Pacific Northwest NATIONAL LABORATORY

UPA™ Universal Parsing Agent

R&D Magazine has named "Universal Parsing Agent" one of the world's 100 most significant scientific and technical innovations for 2007. This invention accepts multiple datasets or streams of information, discovers and extracts information needed by users, and delivers results in their most useful form. UPA provides a flexible, reliable, and scalable solution to information workers' needs in today's high-volume, dynamic data environments.

U.S. DEPARTMENT OF ENERGY

Universal Parsing Agent— More Time Analyzing, Less Time Sorting

PROBLEMS WITH PREPARATION

Information workers using available data extraction tools often spend more time cleaning, sorting, and reformatting data in preparation for analysis than analyzing the data itself. Information workers in government, business, and academia struggle daily with formatting and structuring data for applications. Data often arrives in massive amounts from different sources in various formats, further compounding the problem.

Until now, data extraction tools have been created for a single purpose. When data formats change, users cannot make modifications to the extraction tool. They must rely entirely on programmers to make the necessary changes. Even then, attempts to revise the software can become complicated and cost-prohibitive.

UPA OFFERS AN INNOVATIVE AND COST-EFFECTIVE ALTERNATIVE

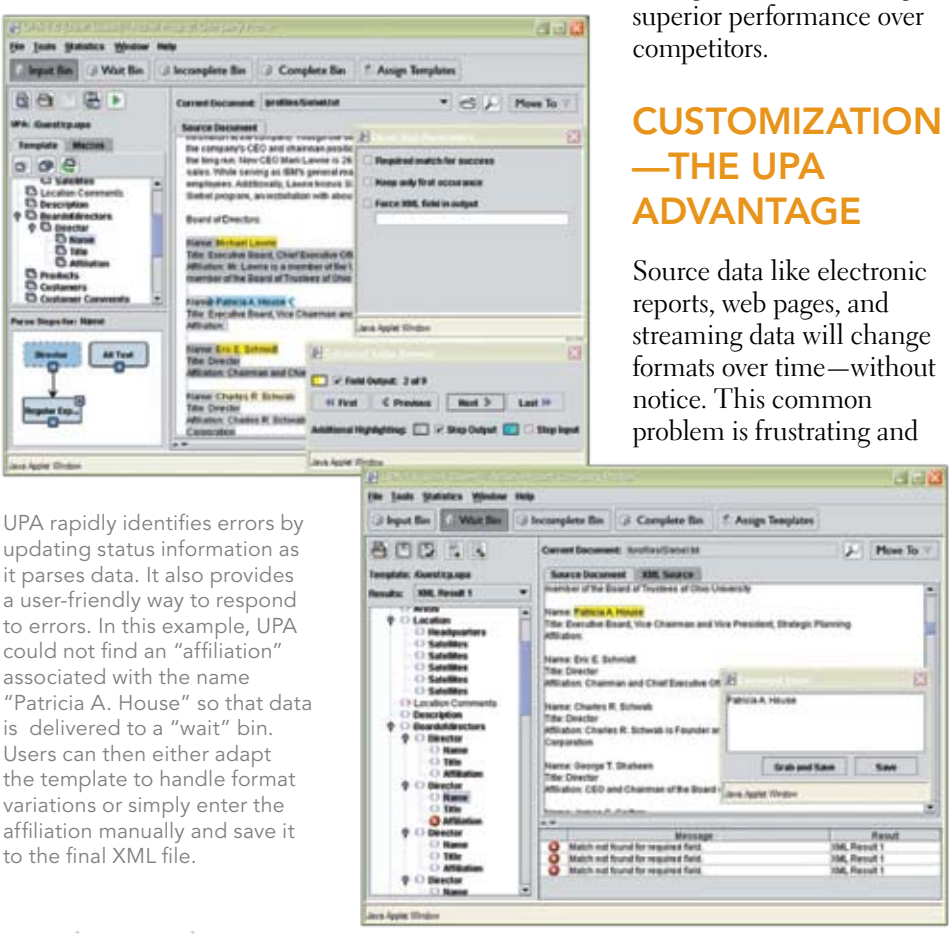
UPA breaks this cycle. UPA empowers users to specify the data they need, instead of relying on programmers. Users tell UPA the exact information they want from each data source by creating a flexible and reusable template that describes the information they want. UPA templates are the software's directions for what data to find, extract, change, enhance, and output.

Templates are like the instructions a researcher would give to a research assistant; they describe what is needed, how it should be presented, "must-haves," alternatives, and even "would-likes." UPA is the research assistant that knows how to sift through the information to find what is needed.

MEANINGFUL RESULTS IN A USABLE FORMAT

UPA outputs information in XML format, which is easily usable by most popular database programs, including Microsoft® Access®, Microsoft® Excel®, and Oracle®. UPA also provides an ideal front-end information extraction and transformation tool for Pacific Northwest National Laboratory's award winning information visualization tools, IN-SPIRE™ and Starlight™. UPA is currently used by information analysts to streamline data flow into these tools by:

- *quickly defining the information elements to draw from raw data feeds and*
- *inserting meaningful tags in the data to expand the usefulness of their programs' analysis results.*



UPA rapidly identifies errors by updating status information as it parses data. It also provides a user-friendly way to respond to errors. In this example, UPA could not find an "affiliation" associated with the name "Patricia A. House" so that data is delivered to a "wait" bin. Users can then either adapt the template to handle format variations or simply enter the affiliation manually and save it to the final XML file.

DIVERSE APPLICATIONS

UPA's capabilities shine in many areas. Businesses that must keep abreast of the competition can use UPA to support their competitive intelligence needs and to create knowledge bases that reduce uncertainty and risk in decision-making.

For example, competitive intelligence analysts in the pharmaceutical industry could create a knowledge base of researchers in their industry who may be supporting other companies. Using conference proceedings as a data source, UPA can parse the proceedings, extracting author names and references from the bibliographies. Once UPA extracts a name, a link is provided to more research papers, more authors, and more information about research in the industry. Harvesting this linked information and extracting names, affiliations, and research areas provides a valuable resource for those developing strategies for maintaining superior performance over competitors.

CUSTOMIZATION—THE UPA ADVANTAGE

Source data like electronic reports, web pages, and streaming data will change formats over time—without notice. This common problem is frustrating and time consuming. Every information consumer knows what happens next—custom programs must be modified, recompiled, debugged, verified, tested, and modified again. UPA eliminates this problem by allowing users to specify the pieces of information they want from the data sources. Source data can then change formats and no changes are equired to UPA. UPA may be used to:

- *create unified data formats for research*
- *prepare historical data for the Internet (for example, converting data from ASCII to XML for web presentations)*
- *parse email survey responses*
- *parse third-party online calendars.*

ABOUT PNNL

Pacific Northwest National Laboratory is a Department of Energy Office of Science national laboratory where interdisciplinary teams advance science and technology and deliver solutions to America's most intractable problems in energy, national security, and the environment. PNNL employs 4,000 staff, has a $855 million annual budget, and has been managed by Ohio-based Battelle since the Lab's inception in 1965.

If you are interested in collaborating with us contact:

Alex Gibson
(509) 372-6086
alex.gibson@pnl.gov
or
Nick Cramer
(509) 375-4728
nick.cramer@pnl.gov
at
Pacific Northwest National Laboratory
P.O. Box 999
Richland, WA 99352
www.pnl.gov

Pacific Northwest
NATIONAL LABORATORY

PNNL-SA-44146