

CREST[®] V2.0

NAA Standard for Classified Advertising Data Technical Overview

Newspaper Association of America
1921 Gallows Road
Suite 600
Vienna, Va. 22182
Phone: (703) 902-1600

Table of Contents

- Executive Summary 3
 - Introduction..... 3
 - Task Force Formation 4
 - Task Force Mission..... 4
 - Initial Results (for Phase 1)..... 5
 - In This Document 7
 - Relation to Previous Versions..... 5
- Overview of Electronic Ad Sharing..... 8
 - The Players 8
 - Markup Classes..... 8
- Issues in Sharing Classified Ads..... 10
 - Unique Ad ID..... 10
 - Generating Universal IDs..... 11
 - Keyword Support for Searching..... 12
 - Improving Search Capability 14
 - Implementing Keyword Tagging..... 14
- Ad Formatting..... 15
- Graphics in Ads..... 16
 - File Formats, Resolution, Size, and Color..... 16
 - Linking Versus Embedding..... 16
- Ad Transport..... 18
- Compound Ads..... 18
- Special Services..... 19
 - Reply Forwarding..... 20
- Contract (Header) Information..... 21
- Extensibility 22

Classification Mapping..... 22

Technology Overview..... 24

 Why Use XML for Classified Ad Markup?..... 24

 The Document Type Definition..... 25

 Document Content Description May Replace DTD..... 25

 Well-Formed vs. Valid XML..... 26

Frequently Asked Questions..... 27

Executive Summary

Introduction

For years, print publishers and their advertisers have experienced a great deal of difficulty in making classified advertising available in electronic media channels. Advertising produced for delivery via a print medium—newspapers, for example—typically could be distributed in an electronic medium only after undergoing several labor-intensive processes. Obstacles to efficient ad publication in electronic media included lack of standardization in the presentation of classified advertising data, which can be deceptively complex, and the wide range of subjects and characteristics that classified ads address.

The rapid emergence and growth of the Internet's World Wide Web as a popular medium for accessing news, information and advertising has highlighted the need for making the aggregation and publication of print classified advertising in electronic media formats easier. Electronic media present many opportunities to classified advertisers and publishers which cannot be replicated in print.

On the Web, for example, auto ads are being combined with searchable blue-book databases. Real estate property listings are presented along with tools like mortgage calculators, realtor profiles, and demographic information. Help-wanted ads are complemented by resume-matching services and career advice. And while the technology is less than perfect, all Web-based ads can be searched electronically; this function alone offers an enormous advantage over traditional methods of publishing.

To more fully leverage their considerable classified advertising resources in electronic media, print publishers and their suppliers need to improve methods and mechanisms by which classified advertising data are collected, stored, transferred and presented to consumers. Establishing a standard for presenting classified advertising data obviously is key to achieving these objectives.

Task Force Formation

The NAA Classified Advertising Standards Task Force was organized to facilitate the electronic exchange of classified ads. The task force is composed of approximately 40 classified advertisers, advertising publishers, aggregators, system users, suppliers and technology experts. Representatives from organizations that either expressed or were expected to have a high degree of interest in the establishment of a standard for classified advertising data exchange—including several nontraditional newspaper industry suppliers—were invited to participate on the task force.

Task Force Mission

The Task Force’s stated mission is, “to establish standards that permit advertisers to provide, and newspapers to share and aggregate, advertising data for publication in media-independent formats.”¹ Standards developed will pave the way for aggregation of classified ads among newspapers on the Internet, as well as enhance the development of classified processing systems. The standards will be non-proprietary and, although intended to assist newspapers and their suppliers in publishing classified ads, they will be freely available to all parties.

¹ Reprinted from the NAA Classified Advertising Standards Task Force Web site at <http://www.naa.org/technology/clsstdtf/>

Initial Results (for Phase 1)

The task force developed a standard that addresses classified advertising from the moment it is placed—whether through another aggregator/publisher or directly from the advertiser—with a publisher intending to make the ad available electronically. Standard information sets, global tracking numbers and common descriptions of the data would permit them to be more easily shared, organized and published through multiple media channels. Though all the ad information might not appear in some media, the additional information collected would allow ads to be fully searchable on the Internet.

The standard does not, by itself, represent a solution to the problem of republishing classified advertising in multiple media. It will, however, provide a common classified advertising data structure that technology providers can use to build better tools and systems for handling these data. With standardized data formats and these better tools, advertisers, ad aggregators and publishers can simplify their workflows and more effectively provide data sought by consumers.

Four components of this standard will allow ads to be used on multiple networks and network appliances, as well as increase opportunities for aggregation. The four components are a standard data format, a standard transaction format, standard text-formatting tags and standard shorthand.

Keeping in mind the goals—unrestricted ad movement, powerful Web-based searching, media-independence, and support from commercially-available software—the task force selected Extensible Markup Language (XML) to develop its standards (see “Why Use XML for Classified Ad Markup?” on page 24).

The task force generated a document type description (DTD) to define XML tags and their proper usage in conjunction with this standard. The DTD has a set of elements, or fields, which describe the product being sold. Some of the fields—such as name and other contact information—are strongly recommended elements, and many more are merely recommended, giving newspapers the standard information they need while also allowing advertisers to be flexible. Because the real estate, transportation and employment categories

represent the bulk of all print classified advertising, the task force decided to tackle these three categories first (in Phase 1).

The DTD incorporates resolutions to a number of issues. These issues include:

- ◆ Tracking down shared ads, for editing or deletion, by a unique identifier;
 - ◆ Associating ad content with keywords, for Web searches;
 - ◆ Supporting the variety of formatting capabilities each publisher offers;
 - ◆ Ensuring quality control in electronic ad transfers;
 - ◆ Supporting special services (reply forwarding, proof copies, tear sheets, sorting) typically offered by newspapers;
 - ◆ Including advertiser, contract, and payment information in ad transfers;
 - ◆ Extending the ad markup language to allow individual publishers to enhance the standard with custom features; and
 - ◆ Mapping classifications between publishers.
- ◆ The group also decided that the following items should be created to facilitate the standard's implementation:
- ◆ A "vision" document or primer to explain to users the standard's purpose and provide examples of how it will look and function;
 - ◆ A DTD implementation model for classified advertising system developers that will provide technical details on XML tag descriptions, definitions and usage;
 - ◆ A DTD implementation model for business and classified advertising professionals that will provide recommendations on data collection; and

- ◆ A program for communicating this standard's availability and acquiring feedback from adopters.
- ◆ In future phases of task force activity, XML tag descriptions and definitions for additional classified advertising categories are expected to be established. Revisions and updates to existing tags will also be explored.

In This Document

This document is intended to fulfill the first bulleted item in the previous list. It summarizes results achieved by the NAA Classified Advertising Task Force during Phase 1 of its standardization effort and familiarizes readers with the most important issues associated with successfully exchanging classified advertising data. The “players” involved in classified ad entry, markup, and transmission are outlined and the technologies used for ad markup and transfer are identified. Potential problems that system suppliers must address are indicated.

A brief overview of XML and the reasons for its selection as the standard's basis is given, followed by answers to frequently asked questions (FAQs).

Relation to Previous Versions

The CREST[®] V2.0 (Version 2.0) format completely replaces the V1.0 exchange format. Version 2.0 continues to embody the philosophy espoused in V1.0 while considerably expanding the utility of the information within each ad instance.

Overview of Electronic Ad Sharing

The Players

The players involved in classified ad publishing and transferring include:

- ◆ the *advertiser* who wants to place the ad;
- ◆ the *ad creator* who can create the ad;
- ◆ the *upstream publisher/aggregator*—who can create or accept the ad and publish it but does not generate XML markup for it; the upstream publisher/aggregator can transfer the ad downstream to additional publishers;
- ◆ the *ad publisher* who can create or accept the ad, generate the appropriate XML markup for it and publish the ad; the ad publisher can transfer basic ad content expressed in XML—along with other required public markup information—to downstream publishers/aggregators; and
- ◆ the *downstream publisher/aggregator* who can take collections of ads with XML markup, arrange them together and publish the XML markup, often in a different media.

Markup Classes

The NAA Standard for Classified Advertising Data includes three classes of markup:

- ◆ *Public Markup—Ad Transfer* information: data to be forwarded to downstream publishers. This information includes the ad content itself, searchable terms to be used for indexing, placement information, a unique identifier, an optional expiration date, mailbox, and an action to be taken by the publisher (create, kill, update, or preview).

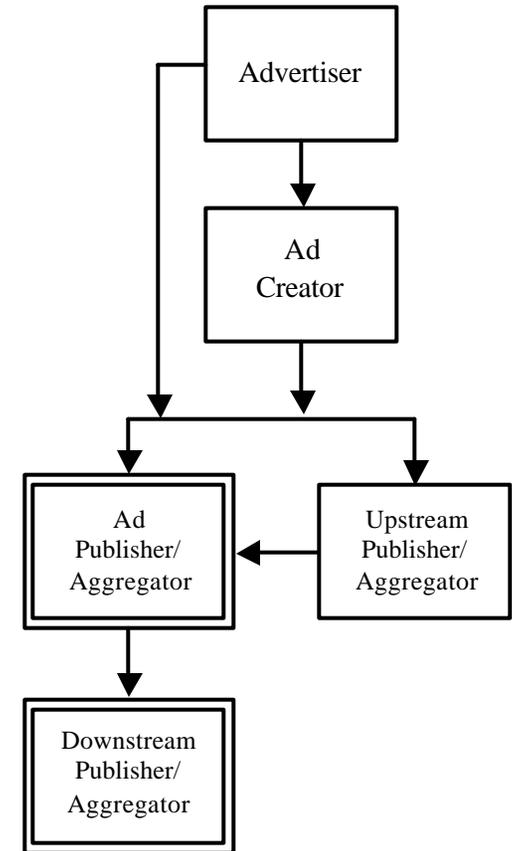


Figure 1: Classified Ad Transfer
Double border represents XML markup

- ◆ *Private Inbound Markup—Ad Origination* information: data required for billing that are not forwarded downstream to other publishers. This information includes account information, comments, and publication-specific claim, sort, width, and zone directions.
- ◆ *Private Outbound Markup—Ad Feedback* information: data returned by the publisher to the ad’s originator that are not forwarded downstream to other publishers. This information includes status, authorization, version, price, rate breakdown, and positioning information, as well as warning and confirmation messages.

In Figure 1, downstream publishers/aggregators receive public markup (ad transfer information) from ad publishers/aggregators for re-publication. In Figure 2, the advertiser provides both public and private information to the ad publisher/aggregator. The private information includes payment arrangements, for example. The public information includes the ad content, formatting, coding, and classification. This information is passed on to the publisher’s audience.

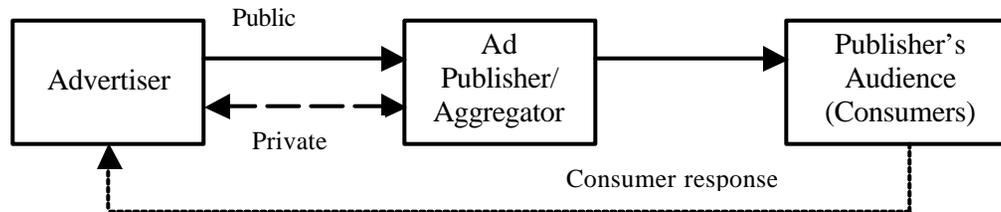


Figure 2: Classified ad markup

Issues in Sharing Classified Ads

The following pages identify some of the issues that are encountered when print publishers and aggregators attempt to share classified ads.

Unique Ad ID

In most cases, local classified advertising systems assign an identifier when ads are submitted, so that the ad can be killed or traced for pickup and editing.

When classifieds are published independently, it doesn't matter that one ad publisher assigns the identifier 1234 to a help-wanted ad on October 6, while another ad publisher assigns the same identifier, 1234, to a real-estate ad on October 10. If these same ads are electronically shared with downstream publishers, duplicate identifiers can create trouble. An operator at one publisher attempting to kill ad 1234 must be able to pinpoint the real-estate ad transmitted by his group, and not a help-wanted ad transmitted by another publisher.

Another reason ads must be uniquely identified is "stereo transfer," which occurs when downstream publishers receive the same ad from more than one source.

An ad taken by print ad publisher 1, for example, assigned the unique identifier "pub1.2468", may be transmitted downstream to print ad publisher 2 and to Web ad publisher 3. Upon receiving "pub1.2468", print ad publisher 2 may transfer the ad to both print ad publisher 4 and to Web ad publisher 3. Although Web ad publisher 3 receives the same ad twice, its system recognizes the redundant identifier and rejects the duplicate. Figure 3 illustrates this example of stereo transfer.

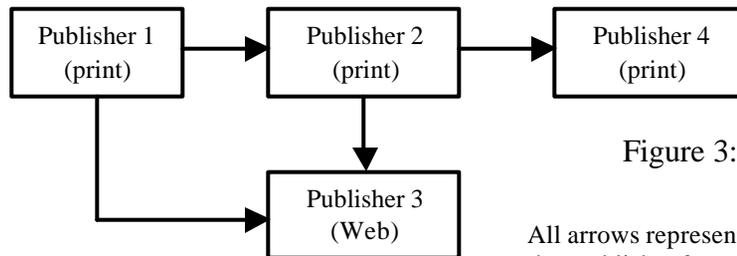


Figure 3: Stereo transfer of an ad

All arrows represent a transfer of the same ad. Note that publisher 3 receives the ad from 2 sources.

Generating Universal IDs

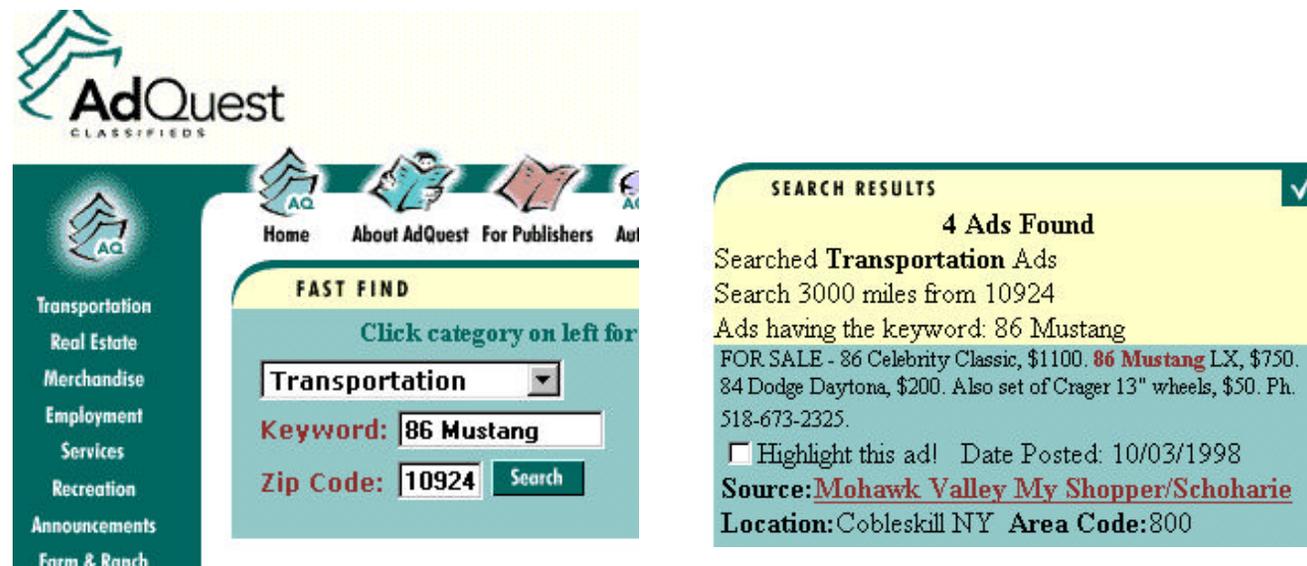
There are several possible methods for ensuring that each ad is assigned a globally-unique identifier, including the following:

1. Each ad publisher or aggregator is assigned a unique prefix (perhaps its own Web domain name, since these names are unique). An ad taken by publisher 1 and issued the locally unique identifier 1234 might be transferred to downstream ad publishers/aggregators with the prefix “pub1”. Pub1.1234 then becomes the ad’s globally-unique identifier.
2. A universal service ensures that no identifier is ever issued more than once. However, this concept prompts the following questions:
 - Who offers this service?
 - How is the number generated?
3. Utilize Microsoft’s method of generating 128-bit Globally Unique Identifiers (GUIDs), which are built from three components:
 - a unique base—the PC’s unique TCP/IP address;
 - the current date and time; and
 - a randomized number

Keyword Support for Searching

Classified ads that originate from print publications and are placed on the Web are not merely electronic duplicates of those print ads. Web publishers add value by making the ads searchable.

Suppose a buyer wants a 1986 Ford Mustang. She goes to a classified Web site, say AdQuest.com. She chooses a category (Transportation), specifies a search keyword (86 Mustang), and identifies her location with a postal code. When she clicks the Search button, AdQuest displays 1986 Mustangs for sale in her region.²



² Reproduced with permission from the World Wide Web at <http://www.adquest.com>. AdQuest Classifieds is a classified advertising service offered by PowerAdz.com LLC.

Some sites allow advanced searches in which the buyer can specify more than one keyword; e.g., mileage, model, year, and so on. The following figure shows another AdQuest search³:

Vehicle Search

Enter Your Zip Code: Distance:

Type of Vehicle
 New Used
 Year Range
 to

Vehicle Make Desired

 Vehicle Body Style
 4WD, only
 Min Price
 Max Price

Model / Special Equipment etc.

Leave this field empty on your first try.
 Narrow list by model, color, option, etc.
 Mileage Range

 Listings per page:

In order to turn print ads into searchable ads, ad text must be coded with database attributes. Ad text may be parsed to extract searchable keywords like *Mustang*, *Celebrity Classic* and so on. Parsing software is commercially available and a number of publishers have developed their own custom software. Parsing, however, is an inexact science. Words may have unique regional—or cultural—meanings, and they may be misspelled. Numbers corresponding to a given item can be spelled out in one ad and displayed as digits in another. Abbreviation and punctuation vary.

Consumers searching classified ads by keyword may need to perform repeated searches, each time altering spelling and abbreviations, in order to find a desired number of ads. In the example above, the keyword “86 Mustang” found ads in which the year was abbreviated

³ Reproduced with permission from the World Wide Web at <http://www.adquest.com>. AdQuest Classifieds is a classified advertising service offered by PowerAdz.com LLC.

with an apostrophe, but not ads in which the year appeared as 1986 or 86, or those in which the year followed the car's model name.

Improving Search Capability

To improve search capability, keywords may be coded along with formatting markup. The operator taking an ad may insert XML tags into ad text, thereby earmarking keywords for entry in a database. Text entered by data element is easier to manage and search. Each element is stored in a separate database field and standardized to enhance lookup.

To illustrate, refer again to the 1986 Mustang search example. If the digits, "86" are stored in a field called *Year*, the year of manufacture (86) won't be confused with the same digits coincidentally appearing inside another ad's telephone number. Furthermore, the year of manufacture can be standardized as 1986, whether it appears in the ad as '86, 86, or 1986. To standardize auto mileage, miles displayed in one ad as 17K and in another ad as 17,000 can both be stored in a field called *Mileage* as 17000.

The use of keyword tagging and standardized ad data fields dramatically improves searching.

Implementing Keyword Tagging

Print publishers' classified advertising operators are accustomed to taking ads by typing the ad content and applying formatting. While keyword-coding ads requires additional effort when performed manually, increasing levels of automation are imminent as technologies for Web searching become more sophisticated.

Furthermore, the process can be reversed. Instead of extracting keywords from ad text, ads may be captured in a database as a series of data elements. From this database, ad copy can automatically be generated. This approach has more appeal when the database already exists. For example, classified aggregators build ads from pre-existing automobile dealer inventories; Realtor.com creates residential home listings from pre-existing MLS databases.

This standard includes provisions for keyword tagging but publisher implementation of this practice is partially contingent on other factors. For example, the resources required for a publisher to perform keyword coding—even after applying currently available automation

tools—still might exceed the financial or business gains that the publisher can realize by providing consumers with enhanced electronic ad search capabilities. In this case, the publisher might decide not to implement keyword coding until either the process of keyword coding requires fewer resources or market conditions improve to provide greater returns on the publisher's investment of resources.

Ad Formatting

Although all print ad publishers support ad formatting in some form, each publisher's formatting capability is unique. Some formatting options that can differ from publisher to publisher include:

- ◆ Standard fonts (e.g., Times Roman, Helvetica)
- ◆ Fonts used for emphasis (e.g., italic bold).
- ◆ Justification options (e.g., centering, tabbed leader dots)
- ◆ Type size
- ◆ Number of columns allowed
- ◆ Images (e.g., logos, halftone photos)

Formatting differences become an issue when ads are shared by several publishers. Advertisers desire some degree of control over the way ads appear, and are accustomed to specifying formatting for a particular publication. What happens when a downstream publisher or aggregator does not support all of the requested formatting?

This standard's DTD can provide the advertiser with a reasonable degree of control because downstream systems will ignore formatting tags that they don't understand—font, type size, and column set size, for instance—and instead apply locally determined formatting to the text inside these tags.

Graphics in Ads

File Formats, Resolution, Size, and Color

There are distinctions between suitable graphics for print versus Web publication. While most print classifieds don't contain graphics, many print publishers now accommodate logos and halftone photographs in classified ads. For example, an auto ad might include the auto dealer's logo; a real estate ad might include a halftone photo of a house. Advertisers typically mail camera-ready artwork to print publishers or supply them with .tif or .eps files via electronic transfer or removable storage media.

On the other hand, ads destined for Web publication often include color images. The Web supports .gif and .jpg image formats but not .tif or .eps. Because of bandwidth limitations, Web publishers may wish to display small, low-resolution images in some cases (in ad hit lists, for example) and larger, high-resolution images in others (allowing users to zoom in on thumbnails for more detail).

For these reasons, the NAA standard's DTD was developed with the flexibility to allow ad publishers to transmit a variety of images from which downstream publishers and aggregators can choose. To assist downstream publishers in choosing the proper image to use in a given situation, the DTD includes XML notation tags. These tags offer a means to identify the inclusion of a specific graphic format and tell the processor which program to use in processing the image. The DTD's notation tags identify various image formats.

Linking Versus Embedding

Ad publishers and aggregators determine the method by which they will make images available to downstream publishers and aggregators. They can choose to:

- ◆ Maintain images on a Web site so that downstream publishers can
 - a) pick up the image and re-publish it, or
 - b) create a hyperlink to the image on the ad originator's Web site;
- ◆ Embed images within the ads they transmit.

Actions associated with each approach and some of the potential pitfalls that they possess are summarized below.

MAINTAINING IMAGES ON A WEB SITE

Within an ad's XML markup, the ad publisher supplies a URL as a reference tag for the image. Downstream publishers and aggregators fetch the image and publish it on their own sites and then substitute their own URLs in the image reference tag.

An advantage to this approach is that downstream publishers do not need to remain dependent on the originator's maintenance of the site. Also, the ad originator's identity is masked from downstream publishers' readers. A disadvantage is that downstream publishers won't know when images have been replaced by newer versions. In most cases, however, new images probably will not be created unless ad text is also updated. The potential for publishing superseded images can therefore be mitigated by updating images whenever ad updates are received.

EMBEDDING IMAGES

When embedding images, ad publishers must represent the graphic in a manner that will allow it to survive processing by XML parsers. Jpeg files, for example, contain binary data that can potentially derail a downstream publisher's XML parser. To avoid confusing the parser, ad publishers can represent the graphic using MIME Base64 content transfer encoding.

The Base64 representation requires 4 bytes in the XML document for every 3 bytes in the original graphic file. The graphic is represented using only A-Z, a-z, 0-9, +, /, =, and white space. Since no < or & is used, the XML parser won't be confused by the Base64 data. MIME is an acronym for "Multipurpose Internet Mail Extensions." Base64 content transfer encoding is fully specified as part of the MIME standard. Details are available at <ftp://ds.internic.net/rfc>.

Ad Transport

The NAA Standard for Classified Advertising Data primarily addresses data exchange and not transport mechanisms for exchanging the data. However, the standard does include provisions for:

- ◆ recognizing that the correct parties are communicating with one another;
- ◆ determining which financial details must be conveyed along with the ad and when this information is required; and
- ◆ identifying if, and how, logo graphics should be transferred.

There are additional issues that, ideally, would be addressed in a future version of this standard or in a separate standard produced by the task force or a similar industry group. Some of these issues include:

- ◆ recognizing whether a downstream user can accept ads;
- ◆ confirming that a transmission was successfully completed and the entire package was transmitted intact; and
- ◆ determining which protocols should be used for electronically transferring ads—http, ftp, or other—or the form of removable storage media (CD, Zip drive, etc.) on which ads should be transferred.

Compound Ads

Compound ads present multiple items (e.g., houses, cars, jobs) together in a single ad. In some cases, compound ads are generated from auto-dealer inventories, auction-house inventories and property inventories supplied by large real estate firms.

Compound ads are popular among advertisers because they usually are larger than other ads and therefore are more likely to attract consumers' attention. They also take fewer lines than

would be taken to run the comprising ads separately, since heading and contact information do not have to be repeated. An example of a compound ad is provided below.

<p style="text-align: center;">PORSCHE</p> <p>'99 911 Carrera Cpe, Black/Black..... NEW '99 911 Carr. Cpe, Arena Red/Beige..... NEW '97 911 Turbo, Silver/Black..... NEW</p> <p style="text-align: center;">LAMBORGHINI</p> <p>'97 Roadster Red/Tan, 100 mi '91 Diablo White/White, 6K mi</p> <p style="text-align: center;">OVER 300 PRE-OWNED EXOTICS IN STOCK</p> <p style="text-align: center;"></p> <p style="text-align: center;">(800) 555-5555</p>

Because electronic media support searching, compound ads must be keyword-tagged just like regular ads. Web ad publishers and aggregators may prefer to split compound ads into separate small ads, each containing the information common to all, since small ads displayed in a list of search results are easier to read on computer monitors than very large ads.

This standard’s DTD supports tagging separate items within each compound ad, distinguishing between common information (like headings, logos, and contact information) and unique items like the individual autos for sale in the above ad.

Furthermore, the DTD provides a means to allow common information to be shared by some, but not all, unique items. In the above ad, the logo and contact information are shared by all of the items; these common elements appear at the bottom of the ad. However, headings are shared only by some individual items. “Lamborghini,” for instance, is a heading shared by only two of the five autos for sale.

Special Services

Many print publishers offer special services—reply forwarding, proof copies and tear sheets, for example—but publishers’ services are not identical. Also, special services are not handled in the same manner by all publications. The NAA standard supports special services but

downstream publishers and aggregators can choose to ignore the tags and extract their contents.

Reply Forwarding

Most consumers of classified advertising have, on at least one occasion, seen instructions to send replies to a postal box number maintained by the ad publisher. For these “blind box” ads, advertisers pay the publisher for the postal box and the forwarding service. Genuine ad replies are forwarded to the advertiser’s address, which remains unknown to classified consumers.

Since individual ad publishers might handle reply forwarding differently, each must be able to provide its own instructions to advertisers. Among the stipulations that these instructions could contain is acceptable reply modes; some publishers might welcome fax, voice mail or e-mail responses in addition to traditional mail replies.

The following elements for reply-forwarding instructions are defined:

- ◆ **Forwarding Element:** Verbiage providing reply instructions, to be substituted with text developed by publishers that offer these services. Ad publishers should be able to specify whether they wish replies held for pick up or forwarded by mail, fax, or in the same mode by which they were received. A means to specify unique verbiage for different publications, along with a unique box number, is also provided.
- ◆ **Mailbox Element:** The mailbox number, assigned by a publisher’s classified advertising system.
- ◆ **Reply Element:** Verbiage providing reply instructions, to be used by publishers that do not offer reply-forwarding services.

Ad publishers may write the ad for multiple publishers at the outset. In this case, they may specify unique reply-forwarding instructions for each publisher separately. Each publisher,

when setting the ad, inserts the verbiage specific to its own publication. (This is an effective way to track replies in order to evaluate which publications elicit the most responses.)

Reply forwarding ads can lead to a number of problems, however, if tags for these elements are not properly used. Ads written for one specific publisher may end up being transferred downstream. If an ad originally written for publisher 1 is eventually transferred to publisher 2, it may still instruct readers to reply to publisher 1's mail box, without indicating that this mail box does, in fact, belong to publisher 1. To minimize confusion, city, state, and postal code information should also be supplied within the ad, so that consumers have a complete address for replies.

Replies may also be daisy-chained. A consumer's reply to a blind box ad accessed from publisher 2, but originally published by publisher 1, would be sent to publisher 2's blind box facility. The reply would then be forwarded from publisher 2 to publisher 1's blind box facility, which would then forward the reply to the advertiser. In this case, publisher 2 never learns the advertiser's identity.

Contract (Header) Information

Billing information required by an ad's original ad publisher might not be transferred to downstream publishers or aggregators, with which the originating publisher often already has standing agreements for ad sharing. Examples of this information include:

- ◆ advertiser account, contact, and payment instructions;
- ◆ optional reference information appearing on the ad invoice
- ◆ publication-specific claim, column, and sorting specifications

Extensibility

Despite the task force's efforts, this standard's DTD probably will not anticipate publishers' every requirement. Publishers therefore require the means to override or add tags to the ad markup language, as needed to suit specific situations. Ad publishers must be able to include unique formatting tags in the ad stream, without invalidating the XML for downstream users.

For example, publisher 1 allows two-color (black-and-red) ad printing while the DTD supports only one color. Publisher 1 should be able to define start and end tags to format the red text. These tags will not be understood by the systems operated by downstream publishers 2 and 3, and therefore will be ignored by them. Ad text inside the ignored tags will then be formatted as plain text.

Valid XML documents identify the appropriate DTD (which, in this case, is the NAA standard's DTD) and can also directly contain DTD instructions. These instructions are processed first—and override—the instructions given in the DTD. This features allows publishers to extend the standard to meet special needs. Ad publishers extending the DTD in this manner should, however, limit their overrides to tag additions and providing the proper context for these new tags. This would assist downstream publishers in:

- ◆ Validating the XML markup; or
- ◆ Ignoring the new tags and being assured that critical tags exist and appear in their expected positions.

Classification Mapping

Publishers use different classification schemes. A typical classification scheme assigns classification numbers in a hierarchy. For example, real estate classifications are grouped together and subdivided by type of offering (e.g. rental, sale, land) and by neighborhood.

Classification schemes are likely to differ in:

- ◆ Classification numbers—“1235 Auto” versus “5642 Auto”
- ◆ Classification titles—“Land for Sale” versus “Lots for Sale”
- ◆ Classification hierarchy—Is type of housing unit below or above neighborhood?
- ◆ Rules for determining where an ad should appear—Is a bicycle sold under “transportation” or under “general merchandise?”

When an ad is transferred from one publisher to another, the new publisher is likely to use a different classification scheme. Given this, which classification should the second publisher use?

Technology Overview

Why Use XML for Classified Ad Markup?

XML (Extensible Markup Language) is a protocol for defining markup languages that establish common syntaxes for structuring data. XML provides a means to transmit and exchange structured data between applications and electronic systems for media-independent display and manipulation.

Each XML element identifies a particular type of information. Unlike display instructions that can be interpreted only by the specific devices for which they are authored, any XML-enabled system receiving XML markup can interpret the elements and determine how to display them appropriately. Using XML markup, classified ad publishers can define the ad text structure, searchable terms, placement information, unique ad identifiers and so forth. Once structured, the ad can be handled by—and tailored for—any device with an XML browser. Because XML is extensible, it can be made to handle every requirement for publishing classified ads across various media.

“By separating structure and content from presentation, the same XML source document can be written once, then displayed in a variety of ways: on a computer monitor, within a cellular-phone display, translated into voice on a device for the blind, and so forth. It'll work on any communications devices that might be developed; an XML document can thus outlive the particular authoring and display technologies available when it was written.

So XML will have a life outside of the Internet, serving the publishing industry at large, for example, and especially people who produce documents intended to appear across multiple media.”⁴

⁴ Reprinted from CNET's Builder.Com site, “Twenty Questions on XML”.

The Document Type Definition

A Document Type Definition (DTD) defines syntax rules for a common XML tag set. It defines the tags that can be used in a document, the relationships between the tags and their order, which tags can be nested, which have attributes, and so forth. In essence, the DTD is an agreement between people exchanging documents on how they will use XML to describe a common document architecture.

The DTD ensures uniform structure and desired order, required for accurate data exchange and interpretation. Because XML does not have one universal DTD—as does HTML—each industry that wants to use XML for exchanging data must define its own DTDs.

This standard includes a DTD, which will be updated and evolve as necessary to address new issues and technologies.

Document Content Description May Replace DTD

Specifications for some new XML technologies are expected to undergo changes before adoption by the World Wide Web Consortium (W3C). DTDs may eventually be replaced by a new Document Content Description (DCD) specification, which recently was submitted to the W3C by IBM and Microsoft Corporation. The proposed specification makes it easier to develop business applications supporting XML.

“DCDs improve on DTDs in the following three principal ways:

“Unlike the DTD, the DCD provides the ability to specify data types. For example, if the value, or content, of a tag is the number 120874, a DCD will let the developer specify whether that number is a date, a time, a time interval, a Boolean value, an integer, a decimal, or some other type of data.

“DCDs will let authors create open content models. The way it is now with the DTD's closed model, an author cannot add tags to a completed DTD. But the DCD will let authors carve out a space for additional tags to be specified at some point in the future.

“DCDs allow new flexibility in letting developers reuse tags. For example, an invoice written using XML could reuse an address tag set within the document. Another XML

document also could use that tag set. Neither of these capabilities exist with the current DTD.⁵

The NAA Classified Advertising Standards Task Force will evaluate the feasibility of adopting the DCD specification—and any other new technology—if it shows promise in offering dramatic net improvements to current data exchange methods.

Well-Formed vs. Valid XML

XML documents are either “well-formed” or “valid”.

A well-formed XML document conforms to the rules of XML syntax, including:

- ◆ XML markup always starts with a left-angle bracket or an ampersand;
- ◆ XML data are always closed with an end tag as in the tag pair `<text> </text>`;
- ◆ Elements and attributes are case-sensitive; and
- ◆ Attributes require quotation marks.

A valid XML document is a well-formed XML document that conforms to a specific Document Type Definition (DTD).

The NAA standard stipulates that XML markup for classified ads be valid; systems generating this markup should confirm its validity.

⁵ “W3C mulls XML spec,” by Paul Festa, Staff Writer, CNET News.com. August 10, 1998.

Frequently Asked Questions

Q: Is the “NAA Standard for Classified Advertising Data” a proprietary standard?

A: No. It is non-proprietary and will be freely available to all parties. The standard was developed using XML to ensure media-independence and to take advantage of commercially-available software.

Q: Who is expected to implement the standard?

A: The standard is intended to benefit any party, including newspapers, that regularly aggregates or publishes classified ads in an electronic medium.

Q: How frequently will the standard be revised or updated?

A: There currently is no set update schedule; the standard will be modified as necessary to incorporate significant improvements.

Q: How will implementers know if a new or updated version of the standard is available?

A: Information about updates will be posted to the NAA Web site at <http://www.naa.org/technology/clsstdtf/>. Developers and implementers interested in receiving word about updates are encouraged to inform NAA by sending an e-mail to Dr. John Iobst at iobsj@naa.org.

Q: When will tags for other classified advertising categories be developed?

A: There currently is no set schedule for developing these tags; they will be generated as needed.

Q: How should implementers submit revisions that they would like to have incorporated?

A: Requested revisions should be submitted by e-mail to NAA's John Iobst at iobsj@naa.org.

Q: Can current or prospective implementers join the task force?

A: The task force is open to any interested parties, who should send an e-mail to NAA's John Iobst at iobsj@naa.org

Q: Is this standard expected to become an international standard?

A: There is a general expectation that the standard will be used internationally. The standard includes no provisions that limit its usage to North America.

Q: Will this standard be reviewed for accreditation by any of the foremost standards bodies, including CGATS, ANSI, ISO or the W3C?

A: Although ISO has standardized some DTDs, in general there currently is minimal interest in standardizing industry-specific document type definitions. CGATS might be interested in adopting the standard at some point in the future.

Q: How does the CREST® V2.0 format relate to CREST® V1.0?

A: The CREST® V2.0 format is a complete replacement for V1.0. Many of the basic concepts remain the same with respect to the exchange of ad materials. The technology for document management has changed sufficiently that current implementations will need to be replaced to support the new format.