Technical Description Document

Cost Estimation Toolkit (CET)

Version 2.4

September, 2008

Greg Hunolt, Bud Booth, Mel Banks, SGT

Funded by NASA Contract NNG05CA99C

Table of Contents

1.0 Introduction

This document presents a technical description of the Cost Estimation Toolkit (CET) developed by SGT, Inc. This document is a companion piece to the Users' Guide for the CET. The purpose of this document is to provide the CET user with an understanding of how the CET performs its cost estimation, while the Users' Guide describes, step-by-step, how to use the CET to produce a life-cycle cost estimate for a new data activity.

This version of the Technical Description Document describes Version 2.4, the September, 2008 version of the CET. The technical description focuses on the CET Estimator (references to the "CET" below can be taken as references to the Estimator tool unless the context is the CET toolkit as a whole. Version 2.4 of the CET includes improved approaches to outlier identification.

The CET is an Excel Visual Basic for Applications (VBA) application. As such the software is contained in an Excel workbook and includes VBA modules and user forms. Each VBA module includes a number of procedures (equivalent to subroutines) and the user forms are Excel programmable user interface features that include VBA procedures.

The CET runs on both PC and Macintosh platforms; it detects which type of platform it is running on and uses user forms and worksheets formatted appropriately for that platform.

The CET also includes worksheets used for a variety of purposes. These include the Activity Datasets entered and maintained by the user. Three worksheets contain the output of the CET; two contain the life cycle cost estimate for a user-described data activity and a third contains a 'quality report' to help the user decide how much confidence to have in he/she should place in the estimate. Worksheets are included to hold the output of the Reviewer tool. Also included are a set of worksheets that are internal to the CET, e.g., used for display backgrounds and to hold intermediate results.

The CET contains two tools, the Estimator which produces the life cycle cost estimates, and the Reviewer which allows users review and fine tune the CET's estimate. The CET Estimator employs the cost estimation by analogy approach, and builds on the SGT general data services provider reference model. The cost estimation by analogy approach requires information about existing data activities that serve as the analogies for a new data activity; i.e. data activities that are sufficiently comparable to the new data activity to permit them to be used as the basis for estimating the life cycle costs of the new activity. The CET contains information for a set of existing comparable data activities (referred to as 'comparables information'). These data activities are not identified, and their information is mapped to the general reference model.

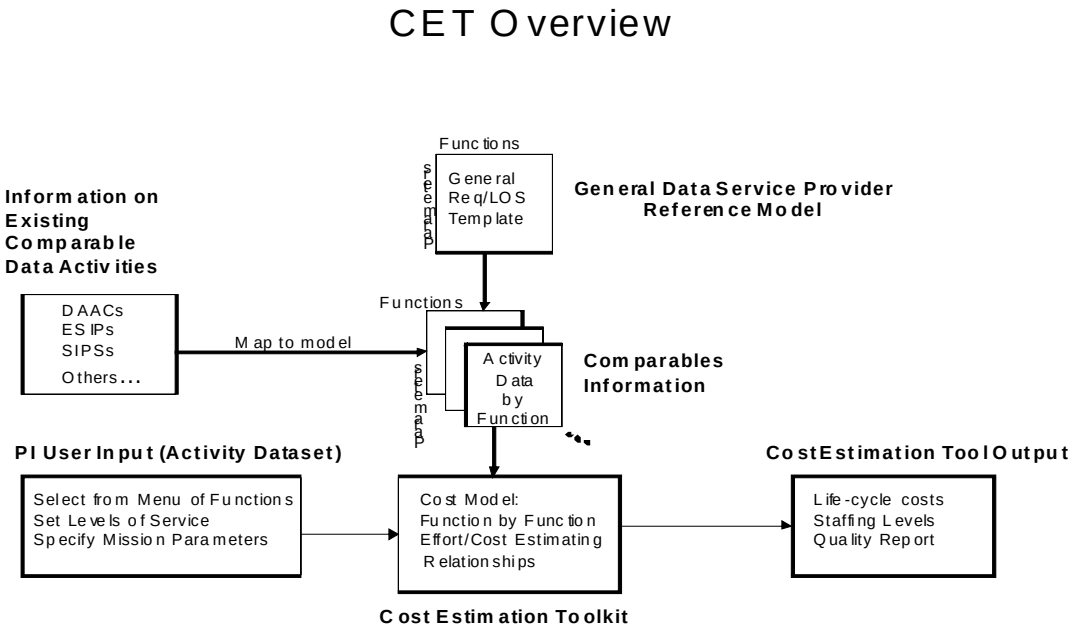
Section 2 provides an overview of the CET. Section 3 discusses the approaches to effort (i.e. staff FTE level) estimation used by the CET.

Background information for the CET was originally provided by a series of working papers developed by SGT as part of the LOS/CE study. These working papers are gradually being replaced as the information in them is updated and added to the Users' Guide and the Technical Description Document. The SGT general data service provider reference model is described in the CET Users' Guide and earlier in LOS/CE Working Papers 3. Working paper 6 is still the current description of logical Data Service Provider types that will be used in a future version of the CET.

2.0 Overview: Cost Estimation Toolkit (CET)

The overview of the CET is illustrated by Figure 1, below. As shown in the figure, the general Data Service Provider Reference Model serves as the underpinning for the CET. The model describes a general data service provider as being comprised of a set of functional areas, each described by a set of requirements and parameters (see the Users' Guide for a description of the functional Areas). Note that the terms 'data service provider' and 'data activity' are used synonymously. The figure shows that information describing existing NASA Earth Science and other data activities (including DAACs, ESIPs, SIPSs, etc.) is assembled, mapped to the function / parameter structure of the reference model, and added to the comparables information that is used by the cost estimation by analogy approach. The result is an internally consistent set of descriptions of a number of different data activities. In each case, the description of a particular data activity will only include those functional areas that are applicable to it (e.g. few sites will perform instrument / mission operations) and for which good information is available (e.g. a minimum set of workload and corresponding effort data).

Figure 1 - Cost Estimation Toolkit Overview



The bottom tier of three boxes illustrates the use of the CET by a PI (Principal Investigator) or other user planning a new data activity. The lower left box illustrates the PI or other user entering information describing a new data activity into the CET. The user enters, for selected functions, information describing levels of service and particular mission requirements (e.g. such as data to be ingested; products to be generated, archived, and distributed; user support, etc.). This set of information that describes the new data activity is called an Activity Dataset.

The lower center box illustrates the CET itself, which accesses the comparables information and proceeds function by function to build a life cycle effort (i.e. staff FTE) and cost estimate for the new data activity. The lower right box illustrates the output from the CET, estimated year-by-

year costs over the new data activity's life cycle, with supporting parameters such as estimated effort (FTE) levels.

3.0 Effort and Cost Estimation Approaches

This section describes the approaches used in the CET Estimator for estimating effort (i.e. FTE levels) and cost for a new data activity. The estimates are based on information about existing activities considered to be comparable with, or analogous to, new data activities. Effort estimation is the heart of the CET, covering all categories of labor over the life cycle of the data activity - conversion of the effort estimates to staff cost estimates is simply a matter of applying labor rates and an inflation rate to the effort estimates. Total costs then are obtained by adding the cost of non-staff items such as computers, media, etc., to the staff costs.

Section 3.1 discusses the general approach to effort estimation. Section 3.2 discusses the approach to cost estimation. Sections 3.3 through 3.15 address each of the functional areas individually.

3.1 Effort Estimation

The objective of the effort estimation process is to produce, for a new data activity, year-by-year estimates of operational, technical, and management staff effort for each functional area applicable to the new Activity Dataset. The effort is expressed in Full Time Equivalents (FTEs), the effort equivalent to that of a single person working full time for a year. This is not literal staffing, since an FTE level of 4 could be performed by four people each working full time, eight people each working half time, or any other combination adding up to the equivalent of four full time people.

The effort estimate for each functional area for a new data activity is based on its planned workload in that functional area, and the relationship between workload and staff effort observed in existing data activities in that same functional area. The workload and staffing information for each functional area for existing data activities is contained in the CET's 'comparables information.'

"Workload" in the context of effort estimation includes not only measures of volumes of data ingested, produced or distributed, or number of products ingested, produced, or distributed, but also levels of service that can drive effort and complexity of operations (e.g. numbers of different interfaces involved,, etc.).

Operational effort covers computer or equipment operators, and other effort that is directly involved with the conduct of the ongoing operation, e.g. production monitoring, quality assurance monitoring, packing and shipping distribution media, archive media handling and screening, etc.

Technical effort associated with a functional area includes engineering or science effort exclusive of direct operations, e.g. science software integration and test, cross-calibration specialists, interface engineering and management. To the degree that a data activity's operation is highly automated, or in some aspects small in scale, operations effort may be performed as a part time activity of technical personnel, while larger activities may include dedicated operational staff.

Section 3.1.1 describes the two step approach taken for effort estimation in the CET. Section 3.1.2 describes the development of EER's (effort estimating relationships) and Section 3.1.3 walks through the estimation process as it is implemented in the CET software.

3.1.1 Overview of CET Effort Estimation Approach

This section describes the "Curve Fitting" approach taken to effort estimation in the CET.

For each functional area, the primary effort categories, technical and operational, are each assumed to be a function of a particular combination of workload parameters (e.g., data volume, number of products). The relationships that are used to estimate effort as a function of workload are called "Effort Estimating Relationships" (EER's), analogous to "Cost Estimating Relationships" (CER's).

The approach includes two distinct steps. In the first step, the CET develops a set of regression equations based its information on existing comparable data activities, and in the second step, it uses those equations in the course of a process for producing a set of effort estimates for a new data activity. The next section, 3.1.2, discusses the first step, development of the equations, and the following section, 3.1.3, describes the estimation process.

3.1.2 Development of Curve Fit Relationship

The CET selects, for each functional area, one or more workload parameters to be used as independent variables in regression of effort on workload. The CET then develops a table of the values of the workload parameter versus the corresponding (FTE level) for each existing activity for which these are available and usable.

The CET uses regression techniques (following "Statistical Methods" by Snedecor and Cochran) to develop the coefficients for a set of seven trial relationships of FTE to workload parameter for each of the selected workload parameters. These relationships are shown in table 3-1 below. In the equations, Y is the dependent variable, FTE; X is the independent workload variable; and a, b, and c are the coefficients computed by regression. The symbol "*" indicates multiplication, and "^" indicated exponentiation (i.e. X^2 is the same as X squared or X*X).

Table 3-1 - CET Effort as f(Workload) Relationships

Linear	Y = a + b*X	
Logarithmic	Y = a + b*lnX	(ln is natural logarithm)
Exponential	Y = a*e^(b*X)	(e is the base of the natural logarithms)
Quadratic	Y = a + b*X + c*X^2	
Square Root	Y = a + b*X + c*sqrt(x)	(sqrt - square root)
Linear-Logarithmic	Y = a + b*X + ln(X)	(ln is natural logarithm)
Linear-Exponential	Y = a + b*X + c*e^X	(e is the base of the natural logarithms)

For the first three relationships, the CET uses single parameter regression of Y's on X's (effort on workload), and for the last four, two-parameter multiple regression; e.g., for the quadratic case the two parameters are X and X^2.

The CET performs the process outlined below for each set of workload parameter – FTE parameter pairs within a given functional area to develop a final set of regression equations to be used in the estimation process (to be described in section 3.1.3).

3.1.2.1 Development of Regression Equations

A. Outlier Pre-Processing:

The CET may perform outlier pre-processing (see notes for each functional area below) prior to the regression computation. The outlier pre-processing technique is called "Cluster Outlier Removal". The CET treats a set of workload parameter, FTE parameter pairs as members of a "cluster" of points mapped to a two dimensional space with FTE and workload value axes. The intent of the cluster outlier process is to identify and remove those members of the cluster that fall furthest from the center of the cluster, up to a given limit (which can vary from functional area to functional area).

The CET computes the average values of the workload parameters and FTE parameters for a given set of a workload parameter – FTE parameter pairs, or "points". The CET then establishes an "average" point for a given set of points as the point with the average value of the workload parameter and the average value of the FTE parameter. The CET computes the normalized vector distance of each point from the average point. This is the square root of the sum of the squares of the difference of the normalized workload value from the average value and the difference of the normalized FTE value from the average value. The normalized workload values are the actual workload values divided by the maximum workload value, and the normalized FTE values are the actual FTE values divided by the maximum FTE value.

The CET then removes from the set of workload – FTE parameter pairs those pairs representing points with the greatest distance from the cluster's computed average point. The number of such outliers removed is restricted by a specified limit.

B. Regression Computation:

The CET computes a set of coefficients for each equation. After computing the coefficients for each equation, the CET uses the new equation and its coefficients to obtain a Y (FTE) value for each X (workload) value, and computes the square of the Pearson Product Moment Correlation Coefficient (a.k.a. "R-Squared"), average absolute error, and standard deviation of absolute error.

The CET then examines the set of relationships it has developed. It first discards relationships that either would yield negative values of workload or that would produce a double value; i.e., two values of Y for a single value of X for X's within the range of the comparables information (which, for example, a quadratic relation might do if the inflection point of its curve falls within the comparables information range of X's.)

C. Outlier Removal:

The CET then performs a second outlier removal process by examining the error associated with each of the original workload, FTE pairs compared to the estimating relationship defined by the regression curve. The concept is that an extreme point might perturb the estimating relationship, causing it to produce poorer results than it would if the outlier were to be excluded. For the CET, an "outlier" is a point (workload, FTE pair) that has an absolute error value that is greater than three times the standard deviation of absolute error for all of the data points. If an "outlier" is found, the CET tests the effect of eliminating it - it temporarily deletes the outlier and re-computes the relationship (using the regression method again). The CET repeats the tests described above for negative or multiple Y values (i.e., more than one FTE value for any given workload parameter value) and if these tests are passed the CET checks to see if the R-Squared

value is improved. If the tests are passed and the R-Squared value is improved, then the outlier is deleted permanently and the new relationship is used. If the negative or multiple Y test is not passed or if the R-Squared value is not improved, the CET reverts to the original relationship.

After the outlier process is completed, if the number of outliers allowed to be removed is not reached, and if the number of comparable activity data pairs is greater than the minimum needed, the outlier test and removal process described above is repeated until either no outlier is found or one of the two test conditions fails (i.e. either the allowed number of outliers have been removed or the number of data pairs is at the minimum needed number). Typically, one or two outliers are removed.

D. Final Curve Selection:

After the sequence above is completed for each of the possible types of curves, the CET selects the relationship with the best R-Squared value.

The final test that is made by the CET is to see if the workload values for the new data activity fall outside the range of the corresponding workload values for the comparable data activities on which the EER was based. If it is, then the CET would be extrapolating beyond the range of the EER for which it was computed. In such a case, the CET, to be conservative, adopts the logarithmic relationship regardless of its R-Squared value. Other relationships can go quickly awry when it is pushed beyond the range of values for which it is computed, while a logarithmic curve tends to flatten. In addition, a safeguard is being tested that would constrain the magnitude of values produced by the forced linear relationship to strengthen the protection against an overly enthusiastic extrapolation when the CET is attempting to estimate a value outside of the ranges of the corresponding comparable information.

For a given functional area, the entire process described above is repeated for each selected workload parameter to yield the final FTE as f(workload parameter) EER for the particular effort parameter (i.e. either operational or technical FTE) and the workload parameter (one of the selected workload parameters for the given functional area).

3.1.2.2 Horizontal Outlier or “Nearness” Test

A function-by-function horizontal outlier or “nearness” test is included in the CET. This test (which may take one of two forms described below) may be performed for each functional area depending on the final “tuning” of CET control parameters. If used for a given functional area, the test is performed prior to the curve fitting process described above.

The “nearness test” is used to select those comparable activities that are “nearest” in workload to the new data activity, so that the life cycle estimate for the new data activity is based on those comparable activities that are most similar to it - i.e., that are the best analogies. Because this is done on a function by function basis, different combinations of comparable activities may make up the set of best analogies for different functions. This allows for the fact that a new data activity might more closely resemble one comparable activity for ingest, another for processing, a third for distribution, etc.

The current form of the “nearness test” is also referred to as the “horizontal outlier” test. It has the same objective as the original “nearness” test, the exclusion of those comparable activities that are most different from the new data activity. For CET Version 2.4, for a given function the CET computes for each activity a set of “nearness” values, one for each of the relevant workload

parameters. Each "nearness" value is the ratio of the comparable activity's value of the workload parameter with the new activity's value. The comparable activities are then ranked in order of 'nearness', using each activity's least "near" value, and the least "near" of the comparable activities are identified as outliers up to the specified horizontal outlier limit. Prior to Version 2.4, only a single workload parameter was used for each function (volume). The current method checks all of the relevant workload parameters, using the most extreme for each function in the comparison across comparable data activities.

The original form of the "nearness test" was as follows: The "nearness" was determined for each function by computing a workload index for the new data activity and the comparable activities, computing a "nearness" parameter for each comparable activity, and accepting those comparable activities for which the difference parameter was less than a threshold value. The workload index value used by the Working Prototype was the natural logarithm of the "work" parameter originally developed for the ESDIS Data Center Best Practices and Benchmark study. The "work" parameter is the sum of the volume in GB handled by the function (e.g. ingested, produced, distributed) and the number of product instances handled divided by 1000. Because of the wide disparity in workload across the comparable activities, the natural logarithm of "work" is used. The "nearness parameter" was the magnitude of the difference between the natural logarithms of "work" for the comparable activity and the new data activity. The CET marked those comparable activities whose "nearness" exceeds the given threshold value.

3.1.3 Effort Estimation Process using Curve Fit Relationship

This section describes the process by which the CET uses the effort estimating relationships it develops between the comparable activities' effort and workload parameters to produce effort estimates for a new data activity.

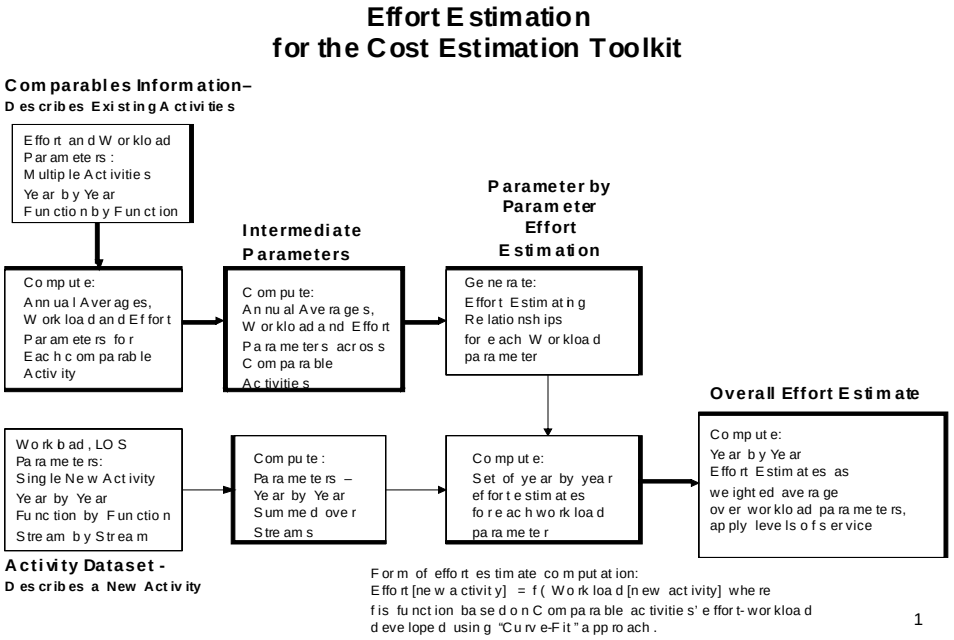
The process operates functional area-by-functional area, producing separate year-by-year effort estimates for each functional area, which are then summed to obtain an overall effort estimate for the new data activity.

This section will describe the process generically, and the functional area sections below (sections 3.3 - 3.15) will describe any variations that exist for individual functional areas.

For each functional area, the process is performed by the CET in the steps described below, as illustrated by Figure 2 below. The process is performed separately for operations and / or technical effort, depending on the functional area.

At the outset of the process, for each functional area, one or two sets of comparable activity workload parameters have been selected to be used in the effort estimation process. One set would be workload parameters to be related to operations effort, and the second set would be workload parameters to be related to technical effort. Note that not all functional areas have both operations and technical effort. (For example, ingest has only operations effort, while processing has both operational and technical effort, and sustaining engineering has only technical effort - see the functional area notes in sections 3.3 - 3.15 below). Relative weights are assigned to each parameter within the set, those that influence operations effort and those that influence technical effort. The weights are based on a 'tuning' of the CET to produce the best possible estimates for a set of independent cases.

Figure 2 - CET Effort Estimation Approach



1. The CET computes intermediate parameters, functional area by functional area, for each comparable data activity provided that usable information is available for the data activity in the functional area. The intermediate parameters are the averages over the data activity's life cycle of each of the workload and staff effort parameters in each of the functional areas included for that comparable data activity. Note that not all activities will perform work in all functional areas, and that even when an activity performs work in a functional area, usable information (effort and some workload parameters) is not always available. The CET uses the information available to it on a functional area by functional area basis.

2. The CET computes intermediate parameters for the new data activity. The intermediate parameters, computed functional area by functional area, are year-by-year sums of workload parameters over the individual streams defined by the user when he/she created the activity data set (ADS) for the new data activity. For example, in the processing functional area the user may have defined four different processing streams which may have different start and stop dates. The workload parameters such as volume per day per stream are annualized and summed to a total volume generated per year for each year the new data activity is in operation.

3. The comparable activities to be used in producing the life cycle cost estimate for the new data activity are selected using horizontal outlier screening.

4. For each comparable data activity selected to be used in computing the effort estimate, the CET computes averages of the parameters selected to be used for effort estimation, using only valid (non-zero) values (an activity might have usable information for the functional area, but might be missing one or more of the workload parameters; it is not required to have all of them present to be 'usable').

5. The CET next computes a set of best curve-fit effort estimation relationships for each of the selected parameters, as described in section 3.1.2 above. Each of these is an equation that can be used to compute an FTE value from a workload parameter value.

6. The CET next computes a set of year-by-year effort estimates, one for each of the selected parameters, by using the year-by-year workload values for the new data activity (i.e. the ADS workload parameter values) as input into the effort estimating relationships. These individual estimates, each based on a single workload parameter, are separate estimates of the full operations or technical effort for the functional area.

7. Adjustments are made to the individual estimates based on a comparison of the data activity value of the parameter and the average value (for the comparable activities which pass the screening) of the parameter; the estimate is increased if the new activity will have a greater workload than the comparable activity average, decreased if it will have less. The magnitude of the increase or decrease is determined by the CET tuning process.

8. A weighted average effort estimate of the operations or technical effort for the new data activity is then computed, using the R-Squared value for each component of the estimate as the weighting factor, thus giving the greatest weight to the relationships that correlate best.

9. The CET then applies the level of service parameters for the functional area. The estimated effort value is increased if the new data activity will perform at a higher level of service than the average for the comparable activities used to make the estimate, or decreased if it will perform at a lower level of service than the comparable activity average. One or more level of service parameters may be used, depending on the functional area. A final overall adjustment is made using a parameter whose value was determined during the CET tuning process.

The projected workload parameters for the new data activity used in the computation described above are provided by the user when the Activity Dataset is assembled. In some cases, averages are computed year-by-year, for example across the ingest streams, product streams, or operational distribution streams planned for the new data activity. The changes in workload expected for the new data activity in the course of its life cycle are preserved.

3.2 Cost Estimation

The development of cost estimates is discussed in two parts, staff costs (Section 3.2.1), and non-staff costs (Section 3.2.2)

3.2.1 Staff Cost Estimation

Staff costs are obtained from the effort estimates by applying labor rates provided by the user. These labor rates are fully loaded (i.e. include all overheads that apply in the user's context), are particular to the user's location, and are assumed to be those effective as of the first year of the life cycle of the new data activity. An inflation rate is applied to the labor rates for the remaining years of the new data activity's life cycle (or the period of years for which a life cycle cost estimate is to be produced).

The CET asks the user to provide five labor rates:

1. Management Staff Labor Rate. This is an approximate overall rate for management and overall coordination of a data activity's work.

2. Administrative Support Staff Labor Rate. This is an approximate overall rate for administrative support for the data activity.

3. Development / Engineering Staff Labor Rate. This is an approximate overall rate for software developers/maintainers, engineering support including system engineering, system administration, database administration, resource planning, network management, facility support, etc.

4. Technical Staff Labor Rate. This is an approximate overall rate for technical and science staff, involved with ESE coordination, activity level coordination, and associated with operational functional areas.

5. Operations Staff Labor Rate. This is an approximate overall rate for operations staff in operational functional areas.

The CET asks the user to provide an inflation rate to be applied to all labor costs, using the mission / project start year as the base year.

3.2.2 Non-Staff Cost Estimation

Individual non-staff items are discussed in the functional area sections below (3.3 - 3.15) where they apply.

3.3 Notes on Ingest Effort Estimation

The ingest functional area includes only operations effort. Any technical-level analysis related to input data is accounted for under the Processing functional area.

Ingest consists only of handling one or more ingest streams. The estimation of operations effort follows the basic "curve -fit" approach outlined in Section 3.1 above, with "tuning" options for "nearness" testing and cluster outlier removal.

Ingest operations effort is estimated as a function of four weighted workload parameters: the number of external interfaces, ingest product types, product counts, and volume. Individual effort components (i.e. the estimates based on individual parameters) are adjusted by comparison of new data activity values to corresponding comparable activity average values.

Level of service adjustments are made to the final ingest operations effort estimate based on the Ingest LOS, Ingest Automation LOS, and Ingest Mode. (See the CET Users' Guide for definition of these.) For Ingest LOS, if the new data activity's Ingest LOS is higher than the comparable activity average Ingest LOS, the effort estimate is increased, and if it is lower, decreased. If the Ingest Automation LOS of the new data activity is higher than the comparable activity average, then the ingest function of the new data activity is more highly automated than the comparable activity average, and the effort estimate is reduced, and if the Automation LOS is lower, the effort estimate is increased. If the ingest function for the new data activity's Ingest Mode indicates that the ingest workload includes any ingest of data received on physical media, the effort estimate is increased, and a minimum level of 1.0 FTE is ensured.

Other workload parameters used in prototype versions of the CET (e.g. ingest formats, counts of formats, format conversions) have been dropped due to a lack of sufficient information. There

was also a fundamental uncertainty in the definition of a distinct 'format'. Even when a formatting system such as HDF is used, the detailed data or product formats are unique for each type. Since a count of product types is already used, a product format count parameter that becomes another count of product types is not needed. Finally, the "work" parameter, an amalgam of products and volume, was dropped as an estimating parameter.

3.4 Notes on Processing Effort Estimation

The processing functional area includes both operations and technical effort, based on comparable activity information.

Processing includes generation of operational product stream(s), non-operational product generation, and reprocessing. A data activity may have operational processing, non-operational processing, or both. If a data activity has operational processing, it may have reprocessing (there is no allowance for reprocessing of non-operational processing).

For the new data activity, reprocessing product counts and volumes are computed for product types for which planned reprocessing was indicated by the reprocessing LOS and reprocessing plan parameters provided for the processing streams. These are added to the operational processing and non-operational processing workload values to produce the new data activity's intermediate parameters.

For the comparable activities, processing workload intermediate parameters include the sum of operational processing, reprocessing, and non-operational processing parameters collected from the activities.

The estimation of operational and technical effort for the processing functional area follows the basic "curve-fit" approach outlined in Section 3.1 above, with "tuning" options for "nearness" testing and cluster outlier removal.

Processing operations effort is a function of three weighted workload parameters: product types generated operationally, count of total products generated, and total volume generated. In each case, these include new operational product counts and volume, reprocessed product counts and volumes, non-operational product counts and volumes. Effort components (i.e. the estimates based on individual parameters) are adjusted by comparison of new data activity values to corresponding comparable activity average values.

Level of service adjustments are made to the final processing operations effort estimate based on the Operational Processing LOS, and/or Non-Operational Processing LOS, and Processing Automation LOS. For the Operational and/or Non-Operational Processing LOS, if the new data activity's LOS is higher than the comparable activity average LOS, the effort estimate is increased, and if it is lower, decreased. If the Processing Automation LOS of the new data activity is higher than the comparable activity average, then the processing function of the new data activity is more highly automated than the comparable activity average, and the effort estimate is reduced, and if the Automation LOS is lower, the effort estimate is increased.

Support has been found in the comparable activity information for technical effort in the processing area, especially where product software is integrated and tested and science QA and validation are performed. Processing technical effort is a function of three weighted workload parameters: product types generated, product types integrated (i.e., where science software from an outside source is integrated and tested and put into production), and product types where

science QA and validation are performed by data activity staff (in some cases the outside source, such as a P.I. team, retains the responsibility for science QA). Processing technical effort components (i.e. the estimates based on individual parameters) are adjusted by comparison of new data activity values to corresponding comparable activity average values.

Level of service adjustments are made to the final processing technical effort estimate based on the Calibration - Validation LOS. If the new data activity's Calibration - Validation LOS indicates calibration - validation effort will be performed, the processing technical effort estimate is increased, and if not it is left unchanged.

As noted in Section 3.3 above for ingest, product formats are no longer used as a workload parameter.

In the future it may be desirable to use the number of production jobs executed as a processing parameter if comparable activity information for this can be obtained. It may be desirable to use a production complexity parameter to produce better estimates, since this can vary greatly, again if comparable activity information for it can be obtained.

3.5 Notes on Documentation Effort Estimation

The documentation functional area includes only technical effort.

Documentation technical effort does not use the general approach described in section 3.1 above, due to the lack of sufficient information. Instead, the base estimate of documentation technical effort is the average technical effort for those comparable data activities for which that information is available.

Level of service adjustments are made to the final documentation technical effort estimate based on the Documentation LOS, User Comment LOS, and Distribution Scope. For the Documentation LOS, if the new data activity's LOS indicates documentation to LTA standard, effort is added to the base estimate, and if the LOS indicates documentation to a current use standard, a smaller amount of effort is added to the base estimate. For the User Comment LOS, if the new data activity's LOS indicates routine use of user comments, effort is added to the base estimate, and if the LSO indicates occasional use of user comments, a smaller amount of effort is added to the base estimate. For the Distribution Scope, if the Distribution Scope indicates public distribution, then effort is added to the base estimate (assuming extra documentation effort is required to support a broad user community).

3.6 Notes on Archive Effort Estimation

The term "archive" in this context refers to the storage of data and products by the data activity whether temporarily (i.e. in "working storage"), for years prior to migration to a true long term archive, or indefinitely as a true long term archive, as indicated by the Archive Purpose parameter.

The CET allows the user to enter parameters describing a pre-existing archive, for example in a case where the new data activity is replacing an existing system and inheriting an archive in place (i.e., without having to do an ingest operation to accomplish migration to a new archive). The user would enter the number of product types, products, and volume for the pre-existing archive, in one or both of two categories, inactive (product types that are archived but are no longer being ingested or generated) or active (product types that continue to be ingested or generated). Pre-

existing inactive product types are additional product types the data activity must handle, while pre-existing active product types do not add additional product types, just more archived products and volume of types also being ingested or generated. A data activity with pre-existing archive must be either one with a multi-year or indefinite plan, i.e. not using only temporary working storage. Not also that by-request distribution draws on the entire available archive, but the CET assumes that 75% of the requests will be for data that is newly ingested or generated.

The archive functional area includes only operations effort. Analysis of the information for comparable activities shows that the archive function is in almost every case highly automated (e.g. the use of robotic storage is ubiquitous) and so there is no distinction to be drawn on the basis of degree of automation and thus no archive automation LOS.

Archive functional area intermediate parameters for both comparable data activities and the new data activity include year-by-year archive volume moved and archive transactions.

Archive volume moved is the year-by-year sum of the ingest volume and processing volume, on the assumption that all products ingested or produced are added to the archive.

Archive transactions are the sum of archive inserts and archive deletes. Archive inserts is the year-by-year sum of all products ingested and produced. Archive deletes is the total number of products removed from the archive each year. The count of archive transactions also includes archive reads made as part of random screening for quality (i.e. to detect archive media problems) according to the archive monitoring LOS.

Archive deletion is included in the current version of the CET. It is based on the retention period parameter associated with ingest and processing streams, which can specify a finite retention for the products included in a stream. It is also based on the reprocessing plan, which may specify deletion of the 'original' data when a new reprocessed version is produced.

In the future, products read from the archive for distribution could be added to the archive volume moved and archive transactions.

The estimation of operations effort for the Archive functional area follows the basic "curve-fit" approach outlined in section 3.1 above, with "tuning" options for "nearness" testing and cluster outlier removal.

Primary and backup archive are included, and 'archive' can be short term temporary 'working storage' as appropriate for an activity.

Archive operations effort is a function of four weighted workload parameters: product types archived, archive transactions, archive volume moved, and archived product count. Effort components (i.e. the estimates based on individual parameters) are adjusted by comparison of new data activity values to corresponding comparable activity average values.

Level of service adjustments are made to the final archive operations effort estimate based on the Archive Purpose parameter. If the new data activity's Archive Purpose is temporary working storage, the archive operations estimate is set to a low effort level consistent with comparable activities that use only working storage.

"Archive media units", cumulative archive volume, and "work" are no longer used as workload parameters.

3.7 Notes on Access and Distribution Effort Estimation

The access and distribution functional area includes only operations effort.

Access and Distribution includes operational distribution stream(s) of one or more product types each, and 'by request' distribution. In all cases, distribution can be by network and/or media. The estimation of operational and technical effort for the Access and Distribution functional area follows the basic "curve-fit" approach outlined in Section 3.1 above, with "tuning" options for "nearness" testing and cluster outlier removal.

For both the new data activity and comparable data activities, access and distribution intermediate parameters include the total year-by-year number of product types distributed, products distributed, and distribution work. The count of product types distributed is the number of product types archived, on the assumption that all products in the archive are available for distribution. The year-by-year totals of products distributed and volume distributed include all products and volume distributed by operational distribution streams and by-request distribution, and include both distribution by network and by media.

Access and distribution operations effort is a function of three weighted workload parameters (as they apply to a given new data activity): total product types distributed, total count of products distributed, and total volume distributed. Access and distribution effort components (i.e. the estimates based on individual parameters) are adjusted by comparison of new data activity values to corresponding comparable activity average values.

Level of service adjustments are made to the final access and distribution operations effort estimate based on the Distribution Means LOS. The implicit assumption is that a typical data activity performs less than half of its distribution (volume) by media. If the new data activity's LOS indicates that distribution is all by network, the access and distribution effort is set to a small level consistent with information for the comparable data activities with only network distribution, an implicit assumption that such distribution is highly automated. If the Distribution Means LOS indicates substantial distribution by media, the access and distribution effort is adjusted up accordingly.

As noted in Section 3.3 above for ingest, product formats are no longer used as a workload parameter, nor is "work".

3.8 Notes on User Support Effort Estimation

The user support functional area includes both technical and operations effort, based on the information about the comparable data activities used by the CET.

For both the new data activity and comparable data activities the CET computes as an intermediate parameter the year-by-year number of users contacted (i.e., the number of distinct users with whom the user support staff make some form of contact, by email, telephone, letter, visit, etc.), and the number of user contacts (the total number of emails, telephone calls, etc., for all users contacted).

For the new data activity, the number of users contacted is the number of by-request distribution users multiplied by the user multiplier parameter (the estimated number of times per year that an

individual by request distribution user will contact the user support staff). The year-by-year number of user contacts is the number of users contacted multiplied by the average contacts-per user-per-year parameter. For example, if 5,000 by-request users are projected for a given year, and 80% of them are expected to contact user support, 4,000 users will contact user support. If the average contacts per user per year is 1.5 (i.e. that, on the average, the 4,000 users would each be expected to contact user support 1.5 times during the year) then 6,000 user contacts are projected for the given year.

For comparable data activities, the number of users contacted and user contacts are based on information provided by the activities.

The estimation of operational and technical effort for the User Support functional area does not follow the approach outlined in Section 3.1 above. Instead, a "nearness" test is used to screen for comparable activities to be used for computation of the user support effort estimates. The 'nearness' parameter is the number of users contacted. Then the average operations and technical effort for the comparable data activities that pass the nearness test are used as the base estimates for the new data activity, with a minimum level assigned if no comparable activities pass the screening.

Level of service adjustments are made to the final user support operations and technical effort estimates based on the distribution scope and user contacts parameters. If the distribution scope indicates public distribution, the base estimates of operations and technical effort are increased, and if the projected number of user contacts is less than the average for the comparable data activities the base estimates of user support effort are reduced. If the distribution scope indicates limited distribution, the base estimates for user support operations and technical effort are increased if the user contacts projected for the new data activity exceed the average user contacts for the comparable data activities, and reduced if they are less than the average for comparable data activities.

3.9 Notes on Implementation Effort and Non-Staff Items Estimation

The Implementation functional area includes both technical effort (no operations effort) and non-staff cost items.

3.9.1 Implementation Effort Estimation

For Version 2.3 of the CET, the estimation of implementation effort, described below in section 3.9.1.1, has been decoupled from the estimation of SLOC (executable Source Lines of Code).

Prior to Version 2.3, the estimation of implementation effort was a two-step process, in which both steps follow the basic "curve-fit" approach outlined in section 3.1 above, including for each step the use of a "nearness" test to select the 'comparable' activities to be used for computation of the estimate. The first step was the estimation of the total amount of new software to be developed, i.e. SLOC, and the second step was the estimation of the total technical effort that will be required to develop the new software. The "curve-fit" approach is used for both the SLOC and effort estimates.

A SLOC estimate is still produced as described in section 3.9.1.2 below for use by in estimating Sustaining Engineering FTE (see section 3.10 below). Section 3.9.1.3 describes how implementation FTE was estimated by the CET prior to Version 2.3.

The estimated total implementation effort is spread evenly over the specified implementation period. Note that continuing implementation effort after the implementation period is assumed to be covered by sustaining engineering (see section 3.10 below).

3.9.1.1 Estimation of Implementation FTE

The CET computes several intermediate parameters used in the estimation of implementation FTE. These are total operations staff for the main operations areas of ingest, processing, archive, and distribution, and measures of volume, products, and product types handled. Total volume and products handled are summed over ingest, processing, archive, and distribution. Total product types handled is the count of product types archived, on the assumption that every type of product ingested or generated is included in the archive.

An overall automation score for the new data activity and for each 'comparable' data activity is computed from the ingest automation LOS, processing automation LOS, and distribution means LOS (note that the archive function is treated as highly automated in all cases). An average automation score for the comparable data activities to be used in computing the implementation FTE estimate is computed.

The estimate of implementation FTE is a function of four weighted parameters: total operations effort for main operational functional areas, total product types handled, total products handled, total volume handled. A single base estimate for the total implementation effort is made (as opposed to the year-by-year estimates of effort described in previous sections). Adjustments are made to the components of the implementation FTE estimate based on comparisons with comparable activity averages.

The automation score is applied to the implementation FTE estimate. If the new data activity has a higher value than the comparable data activity average, the new data activity is more automated than the comparable activity average. It is assumed that this implies more complex software, and hence more implementation effort, so an addition is made to the base estimate of implementation FTE. Conversely, if the new data activity has a lower automation score, then it is less automated, implying less complex software, thus less implementation effort required, and a reduction is made to the base implementation FTE estimate.

This approach does not account for reuse (other than the implicit assumption that reuse by the new activity will be at about the same level as reuse by the comparable activities) or a major system refresh during the life of a new activity. The user is prompted to use the CET's Reviewer tool to make any adjustments needed to account for planned reuse.

3.9.1.2 Estimation of New SLOC Developed

The CET computes several intermediate parameters used in the estimation of new SLOC to be developed. These are total operations staff for the main operations areas of ingest, processing, archive, and distribution; and total work, the sum of ingest, processing, archive, and distribution work. Total volume and products handled are also summed over ingest, processing, archive, and distribution. Total product types handled is the count of product types archived, on the assumption that every type of product handled is included in the archive.

An overall automation score for the new data activity and for each comparable data activity is computed from the ingest automation LOS, processing automation LOS, and distribution means LOS (note that the archive function is treated as highly automated in all cases). An average

automation score for the comparable data activities to be used in computing the SLOC estimate is computed.

The estimate of SLOC to be implemented is a function of five weighted parameters: total operations effort for main operational functional areas, total product types handled, total products handled, total volume handled, and total work. A single base estimate for the total SLOC to be developed is made (as opposed to the year-by-year estimates of effort described in previous sections). No adjustments are made to the components of the SLOC estimate.

The automation score is applied to the SLOC estimate. If the new data activity has a higher value than the comparable data activity average, the new data activity is more automated than the comparable activity average. It is assumed that this implies more complex software, and hence more SLOC, so an addition is made to the base estimate of SLOC. Conversely, if the new data activity has a lower automation score, then it is less automated, implying less complex software, thus fewer SLOC, and a reduction is made to the base SLOC estimate.

The total SLOC estimate is used in the estimation of sustaining engineering effort as discussed in section 3.10 below. Prior to Version CET 2.3, SLOC was used in the estimation of implementation effort as described in the next section.

3.9.2 Implementation Non-Staff Cost Estimation

Non-staff implementation items for which estimates are produced include system purchase cost, COTS software license purchase cost, and facility preparation costs. The estimates for these costs will be spread evenly over the new data activity's implementation period.

3.9.2.1 System Purchase Cost

System purchase cost includes the purchase of all hardware and the operating system and software bundled with the operating system.

The estimate of system purchase price has to take into account the rapid change in price for a given level of capability (e.g. processing power) that is a consequence of the rapid development of computing technology. The CET does this by normalizing comparable data activity information on system purchase costs to a common base year, then producing a base estimate for the system purchase price for a new data activity in terms of the same base year, and finally projecting the base year estimate forward to the planned implementation period for the new data activity.

"Moore's law", which calls for a doubling of capability for a given price (or halving of price for a given capability) every eighteen months, has proven to be a reliable predictor for changes in processing hardware cost. The CET uses a more conservative price reduction factor of 25% per year, which yields a price after 3 years of 42% of the base year price, compared to 25% for Moore's Law (halving in 18 months, twice). The reason for the more conservative factor is that for the CET the system purchase price includes peripherals, which decrease at slower rates, and operating system software, which generally does not decrease.

As an intermediate parameter, the normalized system purchase cost is computed for each comparable data activity for which system purchase cost information is available. The normalized cost is the 'raw' cost adjusted to a base year using the price reduction factor described

above. The CET also computes the total staff count for all of the comparable data activities to be used in making the estimate, and the estimated total staffing for the new data activity.

The estimation of base year system purchase cost follows the basic "curve-fit" approach outlined in section 3.1 above.

The estimate of base year system purchase cost is a function of three weighted parameters: total effort, total volume handled, and total work. A single base estimate for the total base year system purchase cost is made (as opposed to the year-by-year estimates of effort described in previous sections). No adjustments are made to the components of the estimate.

The base year system purchase price is then projected forward to the new data activity's implementation period, and spread over the implementation period (the base year cost is divided into equal portions for each year of the implementation period, and then each implementation period year's cost is reduced according to the CET's price reduction factor (see above)).

3.9.2.2 COTS Software License Purchase Cost

The estimation of COTS software license cost does not use the "curve-fit" approach. Nor is there any allowance for price reduction a 'la a "Moore's law" factor; the cost of COTS software has not been observed to decline.

The CET computes average COTS software license costs for two classes of comparable data activity, small and large, based on whether the data activity's total volume handled is above or below the comparable data activity average volume handled. The comparable activity average volume used is a value computed after deleting data activities with extraordinarily large total volume handled, i.e. greater than 1000 TB per year.

To make the estimate of COTS software license purchase cost for the new data activity, the CET first determines whether the new ADS activity falls into large or small class. If the new data activity is large, then the CET uses the comparable activity average COTS software license purchase cost for large data activities, if it is small, the average for small activities. The cost is then spread evenly over the implementation period.

3.9.2.3 Facility Preparation Cost

Facility preparation cost is the cost of outfitting of existing space, excluding major construction (i.e., the cost of building a new building or adding on to an existing structure.) Included are costs for installation of power, cooling, false floors, partitions, furnishings, etc., to get the space ready to use.

The CET assumes that facility preparation costs can range from $50K to $150K, for small to large data activities, based on maximum staff size. The maximum staff size over the life of a data activity is the total staff for the year when that value is the greatest. As an intermediate parameter, the CET computes the maximum year staff size for each comparable data activity.

In making the estimate, the CET uses the maximum year staff size for the new data activity, and computes an estimated facility preparation cost within the range of $50K to $150K by interpolating between the greatest value of maximum year staff size for a comparable data activity (set to correspond to $150K) and the lowest value (set to correspond to $50K).

3.10 Notes on Sustaining Engineering Effort Estimation

The sustaining engineering functional area contains only technical effort, and is computed for the period that follows the implementation period (usually the operations period but implementation and operations may overlap).

The estimation of technical effort for the Sustaining Engineering functional area follows the basic "curve-fit" approach outlined in Section 3.1 above, including the use of a "nearness" test to select the comparable activities to be used for computation of the estimate.

Sustaining engineering technical effort is a function of two weighted parameters: estimated total SLOC to be maintained (see section 3.9.1.2 above) and the total staff for the main operational functions (ingest, processing, archive, and distribution). Sustaining engineering effort components (i.e. the estimates based on individual parameters) are not adjusted. The base sustaining engineering estimate is the year-by-year weighted average of the sustaining engineering effort components.

Level of service adjustments are made to the base sustaining engineering technical effort estimate based on the sustaining engineering LOS and the automation score. If the new data activity has a lower sustaining engineering LOS than the average for the comparable data activities, the base estimate of sustaining engineering technical effort is reduced. If the new data activity has a higher LOS than the comparable activity average, the base estimate is increased. If the new data activity has a higher automation score than the average for comparable data activities, the base estimate of sustaining engineering technical effort is increased; more automation means more complex software and more effort required to sustain it. If the new data activity has a lower automation score than the comparable activity average, the base estimate of sustaining engineering technical effort is decreased since less automation suggests simpler software with less effort required to sustain it.

Note that while sustaining engineering is intended to include some degree of implementation after the implementation period (i.e., ongoing or periodic enhancement and addition of minor features) it does not include an allowance for major new functions or re-engineering.

3.11 Notes on Engineering Support Effort Estimation

The estimation of technical effort for the Engineering Support functional area does not use the "curve-fit approach". Analysis of the engineering support information for the comparable activities has shown that the data activities fall into two groups, 'large' and 'small' based on their staff levels for the operating activities, such that taking as an estimate of engineering support FTE of a new data activity the comparable activity average for engineering support FTE for the group it fell in, based on the (already computed) estimated FTE for its operating activities, was a much better estimate than the overall CD average. This technique is now used by CET Version 2.4.

Level of service adjustments are not made to the base engineering support technical effort estimate based on the engineering support LOS and the automation score. Analysis of the engineering support information for the comparable data activities showed that there were no usable (i.e. statistically meaningful) relationships between the engineering support LOS and/or automation LOS and engineering support staffing.

3.12 Notes on Technical Coordination Effort Estimation

The estimation of technical effort for the Technical Coordination functional area is based on a table of fixed 'plug values' for each area of technical coordination. Effort (constant over the life cycle) is added for each area flagged as applicable to the new data activity.

3.13 Notes on Management Effort Estimation

Management effort includes three components: activity-level management and coordination (e.g. a data activity manager, a project scientist, etc.); second-level management, i.e. management of functional areas within the data activity; and administrative support.

The "curve-fit" approach is not used for estimating management effort.

As an intermediate parameter, the CET computes the total "working" effort for each of the comparable data activities, and the estimated total working effort for the new data activity. Working effort includes all of the effort except for management. The CET computes averages of activity-level management effort, second-level management effort, and administrative support effort. The CET then computes the ratios of each of these components to the comparable activity average working effort.

The CET then estimates the three components of management effort for the new data activity by multiplying the new data activity's estimated working effort by the appropriate ratio.

3.14 Notes on Miscellaneous Non-Staff Cost Items Estimation

Miscellaneous non-staff cost items are system maintenance cost, recurring COTS SW licensing cost, recurring facility cost, recurring network / communications cost, supplies cost, training cost, travel cost, data purchase cost, and computer services cost.

3.14.1 System Maintenance Cost

Annual System Maintenance cost is estimated as 10% of the original system purchase price, and includes hardware maintenance and operating system fixes and upgrades. The CET uses the base year system cost without reduction - maintenance especially of system software does not decline.

3.14.2 Recurring COTS SW Licensing Cost

Recurring COTS software license cost is estimated as 12% of the original license purchase price.

The CET uses the base year license cost without reduction - the cost of COTS software does not decline.

3.14.3 Recurring Facility Cost

Recurring facility cost includes utilities, facility upkeep, etc., not initial outfitting or furnishing.

The CET assumes a fixed per-FTE rate of 15K$ per estimated ADS staff FTE, based on data from NOAA (14K), LaRC (20K), T/P Facility (11.5K), some etc.

Note - this cost applies every year, on top of facility prep in early years, since as soon as there is staff, there are costs covered by this.

3.14.4 Recurring Network / Communications Cost

The CET estimate for recurring network / communication cost is based on total volume ingested and distributed by network. The CET assumes a T1 gross rate of 1.5 mbits / sec, or 0.19 MBytes/sec. The CET assumes an overall efficiency of 70%, and thus a net rate of 0.13 MBytes/sec, which is 11.23 GB/day (.13 x 86,400 /1000) or 4100 GB/Year (11.23 x 365), or 4.1 TB/Year.

The CET assumes a base year (2003) cost of $3.6K / year per full T1, per 2003 commercial rates, and a future rate cost reduction factor of 25% (note - Grid article claims a nine month halving time!)

The CET assumes a minimum requirement for a site of one full T1.

3.14.5 Supplies Cost

The CET assumes three components to Supplies Cost: General Supplies, including office IT, miscellaneous etc.; Archive Media; and Distribution Media.

3.14.5.1 General Supplies

The CET assumes a base rate of $10K / year plus 1.5K per FTE.

3.14.5.2 Archive Media

The CET assumes a base year (2003) average media unit capacity of 50GB (DLT tape), and cost of $100 per tape, or $2 per GB. This might be a bit high, but tapes will not be 100% full.

The CET assumes a price reduction rate of 15% per year (consistent with ESDSI/SOO Vanessa Griffin's May 2003 email).

The CET skips and zeros out costs if there is no archive function indicated for ADS.

3.14.5.3 Distribution Media

The CET uses ESDIS/SOO Vanessa Griffin’s May 2003 table, which projects a changing mix of distribution media for FY2003 - FY2006, and costs including media, postage / shipping, spread over the DAACs.

The CET assumes the DAAC average cost to be roughly equivalent to an average for the all of the comparable data activities.

The DAAC average cost was 35K in the base year 2003, and the CET assume an annual reduction in media cost of 15%.

The estimate for a new data activity uses this data in conjunction with the Distribution Means LOS as follows:

If the Distribution Means LOS is 1, there is no Media distribution, and therefore no cost.

If the Distribution Means LOS is 2, then there is some Media distribution (less than half of total distribution). The CET uses half the average cost as a base, and bumps it up or down 20% if ADS activity media volume is larger or smaller than the average for all of the comparable data activities.

If it the Distribution Means LOS is 3, then there is mostly Media distribution. The CET uses the full average price as base, and bumps it up or down 20% if ADS activity media volume is larger or smaller than the average for all of the comparable data activities.

3.14.5.4 Training Cost

Training costs are estimated as a function of technical and operations staffing level. A base cost of $1500 per year per person trained is assumed. It is assumed that 75% of the technical and operations staff receives training in each of the pre-operations implementation years, and in the first year of operations. It is assumed that 25% of the technical and operations staff receive training in each of the remaining operating years, covering refresh/update and staff turnover.

3.14.5.5 Travel Cost

The travel cost is provided by the user as an annual travel budget for the new data activity.

3.14.5.6 Data Purchase Cost

The data purchase cost is provided by the user as an annual data purchase budget for the new data activity.

3.14.5.7 Computer Services Cost

Computer services cost can include any form of IT support obtained by the data activity from an outside source; e.g., processing capacity, reproduction and distribution of media, etc., that the user wishes to have included in the CET's life cycle cost estimate. The computer services cost is provided by the user as an annual computer services budget for the new data activity.

3.15 Notes on 'By Request' Distribution Estimation

A simple approach is used to project the number of estimated by-request users per year, the number of requests per year, the number of products and volume distributed by request on media per year, and the number of products and volume distributed by request by network per year. The inputs provided by the user describing the new activity are the peak yearly number of users, average number of requests submitted per year by each user, average number of products per request, and the proportion of the products and volume that will be requested for network delivery. Inputs from the archive functional area are the number of product types, product counts, and total volume that is available for distribution.

The CET uses a simple growth curve for the number of users that projects an initial sharp increase leveling out over the last portion of the operations period, finally reaching the maximum value provided by the user.

A logarithmic equation, Y = .a*ln(X)+b, is used to compute a the fraction of the maximum Y for each year X, where a = .33 and b = .25. The year used in the equation must be rescaled to a 1 to 12 year base that was used to derive the equation, i.e. if the number of years in the operating period was 5, 5 must be mapped to 12 and years 2, 3, and 4 will be mapped linearly between 2 and 11. The Y values (each between 0 and 1) is then multiplied by the maximum expected value of each by request parameter (e.g. number of by request users) to obtain the value for each year.

Then the number of requests and products requested per year are computed from the user-provided averages, and the volume is computed using an average product size from the archive information.

3.16 CET Sensitivity Test

This section describes how the sensitivity test has been implemented in the CET. A three step process is involved. First, while the CET estimate is being produced, each estimating procedure for the operating functional areas (ingest, processing, archive and distribution) computes a range of estimates for different values of each workload parameter. Then, when the estimate has been completed, sensitivity values are computed and stored in the "Sensitive" worksheet and finally the user selects a workload parameter and the sensitivity results are generated and displayed to the user, with this last step being repeated for all of the workload parameters the user wishes to test.

3.16.1 Calculations Made By Each Functional Area Estimating Procedure

The estimating procedure for each operating function (e.g. ingest) compute estimates for the years the function is active. The procedure computes five estimated FTE values for each selected parameter, for each year, applying the sensitivity thresholds to get five parameter values, one (the third) being the neutral value, with two less than and two greater than the nominal value. The estimate for this value is the normal estimate, and only it is used to set the final estimate values.

After computing the set of five estimated values for five values for each year of the given workload parameter, a procedure is called to compute sensitivities. The sensitivity is defined as the fractional departure of the estimate from the nominal value divided by the fractional departure of the workload parameter value from its nominal value. The sets of estimates are averaged over the years (i.e. an average is computed for each of the sets of values for each of the sensitivity thresholds) producing one set of five estimates. Those five values are then used to compute sensitivities (with the sensitivity for the neutral estimate being a dummy value, 0, that is not used). The output of the procedure is a set of five sensitivity values stored in the "Sensitivity" worksheet.

3.16.2 Calculations Made After Estimate is Completed

The overall sensitivity procedure is called after the entire estimate has been completed and the sensitivity results as described above have been obtained for all of the workload parameters for all of the operating functions and are stored in the "Sensitivity" worksheet.

The procedure computes the changes in overall life cycle total activity FTE that result from a range of changes in a workload parameter selected by the user. The user selects the parameter to vary. The range of variation of the workload parameter is from -50% to +100%. Workload parameters that the user can vary are: Ingest - product types, products, volume, external

interfaces; Processing - product types, products, volume, product types integrated, product types QA'd; Distribution - products, volume.

Ripple (i.e. a change in one parameter inducing changes in other parameters) rules apply:
1. A change in Ingest workload ripples to Archive and Distribution (except for External Interfaces, which does not ripple).
2. A change in Processing workload ripples to Archive and Distribution (except for Product Types Integrated or QA'd, which do not ripple).
3. Changes to the operating areas ripple to SLOC and Implementation and Sustaining Engineering.

Ripples to archive from ingest and processing will be done by adding increased workload to archive - i.e., for products, the variation pct is used to compute an increase in products ingested or generated, and the same number of products is added to the archive transaction count - the assumption being that all newly ingested or generated products are archived.

Ripples to distribution are complicated by the fact that there is no direct or clear cut relationship between ingested or generated products/volume and distribution. So assumptions need to be made... and here they are:

Case 1: Site does not generate its own products. Assume that data that is ingested is ingested to be archived and distributed. Assume that the base level ratio of products/volume distributed to products/volume ingested continues to hold as ingest products/volume are varied. So distribution ripple is increase in ingest (fractional change X base level) multiplied by distribution/ingest ratio. In the case of product types, the additional number of new types ingested is added to the number of product types distributed.

Case 2: Site generates its own products. Assume that data that is ingested is ingested as input into the site's product generation process (an over-simplification in some cases to be sure), and so generate no distribution ripple from ingest products/volume. Assume that the site generates products for distribution. Assume that the base level ratio of products/volume distributed to products/volume generated continues to hold as generated products/volume are varied. So distribution ripple is increase in production (fractional change X base level) multiplied by distribution/production ratio. In the case of product types, the additional number of new types ingested is added to the number of product types distributed.

The procedure loops through the sequence of test variations (from -50% to +100% in 10% steps) in workload parameter. For each one it produces a new FTE estimate for the variation in the selected workload parameter (using the five estimate values computed as described above for the five sensitivity thresholds as a basis for interpolation), imposes the ripple rules according to the user's parameter selection. These estimates are then converted to percentage changes, and plotted against the corresponding percentage changes in the workload parameter, thus producing the sensitivity graph displayed to the user

References

1. Cost Estimation Toolkit Users' Guide, September 2008 (for CET Version 2.4)

2. LOS/CE (Level of Service / Cost Estimation) Working Paper 1 - Study Overview and Technical Approach

3. LOS/CE Working Paper 2 - Cost Estimation by Analogy Model

4. LOS/CE Working Paper 3 - Data Service Provider Reference Model - Functional Areas

5. LOS/CE Working Paper 4 - Data Service Provider Reference Model - Model Parameters

6. LOS/CE Working Paper 5 - Data Service Provider Reference Model - Requirements and Levels of Service

7. LOS/CE Working Paper 6 - ESE Logical Data Service Provider Types

8. ESDIS Data Center Best Practices and Benchmark Report, SGT, September 28, 2001

9. Statistical Methods, by George W. Snedecor and William G. Cochran, Iowa State Press, 1989.