Information Services & Use 24 (2004) 121–129
IOS Press

Aggregating the aggregators – An agnostic approach

Rolf Walter and Leif Bjorking
Just-In-Mind AB, Tegnérlunden 3, 111 61 Stockholm, Sweden
Tel.: +46 8 545 940 50; E-mail: rolf@just-in-mind.se; Web sites: www.just-in-mind.com, www.webalerts.net

Abstract. A data warehouse platform for automatic post-processing and post-aggregation of online search results and alerts from online vendors was developed, called Web Alerts. Contents from many information providers are being repackaged "agnostically". The solution, built on the Lotus Notes & Domino groupware, is automatically receiving, parsing, aggregating and displaying a mix of records from Dialog, DataStar, Factiva, LexisNexis, CSA, STN and other vendors in a common web-based solution. This paper describes the design and architecture of the solution, how records are processed and aggregated, regardless of database or online host, and presented in a uniform format with respect to key fields such as Title, Author, Source, Publication Date, Company Name, etc., and the further interactivity offered in the database once the records are aggregated. The web solution may be used, e.g., in business intelligence applications and in corporate Intranets as a current awareness tool for sharing external as well as internal information, offering categorization of records for different organizational units, and personal customisation by creating subsets of records via a search filter. Other features like adding comments, flags, creating own documents, highlighting of original search words for easy perusal, and linking to web sites and patents, have also been added to the solution.

1. Introduction

Lack of internal and external data integration, information residing in "silos" inside companies, federative searching, and many other themes address the many deficiencies in which internal and external information is handled and stored in companies and organizations.

In our company, a need for an automatic post-processing combined with post-aggregation of online search results and alerts was identified 2.5 years ago. In order to harmonize the various record formats from online vendors, and also for different databases within a single vendor, it was decided to design a data warehouse platform. The Lotus Notes & Domino groupware was judged as the most suitable software for this purpose.

Information specialists are often "agnostic" as to the choice of a vendor when searching information. Vendors have different profiles and strengths and weaknesses. The very same article may be indexed with different fields and taxonomies, giving skilled users the opportunities of shopping around for the best vendor in their view.

The traditional way of receiving records from online vendors (Dialog, DataStar, Factiva, LexisNexis, CSA, STN, etc.) has been, apart from the labour-intensive work of copying and pasting from the screen, or exporting to a text file, asking for the records to be sent via an e-mail, or picked it up via an FTP server. This traditional method still requires manual editing of the records.

The web-based solution presented in this paper is an automatic post-processing and post-aggregation platform for information in the fields of business & news, science & technology, patents, trademarks, etc.

For the most recent information on the solution, as it is continuously being upgraded with new contents and features, please cf. to the web site www.webalerts.net.

2. Software used

The solution is programmed in the Lotus Notes & Domino groupware. The first version was made 2.5 years ago in version R5, but to benefit from all new features being added to the Lotus Notes & Domino groupware, upgrading has been carried out on a regular basis. The current version of Lotus Notes & Domino used is 6.5.1. The programming languages used were LotusScript, Java, Notes @Formulas, XML, HTML, JavaScript, CSS (Cascading Style Sheets), etc. Scheduled agents are used for automatizing the processing of records, sending E-mail Summaries to users, and supervising the whole operation of the platform. The users access the information via a web browser, preferably Internet Explorer 5.5 or higher.

3. General outline of the solution – user's perspective

Online searches are carried out, or alerts being set up, with the respective information provider/host. E-mails with records are delivered to, and processed and stored, in a Notes database, in principle every customer having its own database. The user, or the administrator of the platform, only need to order the e-mails to be sent to a dedicated e-mail address for the service, and to abide by some simple rules, such as selecting a record format (tagged or plain text, depending on vendor) and creating the Subject-line of the e-mail to be delivered according to a predefined syntax (to include the Target Database name, and, optionally, a user-defined Category, described elsewhere in this paper).

E-mail delivery was identified as the common denominator, as this delivery mode is presently offered by all vendors. The e-mail's Body contents is automatically being split into individual records (documents) behind-the-scenes by the parser code. Records are displayed to the end-user in a uniform format with Labels, Headings, Tables, etc. in a data warehouse. The end-users may access the information via a Web browser in a user-friendly record format, interface and navigation, with much built-in interactivity.

4. General outline of the solution – hosting and running

A Domino server hosts a Parser Database, an Overview Database and the various Target Databases. The administration of the platform is carried out via a Notes client. E-mails with online search results or alerts arriving to the Parsing Database are parsed via scheduled Notes Agents on a 24/7/365 basis and distributed to Target Databases, every customer having its own database. The Target Database contains a Configuration Document where the administrator of the platform enters customer data and all settings such as version, specific Help information, URL mapping and definitions of record flags. The access rights of the different users are set in the ACL (Access Control List) of the Target Database.

The platform contains templates to handle the different formats in which the online vendors deliver their records. The following set of generic fields are included in every record in the solution: Title, Author, Source, Publication Date, Language, Company Name, Geographic Name and Database Name (the name of the database, e.g., BIOSIS, Derwent WPI, etc.). For patent records, Patent Assignee is treated as Company, and Inventor as Author. With the Arrived Date field, i.e. the date the record arrived

to the database, the records may be sorted, e.g., in reverse chronological order for current awareness purposes. Additionally, a user-created Category generic field may be included, by adding a text string to the Subject line when ordering the online search results or alerts. With the Category field, records may be earmarked according to an organizational unit, a project title, etc., and be used also in the Search Filter to be described later in this paper.

Apart from the "public" generic fields described above, several fields, some of them hidden to the end-user, are added to the records in order to facilitate the tracking of records and the supervision of the operation of the platform: Host Name, Target Database, UserID, DocID (accession number in the host's database, used for duplicate checking), QueryID (or AlertID), Record Format (as defined by vendor) and Database Update (supplied by vendor for alerts).

In an Overview Database, the whole operation of the platform may be supervised. The Overview Database gathers and displays information for every Target Database on the Domino server, such as, e.g., customer data, size of database, number of records stored, last arrived record and when (host, database), traffic statistics, and the access rights for different users.

5. Design goals and challenges

One big challenge in the design of the solution has been to meet the goal of presenting large amount of well-formatted data to the users without long delays.

Since each one of the platform's Target Databases contains thousands of records, and one single record may contain more than 1000 lines of text, a large amount of data needs to be delivered to the users in an efficient way. One primary design goal has been to not keep the user waiting – when a user requests to view a record in the browser the user should be able to see even the largest record directly without delay.

Our approach to meet this challenge has been to design a pre-processor that prepares the records in advance. The steps taken by the pre-processor are:

(1) Assembling all necessary data needed to present the record;
(2) Formatting the data (building Tables, adding URL links, CSS, etc.);
(3) Adding JavaScript code for handling client-based user interactions;
(4) Converting the result of the above into an HTML page;
(5) Storing the HTML representation of the record.

Since the pre-processing is done (and invisibly to the users) at the moment a new record arrives into the Target Database, the result of this approach is that when a user requests a record, the record has already been pre-processed and the only action needed to be performed at the moment of the request is to send the HTML page directly to the browser.

Pros are the great time-savings achieved by eliminating steps (1)–(4) above, resulting in a better user experience. Cons are that extra storage space is needed for storing the HTML representation of the record.

6. Online vendors covered at present

The online vendors covered so far are treated in some detail below.

6.1. Dialog

The leading online vendor, Dialog (accessed via DialogWeb or DialogLink), offers an extensive range of databases in essentially every field. Dialog's tagged format, however old-fashioned it may seem, lends itself perfectly to post-processing. The multivalue delimiter (^) for e.g. Authors and the bar "|" symbol marking the end of a tag, were very helpful in programming the parsing code. The multivalue delimiter (^) is also included in e.g. Tables in patent databases for separating columns, so that nice-looking Tables may be generated. A drawback with the Dialog formats is that the tag names are not always consistent across databases, e.g. the Author tag may appear in different shapes depending on database, such as AU, AU_AUTHOR, AU_AU1, AU_AU2, AU_INVENTORS, etc. All these variants need to be merged into the single generic Author field. The same goes for Source, Publication Date, Company Name, and so on. Due to these inconsistencies, all Dialog databases need separate Templates. Further, all tags are not documented in Dialog Bluesheets. The parsing code will therefore alert the administrator when new tags pop up, so that they may be added to the Template. Hence, no tag, however esoteric it may seem, is ignored.

Dialog does include images for both patents and trademarks. Until recently, images could only be delivered via alerts, but now they are also included when records are delivered via online searches. Records containing images are marked with an image icon in all views.

At present, Templates have been prepared for around 40 of Dialog's databases, covering News, SciTech, Healthcare & Medicine, Patents and Trademarks.

6.2. DataStar

Although nowadays part of The Dialog Corporation, DataStar (accessed via DataStar Web) has an entirely different query language as well as delivery format and record format. The record formats for the different DataStar databases are much more homogeneous than is the case for Dialog. Still, we found it convenient to group the DataStar databases in a couple of Template categories having many tags and features in common: News (many databases from Agence France-Presse Newswires, Italian General News Service and BBC International Reports), Gale Group (Business & Industry, PROMT, Health Periodicals, and half a dozen others) and Adis (Adis INPHARMA, Adis R&D Insight, and a few others). At present, Templates have been prepared for around 25 of DataStar's databases.

DataStar has a tagged format, but not as precise as Dialog. No special delimiters are included in the tags, apart from semicolons for multivalues in e.g. Author tags. As a result, Tables cannot be created, instead, they will have to be preserved as-is in the tags.

6.3. Cambridge Scientific Abstracts (CSA)

CSA has a strong position in the SciTech area offering bibliographic databases in a consistent format. Some of their databases are also available via STN. One single Template is sufficient to cover all of CSA's databases. CSA is offering some attractive features such as bundling databases in the searches, e.g., METADEX is searched in parallel together with "Recent References Related to Materials" and "Web Resources Related to Materials & Engineering", the latter covering thousands of high-quality web sites that are hand-picked and indexed by CSA's editors. The records from these companion databases are also processed and aggregated in the present solution.

CSA's has no special tagged format, instead, their standard format resembles DataStar's tagged format. Parsing CSA's records and transforming them to the common format of this solution does not pose a big challenge. The single Template prepared covers all of CSA's databases.

6.4. Factiva

The service Factiva.com of Factiva (Dow Jones & Reuters) is covering 9,000 sources in 22 languages. Searches may be carried out Online via a web interface, and Alerts (or Tracks, in Factiva's terminology) being set up. For Alerts (Tracks), Factiva.com is also scanning web sites with news contents, apart from newspapers and journals.

Parsing the records of Factiva.com is fairly straightforward, as Factiva.com consists of one single database only, with a total of no more than 30 tags. Some tags, like the RE (region) need to be cleaned up to remove abbreviations and leave only the plain text, e.g. from text strings as "usva: United States – Virginia" or "nz: New Zealand". The objective being to create a as homogenous as possible nomenclature among the aggregated records.

7. Other online vendors and services

The design and architecture of the platform is built in such a generic way that records from any source may be parsed and stored in the data warehouse, as long as they are well-structured and consistent in format.

Other vendors to be included in the future are LexisNexis, STN and Questel Orbit. Later on, major national vendors will also be considered, e.g. Genios which covers many German language sources from Germany, Austria and Switzerland.

Other interesting options to be considered include free alerting services offered by some publishers. The users may access the full text articles if they are subscribers, or purchase the articles online. The results of these alerts could easily be integrated into the solution presented in this paper.

8. Common features for records

8.1. Outline of records

As incoming records are transformed to a common format by the parsing code, the contents may be freely rearranged. Images, e.g., of trademarks and molecules, may be placed at the top of the record, under the Title. Some data of minor importance to the ordinary user such as ISSN, ISBN, Publisher, and many, many others, are assembled at the bottom of records under the heading "Bibliographic Data". In the Footer of each record are shown data such as Host, Database Name and Copyright information. No data from the incoming records are deleted or hidden, but they do are transformed and rearranged with homogenisation and userfriendliness in mind. For record samples, please cf. the information that may be downloaded from www.webalerts.net.

8.2. Cascading Style Sheets (CSS)

The solution contains a total of 15 different CSS classes for Fields, Labels, Headings and Table elements, giving the option to give a look-and-feel to the records and the solution according to a corporation's visual identity guidelines.

8.3. Publication Date

Not all records from the vendors contain a PD-tag (Publication Date). Whenever possible, a Publication Date is extracted programmatically from another tag, normally the SO-tag (Source). The SO-tag may contain, e.g., the string "Mar-Apr 2003", which is transformed to a Publication Date of 2003-03-01. Other strings that may be interpreted in a similar fashion are transformed to a Publication Date.

8.4. Links to web sites, e-mail and patents

Links to web sites, patents at USPTO and EPO (esp@cenet) and e-mail addresses are automatically inserted into the records.

Embedded text strings anywhere in the records that may be interpreted as a URL or an e-mail address are transformed to links opening the user's web browser, or the e-mail client, respectively.

For patent databases, or databases containing patent information (e.g., RAPRA, PIRA and Foodline), the patent numbers are analysed and links created to the respective patent document in USPTO for US patents and to EPO (esp@cenet) for all other countries.

8.5. Highlighting of original search words

Dialog, DataStar and CSA deliver the online search results or alerts along with search strategies. For Dialog this is the default case, but the delivery of search terms may be turned off via the PRO command (Print Results Only). DataStar gives the user the option of actively selecting the search terms to be attached to the delivery. For CSA, the search words are always included. Factiva.com does not deliver the search terms at present.

Highlighting of original search words (if available, and if selected, via arriving online searches or alerts) is carried out automatically. Additionally, and in an additive fashion, other words may be chosen manually for highlighting in the records, either in a global way via the Configuration Document in the Target Database, or individually in the Parser Document in the Parser Database.

To distinguish the highlighting of original search words (or strings) from the Search Filter highlighting (to be described later), they are highlighted in the records with red letters on a green background. The Search Filter highlights with a yellow background. Highlighting of original search words is a feature that the user may turn off via the Tools menu in the web interface of the solution. Highlighting of original search words is working also for Boolean searching, including with parentheses.

8.6. Adding comments and flags to records

Users, if assigned this right, may add comments to a record for others to view, or adding comments themselves. Records having comments attached to it are marked with an icon on the record as well as in all views of the solution's web interface.

A record may also be attached with flags with meanings defined by the administrator in the Configuration Document of the Target Database. A maximum of eight flags in different colours may be used. The flags are shown on the record itself, and in all views of the solution.

8.7. Information about the source database

The experienced user may want to learn more about the characteristics of the vendor database from which a record originates. To that end, at the bottom of the records are located buttons for directly accessing Dialog Bluesheets, DataStar Datasheets and CSA's Factsheets for the respective database.

9. Global features of the solution

9.1. Views for all generic fields

The records may be sorted and viewed in ten different views, by: Arrived Date (default), Source, Company Name, Author, Geographic Name, Language, Title, Publication Date, Database Name and Category. The last field, Category, may be introduced by the customer for internal categorization of records, e.g. with an organizational unit, a project title, etc. Navigation in the views is offered both via a select box and via browsing pages. The last chosen view, and position within a view, is memorized in a cookie.

Records may be viewed in a preview pane, or in a maximized own window. Unvisited records are marked in bold, and visited ones as "read" with plain text. The views include multiple selection of records with check boxes for printing, exporting to Word, setting flags and deletions (administrators only).

Again, for more details please cf. to www.webalerts.net.

9.2. Search Filter

The solution offers a Search Filter for searching by free text, and/or by field(s), as an advanced option. Moreover, a user may have the Search Filter permanently turned "on", to show only a subset of the records, as a personal customisation. Search words are highlighted with a yellow background for easy perusal. The search criteria applied is stored in a cookie, and being memorized the next time the user opens the database. Highlighting is working also for Boolean searching, including with parentheses.

9.3. E-mail Summary

Configurable e-mail notifications are offered as a "push" option. The user does not need to actively check the database for new entries. By subscribing, E-mail Summaries are received continuously as new records arrive, or with a set interval in days. Any number of users in an organization may subscribe for E-mail Summaries, which includes a table with the Titles of all new records, along with links to the individual records, cf. below. For the Editor-in-Chief (to be described later), unpublished records may optionally be included in the E-mail Summary received.

The Email Summaries being sent are formatted and coded by a Java agent, to take advantage of the mail protocols "Multipart/Alternative" MIME-type functionality. This approach makes it possible to send both an HTML and a plain text version of the contents of the e-mail, and letting the functionalities of the receiver's e-mail client decide, without any user intervention, which version to display. This way, receivers with e-mail clients supporting HTML will enjoy a better formatted version with direct links to the records.

9.4. Creating own records with internal information

Users may create their own records with internal information, such as memos and reports. By using the same set of generic fields that are used for the external records, information from whichever origin are being created in the same "mould", and displayed and made searchable in a uniform format.

9.5. Access rights and roles

Depending on the rights assigned to different user categories in the ACL (Access Control List) of the Target Database, users may read or write comments, add/remove flags, add own records, and delete records.

A special and privileged position may be given to an Editor-in-Chief, who acts as a "human filter" for Publish/Hide decisions of arriving records. An Editor-in-Chief may be appointed for e.g. corporate Intranets in current awareness settings. The Editor-in-Chief will see all the records as they arrive, the default being to hide new records, and also be alerted via E-mail Summaries, as mentioned above. Unpublished/hidden records are displayed to the Editor-in-Chief in red colour. The Editor-in-Chief may decide to publish a record, and maybe add flags and/or comments, or decide to delete the record altogether. If approved for publishing, the record will be displayed in black colour to the Editor-in-Chief, and become visible to the ordinary end-users.

10. Future plans for the platform

The platform is continuously being developed along the two main routes Contents versus Features, as has been the case since the development of the platform started 2.5 years ago. Future enhancements are compiled in a project plan, albeit with frequent priority shifts, depending on signals from the market. Among feature enhancements considered are: adding "My keywords" or "My Descriptor Terms", ordering articles (document delivery), links to full-text sources, saving searches, improved consolidation of geographical terms, statistics analyses of e.g. patent assignees and patent classes, introducing SSL, just to mention a few. On the Contents side, the demands from the market will govern this, but the ambition is to cover in a first stage the "big three" among online vendors, i.e. the missing LexisNexis (Dialog/DataStar and Factiva are already included), and later on adding STN, Questel Orbit, Genios of Germany and others.

11. Brand and marketing aspects

The solution presented in this paper is marketed under the brand "Web Alerts" since November 2002. Further information about the solution, including a downloadable pamphlet, may be found at www.webalerts.net.

The solution is operating on a 24/7/365 basis, automatically and "agnostically" post-processing and post-aggregating information from many information providers. The solution may be used in business intelligence applications and in corporate Intranets as a current awareness tool. Among benefits offered are the time savings – no manual reediting work required – thus freeing staff for more interesting tasks. External and internal information may be shared with colleagues, partners and customers in a user-friendly web environment with much built-in interactivity.

12. Short company and author profile

Just-In-Mind AB combines skills in the information profession and software programming. Just-In-Mind AB is an IBM Business Partner at the Advanced level. The authors of this paper hold

the title of Principal Certified Lotus Professional in Application Development. Rolf Walter is affiliated with the Association of Independent Information Professionals (AIIP). For further information, please cf. the web site www.just-in-mind.com.

Acknowledgements

We are grateful to Dialog, DataStar, CSA and Factiva for generously giving us testing accounts for analysing their record formats. Special thanks to the staff of Dialog's Scandinavian office in Stockholm for giving us valuable feedback from the very beginning of the development of the platform.