Problems, Methods and Specialisation

Michael Jackson

(A Contribution to the Special Issue of SE Journal on Software Engineering in the Year 2001)

Draft Version of July 1, 1994
Please do not copy or quote

Abstract

Software Engineering is not a discipline: it is an aspiration, as yet unachieved. Many approaches have been proposed, including reusable components, formal methods, structured methods, and architectural studies. These approaches chiefly emphasise the engineering product: the solution rather than the problem it solves. An approach to understanding and classifying software development problems in terms of problem frames is suggested. In addition to such general approaches, specialisation is essential: the established branches of engineering are all specialisations. Some specialisations have arisen in software development, notably in compiler construction and software for personal computers. More are needed.

1 The Aspiration

The term Software Engineering is usually thought to date from 1968. The report of the first Software Engineering conference [12] explains the background:

“In late 1967 the [NATO Science Committee] Study Group recommended the holding of a working conference on Software Engineering. The phrase ‘software engineering’ was deliberately
chosen as being provocative, in implying the need for software manufacture to be based on the types of theoretical foundations and practical disciplines that are traditional in the established branches of engineering.”

The aspiration was that software engineers would take a merited place among the ranks of civil, electrical, aeronautical, chemical, structural and automobile engineers and their colleagues in the other established branches. By the year 2001 this aspiration, to create and practise a discipline of Software Engineering, will be 33 years old—one third of a century. But the 26 years that have passed so far give little reason to think that the aspiration will have been achieved in another seven years. Software engineering has not become like the other established branches, and it will not become so in the near future.

One reason is simply that it’s difficult. Fred Brooks, in his much-quoted paper [2], said:

“The essence of a software entity is a construct of interlocking concepts: data sets, relationships among data items, algorithms, and invocations of functions. This essence is abstract, in that the conceptual construct is the same under many different representations. It is nonetheless highly precise and richly detailed.

“I believe the hard part of building software to be the specification, design, and testing of this conceptual construct, ...

“If this is true, building software will always be hard. There is inherently no silver bullet.”

Another reason for our failure is that we have added to the essential difficulty of the task by a simple lack of professionalism. In its earliest days, software development grew out of mathematical programming, whose practitioners were mathematicians and physicists. They regarded computer programming as an essentially trivial activity by comparison with their professional work. The boundary between ‘user programming’ and ‘professional programming’ that eventually grew up in the mainframe Fortran and COBOL culture has become blurred in the PC culture. Visual Basic, Paradox and Lotus 1–2–3 are all tools intended for both the most serious and the most casual use. The established branches of engineering do not suffer in this way: there
are no casual builders of motor cars or bridges; there is no do-it-yourself kit for designing steel structures. But in software development it is not easy to draw a clear line between the casual developer and the serious, professional, developer.

As a result, software development is still largely an amateur activity in a very important sense. In a paper analysing the Therac-25 accidents [10], an official of the US Food and Drug Administration (FDA) is quoted as saying:

“A significant amount of software for life-critical systems comes from small firms, especially in the medical device industry; firms that fit the profile of those resistant to or uninformed of the principles of either system safety or software engineering.”

The software for the Therac-25 had errors in simple sequential logic. Its developers were also apparently unaware of the pitfalls associated with concurrent access to shared variables, a problem whose solution, by T Dekker, had been reported in the year of the NATO conference [4].

Another reason for the failure of our aspiration to become engineers is that many theorists and practitioners, even among the most professional, are naïve. The continuing search for panaceas, for universal materials out of which everything can be made, for universal methods to solve all problems, strongly suggests that we have not yet begun to understand the nature of our field. Object-orientation is only the most recent in a long line of proposed panaceas, all claiming unbounded applicability and potency. Even those who complain most bitterly of our lack of professionalism are inclined to exhibit this naïveté. Proponents of formalism, for example, are right to claim that if software development were carried out with more care, more mathematical rigour, and more precision, some of the egregious errors that are so often made would probably be avoided. But the prescription is too readily offered as a complete and universal cure. It is not. Our diseases lie deeper.

2 Prescriptions for Success

Many prescriptions have been put forward to cure our obviously unhealthy state. One strong medicine is the development and use of catalogues of ready-made components, following the example of electronic engineers and some
others. In an invited address at the NATO conference [12], Doug McIlroy said:

“Components, dignified as a hardware field, is unknown as a legitimate branch of software. When we undertake to write a compiler, we begin by saying ‘What table mechanism shall we build?’ Not ‘What mechanism shall we use?’, but ‘What mechanism shall we build?’. I claim we have done enough of this to start taking such things off the shelf.”

Some progress towards this goal has been made in particular environments: libraries of mathematical routines and object classes for GUIs are notable examples. Programming languages come equipped—like C—with libraries of callable functions, or—like COBOL—with repertoires of elaborate statements. And in many areas there is a steady drain of functionality from application programs into the technical environment—the operating system and the DBMS and the communications system. Yet most everyday programming still defies reduction to component assembly. Either the need is too specialised, and the interfaces of the available components don’t fit the context in which they are to be used. Or, perversely, it is too general. How many programmers, even while you read this sentence, are programming linear search? And some of them, no doubt, are programming it wrongly.

Proponents of formal methods, such as Z [19] and VDM [9], offer a different diagnosis and therapy: they point to the often neglected mathematical aspect of software development. Undeniably, computer programs can be regarded as formal things, and can therefore be subjected to mathematical treatment. Formalists want to emulate the practitioners of the established branches of engineering. The structural engineer calculates stresses; the automobile engineer calculates torques and accelerations and wind resistance. The software engineer, they believe, should be calculating too. Calculation would reveal the implicit precondition of an operation, and show that it might be invoked when its results were unpredictable. Or it might show that a proposed refinement was invalid. Obviously, it is absurd to tolerate vagueness, confusion and uncertainty where precise calculation is possible. The Roman engineers would surely not have refused the offer of a better system of arithmetic.

But in many areas of software development the prerequisites for a more mathematical approach are not yet in place. Engineers in the established
branches make their calculations in very well-defined and narrow contexts. An automobile engineer, for example, does not set out to design a new car *de novo*, calculating whether to use a steam engine, whether to use tracks, or articulated legs, or four wheels, or eight, or whether the driver should face sideways or forwards. There is a repertoire of standard designs, evolved over many decades of experience, and each new design is at most a small perturbation from the standard. To combine the chassis and body in one unit, or to set the engine transversely, was a radical departure open only to the most brilliant and daring designers. In practice the design space is very narrow. The engineering calculations are done on familiar components in a familiar configuration.

In most areas, software development has no such established standard designs. The structures offered by formal methods are too general to narrow the design space. The system is seen as a set of operations on a state, or as an assemblage of communicating sequential processes: these are general computational paradigms, and offer almost no guidance to the designer faced with any particular development task. Understanding the problem and structuring the solution is the primary need. Only when that has been done, at least in outline, is the subject matter for calculation available. Rigorous description and calculation must come second. It is then crucially important for some aspects of the task, but it is not always at the heart of the problem.

‘Structured’ [20] and ‘Object-Oriented’ [11] methods are concerned with the problem of designing software structures. In most cases their prescriptions are very general. Every piece of software can be described—at least, to some extent—by a dataflow diagram. Every program can be built—more or less appropriately—in an object-oriented language. The methods associated with these notions give only general guidance in the crucial questions that the designer faces: *What* objects are needed? *What* functions and data streams? The material in which the design is to be expressed gains generality from its recursive structure: the functions in a dataflow diagram can themselves be described by dataflow diagrams; and a subclass in an object inheritance hierarchy can itself be a superclass. But this very generality gives a design space that is too wide for comfort.
3 Problems and Solutions

The architectural approach of Garlan and Shaw [5] moves towards a narrowing of the design space facing the developer. They are concerned to classify and study common architectural styles such as pipes-and-filters and blackboard-connected-processes. Their approach has something in common with the work on programming clichés [16], and with the ‘Pattern Language’ approach [7] that draws on the work of the software methodologist’s favourite architect, Christopher Alexander [1].

All of this work approaches the design task from the side of the software-hardware machine. It focuses on the characteristics and structure of the solution, and so offers a prospect of evolving the kinds of standard design that are found in the established branches of engineering. But there is another side to the development task: the characterisation of the problem to which the software-hardware machine provides a solution.

Johnson, one of the proponents of patterns, says in [8]:

“Alexander focuses as much on the problem to be solved and the various forces on the problem as he does on the solution to the problem. We have a tendency to focus on the solution, in large part because it is easier to notice a pattern in the systems that we build than it is to see the pattern in the problems we are solving that lead to the patterns in our solutions to them.”

Johnson’s point is perceptive. The concentration on solutions is more widespread than may be recognised. Some methods seem to be analysing or structuring the problem when in fact they place all their emphasis on describing a solution. Johnson gives one reason: solution structures are more easily seen than problem structures. Another reason is that a software system is often a partial simulation of its problem domain: the developer is then easily persuaded that a description of the software is a description of the problem. But it is scarcely more so than a description of a car steering wheel is a description of the physiology of the driver’s arm and hand.

Focus on the problem to be solved is implicit in the established branches of engineering, because the materials and techniques, and the repertoire of standard designs, have a narrowly limited applicability. An automobile engineer set to design a bridge would be at a loss. But the materials and
techniques of software development, in those areas where we lack standard
designs, have a very wide applicability. The problem can not be taken for
granted: it must be quite explicitly identified and analysed.

4 Problem Contexts

Responding to Fred Brooks, Wlad Turski [18] pointed out that software
development is concerned with more than a formal computing system:

“But, as we have observed before, software has another aspect:
that of describing properties of an application domain. In this
sense software does not always relate two formal systems [soft-
ware and hardware]. Many computer application domains are
not formal systems at all.

“There are two fundamental difficulties involved in dealing with
non-formal domains (also known as ‘the real world’).

1 Properties they enjoy are not necessarily expressible in any
single linguistic system.

2 The notion of mathematical (logical) proof does not apply to
them.”

The point is that software developers can not ignore the application do-
main: identifying, capturing, understanding and analysing the problem in
its context is an integral and essential part of our concern. In consequence,
software development—unlike hardware development—is about banking and
telephone switching and air traffic control and avionics and compiling pro-
grams and calculating spreadsheets and formatting and displaying texts. In
short, it is about everything that can furnish the subject matter for a program
or system.

At first sight, this might be taken to mean that a software developer
must be—or become—expert in the application domain. But this is not so.
What it means is that the software developer must become expert in those
aspects of the application domain that affect the design and construction of
the software. A domestic heating engineer is not expected to be an architect,
but must be expert at analysing those aspects of a building—the ambient
climatic conditions, the locations of doors and windows, room volumes, wall and roof insulation, and traffic inside a house—that affect the demand for heat and present opportunities and difficulties to the designer of a domestic heating system.

Some attention [15] has been paid to application domains considered generically: to the banking domain, or to missile applications, or to strategic-management-support-systems. And some development methods (for example, JSD [6]) have advocated explicit analysis and description of the particular ‘real world’ for each particular system. Identifying and understanding the ‘real world’ must be a first step towards understanding the problem to be solved. It is the context in which the problem exists.

Consider, for example, this classic problem (adapted from [17]):

“A patient-monitoring program is required for a hospital. Each patient is monitored by an analog device that measures factors such as pulse, temperature, blood pressure, and skin resistance.

“The program reads these factors on a periodic basis (specified for each patient). For each patient, safe ranges for each factor are specified. If a factor falls outside a patient’s safe range, or if an analog device fails, the nurse’s station is notified.”

The problem context is a real world, separable into distinct domains (using the word to denote a distinguishable part of the specific problem context rather than generically to a class of problem context). There are patients; there are analog devices; there is a source of specifications (presumably the medical staff) of safe ranges and monitoring periods; and there is a nurse’s station. The analog devices are connected to the patients; the nurse’s station is not connected to any other part (but will be connected to the machine we are to build). The medical staff will deliver their specifications of ranges and periods to the machine in some way. The problem will be to construct the machine. In software development, the problem is always to construct a machine—but a machine that will fit into a particular context in the world, where its costs and benefits will be felt and evaluated.
5 Problem Frames

Identifying the context of a problem is only a first step towards understanding the problem itself. The ancient Greek mathematicians paid a lot of attention to the study of problems, treating it separately from the related study of solutions and solution methods. An admirable little book by Polya [14] gives an account of their ideas. They classified mathematical problems into *problems to find or construct*, such as:

Given lengths $a$, $b$, and $c$, construct a triangle whose sides are of those lengths.

and *problems to prove*, such as:

Prove that if the four sides of a quadrilateral are equal then its diagonals are mutually perpendicular.

A problem can be characterised by its *principal parts* and a *solution task*. The principal parts of a problem to prove are the *hypothesis*—that the four sides of a quadrilateral are equal; and the *conclusion*—that its diagonals are mutually perpendicular. The solution task is to show that the *conclusion* follows from the *hypothesis*. The principal parts of a problem to construct are the *unknown*—a triangle; the *data*—three lengths $a$, $b$, and $c$; and the *condition*—that the triangle’s sides are of the lengths given in the *data*. The solution task is to construct the *unknown* so that it satisfies the *condition* with respect to the *data*.

The essence of the principal parts is that they are parts of the problem, not of a solution or of the steps towards a solution. This allows a discussion of methods to be cast in the appropriate terms: that is, in terms of the problem. Polya gives methodological recommendations for each kind of problem: ‘Split the *condition* into parts’; ‘Check that you are using all the *data*’; ‘Think of a variation of the *unknown* to bring it closer to the *data*’.

The principal parts and the solution task of a problem form a structure within which the problem can be considered systematically, and an appropriate solution method chosen or devised. Such a structure may be called a *problem frame*. To understand a problem is to have fitted it into an appropriate problem frame by identifying its principal parts and the solution task. Even in the small problems discussed by Polya, the choice of problem
frame and the identification of the principal parts is not always obvious. As
Polya points out, the problem ‘Prove that there is an infinity of primes’ does
not readily fit the pattern of problems to prove; and the problem ‘Show that
there is at least one prime between 6 and 10’ can be treated either as a
problem to find or as a problem to prove.

The idea of a problem frame can be applied quite directly in software
development. We must begin, of course, with the problem context, identi-
fying the relevant parts of the real world and making an initial assessment
of their properties and relationships. In mathematics we can summon up a
rich body of domain knowledge by a single word. We say that the unknown
is a triangle, and we expect the reader to know what that means without
further explanation. We also expect the reader to be familiar with at least
some of the properties of a triangle—that the sum of the interior angles is
180 degrees, that the length of each side must be less than the sum of the
lengths of the two other sides, and so on. But in software development,
the ‘real world’ domains are much richer and more varied than the abstract
domains of mathematics. As Turski points out, they are often non-formal
and their properties may not be expressible in any single linguistic system.
The software developer must therefore be well-equipped to tackle the task
of investigating the relevant properties of these ‘real world’ domains and
formalising them in a variety of appropriate languages.

6 Software Problem Frames

Polya’s two problem types, as he recognises, are insufficient to structure
all problems even of the small class he discusses. Software development
problems are far more various, and we must expect to need many more than
two problem frames. Few problem frames have been explicitly described.
Methods tend to deal in solutions rather than in problems, or to leave the
problem frame partly hidden, implicit in the terminology used to describe
the method. But we can still identify a few software problem frames and
describe them roughly.

The JSD method uses a problem frame appropriate to the development
of information systems. The principal parts of the JSD problem frame are
these:

*Real World*  This is the particular world about which the system is to produce
information. It is a domain in the problem context. It is dynamic: that is, it has a behaviour over time, in which events, and consequent state changes, occur. The Real World is autonomous: that is, its events and state changes are regarded as occurring spontaneously and not as externally stimulated or controlled.

Information Outputs These are the outputs containing the required information about the Real World. They are a domain in the problem context.

Requests These are the information requests made by the users of the information system. They are a domain in the problem context. The Requests are a collection of unstructured streams of time-ordered events. The users are treated purely as a source of these unstructured streams: for example, individual users are not distinguished, and the domain is not regarded as having any internal state.

System This is the machine to be built. It is connected to the Real World, Information Outputs, and Requests domains. It produces the Information Outputs both in response to the Requests and autonomously according to the state and behaviour of the Real World.

Function This is the required relationship that the System must cause to hold among the Real World, the Information Outputs, and the Requests.

The Solution Task in the JSD frame is to construct a System that models, or simulates, the Real World and satisfies the Function.

Another problem frame may be called the Workpiece problem frame. It may be suitable for such applications as word-processing. It has these principal parts:

Workpieces The objects, often textual and graphic documents, that are to be worked on. The Workpieces are an intangible domain of the problem context. They are capable of changing their state, but only as a result of external action: they have no autonomous behaviour.

Operation Requests The requests made by the users of the system for operations to be performed on the Workpieces. They are a domain of the problem context. Like the Requests of the JSD frame, the Operation Requests are regarded as an unstructured stream of events, neither individual users nor domain states being distinguished.
Operations The operations that the users can ask the Machine to perform on the Workpieces. They are a required relationship between the Operation Requests and the states of the Workpieces domain.

Machine The machine to be built. It contains a reification of the Workpieces, and performs the Operations on them in response to Operation Requests. It has no autonomous behaviour.

The Solution Task in the Workpiece frame is to construct the Machine to perform the Operations on the Workpieces in response to the Operation Requests.

A final example may be called the Environment-Effect frame. It has something in common with an approach reported by Parnas and Madey [13]. It is suitable for an embedded system that controls an external domain. The principal parts are:

Environment The domain to be controlled. It has state, and a behaviour that is partly autonomous and partly responsive.

Connection The connection between the machine to be built and the Environment by which the machine can sense and affect states and events in the Environment. It is a domain of the problem context.

Machine The machine to be constructed. Its behaviour must be partly autonomous and partly responsive.

Requirement The domain properties and behaviour—relationships among phenomena of the Environment—that the Machine is to bring about and maintain.

The Solution Task in the Environment-Effect frame is to construct the Machine so that it senses and controls the Environment through the Connection, and ensures satisfaction of the Requirement.

Although these problem frames have been only roughly sketched, it should be evident that they are far from interchangeable. The chosen problem frame must fit the problem. The domains of the problem context must have the characteristics demanded of corresponding principal parts, and must be connected, directly or indirectly, to the machine in the way demanded. The
frame must have appropriate parts to accommodate all the required properties and relationships to be described and considered.

The Workpiece frame, therefore, would be useless for developing a system to control a chemical plant, because the Workpieces domain is assumed to be intangible and inert, but the chemical plant is neither. The JSD frame is inadequate for an embedded system, because its Real World domain is assumed to be autonomous. So there is no principal part—as there is in the Environment-Effect frame—corresponding to the domain properties that are desired as opposed to those that are given.

7 Complexity and Composition

The principal parts of a problem frame furnish the material for a development method. A method prescribes a problem frame, and offers guidance on identifying the domains of the problem context and the principal parts of the frame. It stipulates that certain descriptions are to be made, starting with descriptions of principal parts. It may stipulate an order of description; the various languages to be used; and operations—such as abstraction, composition, transformation, decoration and refinement—by which new descriptions may be derived from those already made. The culminating descriptions, of course, are descriptions of the machine to be built.

A good software development method prescribes a very specific problem frame, and exploits its properties to the full. The known characteristics of each principal part allow the simplest possible language to be chosen for its description. For example, the Real World in a JSD problem can be described in terms of concurrent sequential processes. Known relationships among the domains of the problem context allow descriptions to be composed in simple ways. For example, the Workpieces of the Workpieces problem frame can be described as abstract data types, and the Operations are then simply the operations of the type. Potential difficulties can be categorised and specific solutions offered.

But software development problems are too various and too rich to be captured by any reasonably small set of problem frames. They exhibit complexity, in the sense that more than one problem frame is needed for each problem. The construction of a CASE tool might seem to fit comfortably into the Workpiece problem frame. But if information is also needed about
the progress of the work done using the tool, both the Workpiece problem frame and the JSD frame must be used. Two views must be taken of the one problem, two methods must be applied, and two solutions must be composed. And if the CASE tool must also impose method constraints on the Workers, the Domain-Effect frame must be applied too. The Workpieces, Requests, and Operations then constitute the Environment of this further frame. Even the classic patient-monitoring problem mentioned earlier has a complexity. It might be viewed as a JSD problem (with the addition of the analog devices as a Connection interposed between the Real World and the System). But the need to detect and report failure of the analog devices suggests that they must be treated as the Real World in yet another problem frame.

These complexities, of course, are purely relative to the available repertoire of problem frames and associated methods. Recognising and resolving such complexities is decomposition of problems into sub-problems. But it is decomposition guided by a clear idea of what constitutes a sub-problem: it is a problem for which an appropriate problem frame and an understood method are known.

Traditionally, notions of decomposition have stood as if they were self-sufficient. But they are not. If a problem has been decomposed into sub-problems, then the resulting sub-solutions must be re-composed into one solution. Having divided to conquer, we must re-unite to rule. This consequence of decomposition is less important when solutions are decomposed within a single computational paradigm. Procedure call allows decomposition into a hierarchy of unlimited depth and width; message-passing allows decomposition into an unlimited number of object classes and instances. The subsequent composition is entirely straightforward, within the paradigm.

Decomposition according to more specific problem frames and methods poses a significant composition problem. The same domain—for example, the Workpieces—appears as two different principal parts in two different problem frames, where the associated methods may exploit or rely on different characteristics. There is then a potential difficulty in implementation, to give one domain within the machine an implementation that conforms to both sets of characteristics. This kind of composition exploits what traditional engineers call The Shanley Principle whose relevance to software development was pointed out by de Marneffe [3]:

“... If you make a cross-section of, for instance, the German
V-2 [rocket], you find external skin, structural rods, tank wall, etc. If you cut across the Saturn-B moon rocket, you find only an external skin which is at the same time a structural component and the tank wall. Rocketry engineers have used the “Shanley Principle” thoroughly when they use the fuel pressure inside the tank to improve the rigidity of the external skin!

8 Specialisation

This is the essence of specialisation. The specialist concentrates on problems of a relatively small class, specialising in the techniques for understanding and solving them, including those for composing the solutions to sub-problems. This concentration allows a body of specialised knowledge to be built up, and the appropriate ‘theoretical foundations’ and ‘practical disciplines’ to be brought to bear in a familiar and very specific context.

This is what traditional engineers have done, and is the origin of the established branches of engineering. The aspiration to place the whole of software development alongside the established branches as one more branch of engineering is misconceived. Our aspiration should be rather to develop specialised branches of software engineering, each merit its own place alongside the specialised established branches.

There are already some examples of such specialisation in software. Compiler writing is one notable example. Difficulties specific to the problem area, such as context-sensitive grammars, are recognised and classified. Description transformations, such as the derivation of a LALR parsing table from a grammar, are part of a standard repertoire. The broad structures of different functions of a compiler, such as lexical and syntactic analysis, code generation, peephole and global optimisation, are studied, and ways are devised of composing them into efficient products.

There are pressures to specialise in other software fields. The production of shrink-wrapped software for personal computers is especially subject to strong pressures of this kind. Magazines review competing products, and so lead to a kind of cooperation among producers as they study each other’s products and try to emulate the good features offered by their competitors. Users add a strong pressure towards standardisation, both by demanding interoperability and by insisting on the kind of ‘look and feel’ similarity that
makes it possible for a competent car driver to get into almost any car and drive it immediately.

But in many fields, where each software project is unique, these pressures are not felt. It is an urgent challenge, if we aspire to the status of engineers, to identify, study and embrace the specialisations into which our discipline must be divided. Only specialisation can advance software engineering from an amateur to a professional activity.

9 Acknowledgements

Several people gave me many helpful comments on an earlier version of this paper. I would like to thank them all—John Dobson, Daniel Jackson, Ralph Johnson, and Wlad Turski.

References


[12] Peter Naur and Brian Randell, eds; Software Engineering: Report on a conference sponsored by the NATO Science Committee, Garmisch, Germany, 7th to 11th October 1968; NATO Brussels, 1969.


01/07/94
M A Jackson
101 Hamilton Terrace
London NW8 9QX
mj@doc.ic.ac.uk
jackson@attmail
+44 71 286 1814 (voice)
+44 71 266 2645 (fax)

file: sejrn102.tex