

INFORMATION INTEGRATION TECHNOLOGY: KNOWLEDGE REPRESENTATION AND BAYESIAN STATISTICS FOR PREDICTIVE MODELING*

Deborah Leishman
Statistical Sciences Group
Los Alamos National Laboratory D-1, New Mexico, USA

Laura McNamara
Sandia National Laboratory, New Mexico, USA

Abstract

Complex multidisciplinary projects often lack integrated representations to support a diverse community's problem-solving process. In this paper, we discuss an interdisciplinary approach to knowledge elicitation, representation and transformation developed in the Statistical Sciences group at the Los Alamos National Laboratory. This approach is called Information Integration Technology (IIT), and it combines techniques from Anthropology, Knowledge Representation, and Bayesian statistics to address the complexities of multidisciplinary research. In particular, as shown through a specific example, the underlying methodology used in IIT allows for qualitative representations of complex engineered systems to be translated into Bayesian models to support quantification of system reliability. We use elicitation techniques to elicit qualitative problem-solving structures from scientists and engineers collaborating on difficult R&D problems. The elicited information, in turn, is used to develop ontologies that represent the problem space in a common language for the research team. Iterative cycles of representational refinement, dependency definition and quantification lead to the emergence of predictive statistical models that make intuitive sense to all parties: engineers, elicitation experts, knowledge modelers and statisticians, and are used for quantification of system reliability.

INTRODUCTION

Statisticians are often asked to provide predictive risk and reliability assessments for a wide range of research and development projects. When these projects concern very complex engineered systems such as missiles or airplanes however, statisticians can find themselves challenged in building predictive models capable of integrating multiple types of data, information and knowledge from a wide range of sources.

Statisticians who work in experimental science and engineering fields become quite adept at consulting with research teams to develop a wide range of probabilistic models for decision-making. Traditionally, statisticians have worked fairly bounded pieces of a larger problem: experimental design, for example, or failure mode analysis. This trajectory has resulted in a standard model for statistical consulting in which the clients provide the statistical consultant with a problem definition and some data sources that, in the statistician's mind, lend themselves to a particular class of models. The statistician works an area of the problem, periodically asking clients to clarify some aspect of the model or to provide additional data.

* Distribution Statement A: Approved for public release; distribution is unlimited

The past twenty years or so, however, have seen a trend towards large-scale, complex, multidisciplinary scientific projects that often incorporate experts from a wide range of disciplines, including engineering, biology, physics, computer science, chemistry, and others. The complexity of these problems often requires a greater level of participation from the statistician and a need for more complex modeling

Bayesian models are widely used to combine multiple sources of data to estimate the probability of an event in the future, based on relevant information regarding the occurrence of that event in the past. Although Bayesian models are well suited to addressing complex problems, constructing a Bayesian model requires a great deal of time and information about the problem at hand. The IIT approach was designed to address this problem by using qualitative knowledge representation models such as Conceptual Graphs to represent the complex problem space. Within IIT, these qualitative knowledge models can then be transformed into Bayesian models which are able to specify and quantify aspects such as system reliability. Important within the transformation of the qualitative models to quantitative models is the need to understand the dependencies between parts of the physical system being modeled.

In this paper, we discuss an interdisciplinary approach that combines qualitative and quantitative modeling techniques in an effort to deal with complex problems. This approach, called Information Integration Technology (IIT), was developed in the Statistical Sciences group at Los Alamos National Laboratory to address the complexities of multidisciplinary research. In the following pages, we describe the origins and structure of the IIT approach and demonstrate its use in the development of a hierarchical reliability model for a complex rocket system. The IIT knowledge modeling techniques are of particular interest to Bayesian statisticians, whose problem solving approach often relies on complex hierarchical networks.

INFORMATION INTEGRATION TECHNOLOGY

The diagram shown below in Figure 1 outlines the IIT framework, which we use to derive qualitative knowledge models of a domain of interest, and transform these knowledge models into quantitative mathematical models such as Bayesian networks. The framework specifies the context in which these models are being formulated: for example, a decision-making environment in which they will be used to predict the reliability or performance of a system.

IIT methods and the IIT framework are designed to support the emergence of a comprehensive, quantitative decision support model through developing a set of knowledge representations that serve as a common denominator for all problem owners. In a complex system reliability problem, "problem owners" may include engineers, program managers and sponsors, computer scientists, physicists, technicians, and other experts contributing to the problem. IIT requires the ongoing involvement of a knowledge modeler, who works iteratively among the problem owners, technical experts and statisticians to model the system, its dependencies and probability distributions. The resulting qualitative graphical models provide a comprehensive, representation of the problem space. These representations are arranged hierarchically in interlinked levels of abstraction, the highest of which provides problem owners with an overview of the entire problem space. The hierarchy of specification enables project participants to drill more deeply into the problem while maintaining a consistent logical structure throughout all levels of problem representation. It is important to note that the models developed here are typically different from those that would be produced by the problem owners within normal development of a complex system such as an airplane or rocket. There is often not one overarching model of the system for example and if there is, it is not sufficient to understand all of the complex interactions and dependencies within the system. Such a model is needed however for development of a Bayesian network model of the system.

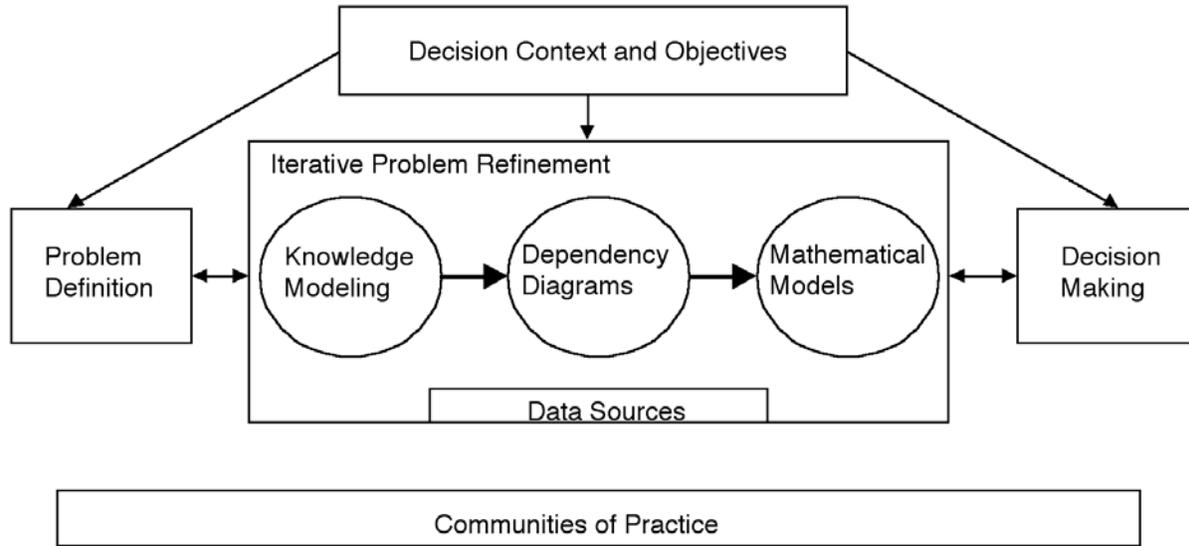


Figure 1. The Information Integration Technology Framework.

The first stage in the IIT method is elicitation of the foundation elements: identifying the communities of practice and/or stakeholders involved in the problem, defining the problem space and the decisions that are to be made by all stakeholders, and documenting the relationship between the stakeholders' objectives and their decisions. Once the problem space is defined, the knowledge modeler begins to work with experts to elicit the conceptual structures they use to work the problem. Using this elicited information, the knowledge modeler develops graphical representations of the problem space using those elements. The knowledge representation used to date in the IIT method is derived from conceptual graph techniques pioneered by John Sowa (1984). As these qualitative representations emerge, the knowledge modeler works iteratively with the problem owners, experts and statisticians to formulate the dependencies between concepts in the knowledge model. Once finalized with the experts, the knowledge modeler and the statistician begin transforming these qualitative representations into quantitative mathematical models.

The resulting mathematical framework is an extremely useful structure capable of combining multiple types of quantitative information to support decision-making in a traceable manner. Doing so requires identifying appropriate data sources to populate nodes in the model, transforming these data into joint probability distributions, and propagating these distributions and their associated uncertainties through the model.

KNOWLEDGE MODELING AND CONCEPTUAL GRAPHS

The conceptual graph model proposed by John Sowa (1984) is a method of representing the mental models that people use to understand the world. This approach combines a mapping to and from natural language with a mapping to logic. A conceptual graph, which consists of concepts and relations connected by arcs, asserts a proposition and takes the form of a finite connected bipartite graph. *Concepts* represent any entity, attribute, action, state or event that can be described in natural language. *Relations* detail the roles that each concept plays, and the *arcs* serve as connectors between the two. These graphs can be written in either a graphical representation or in a linear form to conserve space.

SIMPLE GRAPHS

This section presents parts of the conceptual graph model that form a central core. This includes concepts, relations and the arcs between them. Central to the model is the ability to map the graphs into first order predicate calculus. An example of a simple graph is:

[Cat: #123] -> (State) -> [Sit] -> (Location) -> [Mat]

Which represents "A cat named 123 is sitting on a mat."

CONCEPTS AND RELATIONS

Concepts represent the entities, attributes, actions, states or events found in natural language. In conceptual graph notation, they are shown as square boxes. A concept box has a referent field on the right of the colon. In this way both generic concepts and particular individuals can be referred to. For example, [Person: *] or [Person] both refer to the generic concept contains two fields separated by a colon and shows a concept type on the left of the colon and of Person, while [Person: #123] or [Person: Sam] refer to particular individuals, one named Sam and one named 123. Every generic concept in the graph terminology is existentially quantified. Generic concepts act like variables in logic, while individuals are like constants in logic. Relations in the conceptual graph model specify the role a concept plays and define the relationship between concepts. Relations are shown as circles in the graph notation and can have any number of arcs. For example (Past) is a monadic relation with one arc, (Agent) is a dyadic relation with two arcs and (Between) is a triadic relation requiring three arcs.

A LOGICAL MAP

The conceptual graph model defines the operator \square which maps simple conceptual graphs into formulas in the first order predicate calculus. For these simple graphs, the only logical operators which are needed are conjunction and the existential quantifier. For example, the conceptual graph:

[Cat: #123] -> (State) -> [Sit] -> (Location) -> [Mat]

maps into the following formula when the \square operator is applied:

$\exists x \exists y (Cat(\#123) \wedge State(\#123,x) \wedge Sit(x) \wedge Location(x,y) \wedge Mat(y)).$

Conceptual graphs are usually more concise than logical formulas because arcs on the graphs show the connections more directly than variable symbols. These graphs are often used for systems that perform reasoning, but for our purposes we use them as the qualitative models within IIT.

AN EXAMPLE PROBLEM: THE ROCKET DEVELOPMENT PROGRAM

To illustrate the application of the methods we have developed, we use an example from a research and development program that gathers data on test rockets to analyze their performance during flight and to make modifications to their design as necessary. Throughout this discussion, all engineers and agencies are aliased to maintain controls over Proprietary and sensitive information about the program. We refer to this program as the RDP, or the Rocket Development Program.

The oversight agency for the RDP is a group of engineers located in the southeastern United States; we refer to them as RDPC, or the Rocket Development Program Center. Two other groups of engineers are responsible for building separate sections of the rocket: one group of engineers is building a booster to send the rocket into the upper atmosphere, while the other group designs a test payload for the rocket to carry. In addition, several other sub-contractors and vendors provide parts

and support to each of the two primary engineering agencies. RDPC is primarily responsible for project management, cost controls, and scheduling.

The RDPC program managers came to Los Alamos with a specific problem: how does one develop a predictive reliability model for an engineering system that is still in the design stages? Multiple concerns drove this question: the rocket development program is extremely expensive. Only one or two of the prototypes is built and flown and is usually destroyed in the process; rarely are the engineers able to salvage subsystems for reuse in further iterations of the program. Because each system flown is unique, there is little direct, performance, or reliability data available for parts or subsystems on the test rocket. Hence the program managers had little idea how to make predictions or assess risk areas for the flights.

The goal of the LANL/RDP collaboration was to develop an integrated, full-system, predictive reliability model for an upcoming rocket flight. In developing the model, Los Alamos developed a model framework that captured the critical interactions among the rocket's subsystems during flight. We also elicited and documented the many sources of data and information that the engineers used to build confidence in their rocket before flight. The resulting model combines multiple sources of information in a rigorous, quantitative framework that can be used to identify and weigh potential risk areas to overall mission "success."

QUALITATIVE MODEL DEVELOPMENT

ENGINEERING REPRESENTATIONS AND GOAL DEFINITION

The contracting engineers in charge of developing the rocket are prolific creators of representations: mechanical drawings, electrical layout diagrams, interface control documents, reliability block diagrams, viewgraphs for debating design issues. Not surprisingly, many of these engineers expressed doubt about the utility of creating even more diagrams of their systems. However, while their representations were sufficient for building a test rocket, they were not sufficient for creating a statistical reliability model. As anthropologist Etienne Wegner (1998) has observed, problem solving is a process of devising representations of knowledge around which parties negotiate meaning. Like many engineering communities, the two primary contractors in the RDPC project each assign bounded teams of engineers to work on separate subsystems of the rocket. Engineering representations are used to communicate design requirements across team boundaries. Each iteration results in new, updated representations that capture the current state of knowledge about each of the subsystems required for a functioning rocket. However, at no point in the engineering problem-solving process does the community develop an integrated representation of the rocket's many subsystems as they are intended to work during flight.

To develop a reliability model as a Bayes net, however, the statistician must understand relationships among different elements of the rocket as it works during flight. This is where knowledge modeling becomes a critical step in creating an integrated model, one that captures subtle dependencies among interrelated parts and uses those dependencies to predict states for the overall mission.

The first step in the IIT knowledge modeling process was to meet with the RDP project leaders to identify specific goals for the rocket system, to get an overview of how the rocket would function, to find out which contractors were responsible for the major areas of the project, and to determine the metrics that the RDP project leaders would use to assess the project's outcomes. At the same time, we devised a general set of goals for the statistical model: to support the rocket project by identifying risk areas, and to provide a quantifiable, traceable statement of risk to upper-level managers in RDPC.

TOP LEVEL ONTOLOGY

With an understanding of the goals of the project, the next step within IIT was development of a formal ontology to represent the primary concepts of knowledge in the problem space, and to

understand the network of relationships among those concepts. We first worked with the team to get a very general idea of the system to be modeled and also elicited definitions of success and failure for the RDPC program managers. We borrowed a common aerospace terminology for describing mission outcomes: a “stoplight chart,” which is perhaps more accurately described as a continuum of failure-to-success, represented by red, yellow and green panels. Equally important in this stage was eliciting how the booster and payload builders defined success and failure, so that we could understand how their goals interlocked with RDPC’s goals. We used the same stoplight continuum in elicitation sessions with our experts at each agency. All “stoplight” charts were ultimately combined into a single chart, with all mission goals and states for mission outcomes clearly mapped. In addition, we worked with RDPC to elicit metrics that would determine each of the states for mission success and failure, while eliciting metrics for subsystem performance from the experts at each contracting agency. This information provided the statisticians with a means of quantifying a range of potential outcomes for each of the subsystems in the rocket, and a way to quantify overall mission success and failure.

Working further with the team and their documentation as well as the success failure continuums leads to the development of a first-order ontology, one that mapped at the most basic level the key concepts for the domain “RDP-2 Rocket” and the relationships among those concepts. In the ontology shown in Figure 2, we use a Conceptual Graph representation with concepts as rectangular nodes and relations as circular nodes and arcs indicating directionality among concepts and the relationships that tie them together. This representation is also recognizable to Bayesians statisticians, who use directed acyclic graphs as structures for propagating uncertainties.

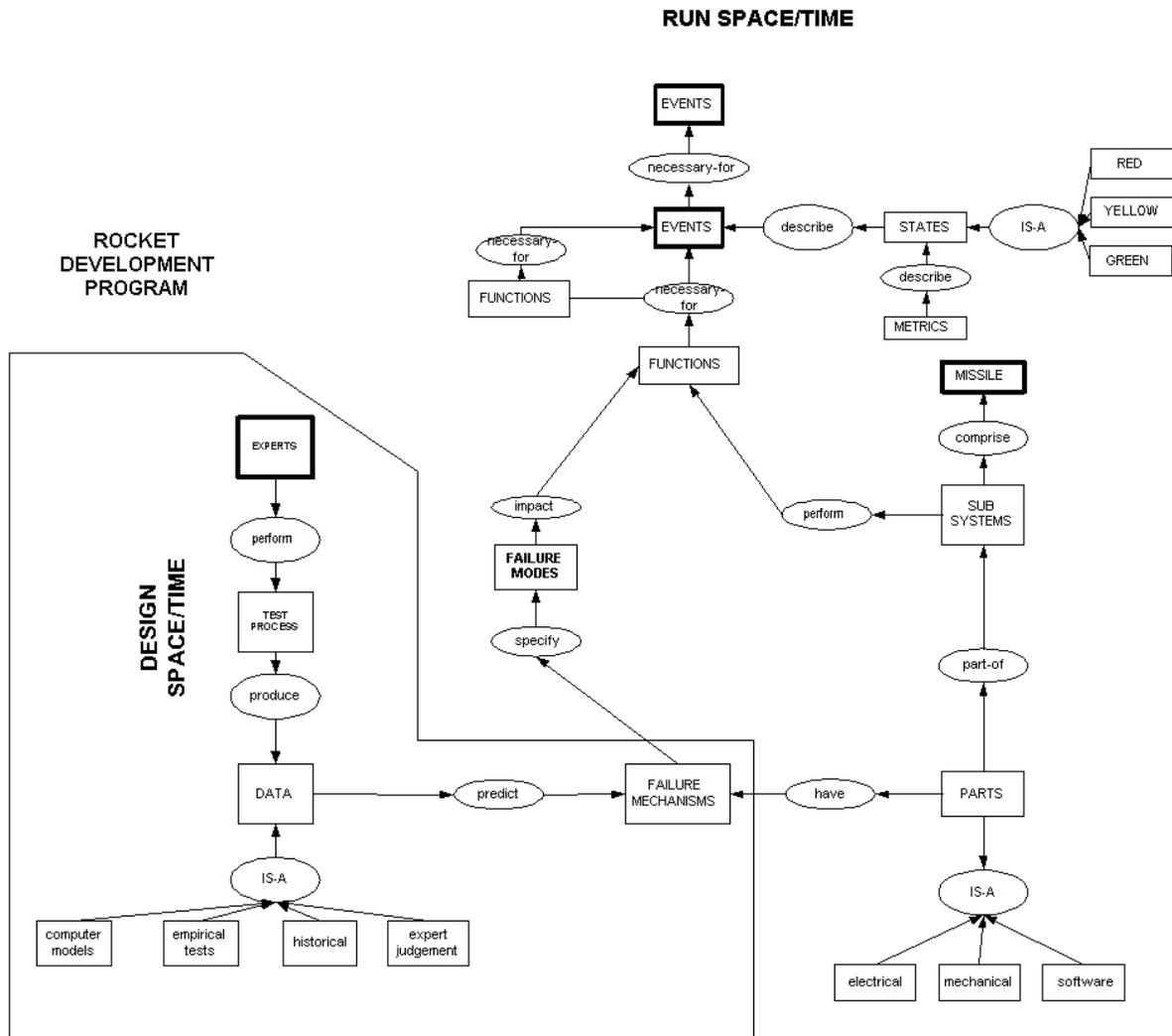


Figure 2. Ontology for RDP-2 rocket model.

Note that the ontology differentiates between two stages in the design process: “design time,” when the engineers are working to plan and build the rocket; and “run time,” which represents the actual functioning of the rocket during flight. Essentially, the knowledge modeler partnered with the engineers in the “design time” area of the ontology to create a statistical model that would be used to predict the reliability and performance of the rocket system during “run time,” the actual flight. Information generated during the design process in the “design time” area of the ontology was used to create a model structure and to gather data to populate the model.

The top-level ontology is a significant point in the IIT method, for it is an elicitation tool that provides a guide for specifying further levels of the domain. In the rocket project, the ontology defines the overall problem structure and in particular defines the relationship between system functions, key events and their relationship to mission success and failure states. Further detailing of the model to lower levels of abstraction will entail asking questions such as: What functions were required in order for a particular event to occur? What parts were required for that function to occur? How could failures in individual parts contribute to failed events?

During the elicitation process, the ontology also guides the development of a hierarchy of representations for the problem, from the most general and abstract representation (the top level

ontology) to the most specific representations (dependency diagrams that detail specific relationships among parts, subsystems and functions). One critical outcome for the representations is traceability from level to level, so that the representations flow in an orderly fashion from the ontology and make intuitive sense to all parties: the knowledge modelers, the statisticians, RDPC, and the builders of the booster and the payload.

EVENT SPECIFICATION

Once the top level ontology was completed we were ready to begin developing detailed representations of its concepts. The first level of specification focused on identifying measurable flight-time events that would act as conceptual waypoints, to discretize the linear flow of the planned rocket trajectory into a series of measurable focus areas for the model. Significantly, the order in this representation of flight events was not a time-ordered linear sequence, but rather a sequence of dependencies as shown in Figure 3. In other words, this level specified the order in which any particular event during the flight could impact, or be impacted, by any other event. Using the success and failure chart in combination with the event specification, the RDP staff could heuristically begin to relate overall mission success to states for any single event, by asking how a red, yellow or green state for a particular event might impact subsequent flight events. In addition, linking of the events to successes and failures and subsequently linkage of events to the functions that support them provides for the specification of the complete model for this system and its performance.

FUNCTIONAL, STRUCTURAL AND DEPENDENCY SPECIFICATIONS

In order to develop the complete model, it is necessary to define subsequent levels of detail that correspond to the top level ontology. In particular, it is necessary to define all functions that support all events, all subsystems and parts that support the functions, etc. This leads to development of two important views of the system, a functional view (all functions) and structural view (all subsystems and their parts) and the inter-relationships or dependencies between them. The next stage then in specifying the full ontology was to focus on each flight event and begin identifying key parts, subsystems and functions. Working with the subsystem engineers, we created the next three levels of specification for each event: a *functional diagram* that detailed only the functions required for an event; a *subsystem-part diagram* that broke subsystems into collections of parts; and a modified *series parallel diagram* that specified the order in which parts in a subsystem work together to perform a function (dependencies).

For each event displayed on the Inter-event Dependency Diagram, we created a representation to detail relationships between functions and events. For example, Figure 4 details the functions that the booster must execute in order for the first stage of the flight to occur. Note that the representation says nothing about the *state* (red, yellow, green) of the functions, or the event itself: the functional drawing simply relates functions to other functions and ultimately to the event, "boosted flight." The relations to the states are captured by the full model, specifically how the events interact and relate to the states.

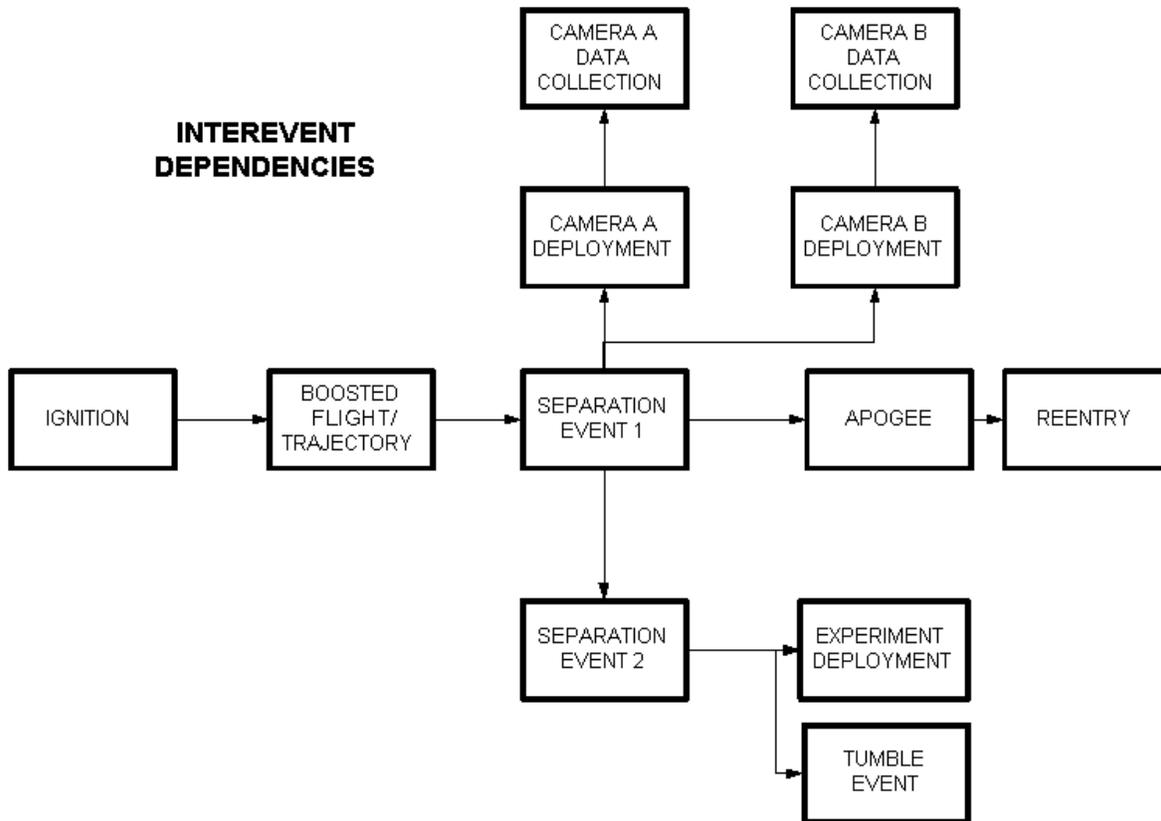


Figure 3. Specification of inter-event dependencies for rocket flight.

The representation above identifies two primary functions for “TR Flight.” These functions include “Data Collection/Vehicle Tracking,” and “Boosted Flight,” which are themselves broken into several sub-functions. These sub-functions, in turn, can be further specified by the parts and subsystems involved in their performance.

Note that in the drawing below, the event “TR Flight” depends not only on a set of nested functions, but also on a previous event in the trajectory, “Ignition.” Given that a rocket flight is an enormously complex set of dependencies, one of the convenient things about this type of representation is that it allows the knowledge modeler to detail only the functions specifically required for the event in question. In other words, while a “Boosted Flight” of course depends heavily on what happens during “Ignition,” those ignition-related functions are detailed in a set of representations for the “Ignition” event and do not need to be re-drawn for “Boosted Flight.”

Not shown in this paper are the next two levels of representational abstraction. *Subsystem-part representations* are graphical inventories of specific parts and the subsystems that house them. It is important to point out that this view provides no information about *how* any constellation of parts performs a function, but rather identifies how specific parts are grouped into subsystems. This is important since functions are the result of individual parts in separate subsystems working simultaneously across subsystems to produce a particular function. This diagram is less a representation for the statistical model than it is a “laundry list” that the knowledge modelers and the engineers use to ensure that all parts are properly grouped into their respective subsystems.

BOOSTED FLIGHT: PRIMARY FUNCTIONS

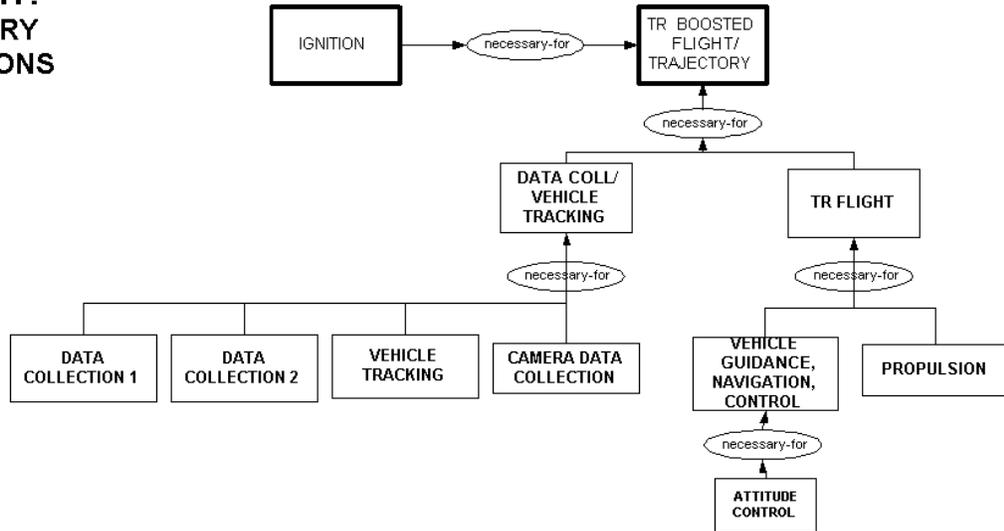


Figure 4. Functional view of event “Boosted Flight”

Dependency knowledge is specified in the next stage of abstraction, a *series parallel diagram* that locates parts within a subsystem and displays the order in which parts function with each other to perform a given function. Most engineering drawings tend to be structural in nature, not functional: in other words, they display connections among parts, rather than describe how parts work together to perform one of more functions. Although we realized that a functional view of the system would be critical for developing any kind of predictive model of rocket performance, that knowledge was not only tacit; it was distributed across numerous individual engineers. Hence it was necessary to elicit and represent this information using the functional and structural specifications described above. This stage marked the beginning of the transition from an engineering understanding of the system, to a statistical dependency model that could be quantified and populated with available data to make predictions about the rocket in flight.

The series parallel diagram was the first step in this transition. This type of drawing is somewhat similar to a series parallel diagram exemplified in a classic reliability block diagram, but with a great deal more descriptive information. Block diagrams simply connect parts to parts in the order that they must perform so that a given phenomenon occurs. The series parallel diagrams we developed followed the structure of a reliability block diagram but contained a great deal more information about the context of a particular part and its functions.

QUANTITATIVE MODEL DEVELOPMENT

DEPENDENCY FORMULATION

Although different kinds of series parallel diagrams provide a wealth of information about how parts and subsystems and functions are linked to events on the rocket trajectory, these diagrams are not sufficient for building a Bayes net. This is because Bayes nets represent *dependencies* among their elements: given what I know about one node in a model, what might I be able to say about nodes whose states depend on that event? The final stage in the knowledge modeling process, then, is to transform the series parallel diagrams into dependency diagrams. The difference between the two is subtle, but critical: Series parallel diagrams specify the linkages among parts related to a function and imply some order to those parts: for example, a power function might be described as, ‘Battery A

feeds power to a PTS, which sends a current to the following electrical components:..” A dependency diagram, on the other hand, describes that same power function as dependent on the performance of Battery A *and* the PTS, and how downstream components’ performance is (at least partially) dependent on that power function.

The most immediate difference between a basic series parallel diagram and a dependency diagram is that subsystems are not represented in the latter. This is because subsystems simply designate the geographical location of parts within the rocket; dependencies exist between their parts and one or more functions. Strictly speaking, no functions depend on a subsystem; however, many functions may depend on the individual parts *within* a subsystem.

In a dependency diagram, we are concerned with specifying three types of information: how functions depend on one or more necessary parts, how the performance of a particular part depends on a particular function (recursive relationships), and how parts may provide redundancy (part A *or* part B is necessary for function X) or single points of failure (part A *and* part B are necessary for function X). These relations among parts and functions specify the dependency structure for a Bayes net. Also, at this point in the IIT method, the knowledge modeler and the statistician are working closely together in the modeling process.

BAYESIAN MODEL DEVELOPMENT

The final transition occurred when the dependency diagram was turned into the Bayes net structure. The diagram shown in Figure 5 is a Bayes net, extracted from the larger rocket model. The statistician built it using the dependency diagram developed. The initial translation can be performed easily from the dependency diagram to the Bayes net, although the knowledge modeler and the statistician do work together to check the Bayes net and ensure that the statistician has specified the right dependencies, labeled the functions and parts correctly, and indicated the proper directionality in the relationship arcs.

The Bayes net is a highly distilled version of the dependency diagram: it eliminates all relationship labels and, at the level shown above, offers no information about subsystem location for any of the parts. Population of the model occurs in later iterations, using the series parallel diagrams for failure (to designate a range of states for each of the part and function nodes), the stoplight charts (to designate states for the mission events), and the series-parallel data diagrams (to identify sources of data for each part and its associated failure modes). The model generates a probability distribution for each event in the inter- event dependency diagram, as well as a final probability distribution for states red, yellow, and green for the entire mission. In addition, the Bayes net allows the user to trace sample paths for different solutions through the states of each node, so that it is possible to connect given outcome for the entire system to the state of any particular node.

CONCLUSION

Multidisciplinary projects often lack integrated representations to support the community’s problem-solving process. It is frequently difficult for project insiders to develop these representations: for one thing, they are focused on meeting the project’s goals. More subtly, insiders often have a great deal of local knowledge about a specific area within a project, but may have difficulty leveraging that into a global view of the problem. Anthropologists and knowledge modelers, on the other hand, are trained to elicit this information and can draw on a wide range of representation techniques to create useful abstractions of the project area.

An interdisciplinary approach to knowledge modeling, one that combines techniques from anthropology and knowledge representation, is particularly helpful in situations where problems are undergoing definition, are emergent, and that involve multiple players from different disciplines and/or geographical locations. When such modeling techniques are paired with quantitative tools from statistics, it becomes possible to develop complex models that can, among other things, enable the integration of multiple, diverse sources of data to estimate performance without testing.

FUTURE DIRECTIONS

Our future directions for the IIT method are to begin using it to integrate diverse information sources for use in more real time environments such as those of a tactical battlefield. In these types of problems information is obtained and must be integrated to develop hypotheses about the situation at hand and provide probabilities and uncertainties to those hypotheses. In this case, the integration must be performed in a real time environment by gathering data in real time and developing hypotheses. However, much of the same basic underlying method used in the problem described above still remains largely intact.

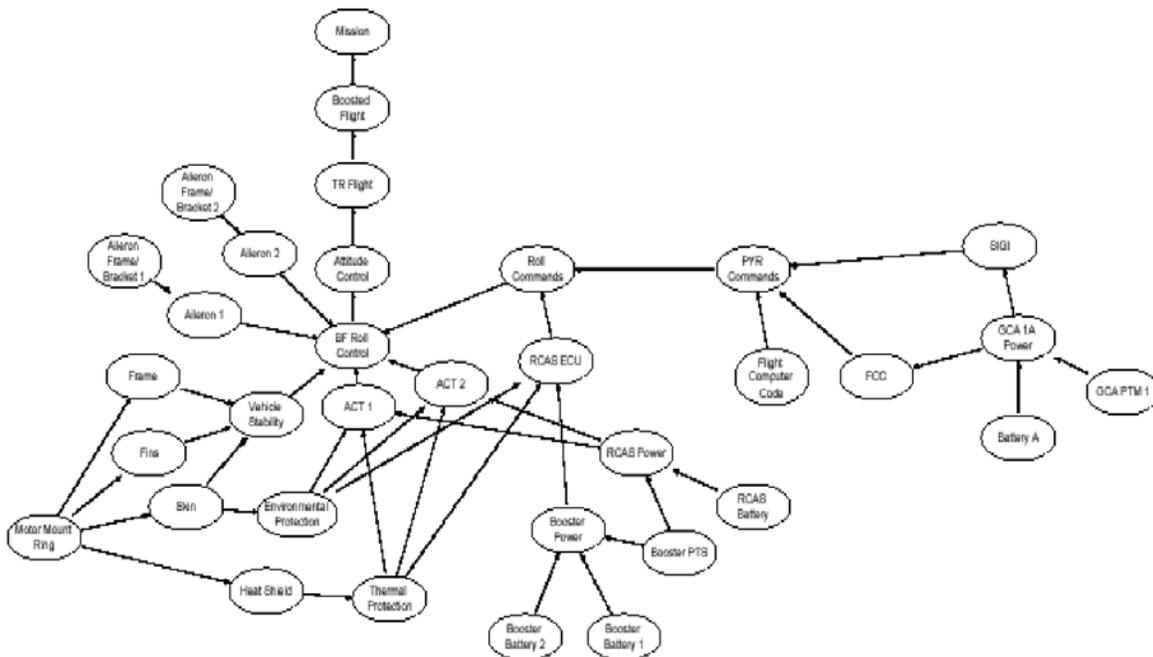


Figure 5. Bayes Net Representation for Roll Control

REFERENCES

- Sowa, John. 1984. *Conceptual Structures*. Addison Wesley Publishing Co. Reading Ma. 1984.
- Wenger, Etienne. 1998. *Communities of Practice: Learning, Meaning and Identity*. Cambridge, UK: Cambridge University Press.