

An Automatic Web Publishing Package for Complex Data Sets

Steve Nickerson
Nickerson & Associates Designed Building Systems Ltd
Ottawa, Ontario, Canada K1Y 3Z4
e-mail: steve@icomos.org

CIPA XVII Symposium - Recife, Brazil - October 1999
CIPA Working Group 6

KEYWORDS

As Found Recording, Information Systems, Database, Archive, HTML

ABSTRACT

The computer revolution has made it possible for Heritage Recorders and Conservationists to gather far more data about a resource, in a much shorter time, than was previously possible. While this is undoubtedly a "good thing" the problem now is how to turn all these disparate data types (photos, sketches, notes, survey data, CAD drawings, database tables, word processing files, etc.) into usable information.

There is a fundamental dichotomy between the data of the "gatherer" and that of the "presenter". From the perspective of a system analyst it might be described as the difference between having a little information for a large audience and having a lot of information for a small one. A museum or publication would fall in the former category, distinguished by the fact that *much of the data has been removed for the presentation*, while the Heritage Recorder occupies the other side of this equation *and must keep everything*.

A Heritage Record contains a vast amount of data that is of interest to only a very small, but important, audience which includes the conservators charged with maintaining the asset, those who may eventually present it to a larger audience and "posterity".

As the data set grows and resources dwindle it has become necessary to create tools to facilitate the organization and analysis of this material to make it as easy as possible for recorders to collect quality data in the first place and to quickly identify areas of the record that may be deficient.

The CARTHTML Publisher is such a tool. It forces the recorders to use a consistent data structure and in return automatically prepares, HTML pages which present the material in an easy to understand format which is accessible to all computers and uses no proprietary software.

This paper will demonstrate this new product, discuss the data structures required to support it and show examples of how it has been used in support of archaeology, art history, architectural and industrial recording, and in preparing the documentation of procedures and training courses.

The Issues

When people speak of a "database" they usually have in mind a computer program of that type whereas the definition in the dictionary is as follows:

database

1. a comprehensive collection of related data organized for convenient access, generally in a computer.
2. See data bank

data bank

1. a fund of information on a particular subject or group of related subjects, usually stored in and used via a computer system.

ie. It is the data, not the software that presents it that is "the database", and it is distinguished by its being "organized for convenient access".

When data is organized by a database program access can be very convenient indeed, but generally this is the case only for those fluent in the program's structure and commands. Those less comfortable with the system often find it frustrating and give up or learn to perform only a few

specific queries. Of course, the material is completely inaccessible to those without the program (*often including those with a different version of the program, a different computer operating system, etc.*).

While the advantages of a database program are generally well worth the learning curve for anyone who will be studying the material in detail such a format is inappropriate as the final resting place for data concerning any resource of historical import. Such material needs to be accessible, not only to as wide an audience as possible but for a very long time which is a much more difficult problem. Using electronic media we probably have very little hope of our work being accessible even 100 years in the future while we base our research on documents and artifacts many times that old.

Data Archiving

We are not going to start carving anything in stone (life expectancy millennia) and probably not much will be put on vellum (life expectancy centuries), or even computer paper (life expectancy decades). However we do can do quite a

bit better than a useful life to be terminated with the next release (or demise) of a database program or operating system.

Suppose that future researchers find our database in an archive somewhere. If it is in a single file (or set of files) from a database program there is very little hope of recreating the data without both the originating program and a computer that can run it. Suppose, on the other hand, that they find our source material - the image files, text files or even simple CAD or database files - there will be a much better chance that they will be able to decipher the contents. There may be thousands of these files but each is a relatively simple problem to convert to whatever systems they may be using at the time and, if one further supposes that the names of all these files conformed to some logical structure it begins to seem possible that some use might be made of the archive.

It would not hurt at all if there was a README file describing the material and its organization and even the printed reports would help some (if the paper still held together and the toner had not separated), but this stuff will never be as complete as the source material. It is, in all probability, the logic itself that would provide the key enabling these future scholars to understand our work and the monuments we were studying.

Data Collection

By far the most difficult task facing those who would protect our heritage monuments is the gathering the initial data, extracting the information from the database for the conservators to use is trivial in comparison. However, because this latter task is already well served by software developed for other purposes we have tended to adopt the existing tools for data access forcing our material into their structures as best we could.

If our big problem is at the stage of data collection it stands to reason that we should optimize our data standards for the collection phase, converting things later, if necessary, to facilitate access. As it happens, the data structure hypothesized above as being appropriate for a long term archive is exactly what we get at the recording stage - a lot of little files in a variety of formats produced by a different authors working at different times and with differing agendas.

The difficulty for both the recorder and the archivist is the same, it is how to organize this material. Essentially the recorder needs to have a place to put it and the archivist needs to know, or be able to figure out, where that was.

Again, optimizing for the recorder, we need to define these structures in advance. It makes no sense and produces a scattered data set to leave this critical step to be tackled, ad hoc, by the people in the field. Field time is precious and once something has been captured it needs to be filed away so the recorders can get on with the next task. Of course, they will need to be able to find it again and nothing is so lost as the item wrongly filed so a clear framework MUST be established, hopefully well in advance, to minimize both the time lost making decisions and the errors inherent in making such spur-of-the-moment judgements.

The Object and File Name

All computer files have a name and all those names start with a string of one or more letters, or numbers, usually followed by an extension. The extensions are generally preempted by the software but even so what is left can easily be turned into an extensive data organization system with just a little effort and foresight.

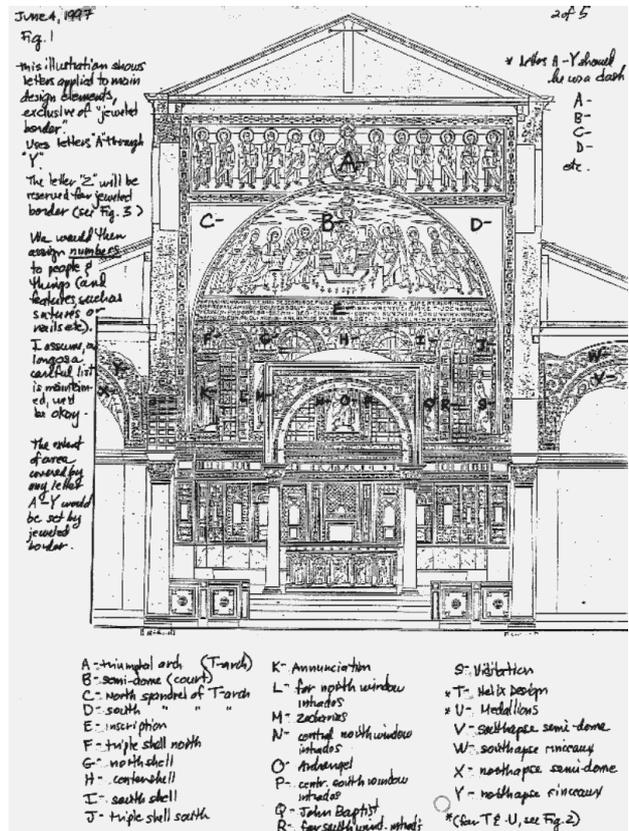


Figure 1 Foresight might look something like this

This exercise, which took the director of the project "a couple of hours" in the comfort of her office probably saved several days in the field. What it does is give names to the objects of study, in this instance the mosaics of the apse of the Eufrosiana Basilica in Croatia. The basic categories laid out in this sketch were expanded to give us unique names for areas as small as a few centimetres square and whether the files were images, sketches, CAD drawings or notes was dealt with by the extension.

The first level of data access, the minds of the users, was also well served because the naming scheme quickly became embedded in everyone's consciousness because they needed a way to communicate with each other and using these pre-established codes provided the path of least resistance. A common language was quickly established and each member of the team could, with a logically derived eight digit code¹, accurately describe the

¹ Only about 30 1-2 digit codes had to be "learned" the next 4 digits defined the location of an area within the whole and the last 2 digits were used for things like multiple images and version distinctions.

location of a feature or quickly find out what information had been collected for any given area.

The mind however quickly becomes overwhelmed by the quantity of the material produced by a modern recording team so keeping track of the overall state of the campaign required some additional tools.

The Data Audit

Our mandate was to measure and map the dome and provide a complete digital photographic coverage at low resolution. Essentially an index that could be used for all work past present and future. This alone would produce a lot of files and, we knew, many more would come in the years ahead so it was vitally important that our part of this effort was not only well organized but complete and of a consistent quality.

Our digital imagery alone occupied 100 megabytes (800 files) and, as was inevitable, there were typos as filenames were entered that rendered files we knew we had inaccessible. As well some scholarly work was being carried out at the same time so many elements would have notes and sketches that needed to be attached, sometimes prepared by people not so intimately acquainted with our naming philosophy. On top of that there was film being used by these people that would have to be kept track of and integrated later.

This was going to be a large and ongoing problem and our solution for it was the "publisher".

The CARTHTML Publisher

What we have developed is a small program that you provide with:

- A directory (including it's sub-directories) where the source material resides.
- A second directory where you want your HTML pages to go.
- A list of text strings that (might) begin file names or file extensions which you want grouped together.
- A name by which the whole set will be known.
- And a bunch of optional stuff that provides subtlety and flexibility.

What it does with the files it finds is determined primarily by the extension and secondarily by configuration files.

- Files with the extension .TXT (and other extensions specified by the user) it turns into HTML.
- For .GIF and .JPG it creates thumbnails and copies of the files (other image types it converts to GIF).
- Database (DBF) files, and queries thereof, it turns into HTML tables.
- It creates links to things like PDF or HTML files.
- It requires that you save CAD, Word Processor and Spreadsheet files in formats that it can understand - DWF, CSV and ASCII text respectively but it will remind you if you have not done so.
- Unknown file types it tries to insert as preformatted text so you will, at least see it and be reminded to deal with it.

- It can also filter out things like backups or other files or directories you specifically do not want to include in the output.

When the software is run it will try to make sense of your data in terms of the rules you have stated and how successful it will be depend on how rigorous your file naming is.

An easy to visualize example would be to send the entire alphabet as starting strings with the root of your C: drive as the source directory. What you would get is an "A" page containing all the files starting with "A" a "B" page etc.

A slightly less simplistic example might be as follows: If you start with the following files:

- abc.txt some text about object "abc"
- abc.sn textual observations by recorder "sn" about "abc"
- abc.dw textual observations by recorder "dw" about "abc"
- abc.gif a scanned sketch describing "abc"
- abc-01.jpg an image of "abc"
- abc-02.jpg another image of "abc"
- abc.dwf a CAD "drawing web format" file of "abc"
- abx.txt some text about object "abx"
- abx.sn textual observations by recorder "sn" about "abx"
- abx.dw textual observations by recorder "dw" about "abx"
- abx.gif a scanned sketch describing "abx"
- abx-01.jpg an image of "abx"
- readme.txt some general comments
- 01011999.sn field notes for Jan 1, 1999 by "sn"

for these files one might use a configuration file like this:

- abc everything pertaining to the object "ABC"
 - abx everything pertaining to the object "ABX"
 - 01 everybody's field notes for January
 - .sn everything written about anything by "sn"
 - .dw everything written about anything by "dw"
- etc.

You could also have a category "ab" if "abc" and "abx" were really variations on the theme of "ab" or a "0101" which would assemble everybody's field notes for only January 1st.

Unused Files

The first check of your data's integrity is the page of "Unused Files". After all the codes you listed as possible beginnings of file names have been exhausted a page is created of all the files that met none of these criteria. There are really only two reasons a file would end up here, either there was a typographical error when entering the file name or it is a category you neglected to specify. Both are easy to fix once you are aware of them and the "Unused" category will disappear once you have done so. In the first example these would all start with numbers or special characters like "-", "~", etc. In the second there would be only "readme.txt" ie. Anything not starting with "abc", "abx" or "01" and which did not have an extension of ".sn" or ".dw"

Of course sometimes you have recorded something truly "unknown" or at least beyond your expertise or

responsibility to categorize. In this case the "Unused Files" flag will be a constant reminder to the whole team that something is unresolved.

Other errors

Slightly more problematic is the situation where you named something "A??" when it should have been "B??". In this case it is the speed of the computer which will help. It takes less than a minute to process 100 megabytes of files² so you can run it as frequently as you make additions to your database and if something doesn't turn up where you expected it you can track it down while it is still fresh in your mind. Failing this you will probably spot it some time when you are working with the "A"s and find an errant "B" among the references there.

Quality Control

A hugely frustrating and expensive situation experienced sooner or later by all Heritage Recorders is the bad or missing photograph or measurement that is not noticed until you have left the site. You will need the full CART package to get any help with missing measurements but CARTHTML can help with some of the other errors and omissions.

For instance, if someone has failed to transcribe their field notes the hole left in the layout of the page for that day will be immediately obvious. Images are particularly well served by this system. They are sorted and laid out in a row allowing for an easy comparison of all those in a set before you quit for the day or even as soon as you download the camera. Problems with exposure, camera movement or coverage are easy to spot while it is still relatively easy to obtain replacements (ie. the scaffolding is still in place).

Another feature with images is the ability to add captions. Any textual information you want to tie to an image can be added by creating a text file with a special extension but bearing the same name as the image. This could contain comments on the condition or materials of the objects in the image or exposure and focal length information or both. By using different extensions but the same name for these different types of information the output can be modified, with a simple configuration change, depending on whether the intended audience is conservationist or technical.

For instance if you have image files named:

- abc001.gif
- abc002.jpg
- xyz003.jpg

with technical notes:

- abc001.tec
- abc002.tec
- xyz003.tec

and explanatory comments:

- abc001.cap
- abc002.cap
- xyz003.cap

² The second time, the first time could take several minutes on a slow machine if there are a lot of thumbnail images to create.

a configuration file entry of "*" and caption definition of "tec" would give you a complete listing of all images with their technical information as captions while changing the caption definition to "cap" would create the same array of images with their explanatory comments.³ It is important to note that whichever option is chosen nothing has changed for the archivist as all the data is still there and is still related by file name.

By keeping multiple configuration files for a single data set you can always have up to date reports organized in different ways without any manipulation of the data itself.

A complete listing of everybody's field notes would look like:

- 01011999
- 01021999
- 01031999

while groupings by recorder would be achieved with:

- .ab
- .dw
- .sb

or a page of all the technical notes

- .tec

Configuration is handled by relatively simple text files which, once tuned to your satisfaction, will work for any project for which you use the same file naming system and want the same reports. In fact if your file naming is completely consistent all you should have to change from project to project is the Metadata and the wording of the index file. They can be created and edited using your favourite text editor or with the help of the recently added, GUI interface⁴.

Conclusions

Of the issues confronting conservationists the one least likely to be resolved by technology is the problem of how to efficiently collect data pertaining to the sites being studied. This is and will always be a time consuming and labour intensive task and it is necessary that things be made as easy as possible for those undertaking it. At least as difficult, but easier to postpone or ignore is the fact that the purpose of all conservation work is not only to preserve the artifact but also to create a detailed archive of everything known about the object in question both to better serve the conservation effort and to provide something for posterity in the event that a calamity befalls the original.

Fortunately both these requirements are well served, in our digital era, by a database consisting of an abundance of small computer files in relatively simple formats related to each other and the objects in question by their names. The drawback to this approach is that the large number of files

³ Actually if you want only images and their captions you also need to change a couple other configuration settings, specifically HtmlOrder=GIF,JPG and NoHtml=*

⁴ Appendix 1. or:
<http://nickerson.icomos.org/carhtml/www/demo/conf.htm>

generated during a detailed recording campaign quickly swamp the ability of the recorder to keep track of it all.

The CARTHTML publisher addresses this problem by working directly with these simple files in their archival formats to produce output in the familiar HTML format that allows everyone - recorders, managers, conservationists and archivists - to view the material in a variety of ways using only public domain software. The original files are not altered and a new presentation format can be prepared from the same raw material whenever standards change by simply modifying the publishing software.

This tool is very much a work in progress and will probably continue to be so indefinitely as it will have to keep pace with developments in information processing so that everyone involved can work in the environment in which they are most productive. It is already a useful tool for testing data integrity, a critical but oft overlooked step in preparing a useful record, and it can help immeasurably with the problems in communication that arise among practitioners of the different disciplines.

In spite of its name and current functionality it should be thought of it not as an "HTML publishing" package but as a reward for keeping your data in order

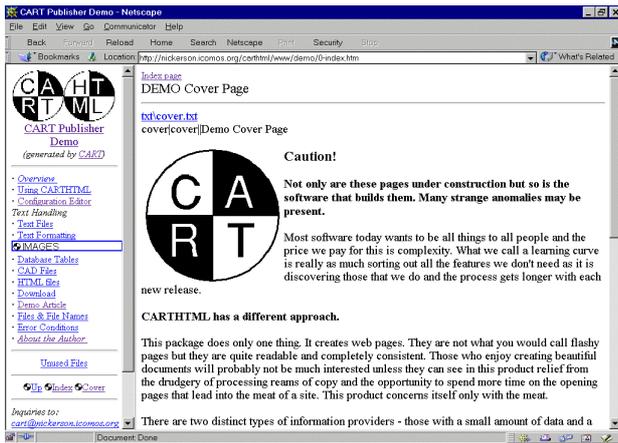


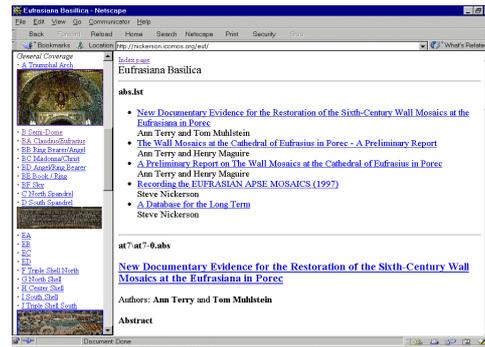
Figure 2 The opening page of the CARTHTML Demo showing the standard layout for indexed data sets.

This article can be viewed online as prepared by CARTHTML at:
<http://nickerson.icomos.org/papers/cipa99.htm>

Numerous other examples of both prepared data sets and work in progress will be found by following the links from:
<http://nickerson.icomos.org/cart/>
 and
<http://nickerson.icomos.org/steve/>

CARTHTML itself is presented at:
<http://nickerson.icomos.org/carthtml/www/demo/0-index.htm>

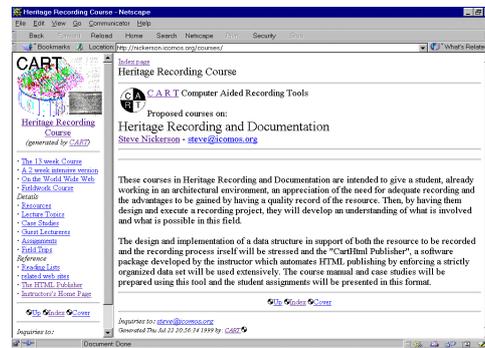
Some projects prepared using CARTHTML



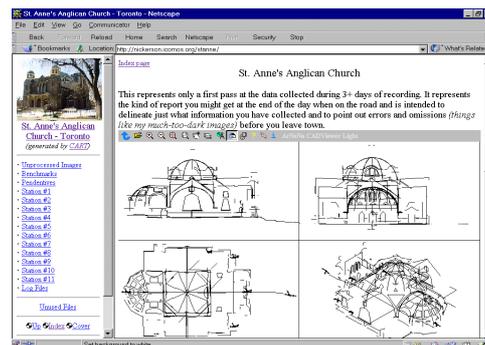
<http://nickerson.icomos.org/euf/>
 The Eufrasiana Basilica - An extensive photographic record and some scholarly reports



<http://nickerson.icomos.org/tebtunis/>
 The Bagnani Archives - Raw data from an old archive laid out so people can see it.



<http://nickerson.icomos.org/courses/>
 Heritage Recording Courses - A series of proposals showing how individual files can be used in multiple ways to achieve different ends.



<http://nickerson.icomos.org/stanne/>
 St. Annes Church - The results of a few days in the field (with a faulty camera)

Appendix 1. The CARHTML Configuration Interface

