Specimina Nova Pars Prima Sectio Mediaevalis XIII (2024),
p. 153–159.

DOI: 10.15170/SPMNNV.2024.13.08

Róbert Acél
acel.robert@pte.hu
University Archives
University of Pécs
Mátyás király utca 15.
7621 Pécs
Hungary

Róbert ACÉL:

Pécs University Almanac: Aims and Results – Technical implementation of the Pécs University Almanac

The Almanac project aimed to create a book and database of teachers with an academic degree from the University of Pécs and its predecessors. To enable data collecting, we first designed and constructed a database system with a web-based client. Then a group of researchers collected the information that was obtainable about the teachers and entered it into the database. The following phase involved data unification and revision.
It took two steps to create a book using the database. It was first exported using a thorough script into Word files, which the editors then looked over and reviewed. After that, a desktop publishing program was used to add the photographs and various appendices to the raw text, creating the final book.

Keywords: database, university history, archontology, twentieth century history, Pécs, Hungary

Although the idea existed before, the realization of the University Almanac started in 2014, when József Bódis, the rector of the university at that time founded the editorial committee. The research aspect of this process had its own challenges, but here I just address the technical side of it, since I was involved in every step.

The goal of the project was to make three volumes, each about a different period of the University's history. These were:

1. The medieval university, institutions of the nineteenth century, the Royal "Erzsébet" University of Pozsony and Pécs (1914–1950),[1]
2. The fragmented institutions of higher education in Pécs (1951–2000),[2]
3. The integrated University of Pécs (2000–).[3]

Since the book itself is supposed to be a kind of data bank, the idea of making an electronic database of it came naturally. Publishing the collected data in a searchable database on the Internet was part of the idea from the beginning. As the first step, the editorial committee (Ágnes Fischerné Dárdai,

[1] LENGVÁRI 2015.

[2] POLYÁK 2017.

[3] ACÉL – GUTAI – SOPONYAI-MÉHES 2023.

Márta Font, Attila Borhidi, István Kajtár, Imre Schneider, and István Lengvári chairman) determined what kind of data should be collected about the subjects.

For researching the first volume, a working group was formed from the staff of the University Archives, and the University Library and Knowledge-Centre. At first, we planned to use a standardized Excel-sheet for data collection based on the specification determined by the editorial committee, and make the book from this, and finally convert the data into a database. We thought about collecting the data into an online database, but it seemed to require too much time.

The Library took upon itself to provide the IT background for the project. With Tamás Markó, head of the library's IT department, István Lengvári, director of the University Archives, we sat down to discuss the possibilities, in August 2014. Tamás Markó ensured us that they can make a suitable multi-user online database for us, with all the things we needed in two months. Since the advantages of going this route were clear (more precise, more standardized data entry, ease of use, greater control, etc.), we chose to wait. Although for unforeseen reasons it took a bit more time, this proved to be the right decision. We had a much easier time collecting and verifying the data this way, and I am sure that the quality of our work is much better.

Our first task was to define the database structure in detail. Although the editorial committee determined what data was needed, we had to clarify and specify it. For example, they defined, that (naturally) the subject's Name should be recorded. So/Thus, a data field was needed for that. But there could be prefixes, e.g. titles of nobility, etc. Those cannot be recorded in the same field, so we needed another field for that. Nationalized names, or other variants were also possible, as so Maiden names, and so on. We went over all the determined and items and thought them over like this.

Then Tamás Markó made a detailed specification document, which served as the base for the developer. Ádám Horváth made the first version, but he left the Library in November. His replacement was Ákos Takács, a great and talented developer, but picking up the pace, and taking over someone else's project takes some time. So/Hence, we could start testing only after the Holidays.

This delay meant, that some colleagues could not wait for it to finish, because their schedules were tight, and they were not available later. They collected their part of the data in Excel. This proved to be handy, as these records served as test material. I entered these into the database in January, finding bugs and inconveniences during the process. There were a few major, and some minor issues.

Namely, we were so focused on getting the data structure right, that we had not thought about how we would navigate through hundreds of records, especially in a multi-user environment. Also, the nature of this work was different from a standard data record job. We did not just enter pre-collected data on the data sheet, saved it and were done with it. It was research. Finding new sources, which contained new information about multiple subjects, was part of the process. A researcher needed to be able to quickly find all of the records that belonged to them, and go over each, to enter a new piece of information that just emerged. So new sorting and filtering tools were needed, and Ákos Takács built them quickly.

There is a saying: "No battle plan survives the contact with the enemy" – the same goes for planning a database, it seems. The data plan needed to be adapted and adjusted to the emerging problems and experiences. Only practice showed, what part was too detailed or not detailed enough, what was inconvenient, and what was unnecessary. The reality of the research often gave us technical problems to solve. An example: we thought beforehand that we would find exact dates of birth and death every time, with no exceptions. But we had to accept, that it was not possible. Now we needed to deal with incomplete dates. A date field in the database could not accept incomplete data. So, we had to add a marker for the dates to indicate what part of the date was missing (and filled with placeholder number) and make every part of the system that delivered that date for a web search, or for export, etc., to display that date in an incomplete form.

These problems and adaptations were finished by the end of January, and after a brief training, the team could start the work. This research phase lasted all spring and was closed as summer came.

A problem became obvious during the early test period. Being consistent with the names of institutions, organizations, and departments is not easy, especially with a team of around 10 researchers. Often different sources wrote the name of the same institution differently. There can be slight changes in the names over time, and the source may use the wrong name for that time period, etc. There are many ways to make mistakes, and most of the time it only shows when one compares the collected data. Therefore, we needed a tool to find those mistakes. And more than that: we needed a tool to correct them. Going over hundreds of records one by one every time we find an inconsistency is not an option. We needed to do it in bulk.

Fortunately, there was time to develop this tool during the research phase. I planned how the tool should look and how it should work, and Ákos Takács built it. This proved invaluable, not only when correcting those inconsistencies, but it served as a rudimentary search engine. The proper and detailed search engine that could be used beyond the scope of simple record management was planned for later; with this, we could do a bit more. For example, this tool made it possible to assemble the chronological list of department heads, which can be found in the appendix of the book.

By summer, all the data had been recorded, and thanks to the aforementioned tool, the major corrections and unifications could be finished in short order. The records were verified by the editors and marked "Ready". The next task was up: making a book out of the database.

By the time we got here, I had time to consider the requirements and the possibilities. We would have to generate a text from the content of the records, as close to the final look as possible. The book would then be finished in Adobe InDesign. But before that, the editors needed to be able to check and correct the generated text in a word processor, preferably MS Word.

I made a one-page 'proof of concept' to find out what works. The first step is to generate a html file by faculty. I chose this format because it is very simple, similar to a plain text file, and it is very easy to piece together by a script. In addition, it is possible to make basic text formatting (bold and italic fonts) in it, and it can be opened and edited in MS Word. Moreover, it is possible to apply text types to different parts of the text, which can be interpreted as styles by MS Word. This last property offered a huge help down the road for the layout editor. The file would be easily converted then to doc format. With that, the editors then could complete the second step, the tedious and meticulous process of checking and correcting the text in a familiar software. After the text is final and ready, it goes to InDesign, where the pictures will be added, and the format and layout will be finalized.

Therefore, I had to write a really detailed and thorough description of how the html files should be assembled, what field goes after what text, down to the last comma and space. Based on this, Ákos Takács then made an export module, that pieced together the text of the chapter of that faculty.

Pictures were also exported, but not as part of the text. For the quality of the final product, it was better to have them separately. The pictures, coming in different types and resolutions, were converted to a maximized size in the same format, and exported next to the html file.

But as everyone knows, even the most meticulously assembled machine-generated text will not be perfect, since there are countless nuances, that cannot be calculated for. That is why our editors had to go over the text, review and fix it where necessary. This time-consuming work fell to Petra Polyák and István Lengvári. They did not only verify the text for generation errors but also checked again for the accuracy of the content. As I mentioned before, all records were reviewed before export, but there is no such thing as too many verifications, and at this stage, errors could pop up that remained hidden before.

This is when the chronological and organizational charts and lists of the departments came very handy. Discrepancies in the dates of commissions of department, faculty, etc. heads were frequent. Sometimes commission dates of different professors contradicted each other. With those lists, we saw where there were overlaps, or holes in the timeline, and we could do some extra research to solve those contradictions. That is why I think that the appendix, that started as a side product of the book, might be the most universally useful part of it. This kind of complete and coherent organizational list cannot be found anywhere else about our university.

This work was finished by early autumn after that came my part as a layout editor. This was surprisingly easy and fast. Since I designed the generation of the export files, I knew what small adjustments would make my job easier down the line. The character-level formatting of the text, and setting up bold and regular parts would have taken weeks, but since we made it part of the export script, this time was saved. Most of the formatting was done by matching styles at the text import, after that inserting the pictures, and adjusting page breaks were nearly the only things to do with the main chapters.

Of course, the pictures gave some work. Besides clarifying their copyright status, they needed some technical attention too. Some of that was done by the Almanac system, it converted the various formats to Jpeg, and maximized the size of the pictures. Naturally, oversized pictures were more rare, often we had to deal with small, and bad-quality photos. We strove to clear and improve them with simple and free-to-use editors to remove stains, folds, stapler marks, etc.

After that, full proofreading came and then the book was ready, it could go to print.

The second book was a whole lot different in regard to research and data gathering, but on the technical side, it was not much different. We had to expand the list of faculties and had to define which faculty belongs to which book. Every subject had a main faculty assigned to them, where they did the majority of their work, and they could have other faculties listed too, where they were also involved. The main faculty was used to define which chapter of which book the subject belonged to. Our various tools had to be supplemented by a "pre-filter", to let us deal with only the records of one book or another.

We needed to change the fields about time of death. With a time period closer to the present, some of our subjects/individuals were still alive, so we needed to accommodate this possibility too. Similarly, the field that listed the reason of departure from the university formerly contained a free text field to list an institute or premade text for retirement and decease. Now some subjects/individuals were still active. Not many though, since by 2017, when the second book was edited, those who were still active mostly spent the majority of their career in the time period of the third book (after 2000). But it was possible, and it happened, so the system had to be ready for that.

The export process was also expanded by a book selector. The script needed only minor tweaks to follow up on some changes in the structure of education in that era and to accommodate the possibility of active and live subjects.

Pictures in this book are more numerous, and they are from more different sources, so clarifying the publication rights was a more complex task, but on the technical side, editing the layout of the book was basically the same as before.

When the second book of the series was ready, we thought we had a large database enough to create a public frontend for it. Up until this point, the database was only accessible to the people who worked on it, after authentication. Now we wanted to make the records of the first two books available and searchable to anyone.

From a researcher's standpoint, any database is as good as its search engine. Therefore, my goal was from the start to make a search page ready to accommodate complex search criteria. I checked out some search applications for inspiration, and the capabilities of the AtoM archival system gave me the most ideas. In this program, you can specify what to search for and in which field, and you can chain any number of such terms together with logic gates (AND, OR, NOT), so we tried to make something similar.

As usual, we made a simple and a complex search option, and the option to just browse through the records. The latter was available by faculties, listing every subject who had worked on that faculty, whether it was their main faculty, or not.

The simple search required a single string to search for, and it looked through all the fields of all records for it. The results were grouped by the field they were found in.

Regarding the complex search, we had to make some compromises. To make a page with an arbitrary number of search terms was too difficult for the development framework we worked with. We had to settle for a fixed number of search terms, namely three of them. It did not seem to be too big of a sacrifice at the time, since we planned to make it possible to start searches based on the results of a former search. Unfortunately, this feature was later scraped due to limited time and resources, but three search terms proved to be more than enough most of the time.

In the complex search, we can choose what field to search in, and with what operator. The operators differed depending on the type of the field. For numerical fields, *less than*, *more than*, *equal*, *not equal* are the operators. For the text fields these operators are: *contains*, *does not contain*, *empty*, *not empty*. Three search terms with these operators combined with logic gates makes it possible to define quite complex search criteria. For example, to list the subjects who got their degree at the Medical University of Pécs, and worked there, but who did not retire from this place. Or subjects who got their degree before the Second World War, but who worked at the University of Pécs or the Medical University of Pécs even after the 1950s. The list of possibilities may not be endless, but it is great enough to satisfy the needs of the vast majority of researchers.

The search results are listed on the screen, and the subject's name, birth date, main faculty, and time of employment are shown, the list can be sorted by the first three columns. Every record can be examined in full detail from here of course. If a researcher needs to download data from more subjects, that is possible too. He can mark records from the search results, or the whole list to export in pdf or xml format. The selected records can be put in a "cart" just like in a webshop. It is possible to put records from multiple searches into the same cart and then export in pdf or xml. The pdf export is basically a custom almanac. It is generated from the selected records by the script used in making the book, with fixed-size photos at the beginning of the text. Of course, without manual verification and editing, it will not be the same quality as the real book, but it is a nice and clear, highly usable format to download.

Our next job is the third part of the Almanac. The changes in data protection and privacy laws gave us much to think about, just as the vastly greater number of potential subjects/individuals. But on the technical side, this will be the simplest, since no real changes are needed.

The Almanac project was challenging, but it shows that with good planning and flexible workflow, it is possible to achieve great results even with very limited experience and resources.

BIBLIOGRAPHY

Secondary Literature

ACÉL – GUTAI – SOPONYAI-MÉHES 2023 — *Pécsi Egyetemi Almanach III. 2000–2019*. [Almanac of The University of Pécs 2. 1951–1999] Ed. ACÉL, Róbert – GUTAI, CSILLA – SOPONYAI-MÉHES, Judit. Pécs, 2023.

LENGVÁRI 2015 — *Pécsi Egyetemi Almanach 1. 1367–1950*. [Almanac of The University of Pécs 1. 1367–1950] Ed. LENGVÁRI, István. Pécs. 2015.

POLYÁK 2017 — *Pécsi Egyetemi Almanach 2. 1951–1999*. [Almanac of The University of Pécs 2. 1951–1999] Ed. POLYÁK, Petra Pécs. 2017.