Analysis of Pacifica Radio Archives Database Catalog

Karen Coyle
Digital Library Consultant
September, 2005

40th
ANNIVERSARY

NATIONAL
ENDOWMENT
FOR THE ARTS
Established 1965

Table of Contents

Introduction

The Pacifica Radio Archives holds over 40,500 radio programs dating from the early days of the Pacifica radio network to current productions. These programs chronicle political and cultural events of the second half of the twentieth century, such as a 1958 discussion on nuclear fallout between Linus Pauling and Edward Teller, and continue into the new millennium with important thinkers of today, like Arundhati Roy speaking at the World Social Forum in 2003. These programs, or new products derived from them, are used as promotional gifts during fundraising, are sold to members of the public, are licensed to producers for use in commercial and non-commercial broadcast products, and are key resources for academic research. Sales of programs and licensing of sound for re-use is a source of funding for the archive and supports the restoration and preservation of the contents of the Archives.

The Archives' web site (http://www.pacificaradioarchives.org) is used to communicate and publicize the existence of the Archives and provides the only public access to its holdings. The Archives use the site to promote programs of timely interest and items of historical importance, as well as to solicit donations and to promote the use of the archival resources in programming taking place on Pacifica and affiliate stations. It contains general contact information and pricing and licensing for use of the program materials. It also contains news on the progress of the preservation and access projects.

Public access to the Archives' catalog database is through the web site and for many programs is the only way for the public to discover the content of the Archives. The quality of the data and the technology used for the database, however, do not meet today's user expectations for searching and display of this type of data. This limits the utility of the catalog database for discovery of Pacifica Radio Archives materials. Improvement of this database and its access methods are key to increasing public awareness of the significant contents of the Archives.

Current Technology

The Archives Catalog Database

The catalog database contains information about each cataloged tape held by the Archives. The database itself is the product of a number of legacy systems solutions employed over more than a decade, and record content does not follow current standards for the encoding of archival metadata. In addition, records created at different times have different content and coding.

Currently, the database is stored in an InMagic Textbase database which is licensed by the Archives for $3000 Per year. This software product allows record creation and retrieval on the Archives' local network for use by staff. Public access to the database is provided through a web site managed by the consulting company (Legal Information Management) through whom the Archives licenses the InMagic software. Although this arrangement has made it possible to provide public access to the database, it has resulted in poor service to the public and possibly to a loss of revenue for the Archive.

Problems with InMagic for Archive Management

The Archives have a need to manage the catalog database of original radio programs and source materials, as well as to keep records of copies of those programs and other products developed by Pacifica using the Archives' collection. The current database contains a single table of data that has descriptive and promotional information as well as some technical and inventory data. Although the InMagic technology could support more sophisticated data, the data itself is inconsistent and often not suitable to the functions of the Archives. In particular, there is not an appropriate structure to store the needed information relating to copies of the sound files in various formats, how those copies were derived from the original, preservation status of the program, and the related item locations. Considerable effort will be needed to make the modifications to the catalog database and these modifications will require related changes to the public interface to the database. The Archives staff feels that these necessary changes will be difficult to realize in the current technology environment.

Problems with InMagic for Public Access

The proprietary nature of the InMagic software means that public access is through the InMagic web software rather than more widely available web site software that is commonly available. The Archives were not able to host their own web site due to these limitations, so hosting has been on a site run by the InMagic representative from whom the Archives license the software. This web site is totally separate from the InMagic database that is housed at the Archive. Updating the web database has been done by exporting the entire

Archives database and sending that electronically to the Legal Information Management web site as a replacement for the data there. For reasons that are yet unexplained, this method ceased working and the web database is now one to two years out of date. Many functions cannot be performed by Archive staff but must be accomplished by the person hosting the database. Archives staff depend on a single individual for support, and have often had the experience of that person not responding when needed. Clearly this solution is not working well for the Archives.

Current Web Site Technology

The general web site for the Pacifica Radio Archive is hosted by a network service provider (JoyceMedia). This service has proven to be reliable and easy to use for the Archive staff and the public. As part of the basic service, the web site allows the creation of databases (using MySQL) and web programming (using PHP). Both of these are very common software solutions that are openly available and have large communities of users. A small database (less than 100 items) of high-profile programs that are available for sale has been mounted on that site using this software solution. Archives staff and local consultants are able to create and modify databases on this site and can create programs that allow searching of the database. The staff would not have been able to create this capability using the InMagic technology given their level of access to the InMagic web site and the restrictions on their ability to make changes.

A transition of the catalog database from the InMagic software to a more open solution using MySQL and PHP would have great advantages for the Archives. This solution would eliminate the cost of licensing InMagic but would not increase the cost of the JoyceMedia web site license as the functions needed are already included in the price paid for that service. It would move the Archives from a proprietary solution where they depend on a single individual for support to a mainstream solution that would allow them to hire consultants from a large pool of technical experts should they need to create programs that are beyond the expertise of Archives staff.

Project Description

The Pacifica Radio Archives information systems have a number of issues that need to be resolved. Among these are:

1. The database that provides vital information on the Archives' inventory contains some poor-quality data that needs to be improved. This improvement will help Archives staff as well as the general public seeking resources that the Archives may be able to provide.

2. The Archives are currently using a proprietary software package that affords little flexibility for service development. This package does not allow Archives staff to make changes they would like to make, and this aspect of the Archives' operations has been frustratingly stagnant for some time.

3. The proprietary software cannot be integrated with the Archives' main web site. This means that potential users must be directed to a different location to search the Archivess holdings. The facilities at that location are not easy to use and do not contain the Archives' "branding" of displays and other features. (See Appendix A)

As in many technology-related projects, there are some functions that can be developed with relatively small amounts of effort, and others that will require more difficult and costly solutions. It is advisable to see the overall project in phases, with the earliest phases consisting of small steps that result in visible gains, and later steps taking on tasks that will provide additional services but may take special effort for completion. Using this analysis, the phases for the full project can be seen as follows:

Phase 1

Cleanup of significant database problems that are hindering public displays and indexing. This phase of the project has been begun, and the part that can be done within the Inmagic system will be completed by October 1, 2005.

This phase was funded by the Nat'l Endowment for the Arts, The Recording Academy (Grammy) Foundation, and donations by listeners and supporters of Pacifica Radio Archives.

Phase 2

Creation of the PR Archives database as a MySQL-based database on the JoyceMedia site, using existing data and existing web software capabilities. This second phase is the subject of the analysis that follows in the remainder of this document.

Phase 3

Additional data cleanup using MySQL facilities and some programming. Some analysis of the cleanup that can be facilitate or completed algorithmically has been undertaken. (Appendix C)

Phase 4

Conversion of the catalog database to a relational format, and addition of fields for technical and location data. (Appendix C)

Plan for Phase 2

With the completion of Phase 1 in the InMagic system, the next step is to migrate the catalog database to the JoyceMedia hosted site using MySQL and PHP. The resulting product will serve both as the administrative database for Archives staff and the public access database on the Archives' web site.

Goals

1. Create the PR Archives database as a MySQL-based database on the JoyceMedia site, using existing data.

2. Develop search capabilities that will facilitate retrieval. (see Appendix B)

3. Set up search and display interfaces that are compatible with the PRA web site look and feel. (see Appendix A)

4. Develop a calendar for the updating of the web database using the output from the current InMagic system.

5. Set up an input, edit, and display interface for Archive staff so that they can migrate entirely off of the InMagic system. At this point, the internal functions will make use of the same database on the JoyceMedia site that is accessed through the public search feature on the web site.

Steps

1. Principle Analyst, working with Archives staff, will specify technical requirements and interface features for both public access and Archive staff functions.

2. MySQL database will be created on the JoyceMedia site. A copy of the PR Archives database has been created in MySQL on an experimental basis during the investigative phase of this project, so our confidence level with this technology is high. Also, the Archives has already a smaller but similar MySQL database on the JoyceMedia site, and is satisfied with its functionality.

3. Programmer consultant will be hired to create the public and internal interfaces.

4. Principle Analyst will perform testing and quality assurance on database and interface functionality.

Project Staff

Pacific Archives Principal

     Adi Gevins

Principal Analyst

     Karen Coyle, Digital Library Consultant

Web Consultant

     Ginna Allison

Archive Staff

     Brian De Shazor, Archives Director

     Mariana Berkovich, Business Manager

     Andrea Hull, Cataloger

Costs

Approximately 50 hours of Principle Analyst time @ $100/hour = $5,000

Approximately 40 hours or Programmer time @ $100/hr = $4,000

Approximately 50 hours of Web consultant time @ $100/hr=$5,000

Grant Administration approximately $1,500

Total = $15,500

Timeline

Note: Current InMagic license extends to April 1, 2006.

Step 1 Creation of MySQL and creation of MySQL database on JoyceMedia site - December, 2005

Step 2 Creation of user interface on JoyceMedia site - February, 2006

Step 3 Creation of database management interface on JoyceMedia site - April, 2006

Appendices

Appendix A - Display of the Catalog Database

Branding, or "look and feel," is important for Web sites. It not only establishes a familiar face for the visitor, it is an aid to navigation in an environment where clicking on a link can take the visitor to an entirely different Internet site. Web sites regularly provide general navigation aids on each page of their site, allowing visitors to understand where in the site they currently are and giving them a quick route to other parts of the site. The Pacific Radio Archives Web site uses both branding and navigation to full advantage.

The Pacfica Radio Archives home page. The home page upper banner, which contains branding for the Archives as well as site navigation, is used on all pages on the site.

The search page for the Archives public catalog is within the main Web site and therefore has the branding and navigation that the site provides.

This is the search screen for the Archives' catalog database, showing its integration with the site.

However, when the results of the user's display appear, the branding and navigation of the Archives web site are entirely missing. This is because the data that is the result of the search is not resident on the Archives Web site. Instead, the user has been unknowingly passed to the InMagic system, which resides on a different Internet server and is not connected to the Archives web site. The results screen is generated by the InMagic system. Within this display, navigation is less than ideal, and the entries are difficult to read. Some of these problems stem from the data itself, but Archives staff feel that better displays should be possible with the current data.

The proposed migration of the database to the Joyce Media site would allow the Archives Web designers to create a display that retains the branding and navigation that is so vital to the quality of the user experience.
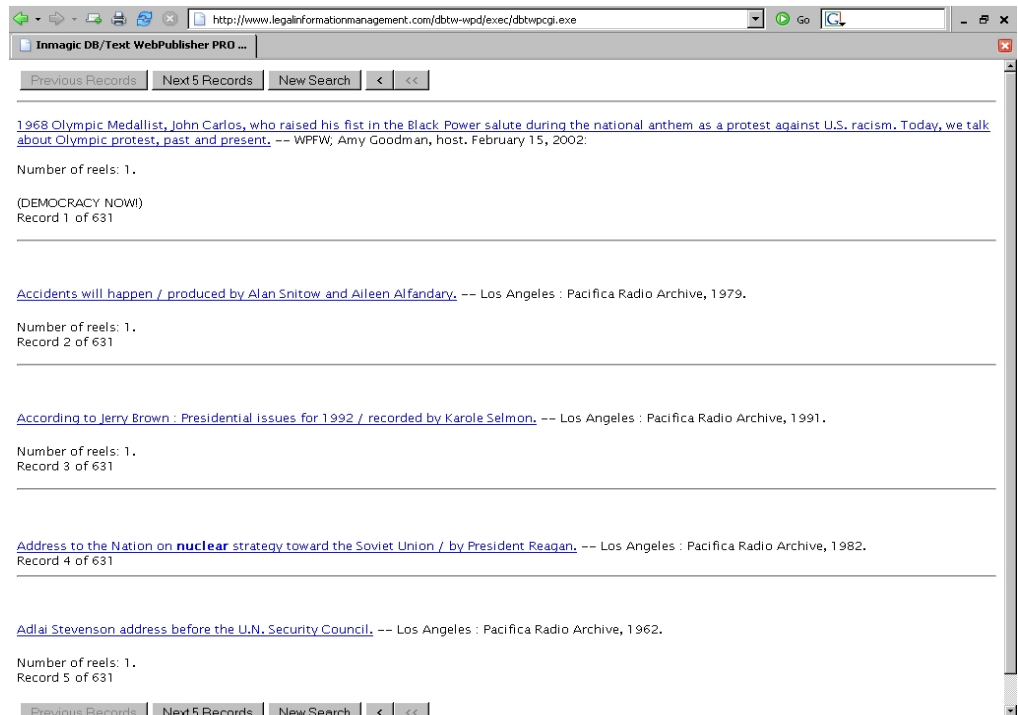
This shows the results of a search in the catalog database, as generated by the InMagic system.

Although the data in the catalog database is less straightforward than the data in the special programs database, it should be possible to provide the navigation and branding of the site, along with an easier to read display, for the catalog database records.

Title:
Accidents will happen / produced by Alan Snitow and Aileen Alfandary. -- Los Angeles : Pacifica Radio Archive, 1979.

Archive number: AZ0192

Documentary on Three Mile Island **nuclear** reactor accident in March 1979. Contains actuality. ACCIDENTS WILL HAPPEN / produced by Alan Snitow and Aileen Alfandary. – Documentary on Three Mile Island **nuclear** reactor accident in March, 1979. Contains statements by public relations men for the utility company (Metropolitan Edison), physicists, politicians, the President, and consumer advocates on what actually happened and what the results were. The last part of the program raises questions about the future use of **nuclear** power. – BROADCAST: KPFA, April, 1979.

Reels: 1. Length: 1 reel (58 min.) : 7 1/2 ips, mono. 58:00. Total: 58:00.

–0– –0–

Recorded on: –0–
BROADCAST: KPFA, Apr. 1979.

Record 2 of 631
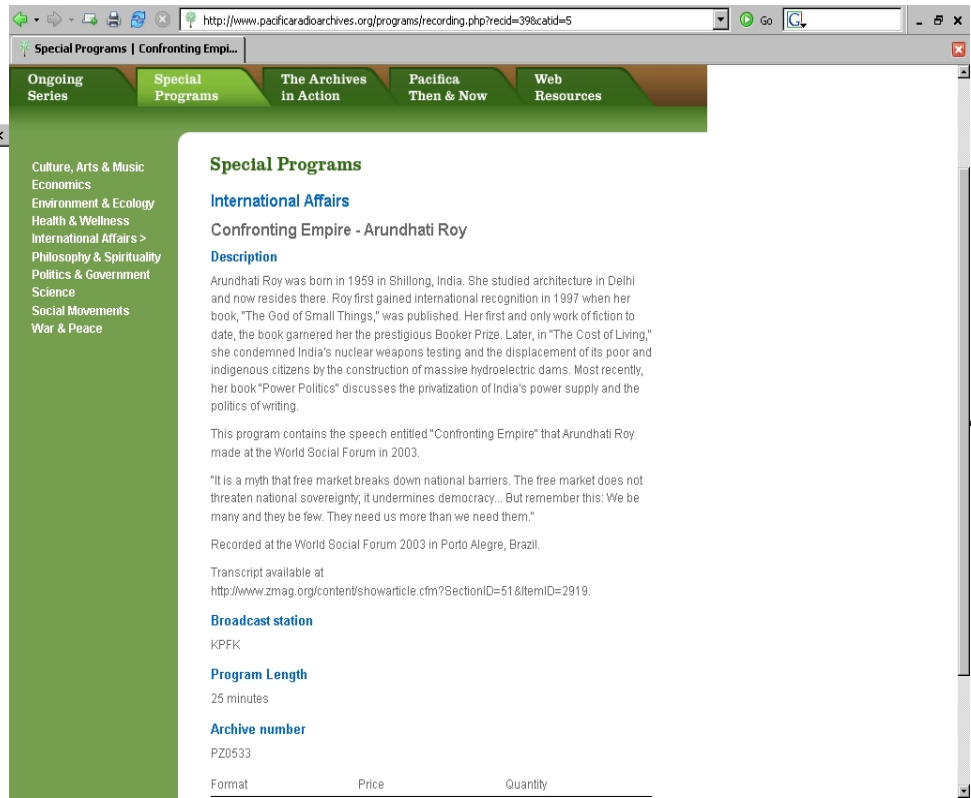
A long display of a single record in the Arcvhies catalog database (upper left) contrasted with a similar display from the database housed on the JoyceMedia web site using MySQL and PHP.

14

Appendix B - Search Capabilities for the Catalog Database

The Archives staff wishes to develop two search capabilities: a basic search that uses a "Google-style" single search box that performs a simple keyword search across all of the text fields in the record; and an advanced search that will allow researchers to be more precise in their retrievals. These two searches are outline briefly below.

<u>Basic Search Fields</u>

- ◆ Keyword

<u>Advanced Search Fields</u>

- ◆ Keyword
- ◆ Title
- ◆ Date
- ◆ Originating station
- ◆ Archive number

Keyword search would be a single search that combines keywords across these fields:

- Title, place publisher date, archive number, series, body, subject-KW, search, content, cast, general

Title search would be a single search that combines keywords across these fields:

- Title, series.

Appendix C - Catalog Database Cleanup

Phase 1 Cleanup

In this phase, cleanup was mainly limited to areas where global changes could be applied to fields.

Remove "-0-" from fields that contain it.

> This was used in earlier systems to indicate a blank field. Later systems use a blank. These characters display in the public display and are confusing. By using a blank, the fields that do not contain data will not display in public displays.:

BEFORE:

> Length: 1 reel (60 min.) : 7 1/2 ips, stereo.
>
> -0-
>
> *Recorded on: -0-*

AFTER:

> Length: 1 reel (60 min.) : 7 1/2 ips, stereo.

Remove captions from the certain fields, such as "RECORDED: " from the "recorded on" field.

> Fields input at some times contain the caption in the field, while other do not, resulting in user displays that repeat the caption:

BEFORE:

> Recorded on: RECORDED: in Albany, California, Mar. 1979.

AFTER:

> Recorded on: in Albany, California, Mar. 1979.
>
> (Note: captions in displays can be modified to better reflect the content of the fields. In this case, the caption "Recorded:" would be a better choice.)

Modify records for the series "Democracy Now!" to use the series title plus the date as the program title for individual programs. Move titles that consist of brief program contents to the "contents" field:

BEFORE:

Title:

Report on Manger Square and trapped Palestinians in Church of the Nativity in Bethlehem ; Human Rights Watch Report On Communal Violence In Gujarat ; Activist COMEDIAN RENO Recounts Her Days After September 11th -- WPFW; Amy Goodman, host. May 3, 2002:

Archive number: PZ0450.155

(DEMOCRACY NOW!)

AFTER:

Title: Democracy Now! May 3, 2002

Contents: Report on Manger Square and trapped Palestinians in Church of the Nativity in Bethlehem ; Human Rights Watch Report On Communal Violence In Gujarat ; Activist COMEDIAN RENO Recounts Her Days After September 11th -- WPFW; Amy Goodman, host. May 3, 2002:

Archive number: PZ0450.155

(DEMOCRACY NOW!)

This phase also did manual cleanup of about two hundred records that had a variety of problems, perhaps due to faulty import routines.

Phase 2 Cleanup

The cleanup changes in this phase will require that some programming be done. Archives staff is able to run programs on their local Windows server, but will also be able to run programs on the JoyceMedia site when the database has been converted to reside on that site. The changes will require a programming language such as Perl or a scripting language like PHP, both of which are available on the JoyceMedia site. It is therefore recommended that this changes take place after the transition to MySQL and PHP on the JoyceMedia site.

Data in all caps:

Detect when a field is all caps, and change to up/low. Ideally, the algorithm should be able to make use of an exception list so that common acronyms like "US" or "IBM" will not be turned to "Us" and "Ibm."

This occurs in Title, Series, Subject, Search

Add field for station of production/broadcast:

Many records have the name of the station in the "publisher" field, but not all do. It is possible to generate a new field that would contain just the station name using the first two characters of the Archive number. For example, archive numbers beginning AZ are for programs produced at KPFA, while archive numbers beginning KZ are for programs produced at KPFK.

Expand "restrictions" acronyms:

The "restrictions" field sometimes contain acronyms that represent certain kinds of restrictions, such as "*IWPOL.*" These acronyms could be expanded only on display, but their meaning is not commonly known, even to Archives staff, so it would be preferable to expand these to easily understood phrases.

Remove carriage returns and extra spaces in "body" fields:

The "body" field as originally developed in a system that generated labels for the Archives' tape boxes. This field has valuable information for display to users because it often contains the only synopsis of the contents of the tape. It also has some redundant information: it repeats the title of the program that is also located in the title field, and it also repeats the broadcast information. Because the field was originally designed as a printed label, it has carriage returns within the text to format it for printing on a particular form. It also uses spaces to separate the data in paragraphs. Because of the way that text is displayed on the web, these spaces and carriage returns do not appear on the web. However, these irregularities do make it difficult to view the fields in the editing screen, and therefore to make needed corrections to the data. Because the fields have a structure, however, it will be advantageous to retain the separation between the portions. This could be done in various ways; below we insert a code (" ") that would be ignored by the web display but that would be visible in the input function.

BEFORE

Views on Chinese invasion of Cambodia and the Cambodia issue before the U.N. SOVIET PRESS / William Mandel.   - SERIES: Soviet scene - William Mandel presents Soviet views on the Chinese invasion of Cambodia, and the Cambodia issues before the United Nations.   - BROADCAST: KPFA, 26 Feb. 1979.

AFTER

Views on Chinese invasion of Cambodia and the Cambodia issue before the U.N. SOVIET PRESS / William Mandel. - SERIES: Soviet scene - William Mandel presents Soviet views on the Chinese invasion of Cambodia, and the Cambodia issues before the United Nations. &npbs;- BROADCAST: KPFA, 26 Feb. 1979.

Phase 3 Cleanup

Phase 3 will need to make changes to the database table structure of the catalog database records. There are two areas where changes are needed: fields with multiple values, and fields relating to copies and multiple versions of the recorded program.

Multiple Value Fields

The current database consists of a single table of data. Within this table, however, there are some fields that contain multiple values. The InMagic database management system handles these multiple values in a way that is not compatible with the standard relational database design that is used by MySQL. In the Inmagic database, each of the multiple values is a separate data element and can be searched separately:

search field:

AMNESTY INTERNATIONAL.

Hinkley, David.

Economic assistance, American — Indonesia.

Human rights — Indonesia.

Indonesia / David Hinkley.

In the MySQL database when it is initially created, these will be in a single field separated from each other by a vertical bar:

search field:

AMNESTY INTERNATIONAL.|Hinkley, David.|Economic assistance, American — Indonesia.|Human rights — Indonesia.|Indonesia / David Hinkley.

This is a typical "one-to-many" relationship that will require a separate table in the relational file design, so that the multiple search fields can be associated with a single bibliographic record in the catalog database.

Creation of Reel and Item Records

Because the database is a single table, programs that are on multiple reels result in the creation of an entire database record for each reel. Although this is suitable for the Archives' inventory management functions, to outside users of the database this appears as duplicate records in their search results. It would be preferable to have a more hierarchical record structure that would have a single record for a program, with related records for individual reels in the program, and with item records for information about individual copies made from those reels. To do this will require additional study of the contents of the database and a strategy for making the change using programming and MySQL features.