nary (to be accurate, the descriptors corresponding to the terms) are included in the document profile. Therefore, the automatic indexing of a book may result in a document profile containing almost all descriptors of the descriptor dictionary. This means that this document will be given as output to any query of any user of the system. Moreover, this means that books do not have to be indexed at all and should just be given to any user as output. The same situation may occur for certain forms of technical references and scientific reports. Clearly, this situation is not satisfactory for the users and, as a result, for the IR developers. This circumstance can also apply to the indexing of journal articles, which make up (according to some estimates) the main part of the document collection that interests users. Although document profiles of articles are not as large as those of books, they still often contain a considerable part of the descriptor dictionary. Therefore, the increase in the document profile due to automatic indexing is a fact. That is why functioning systems with automatic indexing are oriented to indexing abstracts. Nevertheless, some researchers think that even the automatic indexing of abstracts provides document profiles that are too large. But what is a *large* document profile and what size should be judged normal? On what does the evaluation depend? Let us answer these questions and clarify the reasons for the developers' dissatisfaction.

To begin with, automatic indexing is not isolated and independent from other processes in the system. It is only an element in the whole complex of interrelated elements called the IR system. This interrelation is very important. When creating an IR system, or a system of any other nature, developers are oriented (sometimes deliberately and sometimes just intuitively) to certain internal parameters of a specific process in the system. These parameters and their values are chosen according to the parameter values characteristic of other interrelated processes in the system rather than to the features of this particular process. In other words, parameters are selected according to the operation of the entire system as a whole. This is one of the main methodological aspects of system development. It is this proposition that requires matching each process of the system with other processes that are interrelated with it. This, in turn, implies a certain tuning of the system, which is closest to a certain harmony or a certain imaginary balance. Systems analysis includes the following recommendation for achieving such a balance. Because among numerous processes in the system there are those that can be easily tuned and those with rather restricted tuning capabilities, it is useful to adjust (tune) tunable processes to those that are difficult to tune. In other words, one should first implement the processes that are hard to change and then, using the knowledge about their properties, set the required parameter values for the related processes. To show how this can be applied to the creation of an indexing subsystem for the IR system, first we determine which subsystem is interrelated with it. Next we analyze which process should be tuned according to the properties of the other. Finally, we determine what parameter of the tunable process should be changed and what value of this parameter is preferable.

Remember that the quality of the document indexing process is evaluated not according to the document profile constructed, but according to the output generated by the entire system. However, the system output depends not only on the quality of the document indexing but on other issues as well. Another important factor is the quality of the indexing of queries. In the mid-1960s, having conducted an extended experiment with the MEDLARS system, Lancaster (1968) showed that the main cause of retrieval failures is not a poorly constructed document profile, but a poorly constructed query formulation. In other words, the retrieval quality is affected more by the quality of the query indexing than by the quality of the document indexing. Nevertheless, the influence of document indexing results on the system output is quite obvious as is the interaction between document and query indexing processes in the course of the creation of output. In essence, the interaction of document profiles with query formulations follows from the very idea of information retrieval (see Chapter 5). We are, however, interested in the interaction between the very processes of indexing rather than that between the results of these processes; that is, we are interested in whether these processes influence each other. For this purpose, let us consider factors that are commonly addressed in assessing the quality of document indexing. What shortcomings of document indexing are, by common opinion, influencing the quality of the system output?

The common position of developers was probably most successfully expressed by Lancaster. He indicated that "there are two distinct types or failures determined by shortcoming of the indexing process: (1) failures caused by the indexers' errors and (2) failures resulting from the average number of terms attributed to the document in the course of indexing" (Lancaster, 1968). The first arises in manual indexing methods (when the automatic indexing is used, the quality of the term selection depends on the quality of the system descriptor dictionary, as was shown in Chapter 5), whereas the second cause is directly related to all document indexing methods and approaches. Indicating the importance of this problem, Lancaster wrote, "The most difficult problem of indexing methodology for any system is the computation of the average number of terms attributed to the document" (Lancaster, 1968). But why did developers indicate this problem 30 years ago, at the very beginning of IR systems development? The point is that the very practice of IR systems operation showed that the best results of retrieval are determined by certain values of certain parameters characteristic of both the document profile and the query formulation. These parameters are the average sizes of a document profile and a set of descriptors combined by operand AND in the query formulation. Moreover, the values of these parameters are interrelated; that is, if the value (size) of the parameter is changed, for example, for the document profile, then some changes in the query formulation are also required to provide the best retrieval quality. (Note that the best retrieval quality in the new case may be worse than the quality achieved in the previous case.) This means that if we know the real value of the parameter (the average number of descriptors in the set) for one indexing process, then, in