Optical Publishing: Data Conversion/Preparation for CD-ROM Applications

Jeff Szafransky
Hewlett-Packard
Application Support Division
Mountain View, California

INTRODUCTION

Optical publishing and storage technologies, especially CD-ROM, are changing the way information is being distributed and stored. New commercial CD-ROM applications are being introduced each month, and many internal CD-ROM publishing activities are taking place. Hewlett-Packard is involved with optical publishing and CD-ROM with its HP LaserROM support information subscription service.

Many, when first exposed to this technology, inquire as to the cost of the plastic disc itself. After discovering that the disc is relatively inexpensive, they question the cost of the CD-ROM application. This line of reasoning does not address the underlying effort and expense involved in the production of a CD-ROM - data conversion and preparation.

Taking data and information from its existing format to CD-ROM is a complex multi-step process requiring significant expertise in a number of wide ranging fields. Careful consideration must be given to the option of doing some or all of the data conversion/preparation yourself versus contracting with outside service bureaus. In this paper I will describe the necessary data conversion and preparation steps involved in CD-ROM production.

DECIDING ON A CD-ROM APPLICATION

There are many types of information and many ways of distributing it.

New CD-ROM applications are constantly being introduced, and there are many future applications that have not even been thought of yet. A list of some of the current applications appears below. This list is by no means exhaustive, especially when you consider some of the newer multimedia applications being developed.

- Reference Manuals
- Catalogs
- Training Materials
- Product Demos
- Software Distribution
- Directories
- Financial Databases
- Business Archives
- Census Data
- Systems Documentation
- Medical Data
- Dictionaries and other References
- Bibliographic Information

Current distribution methods include paper, on-line services, microfiche, microfilm, magnetic media (floppies and tapes), and the spoken word. An example of the latter would be the traditional training environment with a live instructor.

Selecting CD-ROM to distribute certain types of information involves the analysis of many factors. Needless to say, not all products should be moved to CD-ROM. In making your decision to go with CD-ROM you need to evaluate:

Availability of the Data
Cost for Data Conversion
Storage Capacity Required
Media Costs
Storage Costs
Distribution Costs (CD-ROM vs. On-line, Postage, etc.)
Data Accessibility Requirements
Frequency of Updates
Frequency of Access
Population of target audience
"Market" Demand
Availability of Hardware (PC, CD-ROM Drive, etc.)
Increased usefulness of having data on CD-ROM

SELECTING A DELIVERY SYSTEM

Most CD-ROM workstations are personal computer based with either internal or external CD-ROM drives and a printer. Some systems require a mouse, high resolution graphics displays, and even speakers for audio support. Software includes CD-ROM device drivers and the retrieval software including any necessary decompression or decryption routines. The delivery system requirements are dictated by the application and retrieval software used to access and present the vast amount of data stored on CD-ROM.

Several options are available when deciding on the retrieval software and user interface for the CD-ROM application. You may develop the entire system in-house; develop part of the system in-house using a developers' toolkit from an outside firm; purchase an off-the-shelf system from a vendor; or contract with an outside company for the development of a customized user interface for your application.

Common capabilities of today's CD-ROM retrieval software include windowing, full-text keyword retrieval, browsing, cross references, display and print of both text and graphics, and full help facilities.

See the paper "User Interface Design Methodologies for CD-ROM Information Retrieval" (1) elsewhere in these proceedings for information about the design and development of the HP LaserROM user interface.

DATA PREPARATION OVERVIEW

With an application identified and a user interface selected we may now address the data preparation activities involved with CD-ROM production. See Figure 1 for a high-level graphical representation of the CD-ROM production process. The remainder of this paper will describe each of the data preparation steps in detail.

There are several data preparation companies who are able to perform all of the conversion/preparation steps and provide you with CD-ROM discs. Your expertise, project schedule, staff, resources, and the use of existing tools will need to be evaluated when deciding on the direction to take with the data preparation for your CD-ROM application.

DATA CONVERSION

Data preparation is the most lime-intensive and expensive aspect of CD-ROM publishing. In data preparation the one step that has the potential for accounting for the majority of time and effort is data conversion. Data conversion involves the conversion of text and graphics into a format that is compatible with the retrieval and indexing software.

The data conversion efforts required are dependent on a number of things, including:

- CD-ROM production system data input format
- If the data is currently machine-readable or not
- Format(s) of data that is machine-readable
- Existence of conversion tools

Data Input Format

To allow for efficient and effective processing of all types of data to be offered on the CD-ROM, it is necessary to have all of the data enter into the production process in a consistent format.

Given the retrieval and indexing software used in HP LaserROM, we defined a CD-ROM production system data input format. This format consists of ASCII text files that also provide information to the indexing software about the structure of the data so that the correct fields are indexed and the associated control files are built properly. In addition, the text files needed to contain the codes that tell the retrieval/display software how to format the document for display.

We defined a specific graphic file format for the system. We chose the TIFF (Tagged Image File Format) (2) for HP LaserROM graphics. We also specified resolution of the figures and files sizes.

Machine-Readable Format

You can only deliver data on CD-ROM if the data is machine-readable. So, the first step in the conversion process is to make sure all of the data to be placed on the disc is machine-readable.

Text and graphics are handled in similar ways. They both may be scanned or recreated.

Textual data that is in hardcopy only form is scanned in with the use of Optical Character Recognition (OCR) software creating a straight ASCII file. An alternative to the scanning method is to actually re-key the information into a computer system. One reason re-keying may be used over scanning is if the only hardcopy available is of such poor quality that individual characters cannot be recognized by the OCR software.

Graphic images may also be uplifted from paper copies using image scanners such as Hewlett-Packard's Scanjet. Illegible graphics or those that require better resolution may have to be recreated using a suitable graphics package.

Format of Machine-Readable Data

Assuming that we have all of our data in machine readable format we now must look at the differences between the data formats and the input data format required by the CD-ROM production system. It is quite probable that most, if not all, of the data will need to be converted to the CD-ROM production system input format.

There are varying levels of complexity when it comes to data conversion.

If we refer back to our data input format for text we see that we need to know something of the structure of the data. This allows us to quite easily identify those types of data which will be easier or harder to convert. As a general rule, the more structured the data or the more information about the structure the data carries with it, then the easier it will be to convert to the CD-ROM production system input format.

For example, databases are much easier to convert because of their inherent structure. Programs can be written to extract the appropriate data from the database, insert the codes required by the indexing/retrieval software, and create a file that can be used by the CD-ROM production system. The Software Status Bulletin and Product Catalog data in HP LaserROM originated from IMAGE databases, and we used this programmatic extract-and-format approach. On the other hand, reference manuals and other free-format information presents a large challenge to convert to the right format. The problem arises from not being able to identify specific structures within the text that are required by the indexing and retrieval software.

Structure within text refers to constructs such as document name, chapter or section title, publication date, author, or glossary term. Most documentation formatting languages, such as TDP, are concerned with the format of the information rather than the specific structure or components that make up the document. And, the straight ASCII files, which were created as a result of the scanning process described above, contain no structural information at all. In fact, a majority of the formatting (i.e. font) information is lost in this case.

There is, however, a fairly recent development within the publishing industry and Hewlett-Packard documentation development groups to move towards a descriptive markup of electronic documentation. These powerful markup languages, defined by ISO 8879: Standard Generalized Markup Language (SGML) (3), can be used to organize books and articles by identifying necessary structural information. This structural information can be used by various forms of electronic publishing to produce either a hardcopy product or a CD-ROM.

We have developed the HP LaserROM CD-ROM production system to accept SGML-coded documentation as an input format. This documentation is then converted into the standard data input format at the front end of the data preparation process.

The advantages of this approach are many. A single documentation format may be used for both print and CD-ROM. This reduces maintenance and rework, and is analogous to the reuseable software philosophy currently being promoted in the software engineering community.

Since SGML is a fairly new development, the majority of the installed base of reference manuals selected to appear on HP LaserROM is in other formatting languages. The decision was made to convert this documentation into SGML rather than directly into the CD-ROM production system data input format. Since most of this documentation is related to supported products and is currently being maintained, it makes sense to take advantage of the multiple format publishing possibilities of SGML.

The documentation conversion to SGML is a complex issue given the state of the current documentation and the number of formatting languages currently in use within Hewlett-Packard. Some formats are easier to convert than others because there are some formatting languages which better identify the document structure. This allows for at least partial programmatic conversion to SGML. Considerable amounts of hand conversions by skilled converters take place for the worst case formats and for the final tweaking of programmaticlly converted documentation.

These conversions require the expertise of both software engineers and documentation specialists.

Many of the same conversion issues apply to graphics information as well. There are many graphics formats in use today within the documentation development community at Hewlett-Packard. Formats include HP DRAW, Graphics Gallery, EGS/9000, and paste-up artwork of various formats, to name but a few.

Graphic images scanned in using the HP Scanjet produce TIFF files directly, so these images require no conversion. Other scanners produce various formats which require custom programmatic conversion to the necessary TIFF format. Conversion routines for several of the machine-readable formats have also been developed.

As with text, the differing formats and operating environments (e.g. HP3000, HP9000, Vectra) add to the complexity of the conversion effort and accentuate the multi-discipline expertise required.

Fortunately, there is some data that will be placed on the CD-ROM which requires no data conversion. Data referenced but not presented with the text and graphics need only be transferred to an MS-DOS compatible file. Examples of this type of data include software and non-indexed files. This usually represents a small percentage of the overall data to be placed on the disc, however.

As daunting as the data conversion step looks, it is a necessary first step in the data preparation process. Two fairly recent developments in the area of data conversion can be directly attributed to the growth in optical publishing on CD-ROM. The first is the expansion of existing companies into the data conversion business as well as new entities whose sole purpose is to perform data conversions. The second is the increase in the number of data conversion tools being offered by a host of companies.

INDEXING

Given the amount of data a CD-ROM disc is capable of holding (approximately 600 megabytes), serious consideration must be given to how the user will effectively retrieve the information needed. Performing a serial scan through all of the data would take far too long, so a better way to access the data must be employed. HP LaserROM make use of full-text retrieval technology to search through documents.

Full-text retrieval programs developed for CD-ROM applications use inverted indexes to search for data. The indexing software within the CD-ROM production system creates these indexes. The retrieval system index identifies the location of the information on the disc similar to the way a book index lists the location of information in the book.

The indexing method and the retrieval software are integrally connected because the retrieval software must be able to read and act on the index created during data preparation. Both the retrieval software and indexing method are strongly influenced by the type of data and application. For instance, structured databases and full-text applications are indexed in a different way. The primary difference relates to the multiple fields by which the user wishes to access the data in a database. Full-text applications like HP LaserROM require an indexing method that records the location of every word in every document. This allows the retrieval/display software to highlight the search word when displaying the document, as well as provides the information necessary to perform phrase and proximity searching.

Other components of the document such as chapter titles, authors, publication dates, and the like may also be indexed. Cross-reference indexes are also created in this step.

The indexing step is complicated by the fact that there are parts of the data that do not require indexing. Display enhancement codes are an example of this type of data. The indexing must be capable of being turned on and off automatically during the processing of the data.

Another example where the indexing is temporarily inhibited is for words that you do not want to index, also known as stopwords. Typically, they are (1) articles and prepositions (a, the, of, for, etc.) or other words that add no value to the search or (2) words that appear in so many of the documents that a search on one of these words does nothing to help narrow down the number of qualifying documents. These stopwords are entered into the indexing process via an excluded word list/stoplist file.

Full-text retrieval and its associated indexes require a great deal of disc space. The overhead for certain types of data can be as much as 100 percent. The indexing step, though automated, can also take considerable time depending on the processor used. Despite this, full-text retrieval is the best way to access large amounts of unstructured text and structured database data. So, the indexing must go on!

COMPRESSION/ENCRYPTION

Data compression and encryption of the data on CD-ROM is another very large area which I will only touch on briefly here. I will not go into all the possible methods of data compression or encryption, just some of the reasons for performing these activities and issues involved.

The 600 megabyte storage capacity of today's CD-ROM seems almost infinite to most users, and the technology is progressing so as to make even this figure seem small. But, if you consider that a single graphic image may require more than a megabyte, and that the overhead of indexes could be as much as 100 percent, then you begin to think of how to save some space so you can get your entire application on one disc! This is where data compression enters the picture.

Compression is a process that converts data into a form that requires less space. There are numerous compression schemes for both text and graphics, and these compression schemes may be performed in either hardware or software. Compressed data by definition requires decompression at some time, and this can also be handled in hardware or software.

Information distribution using CD-ROM eliminates the data vulnerability due to unauthorized access over networks and telecommunications systems, but some protection may still be required to protect the data on the CD-ROM disc from unauthorized use. CD-ROM applications can involve highly sensitive data such as defense information or internal company documentation. Or a company may just want to make it harder for someone to do a mass-downloading of all of the information on the CD-ROM. Data encryption is one method of securing data on CD-ROM. Many of the issues involving implementation of data encryption for CD-ROM are the same as those for data compression.

Issues to be considered when evaluating compression or encryption for your CD-ROM application include:

- Is it necessary? (Do you require the extra space or the data security?)
- Is the data a good choice for compression? (How much space will you save?)
- Does the indexing software support the compressed/encrypted data?
- What is the impact on the CD-ROM production system?
- Does the retrieval software support the compressed/encrypted data?
- Is special software or hardware required?
- What is the effect on the retrieval system performance?

You must also evaluate the various compression and encryption methods and select the method that performs best with your application.

DATA LAYOUT

At this point in the CD-ROM production process all of the files are ready and in the form required by the delivery system: data is formatted, indexes are built, graphics are scanned, the data may be compressed and/or encrypted, and all control and support files are present. Data layout is the process of arranging your files and directories in the exact order and location as they will appear on the CD-ROM. The result is known as the disc image, and the final CD-ROM will precisely reflect its structure and content.

The layout of the data onto the disc image usually only needs to be done once, and then each subsequent disc production follows the same layout script. The layout of the data depends on the application, and the primary factors are access time and overall performance of the delivery system. Due to the relatively slow access times of the CD-ROM drive it is necessary to pay particularly close attention to the placement of files on the disc. Generally speaking, files that are accessed together often should be placed in close proximity to each other. Knowing in detail the specific operations of the delivery system allows for a knowledgeable data layout.

Data layout involves the following operations:

1. Collect all files to be placed on the CD-ROM
2. Determine disc directory structure and which files reside in which directory
3. Create the necessary directories
4. Transfer files into their appropriate directories in the disc image

The underlying logical format of most CD-ROM discs is the High Sierra Format or the ISO 9660 format, which is the international standard developed from the High Sierra Format. Most commercially available data prep systems and virtually all service bureaus support these formats.

EMULATION

Once the disc image is made and while the data is still residing on the winchester disc, it is necessary to test the data integrity and simulate how the data will operate with the delivery system. Checking the data integrity involves activities such as making sure all files exist, are the right size, and that they can be opened/executed.

Systems exist that allow for fairly rigorous emulation (or simulation) of the CD-ROM application using the disc image. Some of these systems go so far as to exactly duplicate the seek and read times of many of the CD-ROM drives currently available. This system emulation is the only satisfactory way of verifying that everything has come together correctly and that the application will work once it is placed on the CD-ROM. Waiting until the CD-ROM is created is a risky and potentially expensive alternative.

Checking the results of the emulation may reveal problems with the disc image. Files may be missing or in the wrong location; files may be in an incorrect format; a problem with one of the previous data preparation steps may be uncovered; or the layout of the data may need to be modified.

If a change is required, then we need to go back to the appropriate data preparation step and fix the problem, work through the subsequent steps, and then the emulation is performed again. If the emulation is successful, then we proceed to the next step in the process: premaster tape generation.

The final step in the formal CD-ROM production system is the generation of a premaster tape. The disc image is copied to an ANSI labelled 9-track tape according to one of the CD-ROM mastering plant's input specifications.

PREMASTERING/MASTERING

This section, while technically not part of the data preparation process, is included for completeness. I want to briefly describe the steps involved in the actual CD-ROM disc production.

Very few companies will actually produce their own CD-ROM's. The technology and economics make it much more feasible to go with one of the existing mastering plants.

When the 9-track tape arrives at the mastering facility it is scanned for readability and conformance to format specifications. Mastering facilities do not check logical format, disc layout, or disc directories. This is the responsibility of the group producing the premaster tapes. Whatever data is on the tapes will be mastered to the CD-ROM.

Premastering is the process where the data is transferred from tape to hard disc and the error detection and error correction codes are added to the data. Header information and sync bytes are also added to the data blocks.

Mastering is the process where a master disc (usually made of glass) is produced. The data bits are etched onto the glass using a laser and are represented by pits and lands. A negative image of this disc (usually made of metal) is produced and this becomes the "stamper" which is used to create the CD-ROM discs.

Compact discs are made of clear polycarbonate plastic. Replication of the discs is performed by an injection molding process using the metal stamper to emboss the data patterns into the plastic. Reflective silvering is added covered by a protective plastic coating, and then labels are printed on the discs. Various amounts of packaging may be added at the end of the process. This usually includes the loading of the CD-ROM into either a jewel box or CD-ROM caddy.

The quality control processes within the CD-ROM mastering facilities is quite impressive. Some mastering facilities are so confident about their quality control processes that they guarantee the CD-ROM discs you get will exactly match the data you sent them on tape. Statistics show that a defective CD-ROM is usually the result of incorrect data on the premaster tapes. This highlights the necessity for ensuring that the data sent to the mastering plant is 100 percent accurate.

CONCLUSION

The intent of this paper was to give an overview of what is involved in the creation of a CD-ROM application with regards to data conversion and data preparation. All of the data preparation steps described in this article may be performed either internally or through an ever growing number of service bureaus.

Most CD-ROM software developers are publishing information directly or providing contract development services. In either case the organization most knowledgeable about the customer and the data (the information provider) gives up control of product development and sometimes marketing. As CD-ROM technology has gained acceptance, organizations are seeking to gain control of the development process internally. Providing information managers with their own internal CD-ROM publishing capability enables them to tailor the technology to their business and customers for maximum competitive advantage. Through careful analysis of your application, expertise, schedules, and resources you will be in the best position to determine how to implement an optical publishing solution within your company.

But, let us get back to the original question: "How much does that disc cost?". I would have to answer with something like the following.

"The cost per disc is determined by the following costs:

Data Collection
Data Conversion
Data Preparation
Premastering
Mastering
Disk Duplication
Packaging

plus

Retrieval Software (sometimes charged per disc or CPU)
Documentation
Technical Support
Marketing
Copyright Fees
Licensing Fees

not to mention the cost of development for the retrieval software and the data preparation system, hardware for the data preparation system, ... "

Bibliography

1. Ferguson, Greg. "User Interface Design Methodologies for CD-ROM Information Retrieval" Hewlett-Packard. Presented at 1988 International HP Users Group Conference, May, 1988.

2. Andrews, Nancy and Fry, Stan. "TIFF: An Emerging Standard for Exchanging Digitized Graphics Images". Microsoft Systems Journal. July, 1987.

3. International Organization for Standardization (ISO). ISO 8879-1986(E). Information Processing - Text and Office Systems - Standard Generalized Markup Language (SGML). International Organization for Standardization, Switzerland, 1986.

Other Useful References

Lambert, Steve and Ropiequet, Suzanne (ed.). CD-ROM: The New Papyrus. Microsoft Press, Redmond, Wash., 1987

Ropiequet, Suzanne (ed.), Einberger, John and Zoellieck, Bill. CD-ROM: Optical Publishing. Microsoft Press, Redmond, Wash., 1987

Roth, Judith Paris (ed.). Essential Guide to CD-ROM. Meckler Publishing, Westport, Conn., 1986.

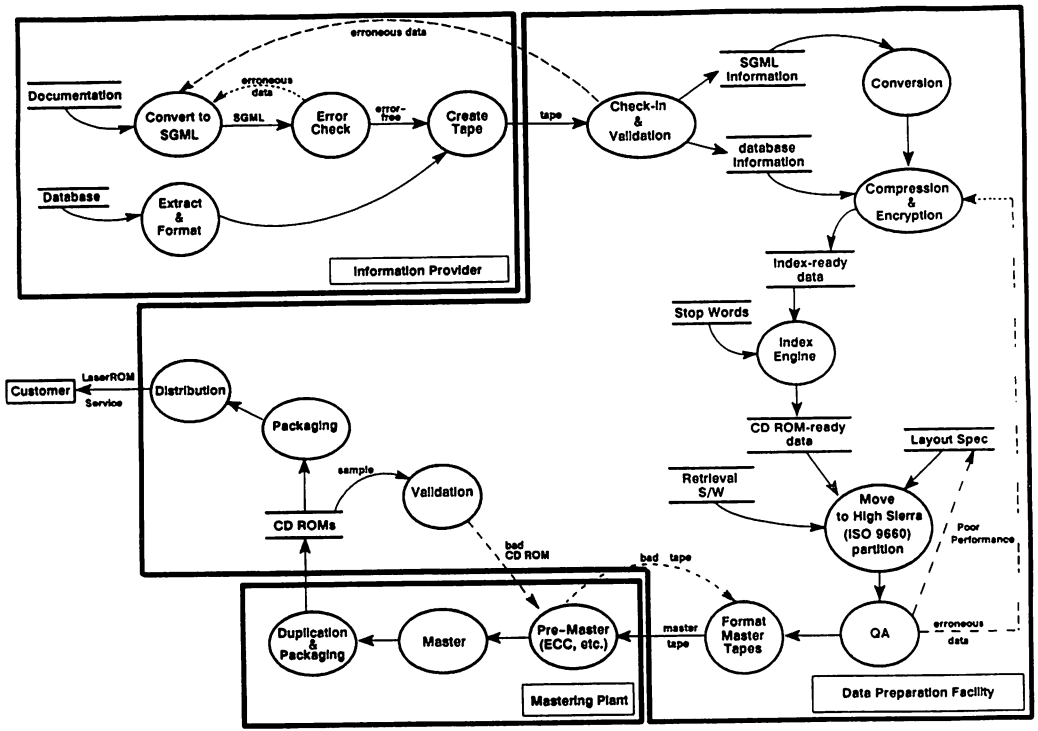

Figure 1 — CD ROM Production Process