Formal Methods to Aid the Evolution of Software

M. P. Ward* K. H. Bennett*

January 17, 2003

Abstract

There is a vast collection of operational software systems which are vitally important to their users, yet are becoming increasingly difficult to maintain, enhance and keep up to date with rapidly changing requirements. For many of these so called *legacy systems* the option of throwing the system away an re-writing it from scratch is not economically viable. Methods are therefore urgently required which enable these systems to evolve in a controlled manner. The approach described in this paper uses formal proven program transformations, which preserve or refine the semantics of a program while changing its form. These transformations are applied to restructure ans simplify the legacy systems and to extract higher-level representations. By using an appropriate sequence of transformations, the extracted representation is guaranteed to be equivalent to the code. The method is based on a formal wide spectrum language, called WSL, with accompanying formal method. Over the last ten years we have developed a large catalogue of proven transformations, together with mechanically verifiable applicability conditions. These have been applied to many software development, reverse engineering and maintenance problems. In this paper, we focus on the results of using this approach in the reverse engineering of medium scale, industrial software, written mostly in languages such as assembler and JOVIAL. Results from both benchmark algorithms and heavily modified, geriatric software are summarised. We conclude that formal methods have an important practical role in software evolution.

1 Introduction

Legacy software may be defined informally as "software we don't know what to do with, but it's still performing a useful job". The implication is that the preferred solution is to discard the software completely, and start again with a new system. This may not be appropriate in all cases, for example:

(i) The software represents years of accumulated experience, which is unrepresented elsewhere, so discarding the software will also discard this knowledge, however inconveniently it is represented;

(ii) The manual system which was replaced by the software no longer exists, so systems analysis must be undertaken on the software itself;

(iii) The software may actually work well, and its behaviour may be well understood. A new replacement system may perform much more badly, at least in the early days. Hence it may be worth recovering some of the good features of the legacy system;

(iv) A typical large legacy software system has many users, who typically have exploited undocumented "features" and side effects in the software. It may not be acceptable to demand that users undertake a substantial rewrite for no discernable benefit. Therefore, it may be important to retain the interfaces and exact functionality of the legacy code, both explicit and implicit;

(v) Users may prefer an evolutionary rather than a revolutionary approach.

*Department of Computer Science University of Durham, Durham, UK

The aims of the work described in this paper are:

(i) To help the expert maintainer in understanding a large legacy system;

(ii) To assist in representing that understanding, using a carefully and formally defined language;

(iii) To automate as far as possible mundane and mechanical tasks, leaving the maintainer to focus on the strategic steps;

(iv) To ensure that the ultimate representation represents the semantics of the source code exactly.

Our ultimate objective is to recover a formal requirements specification for a legacy system, given only the source code (written in a typical third or second generation language).

Given a source program $\mathbf{P}$, the approach is firstly to translate this into an equivalent form $\mathbf{P}'$ expressed in a language WSL, in which all subsequent operations are performed. This wide spectrum language is a key part of our work, as it must facilitate the representation of both low level imperative constructs (e.g. goto's, aliased memory) and also non executable specifications e.g. in first order logic. All transformations are expressed in terms of WSL.

Thus the user starts with $\mathbf{P}'$, and selects a transformation from a library of pre-proven transformations. This is applied, resulting in an intermediate representation $\mathbf{S}_1$. Subsequent transformations may be applied to transform the software into $\mathbf{S}_2$, $\mathbf{S}_3$, ..., $\mathbf{S}_i$. The end point may depend on the need; for example, the user may wish only to perform simple restructuring, or they may need to extract an abstract specification.

Such an approach clearly lends itself to tool support, though the central work of designing the wide spectrum language and proving the transformations has to be done beforehand.

The main focus of this paper is the results of using this approach. Section 2 provides a very brief summary of the theory, though this is available in full detail elsewhere. Section 3 describes the main architectural details of a tool based on the transformation approach.

Section 4 summarises the results of applying this approach to benchmark programs (such as Schorr-Waite graph marking and topological sorting) which both pose particular challenges, because they exhibit a combination of complex control flow and complex data structures, and use data structures for multiple purposes.

Section 5 presents results from using the tool and method on industrial scale software. The problem of maintaining correctness despite using a front end source code to WSL translator is also discussed.

Section 6 presents the conclusions from the results to date.

2 Theoretical Foundation

This project originated not in software maintenance, but in theoretical research, developing a language in which proofs of equivalence for program transformations could be achieved as easily as possible for a wide range of constructs.

WSL is the "Wide Spectrum Language" used in our program transformation work, which includes low-level programming constructs and high-level abstract specifications within a single language. By working within a single formal language we are able to prove that a program correctly implements a specification, or that a specification correctly captures the behaviour of a program, by means of formal transformations in the language. We don't have to develop transformations between the "programming" and "specification" languages. An added advantage is that different parts of the same program can be expressed at different levels of abstraction, if required.

A *program transformation* is an operation which modifies a program into a different form which has the same external behaviour (it is equivalent under a precisely defined denotational semantics). Since both programs and specifications are part of the same language, transformations can be used to demonstrate that a given program is a correct implementation of a given specification.

A *refinement* is an operation which modifies a program to make its behaviour more defined and/or more deterministic. Typically, the author of a specification will allow some latitude to the implementor, by restricting the initial states for which the specification is defined, or by defining a nondeterministic behaviour (for example, the program is specified to calculate a root of an equation, but is allowed to choose which of several roots it returns). In this case, a typical implementation will be a *refinement* of the specification rather than a strict equivalence. The opposite of refinement is *abstraction*: we say that a specification is an abstraction of a program which implements it. See [20,22] and [2] for a description of refinement.

The syntax and semantics of WSL are described in [28,33,36,39] so will not be discussed in detail here. Most of the constructs in WSL, for example **if** statements, **while** loops, procedures and functions, are common to many programming languages. However there are some features relating to the "specification level" of the language which are unusual. Expressions and conditions (formulae) in WSL are taken directly from first order logic: in fact, an infinitary first order logic is used (see [13] for details), which allows countably infinite disjunctions and conjunctions, but this is not essential for understanding this paper. This means that statements in WSL can include existential and universal quantification over infinite sets, and similar (non-executable) operations.

An example of a non-executable operation is the nondeterministic assignment statement (or specification statement) $\langle x_1, \ldots, x_n \rangle := \langle x'_1, \ldots, x'_n \rangle.\mathbf{Q}$ which assigns new values to the variables $x_1, \ldots, x_n$. In the formula $\mathbf{Q}$, x_i represent the old values and x'_i represent the new values. The new values are chosen so that $\mathbf{Q}$ will be true, and then they are assigned to the variables. If there are several sets of values which satisfy $\mathbf{Q}$ then one set is chosen nondeterministically. If there are no values which satisfy $\mathbf{Q}$ then the statement does not terminate. For example, the assignment $\langle x \rangle := \langle x' \rangle.(x = 2.x')$ halves x if it is even and aborts if x is odd. If the sequence contains one variable then the sequence brackets may be omitted, for example: $x := x'.(x = 2.x')$. The assignment $x := x'.(y = 0)$ assigns an arbitrary value to x if $y = 0$ initially, and aborts if $y \neq 0$ initially: it does not change the value of y.

Another example is the statement $x := x'.(x' \in B)$ which picks an arbitrary element of the set B and assigns it to x (without changing B). The statement aborts if B is empty, while if B is a singleton set, then there is only one possible final value for x.

Program transformations may be used to refine a specification to an executable implementation (forward engineering) or to abstract a formal specification from program source code (reverse engineering).

As in a number of related projects (e.g. [3,4,6,26]), we have used the approach of defining a kernel language with denotational semantics, which in our case consists of only four primitive statements and three compound statements. Unlike other work, a purely applicative kernel is not used; the concept of state is introduced within the kernel, using a specification statement which also allows specifications expressed in first order logic as part of the language, thus providing a genuine wide spectrum language. More powerful constructs are defined by definitional transformations i.e. in terms ultimately of the kernel constructs. The use of state as an integral part of the kernel allows the direct modelling of existing imperative programs in the wide spectrum language, thereby simplifying the process of software evolution.

The main novel theoretical contribution lies in the use of infinitary logic in both the kernel language and proof meta language to widen the scope of the transformations it is possible to prove. In particular, it has been possible to develop general purpose transformations for loops and for recursive procedures which can be applied without needing loop invariants.

Details of the theoretical foundations of our work are given in [36]. Examples of the use of the transformation based approach for forward engineering are given in [34,40], and for reverse engineering in [35,39,45]. A survey of work on transformation systems may be found in [42] and also in [27]. The approach may be contrasted with the refinement calculus (e.g. see [19,20,21]) in which the user selects the next refinement step, and in doing so will generate a set of proof obligations, i.e.

theorems which must be proved for the refinement to be valid. Despite the increasing availability of automatic theorem provers, the proof for a large program is still a major activity. For example Sennett [31] describes the implementation of a function which gives rise to over 100 theorems needing proof.

3 The Tool Architecture

The initial prototype of the tool was developed as part of an Alvey project [24] at the University of Durham [41]. This work on applying program transformation theory to software maintenance formed the basis for a joint research project (the ReForm project) between the University of Durham, CSM Ltd and IBM UK Ltd. whose aim was to develop a tool which would interactively transform assembly code into high-level language code and **Z** specifications. The prototype has since been completely redeveloped into an industrial-strength tool which is capable of dealing with medium sized (up to 20,000 lines) modules of source code. Translators have been developed for IBM 370 Assembler and JOVIAL. A COBOL translator is currently under development, and this will enable to tool to assist with migration from Assembler or JOVIAL to COBOL II. Translators for C and ADA are also planned for the future.

A practical system for software evolution has to deal with real programs, not laboratory or toy examples. More specifically, the following requirements were identified:

- The tool must cope with the usual programming constructs and their uses (and abuses) including Gotos, global variables, aliasing, recursion, pointers, side effects etc.;
- It is not acceptable to assume that the code has been developed or maintained using structured methods. Real code must be acceptable, and major restructuring may be required before proper reverse engineering can start. This should be carried out automatically (or semi automatically) by the system.
- Transformations in the library must be proven correct, so that the user can employ them with confidence, but also so that the user does not have to undertake such proofs. The transformations need applicability conditions, and these must be mechanically checked by the tool. In this way, all responsibility for correctness lies with the tool—there are no generated "proof obligations" which the user must discharge before correctness can be guaranteed;
- It must be possible to select a sub-component of a large existing system and to guarantee to preserve the interactions of the sub-component with the rest of the system. This permits attention to maintenance 'hot spots', and also permits a piecemeal approach to evolution;
- The correctness of the implementation must be well established.

The main components of the tool are shown in Figure 1. The core of the tool is the library of proven transformations together with the transformation engine. The transformations in the library were proven before the tool was built. They allow a construct in WSL to be recast into another WSL construct while ensuring that the semantics are preserved. The software maintainer using the tool has only to select a transformation and apply it. He or she does not have to do the proof; the system's transformation engine checks that the transformation is applicable.

However, the first stage is to load the source code into the tool, and this is achieved by the source language to WSL translator as a batch job. The equivalent WSL code is stored internally as an abstract syntax tree (together with ancillary information to aid applicability checking). Further details are given in [7,43] and in Section 5.1.1.

The system is interactive and incorporates a graphical front end, pretty-printer and browser. This allows the programmer to move through the program, apply transformations, undo changes he or she has made, and in special circumstances, edit the program manually: but always in such a way that it is syntactically correct. The system automatically checks the applicability conditions of a transformation before it is applied, or even presented in one of the menus. This means that the

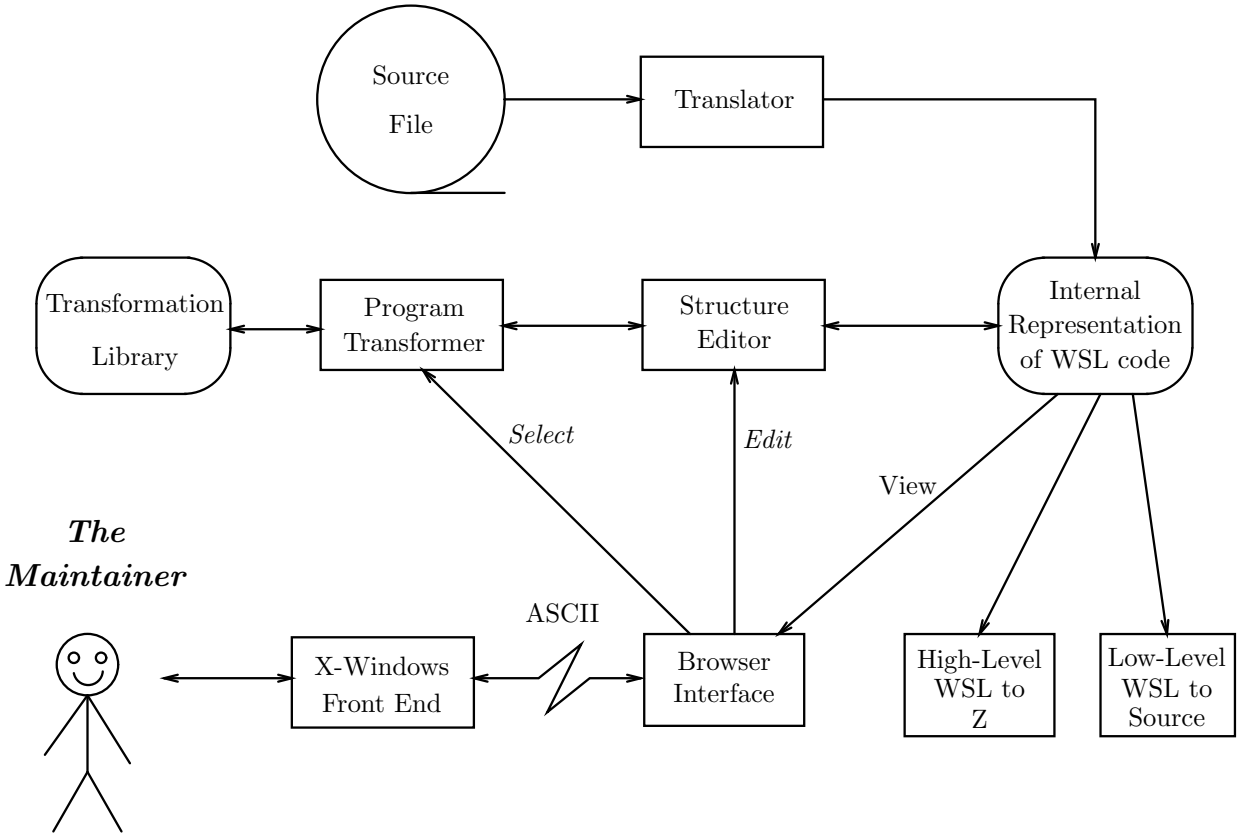

Figure 1: The Architecture of the tool

correctness of the resulting transformed program is guaranteed by the system rather than being dependent on the user. A history/future structure is built-in to allow back-tracking and forward-tracking enabling the programmer to change his or her mind. The system stores the results of its analysis of a program fragment as part of the program, so that re-calculation of the analysis is avoided wherever possible.

The interactive approach means that the programmer is always in control of how much or how little the program is changed. Unlike conventional restructuring tools, the programmer drives the process, and is not simply presented with a restructured program he or she still does not understand. The ultimate shape of the program has been determined by the user, not by some remote tool developer. If at any point the user is unhappy with the latest state of the program, he or she can always undo the last (sequence of) transformation(s).

Presenting the programmer with a variety of different but equivalent representations of the program can greatly aid the comprehension process, making best use of human problem solving abilities (visualisation, logical inference, kinetic reasoning etc).

Note that the theoretical foundation work which proves that each transformation in the system preserves the semantics of any applicable program is *essential* if this method is to be applied to practical software maintenance or reverse engineering. It must be possible to work with programs which are poorly (or not at all) understood, and it must be possible to apply many transformations which drastically change the structure of the program (as in the examples in Section 4) with a very high degree of confidence in the correctness of the result.

Finally, the tool is also capable of computing standard complexity metrics for a selected region of the WSL program, and presenting them in graphical form to show changes with time. Currently, McCabe, structural, size, control flow, data flow and branch-loop metrics may be computed [5].

It has been learned through experience that a user often employs a pattern of transformations, and it is easy within the tool to group such transformations into more powerful single transformations. Currently, such 'super transformations' are added by the tool builder, but as the transformations are represented internally in WSL it would not be difficult to allow the user to do this. The system is constructed as a hierarchy of abstract machines, each of which is formally specified. Additionally, much of the tool is written in either WSL or $\mathcal{META}$WSL, an extension of WSL used for representing transformations. This makes it possible for the developers to use the tool in the maintenance of its own source code.

3.1 Main features

The prototype has been developed using a "rapid prototyping" method, in order that new ideas can be implemented and tested quickly. Over the course of the development, the internal data structures and organisation have been changed several times as new research has shown better ways of doing things. One of the drawbacks of rapid prototyping is that the resulting tool can become unstable: i.e. bug-ridden, poorly-structured and difficult to maintain. To minimise this problem and maximise the flexibility of the tool we developed an "abstract machine" implementation, with formally defined interfaces between the implementation of the abstract machines and the rest of the system. The major components in the system are:

- Internal representation of WSL code;
- Structure editor;
- Transformation library;
- X Windows interface.

Each of these is implemented as an abstract machine with formally defined interfaces. This means that different people can work on reimplementing the different modules without causing integration problems. Another technique we have used is to develop a comprehensive regression test suite in parallel with the development of the system. This has acted as a "trip test": each new version has to pass the test before it is released, and this has prevented many of the problems which can occur with a rapid turnaround of program versions.

4 Results with Benchmark Programs

In this section we describe some of the results of applying our program transformation system to several small but challenging example programs. The examples selected pose particular challenges to the approach, because they exhibit a combination of complex control flow and complex data structures, and use data structures for multiple purposes.

4.1 A Report Generator

The first example is reverse engineering a simple program with a complex control flow. This was taken from a programming textbook [9], although it is a classic example of bad programming style! In [39] we translate the program into WSL and use program transformations to reverse-engineer it to an abstract specification. This gave confidence that the approach could be used on a program with a complex control structure, whose operation was not understood before starting the reverse engineering process.

4.2 The Schorr-Waite Graph Marking Algorithm

Our next example is a forward engineering example: the Schorr-Waite graph marking algorithm [30]. This has acquired the status of a standard testbed for program verification techniques applied to complex data structures. In [40] we present a complete derivation of the algorithm, starting with a mathematical specification of graph marking. We develop a simple recursive algorithm by

applying general purpose transformations, and then apply Schorr and Waite's "pointer switching" technique to develop an efficient iterative algorithm. The central idea behind the pointer switching technique is that when we return from having marked the left subtree of a node we know what the value of $l[x]$ is for the current node (since we just came from there). So while we are marking the left subtree, we can use the array element $l[x]$ to store something else—for instance a pointer to the node we will return to after having marked this node. Similarly, while we are marking the right subtree we can store this pointer in $r[x]$. In [40] we use program transformations to apply this technique to a number of different algorithms.

The transformation approach proved to be a powerful way to prove the correctness of these challenging algorithms. A correctness proof for the algorithm has to show that:

1. The original graph structure is preserved by the algorithm (although it is temporarily disrupted as the algorithm proceeds);
2. The algorithm achieves the correct result (all reachable nodes are marked).

Most published correctness proofs for the algorithm [10,11,17,23,29,32,44] have to treat these two problems together. The methods involving assertions (and intermittent assertions) require an understanding of the "total situation" at any point in the execution of the program. In contrast, our approach makes it possible to separate the two requirements, and thereby to apply the same technique to a number of different algorithms. We have derived several marking algorithms which make use of the pointer switching ides and which illustrate the advantages of transformational development using a wide spectrum language:

1. The development divides into four stages: (i) Recursive Algorithm; (ii) Apply the pointer switching idea; (iii) Recursion Removal; and (iv) Restructuring. Each stage uses general-purpose transformations with no complicated invariants or induction proofs;
2. The method easily scales up to larger programs: for example, the hybrid algorithm presented in [40] is much more complex than the simple algorithms, yet our development uses the same transformations and involves no new ideas or proofs.

4.3 Topological Sorting

Our third example is a reverse engineering problem. In [15], Knuth and Szwarcfiter present an algorithm for determining all the embeddings of a given partial order into a total order. The algorithm is highly unstructured with complex control flow combined with complex data structures. In [38] we restructure and analyse the algorithm using program transformations. The analysis of the algorithm breaks down into several stages, culminating in a formal specification of topological sorting.

1. Restructure to remove some of the control-flow complexity;
2. Recast as an iterative procedure, "abstracting away" the error cases so that recursion removal can be applied easily;
3. Restructure the resulting recursive procedure;
4. Add abstract variables to the program and update them in parallel with the actual (concrete) variables;
5. Replace references to concrete variables by equivalent references to abstract variables;
6. Remove the concrete variables to give an abstract program;
7. Show that the abstract program is a refinement of the specification of topological sorting.

4.4 Polynomials in Several Variables

The previous example exhibited a combination of control flow complexity with highly complex data structures. Our fourth example (also a reverse-engineering example) is a program with complex

data structures (trees implemented as four-way linked structures) and highly complex control flow. Algorithm 2.3.3.A from Knuth's "Fundamental Algorithms" [16] (P.357) is an algorithm for the addition of polynomials represented using four-directional links. In [14] Knuth describes this as having "a complicated structure with excessively unrestrained **goto** statements" and goes on to say "I hope someday to see the algorithm cleaned up without loss of its efficiency". In [37] we use program transformations to manipulate the program, using semantics-preserving operations, into an equivalent high-level specification.

4.5 Conclusions

These and other case studies have demonstrated that the transformational approach can be applied to both forward and reverse engineering problems. During this time, a method and strategy for reverse engineering using formal transformations has been developed and successfully tested. Certain features of the case studies indicate that the approach will scale up to industrial-scale software. This is addressed in the next section.

5 Results with Industrial-Scale Software

5.1 IBM 370 Assembler

Experiments have been undertaken on modules of Assembler taken from real application programs. The majority have been between 500 and 2,000 lines but some have been up to 20,000 lines. These experiments have shown that programs which have been transformed using the tool can be expressed in a form which subjectively is much easier to understand than the original. This applies particularly to real programs which have been modified over many years.

5.1.1 Modelling Assembler in WSL

Constructing a useful scientific model necessarily involves throwing away some information: in other words, to be useful a model must be inaccurate, or at least idealised, to a certain extent. For example "ideal gases", "incompressible fluids" and "billiard ball molecules" are all useful models which gain their utility by abstracting away some details of the real world. In the case of modelling a programming language, such as Assembler, it is theoretically possible to have a perfect model of the language which correctly captures the behaviour of all assembler programs. Certain features of Assembler, such as branching to register addresses, self-modifying code and so on, would imply that such a model would have to record the entire state of the machine, including all registers, memory, disk space, and external devices, and "interpret" this state as each instruction is executed. Unfortunately, such a model is useless for inverse engineering[1] purposes since such trivial changes as deleting a NOP instruction, or changing the load address of a module, can in theory change the behaviour of the program.

What we need is a practical model for assembler programs which is suitable for inverse engineering, and is wide enough to deal with all the programming constructs we are likely to encounter. Our approach involves three types of modelling:

1. Complete model: Each assembler instruction is translated into WSL statements which capture all the effects of the instruction. The machine registers and memory are modelled as arrays, and the condition code as a variable. Thus, at the translation stage we don't attempt to recognise "if statements" as such, we translate into statements which assign to cc (the condition code variable), and statements which test cc. The automatic restructuring and simplification state can usually remove all references to cc, presenting the maintainer with a structured program expressed in **if** statements, loops and actions;
2. Partial model: Branches to register are modelled by attempting to determine all possible targets of such a branch (including all labels and jump instructions which follow labelled

[1] We use the term "inverse engineering" to mean "reverse engineering through formal transformations.

instructions). Each label is turned into a separate action with an associated value (the relative address). A "store return address" instruction stores the *relative* address in the register. A "branch to register" instruction passes the relative address to a "dispatch" action which tests the value against the set of recorded values, and jumps to the appropriate label. This can deal with simple cases of address arithmetic (including jump tables) but may theoretically be defeated if more complex address manipulations are carried out before a branch to register instruction is executed;

3. Self-modifying code: This is not addressed, except for some special cases which are recognised by the translator. In many environments the code must be re-entrant, or is to be blown into a ROM, and therefore cannot be modified. In other cases, the self-modification may be recognised by the translator and may require human intervention to determine a suitable WSL equivalent.

One of the major drawbacks of automatic program restructurers [8] is that complex control structures are replaced by complex data flow structures involving additional flag and sentinel variables with meaningless names inserted by the tool. This does not occur with our tool, and users resolve the underlying structural problems because the transformations make it easy to do so. It is also straightforward to avoid dispersing code that previously was together. This method has the advantage that performance problems and errors which exist deeply buried in heavily modified code become much more easily observable.

The following table demonstrates some sample results for a moderately large (4,500 lines) and fairly complex Assembler module:

	No. of Statements	McCabe Cyclometric	Control Flow/ Data Flow	Branch– Loop	Structural
Translated WSL	4,785	1,299	5,270	1,481	37,168
Automatic Restructuring	4,779	1,299	5,262	1,478	37,134
Simplify some tests	4,775	1,298	5,150	1,476	36,246
Automatic Restructuring	3,372	592	3,711	924	23,140
Remove Dispatch calls:	3,454	577	3,761	970	23,178
Remove Dispatch action	2,920	338	3,101	631	17,616
Remove procedure linkage	2,568	338	2,687	631	16,208
Automatic Restructuring	2,554	334	2,669	631	16,094

A major attraction of the tool has turned out to be the transformations which convert in-line code to procedures, and global variables into parameters. This enables the user to convert a large, unstructured, monolithic piece of code into a main program which calls a set of single-entry single-exit procedures. These transformations alone can make a large difference to the understandability of the code, and prepare it for the recognition of abstract data types.

5.2 Herma Assembler

A more recent project has involved migration of a large (approx 1 million lines) system written in an obsolete 16 bit assembler, currently running under an emulator on a 32 bit microprocessor, to a native C implementation. This has involved writing a herma to WSL translator, applying general-purpose and specialised transformations to the generated WSL code, and translating the result to C using a WSL to C translator. One feature of the source language is that there are no conditional branch instructions. Instead, the "test" instructions set a *skip flag* in the processor, which causes the next instruction in sequence to be skipped and execution continues with the instruction after that (clearing the skip flag in the process). Our translator works on a line-by-line basis: translating each instruction as precisely as possible, while ignoring the surrounding instructions. This means that the WSL representation of each instruction has to test the skip flag in order to decide whether the instruction is executed or skips. This results in a somewhat inflated

McCabe metric for the "raw" WSL. The "Constant Propagation" transformation traces the control flow and eliminated redundant tests of the skip flag (where the flag is known to be clear). Further automated transformations can now be applied to produce a high-level structured program, ready for automatic translation to C. Our first case study consisted of a small (240 lines) Herma module which translated to about 650 lines of "raw" WSL. The largely automatic transformation process reduced this to about 150 lines of WSL which was translated to C. The project is still at an early stage, and it is anticipated that even more compact and efficient C code can be produced with over 90% automation of the migration task.

	No. of Statements	McCabe Cyclometric	Control Flow/ Data Flow	Branch– Loop	Structural
Raw WSL	701	180	746	175	5,654
Constant Propagation	473	54	457	116	3,849
Automatic Restructuring	202	12	184	3	1,847
Some hand transformation	196	11	183	11	1,855

Interactive Assistance with Program Understanding

The first stage in using the tool for program understanding is the translation process. This is an off-line process by which the program under maintenance is translated from the source language to WSL.

Because the WSL has to represent side effects of individual instructions which may or may not be significant to the operation of the program, a single Assembler statement may be translated to more than one WSL statement. Hence it is quite likely that the program's WSL representation will be two to three times as long as the original, since much redundancy will have been introduced. This may seem like a backwards step, but it has a real purpose, namely to ensure that the process of going from the original language to WSL is as simple as possible, with the minimum possibility of error.

Clearly, the translation step must be undertaken "correctly", or else the maintenance engineer will quickly lose confidence in the whole system. However, the problem is not simply one of writing a correct translator from (say) Assembler into WSL, there are two main problems:

1. The Assembler language may not be formally specified. In this case the translation will be from an informal language to a formal language, so "formal methods" cannot be used to prove that the translation is correct;
2. The behaviour of certain Assembler operations, such as self-modifying code, branches to calculated addresses and so on, depends on the state of the entire machine and the interpretation of the bit patterns in each memory location. A complete translation of the behaviour of an Assembler program would therefore have to model the entire state of the machine, including all memory, registers, flags etc., and "interpret" by calculating the next state at each clock cycle. This would be useless for reverse engineering purposes.

Our approach to these problems is to design the translator to be a simple as possible. Each Assembler instruction is translated separately into a sequence of WSL statements which capture all the effects of that particular instruction: including register assignments and condition code setting. The translated program consists of an action system, with a new action started at each of the locations which could be the target of a jump instruction. "Proving" the correctness of the translator now reduces to showing that each individual instruction type is translated correctly, by comparing the WSL against the published documentation for the machine, and the knowledge of experienced Assembler programmers. In fact, such a translator will *define* a formal semantics for the Assembler language.

Of course, such a translation will be grossly inefficient, containing many redundancies (for

example the condition code will be set at every instruction, whether it is needed or not). However, we are now within the formal system, and can make use of all the powerful restructuring and simplifying transformations in order to remove the redundancies with the confidence that the semantics will be preserved. So the next step, after translation, is to apply a whole series of simplification and restructuring transformations. This is carried out automatically, so by the time the maintainer gets to see the translated code it is already close to a high-level language program. The program is then in a state in which the maintainer can do meaningful work on it.

A typical next activity for the maintainer would be to try to establish the flow structure of the program. If the program is in assembler, the flow of control is determined by different kinds of branch instructions, whether they be branches from the bottom to the top of loops conditional upon the value of a register, or branches round sections of code following a test of some condition. In WSL, the branches and labels are represented by an Action System, and a powerful transformation called Collapse_Action_System can be used to remove these branches and labels and replace them with **if** ... **then** ... **else** ... **fi** constructs and **while** or **for** loops, thus revealing the flow in a much more understandable form.

Program Restructuring

When restructuring a program, the most difficult problem is trying to foresee all the side-effects of any changes made. If this is done manually, there is an enormous clerical task in tracing through all possible effects of changes to control flow or ordering of actions. There are tools which can assist with this, but even with such help, it is still a time-consuming and error-prone activity to verify that a restructured program will work as expected.

A fundamental attribute of the Maintainer's Assistant, however, is that its transformations are all mathematically proven to preserve the semantics of the subject program. The programmer can be confident that the program after transformation is functionally equivalent to its original form. Redundant code and variables can safely be removed, 'spaghetti' code can be straightened out, and the program simplified and its maintainability improved. The tool's metrics facility allows comparison of the new version of the program with the original, and can confirm the effectiveness (or otherwise) of the restructuring that has been done.

Specification Abstraction

One of the important features of WSL, from which it gets its name, is that it covers a wide spectrum, i.e. it can be used to represent both low-level operations and high-level specifications. There are specialised transformations which allow the operations in a program to be expressed as specifications. (In this case, these are not really transformations in the strict sense since the specification and the code are not equivalent—rather the code is a refinement of the specification) By this means, the specification of a program can be represented in WSL, and could then if desired be translated to another specification language, such as Z.

With the ability to use transformations to cross levels of abstraction and hence to keep specifications in line with code, the traditional arguments in favour of the "minimum fix" approach to maintenance (i.e. that it is the cheapest and safest) will no longer apply. As confidence and experience with the use of transformations grow, it is envisaged that maintenance will be carried out at the highest practicable level of abstraction, viz. the specification and high level design. Research in this area is currently underway at Durham University.

5.3 JOVIAL

More recently we have constructed JOVIAL to WSL and WSL to JOVIAL translators, and have used the tool for restructuring a number of JOVIAL source modules, ranging up to around 5,000 lines. The two examples in this section were selected for case studies because, despite their fairly

moderate size, they contained some very complicated code which made them difficult to analyse and maintain using traditional methods. The aim was to restructure the programs for ease of maintainability.

5.3.1 Module A

The first example consisted of 861 lines of JOVIAL in a main routine and eight procedures. The module was translated from JOVIAL to WSL, restructured and simplified using the tool, and then translated back to JOVIAL for testing. The following table shows the improvements achieved in terms of various complexity metrics:

	No. of Statements	McCabe Cyclometric	Control Flow/ Data Flow	Branch– Loop	Structural
Before	954	120	845	701	10,371
After	392	92	343	146	5,115

These improvements were achieved by the use of a wide range of transformations, with the choice of transformations was guided by the aim of reducing the complexity as measured by the above metrics. In some cases the immediate application of a powerful transformation such as Collapse_Action_System achieved the desired results; in others, some localised restructuring using some of the more specialised transformations was necessary first. The whole process, including the generation of metrics and call graphs, took about half a day for an experienced user of the tool.

Activity	Step	No. of Statements	McCabe Cyclometric	Control Flow/ Data Flow	Branch– Loop	Structural
Initial	(1)	954	120	845	701	10371
Restructure main routine	(2)	815	120	728	584	9091
Simplify A.S.	(3)	795	116	716	572	8941
Merge calls	(3.1)	791	116	712	568	8901
Collapse A.S (Undone)	(3.2)	1005	188	876	538	10820
Remove recursion	(4)	913	142	838	537	9869
Restructure action body	(5)	812	122	731	539	9023
Collapse A.S.	(6)	783	106	729	537	8963
Create Procedure	(7)	756	98	699	513	8673
Collapse A.S GETTY	(8)	732	98	669	507	8404
Collapse A.S. ProcI	(9)	692	98	627	464	7994
Remove loop	(10)	686	98	624	468	7959
Collapse ProcG, ProcF	(11)	677	98	615	459	7874
Collapse ProcE	(12)	616	98	552	384	7254
Remove loop	(13)	612	98	552	396	7239
Collapse ProcD and simplify	(14)	555	98	497	341	6682
Restructure ProcC	(15)	436	97	383	227	5534
Collapse ProcC	(16)	431	95	374	177	5445
Collapse ProcB and simplify	(17)	392	92	343	146	5115

5.3.2 Module B

The second JOVIAL example consisted of 2,564 lines of JOVIAL. The main sources of complexity in this module were:

1. Heavy use of labels and **goto**'s rather than structured programming practices, which make the control flow difficult to understand; and
2. Multiple exit paths from the program, including exits via calls to closed compound procedures which never return.

The main routine in the module contained the most complex code, but in addition, the procedure ProcI contained some particularly complex code.

A significant amount of simplification and restructuring of the raw WSL is performed automatically as part of the translation process from Jovial to WSL. One very significant area of restructuring is the conversion of closed compound procedures in the original Jovial into pure procedures in the WSL version of the program. Where it is possible for a closed compound procedure to perform a jump to a label which is outside its body, this behaviour is modelled by the setting of a variable in the WSL program which is tested when the corresponding procedure returns and a jump performed to the appropriate label. Thus, procedures always return to the point from which they were called, and any jumps to labels are made explicit. A further advantage is that closed compound procedures now appear in the procedure call graph of a module.

The following table shows the effect of the restructuring process:

	No. of Statements	McCabe Cyclometric	Control Flow/ Data Flow	Branch– Loop	Structural
Before	2,518	847	2,212	1,109	22,986
After	2,222	780	2,138	863	20,832

6 Conclusions

For simple restructuring, skills are needed to identify a simplification strategy and then to select transformations to achieve this goal. However, for acquiring the specification of an existing program, the user also needs to be an expert in software engineering and in the application domain. This confirms the original design objective of providing assistance to the expert maintainer, rather than de-skilling or automating the maintenance task.

We originally thought that all users would want to go from code to specifications. In fact there is a spectrum of requirements, ranging from using the tool for code comprehension and simple restructuring, to reverse engineering from code to a high level of abstraction. Our approach deals with the whole spectrum of requirements.

6.1 A Method for Reverse Engineering

One of the major results from our research, which the availability of a prototype tool has helped to produce, is the development of a method for reverse engineering using formal transformations. The method is based on the following stages:

1. Establish the reverse engineering environment. This will involve a CASE tool to record results, maintain different versions of code, specifications, and documentation and the links between them; together with a WSL code browser and transformation system.
2. Collect the software to be reverse engineered. This involved finding the current versions of each subsystem and making these available to the CASE tool.
3. Produce a high-level description of the system. This may already be available in the documentation, since the documentation at this level rarely needs to be changed, and is therefore more likely to be up to date. The documentation is supplemented by the results of a cross reference analysis which records the control flow and data dependencies among the subsystems.
4. Translate the source code into WSL. This will usually be an automatic process involving parsing the source files and translating the language structures into equivalent WSL structures.
5. "Inverse Engineering", i.e. reverse engineering through formal transformations. It involves the automatic and manual application of various transformations to restructure the system

and express it at increasingly higher levels of abstraction. This is carried out by iterating over the following four steps:

(a) Restructuring transformations. These include removing **goto** statements, eliminating flags, removing redundant tests, and other optimisations. The effect of this restructuring is to reveal the "true" structure of the program which may be obscured by poor design or subsequent patching and enhancements. This stage is more radical than can be achieved by existing automatic restructuring systems [8,18] since it takes note of both data flow and control flow, and includes both syntactic and semantic transformations [1]. We have however had considerable success with automating the simpler restructuring transformations, by implementing heuristics elicited from experienced program transformation users.

(b) Analyse the resulting structures in order to determine suitable higher-level data representations and control structures.

(c) Redocument: record the discoveries made so far and any other useful information about the code and its data structures.

(d) Implement the higher-level data representations and control structures using suitable transformations. A powerful technique we have developed for carrying out these data refinements is to introduce the abstract variables into the program as "ghost" variables (variables whose values are changed, but which do not affect the operation of the program in any way), together with invariants which make explicit the relationship between abstract and concrete variables. Then, one by one, the references to concrete variables are replaced by references to the new abstract variables. Finally, the concrete variables become "ghost" variables and can be removed. See [39] for an example of this process; it is also used extensively in [40]. In general, if our analysis in step 5b is correct then the result of this stage is likely to be in a form suitable for further restructuring.

6. Acceptance test: We now have a high-level specification of the whole system which should go through the usual Q.A. and acceptance tests.

6.2 Why Invent WSL?

For restructuring purposes it is useful to work within a language which has the following features:

- Simple, regular, and formally defined semantics;
- Simple, clear, and unambiguous syntax;
- A wide range of transformations with simple, mechanically-checkable correctness conditions.

No existing programming language which is widely in use today meets *any* of these criteria.

For reverse engineering it is extremely useful to work within a single wide-spectrum language within which both low-level programs and high-level abstract specifications are easily represented.

For migration between programming languages it is important that the transformation system language should not be biased towards a particular source or target language.

These are the considerations which led to the development of the WSL language. The language has been developed gradually over the last ten years, in parallel with the development of the transformation theory. This parallel development has ensured that WSL is ideally suited for program transformation work: the design of the language ensures that developing and proving the correctness of transformations is straightforward and, most importantly, the correctness conditions for the transformations are easy to check mechanically. This last point was important for the success of our transformation system.

We believe that the formal foundations of our language and transformation theory were essential to the success of the project. The practice of implementing any reasonable-looking transformation *without* a formal proof of correctness is very dangerous: the author has discovered errors in trans-

formations published in reputable journals [1], but the errors were only uncovered after having attempted (and failed) to prove that the transformations were correct. Since our tool works by applying a vast number of transformations in sequence, any unreliability in the transformations will have serious repercussions on the reliability of the tool. In practice, the work on proving the correctness of known transformations has been a major driving force in the discovery of new transformations.

An obvious disadvantage of working in a separate language to the source language of the legacy system is that translators to and from WSL will have to be written. Fortunately, for the "old fashioned" languages typical of legacy systems, this is not much more difficult than writing a parser for the language, which in turn is a simple application of well-developed compiler technology for which there is a wide variety of tool support available. In addition, there are three important advantages to our approach:

1. Using a collection of translators for different languages, it becomes possible to migrate from one language to another via WSL. We are currently working on an Assembler to COBOL II migrations: the aim is to produce "high level" COBOL II, not something which looks as though it was written by an Assembler programmer!
2. The second is that the "translator" can be very simple-minded and not have to worry about introducing redundancies, dead code, unstructured code etc. Once we are within the formal language and transformation system, such redundancies and infelicities can be eliminated automatically by applying a series of general-purpose restructuring, simplification and data-flow analysis transformations.
3. Thirdly, our ten year's work on transformation theory can be re-applied to a new language simply by writing a translator for that language. It would be impossible to re-use the development work for a COBOL transformation system in the development of a JOVIAL transformation system. Even a different version of COBOL could invalidate many transformations and involve a lot of re-work.

Based on our results, translation to a formal language is the best way to set about any serious reverse engineering or migration work.

Our work has been criticised by some practitioners for its emphasis on the use of formal methods and formally specified languages. This is odd because the programming language and its support libraries form the basic building materials for software engineering. But no serious engineer would expect to build with components whose properties are not precisely, formally, concisely specified (eg. this beam is specified to be able to take this much load under these operating conditions, etc.) No serious engineer would tolerate "standard" components which differ in an undefined way in their properties and behaviour from supplier to supplier. A serious engineer does not think twice about screwing a nut from one supplier onto a bolt from another supplier: he expects them to fit as a matter of course! A serious engineer expects to have to master a certain amount of mathematics in order to do his or her job properly: differential equations, integration, fluid dynamics, stress modelling, etc. This is far more than the elementary set theory and logic required to understand WSL.

With regard to the "undefined" behaviour of many commercial languages in the presence of syntactic or semantic errors (out of bounds subscripts etc) Hoare [12] said:

> In any respectable branch of engineering, failure to observe such elementary precautions would have long been against the law

This was way back in 1960.

D.L.Parnas at the International Conference on Software Engineering in Baltimore, Maryland in 1993 [25] made the following points on the relationship between software engineers and "real" engineering:

- "Engineering" is defined as "The use of science and technology to build useful artifacts";
- Classical engineers use mathematics to describe their products (calculus, PDEs, nonlinear functions, etc.);
- Computer systems designers should use engineering methods if they are to deserve the name "Software Engineers". This will include the use of mathematics.

6.3 Contributing Factors

We believe that the following main features have contributed to the success of our approach:

- Use of weakest preconditions expressed in infinitary logic;
- Starting with a small, tractable kernel language, extended via definitional transformations;
- Use of an imperative kernel language, with functional constructs added via definitional transformation, rather than a functional kernel language;
- Developing the transformation theory in parallel with the language development;
- Dealing with assembler via simple translation followed by automatic restructuring and simplification;
- Developing an interactive, semi-automatic tool, rather than attempting complete automation;
- Mechanical checking of the correctness conditions at each step, with only valid transformations appearing in the menus;
- Knowledge elicitation: using the prototype and manual case studies to see how the experienced user solves problem, and then implementing these methods and heuristics;
- The use of generic transformations for merging, moving, separating etc.; these are automatically expanded into the appropriate transformation for each situation;
- Rapid prototyping development, with the system organised as a collection of abstract machines with formally defined interfaces;
- Separation of front-end issues into a separate program.

Acknowledgements

The research described in this paper has been partly funded by Alvey project SE-088, partly through a DTI/SERC and IBM UK Ltd. funded IEATP grant "From Assembler to Z using Formal Transformations" and partly by SERC (The Science and Engineering Research Council) project "A Proof Theory for Program Refinement and Equivalence: Extensions".

References

[1] J. Arsac, "Syntactic Source to Source Program Transformations and Program Manipulation," *Comm. ACM* 22 (Jan., 1982), 43–54.

[2] R. J. R. Back, *Correctness Preserving Program Refinements*, Mathematical Centre Tracts #131, Mathematisch Centrum, Amsterdam, 1980.

[3] F. L. Bauer, B. Moller, H. Partsch & P. Pepper, "Formal Construction by Transformation—Computer Aided Intuition Guided Programming," *IEEE Trans. Software Eng.* 15 (Feb., 1989).

[4] F. L. Bauer & H. Wossner, *Algorithmic Language and Program Development*, Springer-Verlag, New York–Heidelberg–Berlin, 1982.

[5] K. H. Bennett, H. Yang & T. Bull, "A Transformation System for Maintenance—Turning Theory into Practice," *Conference on Software Maintenance, Orlando, Florida* (1992).

[6] R. Bird, "Lectures on Constructive Functional Programming," in *Constructive Methods in Computing Science*, M. Broy, ed., NATO ASI Series #F55, Springer-Verlag, New York–Heidelberg–Berlin, 1989, 155–218.

[7] T. Bull, "An Introduction to the WSL Program Transformer," *Conference on Software Maintenance 26th–29th November 1990, San Diego* (Nov., 1990).

[8] F. W. Calliss, "Problems With Automatic Restructurerers," *SIGPLAN Notices* 23 (Mar., 1988), 13–21.

[9] M. Fenton, *Developing in DataFlex, Book 2, Reports and other outputs*, B.E.M. Microsystems, 1986.

[10] D. Gries, "The Schorr-Waite Graph Marking Algorithm," *Acta Inform.* 11 (1979), 223–232.

[11] M. Griffiths, *Development of the Schorr-Waite Algorithm*, Lect. Notes in Comp. Sci. #69, Springer-Verlag, New York–Heidelberg–Berlin, 1979.

[12] C. A. R. Hoare, "The Emperor's Old Clothes: The 1980 ACM Turing Award Lecture," *Comm. ACM* 24 (Feb., 1981), 75–83.

[13] C. R. Karp, *Languages with Expressions of Infinite Length*, North-Holland, Amsterdam, 1964.

[14] D. E. Knuth, "Structured Programming with the GOTO Statement," *Comput. Surveys* 6 (1974), 261–301.

[15] D. E. Knuth & J. L. Szwarcfiter, "A Structured Program to Generate All Topological Sorting Arrangements," *Inform. Process. Lett.* 2 (1974), 153–157.

[16] D. K. Knuth, *Fundamental Algorithms*, The Art of Computer Programming #1, Addison Wesley, Reading, MA, 1968.

[17] R. Kowalski, "Algorithm = Logic + Control," *Comm. ACM* 22 (July, 1979), 424–436.

[18] J. C. Miller & B. M. Strauss, "Implications of Automatic Restructuring of COBOL," *SIGPLAN Notices* 22 (June, 1987), 76–82.

[19] C. C. Morgan, "The Specification Statement," *Trans. Programming Lang. and Syst.* 10 (1988), 403–419.

[20] C. C. Morgan, *Programming from Specifications*, Prentice-Hall, Englewood Cliffs, NJ, 1994, Second Edition.

[21] C. C. Morgan & K. Robinson, "Specification Statements and Refinements," *IBM J. Res. Develop.* 31 (1987).

[22] C. C. Morgan, K. Robinson & Paul Gardiner, "On the Refinement Calculus," Oxford University, Technical Monograph PRG-70, Oct., 1988.

[23] J. H. Morris, "A Proof of the Schorr-Waite Algorithm," in *Theoretical Foundations of Programming Methodology* Int. Summer School, Marktoberdorf 1981, M. Broy & G. Schmidt, eds., Dordrecht: Reidel, 1982.

[24] B. Oakley & K. Owen, *Alvey: Britain's Strategic Computing Initiative*, MIT Press, Cambridge, MA, 1989.

[25] D. L. Parnas, Presentation at the International Conference on Software Engineering, Baltimore, 21st–23rd May 1993.

[26] H. Partsch, "The CIP Transformation System," in *Program Transformation and Programming Environments* Report on a Workshop directed by F. L. Bauer and H. Remus, P. Pepper, ed., Springer-Verlag, New York–Heidelberg–Berlin, 1984, 305–323.

[27] H. A. Partsch, *Specification and Transformation of Programs*, Springer-Verlag, New York–Heidelberg–Berlin, 1990.

[28] H. A. Priestley & M. Ward, "A Multipurpose Backtracking Algorithm," *J. Symb. Comput.* 18 (1994), 1–40, ⟨http://www.dur.ac.uk/~dcs0mpw/martin/papers/backtr-t.ps.gz⟩.

[29] W. P. de Roever, "On Backtracking and Greatest Fixpoints," in *Formal Description of Programming Constructs*, E. J. Neuhold, ed., North-Holland, Amsterdam, 1978, 621–636.

[30] H. Schorr & W. M. Waite, "An Efficient Machine-Independant Procedure for Garbage Collection in Various List Structures," *Comm. ACM* (Aug., 1967).

[31] C. T. Sennett, "Using Refinement to Convince: Lessons Learned from a Case Study," *Refinement Workshop, 8th–11th January, Hursley Park, Winchester* (Jan., 1990).

[32] R. W. Topor, "The Correctness of the Schorr-Waite List Marking Algorithm," *Acta Inform.* 11 (1979), 211–221.

[33] M. Ward, "Proving Program Refinements and Transformations," Oxford University, DPhil Thesis, 1989.

[34] M. Ward, "Derivation of a Sorting Algorithm," Durham University, Technical Report, 1990, ⟨http://www.dur.ac.uk/~dcs0mpw/martin/papers/sorting-t.ps.gz⟩.

[35] M. Ward, "A Recursion Removal Theorem," Springer-Verlag, Proceedings of the 5th Refinement Workshop, London, 8th–11th January, New York–Heidelberg–Berlin, 1992, ⟨http://www.dur.ac.uk/~dcs0mpw/martin/papers/ref-ws-5.ps.gz⟩.

[36] M. Ward, "Foundations for a Practical Theory of Program Refinement and Transformation," Durham University, Technical Report, 1994, ⟨http://www.dur.ac.uk/~dcs0mpw/martin/papers/foundation2-t.ps.gz⟩.

[37] M. Ward, "Reverse Engineering through Formal Transformation Knuths "Polynomial Addition" Algorithm," *Comput. J.* 37 (1994), 795–813, ⟨http://www.dur.ac.uk/~dcs0mpw/martin/papers/poly-t.ps.gz⟩.

[38] M. Ward, "Program Analysis by Formal Transformation," *Comput. J.* 39 (1996), ⟨http://www.dur.ac.uk/~dcs0mpw/martin/papers/topsort-t.ps.gz⟩.

[39] M. Ward, "Abstracting a Specification from Code," *J. Software Maintenance: Research and Practice* 5 (June, 1993), 101–122, ⟨http://www.dur.ac.uk/~dcs0mpw/martin/papers/prog-spec.ps.gz⟩.

[40] M. Ward, "Derivation of Data Intensive Algorithms by Formal Transformation," *IEEE Trans. Software Eng.* 22 (Sept., 1996), 665–686, ⟨http://www.dur.ac.uk/~dcs0mpw/martin/papers/sw-alg.ps.gz⟩.

[41] M. Ward, F. W. Calliss & M. Munro, "The Maintainer's Assistant," *Conference on Software Maintenance 16th–19th October 1989, Miami Florida* (1989), ⟨http://www.dur.ac.uk/~dcs0mpw/martin/papers/MA-89.ps.gz⟩.

[42] H. Yang, "How does the Maintainer's Assistant Start?," Durham University, Technical Report, 1989.

[43] H. Yang, "The Supporting Environment for a Reverse Engineering System—The Maintainer's Assistant," *Conference on Software Maintenance, Sorrento, Italy* (Dec., 1991).

[44] L. Yelowitz & A. G. Duncan, "Abstractions, Instantiations and Proofs of Marking Algorithms," *SIGPLAN Notices* 12 (1977), 13–21, ⟨http://www.dur.ac.uk/~dcs0mpw/martin/papers/prog-spec.ps.gz⟩.

[45] E. J. Younger & M. Ward, "Inverse Engineering a simple Real Time program," *J. Software Maintenance: Research and Practice* 6 (1993), 197–234, ⟨http://www.dur.ac.uk/~dcs0mpw/martin/papers/eddy-t.ps.gz⟩.