Digital Oasis

Requirements Specification

Prepared for **Knight-Ridder, Inc.**

by **Barking Sands Software** and **PressLink**

August 19, 2002
Version 1.06

Executive Summary

Introduction

This document is the result of a requirements investigation which examined the need for archive systems at the Knight-Ridder newspaper companies. The scope of this investigation was purposely constrained to the area where the need is most pressing – the local archiving of photographs; however, the proposed solution is extensible to handle any type of digital data (e.g., text, graphics, advertisements, full pages, video, etc.).

The local photo archive requirements are derived from contrasting the current systems in use today with the vision of an architecture which is flexible enough to meet the needs of many different environments. Today, newspapers must make a choice between purchasing closed systems which are tightly integrated (such as those offered by AP) or tapping into a large, but much less integrated array of software tools on open platforms (such as Macintosh or Windows). Any future system must eliminate this dilemma by being capable of easily exchanging data and messages with other software products on non-proprietary platforms in a "virtually integrated" manner. In addition, it should support standards that are already in place in the publishing industry, and it must offer powerful searching/browsing functionality via an intuitive user interface. Finally, it must be robust enough to handle the large volumes of images which must be stored – without requiring excessive system management.

The benefits of implementing such a system for archiving photos at the newspapers go beyond the enhanced ability to create a better newspaper in a shorter amount of time. It allows local papers to accumulate value in their individual repositories of photos. In the short term, this system will enable quicker response times to special requests from subscribers or other papers. In the longer term, one can envision the local archive as a revenue-generating on-line system, electronically accessible to the other papers, other businesses, and even the public.

Having specified the requirements for this local photo archive, we offer a proposed three-phased implementation plan that fits into the architectural vision referenced earlier. Since many of the pieces of a solution are available today via off-the-shelf products, the first phase of the plan is deliverable by the end of 1993. By concentrating on the "glue" that binds these pieces together, Knight-Ridder can establish an architecture which

allows new functionality and/or replacements for current pieces to be seamlessly integrated into the system at some future date.

Market Opportunities

Although this requirements study targeted users at Knight-Ridder properties, the market for such an archiving system certainly goes far beyond those 28 newspapers. The Advanced Technology group at Gannett, for example, is searching for a similar solution. Information providers such as Knight-Ridder Tribune (KRT), Agence France-Presse (AFP) and Reuters need a local system of their own so that they can effectively handle special requests and build value in their own digital archives. Dialog, VuText, the Knight-Ridder Design Lab, and other newspapers also have expressed interest in such a system.

We believe the local archive solution outlined in this document can compete successfully on price/performance with currently available systems, such as those from AP and AXS. Since the system would offer functionality and performance that meets or exceeds these competitors, and since it could sell for considerably less (perhaps $200/seat for client software or $35,000 for a medium-sized newspaper installation), the market to replace current systems in addition to attracting new business should be considerable.

Project History

In April, 1993, Gannett commissioned RWD Technologies, a system integration company based in Columbia, MD, to undertake a joint project with Apple Computer that would address local photo archiving. RWD produced a requirements document for this project, known as the NEWSworks Image Management System (NIMS), in June. Subsequently, representatives from Apple, RWD, and Gannett met with Dave Bauer of Knight-Ridder and Todd Carter of PressLink to explore Knight-Ridder's interest in such a system and solicit feedback on the NIMS requirements document. Reactions to NIMS, as well as correspondence on the PressLink system among the "Digital Idiots" group, suggests that a more modular, scalable approach that still allows for tight integration of the system components is desirable.

Given this feedback from its properties, Knight-Ridder funded the creation of this requirements specification along with a proposed detailed design for Phase I of the project. During this portion of the investigation, we received input from a variety of people, including:

- Dave Bauer, Knight-Ridder
- Pete Pitz, Knight-Ridder
- Murray Koodish, San Jose Mercury-News
- Craig Morton, Detroit Free Press
- John Van Beekum, Miami Herald
- M. J. Crowley, Philadelphia Newspapers
- Charlie Borst, Knight-Ridder Tribune
- Hugh Moir, Agence France-Presse
- Rich Cates, Apple Computer
- Rick Blair, PressLink
- Roger Fidler, Knight-Ridder Design Lab
- Bob Ingle, San Jose Mercury-News

General Objectives

Business Goals

Lower the cost and time spent to retrieve a photo. Users should have easy-to-use tools at their desks which provide immediate access to the local archive. This access should eventually extend from local on-line and off-line storage to off-site archives.

Leverage existing standards such as the ANPA/IPTC headers and JPEG compression. By adopting a file format that incorporates these standards, the range of data sources and data manipulation tools increases dramatically. By using a file format which binds an image's metadata (ANPA/IPTC text) and thumbnail to the image itself, the flow of information into and out of the archive is greatly simplified.

Streamline current processes. For example, today KRT often requests photos from individual newspapers for its own wire service on an ad-hoc basis. This requires staff at the newspaper to make time to locate the photos and send them to KRT manually. In the future, this process could be expedited by allowing certain personnel at KRT to have electronic access directly to the local photo archive at the newspaper.

Accumulate value in local archive of photos. Today's archives represent an untapped revenue source for Knight-Ridder. Newspapers have the data to become "community knowledge centers" which can be the source for a wide variety of custom publications. In the future, subscribers might dial into a database of pictures, select twelve images, and subsequently receive a custom 1994 calendar in the mail.

Promote innovation and incremental enhancements by standardizing the interfaces to the various modules that comprise the system. As an example, the local archive may become a service that is accessible to users via standard e-mail (a user mails in a request for a certain type of photo, and receives notification by e-mail when relevant photos become available). One can also envision the local photo archive as a component of a larger system which might offer customized newspapers to readers based on their own preferences. These customized papers may be assembled by an intermediary editor who contributes additional value.

Independence from proprietary systems. In a closed system only a finite amount of enhancement and innovation can occur. In addition, it is difficult to take full advantage of data stored in a proprietary system or migrate the data to other systems at some future time.

Network support. Multiple users residing at different physical locations will need simultaneous access to the system.

Desktop-driven client-server approach. The client-server model provides a modular framework in which processing logic and data management functions are distributed appropriately between the user's desktop machine and the more powerful server platform. If systems are designed in this distributed fashion, they instantly become more scalable and adaptable.

Adaptability. The system should be easily configurable by administrators and users to fit the specific needs of their own environments.

Scalability. The system should be available (at an appropriate price point) as a solution for small newspapers as well as large ones.

Ease of administration. Maintenance of the system, including the "aging" of photos to off-line media such as CD-ROM, should be highly automated and intuitive.

Ease of use. The system must fit the way newsroom staff and librarians work. It should not be so complex that only a librarian will take the time to learn and use it, but it should be powerful enough to support the functions that librarians need.

Virtual integration on the desktop. The components of the system that reside on the user's machine should appear to be tightly integrated from the user's perspective. From a systems perspective, however, they should exist as distinct pieces that communicate via a well-defined messaging scheme. Thus, virtual integration allows new or enhanced pieces to be plugged into the system as long as they comply with the published messaging and data standards. Examples of components that can be virtually integrated on the desktop include: mail/messaging, file transfer, encryption/decryption, authentication, authorization, and compression/decompression.

Security. Administrators must be able to define and easily manage who is allowed to search any specific subset of the archive, copy files from the archive, modify metadata about the images, etc.

Current System Analysis

Functionality

Most newspapers are currently using a variety of off-the-shelf tools in order to manage photos and graphics. Packages on the Macintosh platform from vendors such as Adobe, Aldus, Storm, and AXS are used to create, modify, store, and retrieve these images. These packages offer a great deal of functionality, but attempts to make them work together in an integrated system result in frustration and inefficiencies.

On the other hand, many newspapers also use the AP Leaf system (at a cost of up to $120,000 per site) to some degree to accomplish these same functions. AP Leaf was designed to be a tightly integrated system that would satisfy all the photo needs of the paper in one package. Unfortunately, since Leaf is based on proprietary hardware and software, it cannot take advantage of the rich set of media tools and infrastructure available on more open systems like the Macintosh.

Users, therefore, are caught between dependence on AP to enhance its system and keep up with their needs on the one hand and labor-intensive attempts to find the right fit among the many tools on the Macintosh or Windows platforms. As a result, many papers currently use either ad-hoc electronic archives with no real guarantee of success, or they live with paper archives and their inherent limitations. Not surprisingly, most papers end up spending large amounts of money on special photo requests and retransmission fees to the various providers.

All of the current solutions on the market tend to bind the user's data into the fabric of the application, making it more difficult to take advantage of the data in other systems. For example, a newspaper may invest significant resources in adding keywords to images in a collection managed by Aldus Fetch. When one or more images from the archives need to be "exported" from the archive and used elsewhere, the metadata associated with the images cannot easily be extracted. So, the value that was added to the archive in the keywording process is lost once images are taken outside of the archive's domain.

Even when users have access to an electronic archive, such as AXS News Photo Archive, they complain that it is hard to find the right image when searching on the structured fields of the ANPA/IPTC data. This occurs due to the large amount of leeway that photographers have in filling out these fields, and because different providers use varying conventions

when assigning data to them. Predictably, then, usage of these archives is limited primarily to a small percentage of the potential users at the paper.

Administration

Typically, the library staff is left with the task of manually administering the archive and acting as a proxy for other members of the newspaper staff, both of which occupy a significant amount of their resources. If these system management tasks were more automated, librarians could be concentrating on applying their expertise to add value to the archive (e.g., adding content descriptors to the archived items).

Vision of future system

Functionality

In the future, the local archive at a newspaper will not only provide immediate and facile access to the paper's digital warehouse of data, it will also exist as a node on a larger, distributed system available to other newspapers and the public. This type of global system will cause the lines between information providers and consumers to blur. In addition to acting as a consumer of the major wire feeds or other centrally stored data, a local newspaper archive may become a provider to other papers, value-added services, or individuals. The following scenarios illustrate how different people might utilize this archive.

Scenario A: A member of the photo desk staff in Philadelphia is assigned the task of finding a number of potential photos or graphics to run with a breaking story on Bill Gates's resignation from Microsoft. He launches a small browser application on his Macintosh that can function as a client to his local archive. After entering his user name and password, he types "Bill Gates Microsoft" as his query and selects "General News" from a list of potential categories to search. Once he clicks on the "Search Now" button, he sees that the server is now working on his request. Within a few seconds, thumbnails of images begin to appear in his browser window.

Unfortunately, all the images in his browser are black and white photos, and he needs a color photo for the front page of the paper. So, he chooses an option in his browser application that will expand the search to the Reuters and AFP archives, as well as the local archive of the San Jose Mercury-News (since he knows of San Jose's expertise in the computer industry). This time, instead of typing a query, he simply drags one of the thumbnails to the search dialog in his browser. He soon receives a number of images in his browser window, ranked by the degree of commonality with his search image. He selects three of the thumbnails that he especially likes and drags them to a business card icon on his Mac's desktop that represents the e-mail address of the requesting editor.

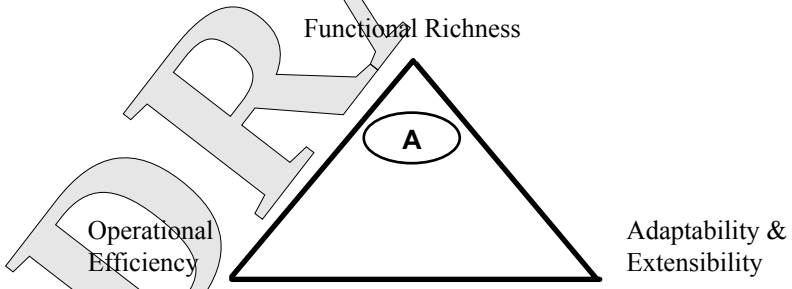
After a few moments, the editor notices a flashing mail icon on his Mac. He double-clicks on the mail message and views in his browser application the three thumbnails sent from the photo desk. He chooses the one he wants on the front page and drags it to the business card icon representing the page assembly person, who will in turn drag it directly into the Quark XPress page that he has been working on. A confirmation mail message

indicating which photo was chosen is automatically sent back to the photo desk.

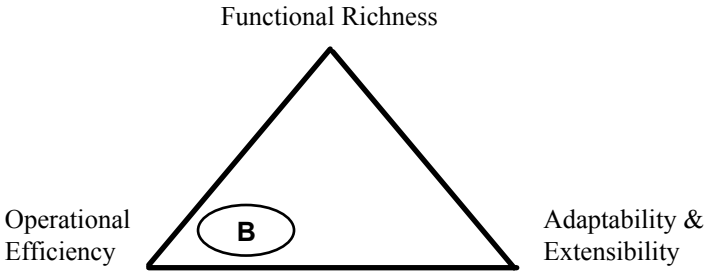
Scenario B: A researcher at Knight-Ridder's new multimedia company is assembling a CD-ROM on the Vietnam war. She brings up her browser application, selects several archives, and submits a query on "the Vietnam war". As thumbnails begin arriving in her browser, she starts to drag the interesting ones into her project window. These photos need to be resized to screen resolution, remove the dust and scratches, sharpened, and saved in PICT format. To accomplish this, she selects all the images in her project window and drags them to an icon representing a previously recorded PhotoFlash script. This script automatically performs all the functions mentioned above.

Optimization of Design Trade-Offs

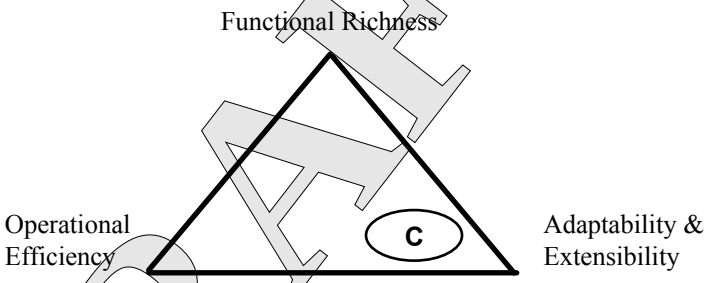
Historically, the three key trade-offs in information systems design – functional richness, operational efficiency, and adaptability – have not been able to be optimized simultaneously. These characteristics define the quality and effectiveness of the system, and they are sometimes depicted in a "trade-off triangle". We compare below the effects of maximizing one of the elements in the triangle at the expense of the other two.



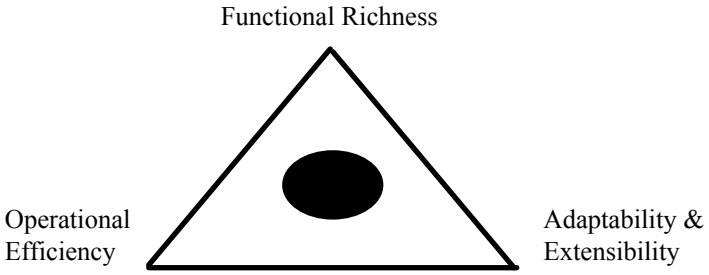
The richer the functionality of a system, the more complex it is to manage and the more difficult it is to adapt and extend in the future. Product A includes many of the features that newspaper archivists need and is packaged in a compelling user interface. Due to its design, however, it may prove to have a high cost of maintenance, and individual newspapers may not be able to adapt it to their specific environments.



Other solutions, such as Product B, aim for low operational costs, compromising functionality and adaptability in the process. Product B is constructed on proprietary hardware and software platforms in an effort to optimize operational efficiency. This custom-built foundation hinders modification and revision and thereby forfeits the usefulness and functionality of the system in the long run.



Product C is positioned as a more generic, and thus highly adaptable, image archiving solution. While these types of systems are initially easier to work with, they usually fail to maximize the power of the machines – they use more computer resources and use them less efficiently than necessary.



In order to achieve a more appropriate balance between these three trade-offs, the proposed system must utilize a different approach in its design and deployment. This approach must incorporate the following qualities:

Scalability - the ability to increase or decrease size or capability in cost-effective increments without software impact or "spikes" in the cost of business operations.

Platform portability - the ability to perform the same business function on different hardware and software platforms without requiring changes to other components of the system.

Distributability - the ability to divide software applications selectively into discrete component parts that are individually portable, yet operate together on the network as a single seamless application.

Sharability - the ability of a software module to be invoked as a utility without any modification. To be sharable, an application must publish its application program interface (API) that identifies the messages it must and can receive, the tasks it performs and how to request them, and the return messages that the requester must be capable of gracefully handling.

Securability - the ability to provide different access privileges to individuals based on the risk classification of data.

Cost-effective Implementation

Since our vision of the future newspaper archive is based on the principles of open systems, modularity, and distributability, the implementation and growth of the system can be driven by the economics of usage. Decisions such as the amount of processing power needed for clients/servers, the physical location of system components, and the location of data in the distributed environment of the global system will all be resolved by

choosing the option that best serves the current users and is most cost-effective.

For example, a newspaper may initially store all of its photos on a single centralized server which is available to all users. Later on, it might be more cost-effective to establish a distributed service for a specific subset of users that account for a high volume of usage on the system. This change in the location of data and handling of system traffic can be accomplished without affecting other components of the system.

Build vs. Buy

"The most efficient development is no development." The proposed architecture supports systems that can produce custom results by using off-the-shelf components wherever possible. Applications are assembled (as opposed to crafted) from these standard sharable software services. Many vendors can provide value-added components that plug-and-play in this environment, thereby encouraging specialization, competition, and innovation.

Opportunities for Added Value

The proposed system concentrates on laying a foundation for converting raw digital data into information – value-added output that is meaningful to humans. By contrast, today's archives do little more than store and manipulate data. People interact with information; they create it, capture it, change it, and use it. By utilizing an archive system that enables added-value products, Knight-Ridder can provide compelling information services to new users.

Benefits

By migrating photos to electronic archives, users will immediately see a number of benefits, such as:

- Universally available images on the LAN, and later on the WAN.
- Decreased retrieval time for an image.
- Eliminate loss through filing errors, etc.
- Photo reprint sales handled more efficiently.
- Reduced photo paper and chemical costs.

By adopting a modular, distributable architecture as the underpinning for the archive system, users gain additional benefits, such as:

- Users can choose the best tool for the job, and be assured of a high level of integration.
- Newspapers can leverage off existing computing resources and infrastructure, thus preserving their investments.
- New resources can be plugged into the system as they become available, thus accommodating the rapid advance of technology as well as changes in business.
- Configurable so that user can automate routine tasks by directly manipulating the user interface.
- Eventual fusion of the local archive with the global system.

Business requirements

Productization

A typical Phase I server system hardware configuration for a newspaper site should include the following:

- An Apple Workgroup Server 95 with 1 GB hard disk, CD-ROM drive, DAT tape backup system, and AppleShare Pro software
- 5 GB RAID-3 storage subsystem

The software will be retailed for less than $2,000 (10 concurrent users) and hardware costs will be less than $35,000 per server. So, for about $37,000 the newspaper's archive would have the capability to store 20,000 to 30,000 compressed photos. Additional storage can be purchased for as little as $2000/GB.

User documentation should be provided in electronic format on the server system (available to all clients). It must be searchable and printable on demand. It will include instructions on each feature of the software as well as task-oriented sections.

In addition , a Software Development Kit (SDK) with accompanying developer documentation should be created in order to enable software from different vendors to plug into the server side of the system.

The system must be easily localizable for international users.

Sale/Distribution

The software will be distributed as a CD-ROM disk, complete with installer, documentation, server and client software, sample photos with pre-constructed full-text indexes.

Installation

The installer software must automatically handle modularity issues (de-installing pieces when necessary, etc.). Clients should be able to install over the network by simply connecting to the server and executing the installer program.

Support

The support needs of the newspaper sites need to be carefully evaluated during the beta seed program and early roll-out of the product. It is likely that some organization with the requisite technical skills will need to be designated as the nation-wide support contact for the system. Support functions should include telephone hot lines, on-site maintenance, and customized development for special needs.

Competitive products

AP Leaf Preserver

AXS News Photo Archive (discontinued)

Aldus Fetch

Multi-Ad Search

Canto Cumulus

Phrasea PhotoBase

Technical requirements

Hardware compatibility

The server system must function on any '040 Macintosh running System 7.1 or above, as well as all configurations of PowerPC. The server machine should also include a CD-ROM drive, tape backup system, and enough redundant hard disk capacity for a minimum of 12 months of photos and graphics (5 - 10 GB).

The client software must function on any System 7 Macintosh (later, Windows, OS/2).

Client software should be able to run on displays at any bit depth.

Software compatibility

The server and client software must be "32-bit clean" and compatible with QuickTime 1.6 or above.

Intended users

Newspaper staff

Photo providers such as KRT

Stock photo agencies

Advertising agencies

Graphics and imaging professionals

Others

Modular components

Modular system so that different interfaces can be used for different functions if desired.

Local archive components can come from any vendor as long as it adheres to the published standards.

Components are scriptable and recordable so that users can configure how they interact with each other.

User-reconfigurable shared services library. Users must have a way to easily open images in a variety of tools and/or perform automated operations on the images - all directly from the browser. This library (a palette of icons representing tools or scripts) should be completely configurable by the user at any time.

Document Types

Many different data formats need to be stored and indexed by the archive in addition to the JFIF format (which includes JPEG image, thumbnail, and ANPA/IPTC metadata). Examples of other formats include: ASCII text, PageMaker, Multi-Ad Creator, Illustrator, PICT, TIFF, QuickTime video/sound, Quark XPress, and Freehand. In general, adding new data formats to the archive should only require the addition of a small, inexpensive component on the server machine and should not affect any other modules (client or server) in the system.

Specific Human Interface needs

Direct manipulation and tight integration with operating system and other applications (drag-and-drop).

Highly efficient, extremely intuitive user interface.

Ability to search the full textual content of any item in the archive by using Boolean operators as well as free-form or fielded queries.

Users should not need to know how the textual content of items has been stored (e.g., what content is stored in fields, what conventions are used as abbreviations in the content, etc.)

Simultaneous searching over any or all directories in the archive.

Ability to create a scheduled search which is executed by the server and the results of which are returned to the client automatically.

Ability to view the items found from a search in a thumbnail browser or a textual list. The items should be sortable by relevance, date, or alphabetically by file name.

Facility to dynamically edit the content of an object in the archive.

Ability to handle groups or batches of information as easily as single files.

Disconnected or remote clients should support all the functionality of the system with the exception of live searching.

Context-sensitive help available where appropriate.

Specific human interface illustrations (screen shots) can be found in the accompanying design document.

Specific API needs

Client components on the Macintosh should implement the AppleEvents Object Model and AppleEvents suites (where appropriate) in order to facilitate inter-application communication.

Client components on the Macintosh should implement the Drag Manager.

Adherence to standards

Metadata (IPTC/ANPA) is bound into the image file. Tools must be available to deal with this metadata in an intelligent and distributed manner.

Compression/decompression will be supported under the JPEG standard.

Emerging full-text query standards such as Z39.50 and SFQL should be continually monitored and supported when feasible.

Interfaces to other systems

Communication via AppleTalk networks (later, TCP/IP).

Integration with OPI output systems.

Integration of 3 major wire feeds directly into archive.

Users must be able to transfer image files from the archive to other systems (e.g., Leaf Desk, Scitex) and back again without losing any metadata information.

Seamless access to local archive, remote archives (via a point-to-point connection) and a global system.

Integration with text archives such as VuText or SII.

The ultimate goal is a unified digital media archive that handles a variety of data types: photos, graphics, pages, ads, text. This would enable clients to view information as it ran in the paper.

Performance

Completely asynchronous in response to user events (i.e., the user can interrupt searches, screen redraws, or any other operation).

Full-text retrieval times will vary depending on current number of simultaneous users, number of items in the archive, length/complexity of the query, number of items found satisfying the query. Retrieval times should, however, stay below 10 seconds for "typical" searches (queries of 1-3 terms returning hit lists of 50 items or less).

Performance levels are maintained with up to 50 simultaneous users on the system.

Security

Access privileges for all users fully configurable from the server.

Access for any directory on the server can be specified as read/write or read-only.

Administration

Simplified posting to local archive (drop box).

Automated hot backup with restoration process.

Automated scheme to age information to off-line media (CD-ROM jukeboxes).

Automated, event-driven tools to manage archive subsets (file system utilities).

Management of access privileges for users and groups, add/remove users and groups.

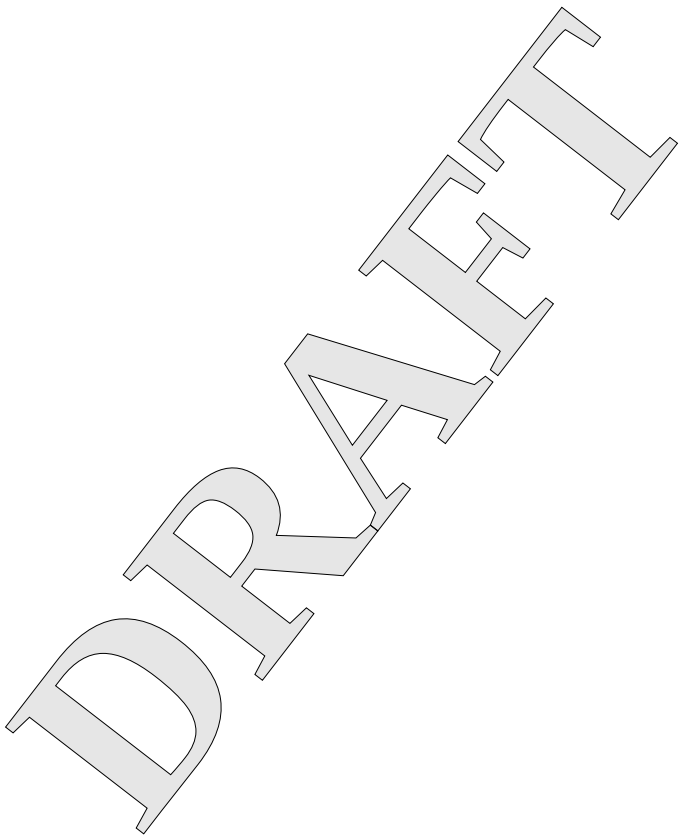
Real-time, automated, incremental indexing of the database with auto-compaction.

Size

Each component on the client machine (with the exception of the photo editing module) should be executable in a low memory partition (less than 1 MB).

Each component on the client machine should require less than 1 MB of disk storage.

Software on the server should be executable in an 8 MB partition.



Phased Implementation

Below is a summary of the preliminary phased implementation plan for the system. Reference the accompanying design document for a detailed description of the system's components and their methods of interaction.

Phase I

In the first phase of the project, to be completed by the end of 1993, several off-the-shelf packages will be integrated on the Macintosh platform using AppleSearch to give users a full-text retrieval system for accessing photos in the local archive.

On the server, AppleShare Pro and AppleSearch 1.0 will provide a secure, fully indexed archive which can be accessed by System 7 Macintosh clients on any AppleTalk network.

On the client, users can combine several different tools: photo editing packages (Adobe PhotoShop, Storm PhotoFlash, IronMike plug-ins), page layout packages (Quark), and the IronMike Browser. From the browser, users will be able to request both real-time and scheduled searches on the server. Once images are viewed in the browser, they can be dragged (singly or in groups) to a "service" on palette of icons. This will trigger the appropriate application to open the image.

In addition, "transnormalizer" software will be available to convert the major wire feeds into a normalized form (including JPEG image, thumbnail, and ANPA/IPTC metadata), which can be easily indexed on the server and viewed in the client applications.

Phase II

In Phase II, the server platform will change to PowerPC (expected to ship in March, 1994). This will boost the performance of the server and allow more flexibility in the types of services that can be provided. Windows clients will also be supported at this time.

Additionally, AOCE services will be integrated into the client software to provide mail/messaging and enable workflow configuration.

Word services will be available from the server for functions such as spell-checking or synonym generation (during the authoring of image metadata).

Phase III

Phase III will fuse the local archive with a global network to allow a more distributed system discussed in the Vision section earlier. At this point, it may be more economical to store the major wire feeds centrally on the network or at the information provider's own site, instead of at each newspaper.

Acknowledgements

In addition to interviews with the persons mentioned on page 2-3, information from the following sources was used in creating this document:

- *Vital Fundamentals*, Apple Computer, Inc.
- *The Delivery and Distribution of Information in a Client/Server World*, a report by Arthur D. Little, Inc.
- *Electronic Commerce in the Media Industry*, a report by INPUT