TYPE OF PROPOSAL: paper
TITLE: DiCoMo: A cost estimation model for digitization projects
KEYWORDS: document digitization, cost estimation, digital libraries

AUTHOR: Alejandro Bia
AFFILIATION: Biblioteca Virtual Miguel de Cervantes, University of Alicante
E-MAIL: alex.bia@ua.es

CONTACT ADDRESS: Biblioteca Virtual Miguel de Cervantes, Universidad de Alicante, Apdo. de correos 99, E-03080, Alicante, España
FAX NUMBER: +34-965909477
PHONE NUMBER: +34-600948601

DiCoMo: A cost estimation model for digitization projects

INTRODUCTION

The estimate of costs of a digitization project is a very difficult task, to say the least. It is difficult to make exact estimates beforehand due to the great quantity of unknown factors. However, the common practice when signing digitization agreements is to set a firm commitment before beginning the works so much concerning economic costs as in terms of time lapses and deadlines. In the same way as for software development projects an incorrect estimate of times and costs produce delays.

Based on methods used for cost prediction in Software Engineering like CoCoMo (Constructive Cost Model; Boehm, 1981) and Function Points (Albrecht, 1981), and using historical data gathered in almost three years of existence of our digital library, we have developed a model for digitization cost estimates (DiCoMo, Digitization Cost Model) for text digitization projects in general.

This method can be adapted to different production processes, like the production of digital texts using scanning plus OCR and human proofreading, or the production of digital facsimiles (scanning without OCR). The estimate done a priori is improved as the project evolves by adjustments done with real data obtained from previous stages. Each estimate is a refinement obtained as a result of the work developed so far.

FACTORS THAT AFFECT DIGITIZATION COSTS

There are many factors that influence the cost of production of a digital object. Both these factor and their effect on costs are difficult to be determined and have to be carefully studied. Among them, we can highlight the individual capacities of the persons assigned to the project and their familiarity with the specific characteristics of the work to be digitally published, its complexity, size, demanded level of quality, the technology used, familiarity with the computer tools to be used, etc.

Main Factors that influence digitization costs:

- Size of the material to publish
- Complexity of the task
- Individual capacity of editors, correctors and scanner-operators
- Special quality requests
- Technological level of the environment

The assigned time impacts mainly in the quality of the product obtained which is notably lowered when the times assigned are unreally short and they force the technicians to work under excessive pressure. This is particularly true for the correction and editing process, where text output from OCR has to be carefully proofread and corrected. This is a delicate craft that takes time and cannot be done under excessive pressure since when it is not properly done obliges to further revision and corrections that have a very negative impact in costs with a final result that turns to be worst than the time initially saved.

Next each one of these factors is described.

Size of the material to publish

Digitization projects, compared to software design projects, have the advantage that we know beforehand the size of the work to be done (namely the number of pages or words to digitize).

The first and easiest way to determine the raw size of a text to be digitized is to count the pages. But pages are not equally dense for all books. We can have an approach to the density by counting the words that fit in a standard page, or the words that fit in a fixed size window, and then assuming that the rest of the pages are similar in this respect.

To count individual words would be more accurate, but not a practical approach at all. Anyway, after the OCR process takes place we will obtain a text file, with errors of course, but nevertheless a text file where we can automatically count the number of words or the size in bytes to have a better measure of the size of the proofread and correction work that comes ahead. So this second measure of the size serves to adjust the initial estimates for higher accuracy.

Complexity of the task

This is by far the most significant modifier concerning cost estimates. In fact, there are many complexity factors that affect each stage of the development process. In the case of the correction stage, which we consider the most critical one, there are various factor to take into consideration:

- the type of text (prose, verse, drama in prose, drama in verse, dictionary)
- footnotes if the number is too high
- quotations in foreign or classical languages (if too many)
- the complexity of the author style and vocabulary
- the quality of the OCR output (few or lots of errors)
- the legibility of the paper copy used as original to produce the digital one from

Concerning markup, complexity varies according to the number and difficulty of the tags to add. Drama, with the need of a *castlist*, *speaker* and *speeches*, require an additional amount of tagging. Verse with split lines is another case of added complexity since special care need to be taken of attribute values where we code which part (initial, middle, or final) of the split line of verse we are tagging.

In the case of manuscript digital facsimile production, a case of increased complexity is when we have to work on rare and valuable originals that have to be handled with special care (wearing rubber gloves for instance) and using a digital photographic camera instead of a flat bed scanner. On the contrary, digitizing unbounded pages using a flat scanner with automatic page feeder would be the easiest case.

Individual capacity of the technicians

In the computer programmers world, individual productivity has been measured extensively. Harold Sackman and collaborators, carried out an experiment in 1968 where they made evident that performance differences registered in individual programmers were much bigger that those attributed to the effect of the working environment. The difference between the best and the worst performance was very high, being the experience a decisive factor. In a later experiment, Sackman observed a variation in the productivity of as much as 16 to 1 (Sackman, 1968).

DeMarco and Lister also discuss the effects of a well integrated group to enhance productivity in their book Peopleware (Demarco+Lister, 1987).

Now back in the field of electronic editing, the results that we have measured comparing correctors performance in tasks of correction of texts output from OCR show us differences in productivity of as much as three to one, depending on the individual ability and speed of correctors. Variations in productivity of this magnitude is significant for cost estimates, so including a parameter to adjust the estimates according to individual skills of the workers assigned seems to be necessary.

Special quality requests

In the case of digital text production, to produce a modernized version from an old text takes additional time and effort. Madison markup for the transcription of a manuscript is another example of additional requested complexity.

So is the case of making high legible digital facsimiles of ancient manuscripts, where special care and fine-tuning of the scanning properties may be needed.

Technological level of the environment

This is a relevant issue when using different technologies or migrating from old to new production tools. When the environment is stable and well known, and the estimate equations are well adjusted for it, there is no need to care of this issue.

COST ESTIMATION MODELS

According to R. Fairley (Fairley, 1985), inside most organizations, the estimation of production costs is usually based on past experiences. Historical data are used to identify the cost factors and to determine their relative importance within the organization. The above-mentioned, is the reason for current project's production costs data to be stored for later use.

We can classify cost estimation methods in two broad categories depending on whether the approach goes from the general to the specific (top-down) or the other way around (bottom-up).

Top-down models first consider the costs at the most general level, being usually based on the exam of the costs of previous similar projects.

Bottom-up models first consider the development cost of each module which are then summed up to obtain the total cost. This technique underlines the costs associated with the independent development of each module or individual component of the project and we believe that it is the most appropriate for projects of digitization of literary works.

It can lead to errors if the time spent in some parallel tasks like version control, backup copies, digital preservation and quality control is not taken into account. The bigger the project, the more important these factors become.

Expert judgement

A top down technique frequently used to estimate costs is based on expert judgement. Expert judgement is based on experience, on previous knowledge and on the commercial sense of one or more individuals inside the organization (Fairley, 1985).

According to Fairley, the biggest advantage of expert judgement which is experience, can end up being its weakness. The expert can trust that a project is similar to a previous one, but it can happen that he/she has forgotten some factors that make the new system significantly different. To compensate for this, and for a possible lack of experience in a particular project type, a group instead of individual experts are used to try to arrive to a consent. The purpose is to minimize individual flaws and the lack of familiarity with some kind of projects, neutralizing personal tendencies.

The DELFI method

The DELFI method is a group expert judgement method that tries to minimize the interpersonal influences in a groups that may spoil the final estimation.

The DELFI technique was developed at the Rand corporation in 1948, with the purpose of obtaining the consent of a group of experts without having the negative effects of group meetings (Helmer, 1966). This technique has been adapted to estimate costs in the following way (see (Fairley, 1985)):

1. A coordinator provides each expert the documentation with the description of the project and a form to write his/her estimate.
2. Each expert studies the definition and anonymously determines his/her estimation. The experts can consult the coordinator, but not other expert.
3. The coordinator prepares and distributes a summary of the estimates done so far.
4. The experts carry out a second round of anonymous estimates, using the results of the previous one. In the cases that an estimate differs much from the others, an anonymous justification by the expert that made it may be requested.
5. The process is repeated as many times as it is considered necessary, avoiding group discussions.

It is possible that after several rounds of estimates there is no consent. In this case, the coordinator will study the causes of such disagreements and try to solve the differences, sometimes by adding new information.

In our case we have preferred not to use this type of expertise based estimation methods, but to take advantage of the expert knowledge to determine the factors that have a significant influence on production times. For digitization costs estimation we preferred to use bottom-up methods, preferably algorithm-based ones, as we explain below.

Work break-down structures

Work breakdown structure or WBS, is a top-down method that helps to plan a project. A work breakdown structure is a hierarchical flowchart where the different parts of a project are established reflecting a hierarchy of products and processes.

A WBS flowchart of processes identifies the work activities and their interrelations. Using the WBS technique, the total cost of the project is calculated by adding the costs of the individual components in the flowchart.

Algorithm based cost models

With algorithm-based cost models the costs are calculated by adding the costs of each of the modules or subparts of the project in a bottom-up fashion. The constructive cost model or COCOMO is a cost model based on algorithms described by B. Boehm (Boehm, 1981).

According to Pressman (Pressman, 1988), the equations for Basic-COCOMO are the following:

E = a KLOC^^b

M = c E^^d

where **E** is the *effort* applied in persons-month, **M** is the development time in chronological months, and **KLOC** are the estimated number of lines of code (in thousands). The factors **a**, **b**, **c** and **d** are from a table given by B. Boehm. This values are for estimating software production costs, and are irrelevant for our purpose. We will only take the spirit of COCOMO and adapt it to estimate digitization costs.

Intermediate-COCOMO takes the following form:

E = a KLOC^^b EAF

where a new *effort adjusting factor* is added. This factor is obtained by first evaluating a set of complexity factors and then extracting the value EAF from a table. Basically, EAF is a value close to 1 that adjust the resulting effort calculation according to the overall complexity determined by various features of the project. Is usually a value between 0.70 and 1.65.

In our digitization cost model (DiCoMo), we use a basic equation similar to Basic-COCOMO, but with an added fixed value **f**:

H = a P^^b + f

We do not have to estimate the size (in pages) **P** which is known beforehand. We directly calculate the number of hours **H**. We add a fixed value **f** which stands for the fixed time necessary for the type of task considered, independent of the size of the task. An example of this is the time needed to adjust the scanning software parameters, which is a fixed time which does not depend on the number of pages to be scanned later.

For an example of this Basic-DiCoMo approach, see figure 1, where an estimation curve (thick line) approaches real data spots (black squares) that represent real measures from the correction stage. The thin straight line represent a linear approach to the spots, while the curve represents the following estimation equation which gives us the estimated number of hours to correct a text given the number of pages:

H = 0.0653 P^^2,091 + 8

Correction-process data for July 2001 (Hours vs. Pages).

Figure 1: Correction-process data for July 2001 (Hours vs. Pages).

For instance, a standard-complexity book of 100 pages will take about 59 hours of correction and markup according to this estimation.

This simplistic approach doesn't take into account the fact that different literary works have different degrees of difficulty at the time of proofreading. This differences in complexity is due to several facts. We have detected the most important ones and added a complexity adjustment factor **c** which leads to our Intermediate-DiCoMo equation:

H = a c P^^b + f

Procedure for the estimate of costs using DiCoMo

1. identify all the subprocesses and all the units (books or other literary works) to be processed.
2. measure the size of each unit and establish the production steps it will undergo from the corresponding processing workflow.
3. specify the complexity factors for each unit.
4. calculate the time each unit will take, as the sum of the estimated times for each production step it will undergo (use the adequate equation for each step together with the corresponding complexity factors).
5. calculate the total time of development for the project.
6. compare the estimate with another, perhaps a top-down one like the *DELFI* technique or *expert-judgement*, identifying and correcting the differences in the estimate.

This version of DiCoMo is a simplified equivalent of the intermediate COCOMO model (Boehm, 1981) used for software development estimates adapted in this case for digitization projects estimates.

Many studies have attempted to relate size oriented methods like COCOMO and function oriented methods like function-points (Albrecht+Gaffney, 1983). We take from the *function points* model of Albrecht (Albrecht, 1981), the idea of modularization according to functions. In our case we consider each production step a functional unit, for which the estimation equation is applied.

CONCLUSIONS

We have designed a cost estimation model for digitization projects based on known software engineering cost models. These methods allowed us to predict the overall time a digitization project would take within a 20% error range.

Digitization projects, compared to software design projects, have the advantage that we know beforehand the size of the work to be done (namely the number of pages or words to digitize). In software engineering we can only guess the total number of lines of code or function points a project will take, and the certainty of the cost estimates will depend largely on this preliminary size estimate.

We verified that the model we propose works well in practice, and can be easily applied to different digital production processes. But the cost equation needs to be fine-tuned in advance for the model to be applied. This requires two things to be done first: on one hand the main objective factors that affect the time required to do the work must be determined and weighed, and on the other hand sufficient historical data must be gathered to fine-tune the parameters of the cost equation. Having this information a cost-equation for the specific production process is easily obtained.

Good expert knowledge of the process facilitates the fine-tuning task and allows for better estimation equations. Nevertheless, the cost-equations can be dynamically improved by re-adjusting the parameters with the new data feed-back from new projects. In this way the estimation model can be incrementally improved.

The production process of digital text books.

The production process of digital text books.

The production process of digital facsimile books.

The production process of digital facsimile books.

Automatic transformation: the one-source many-uses principle.

Automatic transformation: the one-source many-uses principle.