Knowledge engineering - science into practice: putting knowledge where it counts

B. Marshall, G.D. Lyon, A.C. Newton & J.W. McNicol

In the SCRI annual report of 1993, we wrote about the opportunities opened up by flexible modelling, which addressed knowledge and uncertainty. A point then, which is equally important today, is the ability to reveal the wealth of strategic knowledge previously hidden in the 'black boxes' of crop models constructed in procedural languages such as FORTRAN and PASCAL. There are two types of knowledge, domain knowledge or facts, and strategic knowledge or know-how. It is in this second aspect that flexible modelling has its greatest impact, enabling the experts' know-how to be passed on to the client in forms that are understandable and context specific. Strategic knowledge is knowing how best to use the facts to make informed judgements either in short term tactical decisions or in longer term planning. Eight years on, and we have reached the market place with two products: mapp™, the Management Advisory Package for Potatoes for immediate decision making, and PCN, a longer-term planning package for the control of potato cyst nematode.

More than 5M tonnes of raw potato are produced each year in the UK. The future of the UK potato industry depends on quality. The average value of the raw product is around £40 per tonne but this can be easily doubled or trebled by achieving the right size grades of quality potatoes for baker and pre-pack markets and can even be as high as a ten-fold increase in the specialist markets of salad potatoes. Failure to achieve the right size of tuber alone is estimated to cost the industry £24M per annum. mapp™ helps to match seed rates to expected yield and market, tracks crop development from planting to burn down; manages weather data; helps decide herbicide application; enables an effective irrigation strategy; helps to limit common scab; predicts yields and changes in tuber size distribution; shows the relation between burn-down dates and profits; uses the grower's management data and sample digs to predict values that are unique to the specific crop; covers over a hundred commercial cultivars; applies to all market outlets; and provides context-sensitive help (Fig. 1).

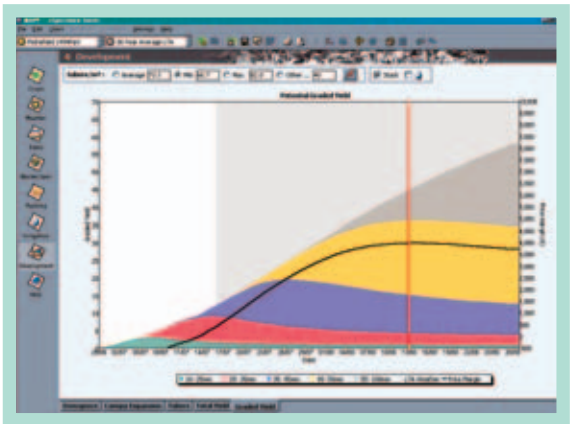

Figure 1 mapp™, the Management Advisory Package for Potatoes, predicting the development of graded yields in a seed crop with time (pale blue, <25 mm; pink, 25-35 mm; blue, 35-45 mm; yellow, 45-55 mm; grey, > 55 mm).

This user-friendly software package has made available to the industry nearly two decades of research into the physiology of potato growth, development and yield. This has been possible by combining the skills of crop and environmental physiologists, mathematical modellers and staff of the Department of Artificial Intelligence and the Artificial Intelligence Applications Institute at the University of Edinburgh with the experience of the potato industry itself. Windows-based programming provides the user-friendly interface that the customer has come to expect from modern software, and the framework for integrating contrasting programming techniques. Databases store the weather data, both current and long term average for predicting the future of the crop, as well as the husbandry data and, most importantly, the information gathered by direct observation of crop performance. Models, tried and tested in the field throughout the UK, of the evolution of tuber size distribution and water-limited yield are implemented in Visual C++. Then CLIPS, a rule-based system, provides the 'intelligence' or know-how to answer a given question and only prompt the user for missing information that is essential to answering that query. Visual Basic provides the Windows-based framework.

A multi-disciplinary team is essential and its diversity presents challenges for both coordination of and com-

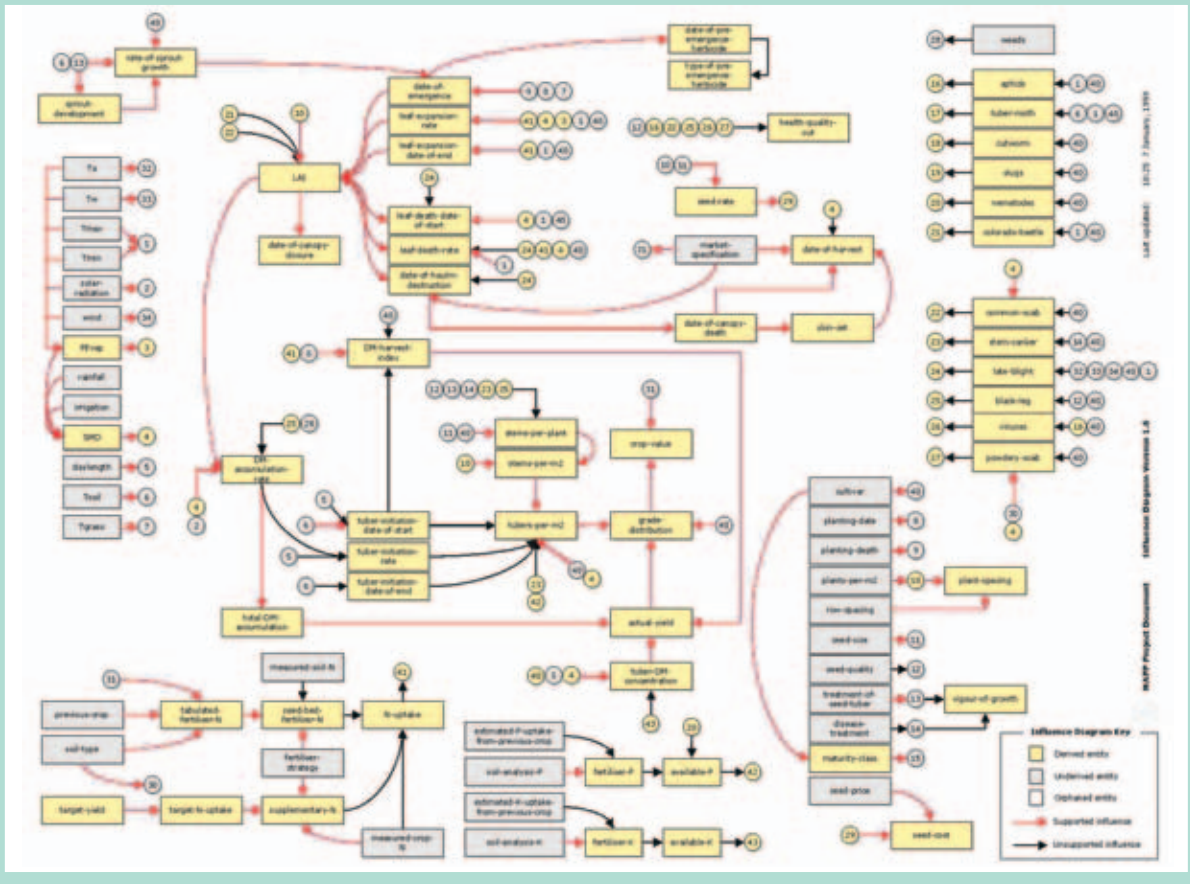

Figure 2 The Active Influence Diagram for the mapp™ project - the pivotal tool for coordination, communication and design - showing the inter-dependencies of key crop characters, weather and husbandry (red arrows indicate those influences currently addressed and black arrows are for future consideration). For clarity, not all arrows are shown in their entirety and circles with the same number indicate that they share a common connection. Clicking on any character (box) invokes a query to an underlying database that reveals the immediate and all secondary influences on the character.

munication within such a project. The creation of an Active Influence Diagram became pivotal (Fig. 2). It provides the discipline for defining the project, the ability to communicate ideas accurately and efficiently between team members, eliminates jargon and documents the structural specification. Using software developed in-house, the diagram automatically generates a database of inter-dependencies between the variables along with the explanations of each variable. The CLIPS rule-base is then constructed directly from this database. Design is an iterative process; its visualisation and automation through the active influence diagram was a major benefit to the project.

The PCN model (Fig. 3) is an example of another expert system being developed at SCRI to aid in the longer term management of the white potato cyst nematode (PCN) which is reaching epidemic proportions in which approximately 60% (and increasing) of

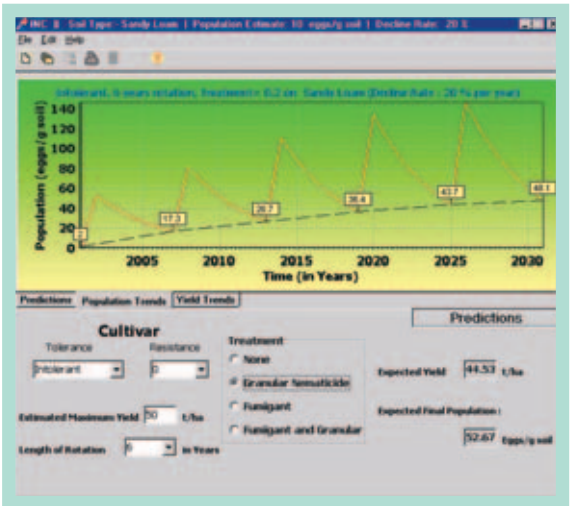

Figure 3 PCN expert system: predicting population trends in white potato cyst nematode and impact on yield for a user-specified management strategy.

potato land is infested and is currently costing the industry an estimated £50M per annum. The system shows that this situation will continue unless there is more effective integration of existing control measures.

Rotations have shortened with the introduction of potato varieties resistant to the previous nematode pest, the golden PCN, in the late 1960s. It takes approximately 30 years, typically five potato crop rotations, for very small populations to increase to damaging proportions. 94% of potatoes grown in the UK are susceptible to the white potato cyst nematode. There are no fully resistant potato varieties commercially available and those that do have a degree of (partial) resistance have limited market outlets. Granular nematicides appear to be less effective at controlling the white PCN than the golden PCN species. The growing population of white PCN can go undetected until it is too late. This new computer-based model assists the grower to predict the effect of the management strategy that they intend to implement and thus reduces the risk. It also highlights the need for good sampling strategies and then makes effective use of the data collected. Although less complex than mapp™, the development of a user-friendly expert system for PCN management has also benefited from the development of a Windows-based programming environment.

Reflecting back to the article published in the 1993 Annual Report, the incorporation of the effects of uncertainty into models for risk management has been slower to reach the market place than anticipated. Uncertainty exists both in the formulation of a model as well as in its inputs. Mathematical models are frequently constructed using some form of 'best fit' relation with estimates of the associated parameters. The fact that the observed data does not lie precisely on this relation, rather it is randomly distributed about the relation, is often ignored. Ideally, this should be captured in the model by, for example, including both the best estimates of the parameters' values and a measure of their uncertainty. The uncertainty of future weather also plays an important part in forecasting the performance of a crop. The availability of local weather records is often limited both in the number of years and detail. Accordingly, weather generators have been developed to reproduce the mean and variability of the key weather characters. They are calibrated for the specific location and rapidly generate a set of variable seasons representative of the locale. They are undergoing further refinement; capturing the correlations between the variables and paying special attention to the frequency of wet days and the distribution of rainfall amounts.

The incorporation of uncertainty has been slower because software for constructing models that capture this uncertainty or probability had to be written in-house, from first principles. Now there is both commercial and free software available for the construction of probability models, e.g. Bayesian Belief networks or 'Causal models', that do not require specialist programming skills to use, and their utility has been enhanced. Originally, probabilistic relations between a set of variables could only be represented by a set of joint probabilities. This meant that all variables had to be converted into a set of qualitative ranges, e.g. 'very low', 'low' … 'very high', before joint probabilities could be assigned. As the number of discrete ranges per variable increased, so the requirement for computer memory increased exponentially. Modern software has removed this constraint. Relations between variables no longer have to be made discrete and the belief in the value of the associated parameters can take any form of continuous probability density function. In a recent piece of research, using this modern software, we have produced a probabilistic model of the factors influencing

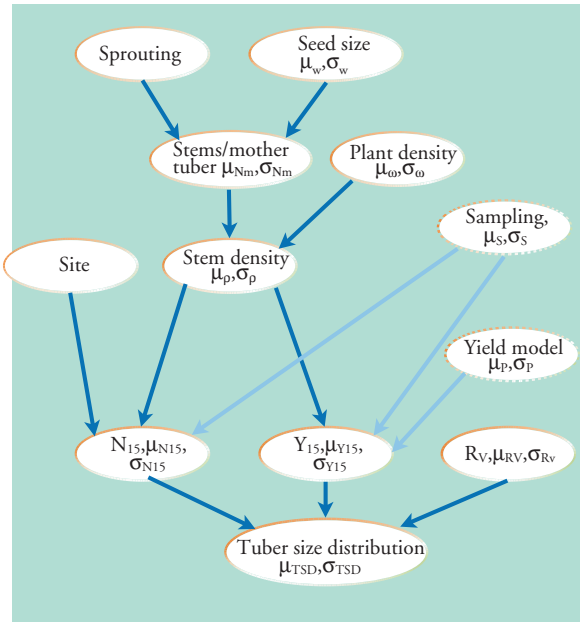

Figure 4 Simplified causal graph of factors affecting tuber size distribution. The parameters μ and σ indicate that the variables have both a mean and variability. Likewise, the parameters that define the relations between any two nodes (ellipses) have both means and variances associated with them, which together capture both the best-fit relation and the variability about them.

tuber size distribution (Fig. 4). It requires further testing before considering release into the market place. Nevertheless, it can already provide valuable quantitative insights into where the greatest causes of uncertainty in its predictions lie.

The demand for knowledge engineering, including probabilistic modeling, is growing rapidly in many biological research areas. For example, in pathology where in the last few years many new genes have being identified, from many different host-pathogen interactions, revealing both common processes and unique pathways. Any model designed to provide an integrated understanding of pathogen recognition and defence response must include spatial and temporal aspects to the expression of these genes. The recent development of Expressed Sequence Tag (EST) technology, together with enrichment processes such as Suppression Subtraction Hybridisation (SSH), means that the discovery of genes induced in response to treatment such as pathogen recognition is outpacing our ability to carry out research to understand their function. In order to build such an understanding, many isolated host-pathogen recognition models have been published; most of these became outdated quickly by new information and proved too inflexible to update. Furthermore, many new gene discoveries simply can not be integrated with these models and require new speculative relationships to be devised. These speculations are being used to drive new experimentation, so it is essential that they are able to incorporate all the most recent data. Currently there is no such integrative mechanism. We have drawn together data from our gene discovery programmes in a spatial diagrammatic representation of pathogen recognition in potato (Fig. 5). Gaps in our understanding of plant-pathogen interactions have been filled from the published literature on other systems, including animal pathogens and other informed speculation. However, the diagram has become complex, cannot be systematically queried, and has no temporal dimension. Furthermore, it is difficult to distinguish between the sources of information. Nevertheless, it continues to be the focus of informed discussions, which have generated many testable hypotheses.

Another important potential application is the prediction and tracing of contaminants, both in ecological

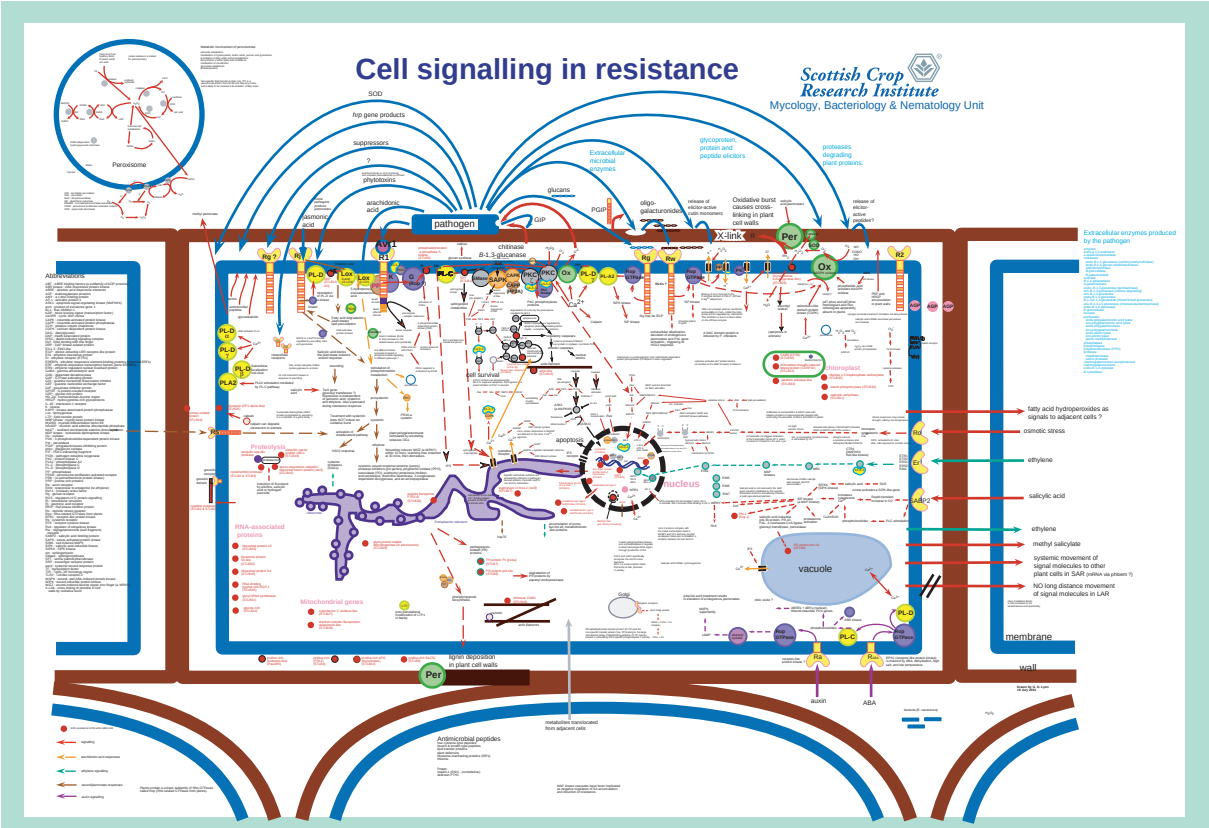

Figure 5 A diagrammatic representation of some known and postulated relations between molecules involved in signalling events associated with the response of potato to infection by a plant pathogen.

systems and in food. For instance, SCRI's MAFF-funded work with the Central Science Laboratory on modelling oilseed rape feral populations could be the forerunner of software of equal value to ecologists, agronomists and food standards inspectors. Models based on a partial understanding of a system, such as the host-pathogen response, are used to promote debate about the nature of the system. Informally represented models, e.g. where we rely on a subjective understanding of a diagram, are limiting in the forms of analysis that we can apply to them, and in the extent of information to which they can reliably connect. We need automated techniques, which stimulate exploration and hypothesis formation while adding the precision and objectivity necessary to integrate knowledge from diverse and rapidly expanding knowledge sources. These are some of the challenges facing knowledge engineering and biologists working together.

Developing these software applications requires the skills of many people. In particular we would like to acknowledge the inputs of M Elliot, DKL MacKerron, MS Phillips and DL Trudgill at SCRI, and R Rae , D Robertson and J Tonberg of the University of Edinburgh.