LEVEL

AD A109201

PRESENTATION TO THE

THE SIXTEENTH ANNUAL

DEPARTMENT OF DEFENSE COST ANALYSIS SYMPOSIUM

TITLE: A LIFE CYCLE COST DATA BASE FOR AUTOMATED DATA SYSTEMS

AUTHOR: MR. WALTER J. HOULETTE

HQ AFLC/ACMCR

12 16

012300

HOSTED BY THE COMPTROLLER OF THE ARMY

SHERATON NATIONAL HOTEL

ARLINGTON, VIRGINIA

OCTOBER 4-7, 1981

A LIFE CYCLE COST DATA BASE
FOR AUTOMATED DATA SYSTEMS

Mr Walter J. Houlette
Cost Analyst
HQ AFLC/ACMCR
Wright-Patterson AFB OH

07 JUN 1981

A LIFE CYCLE COST DATA BASE

FOR AUTOMATED DATA SYSTEMS

Mr Walter J. Houlette

Cost Analyst

HQ AFLC/ACMCR

Wright-Patterson AFB OH

INTRODUCTION

When it is necessary to project the cost of a new automated data system (ADS), are your analysts using a base of data derived from the development of your systems? Or are your estimates being made, based upon data from a source outside your organization? Several data bases are made up of data from a wide range of sources. Using these creates many questions, such as: are the data really pertinent, what bias is introduced by using somebody else's data base, and could the impact of biased data be identified?

We often hear that there is no widely accessible base of data that is usable to estimate software costs. The literature is replete with statements lamenting the lack of ADS cost data, but most of the recently developed cost models require the input of historical cost information, or estimates of the same, in order to obtain usable results.

~~DEVELOPING A DATA BASE~~

At AFLC, we took a straight forward approach to solving this problem. We found that we really didn't have a lack of pertinent cost information. Our problem was that the information had been developed for other purposes. The data were there waiting to be used, but had never been assembled and arrayed so that they could be used to estimate ADS costs. The tools were in the tool box, but they were labeled for specific uses. So we set out to broaden the use of data that were already available in some of our ADS management systems. We used a small committee of data processing and cost analysis professionals. Their task was to identify the information needed to estimate ADS costs, analyze that data to verify that it was usable for this purpose, and to initiate the action to develop a life cycle cost data base. Figure 1 reveals the membership of our committee.

Figure 1

HQ AFLC ADS Life Cycle Cost Committee

2Lt Jill Shulas - LMP
Mr Eugene Hammond - LMD
Mr Don Slape - LME
Mr Bob Echols - LMO
Mr Frank Perry - LMT
Mr Tom McFadden - LMV
Mr Vernon Zeckser - LMZ
Mr Steve Passage - ACM
Mr Bob Golen - ACM
Mr Ralph Ta, .or - ACM
Mr Bill Reid - LMV
Mr Walt Houlette - ACM - Chairman

The Committee found data in four of our management systems that were probable indicators of ADS costs. From the Command ADS Authorization and Utilization Management System (DSD P040E) we could obtain the systems identity, program language, lines of delivered code, number of programs, number of utility programs, number of input interfaces, and number of output interfaces. From the AFLC Computer Product Management System (DSD P019) we could obtain the number of reports each system provides. The Simulation, Modeling and Computer Performance Evaluation/Analysis System (DSD K053A) contains the number of files in a data system. And the DAR (Data Automation Requirement) Resources Management System (DSD P007) contains the amount of manpower used to develop, enhance, and maintain a system. Armed with these sources of information, the committee set out to determine if the data were usable for estimating ADS costs.

THE TEST PROGRAM

Our test program began with a sample of information from P040E, P007, and P019. We gathered data concerning the development of 80 systems. The data elements analyzed were: development man-hours, lines of code, number of reports, number of programs, number of utility programs, number of input interfaces, and number of output interfaces. We used a step-wise multiple linear regression analysis to determine the fit of these data elements. Table I displays the correlation matrix for the initial analysis. We noted few acceptable correlations from this sample of data. We pondered these results and decided to perform another test. So we selected another area where we have a source of current information; a well managed conversion project. There were data available about forty-six systems that had been converted. We used the same step-wise multiple linear regression technique. Table II portrays the correlation matrix for the conversion analysis. It reveals very few acceptable relationships between the variables. So we set out to determine why there were such poor relationships among the variables we had selected for analysis.

TABLE I

CORRELATION MATRIX SYSTEMS DEVELOPMENT

n = 80

	Man-Hours	Lines of Code	Reports	Programs	Utility Programs	Input Interfaces	Output Interfaces
Man-Hours	1.000						
Lines of Code	0.269	1.000					
Reports	0.361	0.393	1.000				
Programs	0.183	0.545	0.553	1.000			
Utility Programs	0.141	0.450	0.236	0.761	1.000		
Input Interfaces	0.337	0.237	0.141	0.123	0.288	1.000	
Output Interfaces	0.536	0.243	0.220	0.125	0.234	0.649	1.000

TABLE II

CORRELATION MATRIX - CONVERSION PROJECT

n = 46

	Man-Hours	Lines of Code	Reports	Programs	Utility Programs	Input Interfaces	Output Interfaces
Man-Hours	1.000						
Lines of Code	0.695	1.000					
Reports	0.309	0.347	1.000				
Programs	0.493	0.575	0.527	1.090			
Utility Programs	0.493	0.487	0.368	0.810	1.000		
Input Interfaces	0.201	0.087	-0.077	-0.045	0.113	1.000	
Output Interfaces	0.341	0.142	0.013	0.007	0.157	0.607	1.000

We used the data from the conversion project because of its controlled and up-to-date nature. We began with a review of the number of different types of systems included in the conversion project. Ten different types of systems were observed. This caused us to look at the systems conversion procedure in the same manner a quality control engineer looks at an industrial process. We found that for a given type of system, essentially the same people worked on the project. They used the same program language, the same automated data processing equipment, general programming structure and coding procedures. Armed with this insight, we reviewed the data in the Systems Development sample. We found essentially the same information in the development sample that existed in the data about systems conversion. This indicated that a situation existed that could be considered to be a constant chance cause system. To verify this, we selected the data about maintenance systems from both files of data. We performed the same type of step-wise multiple linear regression analysis of these data sets. Tables IA and IIA portray the correlation matrices for the these analyses.

TABLE IA

CORRELATION MATRIX - SYSTEMS DEVELOPMENT

MAINTENANCE SYSTEMS

n = 10

	Man-Hours	Lines of Code	Reports	Programs	Utility Programs	Input Interfaces	Output Interfaces
Man-Hours	1.000						
Lines of Code	0.741	1.000					
Reports	0.934	0.668	1.000				
Programs	0.935	0.891	0.874	1.000			
Utility Programs	0.879	0.736	0.961	0.865	1.000		
Input Interfaces	0.914	0.701	0.988	0.877	0.981	1.000	
Output Interfaces	0.901	0.844	0.895	0.919	0.960	0.928	1.000

TABLE IIA

CORRELATION MATRIX SYSTEMS CONVERSION

MAINTENANCE SYSTEMS

n = 8

	Man-Hours	Lines of Code	Reports	Programs	Utility Programs	Input Interfaces	Output Interfaces
Man-Hours	1.000						
Lines of Code	0.955	1.000					
Reports	0.909	0.899	1.000				
Programs	0.930	0.985	0.913	1.000			
Utility Programs	0.807	0.912	0.700	0.909	1.000		
Input Interfaces	0.505	0.411	0.221	0.384	0.413	1.000	
Output Interfaces	0.791	0.802	0.549	0.745	0.808	0.698	1.000

The results of these analyses were encouraging. The high coefficient of correlation displayed in Tables IA and IIA indicated that five of the six variables should be good indicators of the man-hours required to develop or convert automated data systems. Analysis of the residuals indicated that, in each case, they were normally distributed with a mean of zero and acceptable standard deviations. Also, there were no adverse patterns in the distribution of the residuals. Both analyses indicated that the regression equations were usable for predicting the cost of developing and converting Maintenance Systems.

We expanded our analysis program to include other variables that might be usable. In one analysis, we considered a total of fifteen variables. But, the tests indicated that the one additional variable needed was the number of files in a system. As a result of these tests, we became confident that we could assemble systems development and systems conversion data that would permit the projection of the pertinent manpower requirements and costs. The committee recognized that continuous analyses of this data would be required because there are new data systems being developed and others being dropped from use. So we are tracking the progress of systems now being developed and converted to assure that predicted costs and actual costs are comparable within acceptable limits.

EXPANDING THE DATA BASE

Another cost estimating problem faces the analyst: the cost of preparing requirements documents for data automation projects. These are primarily functional area costs, but they do include some ADP personnel costs. These costs accrue during the early phases of ADS development.

Our DAR Resources Management System (P007) accumulates man-hour expenditures for each step of systems development. The ADP steps are entitled Evaluation, Design, Coding and Testing, Documentation, and Implementation. We have analyzed these data and found that there is good correlation between the amount of resources used in each phase, and the total development resources.

Our next step was to determine if the functional area manpower requirements were related to the ADP manpower requirements. We didn't have a file of functional area data pertaining to systems development, but we did have some project officers who were willing to provide useful information. Our more recent economic analyses have included the cost of doing the work required before ADS design begins. Our project officers aided by providing the records of time spent in each phase of requirements document preparation. We were able to acquire this data for twenty-seven ADP projects. Table III displays the correlation matrix for these data.

TABLE III

SYSTEMS REQUIREMENTS RESOURCES

CORRELATION MATRIX

n=12

	Feasibility Study	Economic Analysis	DAR	Functional Description	Data Project Directive	Data Project Plan	Total Man-Hours
Feasibility Study	1.000						
Economic Analysis	0.862	1.000					
DAR	0.984	0.922	1.000				
Functional Description	0.752	0.515	0.639	1.000			
Data Project Directive	0.271	0.707	0.423	-0.165	1.000		
Data Project Plan	0.382	0.186	0.418	-0.110	0.036	1.000	
Total Man-hours	0.956	0.785	0.901	0.907	0.136	0.154	1.000

We combined the functional area information with that from P007 to get a span of data from conception to the completion of systems development. The data elements used in this analysis are expressed as man-hours to accomplish these functions: Feasibility Study, Economic Analysis, Data Automation Requirement, Functional Description, Data Project Directives, Data Project Plan, Subtotals for Requirements Development, Design, Coding and Testing, Documentation, Implementation, Systems Development Subtotal, and a Grand Total.

We used the step-wise multiple linear regression analysis technique to analyze the functional area and development file. We found good correlation values as are shown in Table IV. The analysis reveals that knowledge of the hours used to develop the feasibility study can be used to predict the resources required to develop the economic analysis, and functional description of the system. The test also indicated that feasibility study hours can be used to predict the time required to develop the ADS, as well as the total time required to prepare requirements documents and develop the system. This test opened another door: it indicates that there is a relationship between all the steps required to conceive and develop an automated data system. Therefore, with a larger base of data, we should be able to project the amount of manpower required to perform each major task required to develop an ADS. This will permit us to estimate the total resource requirements and apportion the resources to the major tasks.

TABLE IV

SYSTEMS REQUIREMENTS AND DEVELOPMENT

CORRELATION MATRIX

n=27

	FS	EA	DAR	FD	DPD	DPP	SUB TOT	DSN	C&T	DOC	IPL	DEV	GTOT
FS	1.000												
EA	0.894	1.000											
DAR	0.980	0.930	1.000										
FD	0.660	0.387	0.545	1.000									
DPD	0.543	0.848	0.648	-0.111	1.000								
DPP	0.534	0.475	0.570	-0.093	0.435	1.000							
SUBTOT	0.958	0.799	0.906	0.840	0.378	0.346	1.000						
DSN	0.711	0.461	0.647	0.861	0.008	0.201	0.841	1.000					
C&T	0.815	0.507	0.730	0.743	0.041	0.491	0.838	0.761	1.000				
DOC	0.859	0.621	0.785	0.688	0.224	0.536	0.857	0.855	0.880	1.000			
IPL	0.594	0.418	0.603	0.126	0.216	0.835	0.449	0.332	0.728	0.627	1.000		
DEV	0.847	0.544	0.767	0.479	0.081	0.519	0.865	0.849	0.981	0.935	0.726	1.000	
GTOT	0.877	0.589	0.800	0.770	0.127	0.509	0.897	0.860	0.976	0.940	0.702	0.997	1.000

NOTE: A Table of Abbreviations is shown at Attachment 1.

Another portion of the life cycle cost of an ADS involves the maintenance of a system. The DAR Resources Management System (P007) records the amount of resources used to maintain each ADS. The data recorded are man-hours used to evaluate the problem, and to design, code and test, document, and implement the correction.

TABLE V

MAINTENANCE OF OPERATING SYSTEMS

CORRELATION MATRIX

n=36

	DSN	C&T	IPL	DOC	TOT
DSN	1.000				
C&T	0.348	1.000			
IPL	0.797	0.439	1.000		
DOC	0.543	0.631	0.697	1.000	
TOT	0.835	0.715	0.882	0.866	1.000

TABLE VI

MAINTENANCE OF OPERATING SYSTEMS

CORRELATION MATRIX

LOGISTICS SYSTEMS

n=8

	DSN	C&T	IPL	DOC	TOT
DSN	1.000				
C&T	0.470	1.000			
IPL	0.861	0.399	1.000		
DOC	0.858	0.823	0.694	1.000	
TOT	0.836	0.732	0.863	0.955	1.000

When the maintenance data were analyzed by type and system, we found that there is a good basis for estimating these resource requirements. The initial tests indicate that we can predict how much manpower will be required to maintain a system. Tables V and VI display the correlation matrices of an overall sample of 36 systems and the 8 Logistics Systems which were a part of that sample. These small samples indicate that we should get good results when we compare the data from a specific set of like type systems through each major phase: Requirements, Development, and Maintenance.

USING THE DATA BASE

When a new system must be developed the project manager must analyze the requirement. His initial task is to determine the need for a system and justify it to his superiors.

At this point in time, very little is known about the system. But there are a few facts available that will help the project manager. The most important fact is that the manager needs reports. He knows what reports are provided by his present system and what additional information is required for his management process. Contact with his data automation manager would identify the system development team, the program language, and computer to be used. The type of system is usually identified to the organization that has the requirement (i.e., Maintenance, Engineering, Logistics).

So we find that our project manager really has a lot of information at hand, and it can provide him with much more. Let us assume that our project manager works in the Maintenance Management function and that the new system he desires must provide 50 reports. He can now call a resources analyst and obtain the following information from the ADS Life Cycle Cost Data Base.

The Correlation Matrix for the Development of Maintenance Systems is displayed below.

TABLE VII

CORRELATION MATRIX

DEVELOPMENT OF MAINTENANCE SYSTEMS

n=18

	Man-Hours	Code	Programs	Utilities	Reports	Files	Output Interfaces	Input Interfaces
Man-hours	1.000							
Code	0.848	1.000						
Programs	0.851	0.830	1.000					
Utilities	0.235	0.512	0.447	1.000				
Reports	0.930	0.856	0.[illegible]3	0.355	1.000			
Files	0.915	0.800	0.849	0.445	0.936	1.000		
Output Interfaces	0.431	0.386	0.550	0.374	0.464	0.614	1.000	
Input Interfaces	0.692	0.508	0.618	0.307	0.661	0.824	0.813	1.000

Review of the correlation matrix indicates that the number of reports a system will have is an acceptable predictor of the ADP man-hours required to develop a system, the number of lines of code that will be delivered, the number of programs that will be used, and the number of files the system will have. The number of reports will also provide a marginal estimate of output interfaces and input interfaces, but a poor estimate of the number of utilities that will be used. So let's take a look at the regression equations and determine how large the system is going to be.

Man-hours = 426.71 + (171.04 * 50) = 8979

Lines of Code = 519.94 + (811.64 * 50) = 41,102

Programs = 5.7 + (0.72 * 50) = 42

Utilities = 7.05 + (0.22 * 50) = 18

Files = 8 + (2.4 * 50) = 128

Output Interfaces = 0.98 + (0.03 * 50) = 2

Input Interfaces = 0.81 + (0.05 * 50) = 3

The data base has permitted us to estimate the parameters of the new system when we merely knew that it must output 50 reports. This information permits the project manager to know something about the work load that will be assigned to the data system design team.

There is another set of information that we can provide to the project manager: the amount of work required to prepare requirements documents. In Table IV we noted that total data system development hours are a good predictor of man-hours required to develop the requirements documents. Using the equations from the analysis of the data file that output Table IV, the following manpower requirements are identified:

Task	Man-hours
Feasibility Study	342
Economic Analysis	1,056
Data Automation Requirement	714
Functional Description	781
Data Project Directive	282
Data Project Plan	252
Total	3,427

At this point, the project manager has an estimate of 3,427 man-hours to develop requirements documents and 8,979 man-hours to develop the data system. And the data base analysis was available for two reasons: the project manager knew he needed 50 reports and there was an accumulation of information about Maintenance Systems in our data base.

As work on the system progresses, the project manager and the development team will learn more and more about the system. After the Feasibility Study is completed, more finite knowledge about the numbers of programs, files, interfaces, and the use of utilities will become available. These improved values can be used to improve the estimates of manpower required to perform each subsequent task. By the time the Functional Description is completed, a very firm estimate of the costs of the project should be available. Also, a very good method of tracking actual costs and comparing them to projected costs will be available. This comes about because each project record in the data base will be updated every two weeks.

So the project manager will require only a few data elements to receive the benefits of the data base. As his project progresses, he will have the ability to obtain updated estimates and evaluations of actual performance compared to that projected. The ability to prepare milestone reports about the project will be enhanced. And a life cycle management capability will be available for all major data automation projects, as required by DODI 7920.1

THE DATA BASE

Our data base will provide the information needed to estimate the resources required to conceive, develop and maintain an automated data system. The data base includes the information about resources used in the conceptual phase as well as the systems development and maintenance phases. Systems operating costs are usually obtained from computer operations reports. So we will, with implementation of the data base, be in a posture to provide an estimate of the resources required to analyze the need for a system, document the requirements, estimate the cost of resources, and allocate the resources over the life cycle of the system.

The data base will provide the inputs to most types of cost models, depending on the needs of the estimater and the user. It will provide the most up-to-date information that can be documented for developing cost estimating relationships. The data base will contain files of systems development information, systems conversion information, and systems maintenance information. It will also contain a file of systems enhancement information. The Committee has found that the information about DARs used to enhance a system or to make small changes are also usable to project the cost of these lesser tasks.

Our data base project revealed that the information needed to project ADS costs was available in our management systems. Our current effort involves bringing this information together into a set of files. The small DARs have been initiated to provide the data. A DAR has been submitted to design the arrays of the data base and provide the means to sort the data and analyze it. The outputs of these analyses will be sets of resource estimating equations. Although only minimal information about a new system may be available, an estimate of the resources needed to develop and maintain it can be made. As more knowledge of the system becomes available, additional estimates of resource requirements can be provided. By the time data system design is completed, a very good estimate of resources requirements and costs can be provided. The burden of developing feasibility study cost estimates and economic analyses should be lightened considerably by having a sound base of resources information. The data base should provide the ability to provide resources projections in minimum time, since the resource estimates will be automated. The result will be the ability to test and validate resource estimates at any point in the system's life cycle.

With the implementation of the use of this data base, resource projections will be based upon the history of similar systems developed by AFLC. The information used as the basis of our projections will be pertinent to our new efforts, and not be biased by the use of "rules of thumb" or data from an unknown source. If there is some sort of bias in the information in our data base, it will be found during data base analysis and the cause will be identified.

The task of estimating resources requirements and costs is quite involved. The involvement can be reduced in magnitude by the identification of sources of information. Many data systems contain information that may be used for purposes other than that intended for the original system. Our Committee did not have to develop a single new data element to provide the ADS life cycle cost data base. Every data element existed in another data system. It is probable that most organizations could do what we have done at AFLC. The relationships that we have identified will probably hold true for any software development organization. We believe the key factor is to use your own data to project your resource requirements for ADS projects. Avoid the all inclusive bases of data, axioms, and rules of thumb because they probably don't fit your specific situation. Information about what occurred in your software development organization when you built systems previously provides the best basis for estimating resource requirements for your new systems. In short, build your own data base, keep it current, and use it.

TABLE OF ABBREVIATIONS

ADS	Automated Data Systems
C&T	Code & Test
DAR	Data Automation Requirement
DEV	Developmentation
DOC	Documentation
DPD	Data Project Directive
DPP	Data Project Plan
DSN	Design
EA	Economic Analysis
FD	Functional Description
FS	Feasibility Study
GTOT	Grand Total
IPL	Implementation
SUBTOT	Subtotal
TOT	Total

Atch 1