Re: Q: digital image databases

Maximilian Schich

Dear Natasha Goldman, dear H-Arthist,

As the question of database structures for (digital) images concerns us all more or less, I would like to contribute my modest knowledge on that subject. This answer does not deal with one special program.

In this e-mail I'll provide information on six areas:
1. Possible database structures
2. The role of IDs and norm-data
3. Image quality and storage
4. The nature of available services (image sellers, content brokers...)
5. Where to get information, how others do it
6. Some words on money and ideology

1. POSSIBLE DATABASE STRUCTURES
First, the problem of database structure should be treated independent from the system used, or even more important from the system you would like to use.
Structure is a property of the data. Therefore it should not be distorted by some kind of tool, which is not adaptable to the data.

Maybe it is too exaggerated, but typically there are two kinds of database structure dealing with (digital) images. The first is what I would like to call the phonebook-structure. The second kind is the complex research database.
The choice what to use should be dependent primarily on the data you already have or you would like to collect.

The phonebook structure is usually very flat, which means that there are very few hypotactical relations inside the data. As a consequence the data itself is often very redundant.
The phonebook-structure is very similar to a card catalogue in an old fashioned library or to the stickers you will find on slides in an old fashioned visual resource collection. Databases using that kind of structure usually serve as a finding aid for the patrons of a slide library or something similar.
A great percentage of phonebook-applications use very simple, often off the shelf database products and are very happy with that. Concerning the

definition of the structure, standards are used widely, the VRA-Core and the Dublin-Core being the most well known. Recently more sophisticated systems have been developed, which still depend on the very flat phonebook-structure. They serve different purposes like multi collection search, very comfortable image viewing or building presentations for a lecture...
More than that, the retrieval tools you can use on your data get more sophisticated every day.

The phonebook structure is most likely your solution, if you are building a visual resource collection from scratch and your main purpose is providing images to your patrons.
It is definitely not your solution if you have data already or would like to work with data, which is more complex than the phonebook-structure.

In that case the complex research database is your solution: Many of the flat structures describe the image and the object depicted on the image as one entity. This habit is similar to the label on an old fashioned slide, where you could read "Bernini, Apollo chasing Daphne, 17th Century".
For a phonebook-application this is sufficient and widely used (for e.g. at prometheus-bildarchiv).
Outside the phonebook this practice does not make sense.
First, the image is not equal to the object depicted. Bernini is not the photographer. The slide is not dated in the 17th century.

Moreover it is impossible to describe more than one object depicted in one image in such a flat structure: Consider a slide with two sculptures, one ancient, the other by Bernini. How would you enter the date information in a single field without dating Bernini¹s sculpture to antiquity and without repeating information entered in other fields?

Many collections dealing with (digital) images already have a lot of data. The structure of this data is as diverse as human cognition. Adapting this data to any standard of the day is in any case a minimal solution. Like human cognition, the data cannot be standardized. If possible, the structure inside the data should be kept. The downsizing of information to any standard of the day can be automated in most cases for secondary purposes. What seems complex at first is much better maintainable with a complex system than with any "simple" solution. Even the most simple field definitions can be distorted by wrong entries.

Dates, attribution of authorship, even the identification of the work itself are subject to scientific opinion. A lot of the data collected for centuries already reflects this, by including citations and other useful information. This good practice should apply for electronic tools as well. Let me explain this in a real world example:
At the Munich Glyptothek there are separate card catalogues concerning

object information, photographic documentation, bibliography and provenance information. Each of them dates back to 1850 or even earlier. The object information contains notes on the personal opinions of the Glyptothek staff. This information is very precious, as it contains opinions which were never published.
The bibliography does not only contain the references to the literature, but even the date and style the author attributes to the specific object. The catalogue on photographic documentation is an important source for the historian of photography as well as for the archeologist, as it contains information on the photographers.
All these information has been incorporated into the electronic archive. A large percentage of this information would have been lost using a flat standard, like for e.g. the Dublin Core. The originators of opinion would now be untraceable.

The solution for the complex database structure is the semantic web, in which all dimensions of data can be normalized as desired.
It is not limited to a specific structure or rules, which you will find most likely in any relational or object oriented application off the shelf. The connection of the semantic web to a flat structure is very easy, as the flat data can be extracted from the complex data on the fly. The advantage is that not only today¹s phonebooks can be provided with your information, but also the systems of the future, which are able to transmit more complex information.

Guidelines for a semantic web data structure are provided by the ICOM/CIDOC-CRM initiative. (see http://cidoc.ics.forth.gr/)
The system used at the Munich Glyptothek is DYABOLA, an application which is capable of addressing all the subtle complexities appearing in research databases, museum inventories, excavation databases and library catalogues. It is available free of charge for interesting projects. (see www.dyabola.de)
The structure can be adapted to the data by the researcher without programming. The downside is that the researcher has to understand the structure.
This however is in the nature of the application, as structure is a property of the data. It isn't the business of some specialist in the information science department. It's the business of the researcher.

2. THE ROLE OF IDs AND NORM-DATA
IDs are very important. Every Item in your collection and every other normalized dimension of your data (person, location...) should be provided with some kind of unique identifier. This is very important, because otherwise nobody is able to cite or refer to your data.
Still there are large visual resource collections containing 700'000 photos without such IDs. As a consequence you will end up searching photographs for

hours, as they change their location inside the collection frequently over the decades. In an electronic environment missing IDs generate fuzzy results, which is as annoying as running through the collection for several hours.
Providing IDs is very simple. Most likely every item in your database has a record number or something similar. Make sure that this number is visible to your patrons. A combination of the record number with some kind of URL or DOI (digital object identifier) will make your records referable from anywhere in the world. The whole identification should be legible very easy, as the world comprises sheets of paper also, not only machine-read anchor-tags in some sort of hypertext.

Norm data like artist names, geographic locations, and so on are useful, especially when communicating your data to patrons via some kind of broker (see below).
However there are two issues to be cautious. First there is not and hopefully won't be any kind of brazilianyte "central services".
For artist names you could choose the already established ULAN or AKL. Alternatives are the PKNL (prometheus), which was released last week or the this and that list, which will be issued next year. Incorporating one or all of them into your data is an effort which has to be considered. It can be automated to some extent, but proper control means some manual work. It can be useful, as normalizing artist names is serious intellectual work and it does not make sense to do the same work all over again.
Second caution: All the norm-data will be useful to identify the well known artists, locations, etcŠ. When comes to the long shallow tail of the data, the norm-data won't be very useful. In a mathematical sense your data is not normal very likely.
Consider "The guy from Naples who paints sheep very similar to, but in my opinion is not 'the guy from Naples who paints sheep (ULAN-No. this and that)'". In that case it does not make sense to send an e-mail to the AKL or ULAN. Maybe you change your mind next week and the name you have provided does not make sense any more. Working with renaissance drawings or baroque painting this can be very important, as you will end up with up to 70% anonymous or loosely attributed material.
Adding the ULAN-Number to an attribution does not reinforce the attribution. To be sure, adding the ULAN-dataset of Piranesi to a drawing does not make it more piranesian.
In the end data which is annotated with norm data can be brilliant or crap. Data can be brilliant too, not using norm data at all.

3. IMAGE QUALITY AND STORAGE
Go for the best quality you can get. Display technology will improve to an enormous extent in the next years. Today¹s state of the art displays for laptops have a resolution of 1920 x 1200 pixels. In next generation video this resolution will be 16(!) times higher.

If image sellers provide you with high resolution images they deliver 3000 x 2000 pixel, the scanning standard of a regular 36 x 24 mm slide. To be precise this is only "high resolution" in comparison to a 1980s home computer.
Compared to a "Lichtdruck" by Brunn-Bruckmann from 1905 this resolution is lousy. Printed in that quality, the hi-res scan won't be any larger than a few square inches. The hi-res images are definitely not the material Wölfflin and Warburg would have chosen.

Today high quality printing means 550 to 600 dots per inch per color with up to 10 colors (forget RGB). That means what you can see on a regular book page is light-years from the best beamers available.

With an eye to the future scans should contain as much information as possible. Storage cost is not a problem any more.

The images can be archived in the original resolution and either be downsized to the specific use on the fly or be held separately in a lower resolution. It may be useful for that application to make use of a separate image server, which basically functions like a file-server and is independent from the database system containing the meta-data. On such servers the images are stored according to their ID in the database. The server delivers the image in the preferred form on the fly (resized, filtered ...). To separate the database from the image server is useful, as image databases providing both are a mixture of average solutions on both sides more often than not. If possible go for a media server which can handle not only images but all kinds of media.

Make sure that not only the collection database has access to the images on the server. The image server can also be used as a source for student's or researcher's projects, which use other database systems or simpler solutions like static html web pages. With a solution like this you could manage an inventory of all relevant images at your institution and in the same time others would be free to experiment, all while using the same images.
With that solution, lecture systems (power point, acrobat or proprietary) do not have to be standardized.

4. THE NATURE OF AVAILABLE SERVICES
Image sellers are companies which deliver slides and digital images often including phonebook-style metadata (MARC-format, VRA-CORE...). The most well known in the U.S. are Saskia and Davis. There is no need to mention, that the material provided by these companies is much better than most of the copy-stand photography taken from books.

Content brokers are services, which communicate your images to a wider audience and/or deliver images from other sources (via a collective interface). Most of the time, a flat data standard is used in the operation.

In that case complex data from your database will be simplified. Some information will be lost. On the other hand, your information reaches a wider audience.
Being a member of some sort of brokering service will also enrich your collection, following the rule give 500 images and get 100'000. Subscribing to a brokering service, should depend on your purpose, as data quality is differing. Some providers like AMICO specialize on quality images and data from museums and other collections. (see www.amico.org)
Other services like Prometheus-Bildarchiv in Germany also focus on connecting the existing slide libraries of the member institutions, which are mostly university departments. Here the images are often copy-stand photography, but nonetheless the service is more useful than any single slide library.

As with any other system there will be no exclusive solution. In the recent 15 years there have been many attempts to establish some kind of "central services". As with standards, new services are established frequently, CollectionConnection mentioned by Geert Souvereyns being the latest example. In the end there won't be one single "central service". Even google catches only a small part of the known web.

5. WHERE TO GET INFORMATION, HOW OTHERS DO IT
The problem of how to collect, maintain and provide (digital) images is one of the most important tasks of art research, especially in an institutional framework. It is well known for a long time, that collaboration between institutions and individuals is crucial. However as in any other field of activity it is very difficult to get bias-free information. Like in politics there are different lobbies. Therefore when building a new database, it is useful to consult all of them.

The group you belong to most likely according to your question is the art information professional, which includes art librarians and visual resource curators. Your duty is conservation and providing patrons (i.e. researchers) with material.
Most likely you will focus on developing standards of good practice and reduce work by symbiotic collaboration with others.

The second group is system providers, i.e. people and companies producing database systems and other applications. They will most likely tell you that their way to do it is the only one and that all other systems are crap. Be aware of the open-source guy at the coffee-machine. He will be the most annoying, as his system is not a tool but a religion. I personally use a lot of open source products too. Open source software is a very good thing. But in the end it is a tool just like the bone or the space station. Remember that the structure is in the data. The system has nothing to do with it, if it is capable to reflect the structure. (by the way: open source has nothing

to do with open access)

The third group is service providers, i.e. image sellers and content brokers. They are a very valuable source of information, as they work with a lot of data. Usually they can tell you more about structure than the theoretic guy from the information science department. On the other hand they use standards which might be too strict for a complex application.

The fourth group is the most important. It is the group of the user, i.e. the researcher. This is the group where you and I belong to personally, not as a member of an institution. As a lobby we are underrepresented. Therefore our demands are the greatest. What we want is unlimited bandwidth, as George Lucas once put it. Our duty is to take all the tools and all the information and play a symphony.
We want to know all the complexities of any data. Search and browse it with different tools. Get the best image quality. Rip, mix, and burn it. Remix it. Sample it. Scratch it. And data mine it.
We do that in order to extract new information from the data, to produce new data, to verfiy the data, to falsify the data and find out unknown properties of the structure.
You can be sure that our way does not comply with any standard. Most of the predefined standards will fall in the progress of our research. As William J.T. Mitchell pointed out recently, art research is science. And science, according to Paul Feyerabend, is basically an anarchic enterprise.
Therefore the data should be as open as possible.

Below there are some references, belonging to the first group, that of the art information professional.
I won't contribute the addresses of the system and service providers, as this field is constantly growing, and any listing would be idiosyncratic. Most of them advertise. You will find out, that some of the best providers are subject to damnatio memoriae among some other groups of people. The reason for this is that they do things, the people you talk to, would like to sell or put on their own list of honour. Concerning this phenomenon you should be aware of constellations where the first three groups named above mix. Institutions should provide bias free access to all information. This duty does not comply with the development of an exclusive system or one single service channel.

Advice for the art information professional:
In the U.S. the Visual Resource Association (VRA), a collaboration of visual resource curators from all over the country, is the most important source of advice. Here, hundreds of members from small and large Institutions exchange their Know How and develop guidelines of good practice.
(see http://www.vraweb.org)

A very interesting study by on visual resource collections in the U.S. and

their dealing with digital images is ³Susan Craig: Survey of Current Practices in Art and Architecture Libraries. in: The Twenty-First Century Art Librarian. Haworth Information Press (December 1, 2003) p. 91-108²
(see www.amazon.com)

In Europe there is no official profession equivalent to the visual resource curator as far as I know. Normally the same job is done by librarians or art historians. However as in the U.S. the maintenance of visual resource collections is related very closely to the field of art librarianship. It seems in fact, that the two fields are merging in some extent.
(see final draft 3/10/05 at http://www.uflib.ufl.edu/afa/pdc/coredata.htm)
A lot of countries in Europe have an association equivalent to the Art Libraries Society of North America (ARLIS/NA).
(see www.arlisna.org)
In Germany this is the "Arbeitsgemeinschaft der Kunst- und Museumsbibliotheken", which provides a comprehensive List of similar Institutions.
(see http://www.akmb.de/web/html/links/fachverbausland.html).

It is important however, that the art librarians¹ point of view is characterized by the handling of library material. Books and articles usually have a title, an author, a date of publication and a location of publication. Therefore it is possible to standardize the data to some greater extent.
Works of art, which are the real subject of most (digital) image databases, do not have these clear cut properties usually. Therefore you should not believe everything they tell you about standards.

6. SOME WORDS ON MONEY AND IDEOLOGY
The adaptation of a database system to your own needs is an intellectual piece of work. It does not matter if the system is open source or some other product - somebody has to do the work.
If your purpose is very similar to someone others, chances are good, that you will get a working system at a very low expense. If the adaptation is more complex and individual, it will cost some time and/or money no matter whether the system used is a commercial product or a piece of open source software - somebody has to do it.
With open source, we have to be aware, that our purposes are in no way as widespread as bittorrent or mozilla firefox. Therefore chances are that there are very few programmers working just for fun on your special problem of the day. In the end you will end up paying for the work.

In fact there are many models which can be successful. You can buy a standard application off the shelf, use open source software or even write your own application. Paying some commercial provider can be much cheaper than hiring a self-appointed specialist, who is going to invent the wheel

again.

Asking people with a lot of experience in the field will certainly cost some money. On the other hand, it might keep you from making mistakes others have made, which will most certainly cost a lot more.

The problem of "dependence on exclusive solutions" is a thing which has in no way to do with the database system or the licensing model used to spread it.
You should insist, however, that the data is exportable from your system without the help of a programmer.
Given that, you are save if the company goes bust or if your three open source programmers are unreachable, because they got a well paid job at the NSA.

For a funny account on open source and ideology in Europe see Bruce Sterling in WIRED 11.09
(http://www.wired.com/wired/archive/11.09/view.html?pg=4)

With kind regards,

Maximilian Schich.
maximilian@schich.nu
www.schich.nu

(The author is art historian; his continuous professional engagement in art information dates back to 1996)