DQ OPTIONS: EVALUATING DATA QUALITY PROJECTS USING REAL OPTIONS

(Research-in-Progress: Cost Benefit Analysis of IQ Improvement)

Monica Bobrowski
Pragma Consultores, Argentina
mbobrows@pragmaconsultores.com

Sabrina Vazquez Soler
Pragma Consultores, Argentina
svazquez@pragmaconsultores.com

Abstract: Data plays a critical role in organizations up to the point of being considered a competitive advantage. However, the quality of the organizations' data is often inadequate, affecting strategic and tactical decision making, and even weakening the organization's image. Nevertheless it is still challenging to encourage management to invest in data quality improvement projects. Performing a traditional feasibility analysis based on ROI, NPV, etc., may not capture the advantages of data quality projects: their benefits are often difficult to quantify and uncertain; also, they are mostly valuable because of the new opportunities they bring about.

Dealing with this problem through a real options approach, in order to model its intrinsic uncertainty, seems to be an interesting starting point. This paper presents a methodological framework to assess the benefits of a Data Quality project using a real options approach. Its adequacy is validated with a case study.

Key Words: Data Quality; Real Options; Project Evaluation

INTRODUCTION

Quality control and management have become competitive needs for most businesses today. Approaches range from technical, such as statistical process control, to managerial, such as quality circles. An analogous experience basis is needed for data quality.

In practice, data quality thereof is low, which weakens the company's position for strategic and operative decision-making in detriment of its image before clients. However, it is hard to justify the need to invest in projects aimed at improving data quality.

The question then is how to justify a preventive approach to these issues? The first approach would imply carrying out a classic feasibility analysis, using ordinary techniques: NPV, PI, IRR. Nevertheless, there are many limitations when applied to the analysis of quality investment projects: the benefits of this kind of projects are usually difficult to quantify economically, basically because they are not direct: they are related to the opportunities they bring about. In addition, part of the economic impact is associated with "prevented problems" cost saving which are difficult to measure, and is not capture by traditional indicators.

Within this context, it seems interesting to use a real options approach ([6], [1], [7]), to model the uncertainty that exists with respect to the subsequent decision. This model also allows to capture the essence of the NEAT methodology ([4]), which presents the need of a diagnosis to assess, based on its output, the convenience of implementing a corrective improvement action on the data and also to establish specific improvement expectations. This model would allow assessing the best investment that an organization can make to improve its data, considering the performance evolution of the quality investment and the future benefit expectations.

There are records of the use of the real options model to assess different software engineering projects ([5], [13]). However, their use to assess the benefits of quality investments has not been studied yet. We believe this model offers an interesting potential that deserves exploration.

The objectives of this paper are:

- To define a methodological framework to assess the benefit of a data quality improvement project by using real options.
- To validate the model proposed by means of a case study.

Note: in this work, we use data and information interchangeably, meaning both raw and processed data. Although there are some differences, they are of no significance in this context.

IT'S VISION OF QUALITY

In recent decades, software has grown to become a vital part of most company's products and services. As pointed out in [9], such growth brings about the responsibility to establish the contribution of software on the organizations' income.

Many of the problems that come up when using poor quality data are well known to software engineers. The NEAT methodology provides a systematic way to determine data quality in order to produce an improvement plan. The output is a diagnosis of the present state of the data quality and an improvement plan that comprises both corrective and preventive actions (in order to maintain the level of quality finally achieved). In particular, NEAT bases its approach on the GQM framework ([8]) for metrics definition.

Still, there are no serious studies aimed at providing a framework to analyze the convenience of investing in data quality improvement. In general, organizations work when they find they have very poor data (lawsuits lodged by clients, returned post, networks that do not match reality, etc.), and try to correct such data because they have no other choice (or they stop using them if they are irreparable). The analysis is ad-hoc and generally speaking only the initiative's cost is assessed, which submits the decision to the resulting amount (high or low) ([10], [14]).

REAL OPTIONS

Stewart Myers coined the term "real option" when developing the idea that financial investments generate real options ([7]). Myers claimed that the valuation of investment opportunities using the DCF traditional

approach disregarded the value of options arising in risky and uncertain projects. This idea was later expanded to any kind of investment decision and corporate budgeting.
According to [1], the real options model is a line of thought which comprises three main components rather useful to managers:

- Options are contingent decisions: an option is the opportunity to take a decision once an individual sees how events are developing.
- Options valuation is aligned with financial markets valuation: the approach uses data and concepts from financial markets to assess complex payments in various kinds of real assets.
- Thinking about options may serve to design and manage strategic investments proactively.

As a matter of fact, not all investment decisions deserve the use of real options ([1]). In some cases, the investment is clearly good or bad, and an analysis based on real options would not change the decision. However, many of such decisions fall into a gray area that requires thorough assessment. A real option analysis becomes necessary when:

- There is a contingent investment decision
- Uncertainty is such that it is convenient to wait and gather more information
- Value seems to lie on future growth possibilities than on direct cash-flows
- Uncertainty is such that flexibility becomes important
- Updates will take place during its development

Real options valuation

In the beginning, real options valuation models assumed that costs were deterministic, while, in practice, costs, as benefits, tend to be uncertain. The time necessary to complete the project is usually uncertain as well. These features are characteristics of the real options model, the valuation of which must incorporate static product life cycles and variable cost structures.

The application of the real options valuation has expanded to appraise intangible assets investments, such as acquisition of knowledge or information and intellectual property, which are usually called virtual options.

Let’s see the assumptions of the Black&Scholes model that are not met in the case of real options:

- The project’s volatility is not constant throughout time
- There is no final expiration date for the option
- Both the underlying asset value and the exercise price (i.e., the project’s development cost) behave stochastically
- Payoffs are not normally distributed
- Real assets do not follow a “random walk”

Given its discrete nature, in the case of the binomial model, the evolution of parameters can be monitored on a step-by-step basis while “unforeseen” changes can be observed. However, this also hinders the application of the model since, given the stochastic nature of many of its parameters in the case of the real options, it may be necessary to analyze several periods, making the construction of the binomial tree difficult.

Real options in IT projects

The application of real options to the information technology field has increased in recent years, the main reasons being twofold:

- A renewed need to justify the convenience of investing in IT.
- The boundaries of traditional project assessment techniques to model an IT investment properly.

[13] Presents an approach to assess design decisions within the context of software development. Authors suggest assessing these principles integrally within the framework of an options analysis, to appreciate their contribution to the project's value. In [2], Benaroch suggests managing IT investment risk within the framework of real options. The author together with Kauffman presents in [3] the application of valuation through real options on an IT investment project: the analysis of the right time to install POS (Point of Sale) debit services in the Yankee 24 banking network in New England. In [12], Schwartz and Zozaya-Gorostiza propose two models to assess IT projects based on whether they imply infrastructure acquisition or development. Authors also suggest a homogeneous framework to incorporate both types of projects.

Staged options

Many projects break down in a number of sequential stages where each step is based on the successful completion of the previous step and the management's possibility to assess the project in each stage ([7]). The benefit of a staged option will only be appreciated once all its stages are fulfilled. Some examples are: investments in new technologies and in R&D.

The assessment of this kind of option depends on the knowledge on the costs and benefits' stochastic processes:

- If they are known are known (or assumed), known closed-form valuation methods may be used.
- If they are unknown, the binomial model offers a viable assessment option.

ASSESSMENT OF DATA QUALITY PROJECTS USING REAL OPTIONS

We present a methodological framework to apply the concept of real options to the assessment of data quality projects. In order to illustrate the concept, we will use the NEAT methodology, identifying its various stages.

There are different types of assessments that can be performed on a project of these characteristics:

- To assess the convenience of a specific investment.
- To assess the maximum convenient investment within a certain framework.
- To analyze different scenarios to establish which one justifies a certain investment.

Moreover, this study can be repeated while the project progresses and uncertainty decreases.

For the application of the real options model, it is essential to identify the sources of uncertainty, which are characteristics of this type of projects:

- State of data –Theoretical scope of the improvement.
- Cost of the improvement.
- Real improvement.
- Contingent projects: some projects may depend on the actual quality achieved and therefore, not be convenient if the improvement does not reach certain levels.
- Benefits to obtain from the projects to carry out.

Although the numerical analysis is vital to draw a conclusion, we do not disregard the fact that many of the values used are predictions (with a higher or lower degree of certainty), estimations and even desires. Hence, we will be more interested in establishing a homogeneous comparison framework of investment alternatives based on the use of real options.

Why use real options?

As mentioned in previous chapters, it is difficult to assess the convenience of a quality investment project (whether of software or data) just by looking at the cash flow directly associated. In this type of project we can highlight the following features:

- They let the organization be ready to carry out actions that they would otherwise not be able to do.
- A great part of the benefits is qualitative and difficult to quantify (what is the value of a robust application?).
- Benefits are not always direct; quality improvement projects establish the starting point to many different initiatives that may or may be not carried out, and they open new opportunities to organizations that were no even considered before. For instance, a flexible software design allows considering software evolution, which would be costly or even impossible otherwise.
- Although they open new opportunities, the benefits still are uncertain (the quality may not raise the expected level, the timing is wrong, etc.).
- There are some economic issues to consider when investing in quality. Not every investment is cost-effective.

Also traditional techniques (DCF, VPN, ROI, IRR) are limited:

- They do not capture the possibility to change the investment sequence.
- They do not consider the option to abandon a project.
- They deal with deterministic and known costs (which is not always the case).
- The discount rate and the future cash flows may be arbitrarily determined, affecting the accuracy of the computation.

These factors make thinking in terms of real options seem a natural alternative for these projects.

Methodological framework

Now, we will propose an assessment life cycle of a DQ project.

This methodology is absolutely simple and general. However, we have listed some basic considerations to make its application successful:

- As in every IT project, the key issue is to understand all the requirements correctly. In this case, this means understanding the project to be assessed, which components follow stochastic processes and which are deterministic, and what all the possible sources of benefit are.
- Future benefits are associated to projects, the costs of which are independent from the data improvement costs.
- The design of the solution must contemplate the possibility of assessing more than one option, whether due to different future projects, diverse investment alternatives, different stages, etc.
- The choice of the assessment model is important to ensure the correctness of the analysis. Nevertheless, given a highly uncertain context with little experience in the use of the technique, it is convenient to promote simpler models (considering their limitations) and, based on the experience, search for models that can better adjust to the problem under resolution.
- Analysis of the output is vital, not only for the project under assessment but also for the lessons learned which would enable to improve the application of the model in the future.

- It is convenient to repeat the analysis while the project progresses and becomes more certain in order to improve decision-making.

A real option for a DQ project based on the NEAT methodology

In order for the proposed methodology not to be so abstract, considering that we are presenting it for a specific type of project (NEAT projects), in this section we will describe how a real option with these characteristics should be produced. However, this does not mean that under different circumstances another type of option would be more convenient. This section may be considered a *partial instance* of stages 1 and 2 of the methodology presented.

Some considerations:

- We leave this last stage of the methodology, monitoring, out of the scope of the projects to be assessed for the sake of simplicity.
- The benefits obtained from this type of project are associated to projects to be carried out in the future, which have their own cost.
- An improvement plan may consist of several tasks, which are optional for the organization, each of them with its corresponding cost and improvement expectation.
- In the case of poor quality data, the company may be losing money for that reason (as we will show in the case study). In this case, the benefit of the data improvement will be to prevent such loss and therefore it will be direct and positive.

At this point, we will present a two-stage analysis that allows to carry out a preliminary assessment to determine if it is worth to implement a data quality improvement project within the NEAT framework and then, a more detailed analysis that covers specific tasks to be carried out. The application of the model adds a few requirements to the methodology, which will be explained later. In the next chapter, a complete case study is presented.

First stage

A first analysis to carry out before the elaboration of the diagnosis may be to consider under what context the diagnosis can be justified. Even without knowing the improvement cost, this means assessing different possibilities with various probabilities, considering probabilities also for the success of the improvement and considering different contingent projects. In this case, and to make matters simple, it is recommended to apply the binomial valuation model and produce an option for each improvement plan cost.

Second stage

The second stage is aimed at carrying out an in-depth analysis of the convenience of the improvement. Here, the steps to follow are similar to the first stage; except the improvement is already broken down and diverse tasks can be combined. It is necessary to establish:

- Different improvement configurations (if they exist)
- Cost and probability of the possible success of each improvement configuration
- Possible cost and probability of success of the project/s and determination of the minimum quality requirements, to establish under which improvement configuration they are feasible.
- Project/s' benefits

For this assessment, we also suggest using the binomial model as proposed by Brach in [7]. It is worth pointing out that this approach, although simple, poses an essential constraint: it does not allow capturing the possibility for the success of an improvement action (or of the plan as a whole) to be partial.

CASE STUDY

In this chapter, we will apply the methodology defined in the previous sections to evaluate a real DQ project. At the present time, the project is in the lessons learned phase, so the analysis allows the upper management to assess its progress.

Actual names and figures were changed in order to preserve confidentiality. These changes don't affect the analysis outcome.

Problem specification

International Petroleum is carrying an information quality assurance program for its geology and geophysics (G&G) information. This program relates with a parallel project whose goal is the classification of the physical data.

The aims of the project are:

- Perform an information diagnosis
- Ensure that the information classification has a minimum quality level
- Ensure that the tools used are adequate
- Standardize the data loaded in the different applications
- Implement a data quality improvement process

Solution design

To evaluate the case, we will apply the proposed two stages. In this section we will design the components that will take place in each of the evaluations.

The following list shows data needed in both stages (money figures in MM US Dollars):

- Risk free ratio: 7%
- WACC (capital cost) 13.5%
- Diagnosis' cost: $0.1
- Diagnosis' length: 2 months

The company only identified the physical data classification project. This is a high yield project. However, the project success depends on having high quality information.

Project Info

- Diagnosis' cost: $0.1
- Length: 5 years
- Scenarios probability: 50%
- Total Cost: $4.21

Project Yield

- Optimistic
    - NPV: $9
    - Revenue: $15
- Pessimistic
    - NPV: $0.5
    - Revenue: 5.35

First Stage

In this stage of the analysis we need to know the investment that the company wants to make in data quality. This level of investment will allow the evaluation of the project feasibility even if the total cost of the data improvement program is still known.

Improvement Project

- Length (diagnosis + improvement): 1 year
- Success probability: 60%

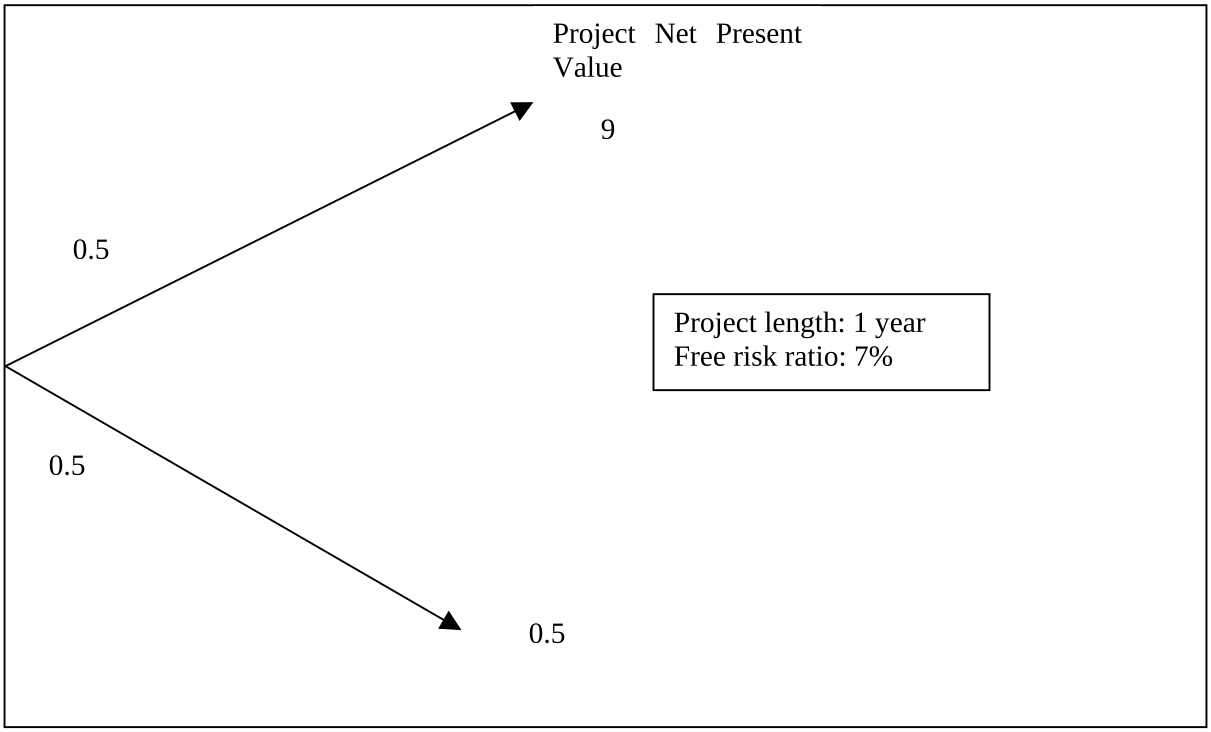

Figure 1 – Project Feasibility

Second stage

This analysis is done after the diagnosis is completed and the improvement plan tasks are identified and estimated. This first analysis will be similar to the previous with the addition of the improvement cost and success probability. A later analysis will identify sequential tasks.

Required data:

- First Stage

  - Length, cost and success probability
- Second Stage
  - Improvement plan (activity based)
  - Activities dependency

Diagnosis result

The following issues were detected during the analysis:

- Problems in roles and responsibilities definition
- Lack of data loading criteria definitions
- Data inconsistencies between different applications

Improvement plan

- Data loading criteria definition and implementation
  - Cost: 1.2
  - Success probability: 70%
  - Length: 4 months
- Data cleansing
  - Cost: 1.5
  - Success probability: 50%
  - Length: 3 months

Global improvement plan

- Cost: 4
- Length: 10 months
- Success Probability: 60%

(Success of the improvement plan depends on the success of the data cleansing and, in equal parts, on the success of the remaining task)

The following diagram shows the first stage design:

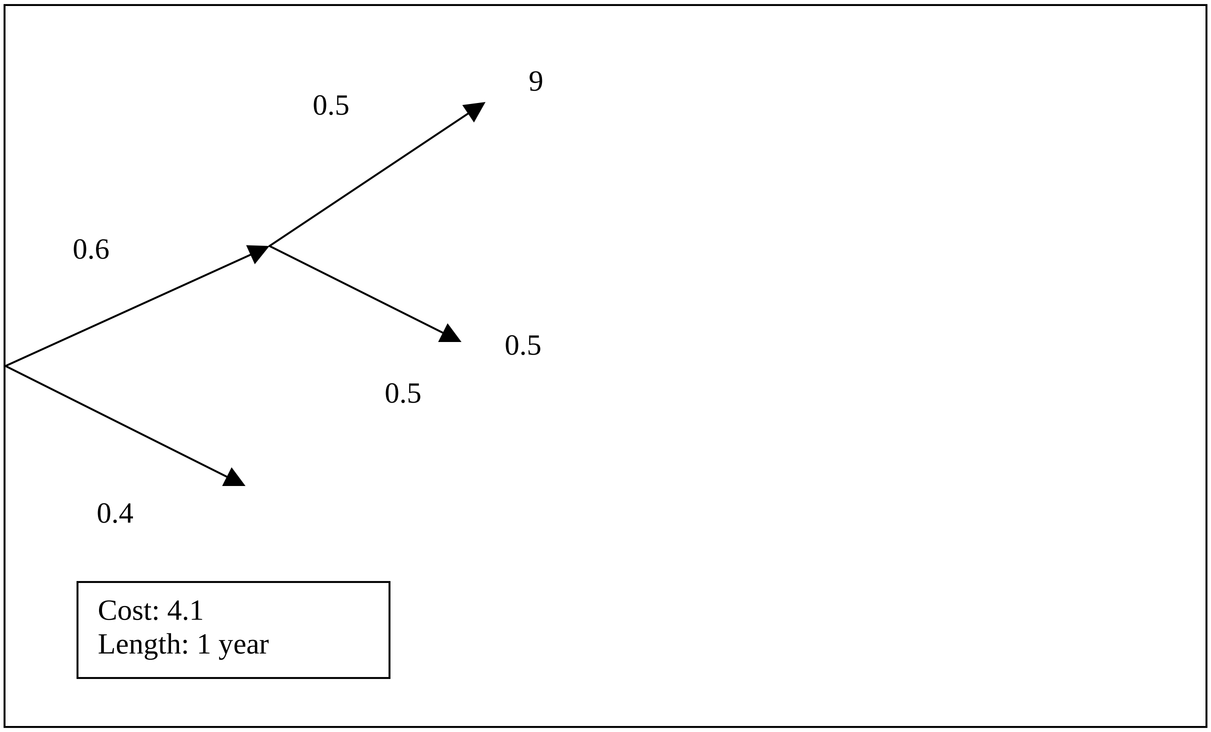

Figure 2 – Overall Improvement

This graph represents the overall improvement project: The first node shows that the improvement will succeed with a 60% probability. If the improvement fails, the overall project will be abandon. If it succeeds, the physical data classification project will be executed, with a 50% success probability.

Implementation

To calculate the value of the options a binomial model was used. All the necessary functions were defined in a Microsoft Excel spreadsheet.

First Stage

Project NVP
$V = (q_{max}*S_{max}+q_{min}*S_{min})$
$V = (0.5*9+0.5*0.5)=4.75$
$\sigma= 4.25$

Risk free probability[1]:
$p = ((1+r_{free}*V)- S_{min}) / (S_{max} - S_{min})$
$p = ((1.07*4.75)-0.5) / (9-0.5) = 0.539$

Critical investment level:

[1] [7] uses this formula to calculate p. This formula is less complex than the used to evaluate financial options. P represents the value that q should have in a risk free scenario.

$K = ((p * S_{max} + (1-p)* S_{min})/ (1+r_{free})^t) -K*(1+r_{wacc})^t)= 0$
$K = ((0.539*9 + (1-0.539)*0.5)/1.07) / 1.135= 4.18$

Second Stage

To evaluate the chained options we will use these formulae.

First analysis:

The previous node evaluation is completed in the same way. First we calculate the node NPV assuming that the immediate node was successful, subtracting the maximum value for the WACC for the number of periods (15 in this case). The minimum value is always 0, because if the task fails the project is discontinued). The expected value is calculated using the success probability. In this case:

Stage	NVP	q (success)	length	Cost	Option Value	P
Project	15 (max) 5.35 (min) 10.175 (exp)	0.5	1 year	4.21	5.96	0.573
Improvement + diagnosis	13.2 (max) 0 (min) 5.4 (exp)	0.6	1 year	4.1	1.27	0.435

Max Value:
$S_{max} = S_{max'} /(1+r_{wacc})^t$
Where $S_{max'}$ is the maximum value of the immediate node
$S_{max} = 15/1.135=13.2$

Project NPV:
$V = (q_{max}*S_{max}+q_{min}*S_{min})$
$V = (0.6*15+0.4*0)=5.4$

Risk free probability:
$p = ((1+r_{free}*V)- S_{min}) / (S_{max} - S_{min})$
$p = (1.07*5.4) / (13.2) = 0.435$

Option Value:
$C = ((p * S_{max} + (1-p)* S_{min})/ (1+r_{free})^t) –K$

$C = ((0.435*13.2 /1.07) -4.1= 1.27$

Second analysis

Stage	NVP	q (success)	length	Cost	Option Value	P
Project	15 (max) 5.35 (min) 10.175 (exp)	0.5	1 year	4.21	5.965	0.573
Data cleansing	13.2 (max) 0 (min) 4.48 (exp)	0.5	3 months	1.5	3.225	0.363
Dataflow, roles & responsibilities	12.78 (max) 0 (min) 2.76 (exp)	0.7	4 months	1.3	1.6	0.231
Data loading criteria	12.22 (max) 0 (min) 1.95(exp)	0.8	3 months	1.2	0.734	0.171

Analysis

After the first stage we can conclude:

- The diagnosis is worthwhile
- If the improvement costs are at most 4 it is convenient to proceed

The second stage gives as a more detailed view:

- It is still valuable to carry out the improvement
- We can observe how the option value increases as long as intermediate milestones are successful and uncertainty is reduced.
- Our model forces staging the improvement project, defining checkpoints to decide whether to abandon it, and consequently lowering the potential loss.

Conducting a NPV analysis (setting the expected DCF):

$NPV = -4.1 + (4.75/ 1.135) = 0.08$

Although the NPV is positive (which is not always the case), based on this number we would not be able to decide whether to abandon the project if one of the stages doesn't succeed. Also, the value is relatively low, so we could be tempted to discard the project. We see that the NPV analysis may force us to discard valuable projects, but also to proceed with a project without reexamining the decision. Finally, the value of the option is substantially bigger then the NPV, because the option analysis captures the "flow of time" that reduces uncertainty and helps to decide whether to abandon or continue.

Conclusions

After developing the case study we can conclude:

- The proposed analysis gives us a global view of the whole project and how much we would be willing to invest in it, considering future expectations.
- The binomial valuation model facilitates the definition of simple spreadsheets to perform the calculations, even when specific software is not available. However, tracking the valuation is a complex process.
- The proposed model forces the definition of go/no go stages, which add flexibility to the management decision process, allowing taking risks and delaying decisions in a controlled framework.
- It would be valuable to enrich the model adding sensibility analysis, scenario analysis, project combinations, investment portfolios, etc.
- The proposed model shows some advantages over traditional indicators, showing some limitations of the classical view. However, more research in this field is needed.
- The chosen valuation method considers neither partial success, nor the possibility to parallelize stages. The steps are sequential and the results are success or failure.
- More experimentation with complex cases is needed.

CONCLUSIONS

This work was aimed at defining a methodological framework to assess the benefit of a data quality improvement project using real options and to validate the proposal with a case study. We have drawn the following conclusions:

- In spite of the relevance of the information, it is difficult for organizations to find an economic justification to invest in data improvement.
- Traditional techniques are very limited, since they do not consider the possibility of changes in the investment sequences, they consider deterministic and already known costs, and may have a certain degree of arbitrariness in the choice of a discount rate, in addition to the determination of future flows.
- Real options are a suitable alternative to reason about quality projects in general and data quality in particular, since they allow to consider uncertainties in terms of costs and benefits, flexibility to decide whether to move forward with the projects, different open opportunities, etc.
- We have proposed a methodological framework to use real options for assessment of data quality projects. The framework is general and we have instanced it for this type of project specifically. We have not made emphasis on the valuation method, since we rather prioritize the underlying reasoning model.
- The application of the methodology to a concrete case allowed proving its simplicity, its easy implementation and also some of its constraints (binary result of activities, arbitrary estimation of some probabilities, etc).
- The methodology proposed does not resolve the difficulties when trying to quantify benefits and opportunities. Moreover, although possible, the analysis of different scenarios may be highly complex and bothersome.

REFERENCES

[1] Amran & Kulatilaka: *Real Options*, Harvard Business School, 1998

[2] Benaroch, M: *Managing information technology investment risk: a real options perspective*, Forthcoming in Journal of Management Information Systems, 2002

[3] Benaroch & Kauffman: *A case for using real options pricing analysis to evaluate information technology project investments*, Information Systems Research Vol 10 N. 1, 1999

[4] Bobrowski, Marré & Yankelevich: *A NEAT Approach for Data Quality Assessment*, In Information & Database Quality - Calero, Genero & Piattini (eds.), Kluwer, 2001

[5] Boehm &Sullivan: *Software Economics , A Roadmap*, ICSE 2000 Invited Paper, 2000

[6] Brealey & Myers: *Principles of Corporate Finance, 6th edition*, McGraw-Hill, 2001

[7] Brach, M.: *Real Options in Practice*, Wiley & sons, 2003

[8] Fenton N. , Pfleeger S.: *Software Metrics, A Rigorous and practical Approach*, 2nd ed, Fenton & Pfleeger- Brooks/Cole Pub, 1998

[9] Kitchenham & Pfleeger: *Software Quality – The elusive target*, IEEE Computer, January 1996

[10] Loshin, D: *The cost of poor data quality*, DM direct, 2001

[11] Redman, T.: *Data Quality - Field Guide*,Digital Press 2001

[12] Schwartz & Zozaya-Gorostiza: *Investment under Uncertainty in Information Technology: Acquisition and Development Projects*, 2001

[13] Sullivan, K.J, P. Chalasani, S. Jha, and V. Sazawal: *Software design as an investment activity: A real options perspective* - in Real Options and Business Strategy: Applications to Decision Making, L. Trigeorgis, consulting editor, Risk Books, December 1999

[14] Trillium Software: *The ROI of Data Quality*, 2002

Acknowledgements: We'd like to thank Daniel Yankelevich, Martin Patrici, and Dafne Bobrowski for their helpful comments. We also like to thank Natalia Del Aguila, form Universidad Torcuato Di Tella, for her helpful guidance during the whole work.