SOFTWARE DATA LIBRARY

run under the U.K. Alvey initiative

PURPOSE: to set up a library of software engineering data

This talk discusses how the SoftWare Data Library project (SWDL) intends to motivate data collection within data providing organisations.
Data Collection with SWDL - Why?

a) Medium and Long-term Advantages

- access (eventually on-line) to a large volume of software engineering data
- Regular SWDL newsletter giving summary statistics of the database contents
- SWDL will process requests for in-depth analysis of your own data or the general data pool
- insight into software engineering concepts
- access to prototype tools being developed using SWDL data.

b) Short Term Advantages

- consultancy about data
- consultancy about data collection
- extensive support for data collection (methods, manuals, forms and graphs)
- immediate statistical analysis
- consultancy about data analysis techniques

However a project manager will not begin collecting data unless they see it as giving immediate advantages to their project.
LIBRARY TOOLS FOR DATA COLLECTION

Data Providers will receive

- on site design of prototype forms and graphs to suit each project of the provider (LIBRARY representative with portable PC and printer)
- quick high-quality production of the customised forms and graphs, with symbol stickers, templates etc.
- a manual on data collection with the following sections

1) Short-term and Project Control Advantages of Data Collection
2) Examples of Data Collection and Use
3) Comparability
   (the information needed to permit comparison of data recorded in different environments)
4) Procedures for collecting, transmitting and validating data and the time and effort required by these
5) Confidentiality, Privacy, Integrity and Security

(The rest of this talk discusses the ideas of section (1).)
(Presentations on the other sections are available.)
Short-term and Project Control Advantages of Data Collection

Data Collection in a given area
- enhances understanding of that area
- enhances management control of that area

To obtain these benefits you need
- the data
- a method of extracting the meaning of the data

LONG TERM
1) elaborate analysis using data from many completed projects

INTRA-PROJECT CONTROL
2) quick, robust analysis of the data collected to date from one project.

SWDL data collection techniques aim to combine data recording with quick, graphical interpretation.
Data Recording Instruments

**GRAPHS**

**Advantages**
- high ratio of information to recording symbols
- reduces the number of processes involved in recording and using data
- instant analysis of data
- compact display of widely-varying amounts and types of data
- useful to record summary data on even if the data is already held on forms for clerical reasons
- make data comprehensible

**Disadvantages**
- possible problems of clutter, smudging, etc.*
FORMS

Advantages
- necessary as a basis for plotting many different graphs
- no need to think about spatial relationships

Disadvantages
- data must then be plotted or analysed by computer to be understood

All data can be recorded on forms or on graphs or both. Each SWDL graph has a corresponding form for recording the same data. In addition there are many analysis graphs.

Data providers choose the mix they want and request the axis scales, sizes, etc. suited to each project.

MACHINE-HELD DATA
- negates 'id comments above
- this is the direction in which data collection should go
Graphical Analysis of Data

1) Advantages of Graphical Analysis Techniques

Robustness

- software engineering data routinely violates the assumptions of classical statistics
- methods that make "hidden assumptions" about the data are very untrustworthy

User-Friendliness

- a graph is more meaningful than a table of numbers
- computer graphics can show dynamically how an analysis method works
- a series of graphs will be accepted in places where a table of numbers marked with an intimidating name causes alarm and despondency
Figure 4.2 The four scatter plots all have correlation coefficient equal to 0.70.

Figure 4.3 The four scatter plots all have correlation coefficient equal to 0.70.
2) The Nature of Graphical Analysis

Quick graphical analysis for project management consists of

- identifying the normal trend in the data
- detecting anomalous (i.e. outlying) points

The organisation has a software development process which produces software of saleable quality. In the long-term they may want to change their software development process due to LIBRARY analysis. In the short-term they want to

- understand it (for project prediction)
  done by identifying the normal trend
- recognise when it has gone wrong (for project control)
  done by outlier detection.
Figure 5: Complexity against size of module for subsystem Z
FIGURE 1  Relationship Between Program Size and Number of Errors

Number of Errors Per Program

Size of Program in lines of Code
number of errors

Program size (lines of code)
3) Methods for Analysis

These techniques can be applied on site by the visiting LIBRARY representative (using portable PC)

UNIVARIANT

i) Boxplots

BIVARIANT

ii) resistant line plotting

iii) robust bivariant regression

The techniques below would be used at the LIBRARY site to prepare a detailed report on the provider's data

MULTIVARIANT

iv) principal components – use for grouping data, e.g. into dimensions related to size, dimensions related to effort, etc..

v) Clustering – look for groups within the data items within which trends and outliers may be sought.

vi) Multivariant regression (when possible) – decide on a predictor and examine residuals.

vii) CART (Classification and Regression Trees) – still at an experimental stage
Figure 3: Boxplot for H/E code instructions for subsystem 1.

Changes < 5

XXX X

Changes ≥ 5

0 800 1600
RELATIONSHIP BETWEEN APTITUDE TEST SCORES AND SUBSEQUENT PROGRESS IN COMPANY