

Vocal Music Database Exchange Format

Introduction

I'm dealing with vocal music related data for quite a while now. It started back in 2002 when I wrote a web based administration backend for CASA to maintain addresses, groups, memberships and the famous arrangement library, followed by a full integration of the www.casa.org website in 2004.

Main goal for me was to link as many information together instead of keying it in multiple times.

In the following years I developed this approach for my own website project (www.acappella-online.de) also, in first place in my own PHP system, in 2007 replaced by a system based on Drupal.

The main challenges I faced every single time were the following:

- 1) How to link what data together to avoid as much (inconsistent) data input and generate maximum value for the reader?
- 2) How do I encourage groups and agencies to provide their data for my records?

For 1) it was beneficial that I'm writing software for ages and I know quite well how to normalize data to avoid redundancy. In addition I checked other websites with such information and analysed what were the missing pieces, pitfalls and other circumstances that makes it difficult to deal with the data.

Just two examples in the following:

- if the group names are mentioned just in the description of the event, you cannot generate a list of concerts of that group.
- if the location is not given, you cannot search for events in your region

For 2) it is important to understand that group already have a lot of places in the www they have to maintain. Think about the group's website, Facebook, twitter, MySpace, YouTube and more. Why the heck should they key it in another time on my page??

I solved that problem in the beginning in way that needed a lot of individual programming: As in those days most of the websites were manually programmed and usually just having simple HTML structures, I wrote Site grabbing robots that visited the concert schedule pages of the groups, extracted the event data and pushed into my database.

As this approach was not maintainable with a growing number of Group Sites I decided to stop it when I moved to my Drupal driven page.

The now remaining argument (coming back to my initial question) why people should maintain data on my platform is: it has a high visibility in the scene and at Google and other search engines.

Having read this you may ask: *"And?? What's the problem of this guy now???"*

I can tell you what it is: in our global world of the Internet nothing is local or regional anymore. People from the US and Asia are visiting European vocal festivals (and vice versa), groups touring worldwide, fans search for information more globally and much more. This means that the problem is not just how to convince people to key in data on the Website of vocal music in one (the home) country. In addition it is



potentially worth to be present in the database of organizations representing the continent as well as global information stores for vocal music. (Honestly, this is not restricted to vocal music, but I personally will focus on it)

Status Quo

At the moment I'm designing the database and information store structures for EVA (european voices association), a non-profit organization to foster vocal music in Europe - see www.europeanvoices.org - and I see myself back facing my all-time favourite problems of data exchange as lined out above.

After thinking about it for a while I decided to do a 3 phase approach to nail it down:

- 1) Write down the generic concept of a global data exchange format for vocal music data based on the analysis of needs in the user base
- 2) Publish the concept and discuss it with organizations and website owners worldwide and finally decide on a format.
- 3) implement it for EVA, potentially write data connectors/tools to be used by other website owners

Kicking Off Phase 1

What you usually do first when you start a software design is a stakeholder analysis to find out who is potentially interested the outcome of your project. If you are clear about your stakeholders, you get a much better view on the different angles the results will be looked at later and it helps you a lot during your requirement definition.

Let's have a look at every single stakeholder to identify the expectation and derive the needs for the concept. I have focussed on mayor interests only to make the list not too long and too much diversified.

- The Artist/Group
 - o Wants to spread the word information and activities
 - o Wants to find teachers and business partners
- The Fan
 - o Wants to get information about his favourite group
 - o Wants to get information about activities in his area
 - o Wants to get information about activities and groups when travelling
 - o Wants to get information about organizations and festivals
- The Teacher
 - o Wants to spread the word about his offerings, skills and activities
 - o Wants to get information about artists interested in his offering
- The Company
 - o Wants to spread the word about its offerings
 - o Wants to get information about artists interested in its offering
 - o Wants to get information about events to support, take part or offer services
- The Organization
 - o Wants to provide an as complete as possible view on their core interests in the area it covers
 - o Wants to provide information beneficial to other stakeholders

- The Media
 - o Wants to get information about interesting activities to publish

If you drill those needs down you can see that the basics can be categorized into 2 types of information: Address and Calendar

Address	Calendar
Group Information	
Group Event	Group Event
Teacher Information	
Workshop	Workshop
Booking Agency	
Event/Group Management	
Recording Studio	
Equipment vendor	
Festival	Festival
Organization	
School	
	Media Appearance

Only the area of media (CDs/DVDs/MP3/...) is not represented in one of the 2 main buckets. In addition I have to say that for sure the for example the group information consist of much more than the address, but the address is an important part.

If we have a look at the identified data blocks in the first go, it seems quite simple and straight forward. We can put them all in an address database and a calendar and we are done. But this is valid for the first look only. If you dig a bit deeper you can identify the 3rd important component needed: linking information.

What does it help if I have the group information and not having it linked to the concerts of the group? Would I have to key in the Group information for every single event? What does it help to have a festival listed and not being able to list the concerts and workshops during that Festival? Linking information is the key!

Below an exemplary link list between data objects:

Object	Link	Object
Artist/Group	Represented by	Booking Agent
Artist/Group	Represented by	Management
Concert	Performed by	Artist/Group
Concert	Happens at	Festival
Workshop	Held by	Teacher
Workshop	Happens at	Festival
CD/DVD	Recorded by	Artist/Group
Web Link	Shows	Artist/Group

With this information I'm able to design the basic data model with all the information to be exchanged, BUT:

The creation and transport of data is not the biggest challenge!

The biggest challenge in that environment is the definition of uniqueness and the so called “leading system”.

To illustrate the problem imagine the following situation:

There is a Group XYZ added into a database in the US. At the same time a group XYZ is added into a database in Europe. What defines the valid entry? Is that the same group or two different groups? In case it is the same group, what do we do if in both places changes were made? What change is the correct one? In case they are two different groups, how do we make sure that the events attached to the US group XYZ are not moved to the European XYZ when a data transfer of concerts between both databases takes place?

Honestly I haven't found the golden bullet yet. There are multiple possibilities I will think through during the further development of the concept, just as a heads-up a few possible approaches:

- One big global central database acting as the master and everything is added and edited there and just the output will be provided to the local databases.
- Globally defined unique identifiers per database instance to identify the source of information and to make sure to update the right record
- Clear matching criteria to identify duplicates. Clear process to handle those exceptions, like: no overwrite of duplicate and alert importing party about the duplicate.
- Mark imported data (origin other than own data) as read only. Updates can only be made on the source side.
- Enable the owner of information (e.g. the group) to decide what database holds their “master” of data.

Have one global database instance doing the importing, consolidating and data cleansing. This database is used as the communication master (upload to master to this DB only, download to local DBs from this DB only)

- The group(s) maintain the data on their own website, are able to provide the data in the exchange format (RSS newsfeed-like) and register this feed with all other databases (CASA, EVA, Festivals, ...) where they want this data to be present

Ok – let's make it happen. This is definitely not a one man show. I see it as an effort of the community and I invite everybody interested to become a part of this journey. May it be as a contributor to the fields needed in the database, objects of interest I forgot to include. May it be as a potential stakeholder I missed and explain what you see as your explicit needs and benefits. I'm open to any comment, any help and support to make this happen.

I will now start writing down the object structures in detail as the basis for further discussion.

To summarize the advantages and benefits if we get this off the ground:

- Maintain once, but be present in multiple places. Reduces a lot of effort for groups, artists, teachers, website owners
- Exchange of information and datamining for the big organizations and websites. With this concept it is possible to generate the most complete data pool EVER for vocal music.



- Useful information filtering or searching/finding as all data using this format have a guaranteed data quality and content. All interested people can search their area, their country or region.
- Potentially also better usability across the participating platforms as with similar datastructure must likely the access to the data is comparable too.

All in all there are tons of benefits for all of us being in the vocal music scene and it will allow us to reach a level of service for the community we never had before (and maybe no other art form has).

Feel free to get in contact with me in case you want to comment, participate, help or just give some input to incorporate in this piece of work. You can reach me at volker@europeanvoices.org

Volker