Affiliate Datafeed Project – Updated July 9, 2008

PLEASE NOTE: To be considered, your reply <u>must</u> address the specific "requirements" below. I will <u>ignore</u> "boilerplate" responses. Don't ask for "a more detailed specification." Read this whole document!

<u>Project Overview:</u> I am seeking to create a sort of "affiliate content management system" for affiliate datafeeds, to streamline my "affiliate-role" work (http://www.MarkWelch.com/affiliate-role.htm). I am seeking to work with an individual or team who can help me to design and implement this project. Please read this entire document before responding.

My goal is to produce a system that includes some of the functionality currently offered by:
- web services like GoldenCan, PopShops, and DataFeedFile
- software like Datafeed Studio (see a list of similar software
     at http://www.webref.eu/affiliate-marketing-shopping-development-tools.php)
- "merge" software such as WebMerge and Cusimano scripts
but adding greater flexibility and adding "content management" functionality.

The system will need to include:

- a relational database capable of managing 10-15 million product records (probably MS SQL Server or MySQL);

- scripts to import datafeeds from many hundreds of merchants, and to "clean up" datafeeds;

- a user interface to manage and edit content, including selecting products or search queries, writing original articles, re-writing product descriptions;

- tools to auto-generate static content pages (this might be done with WebMerge or other existing tools);

- tools to publish ads through Google's AdWords API.

I do not expect all components to be completed in one phase. However, all phases must be anticipated while doing work.

Requirements:

(1) You and your development team must be able to communicate clearly and effectively in English. You and your development team must be located in an English-speaking country (US/CA/UK/AU/NZ).

*After wasting <u>many</u> hours with incompetent, time-wasting Indian development firms, **<u>I will not consider offshore</u>** development from India, China, or Eastern Europe. I know that there are many competent, ethical software-development firms in India and elsewhere, but <u>I don't have the time</u> to screen out the others, so I cannot consider any.*

(2) I have more experience with **Windows** server environments, and I recognize that I would face an incredible learning curve if the project were implemented in Linux. Accordingly, while I will consider any possible solution, I will absolutely never agree to pay hourly fees for work done on a Linux/LAMP solution.

(3) Your reply must <u>not</u> include Microsoft Word attachments (I don't open potentially-dangerous attachments from strangers). Try using plain text or a PDF file.

(4) You must understand the structure of **affiliate datafeeds**, and be aware of common datafeed problems. If you're not familiar with datafeeds and existing tools to manipulate them, explore these ABestWeb discussions:
   a. http://forum.abestweb.com/showthread.php?goto=newpost&t=106957 (my recent experience)
   b. http://forum.abestweb.com/showthread.php?t=60200 ("bulletproof datafeed sites" thread)
   c. http://forum.abestweb.com/showthread.php?t=93877 ("non-unique product IDs in datafeeds")
   d. http://forum.abestweb.com/showthread.php?t=103423 ("bad characters in datafeed")
   e. http://forum.abestweb.com/showthread.php?t=79080 ("duplicate descriptions")
   f. http://forum.abestweb.com/showthread.php?t=86582 (inserting datafeeds to MySQL)
   g. http://forum.abestweb.com/forumdisplay.php?f=188 (ABW Datafeed Forum)

   If your reply does not identify specific knowledge or experience related to working with affiliate datafeeds, I will delete it without any reply.

(5) Some experience with "**content management systems**" (used to manage the writing, editing, review, and approval process for large amounts of web content from multiple sources) is very helpful.

(6) Experience with the Google **AdWords API** (not other Google APIs) is helpful.

Compensation:

- **Hourly**: I would prefer not to pay someone to "do all the work for me," but instead to work with me so that I fully understand the system and can make changes myself.   For your advice and assistance in helping me to build a system that I understand and can modify, I expect to pay a reasonable <u>hourly</u> rate, commensurate with your experience and skill level.

  I will pay hourly only for help building a Windows-based solution. *Do not respond with a proposal for a LAMP solution which includes any reference to hourly fees.*

- **Other**: If you are interested in "building the system yourself" (possibly even "productizing" this system for resale to others) then we could discuss other financial arrangements (this might include negotiated payment amounts for specific project components, possibly with some revenue-sharing if the system is resold to others). I will not consider any solution that I cannot adapt and modify to meet future needs.

- **Budget**: I do not have a fixed budget for this project. I have already spent several thousand dollars on "failed efforts." If you seek up-front payment of $5,000 or more before doing any work, please don't contact me.

Here is my current overview of the project (which certainly will require a lot of refinement).

Phase One: Database Architecture & Preliminary System Specifications

- Identify tables, fields, relationships, indexing.
- See http://www.markwelch.com/datafeed-project-tables.xls (each table schema is described in a separate tab)

  o There will certainly be a separate "raw datafeed" table used during import, which will later be merged into a separate "products" table. The structure of the "raw datafeed" table will probably be different for each affiliate network or data source (e.g. shareasale_raw_datafeeds, cj_raw_datafeeds, etc.) In theory, the "raw datafeed" tables are temporary – but in practice it's likely that there will be data constantly flowing through them.

  o The "products" table will need to include "imported-and-translated" content – but a number of these fields will be subject to an "override" with alternate text (most often, rewritten product titles and descriptions. The "override" data might be stored in the same table, or possibly a separate table. There will be "override" data only for a very small fraction of all product records (perhaps 1%). In general, most "override" events will be effectively be executed as part of the substitution/translation phase, to reduce the need for manual review when changes occur to product records.

  o A "merchants" table will be used to to identify which merchants' datafeeds will be imported, and which channels of distribution are permitted for each merchant's products. (For example, some merchants may prohibit promotion via AdWords, or may impose keyword bidding restrictions.)

  o There will also be "content tables," to define content to appear on web pages (including tables that link products to the pages where they will be displayed, and specifications for modifying which merchants and products will be included or excluded from particular pages or content areas, probably using separate tables to identify groupings of merchants, products, and content pages.

  o Other tables will define AdWords text ads, and track the performance of content pages and AdWords advertising.

- Choose technologies, versions; identify software licenses required.

Phase Two: Datafeed Import

A crucial element of this project is a program or script that can import and read "datafeed" files (tab- or pipe-delimited text files) and store them into a temporary database table.

The database will contain (and some kind of user interface must be devised to access) information about each merchant and about each datafeed format. Datafeeds for each merchant will be updated "automatically" (checking for updates daily, and automatically downloading and importing updated datafeeds when available) or "manually" (by user instruction).

2A: ShareASale Datafeed Contents:

The first sets of "datafeed" files are 300+ pipe-delimited files containing nearly 2 million product records from 300+ separate ShareASale merchants. Soon after, I will add datafeeds from other systems, which will have different field names and field orders, and will probably raise different validation issues. These may include:

> Affiliate Fitness / Affiliate Crew, AffiliateFuture, Affiliate Window, AllPosters.com and Art.com, Amazon Affiliate Web Services, AvantLink, Bridaluxe (formerly Red Galoshes), Buy.at, ClickBank, ClickXchange, ClixGalore, Commission Junction, Google Affiliate Network (formerly DoubleClick/Performics), Kolimbo/Kowabunga, LinkShare, PepperJam, and/or TradeDoubler.

Format of ShareASale Datafeed:

There is no "field-name" row in the ShareASale datafeeds (but here's what it would look like, if it were present):

ProductID|Name|MerchantID|Merchant|Link|Thumbnail|BigImage|Price|RetailPrice|Category|SubCategory|Description|Custom1|Custom2|Custom3|Custom4|Custom5|LastUpdated|status|manufacturer|partnumber|merchantCategory|merchantSubcategory|shortDescription|ISBN|UPC

Each ShareASale product feed is "pipe-delimited" with 26 fields, in this order. Many of these fields are often blank in many datafeeds. I've guessed at the best data-type for each field.

1.	ProductID	Long Integer
2.	Name	nvarchar(255)
3.	MerchantID	Integer?
4.	Merchant	nvarchar(50)
5.	Link URL	nvarchar(255)
6.	Thumbnail URL	nvarchar(255)

```
 7. BigImage URL          nvarchar(255)
 8. Price                 money                    24.95
 9. RetailPrice           money                    29.95
10. Category              nvarchar(50)
11. SubCategory           nvarchar(50)
12. Description           nvarchar(indeterminate-length)
13. Custom1               nvarchar(255)
14. Custom2               nvarchar(255)
15. Custom3               nvarchar(255)
16. Custom4               nvarchar(255)
17. Custom5               nvarchar(255)
18. LastUpdated           Date-Time                2007-01-22 22:57:07.170
19. status                nvarchar(50)
20. Manufacturer          nvarchar(255)
21. PartNumber            nvarchar(255)
22. MerchantCategory      nvarchar(255)
23. MerchantSubcategory   nvarchar(255)
24. ShortDescription      nvarchar(255)
25. ISBN                  nvarchar(25)
26. UPC                   nvarchar(25)
```

The script will need to periodically check remote servers for new versions of each datafeed, and then auto-download and import the new datafeed.

I've posted four sample datafeeds at http://www.MarkWelch.com/dfsamp1.zip

Some specific implementation issues:

- Product descriptions [12] have "no character limit" in the source data. We will probably choose to truncate "long" descriptions to a fixed maximum-length (probably 2,000 to 4,000 characters), but the remaining columns must be parsed correctly.

- Some merchants provide datafeeds in which the ProductID field is not unique. The URL columns (5,6,7) are also not consistently unique for a merchant. For those merchants whose ProductIDs are non-unique, a different field will need to be specified as the "index value" to be used in matching during update cycles.

- The "status" field (column 19) may contain several different values (some numeric, some text). Just because one merchant inserts only numeric values into this field, or another inserts values like "In Stock" or "Out of Stock," cannot be relied upon while choosing a data type.

- The ISBN [25] and UPC [26] columns may contain data in different formats (ISBN-10, ISBN-13, with or without dashes).

- During import, the system must count and report the number of "successful" and "failed" import rows.

- Assume that there **will** be invalid characters in text fields, which will need to be substituted before storing to SQL.

- The "parsing" or "mapping" of each datafeed may vary, because different merchants may assign different types of values to certain fields; some data (especially the "custom" fields) will probably be mapped to other fields within the product database. This will probably require special parsers for certain ShareASale datafeeds (and probably mapping structures for each network-datafeed format).

2B. Data Validations:
- Link, Thumbnail, and BigImage fields should start with http://
- Date-Time should be current era
- RetailPrice should never be greater than Price
- All prices must be > 0 (or null)
- Null values are OK for all fields except ProductID, Name, and Link.
- Assume that there **will** be invalid characters in text fields, which will need to be substituted before storing to SQL
- Date values do NOT need to be full precision – it's OK to round or truncate to "hour"

2C. Substitution or Parsing:

- There will be a <u>large</u> user-defined set of text substitutions, which will need to be implemented before comparing "new" to "existing" product data. This might occur during the import phase, or as a process after the temporary table has been created.

- Some examples:
    o Replace "Niker" with "Nike®" and " 1/2 " with " ½ "
    o Remove "FREE SHIPPING FOR ORDERS OVER $75"
    o Remove all HTML and URLs from description & title fields
    o Replace "userid=YOURUSERID" in URLs with numeric value

2D: Matching and Inserting Records into Product Database:

- **Discontinued Items:** The datafeed will identify the "complete" list of available SKUs for a merchant. If the database includes other SKUs (discontinued items), then:

    o If the SKU has any "associated content," including custom categories or product descriptions, or is otherwise referenced in any resource, the "discontinued item" should be retained in the database, but flagged as not-currently-available.
    o If the SKU is "completely unused," then it can be deleted from the database.

- **New Items** must be inserted into the product database.

- **Unchanged Items:** The "last-updated" record in the database must be incremented.

- Changed Items:

    o If the previous product data is not "in use" (identified as having already propagated to edited/processed/published content) then the previous record can simply be overwritten.

    o If the only change is an adjustment to price (within certain parameters), or to an image or link URL, then the old record may be overwritten with the new data.

    o All other changed items which are "in use" must be inserted into the product database, and the pre-existing record must be flagged as "changed" (the database will then contain at least two records for this product, which will eventually need to be "reconciled").

Phase Three: Reconciling/Managing Changed Records:

- In many cases, the review will simply be "approval" – allowing the new record to overwrite the old record for a product. (This must include options to "approve change to this record" as well as "approve change to all records with similar changes.")

- In other cases, records will be flagged for manual review and processing (for example, changes to a product title or description will require review to

determine if corresponding content should be changed; this full-review will be a separate phase).

Phase Four: Editing and Producing Content

4A: Creating Content
- Edited Product Titles, Descriptions
  - There can be multiple variations for each product!
- AdWords ads
- Keyword lists
- Content/Landing Pages (introductory text, sidebars, etc.)
- Directory Pages
- The vast majority of products in the Product Database will not have any associated content, and will appear only in 'search results' (such as directory pages or product-comparison pages).
- The system must be able to create and manage tracking codes (URLs with parameters) for each unique content component.

4B: Managing Updates:

- Pausing AdWords campaigns with unapproved changes or unusual performance data;

- Review and edit content associated with changed records and paused campaigns (including bulk approval of "similar changes," when possible);

- Merge records after review.

Phase Five: Generating Web Site Content Pages

- Static web pages: The system must be able to "publish" content as static web pages; this will include production of site-map pages.

- Initially, these files would be published to local directories.

- Over time, the system should be able to publish content to remote servers.

- Dynamic content: Ideally, the system should also be able to generate "templates" that allow creation of dynamic web sites (for example, searchable product-comparison or price-comparison systems); this is not an immediate priority.

Phase Six: Interfacing to AdWords API

- AdWords ads must be uploaded to Google's AdWords system (and later, to Overture/Yahoo, MSN AdCenter, and other PPC-search text-ad systems).

- In the earliest phases, the system might merely generate CSV or tab-delimited files for import into the AdWords Editor.

- The ultimate goal will be to interact directly with the AdWords API, including downloading statistical data, managing A/B tests, and automatically adjusting bids based on ROI data.

To follow up on this project outline, please contact Mark Welch (MarkWelch at MarkWelch dot com, telephone 925-462-8483.

Phase Seven: Interfacing to Merchant & Network APIs

- Eventually, the system should be able to import merchant performance and sales data from affiliate networks and individual merchants (possibly using APIs, but in some cases possibly initiating secure logins, and then downloading and parsing http web pages).

- The system may also need to import data from server logs.

- Performance data will be used to optimize content pages and AdWords campaigns.

User Interface Screens/options

 + Create/Edit Parser
 + Create/Edit Parser-Map (to add a "minor" variation, such as mapping Custom1 field, for a parser).
 + Create/Edit Datafeed (merchant name, network, ID, datafeed location, parser)
 + Import Menu (import now, pause/resume automatic import)
 + Create/Edit Merchant Information (edit all fields in 'merchants' table)
 + Create/Edit Merchant-Site-Review data
 + Create/Edit Merchant Group [list of merchants included in a task]
 + Review Duplicate Products (approve changes, reject changes, edit changes)
 + Create/Edit Substitutions (add, change, delete substitution strings); menu should include buttons to "execute [current substitution, all substitutions] on existing data"
 + Create/Edit Site (domain name, page-group(s), page template)
 + Create/Edit Page-Group (directory level, categories, page template)
 + Create/Edit Page Template (Columns, Font Sizes and Styles, etc.)
 + Create/Edit Page (define title, meta tags, intro text, etc.)
 + Create/Edit Alternate Product Title/Description
 + Create/Edit Product-Display Format (single-product; examples below)

Example Product Display Formats:

Accu-Chek Active Blood Glucose Meter Kit

$18.99

Diabetes Monitoring Kit Uses Very small sample required. Easy to use. Fingertip or forearm testing. Results in five seconds. Contents: Accu-Chek Active Blood Glucose Meter (battery included); adjustable Accu-Chek Softclix Lancet Device with lancets; easy-to-use instructions (English and Spanish); carrying case; self-test diary. Sold separately: Accu-Chek Compass and Accu-Chek Pocket Compass Software provide easy-to-read graphics and charts. Accu-Chek Active test strips and control solutions s... [Read more]
Merchant: Hocks.Com On-Line Pharmacy

FreeStyle Lite Test Strips 50/box

FreeStyle Lite Test Strips 50/box

$36.99

For Use with FreeStyle Lite Meter Only! Directions PLEASE READ THE PACKAGE INSERT BEFORE USING THIS PRODUCT. Do not use if expiration date has passed. Do not refrigerate or freeze. Store at room temperature below 86 degrees F (30 degrees C). Do not use if box seal is broken or missing. Indications No coding required! Virtually pain free testing. Use only with the FreeStyle Lite System. For self testing. Ingredients Chemical Composition: PQQ Glucose Dehydrogenase > or = 1.0 Unit; Other Ingredients (Buffer, Mediator, Etc): > or = 0.01 mg Warning For in vitro diagnostic use.
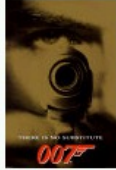Merchant: Hocks.Com On-Line Pharmacy

New York Yankees Lou Gehrig Caricature White T-Shirt
$20.00
[Read more]
Merchant: Distant Replays

Baby Planet Solo Sport Stroller - Crimson
$189.99
Merchant:

BabyDirect.com

James Bond Poster
Buy for $6.99
Framed   Mounted

From Russia With...
Framed Art Print
Buy for $46.98

GoldenEy
Buy for
Framed

Travel Poster
$25.00
HIGH - classic reproductions of vintage posters and artwork. Custom printed using the finest papers canvas and inc [Read more]
Merchant: RareArtShop.com

Designer Style Cable Ring with Garnet CZ
$28.00
Delicate gold accents set this Designer Style Cable Ring with Garnet CZ apa [Read more]
Merchant: Centerbrook Sales

Contact Information: If you skipped directly to this last page, without reading the full document, then **please don't contact me**. My contact information appears above "Phase Seven," above. *I am using this "trick" to "hide" my contact information because every time I post to Craigslist, I am flooded by "boilerplate" replies, mostly from offshore development firms.*