Chapter 34

A Publication Database for Research Documentation and Performance Evaluation

KARL RIEDLING[1] and SIEGFRIED SELBERHERR[2]
[1] Technical University Vienna, Institute of Sensor and Actuator Systems, A-1040 Vienna, Austria. [2] Technical University Vienna, Institute for Microelectronics, A-1040 Vienna, Austria. E-mail: karl.riedling@tuwien.ac.at, siegfried.selberherr@tuwien.ac.at

Particularly in technical sciences with interdisciplinary aspects, established publication collection systems are not well suited for the documentation of the research output of entire academic organization units, and even less for evaluation purposes. To support allocation of resources dependent on publication output, the Faculty of Electrical Engineering and Information Technology at the Technical University Vienna therefore decided to custom-design a publication database, which the entire university later adopted. This Web-based database supports a wide range of publication types and features simple extraction of lists and counts of publications based on a variety of query criteria. The database provides public search facilities and web services that dynamically create publication lists and records, and exports its contents to the university library system; furthermore, it supplies all publication-related evaluation results of our university, which also affect resource allocations. Therefore, the database serves both as research documentation and as evaluation tool.

INTRODUCTION

Publications are one of the main “products” of research organizations. Accordingly, the publication lists of researchers and research groups not only offer valuable insight in their current activities and document their research results; the publication output is also commonly used as a measure for the quality of scientific work. However, manually maintained publication lists tend to be outdated, and evaluation data provided by researchers are not in all cases accurate and verifiable. Both aspects call for the use of databases with some kind of built-in quality control, which also offer a variety of search and query facilities.

In many scientific fields, there are internationally recognized publication collection systems that completely cover their respective areas. These systems, e.g., Inspec [1] or Compendex [2], offer a vast amount of information and are excellent sources for scientific research. However, they may or may not cover the entire spectrum of the interdisciplinary work often done in engineering sciences. Although they permit to search for the publications of a particular author, they usually have no provisions for reliably extracting publication lists or counts for groups of scientists or entire organization units, which typically is required for evaluation purposes and for the documentation of the research results of a particular organization. In addition, a complete account of the work performed at an academic institution also comprises less "official" publications like theses or reports, which are by design ignored by standard publication collections.

Frequently, the wish of university authorities for reliable evaluation data is the incentive for establishing sustainable solutions to these problems. In our case, the Faculty of Electrical Engineering and Information Technology at the Technical University Vienna decided to custom-design a publication database, with the intention to allocate resources dependent on reliably determined publication output. The emphasis of the project was and is on the documentation of the *own* scientific work; there never was the intention to compete with the publication collection systems mentioned above. It was clear, however, that researchers accept only very reluctantly an instrument that serves exclusively as an evaluation tool, with a consequently detrimental impact on the quality of data in this database. Therefore, it was a major design aspect to have a publication database that is capable of making full use in every perceivable direction of the information contained, even in the initial stage of this project which started in spring 1999.

Since a quick solution was required, we chose *Microsoft Access* for the prototype version of this database. This prototype became operational after only a couple of months of development; it was introduced faculty-wide in late 1999. Some severe drawbacks of *Access* in a multi-user environment, and the very limited access to the contents of the database, which effectively precluded its widespread use as research documentation system, led to the development of a Web-based database solution with a LAMP (Linux – Apache – MySQL – PHP) approach. Based on the concept, and our experience meanwhile gathered with the *Access* prototype, a group of four students wrote the initial code of the Web-based database. The Web-based version became available in mid-2001, almost two years after the *Access* prototype had been ready for use, and after 13 version releases of the *Access* database. We migrated more than 3600 publication records from the *Access* prototype to the Web-based publication database.

From the very beginning of this project, one single person, the first author of this paper has been executively in charge of the architecture, implementation and programming of the publication database. Meanwhile, the volume of the software – currently almost 48,000 lines of PHP source code – has grown by a factor of more than six due to the implementation of a wealth of additional functions and improvements in close to 70 major and minor releases. This expansion was partly due to additional evaluation functionality required by law or the university authorities, but to a greater degree because of functions to provide enhanced usability and "added value". Since the software met the expectations of the university authorities, the entire university adopted it in mid-2002. It now provides all publication-related evaluation data and many aspects of the university's research documentation.

THE CONCEPT OF THE PUBLICATION DATABASE

Two possibly conflicting requirements determine the design of a system like the publication database: the completeness of the data held in the database, and the ease of use for the intended users. Information in the database has to be as comprehensive and detailed as possible to allow for all conceivable queries, particularly because evaluation queries also have to take into account the types and quality of the publications. We designed the database to allow the institutes to enter their publication data themselves, which results in total freedom with regard to which publication types, and which details of publications the database can hold. However, most users who create entries into the database are not familiar with the advanced aspects of bibliography; this precludes a "full-blown" bibliographic system. It demands, in contrast, a flexible approach where only self-explanatory fields essential for identifying and verifying a publication need to be filled in, while optional fields are available for additional information such as abstracts, keywords, or links to electronic versions of the publication. It also requires a user interface that makes it easy for non-scientists to work with the system. In many instances, secretaries are responsible for the data maintenance in the database. This also distinguishes our publication database from the large publication collections, which primarily are intended for use by scientists or librarians.

Instruments that only serve the purpose to collect statistical data are generally not well accepted. Therefore, we decided to build enough knowledge management system facilities into the publication database to provide sufficient additional scientific benefit to its users and hence improve its acceptance. An advantage for all users is to extract their own publication lists, even dynamically for use on a web site. The standardized reference format that greatly facilitates the preparation of project applications or departmental reports is an additional benefit in creating publication lists from a database. Furthermore, the visibility of the own work is improved for external visitors who are able to freely search for information in the database, and data export into other research documentation or library systems is possible. Last but not least, a financial profit resulting from a publication-dependent allocation of resources is a very important benefit, in any case for the successful groups.

The database must support a wide range of publication types, including less "official" publications like internal reports or academic theses, to allow both an operation as a knowledge base and to determine evaluation data. It must provide the possibility to search for information and permit a simple extraction of counts and lists of publications based on a variety of query criteria. It must allow selection, grouping, listing, and rating of publications according to their types and properties, and according to various attributes of their publication media. This implies a genuine relational database structure, where each item of a publication entry is located in an individual field of a database table.

A publication jointly written by several authors affiliated with different organizational units is supposed to appear in the publication lists or evaluation data of each of the authors, and of each of the units to which the authors belong. To allow the selection of all publications of a particular group or institute, the names of persons must reside in a separate table of a relational database. This table of persons is linked to the table of publications, and has references to the groups and institutes to which these persons belong (Figure 1).

A consequence of this approach is that users must select the names of the authors from a list during the creation of a publication entry. For reasons of uniformity, the same applies to the names of editors of books or conference proceedings, of the reviewers or supervisors of doctor's or diploma theses, and of other persons involved in publications of some special types. Obviously, users must be able to add new records to the persons table when creating a publication entry.

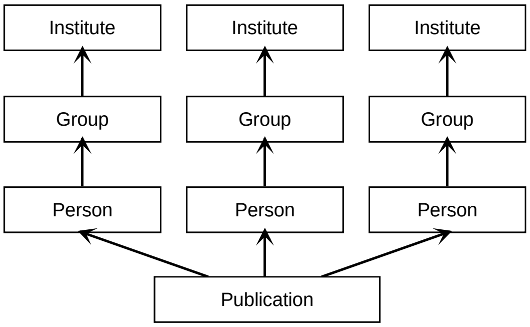


FIGURE 1
HIERARCHIC ORGANIZATION OF PERSON RECORDS

The maintenance of the information required for judging the quality of publications in evaluation schemes should be as easy as possible. It would not make sense to have the SCI (Science Citation Index) status of a publication or the impact factor of the journal in which it appeared entered separately for each publication. These are properties of a "publication medium" (e.g., the journal), which properly belong into a publication medium record (Figure 2). Similar to the names of authors, a suitable publication medium record has to be selected from a list; media are added to this list if they are not yet in the database. It should also be possible to tie together publication media with similar properties and regard them as belonging to a specific "media type" that, in turn, determines their "weight" in an evaluation. For example, "journals listed in the SCI with an impact factor greater than 1" may constitute a particular media type. Journals and, e.g., conferences obviously cannot share media types; they therefore constitute different "media classes". The media classes recognized in the publication database are journals, publishing houses (for books and contributions to books or proceedings volumes), events (for talks or poster presentations at conferences or other scientific meetings), and patents. The publication media concept is not used for some publication types like academic theses or internal reports.

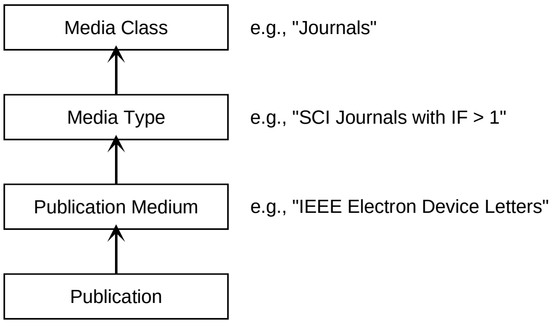

Media Class
e.g., "Journals"
Media Type
e.g., "SCI Journals with IF > 1"
Publication Medium
e.g., "IEEE Electron Device Letters"
Publication

FIGURE 2
HIERARCHIC ORGANIZATION OF PUBLICATIONS

For evaluation schemes, our publication media concept greatly facilitates the quality control of the data: Instead of looking at classifications in hundreds of publication entries, only the classifications of the publication media require checking. Particularly in the case of journals and publishing houses, the number of publication media grows only slowly after an initial phase, and it is easy to look up these newly added media in the proper databases, under certain circumstances even automatically.

Different types of publications imply different items of information in their database records, and different output formats. For example, contributions in a printed proceedings volume usually have consecutive page numbers, while those on a proceedings CD often do not. It makes sense, therefore, to define “publication types.” A publication type determines not only the number and meaning of data fields and the output data format; it also controls the media class to which the publication media offered for selection must belong.

This structure results in the ER (*entity-relationship*) diagram shown in Figure 3, which is a simplified representation of the actual table structure of the publication database. Currently, the database comprises about 50 tables most of which are related with one another.

Figure 3 does not show the numerous tables that hold auxiliary information such as the formatting of the reference output, the grouping of publication types in publication lists, or the evaluation queries and results, and it also shows only one relation that determines the “owner” of a publication record (i.e., the person who created the entry). All tables that can be modified by regular users hold similar fields in addition to “owner” fields that permit to determine the last person who changed the record. The core table structure as shown in Figure 3 has remained unchanged since the beginnings of the database; however, added functionality necessitated many new fields in some of these tables, and additional auxiliary tables.

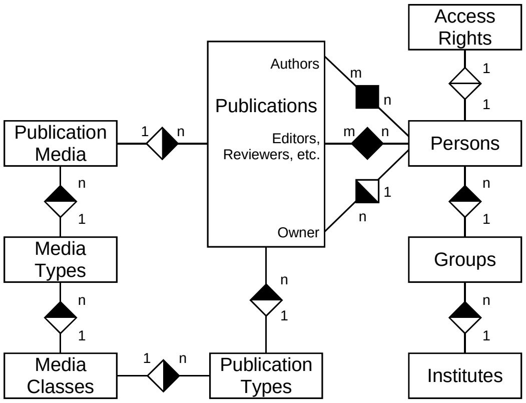


FIGURE 3
SIMPLIFIED ER (ENTITY-RELATIONSHIP) DIAGRAM OF THE PUBLICATION DATABASE

Relational databases for holding publication data are, of course, state of the art. Our database, however, distinguishes itself from other publication collections by the depths of the hierarchies of publication and author records. Such a structure is indispensable for ranking publications by their quality with an arbitrarily fine granularity, and for reliably carrying out queries for the publication output of persons, small groups, institutes, and entire faculties. In contrast, the large publication collections with the essential purpose to provide scientific information have no need for such a detailed structure.

THE IMPLEMENTATION OF THE PUBLICATION DATABASE

General Software Structure of the Publication Database

A Web-based approach appeared favorable over any other client-server solution for the replacement of *Access* whose shortcomings in a multi-user environment dictated a different concept in any case:

- The lifetime of client software is often determined by the lifetime of the operating system or application under which they have to run, usually just a few years. We were looking for a sustainable solution for the publication database.

- There is a wide range of hardware and operating system platforms at a university. This practically rules out dedicated LAN-based clients.
- Web-based systems allow concentrating data processing at the server side, where a well-controlled, well-known and secure environment exists.
- In a Web-based system, client and server software are practically independent from one another. An upgrade on the client side does not imply an upgrade of the server, and vice versa.
- Maintenance should be easy. Upgrading a Web-based system requires no software distribution to the clients.
- Using the database as a knowledge base implies external access to the publication information. This calls for a web interface anyway.
- A web interface also allows the implementation of web services. This can facilitate an integration of the database with other related systems with a possibly widely differing software structure.

In general, using conventional web browsers as clients and the HTTP or secure HTTP protocols for transport makes the database platform-independent and worldwide accessible. Browser-independent programming is mandatory since university staff tends to use a variety of browsers, including some “exotic” species.

For primarily financial, but also for technical reasons, we chose a LAMP structure for the database server with client-based JavaScript for local pre-processing. The program structure chosen keeps most of the processing in the server-based PHP code. This facilitates software management and provides a secure and reliable processing environment. In particular, all potentially security-related functionality resides in server-side PHP. Most of the JavaScript code in the publication database serves only to enhance the usability of the user interface, for example by presetting certain form elements after modifications of other elements. Another important JavaScript-based feature is to check the completeness of an input form for a quick feedback to the users if required data fields are missing. (In order to prevent errors caused by faulty browsers, the server subsequently checks the syntactical correctness of the entered data for a second time.) Although the client-side code uses only the most established JavaScript features, problems with some browsers required a conversion of the initially rather extensive client-side JavaScript data pre-processing code into PHP code wherever possible. The introduction of new browsers calls for repeated testing of the JavaScript and HTML rendering functionality. Occasionally, browser bugs or a non-standard browser behavior made code modifications necessary. With one exception – the display of Greek characters – no browser-dependent programming was needed, though.

Various entry points permit access to the publication database:

- An “administration module” for data input and maintenance with authenticated admission;
- several interactive public interfaces that allow searching for publications and/or creating publication lists of persons, groups, or institutes, with optional restriction to arbitrary time ranges and publication types;
- functions that dynamically create HTML pages with publication lists defined by their call parameters in a widely adaptable design for the inclusion on other web sites;

- features to export publication data in HTML, ASCII text, TeX or XML formats; and
- simple web services that prepare data output in various formats on demand, based on diverse dynamically chosen selection criteria.

The publication database not only provides web services, it also invokes web services offered by other systems. This approach allows portable and platform-independent real-time data exchange with other databases and results *de facto* in an integration of research-related data collections.

The administration module requires client-side JavaScript and, in its latest versions for its full functionality, a JavaScript 1.3 capable browser, e.g., at least *Internet Explorer* 5.5 or *Netscape* 6. (With some restrictions, even older browsers such as *Internet Explorer* 5 and *Netscape* 4 are sufficient.) In contrast, the public interactive interfaces are also operational without JavaScript, although they have a smoother user interface on JavaScript-enabled browsers; in fact, even *lynx* can work with the public interfaces. Currently, the administration module supports German only, but it permits the creation of publication lists in English and German. The public interfaces are in English and German, and the web services likewise provide bilingual data where necessary.

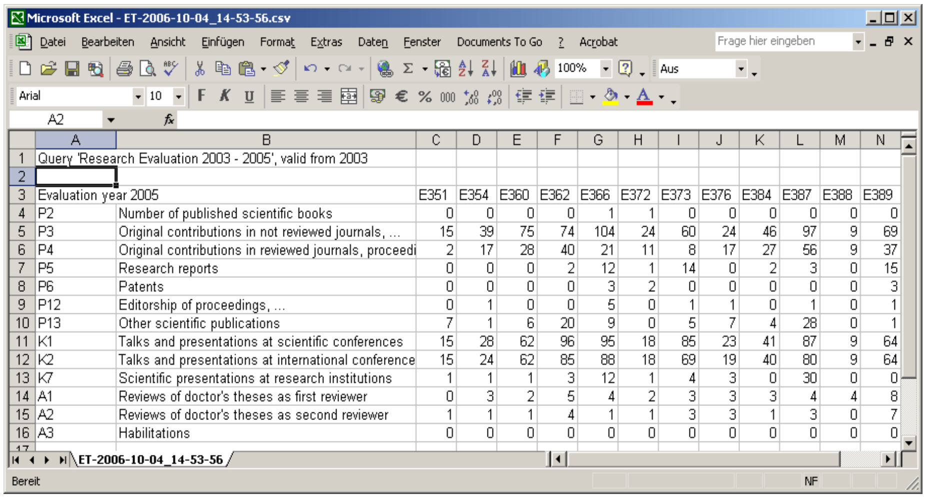
The Administration Module

The authenticated administration module of the database features multi-level access privileges. Users can have permissions to edit their own publication entries, those belonging to their groups, or to their institutes. “Their own” can be understood as entries created by the current user, plus all entries in which this user appears as an author. These rights may extend analogously to the publications of group or institute members. Administrators can edit any entry in the database, plus administrative parameters. Separate privilege attributes permit users to change evaluation-specific parameters or to perform resource-intensive complex evaluation queries. Since permissions for editing a publication also depend on the relation of the user of the administration module to at least one of the authors, the table where the access rights are stored is closely linked to the table that holds the names of persons appearing in publication entries (see Figure 3).

The administration module allows not only the maintenance of the publications table, but essentially of all the tables shown in Figure 3, plus a number of tables not shown there. It has facilities for the on-demand creation of publication lists or export files in various formats (HTML, ASCII text, TeX or XML). Since all kinds of administrative queries must be possible, there are many more adjustable parameters than in the public interfaces that control the search for and selection of publication records. These parameters may pertain to the contents and properties of the publication records as well as to information associated to them such as various legally required or internally defined classification schemes.

The database supports the addition of keywords and abstracts in English and German into the publication records, and permits to upload files of electronic versions or to reference them via web links. Actually, users may upload or reference three files for each publication record: A publicly visible version that is feasible, if there are no copyright restrictions to a publication, a “hidden” version, plus an additional file possibly needed for the validation of the publication record. The latter two files are only accessible from within the administration module. This allows the validation of publication entries using copyright-protected electronic versions that must not be made publicly visible.

The publication database creates statistics and evaluation data according to two different schemes. One scheme accounts for the “official” evaluation algorithms, essentially simple counts of publications in specific categories. An experimental algorithm [3] allows a more detailed weighing of publications. The statistical inquiries the publication database must support are frequently rather complex and may consist of a large number of different individual database queries which must be repeated easily and reproducibly for a large number of organizational units. (An important official Austrian research evaluation involves almost 100 separate database queries that are repeated for each of the more than 80 institutes of our university.) A special function in the administration module allows easy definition and modification of the queries; they are stored in special database tables. Simple queries may consist of an arbitrary number of about 35 conditions, which are AND-combined, and select publications that belong to one of a set of specified publication and media types. The conditions may pertain to properties of the publications, of the publication media, or of the authors. Complex queries are an OR-combination of several simple queries. Only administrators may edit the queries, but any user of the administration module can inspect them and carry them out one by one. For selected users, a special page is available which allows bulk execution of a set of queries applied to a range of organizational units; the results of such queries are available in a CSV format compatible with, e.g., *Microsoft Excel* (Figure 4).



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Query 'Research Evaluation 2003 - 2005', valid from 2003													
2														
3	Evaluation year 2005		E351	E354	E360	E362	E366	E372	E373	E376	E384	E387	E388	E389
4	P2	Number of published scientific books	0	0	0	0	1	1	0	0	0	0	0	0
5	P3	Original contributions in not reviewed journals, ...	15	39	75	74	104	24	60	24	46	97	9	69
6	P4	Original contributions in reviewed journals, proceedi	2	17	28	40	21	11	8	17	27	56	9	37
7	P5	Research reports	0	0	0	2	12	1	14	0	2	3	0	15
8	P6	Patents	0	0	0	0	3	2	0	0	0	0	0	3
9	P12	Editorship of proceedings, ...	0	1	0	0	5	0	1	1	0	1	0	1
10	P13	Other scientific publications	7	1	6	20	9	0	5	7	4	28	0	1
11	K1	Talks and presentations at scientific conferences	15	28	62	96	95	18	85	23	41	87	9	64
12	K2	Talks and presentations at international conference	15	24	62	85	88	18	69	19	40	80	9	64
13	K7	Scientific presentations at research institutions	1	1	1	3	12	1	4	3	0	30	0	0
14	A1	Reviews of doctor's theses as first reviewer	0	3	2	5	4	2	3	3	3	4	4	8
15	A2	Reviews of doctor's theses as second reviewer	1	1	1	4	1	1	3	3	1	3	0	7
16	A3	Habilitations	0	0	0	0	0	0	0	0	0	0	0	0

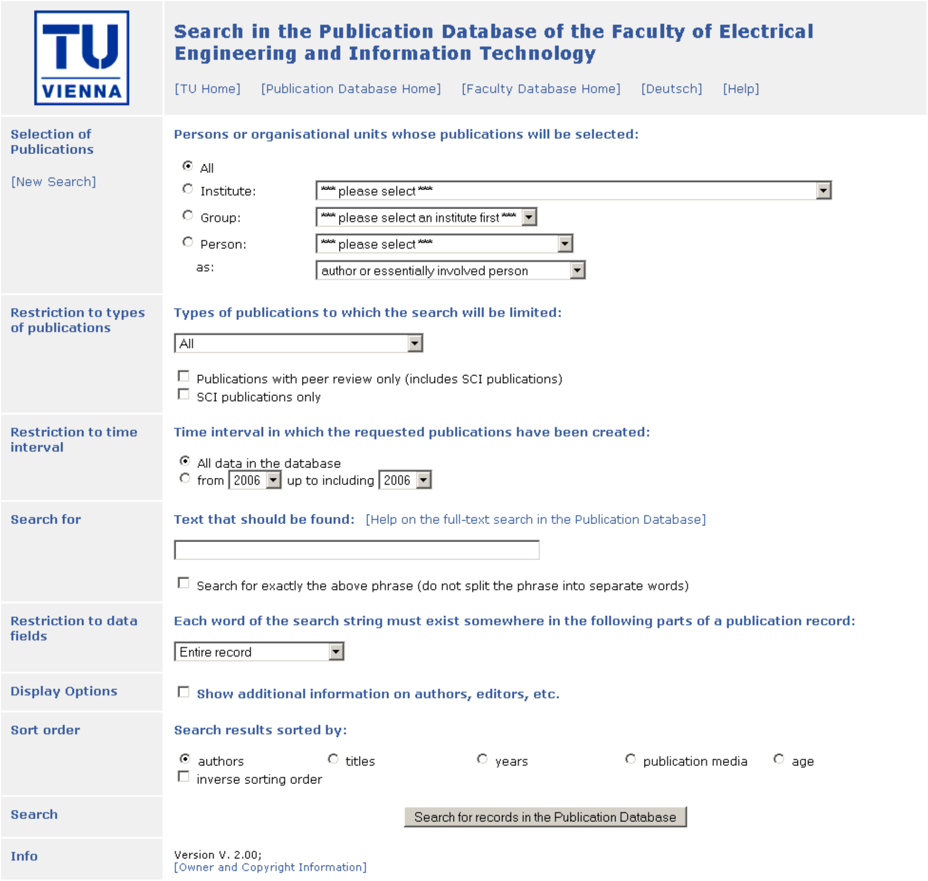
FIGURE 4
RESULTS OF A STATISTICAL QUERY FOR ALL INSTITUTES OF A FACULTY IN AN *EXCEL*-COMPATIBLE CSV FILE

The design concept that permits maintaining an unlimited number of groups of evaluation or statistics queries proved to be extremely beneficial: Not only do the legally required evaluation schemes change repeatedly, there are also several other statistical inquiries that may apply to the data of one faculty only or of the entire university. Having a set of versatile queries at hand, and having the possibility to define new queries easily if needed, reduces the time involved for answering specific questions from weeks to hours.

Additional functions of the administration module comprise various database maintenance and integrity testing functions; functions for extracting evaluation data in a special format; and a tool to create URLs for inclusion on other web sites that request a certain selection of publication data from one of the web services of the database. While the URL generator is available to all users of the administration module, only administrators or specially privileged users may access the other functions.

The Interactive Public Interfaces of the Publication Database

The interactive public interfaces permit to set a variety of query conditions, and generally create human-readable lists of the matching publications in HTML format. In addition, the administration module and the web service functions also support ASCII, TeX, or XML-based output.



Search in the Publication Database of the Faculty of Electrical Engineering and Information Technology

[TU Home] [Publication Database Home] [Faculty Database Home] [Deutsch] [Help]

Selection of Publications [New Search]	Persons or organisational units whose publications will be selected: ● All ○ Institute: *** please select *** ○ Group: *** please select an institute first *** ○ Person: *** please select *** as: author or essentially involved person
Restriction to types of publications	Types of publications to which the search will be limited: All □ Publications with peer review only (includes SCI publications) □ SCI publications only
Restriction to time interval	Time interval in which the requested publications have been created: ● All data in the database ○ from 2006 up to including 2006
Search for	Text that should be found: [Help on the full-text search in the Publication Database] □ Search for exactly the above phrase (do not split the phrase into separate words)
Restriction to data fields	Each word of the search string must exist somewhere in the following parts of a publication record: Entire record
Display Options	□ Show additional information on authors, editors, etc.
Sort order	Search results sorted by: ● authors ○ titles ○ years ○ publication media ○ age □ inverse sorting order
Search	Search for records in the Publication Database
Info	Version V. 2.00; [Owner and Copyright Information]

FIGURE 5

FUNCTION FOR THE INTERACTIVE SEARCH IN A FACULTY PUBLICATION DATABASE

Various query functions in the interactive interfaces permit restricting a search to entries meeting certain conditions, e.g., the affiliation of at least one author or essentially involved person to a particular organizational unit; publication years; publication types, and some more (Figure 5). For most publication types only the affiliation of the authors matters; for some, such as academic theses, an entry is displayed, if either the author or the supervisor of the thesis is the selected person or belongs to the unit chosen. All interactive interfaces provide full-text search functions. The full-text search may optionally take into account the entire record including abstracts etc., or only certain fields of the record. The search string may be used literally, or be split into separate words all of which must appear in a publication record.

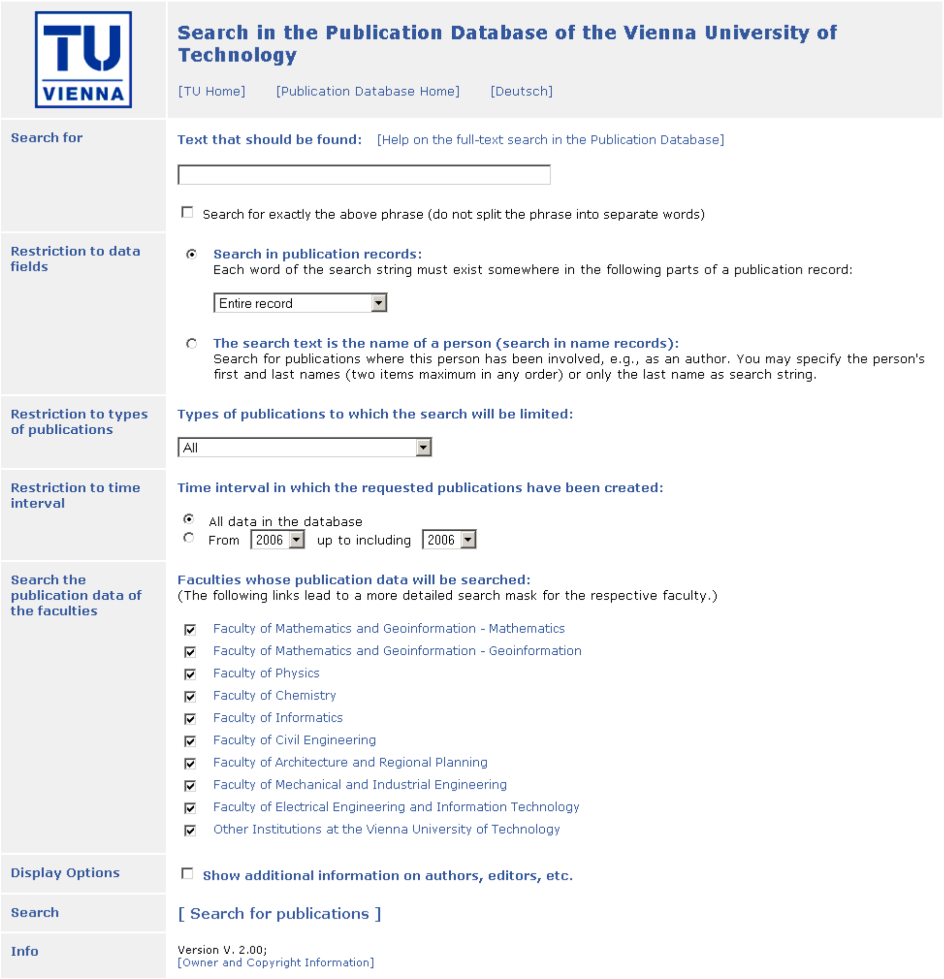
The interface shown in Figure 5 permits access to the database of one particular faculty. For reasons detailed below, we decided to dedicate one separate database to each faculty when the publication database was introduced university-wide. This made a portal necessary that transparently searches all databases in turn (Figure 6). The only purpose of this “global search” function is to allow a full-text search in the database, as opposed to the search function for a faculty database (Figure 5), which also may be used to obtain publication lists for organization units or persons. Therefore, the “global search” has no facilities to limit its selection to a certain organization unit or person; it allows, however, the exclusion of entire faculty databases from the search to restrict its output to the publication records most relevant to the problem in mind. The search algorithms used here are essentially the same as those of the search page of a faculty database.

Faculty Databases and Global Functions

As said above, we decided to implement one separate copy of the database for each of the faculties. The resulting ten databases reside on the same physical server; they are accessed via the virtual web server concept of Apache. Although the maintenance of ten separate databases requires more effort, several reasons favored the solution chosen:

- Users need not enter publication data for a faculty other than their own. It does not matter to them whether they log into a university or a faculty publication database.
- Evaluation and most research documentation data are gathered on a faculty or institute base. Splitting the database in the way chosen does not constitute a problem for these applications.
- Faculties may want to use individual configurations of the database. This is easier to implement in separate copies.
- Users must select author and publication media names from lists of already registered records. These lists grow rapidly: In the database of the Faculty of Electrical Engineering and Information Technology, which holds the publications from 1996 on, there are currently about 6,000 name and 3,400 media entries (for more than 11.000 publications). Using only one database for the entire university would increase these numbers by a factor of 4 to 5, which makes selecting suitable name or media entries from lists ambiguous or at least impractical.
- The drawback of having to search in several databases could easily be resolved by introducing the “global search” portal shown in Figure 6. There are several other global functions for certain administrative tasks.

Apart from a single file that defines the (very few) configuration parameters specific to one particular faculty database, all copies of the database use the same set of PHP, HTML, and image files. This reduces software updates to a copy operation in batch mode, and hence makes them rather straightforward.



TU VIENNA	**Search in the Publication Database of the Vienna University of Technology** [TU Home] [Publication Database Home] [Deutsch]
Search for	**Text that should be found:** [Help on the full-text search in the Publication Database] ▭ □ Search for exactly the above phrase (do not split the phrase into separate words)
Restriction to data fields	● **Search in publication records:** Each word of the search string must exist somewhere in the following parts of a publication record: Entire record ▾ ○ **The search text is the name of a person (search in name records):** Search for publications where this person has been involved, e.g., as an author. You may specify the person's first and last names (two items maximum in any order) or only the last name as search string.
Restriction to types of publications	**Types of publications to which the search will be limited:** All ▾
Restriction to time interval	**Time interval in which the requested publications have been created:** ● All data in the database ○ From 2006 ▾ up to including 2006 ▾
Search the publication data of the faculties	**Faculties whose publication data will be searched:** (The following links lead to a more detailed search mask for the respective faculty.) ☑ Faculty of Mathematics and Geoinformation - Mathematics ☑ Faculty of Mathematics and Geoinformation - Geoinformation ☑ Faculty of Physics ☑ Faculty of Chemistry ☑ Faculty of Informatics ☑ Faculty of Civil Engineering ☑ Faculty of Architecture and Regional Planning ☑ Faculty of Mechanical and Industrial Engineering ☑ Faculty of Electrical Engineering and Information Technology ☑ Other Institutions at the Vienna University of Technology
Display Options	□ **Show additional information on authors, editors, etc.**
Search	[Search for publications]
Info	Version V. 2.00; [Owner and Copyright Information]

FIGURE 6

INTERACTIVE SEARCH IN THE ENTIRE PUBLICATION DATABASE ("GLOBAL SEARCH")

EXPERIENCE WITH THE PUBLICATION DATABASE AT THE TECHNICAL UNIVERSITY VIENNA

The quality and reliability of the data collected in the publication database depend on two important factors, user acceptance and quality control by organizational and technical means. Bad user acceptance results in careless entries that in turn demand more convoluted quality control mechanisms. Providing sufficient additional benefit to the users to increase their acceptance of the system therefore enhances the effect of the quality control procedures.

User Acceptance

Users at the university initially regarded the publication database as an instrument designed to increase rather than reduce their workload. Their first reactions to its introduction ranged between suspicion and hostility. It was important to point out to them that in future all publication-related evaluation data would come from the database, thus sparing them several such surveys per year. Furthermore, there was the increased visibility of their work, and the additional benefit of on-line publication lists and queries. Derived from the database data, the Faculty of Electrical Engineering and Information Technology introduced a financial bonus for institutes and first authors of high-quality publications, which was not only a strong incentive for publishing and officially documenting published work, but also a tangible benefit that made the reception of the database much more favorable. This bonus provably enhanced the data quality and completeness.

There are obviously “cultural” differences between the faculties even at a university with exclusively technical orientation, which resulted in widely differing expectations and wishes, when we introduced the database university-wide. There were common requests from institutes with existing publication collections of some kind for an import function. We developed a tool for data import in a variety of formats; however, it was hardly used when it became available. Still, many institutes have voluntarily entered their earlier publications manually meanwhile to obtain complete publication lists for their web sites although the university only required the registration of new publications to the database.

After the initial opposition had subsided, people quickly learned to take full advantage of the database. As the number of publication entries grows, more and more institutes and groups use the database as a source for publication lists displayed on their web sites. The database provides these lists through a number of web services. Lately, the XML service has found increasing acceptance by groups that not only process the XML data for custom-designed output on their web sites, but also want to create publication references in formats not directly supported by the publication database, such as BibTeX.

User Interface

In addition to proper fault-free operation of the publication database and the implementation of new functionality necessitated by law, by the university authorities, or by its author’s wish to make optimal use of the data in the database, the usability of its user interface has been the most important design issue. Often, seemingly insignificant features greatly facilitate work for the users, e.g., the possibility to limit searches to entries that still require some kind of action, or to sort entries by age (with the latest on top of the selection list). Occasionally, log-files allowed insight into user behavior, which resulted in a re-design of some functions. It also took plenty of real-user experiences to find a proper strategy, when to warn users that they were using selection restrictions to the data they were operating on, and when not to bother them with a warning popup. In some cases, program messages that appeared clear enough to a large part of the users needed re-wording, because some groups of users – apparently with a slightly different faculty “culture” – consistently misunderstood them. Feedback from the users is generally taken very seriously; it has greatly contributed to the user-friendliness of the database. The fact that people with a wide range of backgrounds – from secretaries to scientists – have to work easily with the database imposes very high demands on the clarity and usability of its user interface, which also distinguishes it from the large publication collections, which usually address specialists only.

Quality Control

Several automated features and human actions guarantee high data quality, which is of equally high importance for both research documentation and evaluation purposes: Algorithms test for the proper contents of required fields and check for duplicates of new or existing entries. The latter is particularly important because several groups may attempt in parallel to enter the same publication that they had created jointly. The database reports a possible duplicate, if at least two of four properties – titles, lists of authors, publication media, and page numbers or counts – match for two entries with compatible publication types. Lists of authors are created by selecting names from a database table; it is sufficient to test them for identity, which also applies to page numbers or counts. Title and media name strings, in contrast, may differ even for genuine duplicates due to typing errors or abbreviations; a simple test for identity is not sufficient in this case. A naive approach that considers titles as matching, if one title string completely contains the other, was more efficient in finding duplicates than the test for identity, but still far from satisfactory. Therefore, we introduced a “similar string” algorithm for the comparison of titles and media names. The most efficient approach to search for similar strings in PHP [4] is the Levenshtein algorithm [5, 6], which returns the number of characters that have to be added, changed or removed to transform one of the strings into the other. A Levenshtein distance of less than a string length dependent limit constitutes a match. However, the Levenshtein algorithm is rather resource consuming, which matters even more, because – in contrast to the simpler approaches described above – it must be implemented in interpreted PHP program code, rather than in (faster) SQL queries. The publication database uses a smart restriction to those publication types and publication years where duplicates might perceivably exist to make the performance of this algorithm acceptable for routinely use. In fact, the duplicate tests performed when a publication record is stored after editing are not even noticeable as a delay. Reports of duplicates are only warnings without automatic consequences; the decision whether reported possible duplicates are real ones, and which consequences have to be taken, is left to the user or administrator who initiated the check.

In addition to the automated tests, a specifically assigned person validates entries of print publications based on submitted reprints that may optionally be in electronic form. Finally, a group of senior researchers checks the semantic correctness of publication entries and their proper association to media types.

The Publication Database as a Knowledge Management System

After four years of university-wide operation, the database holds an impressive amount of information. By April 1, 2007, there were 55,000 records for publications, 28,000 records for persons, and 19,500 records for publication media, with a yearly growth of 10,000 publication entries. About 42,000 search operations or requests for publication lists are carried out per month, and there are 350,000 visits and 5,400,000 page hits at the database web site per year.

We already mentioned that many researchers and institutes at our university obtain the publication lists displayed on their web sites and included in their reports or project proposals from the publication database. While some institute web sites simply provide a link to one of the web services of the publication database, other web sites request, store and process publication data for embedding them in their own pages. A considerable fraction of the visits to the publication database is by institute web sites. The public interactive interfaces of the database are obviously well accepted. Since the database documents our university's work only and there is no intention to compete with the large publication collections, we consider the current rather simple search facilities sufficient.

Furthermore, the university library periodically imports the data collected in the database into their own library system [7]. In addition to the basic publication reference data, the university library receives the contents of the abstract and keyword fields, and the references to public and hidden files. Abstracts and keywords are transferred into the library system, and referenced files are copied to a literature server where appropriate. In addition to serving as a knowledge management system on its own, the publication database also acts therefore as a knowledge collection tool for the university library.

The publication database is one of several systems at the Technical University Vienna that document various aspects of research and teaching. For historical and technical reasons, these systems are separate from one another, but they are closely interconnected. For example, the publication database permits to associate publications with projects, which reside in a separate database. Web pages or web services on either side allow displaying projects linked to a particular publication, and vice versa. The publication database has to maintain its own tables for authors and users, because about 80 percent of the person entries in the publication database refer to external authors rather than university staff. However, it obtains staff IDs from the university's staff database via a web service that is invoked if a person is declared a member of an organizational unit of the university when the entry is made in the publication database. The concept of using separate but strongly interoperating databases for separate tasks, rather than a large unified database, has the advantage that the individual databases can be uncompromisingly optimized, and, if necessary, upgraded or replaced without much adverse effect on the entire system. From the point of view of external visitors, the interoperating databases behave like a single complex database.

The Publication Database as an Evaluation Tool

The design of the publication database allows a multitude of evaluation queries, involving a large variety of query conditions. The official Austrian evaluation schemes tend to be complex, sometimes with previously unexpected criteria. Even the most complex queries are possible at “key press”, reliably and reproducibly, for any range of organization units, once provisions have been made for entering all required details, the database queries have been defined, and the publication data have passed the quality control mechanisms described above. Between three and six different sets of statistical queries are routinely in use. It normally takes just a few minutes to generate the data that would have involved the cooperation of hundreds of scientists, and hundreds of hours of their combined working time, without the publication database.

CONCLUSIONS

The publication database presented in this paper has been in use at the Technical University Vienna for seven years, first at the Faculty of Electrical Engineering and Information Technology only, later at the entire university. It has gradually grown during this time from a stand-alone evaluation instrument with the facility to generate publication lists to a comprehensive knowledge base for publication data that closely interacts with several other related databases. University institutes, external visitors, and, last but not least, robots of search engines increasingly make use of its facilities, which contributes to an enhanced visibility of the database contents in the scientific community, and to a growing acceptance by researchers at our university. Meanwhile, the publication database has spread beyond the Technical University Vienna: At the end of 2006, we implemented it at the *Austrian Research Centers* (ARC), a commercial research institution, where it replaces a Hyperwave-based [8, 9] application whose structure had made it unsuitable for the current evaluation requirements. More than 11,000 publication entries have been successfully migrated from the old to the new publication database.

REFERENCES

1. "INSPEC – Bibliographic Database for Physics, Electronics and Computing Research," The IET, *http://www.iee.org/publish/inspec/* (accessed May 20, 2007).
2. "Ei COMPENDEX," Elsevier Engineering Information, Inc., *http://www.engineeringvillage2.org/* (accessed May 20, 2007).
3. K. Riedling, "Design and Implementation of a Publication Database for the Vienna University of Technology," *Proceedings of the Vienna International Conference on eLearning, eMedicine and eSupport (VIEWDET 2003),* Vienna, Austria, 2003.
4. G. Hojtsy (ed.), "PHP Manual," PHP Documentation Group, *http://www.php.net/docs.php*. Accessed May 20, 2007.
5. V. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals," *Soviet Physics Dokl*., Vol. 10, 1965, pp. 707– 710.
6. A. Bogomolny, "Distance Between Strings," *http://www.cut-the-knot.org/do_you_know/Strings.shtml* (accessed May 20, 2007).
7. H. Hrusa, Ch. Kirschner, F. Neumayer, "Datenimport aus der TU Publikationsdatenbank in den Aleph-Bibliothekskatalog," *Mitteilungen der VÖB*, Vol. 58, No. 2, 2005, pp. 21 – 29.
8. "Hyperwave," Institute for Information Systems and Computer Media, Technical University Graz, *http://www.iicm.tu-graz.ac.at/liberation/library/reports/rp_feedback/n39/n71* (accessed May 20, 2007).
9. "About Hyperwave," *http://www.hyperwave.com/e/about/* (accessed May 20, 2007).

Karl Riedling was born in Vienna, Austria, in 1948. He obtained his diploma and doctor's degrees in Electrical Engineering at the Technical University Vienna in 1972 and 1979, respectively. He assisted in the realisation of the first Austrian university-based semiconductor technology laboratory at this university, where he worked on the technology of indium antimonide, and on the application of ellipsometry in microelectronics technology. As a postdoc at IBM East Fishkill in 1980/81, and as a visiting scientist at Arizona State University between 1985 and 1987, he contributed to the automation of the Czochralski growth processes for silicon and gallium arsenide single crystals, respectively.

In 1988, he obtained the *venia docendi* for *Technology of Microelectronic Components*, and is now an Associate Professor. His current research activities are in the area of the presentation of scientific results.

Siegfried Selberherr was born in Klosterneuburg, Austria, in 1955. He received the degree of Diplomingenieur in electrical engineering and the doctoral degree in technical sciences from the Technical University Vienna in 1978 and 1981, respectively. Prof. Selberherr has been holding the *venia docendi* on Computer-Aided Design since 1984. From 1988 to 1999 he was the Head of the Institute for Microelectronics. From 1998 to 2005 he served as Dean of the Faculty of Electrical Engineering and Information Technology. His current research topics are modeling and simulation of problems for microelectronics engineering.