Thoughts on a Distributed Web-Portal For World-Wide Collaboration Among Architectural Archives and Historians

Dr. Bernd Kulawik MA, Bern, Switzerland

June 8, 2009

Abstract

The paper presents a proposal for a co-ordinated (development of a) web platform for the presentation of and collaboration among architectural archives, architectural historians and architects world-wide. This platform should be organised on the basis of a distributed database-supporting and -supported Content Management System that allows users to add information in a referenced review-based workflow system, contribute to information added by others and comment on topics. In addition, the system could and should be used as an intranet for the participating institutions where they can also document and publish information about their collections and other resources. The distribution of the system would allow to have a central entrance portal — like http://www.a2w2.net — hosted by one of the participating institutions (or with any web-service provider world-wide), while the separate web-areas of the different institutions could be hosted on their own servers — in this way avoiding security and copyright problems: As long as a resource is published by the institution that owns these rights, it can control the access to this information at any given time and log any access from the outside. — To keep such a system available and make it a durable resource on the web, it is of the highest importance to have it based on free and open web-standards and to use a software that does not depend from a special manufacturer whose decisions about the future of the software could bring to a halt such a project with potentially hundreds (if not thousands) of people involved. In addition, the software should allow barrier-free access to any information and fullfill strict requirements regarding safety and version control in the review workflow. — The system proposed here to start such a portal is the award-winning free and open source Content Management System Plone. It is developed by more the 2'000 programmers and interface designers, and fulfills all the above mentioned requirements. Though this can be said about some dozen other CMS, reasons are given why Plone would be the best solution.

1 Introduction

Being an architectural historian myself who had to travel almost through Europe and North America to access the sources for his research, I know the limitations for researchers very well — but also the requirements of any archival institution, because I also worked at historical archives myself. On the one side, that of the single user/researcher, expanses for travel, photographs or scans of the sources needed for our research, can amount to (dozens, if not hundreds of) thousands of Euros over time — usually financiated not by the researchers themselves but by their institutions, universities or by funding organisations. In addition, many researchers need access to and use the same sources — causing a lot of multiple work efforts by different people doing almost the same work again and again. In addition, even the information already available on the web is scattered over hundreds of websites maintained by institutions, research groups or single persons.

Archives, on the other hand, need to document their treasures and (should) want to present them on the web, because what is not on the web, "virtually" almost does not exist anymore nowadays. But they still need to control the access to their sources and, e. g., the usage and proliferation of images and other potentially copyright-protected material. But usually, archives do not have the resources and the manpower to collect all the information about their treasures in databases AND present them on the web. Some archives started campaigns, sometimes in co-operation with research institutions, universities or companies (and lots of money) to develop their own solution to these problems. So, it would be of great usage and importance to have a solution for these problems and requirements. In my paper I want to suggest such a solution and also point out, why I do not think that existing projects fulfill the requirements.

2 The user's/researcher's point of view

Imagine an architectural historian searching for information about a (historical) building or architectural project: Usually, not only one architect has been involved and not all persons involved are known by their real or only the same name. For instance, during the Renaissance and Baroque, many architects used several names over time — and obviously, these names do not appear in the documents in a unified form. But even in later times, architects may have changed names or have been collaborators in different bureaus where their contribution to a special project often cannot be identified by names. Also, there may be material like drawings, written sources or financial documents that are still not identified as belonging to this special project. In addition, it is often not known, where all the material is located today, because archives where

transferred, united or — in the case of private collections — dispersed over several collections even world-wide.

So, where does our architectural historian start to search for information? Because of the described situation, it is impractical to ask all possibly relevant archives for information directly by letter or e-mail, especially when it is not known, where all the material could be today. Usually, one starts the search on the basis of already published information. But another — and today much more often used — starting point would be ... Google or other search engines on the web. Here another problem arises: Most of the databases used in archives are not available on the web, or ar — at least — not searchable because their content cannot be indexed by the search engines.

The usual solution is to restrict the research to a set of (known) sources from a very limited number of archives. Though this still may (and usually does) lead to comprehensive studies and interesting results, it is far from the optimum that — in my point of view — would require the usage of *all* surviving sources. And, today, this is not an utopian vision but could be achieved by collaboration.

Another aspect, already mentioned, is the question of funding: When researchers travel to an archive, this is usually funded by an institution that is also involved in funding archives or supporting archival projects. And this money can only be spent once.

In the archive, the researcher has a limited amount of time to get an overview of the material available, select the relevant sources and study them in detail. Usually, our researcher will collect much more information than will be used in the final publication. So, a lot of work for finding and transcribing sources is virtually lost. Therefore, and in addition, most of the material is used by different persons several times.

So, a lot of money and time is spent to find the information that could already be available — and then, by guessing that some more information might be at places where other sources are already known to exist, through personal investigation the amount of potentially available information is increased — but in the end only a (usually small) part of all the information gathered is really published.

The solution suggested to this problems for researchers would be some sort of meta-archive that could collect information from (potentially) all archives and make it searchable in a comfortable way.

3 The archives perspective

As already mentioned, archives themselves usually do not have the manpower to examine all of their sources in detail — and even if they have these possibilities at hand, they do not have the specialised knowledge that single researchers have accomplished

over years in their subject, and therefore it is not only difficult to identify interesting information, but it is also difficult to present all this information on the web. (I hereby suppose that archives are interested in the availability of the information they contain ... and not in restricting access.) Much of this work is done by the specialised researchers visiting the archive.

Since the 1980s, some archives started to use databases to collect their information — and there are even some cases, where these databases now contain astonishing amounts of information and are even available online, like the collection of drawings at the Louvre. But these databases usually are so-called relational databases that create "relations" between sets of information. These sets as well as the sort of relations has to be preset before the work begins and usually cannot be adjusted to new needs easily. So, to use these databases, one has to know their structure or at least the guiding ideas behind the chosen system of information storage. And, of course, these systems usually are not the same in different libraries. They may be similar, but — as long as no standardised software solution is used — it is not easy to find similar information in different databases. But most important is the fact, that this information is usually only accessable through special user interfaces on the web-sites of these institutions and that it is therefore not indexed by search engines. So, the researcher has to visit every single database online or at the institution ... and has to suppose, that all the information relevant to the research subject is stored in the database. Even in this case, e.g., misattribution of sources could limit the success of research on any level.

4 Basics for a Solution

A solution I want to propose for these problems and difficulties is the creation of a common web-portal for architectural archives and all institutions and persons that are interested in a collaboration to make as much information available as possible.

4.1 Requirements

The requirements can be grouped according to structure, technology, software and community:

4.1.1 Structure

1. Single institutions should be able to maintain their own web-space inside the portal, i.e. a distributed structure is needed.

2. This distributed storage should appear to users as a single one.

3. Access to different areas should be regulated by a common policy, i.e. there should

be open, semi-open and closed areas so that information can be collected inside an institution and made available to (registered) users or the public whenever needed in separate portions, e.g. to protect ungoing research.

4. Information should be stored in a structured way that is as close as possible to the "natural" way users — researchers and archivists — deal with material and information, i.e.:

   (a) the structure should be as "flat" as possible and understandable to researchers and archivists

   (b) single object should be treated as "objects"

   (c) the access or "path" to these objects should be similar to the path in the "real world" that one uses to access an object

   (d) The structure of the data/information should be traceable with "human-readable" URLs, i.e. instead of
   "http://www.archive.org/f124?=&&search:=xyc&&?3927dl710ed72e"
   something like:
   "http://www.archive.org/riba/collection_a/volume_10/sheet_54/verso"

   (e) non-objects, like meta-information, should be connected to all relevant objects

   (f) as long as an information may be subject to research or new interpretations (like dating, attribution to a person or project etc.) this information should not be "part" of the object and especially should not be used for the structuring of the objects. [I.e.: A drawing should be found under its "objective" or "impartial" signature, not under a "subjective" attribution to a person that is usually based on interpretation and therefore — in principle — subject to change.]

4.1.2 Technology

1. The data storage should not be centralised in one place, or at least should be mirrored in distant locations.

2. Separate data storage for separate institutions/archives would be preferable.

3. Data in all storages should be accessable from the entrance portal as if it were concentrated in a single place.

4. Data in all storages should be searchable over the internet and therefore accessable for search robots.

5. All data should be connectable over the internet as well as over "internal" links that remain "intact" even if the "location" of an object in the database is changed (i.e.: the URLs should be mapped onto Unique IDs that do not change so that an internal tracking UID system could find the object at any time).

6. It should be possible to manage the access and manipulation of information in a workflow-based regime, so that potentially all users registered with their real name could insert information and could comment and/or control new information.

7. It should be possible to collect data in "virtual" pools like working groups consisting of collaborators from all over the world working on a special topic or like persons, projects etc. so that all information available in the system could be united virtually at one point that offers more information than a simple (or advanced) search.

8. Discussion of single entries, any given information or more general topics should be possible.

4.1.3 Software

1. The software should be open source and free to avoid dependance from one single software vendor.

2. The software should not be platform / operating system dependent.

3. The software should support open standards to not exclude future migrations.

4. The software should allow distributed computing and distributed storage.

5. The Software should have an active developing community with contributors from (almost) all over the world, so that special help for the further development of the software as well as training for users is available everywhere.

6. The software should be usable after short training.

4.1.4 Community

All this requires a well-organised, but potentially open community: interested persons should easily be able to become members of the community: For simple membership, only the real name should be required (no pseudonyms) and contact address and (maybe) some information about the motivation for the research. The community should delegate their main decisions to a committee that should operate in an open, preferable web-based way to allow all members to follow discussions and the process of decision-making.

4.2 Proposal for a Software Solution

All these requirements should sound familiar to those working with collaborative web-based software like, e.g., Wikipedia, or larger Web Content Management Systems. And in fact, many modern systems fulfill these requirements — so it is somehow a question of "taste" to decide which software soluten to use. An argument against a wiki-based system is the "unstructured" way in which information is stored: These systems usually (and deliberately) allow to put information anywhere — so users are more likely to make mistakes by leaving the aforementioned "paths" to the objects. But Wikipedia's software is known to be reliable and stable and able to run on in a distributed way over datacenters world-wide, usually separated by languages.

Content Management Systems exist in very different "flavuors", but most of them use "crypted" URLs because they put the information needed for a database search into the URL: Therefore, as long as a search engine does not "know" how to fill in a search form, the information can not be indexed. Also, these system do not allow the aforementioned "readable" URLs and "natural paths".

These requirements, therefore, limit the range of available software to a few solutions, under which I personally know and prefer two Content Management Systems based on the free, open source Application Server ZOPE = Z Object Publishing Environment, using an object-oriented database and programming language (Python) which can be combined with all major relational databases. So, for instance, data from a relational database could be automatically collected into ZOPE's own database and made available online. With ZEO (Zope Enterprise Objects) a tool and solution exists to distribute the content of a central ZOPE server onto several "proxies" that can be used to create the "distributed" structure of a network of institutions, each one maintaining its own database.

While one of the two solutions, ZMS (= Zope-based content Management System) is for rather smaller solutions and institutions, the other one, Plone, is used in large projects and companies with hundreds of collaborators and dozens of institutions or

sub-entities. In addition, its usage is very simple and requires — for the single user — a training of less than an hour to be able to add information into the system. It also runs on small machines like Laptops as well as large Servers: Therefore, for instance, it would be possible to create "local" clones of the entire community system (as long as large, memory-consuming information like images is stored separately and loaded into the right place when required).

5 Already existing Web-Resources

Of course, my proposal for a collaborative information portal on architectural archives is not the first. At least two existing sources on the web come to mind: ArchitecturalArchives.net and Mace-project.eu — BUT:

5.1 architecturalarchives.net

- ▷ is explicitly restricted to European Archives
- ▷ The website is contained inside a "frame" — technically spoken, i.e.: all information is "locked" inside this frame but *not* directly accessable from the outside via URLs or web search ... and this means: it is almost invisible, even though it's on the web ...
- ▷ And: The total amount of information offered on this site is almost negligible.

5.2 mace-project.eu

- ▷ MACE tries to establish a set of architectural categories and terms on architecture in its database and on the web.
- ▷ This set cannot be translated into other languages, and it is not consistent.
- ▷ Again, all the information is encapsulated or "locked" inside a frame — and therefore not accessable and retrievable from the web.
- ▷ In addition, most of the information seems to be collected from outside the portal, i.e: if the source — e.g. a picture from "flickr.com" — disappeares, the information in the MACE portal itself disappears or is at least fragmented.
- ▷ The usage of the website is not simple, its structure even could almost be classified as "confusing".