Enhancing Existing Content By Transforming Books to Databases

Wendy C. Robertson
University of Iowa

© 2011 Wendy C Robertson

Comments
Includes speaker's notes.

Hosted by Iowa Research Online. For more information please contact: lib-ir@uiowa.edu.
At the University of Iowa Libraries, we have embarked on a variety of publishing initiatives. We also have converted a few print texts to databases, working with an in-house programmer and web developer. These are resource intensive projects, so we do few of them and only for items that are institutionally important. In each case, we are re-envisioning the print text to see how it can be most effectively used online and as such are not merely hosting a previously published PDF but creating a new way to interact with the material.

I will not be covering technical details of how we made the databases but instead will talk about issues and considerations that have come up in the process. In the discussion following, I hope to learn if others are doing projects like these and issues you have encountered.
Some the considerations we try to address when planning the project are:
What is the desired functionality for the online version?

What is the desired or needed functionality for the online version?
Just because it is possible, doesn’t mean it is necessary.
Just because it is desirable doesn’t mean we will have the time to implement it.
What are the features of the print?
What features are necessary to retain?
What features are relics of print’s limitations?
Some of the issues we are still working out are:

How true to the original should we be?
Does all formatting in the original need to be retained?
How do we deal with minor errors in the original?
How can we communicate most effectively with our programmer and web design team?

How can we communicate most effectively with our programmer and web design team? We only get a small amount of these colleagues’ time so we need to be as clear and detailed as possible from the outset.
We have a close partnership with the University of Iowa Press, with several projects coming from that partnership. We have converted one of their reference books to a database each year since 2009, launching at our summer book festival.
Our first project was *The Biographical Dictionary of Iowa*. The database was created a year after the print was published, giving us a clean, digital file to work from. Unfortunately, the text was not structured and we lacked software that might have made the transition easier. My former colleague, Joanna Lee, manipulated the data, primarily using XML. This project required her to learn a lot and she used this knowledge on other projects and to teach the rest of us more about XML, showing how these projects can have unanticipated benefits.
The print is sorted by individual's last name, with contributor and topical indices. The database includes these, as well as a chronological sort and full text searching. The chronological sort would not be possible in a simple PDF.
Having a single page per person makes retrieving a cited reference simpler than in a PDF. However, we don’t make it clear how to cite the entries.
Having a single page per person makes retrieving a cited reference simpler than in a PDF.
This lack of a clear citation is particularly noticeable in Wikipedia references. I hope we will add this information in the future.
The print uses bold to identify names of other people with entries in the book. We could not retain this feature in the online version. Ideally these would be links to the other entry. This loss of functionality was deemed acceptable given the other gains of the database and our time constraints.

We learned a lot from the process, including how to do things more efficiently next time, but perhaps most importantly, we learned it was possible for us to do a project like this.
Our 2nd database project was *The Vascular Plants of Iowa*. Joanna Lee also led this project.
The book was published in 1994 and in large part is a print version of a 1980s database. Unfortunately, that database no longer existed so the book needed to be scanned, OCRd and turned into structured tables. When entries are in context on a page, certain shortcuts can be taken with information. In this case, the family names are headers and the genus is only spelled out once. The database needed to use the full genus name for each entry. Italics are important in botanical notation so we needed to retain this formatting.
The database allows browsing by name, region and a full text search.
Each entry gives complete information, so you don’t need to follow the headers and abbreviations.
The book has an index by taxon names and synonyms. The index indicates the family and the major taxon, which is how the book is arranged. As a non-botanist, I find the book awkward to use.
The database allows you to drill down by classification and has a bread crumb trail which allows you to jump back to other levels.
The book includes a listing by common name. Many of the names are inverted in the index but not the entry. We missed this important piece of functionality and had to invert the names in the index near the end of the project when our UIPress partners pointed out the error. We had been looking at the entries individually but hadn’t really thought about how people would use this resource.
We turned the book’s bibliography into structured data. The entries display exactly as they appear in the text but the page code has both COinS and open URLs using the WorldCat registry.
This data uses complete titles and has data omission corrected. COinS is particularly useful with LibX and for pulling references in Zotero; the WorldCat URL is for people without LibX.
Our current project is *An Illustrated Guide to Iowa Prairie Plants*. 
The data is now in Excel and I have mapped the relationships between the tables. I included an inverted form of the common name for indexing and a link to the Vascular Plants data, so these two databases can link to each other. I corrected two obvious typos but otherwise have made no changes to the text. We are very cautious about making such changes because we want to be true to the original but we want the database to be correct.
The maps are a challenge. I am unsure how we will deal with them other than simply as images. I would like a reasonably streamlined method to turn the dots into county names for all 200+ maps. This would allow us to search by county, which would be a great feature.
We have not yet discussed desired functionality with our library botany subject experts which could result in additional work on the tables. We are unsure how important mobile access will be. We don’t know if we should add controlled terms based on the free text fields in order to filter results. We may want to have words link to their glossary definition.
Another project, launching this week, is a digitized guide to the Henry A. Wallace microfilm collection of correspondence held in 3 institutions. The microfilm of our collection has been completely digitized, but that for materials in the Library of Congress and the Roosevelt Library has not been digitized. The database indexes all 3 collections and links to our digitized content.
The original is one alphabetical list, with many items appearing twice, both under the sender and the recipient. Frequently occurring names are listed as initials. The name on the left is inverted, while the one on the right is not. Any letters to or from Wallace only appear under the correspondent’s name. Our students did quite a bit of OCR cleanup, correcting characters and fixing the table structure in Excel. My colleague Mark Anderson and I did further cleanup on the data.
The database has one entry for each item because we can search on both names at once. The pull down list will show you the number of items by sender or recipient and the results return both. Having a single entry per item meant flipping names and removing the duplicates. The entries were not always identical because sometimes an informal form of name or a misspelling was used on the original correspondence and the index included both forms. We needed to retain these intentional duplicates, but correct additional OCR problems found during the de-duplication. We also found a few typos in the original index and when they were on entries for locally held items, we corrected these errors.
Before sending the structured data to our programmer, we tried to envision how the database would be used.

The following are examples of the types of searches we envision:

- I want to find all letters TO J Edgar Hoover
- I want to find all letters written between Henry Wallace and Nicholas Roerich
- I want to find anything written by Wallace on April 12, 1945 (when FDR died)
- I want to find everything written in January or February 1945 when Wallace was proposed as Sec of Commerce
- While browsing the authors, I find a name I didn’t expect and then want to find everything to/from Ralph I Coryell
- I want to find all the letters to/from anyone with the last name Kiernan
- I want to find all the letters to/from Mordecai [last name forgotten by me]
We listed in detail the desired functionality, in priority order, so that they could focus on important items. In this case, we decided mobile access was important since people could use the database while sitting at a microfilm reader. We also pointed out issues with the data that might affect design. After we sent this information to our web development group, they requested that we give them mockups. In discussion with the developers, we shifted some of these specifications and designs.
We decided to limit specific searching by date range to years. As with all the results, they display in date order but can rearranged by clicking on the column heading.
We allow searching by specific date in ISO format in the general search box. We decided few people would need this option but we wanted it to be available.
We are working out exactly what we should give to our programmer and web developer in order to streamline the process on future databases. Our list currently is:
Provide structured data with clear relationships, preferably structured as tables, with unique IDs.
Clearly list desired functionality, in priority order, explaining why features are necessary, giving examples of expected use.
Create basic mockups of ALL search and results screens so that they understand how you expect the database to look and function.
We consider all these projects to be a success. They are time intensive, but are worthwhile for important items in the collection. They broaden our expertise and provide greater exposure for our Digital Services department. In this way the library has moved beyond collecting, distributing, and preserving content into enhancing content and keeping our important materials accessible in the digital world.
wendy-robertson@uiowa.edu

http://ir.uiowa.edu/lib_pubs/69/