Open a document.
3. From the list of template components, select the template component you want to customize.



The component will highlight and the default Parse Step boxes will display.



4. In the document, highlight the month (February in this example) with the cursor. The Select a Parse Step Window will display a set of options.



Pattern- name:value (Day)
Pattern- singlename:value (Day)

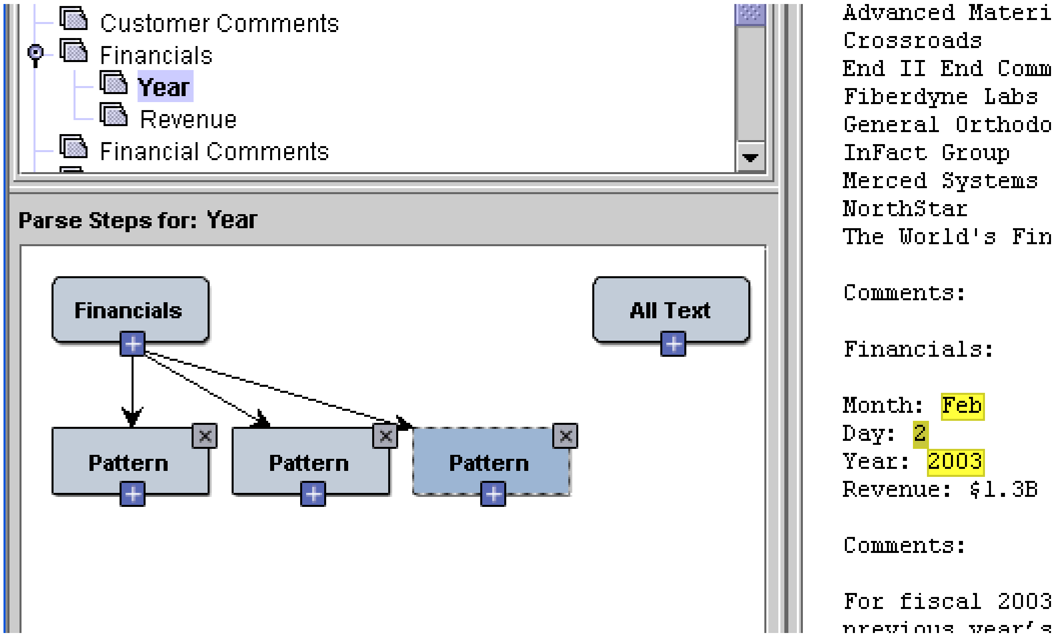
Recommendation by UPA.

Pattern

```
\bDay[:=][ \t]*?(\S.*?)?\s*\r?(?:\n|$)
```

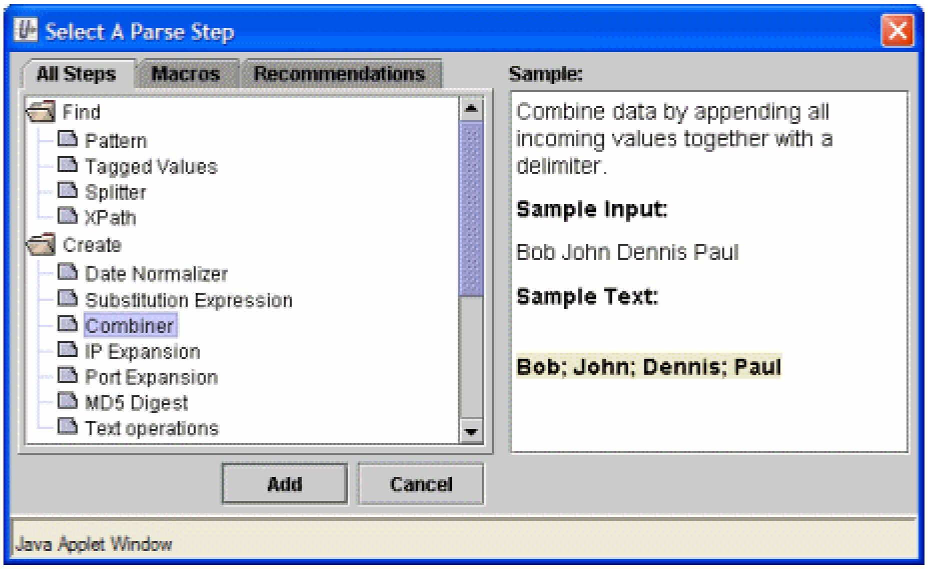
- Using group : 1
- is NOT Case sensitive

5. Click the Add button. The new parser pattern box will be added to the Parse Steps list.
6. Repeat steps 4 and 5, adding a pattern for the day and the year. The Parse Steps list will now show three patterns, each containing a separate date element.



```
Advanced Materi
Crossroads
End II End Comm
Fiberdyne Labs
General Orthodo
InFact Group
Merced Systems
NorthStar
The World's Fin

Comments:

Financials:

Month: Feb
Day: 2
Year: 2003
Revenue: $1.3B

Comments:

For fiscal 2003
previous year's
```

7. On the first Parse Step Pattern box, click the plus ⊞. The Select a Parse Step window will display.



Sample:

Combine data by appending all incoming values together with a delimiter.

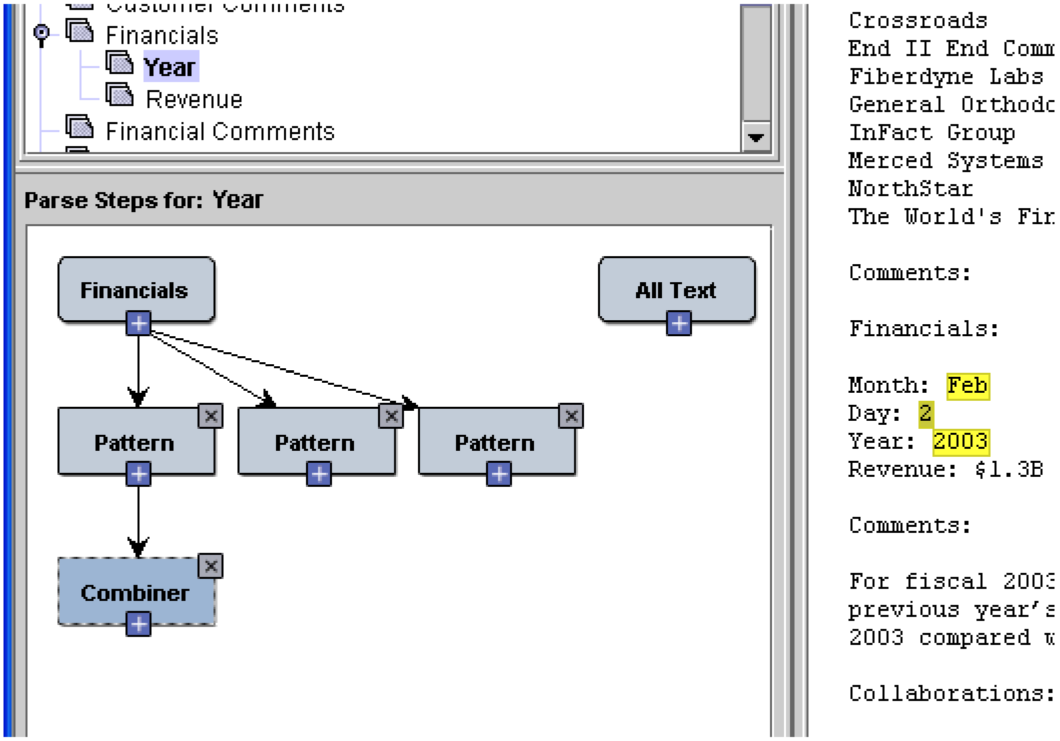
Sample Input:

Bob John Dennis Paul

Sample Text:

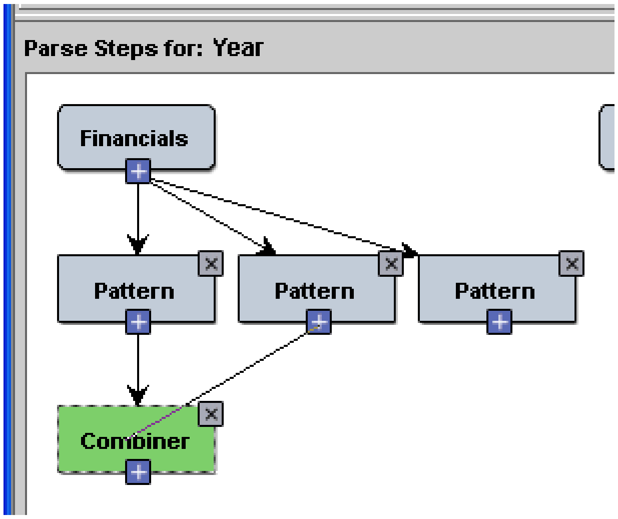
Bob; John; Dennis; Paul

8. For this example, select Combiner from the list of Parse Steps. Combiner examples will display in the Sample area on the right.
9. Click Add. The Combiner box will display in the Parse Steps list.

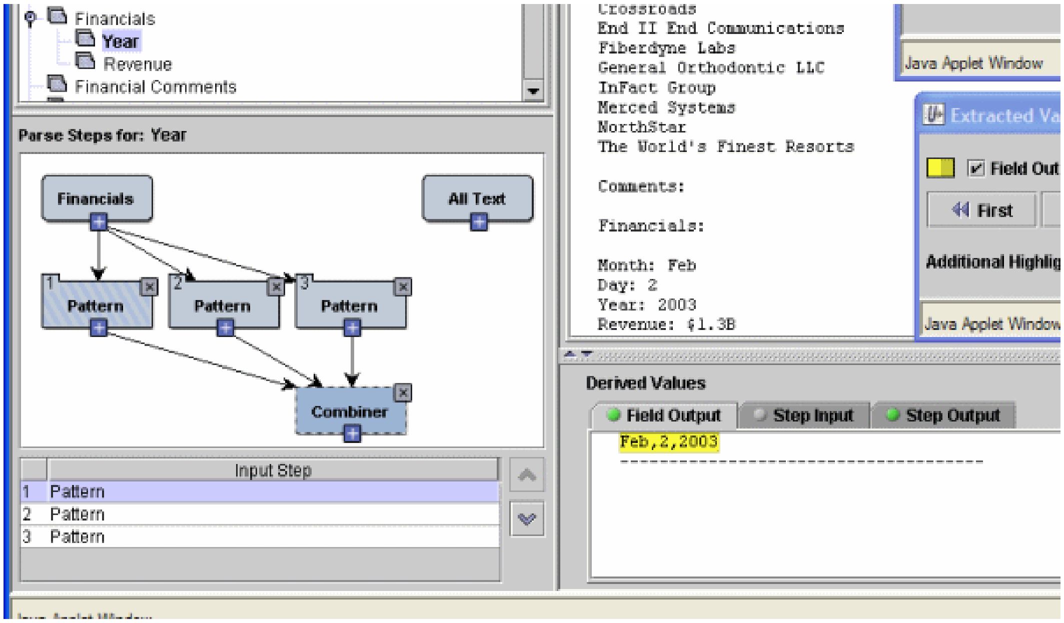


```
Crossroads
End II End Comm
Fiberdyne Labs
General Orthodc
InFact Group
Merced Systems
NorthStar
The World's Fir

Comments:

Financials:

Month: Feb
Day: 2
Year: 2003
Revenue: $1.3B

Comments:

For fiscal 2003
previous year's
2003 compared w

Collaborations:
```

10. Click in the second Pattern box and drag a join line from the Pattern box to the Combiner box.



11. Click in the third Pattern box and drag a join line from the Pattern box to the Combiner box.



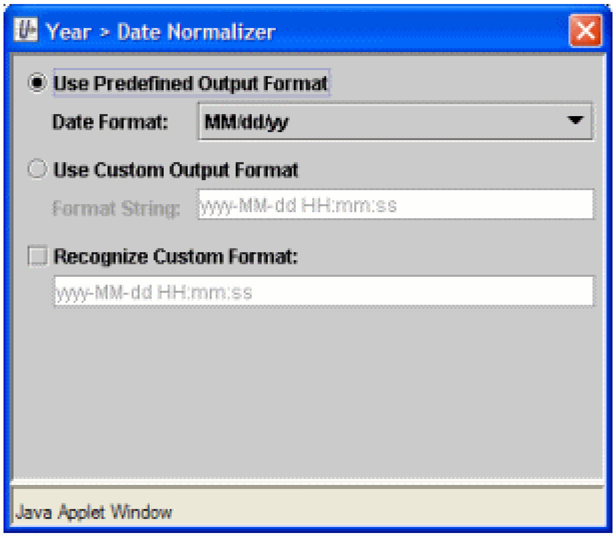
	Input Step
1	Pattern
2	Pattern
3	Pattern

```
Crossroads
End II End Communications
Fiberdyne Labs
General Orthodontic LLC
InFact Group
Merced Systems
NorthStar
The World's Finest Resorts

Comments:

Financials:

Month: Feb
Day: 2
Year: 2003
Revenue: $1.3B
```

```
Feb,2,2003
---------------------------------------
```

The three elements of the date patterns will now be joined in the Combiner box. You can view the Field Output in the Derived Values panel. The order of the defined patterns display in the Input Step panel. You can change the order of the patterns by using the up and down arrows on the side of the panel.

12. On the Combiner box click the plus ⊞. The Select a Parse Step window will display.
13. From the Select a Parse Step window, select Date Normalizer and click Add.
14. The Date Normalizer box will be added to the parse steps. The output from the Combiner will now flow into the Date Normalizer. While the Date Normalizer box is selected, the Date Normalizer window will display the options available.



● Use Predefined Output Format

Date Format: MM/dd/yy

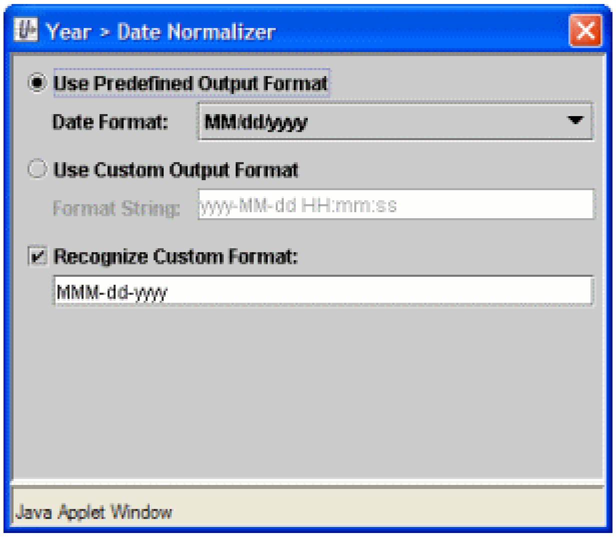
○ Use Custom Output Format

Format String: yyyy-MM-dd HH:mm:ss

□ Recognize Custom Format:

yyyy-MM-dd HH:mm:ss

15. Click the Recognize Custom Format box and enter the format in which you want the combined date to display. For this example, enter MMM-dd-yyyy.



Year > Date Normalizer

● Use Predefined Output Format

Date Format: MM/dd/yyyy

○ Use Custom Output Format

Format String: yyyy-MM-dd HH:mm:ss

☑ Recognize Custom Format:

MMM-dd-yyyy

Java Applet Window

These results will be saved in the XML output file when the parser processes all documents in the Input Bin.

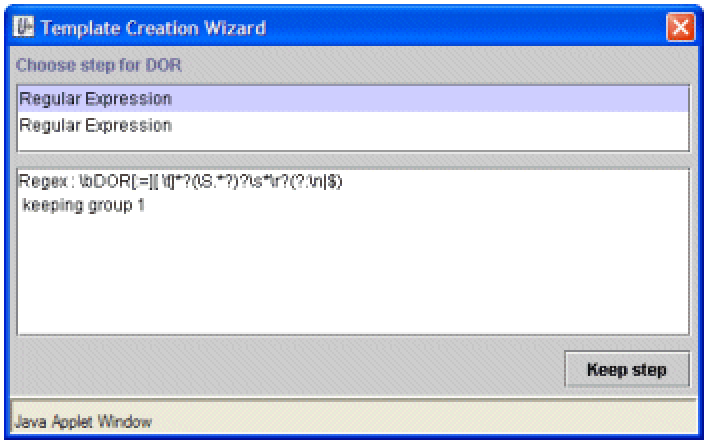
16. Click Save Template to save your additions to the template.

Generating a Template Automatically

UPA provides a tool which will generate a basic template for you, given structured input. To generate a template automatically

1. Select and open a text document. The UPA Input Bin template builder will display, along with a source document.

   Or, click the Create a New Template button . The UPA Template Builder will display a new template.
2. On the Template tab toolbar, click the Automatically Recommend Steps button . The Template Creation Wizard window will display.
3. Click Begin Processing. The Template Wizard will begin stepping through the source document.



```
Regex : \bDOR[:=|[ \t]*?(\S.*?)?\s*\r?(?:\n|$)
 keeping group 1
```

4. When the wizard finds a spot that could be parsed with more than one possible expression, it will pause and display the possible parsing expressions you could use.
5. Click on each expression in the list. The full expression offered by the wizard will display in the bottom field.
6. Select one of the expressions and click Keep Step. The wizard will continue to go through the document.
7. Continue to select between expressions as necessary. The wizard will display the Completed window.
8. Click Done to exit the Template Creation Wizard.
9. The newly defined elements of the template will highlight in bold in the Template area of the Input Bin window.
10. Finish the template by adding parse steps to any other template elements not defined by the wizard.
11. Click Save Template or Save Template As to save your template.

Parse Steps Reference

To process your documents, you must first build a parser sequence into the template. A parser sequence can be made up of the following elements.

- Find
- Create
- Filter
- Validate
- Decision

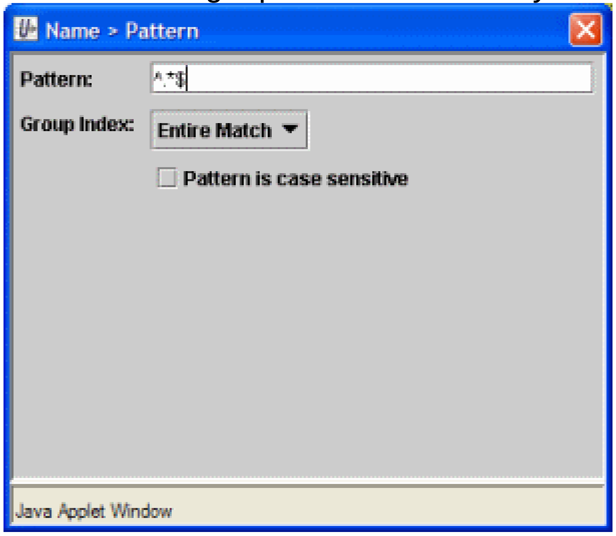
Each element performs a set task. Elements can be combined in many ways to find very specific patterns of information.

Find

Find uses regular Perl expressions to seek out and tag parts of a document that have set components as a consistent part of their structure.

Pattern

Find data using a pattern described by a Perl 5 regular expression.



Name > Pattern

Pattern: ^.*$

Group Index: Entire Match

□ Pattern is case sensitive

Java Applet Window

- Pattern - A regular expression string which describes the pattern of text you are searching for.
- Group Index - a pattern can have parentheses in it, each parenthetical pair defines a group. The group index is the specific group you would like the parser to keep.

Check the Pattern is Case Sensitive box if you need the pattern to be case sensitive. For example if you are searching for “Cat”, the system will not select “cat”.

Pattern Example

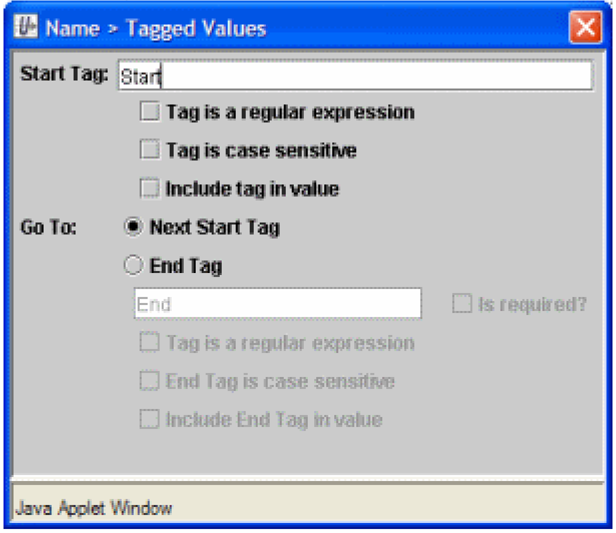
\bDOCCLASS[:=][\t]*?(\S.*?)?\s*\r?(?:\n|$)

The above example pattern will match the string "DOCCLASS" followed by a : or = character, plus 1 or more spaces or tabs. The parentheses are around the value. The value ends at the end of the line.

Parentheses are referenced by "group numbers", counting from the inside of the parentheses. Exception parentheses of the form (?:) are considered non-grouping parentheses and do not count toward the "group numbers".

Tagged Values

Define a word or phrase of text (a string of characters) to serve as a 'tag' which the parser will use as a marker. During processing the parser will select text within the markers you have defined. To set up tagged values you must define a Start Tag and a Go To.



Name > Tagged Values

Start Tag: Start

☐ Tag is a regular expression

☐ Tag is case sensitive

☐ Include tag in value

Go To: ● Next Start Tag

○ End Tag

End ☐ Is required?

☐ Tag is a regular expression

☐ End Tag is case sensitive

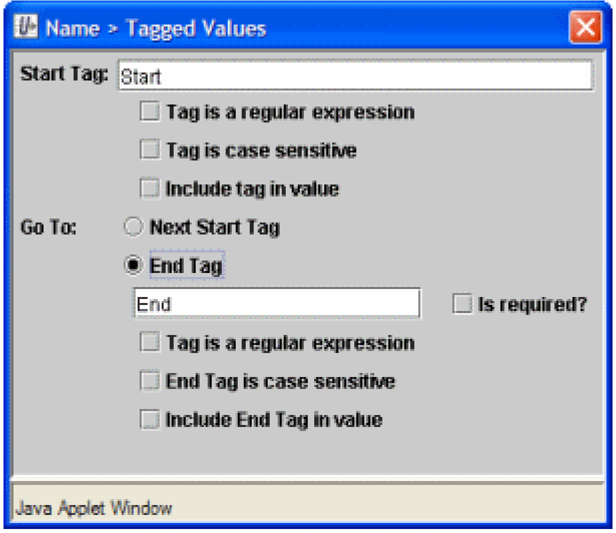
☐ Include End Tag in value

Java Applet Window

Start Tag - a string which describes the beginning of the text block you are searching for. A start tag can be any one of the following.

- Tag is Regular Expression - the start tag will be interpreted as a regular Perl expression.
- Tag is Case Sensitive – check to indicate the start tag is case sensitive. For example if you are searching for “Cat”, the system will not select “cat”.
- Include Tag in Value – check to include the start tag in the extracted text.

Go To - select one of two options for describing the end of a tagged value, the Next Start Tag and the End Tag.



Name > Tagged Values

Start Tag: Start

□ Tag is a regular expression

□ Tag is case sensitive

□ Include tag in value

Go To: ○ Next Start Tag

● End Tag

End □ Is required?

□ Tag is a regular expression

□ End Tag is case sensitive

□ Include End Tag in value

Java Applet Window

- Next Start Tag - each time a start tag is found it marks the end of a tagged value and the beginning of the next.
- End Tag - the string which describes the ending of the text block you are searching for.

You can also refine your selection by checking one of the following options.

- Check the Is Required? box if an end tag must be found in order to extract the tagged value. If you leave the box unchecked, the text will be extracted up until the next start tag or end tag.
- Check the Tag is Regular Expression to have the end tag interpreted as a regular expression.
- Check End Tag is Case Sensitive to indicate the end tag is case sensitive. For example if you are searching for “Cat”, the system will not select “cat”.

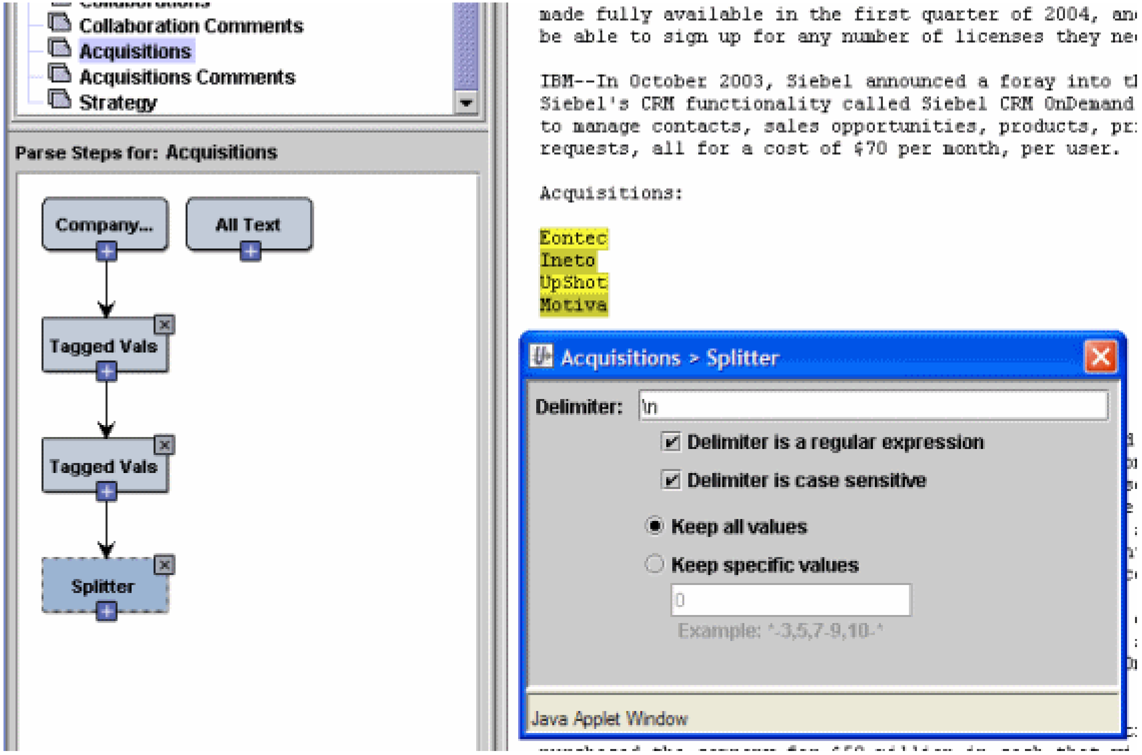
- Check Include End Tag in Value to have the end tag included in the extracted text.

Splitter

Use the Splitter to split data into separate elements based on a delimiter you define. A delimiter is the set of characters in between each of the values you wish to extract. A delimiter can be one character, like a comma or colon, or it can be a word, set of words, or mixture of characters. To define a splitter, type the delimiter characters in the Delimiter field.

- Check the Delimiter is a Regular Expression box if the delimiter will be interpreted as a regular expression.
- Check the Delimiter is Case Sensitive box if you need the delimiter to be case sensitive.

You can check none, one, or both boxes.



```
made fully available in the first quarter of 2004, an
be able to sign up for any number of licenses they ne

IBM--In October 2003, Siebel announced a foray into t
Siebel's CRM functionality called Siebel CRM OnDemand
to manage contacts, sales opportunities, products, pr
requests, all for a cost of $70 per month, per user.

Acquisitions:

Eontec
Ineto
UpShot
Motiva
```

Acquisitions > Splitter

Delimiter: \n

☑ Delimiter is a regular expression

☑ Delimiter is case sensitive

● Keep all values

○ Keep specific values

0

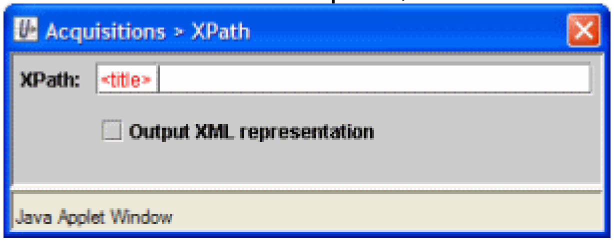
Example: ^-3,5,7-9,10-^

Java Applet Window

- Select Keep All Values to indicate all the values split out of the incoming value are output values. This means all values will be kept and passed on to the next step in the parser, or kept in the XML output file.
- Select Keep Specific Values to keep only the values specified as output values. You can specify numbers, ranges, open-ended ranges, and any combinations of the three.

XPath

Use xpath to find and collect parts of a document tagged with XML embedded in the data. To use this option, the document data must contain valid XML tags.



Acquisitions > XPath

XPath: <title>

☐ Output XML representation

Java Applet Window

Check Output XML Representation to have the data output marked up with XML tags. If you do not, each value from the collected XML elements becomes output.

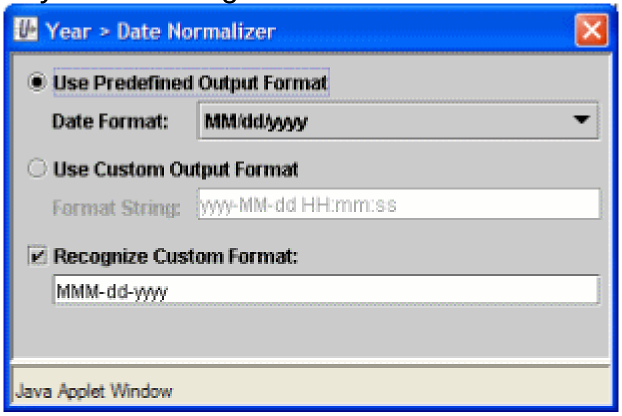
More information on XPath can be found here:
http://www.w3schools.com/xpath/default.asp.

Create

Create enables you to create parser elements that turn data found in documents into other forms more useful for further processing. Use the Create elements to standardize dates, substitute one form of data for another, combine pieces of data, and expand some data.

Date Normalizer

Use the Date Normalizer to save dates found in documents to a standard format of your choosing.



Year > Date Normalizer

● Use Predefined Output Format

Date Format: MM/dd/yyyy

○ Use Custom Output Format

Format String: yyyy-MM-dd HH:mm:ss

☑ Recognize Custom Format:

MMM-dd-yyyy

Java Applet Window

- To use a standard date format, select Use Predefined Output Format. After you select the option, choose a date from the Date Format drop-down list of predefined date output formats.
- Select Use Custom Output Format to define a specific date output format. After you select this, fill in the Format String field with the date and time pattern you want.

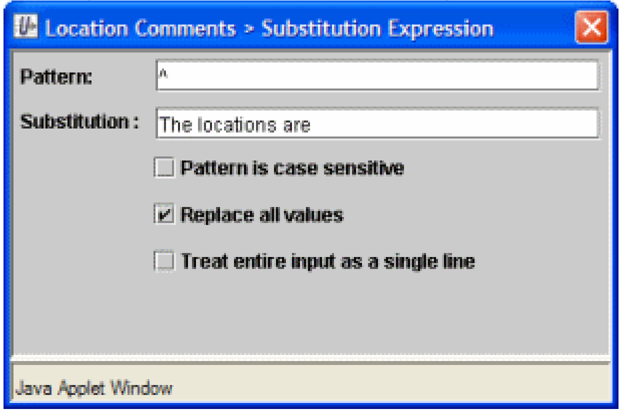
Check Recognize Custom Format to have your custom date and time pattern recognized by the parser in addition to the predefined formats.

Letter	Date or Time Component	Presentation	Examples
G	Era designator	Text	AD
y	Year	Year	1996; 96
M	Month in the year	Month	July; Jul; 07
w	Week in the year	Number	27
W	Week in the month	Number	2
D	Day in the year	Number	189
d	Day in the month	Number	10
F	Day of the week in month	Number	2
E	Day in the week	Text	Tuesday; Tue
a	Am/Pm marker	Text	PM
H	Hour in the day (0-23)	Number	0

`k`	Hour in the day (1-24)	Number	`24`
`K`	Hour in the Am/Pm (0-11)	Number	`0`
`h`	Hour in the Am/Pm (1-12)	Number	`12`
`m`	Minute in the hour	Number	`30`
`s`	Second in the minute	Number	`55`
`S`	Millisecond	Number	`978`
`z`	Time zone	General time zone	`Pacific Standard Time`; `PST`; `GMT-08:00`
`Z`	Time zone	RFC 822 time zone	`-0800`

Substitution Expression

Use Substitute Expression to define a Perl regular expression which the parser will replace with a substitute expression.



Location Comments > Substitution Expression

Pattern: ^

Substitution : The locations are

□ Pattern is case sensitive

☑ Replace all values

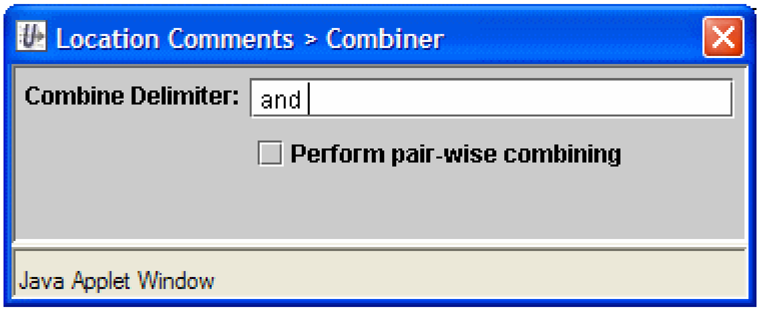
□ Treat entire input as a single line

Pattern - a pattern described by a Perl 5 regular expression which describes the pattern of text you are searching for.
Substitution - the characters comprising the replacement text.

- Check Pattern is Case Sensitive to indicate the pattern you want to replace is case sensitive. For example if you are searching for “Cat”, the system will not select “cat”.
- Check Replace All Values if you want all matches of the pattern to be replaced with the substitution string, otherwise just the first match the parser finds will be replaced.
- Check Treat Entire Input As a Single Line to replace text that may wrap to more than one line in a document.

Combiner

Enables you to combine data by adding the collected data together and separating the data chunks with a delimiter such as 'and' or a comma.



Combine Delimiter: and

□ Perform pair-wise combining

To create a combiner, enter the delimiter in the Combine Delimiter field. The delimiter is the string that is inserted between the incoming values.

Check Perform Pair-wise Combining to combine values from 1 branch (one set of items) with the same indexed value of the other branch (second set of items). For example:

```
                 Branch 1 Branch 2
                 ----------- ---------
        values->  "Fred"  "George" "Smith"   "Jones"
        results->  "Fred Smith", "George Jones"
```

IP Expansion

Use IP Expansion to expand partial IP addresses of the form 127.0.0.x to a list of individual values in the format 127.0.0.0, 127.0.0.1 ... 127.0.0.255 and addresses in the form 127.0.x.x to a list in the format 127.0.0.0, 127.0.0.1 ... 127.0.255.255.

Note: expansion is only supported for the last 2 octets of an IP address.

Port Expansion

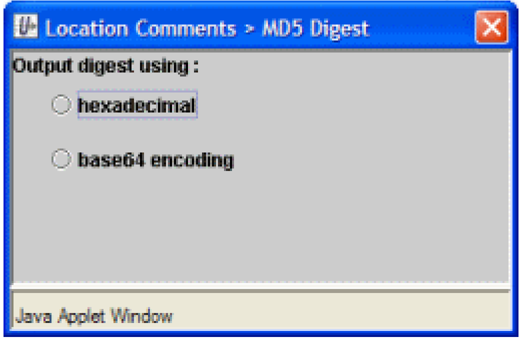
Use Port Expansion to expand the range of common port numbers 25-30 into the individual values 25, 26, 27, 28, 29, 30.

MD5 Digest

You can use MD5 as a convenient way to generate a 'unique' string/key. Message Digest number 9 (MD5) is an algorithm which will generate a 128-bit cryptographic message digest value (also known as a hash or checksum). This hash is generated from the input of a document or file. The hash is used as a digital 'fingerprint' to verify the contents of the document have not been changed. If the file is modified, the checksum changes, signalling the file has been altered from the original version.

For example you can use MD5 to generate a unique key for a document before putting it into a database.

For further information see: http://en.wikipedia.org/wiki/MD5.



Output digest using :

○ hexadecimal

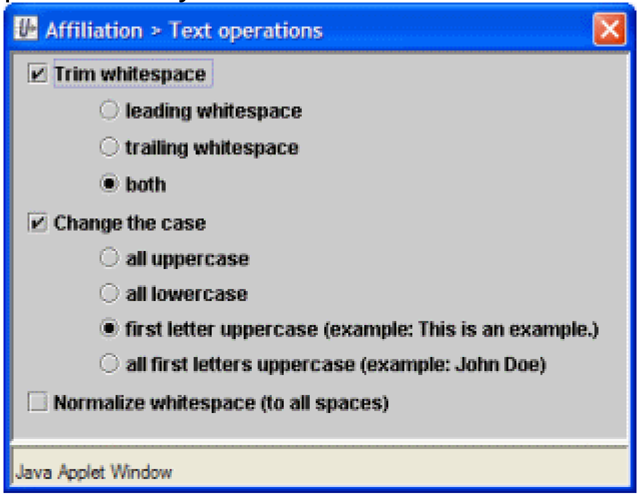
○ base64 encoding

To generate an MD5 digest, select from the following output digest options:

- Hexadecimal - output the format using hexadecimal characters.
- Base64 encoding - create output using a base64 encoding.

Text Operations

You can perform certain common actions on text (strings) in the documents processed by UPA. You can select from the following options.



☑ Trim whitespace

○ leading whitespace

○ trailing whitespace

● both

☑ Change the case

○ all uppercase

○ all lowercase

● first letter uppercase (example: This is an example.)

○ all first letters uppercase (example: John Doe)

□ Normalize whitespace (to all spaces)

- Trim whitespace - Whitespace is considered : tab, carriage return, newlines, spaces, and formfeed characters.
  - Leading whitespace -- remove whitespace characters from the beginning of the string of characters:
    " Fred " becomes "Fred "
  - Trailing whitespace -- remove whitespace characters from the end of the string of characters:
    " Fred " becomes " Fred"
  - Both -- remove whitespace at both ends of the string of characters:

    " Fred " becomes "Fred"

- Change the case

- - All uppercase - change "Fred" to "FRED".
  - All lowercase - change "Fred" to "fred".
  - First letter uppercase - change "the cow is brown" to "The cow is brown".
  - All first letters uppercase - change "the cow us brown" to "The Cow Is Brown".

- Normalize whitespaces - replace any sequence of whitespace characters with a single space. This is useful when you need to store content with newlines removed or for searching purposes.
  For example :
  " The brown cow said,
  "How now brown cow!"
  The fox smiled".
  would result in: " The brown cow said, "How now brown cow!" The fox smiled."

Country Codes

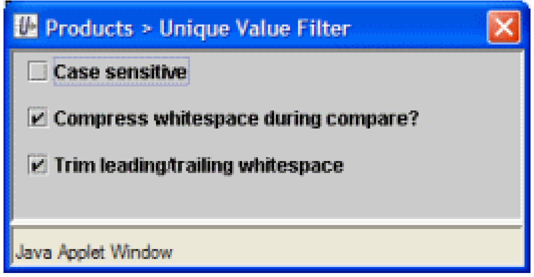
This parse step enables you to convert from a country name to a two letter country code. This step uses the server resource file country_codes.xml.

Filter

The filter options enable you to filter out certain selected items from your data.

Unique Value Filter

Use the unique value filter to remove duplicates of values in your documents. You can select from the following options.



Products > Unique Value Filter

☐ Case sensitive

☑ Compress whitespace during compare?

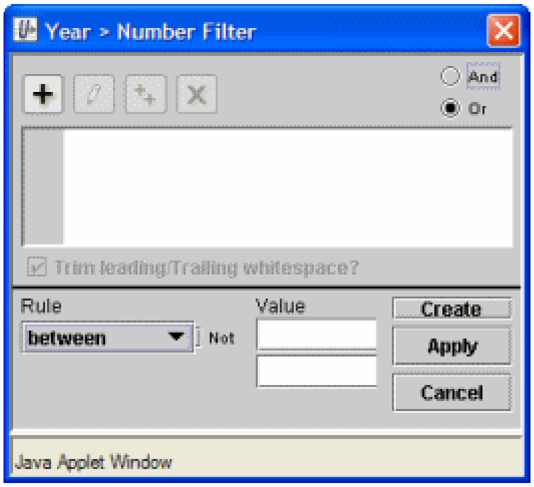
☑ Trim leading/trailing whitespace

Java Applet Window

- Case sensitive - check this option to create a case sensitive filter. For example, "Fred" will not equal "FRED".
- Compress whitespace during compare - when this option is checked, "The brown   cow" will be considered a duplicate of "The brown cow".
- Time leading/trailing whitespace - when this option is checked, " Fred   " will be considered a duplicate of "Fred".

Number Filter

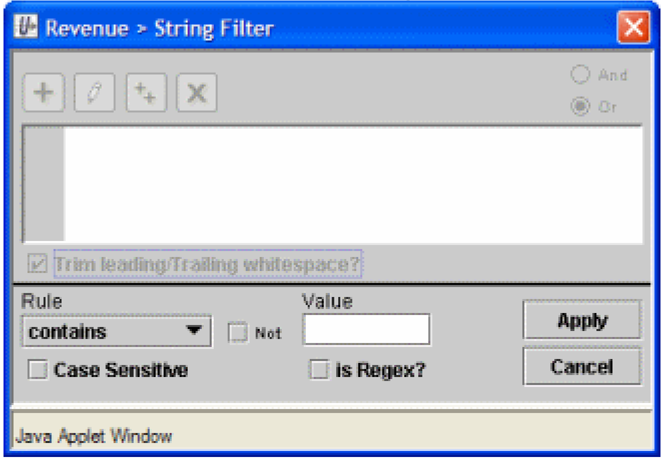
Use the number filter to remove values based on numeric rules such as "is a number" and "between values".



Rule: between

String Filter

Use the string filter to remove values based on string rules such as "contains", "starts with", "ends with", "equals", and "is at least length".



Rule: contains

Validate

You can have the parser validate any items it finds. For example, if all results must begin with a capital letter, the parser will flag the result if it begins with a lower case letter. It will then place the document in either the Wait or Incomplete Bin depending upon the rules you set up for document processing.

Number Validator

The parser will perform a check based on numeric rules such as "is a number" or "between values"

String Validator

The parser will perform a check based on string rules such as "contains", "starts with", "ends with", "equals", and "is at least length".

Decision

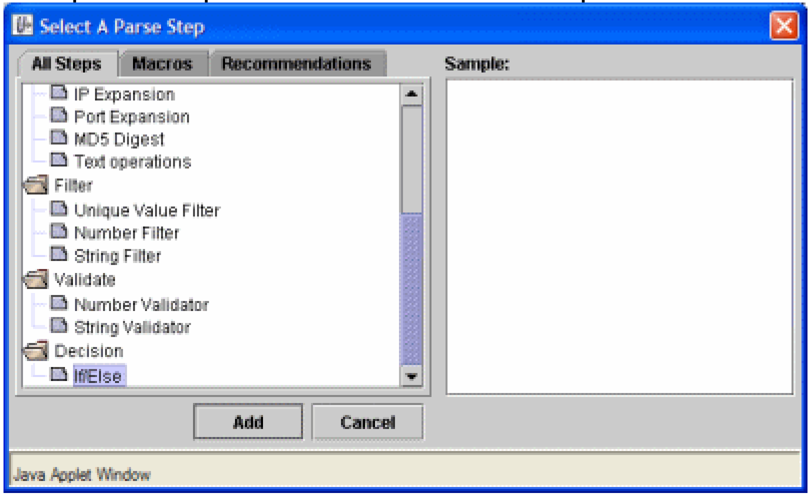
You can build decisions into the parser template by adding an If/Else parser step. Creating an If/Else decision enables you to build a decision point into the template.

For example, if you want to perform a special action on the name Patricia, and no special action on other names, you can set up an If/Else to handle that split in parser actions. Once the parser finds the name Patricia, it will perform one set of actions, while all other names will take a different route and will have a different action performed on them.

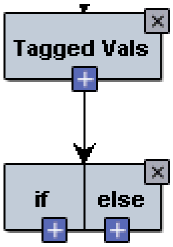
Decision If/Else

To add an If/Else step perform the following procedure.

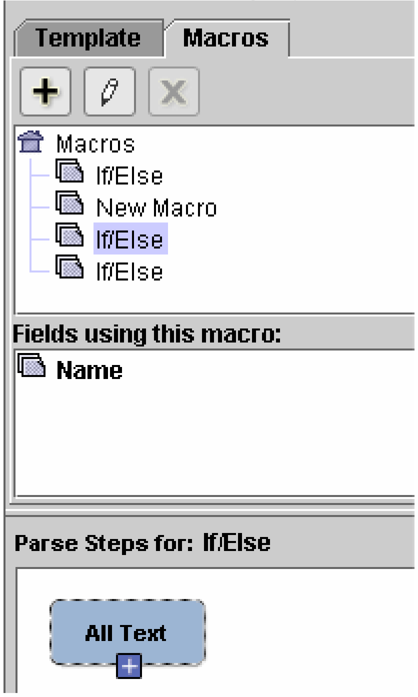
1. Under the parse step where you want the if/else, click the plus to add a new parse step. The Select A Parse Step window will display.



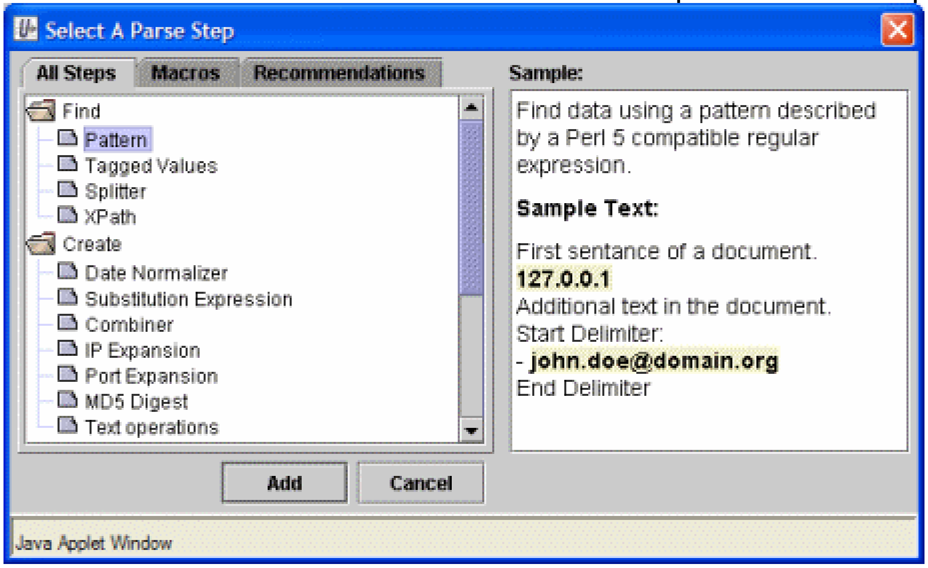
2. From the All Steps tab select If/Else and click Add. The If/Else box will be added to the parse steps under the prior parse step.



3. To set the conditions for the if/else, double-click on the step box. The macro tab will display. This is normal for setting up an if/else process.



4. Click the plus symbol on the parse step box to select the parser conditions for the 'If' side of the statement. In this example, we would click the plus under the All Text box. The Select A Parse Step window will display.



Sample:

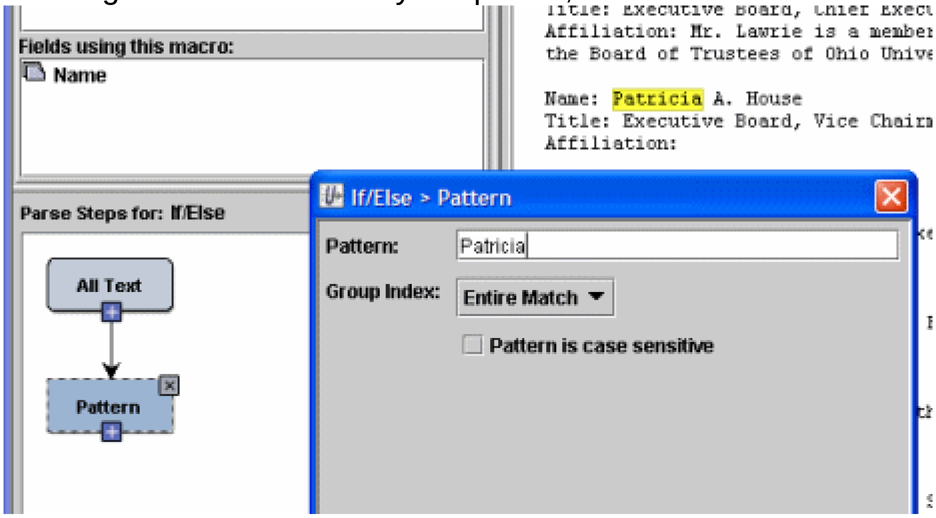
Find data using a pattern described by a Perl 5 compatible regular expression.

Sample Text:

First sentance of a document.
127.0.0.1
Additional text in the document.
Start Delimiter:
- **john.doe@domain.org**
End Delimiter

5. Select a parse step and click Add. In our example, the Pattern parse step is selected.

6. Define the true condition for the if/else. In our example, if the word pattern or string 'Patricia' is found by the parser, the condition is true.



Fields using this macro:

Name

Parse Steps for: If/Else

```
Title: Executive Board, Chief Execu
Affiliation: Mr. Lawrie is a member
the Board of Trustees of Ohio Unive

Name: Patricia A. House
Title: Executive Board, Vice Chairm
Affiliation:
```

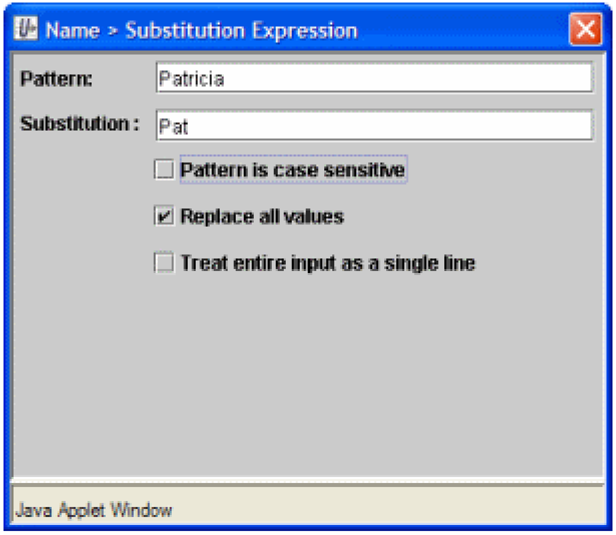
If/Else > Pattern

Pattern: Patricia

Group Index: Entire Match

□ Pattern is case sensitive

7. Click the Template tab and return to the Parse Steps list. Now that you have set up the condition you want, you can set up the operations you want the parser to perform once a match is either found or not found.
8. Click the plus under the if box, select and define the parse step. This step will be performed if the condition is true. In our example, if Patricia is found in a document, the name Pat will be substituted for Patricia.



Name > Substitution Expression

Pattern: Patricia

Substitution: Pat

□ Pattern is case sensitive

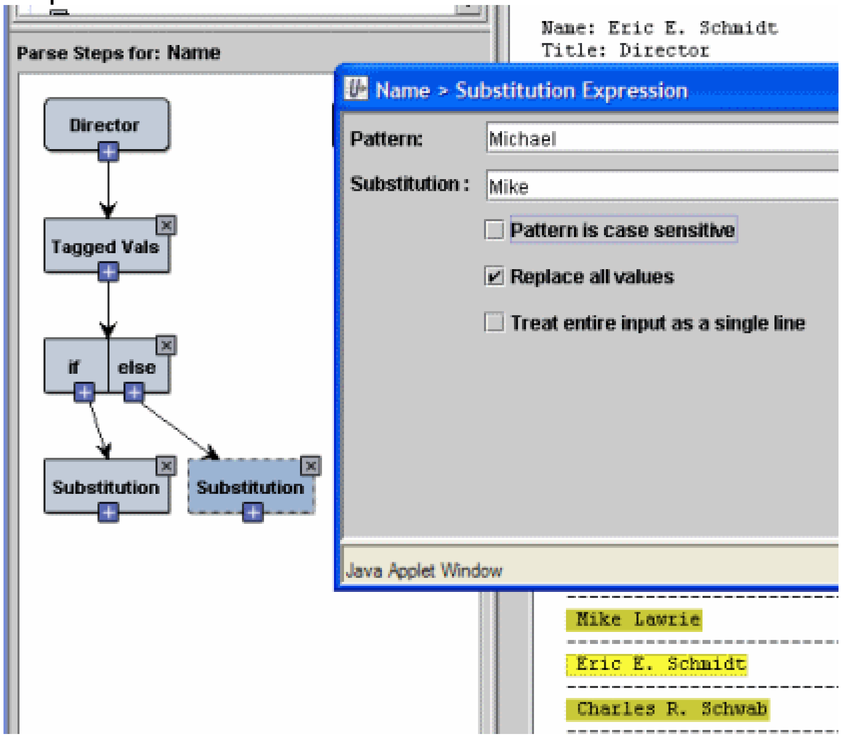
☑ Replace all values

□ Treat entire input as a single line

Java Applet Window

9. Click the plus under the else box, select and define the parse step. This step will be performed if the condition is false. In our example, if any string other than Patricia is found in a document, the alternative action will

be performed.



```
Name: Eric E. Schmidt
Title: Director
```

Pattern: Michael

Substitution : Mike

□ Pattern is case sensitive

☑ Replace all values

□ Treat entire input as a single line

```
Mike Lawrie
Eric E. Schmidt
Charles R. Schwab
```

A further example would be setting up an if/else decision similar to the following.

Beginning with the starting values: "The quick brown dog" and "The slow blue snail", you can set up the if/else decision to act on the patterns.

> If the Pattern --> brown, change to black
> Else perform the Else Step if the Pattern -->blue|red|yellow, change to white

As the parser moves through these steps it will perform either one substitution (changing brown to black) or another substitution (changing (blue|red|yellow) to white).

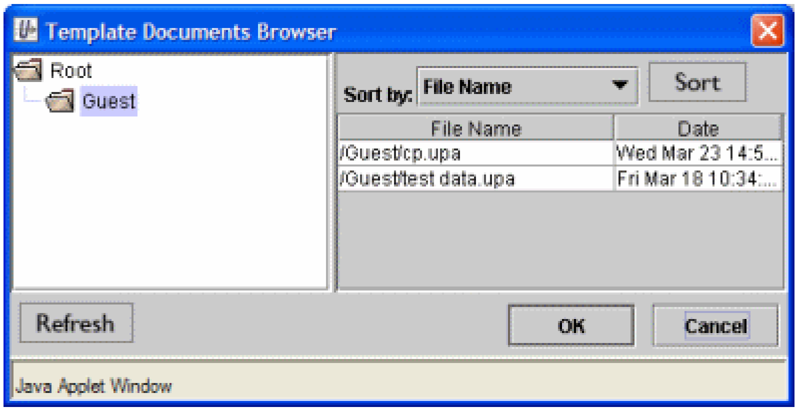
Results "The quick black dog" and "The slow white snail".

Loading a Template

Once you have started UPA, you can load another template at any time. Templates are the heart of UPA, enabling the software to parse a document into useful data chunks. To load a template

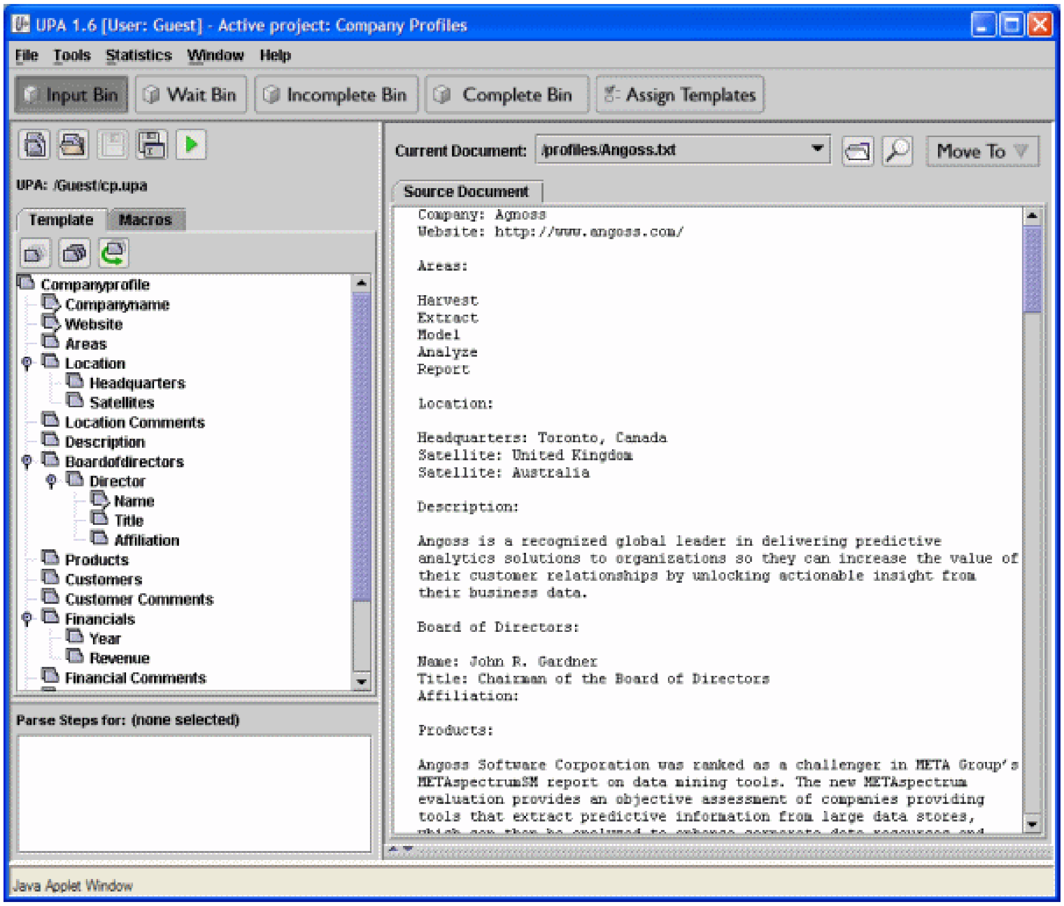
1. From the UPA main window File menu, select Load Template.

   Or click Open Template .
   The Template Document Browser window will display.



File Name	Date
/Guest/cp.upa	Wed Mar 23 14:5...
/Guest/test data.upa	Fri Mar 18 10:34:...

2. You can expand and navigate UPA directories displayed on the right side of the Templates Document Browser. Select a sub-directory to display the list of stored templates.
3. Select a template from the list and click Ok. The template will display in the UPA main window.



```
Company: Agnoss
Website: http://www.angoss.com/

Areas:

Harvest
Extract
Model
Analyze
Report

Location:

Headquarters: Toronto, Canada
Satellite: United Kingdom
Satellite: Australia

Description:

Angoss is a recognized global leader in delivering predictive
analytics solutions to organizations so they can increase the value of
their customer relationships by unlocking actionable insight from
their business data.

Board of Directors:

Name: John R. Gardner
Title: Chairman of the Board of Directors
Affiliation:

Products:

Angoss Software Corporation was ranked as a challenger in META Group's
METAspectrumSM report on data mining tools. The new METAspectrum
evaluation provides an objective assessment of companies providing
tools that extract predictive information from large data stores,
```

Input and Incomplete Bin Toolbars

The UPA Input Bin uses three toolbars to help you perform template building tasks. These toolbars are

- Input Bin Main Toolbar
- Template Tab Toolbar
- Macros Tab Toolbar

Input and Incomplete Bin Main Toolbar

Button	What it Does
📄	Create a new template.
📂	Opens an existing template.
💾	Save the current template and any changes.
💾	Save the current template to a new file name.
▶	Parse the open document with the current template.
📁	Revise the document list, add or delete documents in the list.
🔍	Find a word, phrase, or item in the current open document.

Template Tab Toolbar

Button	What it Does
▨	Hide or Show selected fields in your template.
▨	Toggle any hidden fields.
↻	Automatically recommend template parsing steps.

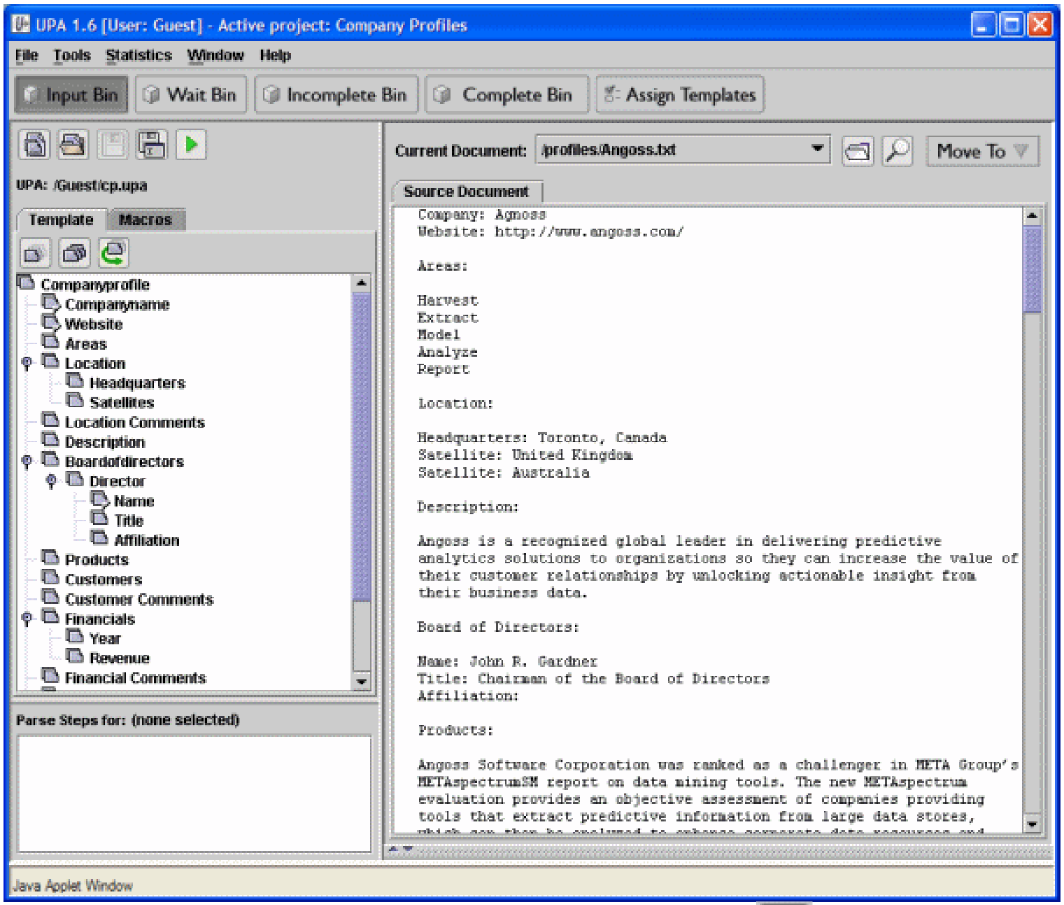
Macros Tab Toolbar

Button	What it Does
+	Add a new macro to your template.
✎	Rename a selected macro.
✖	Delete the selected macro.

Modifying Input Templates

To modify an existing template, you will need to open a previously created template, modify the parsing patterns and save the template. To modify a template for document processing, start UPA from your internet browser and perform the following procedure.

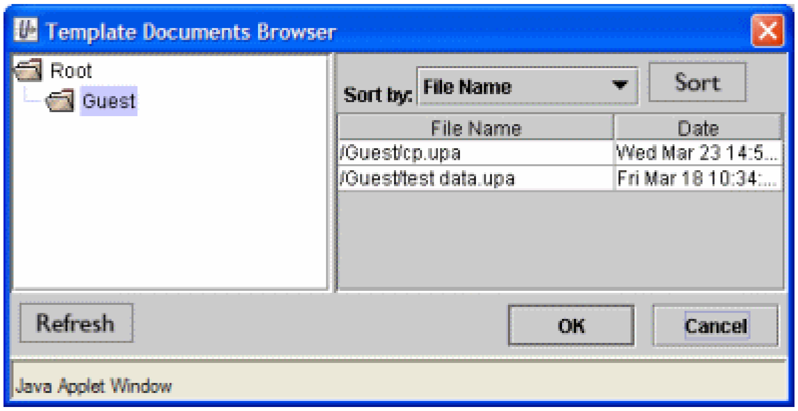
1. Select and open a text document. The UPA Input Bin template builder will display, along with a source document. The UPA Template Builder will initially display a new template.



Current Document: /profiles/Angoss.txt

```
Company: Agnoss
Website: http://www.angoss.com/

Areas:

Harvest
Extract
Model
Analyze
Report

Location:

Headquarters: Toronto, Canada
Satellite: United Kingdom
Satellite: Australia

Description:

Angoss is a recognized global leader in delivering predictive
analytics solutions to organizations so they can increase the value of
their customer relationships by unlocking actionable insight from
their business data.

Board of Directors:

Name: John R. Gardner
Title: Chairman of the Board of Directors
Affiliation:

Products:

Angoss Software Corporation was ranked as a challenger in META Group's
METAspectrumSM report on data mining tools. The new METAspectrum
evaluation provides an objective assessment of companies providing
tools that extract predictive information from large data stores,
```

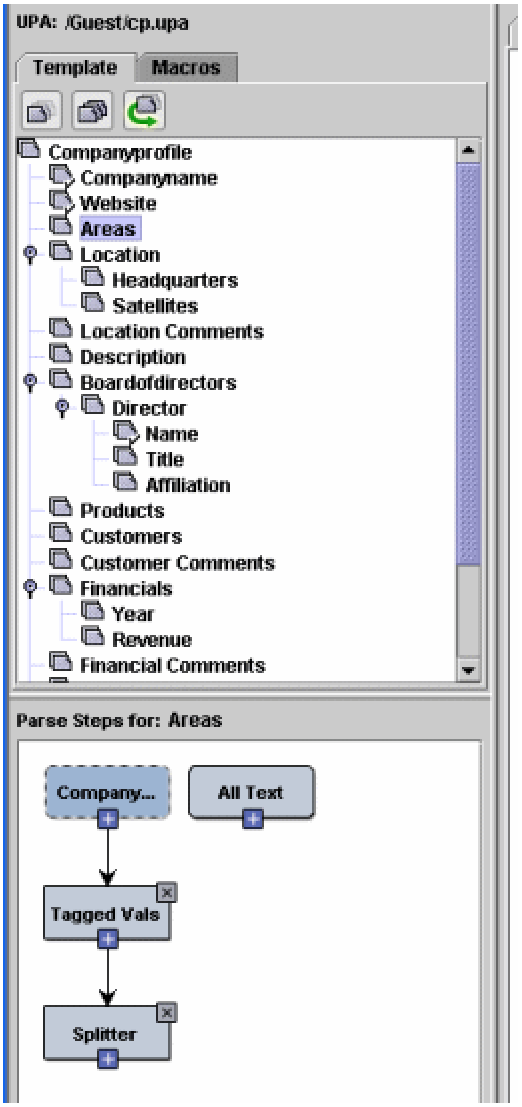
2. To modify an existing template, click Open Template . The Template Documents Browser will display.



Sort by: File Name

File Name	Date
/Guest/cp.upa	Wed Mar 23 14:5...
/Guest/test data.upa	Fri Mar 18 10:34:...

3. Select the template you want to modify from the list and click Ok. The template parser elements will display in the Template tab.

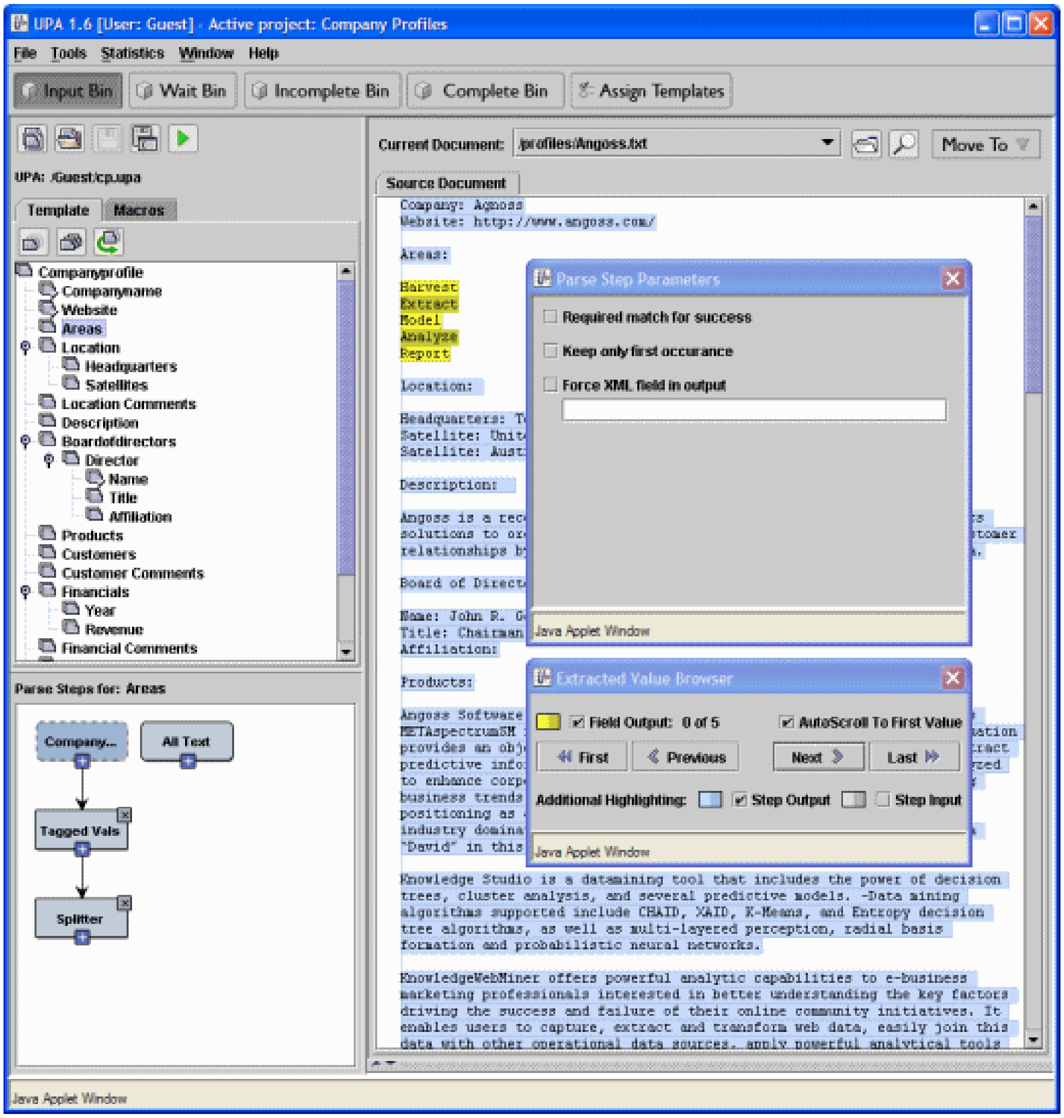


UPA: /Guest/cp.upa

```
Companyprofile
    Companyname
    Website
    Areas
    Location
        Headquarters
        Satellites
    Location Comments
    Description
    Boardofdirectors
        Director
            Name
            Title
            Affiliation
    Products
    Customers
    Customer Comments
    Financials
        Year
        Revenue
    Financial Comments
```

Parse Steps for: Areas

4. Select an item from the list of template parser elements. The parse steps, if any, will display in the Parse Steps for: [parse step name] area. If no parse steps have been defined, the default parse steps will display.



Current Document: /profiles/Angoss.txt

Parse Step Parameters:
- □ Required match for success
- □ Keep only first occurance
- □ Force XML field in output

Extracted Value Browser: ☑ Field Output: 0 of 5 ☑ AutoScroll To First Value

Additional Highlighting: ☑ Step Output □ Step Input

```
Company: Agnoss
Website: http://www.angoss.com/

Areas:

Harvest
Extract
Model
Analyze
Report

Location:

Headquarters: T
Satellite: Unit
Satellite: Aust

Description:

Angoss is a rec                                          rs
solutions to or                                          tomer
relationships b                                          s.

Board of Direct

Name: John R. G
Title: Chairman
Affiliation:

Products:

Angoss Software                                          e
METAspectrumSM                                           ation
provides an obj                                          tract
predictive info                                          rzed
to enhance corp                                          y
business trends
positioning as
industry domina                                          a
"David" in this

Knowledge Studio is a datamining tool that includes the power of decision
trees, cluster analysis, and several predictive models. -Data mining
algorithms supported include CHAID, XAID, K-Means, and Entropy decision
tree algorithms, as well as multi-layered perception, radial basis
formation and probabilistic neural networks.

KnowledgeWebMiner offers powerful analytic capabilities to e-business
marketing professionals interested in better understanding the key factors
driving the success and failure of their online community initiatives. It
enables users to capture, extract and transform web data, easily join this
data with other operational data sources, apply powerful analytical tools
```

5. Modify the parse steps. For basic instructions on working with parse steps, see Creating New Templates and Multi-Step Parsing.
6. Use the Template Toolbar buttons to work with the template. You can
   - Save the modified template
   - Save the modified template to a new file name
   - Open another template to modify
   - Create a new template

- ▶ Test the template by parsing the current document.

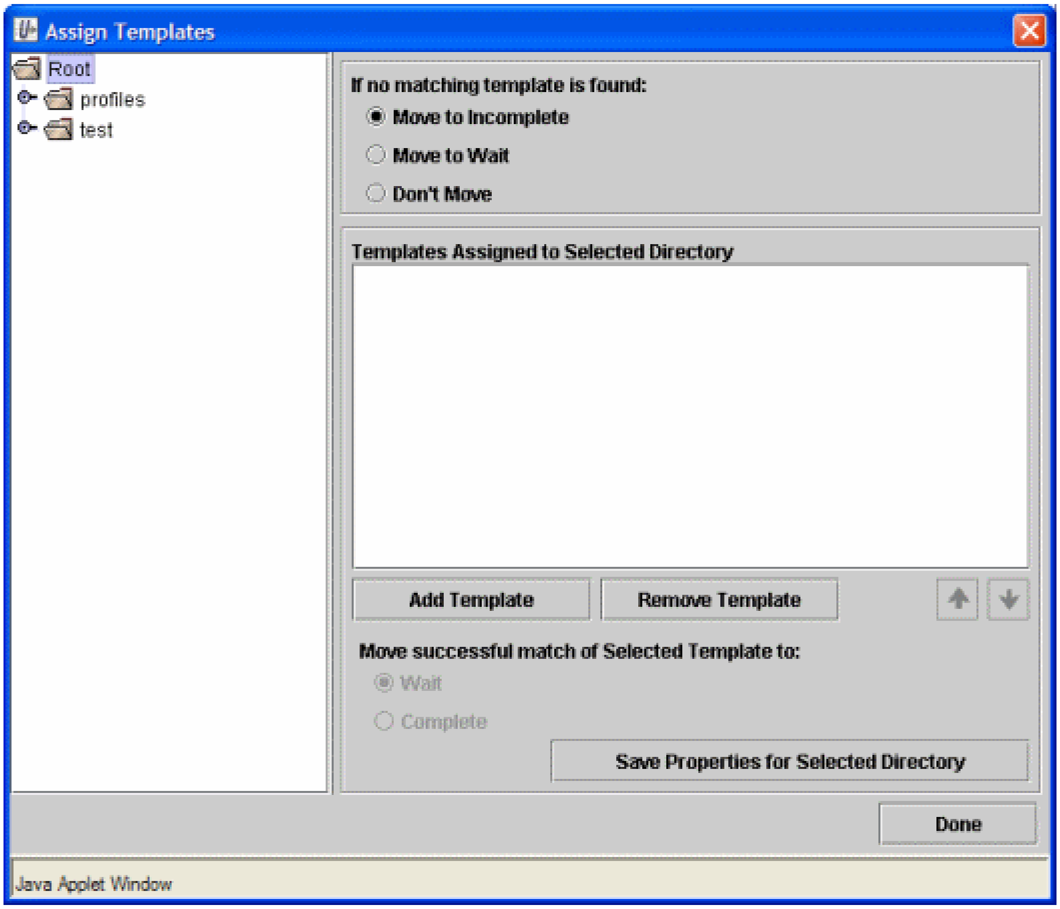
7. After modifying the template, click Save or Save As to keep your work. To discard your changes, open a template without saving your current changes.

Assigning or Removing Templates

When new templates are created, they need to be assigned to the Input Bin document directory. Once a template is assigned to Input, it will be used for the batch processing of any document in the directory. To assign a template, start UPA. From the UPA main window, perform the following procedure.

1. From the UPA main window File menu, select Assign Templates.

   Or click Assign Templates [Assign Templates] on the UPA main window toolbar.
   The Assign Template(s) window will display.

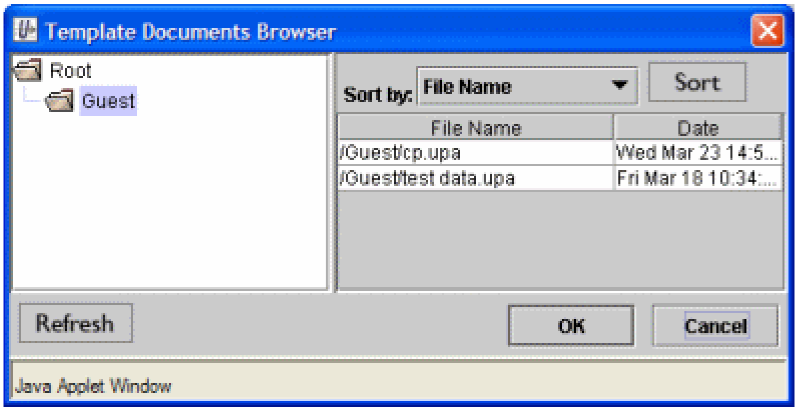


If no matching template is found:
● Move to Incomplete
○ Move to Wait
○ Don't Move

Templates Assigned to Selected Directory

Move successful match of Selected Template to:
● Wait
○ Complete

2. Select and open a directory by clicking on it. Any templates already assigned to a directory will display in the Templates Assigned to Selected Directory window.
3. Add a template to a directory by clicking Add Template. The Template Documents Browser window will display.



File Name	Date
/Guest/cp.upa	Wed Mar 23 14:5...
/Guest/test data.upa	Fri Mar 18 10:34:...

4. Browse the directories by clicking on them. Any template files in the selected directory will display in the file list.
5. Select a file from the file list by clicking on it.
6. Click Ok. The selected template file will be added to the Templates Assigned to Selected Directory area of the Assign Templates window.
7. In the Assign Templates window, use the up and down arrows to change the order of templates in the list. The UPA parser will process documents using the first template in the list, then the second, and so on.
8. To remove an old template from the list, select the template and click Remove Template.
9. Click Save Properties for Selected Directory

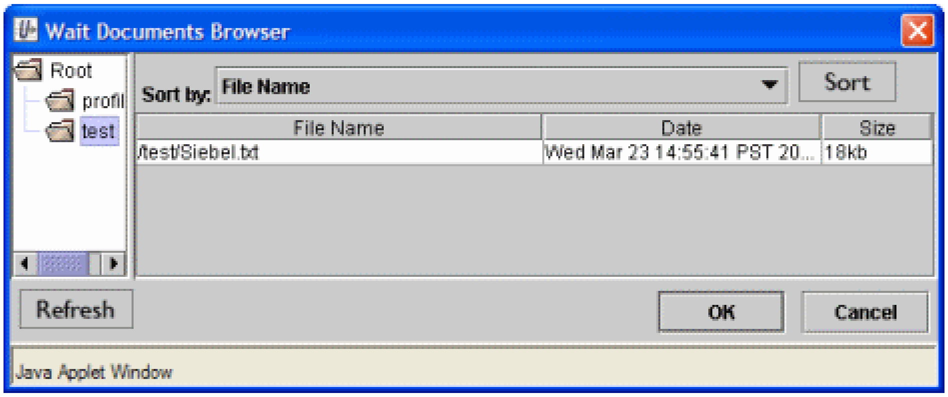
to save the added template file to the selected directory.

Selecting Wait and Incomplete Documents

Documents requiring further work are stored in the Wait and Incomplete bins. Wait Bin documents have been processed and are waiting for user approval, while Incomplete Bin documents have been processed, but did not complete the processing successfully. Wait Bin documents do not necessarily need any hand correction by you, but you can review the document and enter additions and corrections if needed. Incomplete Bin documents require some level of correction by you before they can be processed successfully and moved to the Complete Bin.

Selecting Wait Bin Documents

1. From the UPA main window, click on Wait Bin Wait Bin. The Wait Documents Browser will display.



File Name	Date	Size
/test/Siebel.txt	Wed Mar 23 14:55:41 PST 20...	18kb

2. From the left hand side of the Documents Browser, open a folder and select a document from the bin.

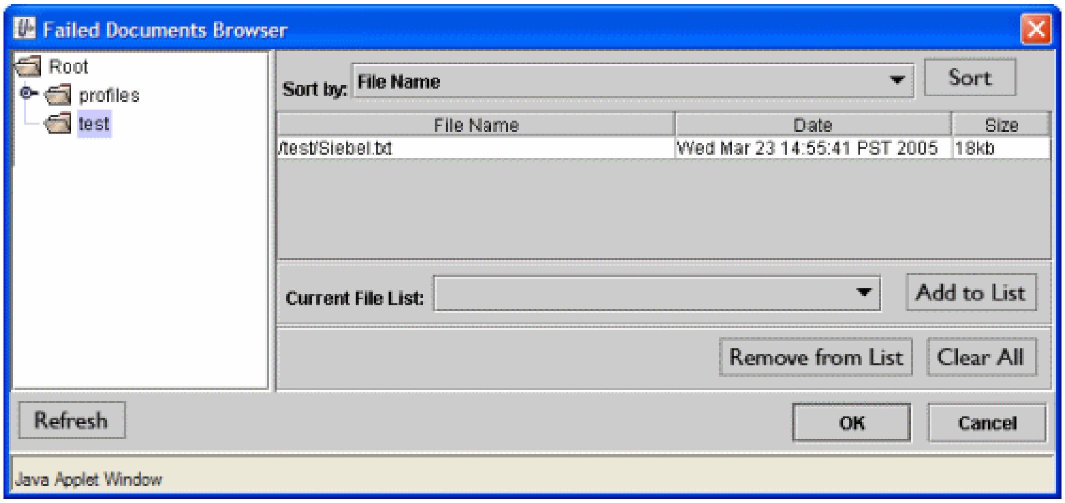
Tips:

- You can select only one document at a time.
- You can open a document by double-clicking on a file name.
- You can open documents in other directories by using the directory folders on the left side of the Document Browser window.

3. Click OK. The selected document will display in the Current Document work area.

Selecting Incomplete Bin Documents

1. From the UPA main window, click on Incomplete Bin . The Failed Documents Browser will display.



Sort by: File Name

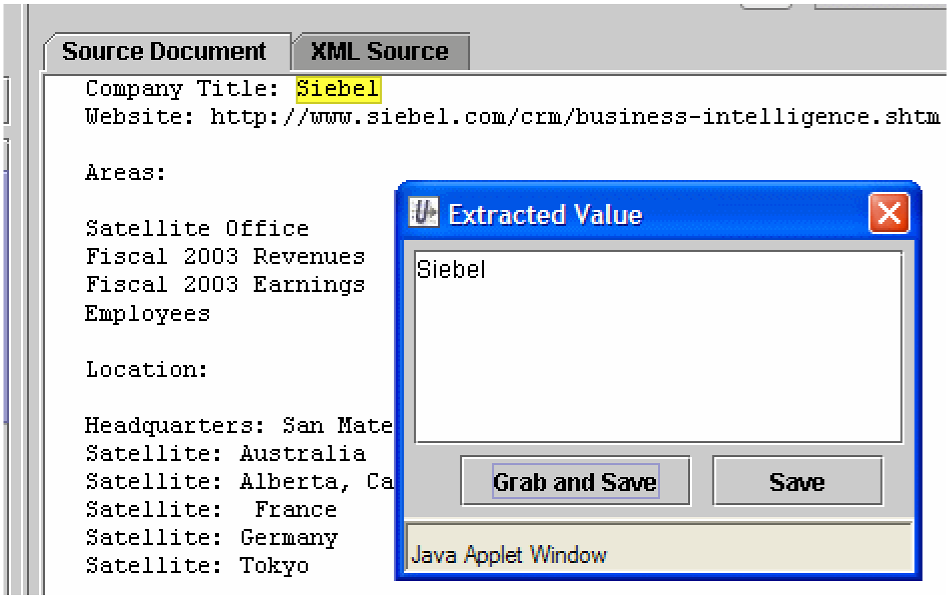
File Name	Date	Size
/test/Siebel.txt	Wed Mar 23 14:55:41 PST 2005	18kb

2. From the left hand side of the Documents Browser, open a folder and select one or more documents from the bin.

   Tips:
   - You can select multiple documents by holding down the Ctrl key and clicking on all the documents you want in the bin.
   - You can add documents to the list by double-clicking on a file.
   - You can open documents in other directories by using the directory folders on the left side of the Document Browser window.
3. Click Add to List. The selected documents will be placed in the Current File List.
4. To see the documents on the list, click the drop-down button on the Current File List.
5. To remove a document from the list, select the file from the drop-down menu and click Remove from List.
6. To remove all the selected documents from the list, click Clear All.
7. To begin working with your selected documents, click OK.

Extracted Value

While viewing the Wait Bin and Complete Bin documents, the Extracted Value windows will display the XML extracted data for each item parsed by UPA.



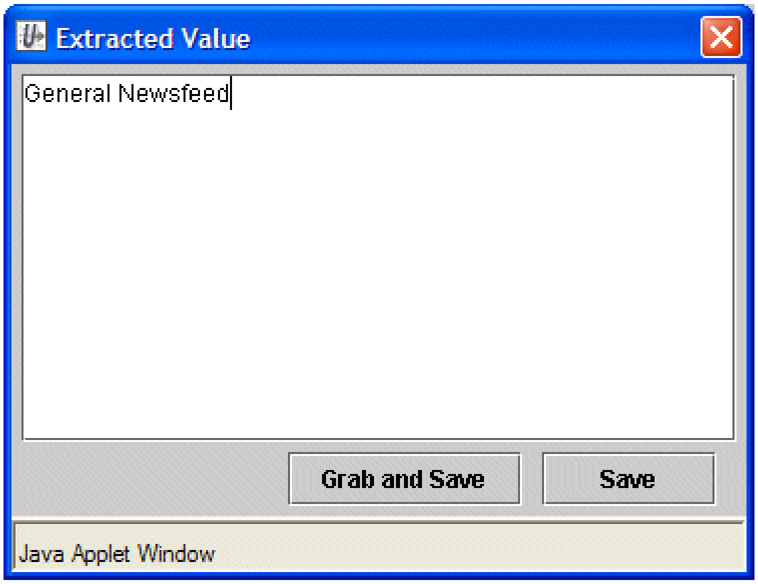
```
Company Title: Siebel
Website: http://www.siebel.com/crm/business-intelligence.shtm

Areas:

Satellite Office
Fiscal 2003 Revenues
Fiscal 2003 Earnings
Employees

Location:

Headquarters: San Mate
Satellite: Australia
Satellite: Alberta, Ca
Satellite:  France
Satellite: Germany
Satellite: Tokyo
```

Siebel

In the Complete Bin you can only view the XML extracted data, while in the Wait Bin you can use the Extracted Value window to add data from the document to the XML data.

When working with the document in the Wait Bin, you can use the Save Extracted Value window to update the currently selected field in the XML Result tree. In this way you can add data that may be missing. To update the data

1. Select a field from the XML Result tree.
2. Enter the data you want for that field in the Extracted Value dialog and click Save.



General Newsfeed

Or

3. Select text from the document and click Grab and Save. The selected text will be added to the selected field's XML result.

Working with the Wait and Incomplete Bin

UPA enables you to work with documents at various stages of processing. All documents that complete the initial parser processing are moved to the Wait Bin along with the generated XML markup files. Documents that do not process successfully are moved to the Incomplete Bin.

Working with the Wait Bin

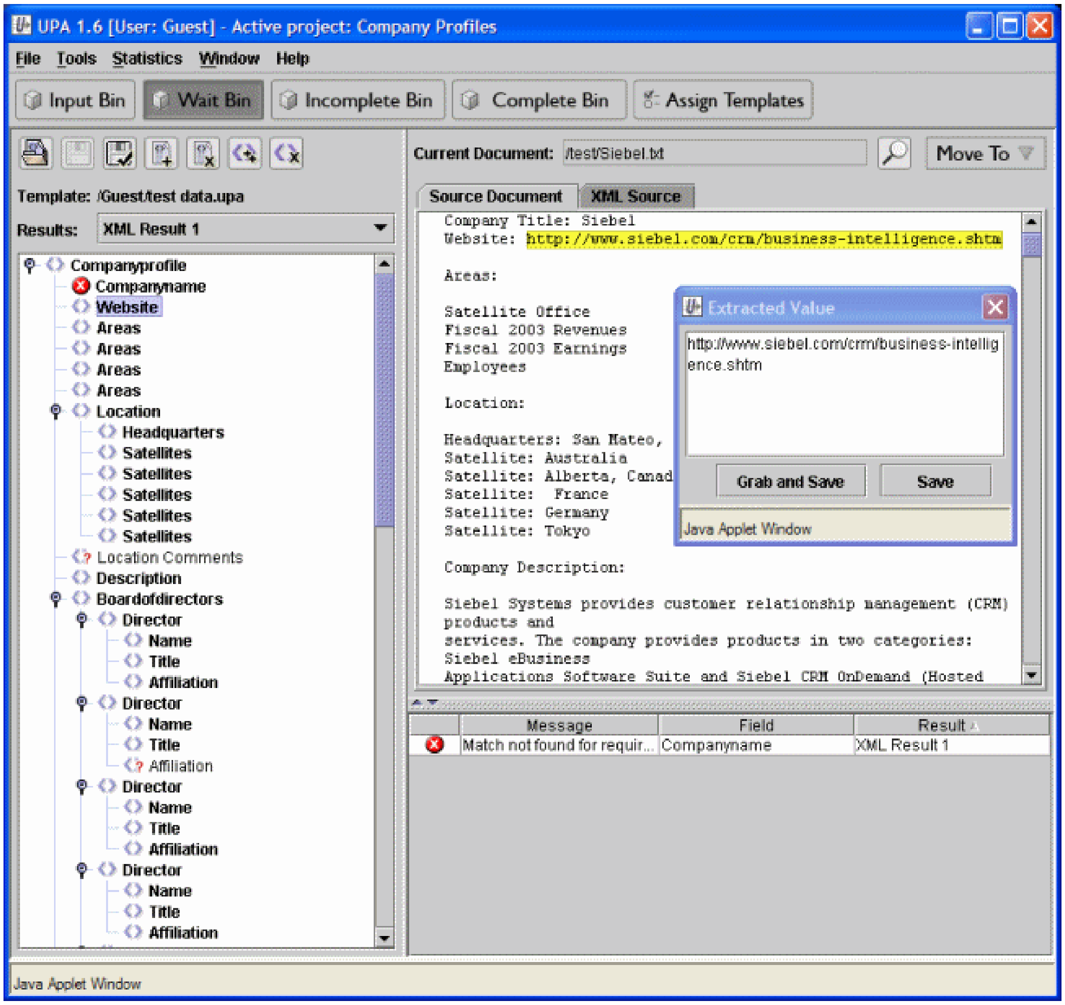
Once the documents are in the Wait Bin, you can perform the following tasks:

- Open a Wait Bin document
- Check and correct the XML for each field
- Save the XML to the database
- Move the document back to the Input Bin
- Move the document to the Incomplete Bin for further work
- Move the document and associated XML to the Complete Bin.

To work with documents that have been processed and are now in the Wait Bin, perform the following procedure.

1. From the UPA main window, click on Wait Bin

, and select a document from the Wait Documents Browser. The selected document will display in the Current Document area. The initial view will be of the document Source Text tab.

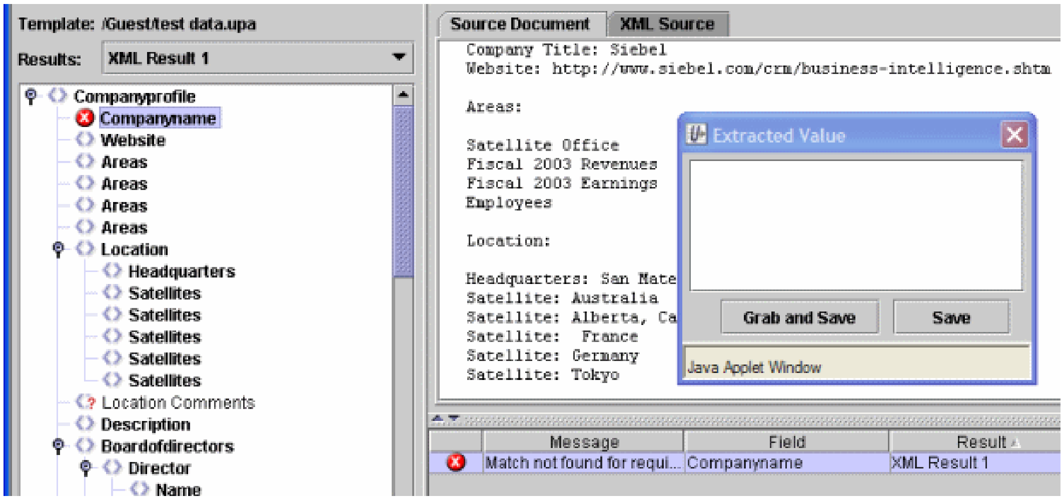


```
Company Title: Siebel
Website: http://www.siebel.com/crm/business-intelligence.shtm

Areas:

Satellite Office
Fiscal 2003 Revenues
Fiscal 2003 Earnings
Employees

Location:

Headquarters: San Mateo,
Satellite: Australia
Satellite: Alberta, Canad
Satellite:  France
Satellite: Germany
Satellite: Tokyo

Company Description:

Siebel Systems provides customer relationship management (CRM)
products and
services. The company provides products in two categories:
Siebel eBusiness
Applications Software Suite and Siebel CRM OnDemand (Hosted
```

Extracted Value: http://www.siebel.com/crm/business-intelligence.shtm

	Message	Field	Result
⊗	Match not found for requir...	Companyname	XML Result 1

2. The icons in the Results list will reflect the status of the parsed data for each field. The data problem icon ⊗ indicates that while data was found and saved by the parser, it did not necessarily meet all the criteria defined in the template. Check the XML fields by clicking on the field in the Results list and confirm that the information stored by UPA in XML format is correct. The XML fields will display in the Extracted Value window, and a message will display in the Message area.


Template: /Guest/test data.upa

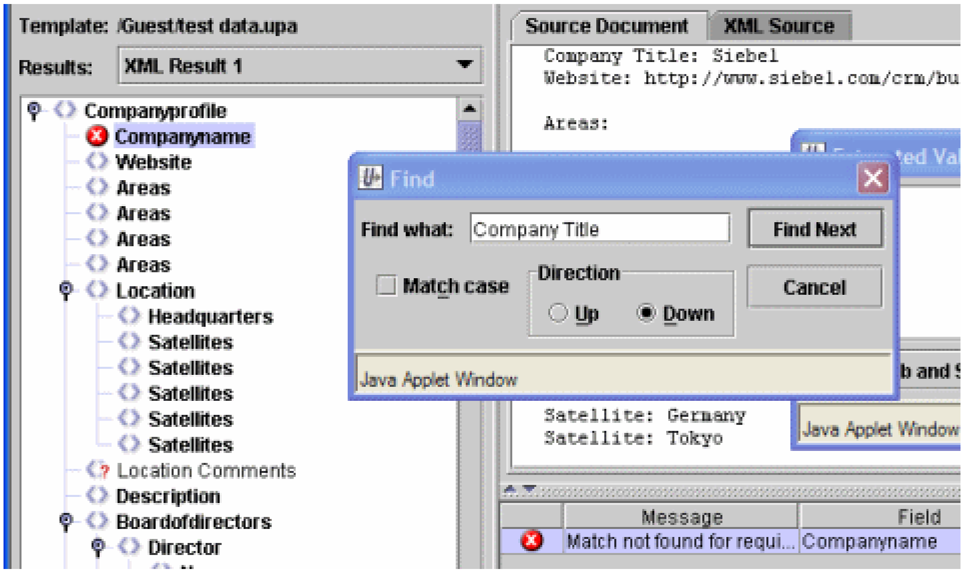
Results: XML Result 1

```
Company Title: Siebel
Website: http://www.siebel.com/crm/business-intelligence.shtm

Areas:

Satellite Office
Fiscal 2003 Revenues
Fiscal 2003 Earnings
Employees

Location:

Headquarters: San Mate
Satellite: Australia
Satellite: Alberta, Ca
Satellite:  France
Satellite: Germany
Satellite: Tokyo
```

Message	Field	Result
Match not found for requi...	Companyname	XML Result 1

3. If the stored XML result value is wrong or missing you can correct the XML if necessary. To correct the XML
   A. Select an XML field from the Results list. The Extracted Value window will either be blank, or will show the value you want to correct.

   B. To find a value in the XML window, click Find In Document . The Find window will display.


Template: /Guest/test data.upa

Results: XML Result 1

Find what: Company Title

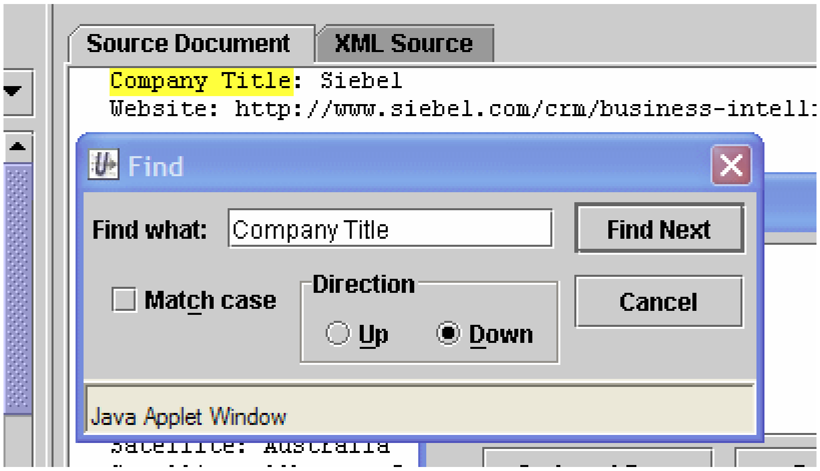
□ Match case

Direction: ○ Up ● Down

```
Company Title: Siebel
Website: http://www.siebel.com/crm/bu

Areas:

Satellite: Germany
Satellite: Tokyo
```

Message	Field
Match not found for requi...	Companyname

   C. Enter a word or any other value you want to search for and click Find Next. Any occurrences of the word or phrase will display in highlights.



```
Company Title: Siebel
Website: http://www.siebel.com/crm/business-intell:
Satellite: Australia
```

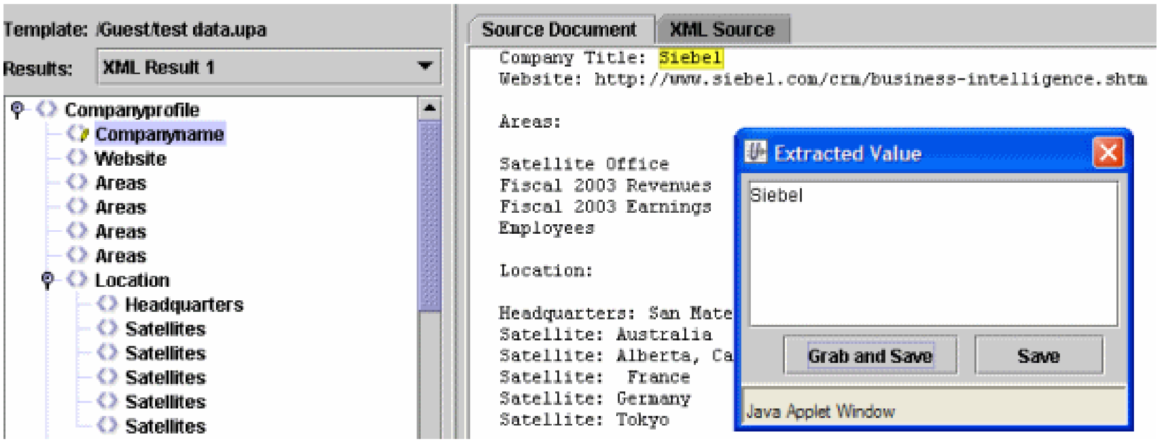
Find what: Company Title

☐ Match case

Direction: ○ Up ● Down

D. In the Extracted Value window, type in the value you want in the selected field and click Save.

E. Or select the missing value from the document by highlighting it with the mouse, and click Grab Value in the Extracted Value window. Your entries will display when you select the field from the Results list. The results field will highlight in bold and a hand-entered icon will indicate you corrected the entry for that field.



Template: /Guest/test data.upa

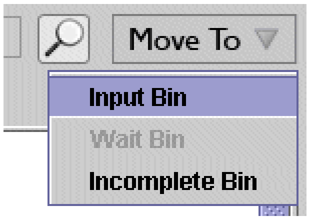
Results: XML Result 1

- Companyprofile
  - Companyname
  - Website
  - Areas
  - Areas
  - Areas
  - Areas
  - Location
    - Headquarters
    - Satellites
    - Satellites
    - Satellites
    - Satellites
    - Satellites

```
Company Title: Siebel
Website: http://www.siebel.com/crm/business-intelligence.shtm

Areas:

Satellite Office
Fiscal 2003 Revenues
Fiscal 2003 Earnings
Employees

Location:

Headquarters: San Mate
Satellite: Australia
Satellite: Alberta, Ca
Satellite:  France
Satellite: Germany
Satellite: Tokyo
```

Extracted Value: Siebel

4. Once you have reviewed the XML fields, you can save the document to one of two areas:
   - Input Bin - for reprocessing, perhaps with a different template
   - Incomplete Bin - for further correction or for creation of a new template
   - Complete Bin - no further work required, XML and document stored to the database archive.
5. To move the document back to the Input area, select Input Bin from the Move To drop-down list.



6. To move the document to the Incomplete area, select Incomplete Bin from the Move Document To drop-down list.
7. To move the document to the Complete area, click Save XML Results to Complete 💾✓.

As you work on the document, the icons displayed in the template area on the right hand side of the UPA main window will change.

Wait Bin Toolbar

The Wait Bin toolbar displays the following buttons which you will use to work with the document:

Button	What it Does
📂	Opens an existing file in the Wait Bin.
💾	Save the current XML results and any changes to the template.
💾✓	Save the XML results to the Complete Bin.
📄+	Create a new XML result document.
📄✕	Remove an XML result document.
<+	Duplicate the selected XML field.
<✕	Remove the selected XML field.
🔍	Find a word, phrase, or characters in the document.

Working with the Incomplete Bin

UPA enables you to work with documents at various stages of processing. All documents that complete the initial parser processing are moved to the Wait Bin or Complete Bin along with the generated XML markup files. Documents that do not process successfully are moved to the Incomplete Bin.

Note: Failed documents are typically placed in the Incomplete Bin, but you can set an option in the Assign Template window that will put failed documents in the Wait Bin.

The Incomplete Bin contains all files where no direct matches were found in any available templates. You can view these files and modify or create new templates to reprocess any failed files.

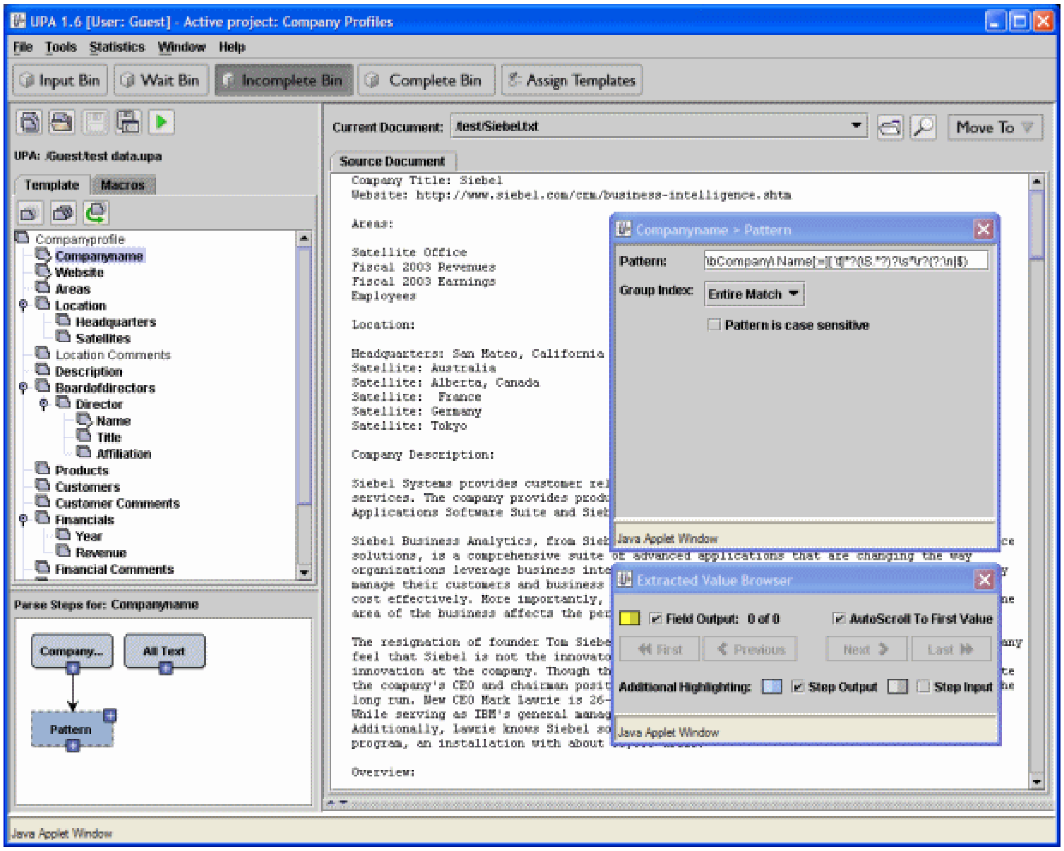
Once the documents are in the Incomplete Bin, you can perform the following tasks:

- Open Incomplete Bin documents.
- Update the template to work with the failed documents.
- Create a new template to work with the failed documents.
- Move the document back to the Input Bin for processing with new or updated templates.
- Parse the document with the new or updated template.
- Move the parsed document to the Wait Bin for further work on the XML.

To work with documents that have failed to complete processing and are now in the Incomplete Bin, perform the following procedure.

1. From the UPA main window, click on Incomplete Bin

, and select a document from the Failed Documents Browser. The selected document will display in the Source Document area. The defined elements of the template will display under the Template tab.



UPA: /Guest/test data.upa

Current Document: /test/Siebel.txt

Parse Steps for: Companyname

```
Company Title: Siebel
Website: http://www.siebel.com/crm/business-intelligence.shtm

Areas:

Satellite Office
Fiscal 2003 Revenues
Fiscal 2003 Earnings
Employees

Location:

Headquarters: San Mateo, California
Satellite: Australia
Satellite: Alberta, Canada
Satellite:  France
Satellite: Germany
Satellite: Tokyo

Company Description:

Siebel Systems provides customer rel
services. The company provides prod
Applications Software Suite and Sieb

Siebel Business Analytics, from Sieb
solutions, is a comprehensive suite of advanced applications that are changing the way
organizations leverage business inte
manage their customers and business
cost effectively. More importantly,
area of the business affects the per

The resignation of founder Tom Siebe
feel that Siebel is not the innovato
innovation at the company. Though th
the company's CEO and chairman posit
long run. New CEO Mark Lawrie is 26-
While serving as IBM's general manag
Additionally, Lawrie knows Siebel so
program, an installation with about

Overview:
```

Companyname > Pattern

Pattern: \bCompany\ Name[:=]['"]{"?(\S.*?)?\s*\r?(?:\n|$)

Group Index: Entire Match

☐ Pattern is case sensitive

Extracted Value Browser

☑ Field Output: 0 of 0 ☑ AutoScroll To First Value

Additional Highlighting: ☑ Step Output ☐ Step Input

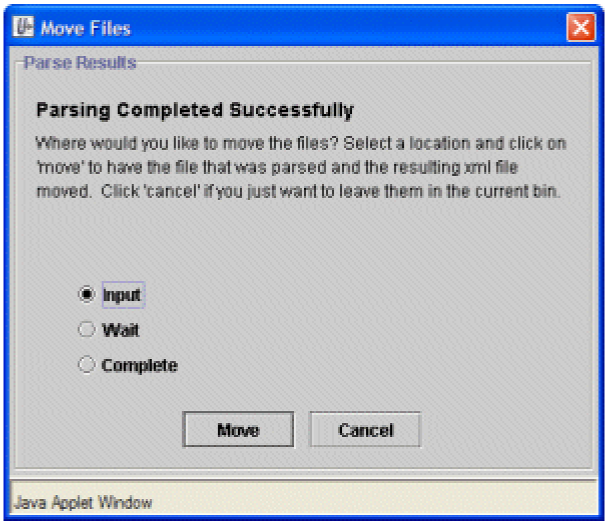
2. Edit and adjust the template fields by following the same procedures you use to create Input templates.

3. Click Save Template As 🖫 to save the template to a new file name.

   Click Save 🖫 to save your changes to the current template and overwrite the old version.

4. To test the updated template, click Parse Current File ▶. The UPA parser will use the template to parse the currently open failed document. A message will display confirming parser success or failure.

5. When the parser completes successfully, you will be given the options to move it to either the Input, Wait, or Complete Bins.



Move Files

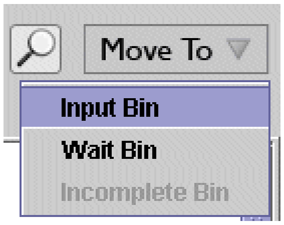
Parse Results

Parsing Completed Successfully

Where would you like to move the files? Select a location and click on 'move' to have the file that was parsed and the resulting xml file moved. Click 'cancel' if you just want to leave them in the current bin.

● Input
○ Wait
○ Complete

6. Select one of the options and click Move. If you do not want to move the parsed document right now, click Cancel.
7. Once you have an updated template, move the document back to the Input area for reprocessing by selecting Input Bin from the Move To drop-down list.



If you have already parsed the document successfully and did not elect to move it right after parsing, move the document to the Wait Bin by selecting it from the Move Document To drop-down list.

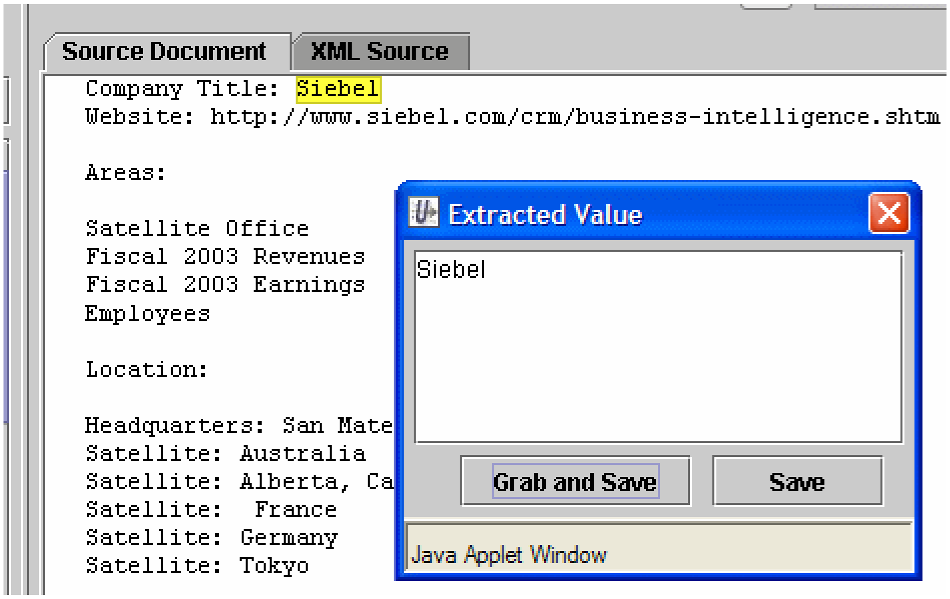
As you work on the document, the icons displayed in the template area on the right hand side of the UPA main window will change.

Incomplete Bin Toolbar

The Incomplete Bin toolbar displays the same buttons as the Input Bin Toolbar.

Extracted Value

While viewing the Wait Bin and Complete Bin documents, the Extracted Value windows will display the XML extracted data for each item parsed by UPA.



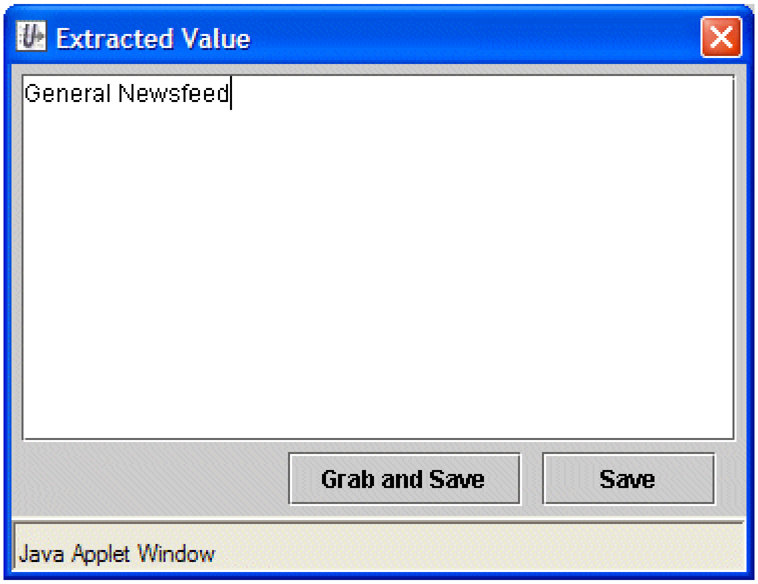
```
Company Title: Siebel
Website: http://www.siebel.com/crm/business-intelligence.shtm

Areas:

Satellite Office
Fiscal 2003 Revenues
Fiscal 2003 Earnings
Employees

Location:

Headquarters: San Mate
Satellite: Australia
Satellite: Alberta, Ca
Satellite:  France
Satellite: Germany
Satellite: Tokyo
```

Siebel

In the Complete Bin you can only view the XML extracted data, while in the Wait Bin you can use the Extracted Value window to add data from the document to the XML data.

When working with the document in the Wait Bin, you can use the Save Extracted Value window to update the currently selected field in the XML Result tree. In this way you can add data that may be missing. To update the data

1. Select a field from the XML Result tree.
2. Enter the data you want for that field in the Extracted Value dialog and click Save.



General Newsfeed

Or

3. Select text from the document and click Grab and Save. The selected text will be added to the selected field's XML result.

Index

Distribution

No. of Copies

Onsite

7 Pacific Northwest National Laboratory

W. E. Cowley	K7-22
N. O. Cramer	K7-22
A. G. Gibson	K7-28
R. T. Scott	K7-28

No. of Copies

M.A. Whiting	K7-22
C. Winters	K7-63
Information Release Office	K1-06