Implementing an Intelligent Multimedia Presentation Planner using an Agent Based Architecture

Todd Andrews, Michael Broughton & Dominique Estival
Command and Control Division (C2D)
Defence Science and Technology Organisation (DSTO)
Tel: +61-08-8259-5555
{Todd.Andrews, Michael.Broughton, Dominique.Estival}@dsto.defence.gov.au

ABSTRACT
This paper describes the implementation of an intelligent Multimedia Presentation Planner (MPP) in a multimodal dialogue system. Following the development of an architecture based on the Standard Reference Model, designed specifically for FOCAL, this implementation has been integrated with an earlier spoken dialogue system. The design now ensures that the framework is portable to other multimodal environments.

Keywords
Multimedia presentation, agent-based dialogue systems

1. Introduction
In this paper, we describe the implementation of an intelligent Multimedia Presentation Planner (MPP) in a multimodal dialogue system. Following the development of an architecture designed specifically for FOCAL, this implementation now ensures that the framework is portable to other multimodal environments, such as the LiveSpace and Interactive Collaboration Space laboratories also in use at DSTO. The paper focuses on the presentation side of the system, but shows how the MPP is part of a complex agent-based architecture which allows for a variety of input and output modalities. After the brief description of FOCAL given below, Section 2 is a general overview of the MPP and Section 3 describes in detail each of the five agents, giving examples from a simple presentation. Section 4 goes through the more complex example taken from the scenario serving as the testbed for FOCAL development. Section 5 concludes by sketching further work.

1.1 FOCAL
The Future Operations Centre Analysis Laboratory (FOCAL) at Australia's Defence Science and Technology Organisation (DSTO) is aimed at exploring new paradigms for achieving situation awareness (Endsley, 1995) in military command and control systems. In support of this, new media types are being developed that include Virtual Advisers, 3D battlespace visualisations, 3D abstract data representations and Virtual Video technologies (Wark et al., 2004). These are planned to be utilised in combination with more established media technologies, such as static and video imagery, and spoken and written text. In addition to the more unique media types, FOCAL's primary display comprises a 150º field-of-view, large spherical screen capable of displaying a stereoscopic semi-immersive environment. FOCAL implements a heterogeneous multi-agent architecture based on the CoABS (Global InfoTek, 2003) infrastructure.

The agents are implemented in ATTITUDE, which enables representation and reasoning with uncertainty and reasoning about multiple alternative scenarios (Lambert, 1999). The dialogue system takes advantage of ATTITUDE's ability to reason about propositional attitudes (Estival et al., 2003). In this paper, we will focus on the presentation side of the dialogue system and we will not discuss the input modalities, but spoken and typed input are already integrated and other input modalities, such as gesture, are being added.

FOCAL requires that an MPP fulfil at least two complementary roles, that is, be able to generate presentations suitable for a brief, and be able to generate suitable responses to interactive questioning. A collaboration between CSIRO and DSTO resulted in CSIRO proposing a suitable framework for FOCAL's MMP (Colineau and Paris, 2003; Paris et al., 2004). The proposed framework utilizes Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) and extends the standard Reference Model proposed by Bordegoni et al (1997). Examples of other MPP systems include André and Rist (1990, 1993), Maybury (1993), Feiner and McKeown (1997), Bateman et al. (1998).

1.2 Integration with FOCAL
The FOCAL architecture diagram (Wark et al. 2005), given in Figure 1, shows the integration of the MPP into the FOCAL system. Of particular interest are the communication paths with the Dialogue Manager, Query Manager and Information Fusion Agents, which are briefly described below.

1.2.1 Dialogue Manager
The Dialogue Manager (DM) generates a communicative goal for the Multimedia Presentation Planner to achieve, given the communicative acts it receives from Multimodal Input Processing (MIP). For each possible interpretation of the communicative act, it generates the possible dialogue acts and types of system responses, and selects the best combination. In future, dialogue ambiguities and errors will be resolved by further sub-dialogues.

1.2.2 Multimedia Presentation Planner

The Multimedia Presentation Planner (MPP) attempts to achieve the communicative goals assigned by the DM. It interacts with the Query Manager to interpret the semantic content of the communicative goals, retrieve any information needed, and/or accesses local multimedia databases to retrieve multimedia clips that satisfy the communicative goal.

1.2.3 Query Manager

The Query Manager (QM) interprets the semantic content of communicative acts/goals passed to it, from the DM, MPP or semi-autonomous IF processes, by referencing ontologies, and retrieves appropriate information from distributed static or dynamic information sources when necessary. The QM can interact directly with the DM to resolve query ambiguities or errors when required. The ontologies used by the QM are also used when creating the grammars and lexicons used for speech recognition and natural language processing (Nowak et al., 2004).

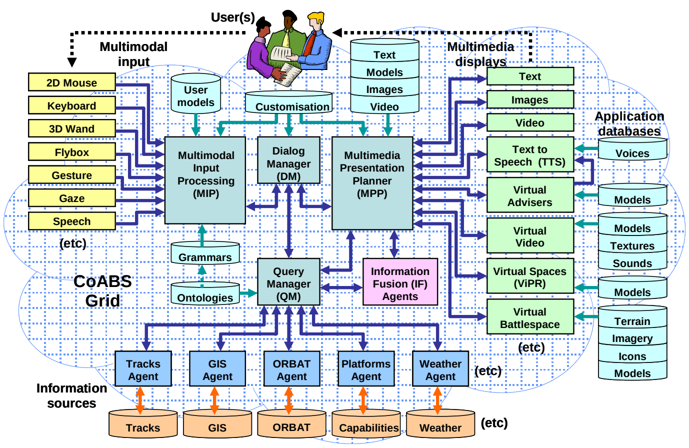

Figure 1. FOCAL Architecture

2. The Multimedia Presentation Planner

The Multimedia Presentation Planner (MPP) Agent is responsible for selecting and displaying media items for a multimedia presentation. In order to select the appropriate media items, the MPP is passed a communicative goal from the Dialogue Manager Agent, represented as a formatted string. The MPP extracts the appropriate information from the string and uses resources such as the Query Manager Agent to develop a selection strategy. Next, the MPP makes a decision on the spatial and temporal aspects of the presentation layout based on the specific media to be used and also some rules and templates that already exist. It produces a presentation strategy and each media item is realised by either being generated in real-time or located from a database. The multimedia presentation is then displayed on the screen and any other medium that is required.

As shown in Figure 2, the MPP is made up of four Agents corresponding to the layers of the Reference Model (Paris et al., 2004): Content Layer, Media Selection Layer, Presentation Layer and Distribution Layer. The Local Rendering Agent (LRA) can be seen as a fifth layer of the MPP, but it is slightly different because, unlike the other four agents, the LRA is not independent of the system it is running on.

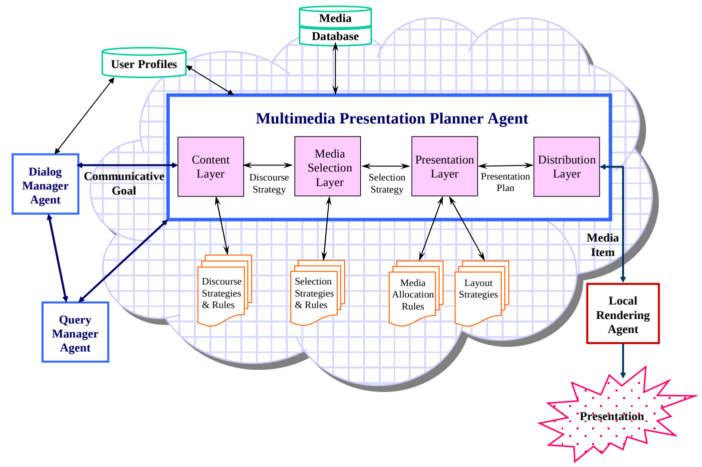

Figure 2. The MPP agents

3. The MPP Agents

The input to the MPP Agent is a communicative goal sent by the Dialogue Manager. For example, a communicative goal could be a request for a statement in response to an explicit query made by a user, for example: *What is Melbourne?*

3.1 Content Layer Agent

The aim of the content layer is to take a communicative goal from the Dialogue Manager as input, then process it and produce a discourse strategy. The discourse strategy is a representation of the information the multimedia presentation will be presenting to the user. Many aspects need to be considered when processing the communicative goal and constructing the discourse strategy. For example, the length of the presentation, how much detail should be included, the context and stage of conversation, and also the user's preferences.

3.1.1 Implementation

The Content Layer Agent (CLA) receives a communicative goal containing a content expression. The content expression is used to find a discourse strategy from the CLA's knowledge base by adequately matching the aims and constraints of the communicative goal. There may be a number of discourse strategies that would make a successful match with the input communicative goal. The way the CLA selects the most appropriate discourse strategy is by allocating discourse strategies a score. The discourse strategies are ranked from the most simple to the most detailed and the CLA iterates through them, simultaneously from the most simple through to the most complex and from the most complex to the most simple. The most simple discourse strategy will have the lowest score while the most detailed discourse strategy will have the highest score. The CLA will select the highest scoring discourse strategy.

While iterating through the possible discourse strategies from the highest scoring down, the process is stopped as soon as a successful match is made and that discourse strategy is chosen. It is important to note that the most highly rated discourse strategy will not always be a match, because a successful match is based on factors such as the length of the presentation and user profiles. Therefore a match might be somewhere in between the longest, most detailed discourse strategy and the most basic discourse strategy. During the main iteration from the most complex discourse strategy down, there is a second iteration from the lowest scoring up. Unlike the first one, this iteration will not finish once it finds a match, because it should find a match on the first try with the most generic discourse strategy. However, the successfully chosen discourse strategy will be the last match made before a specified timeout value. This second iteration is used as a fallback mechanism in case there are too many to get through. If the high-to-low loop has not succeeded after a certain time, then the best lowest scoring routine will be used.

3.1.2 Discourse Strategy Example

The discourse strategy provides an answer to a query, or a reply to a statement. It may also provide additional information in order to provide a longer presentation as the communicative goal requests. For example, in the case of the *What is Melbourne?* user query, the most simple discourse strategy would have been to reply, "`Melbourne is a City`". However a more complex discourse strategy might contain some additional information that is provided by the ontology such as "`Melbourne is a city. It has a population of over 3 million`."

3.2 Media Selection Layer Agent

The Media Selection Layer Agent (MSLA) is responsible for selecting the media items that are to be displayed in the final presentation. The MSLA is passed a discourse strategy as input from the CLA. It is from this discourse strategy that the MSLA will need to query the appropriate information sources to find all the relevant media items available. Once this has been done, a decision must be made as to which media item to use. The MSLA will need to use the Media Database to lookup any images, videos, and documents that may be available on a particular concept or topic. It will also ask the Query Manager Agent for additional information from the ontology or other information sources, such as geospatial information (GIS). Each presentation will differ based on the media items selected and media types used. Media items will be selected according to temporal and spatial criteria depending on a number of factors, such as relevance, time constraints on the presentation, visual and logical coherence, complementary media items, presentation history and also user profile considerations such as assumed knowledge and media preference.

3.2.1 Implementation

The MSLA interacts with the Media Finder agent to form a list of relevant media items for a particular content expression or possibly even a keyword. The agent can also generate ATTITUDE expressions that do not come from the Media Finder database and use them as input. For example, it can decide to display a location in a virtual battlespace environment. The MSLA is also responsible for Language Generation and it can generate simple sentences that correspond to the relationship and arguments from the content expression triples. These content expressions are binary relations that represent knowledge from the domain ontology by using two arguments that can be concepts or instances of concepts, and specifying the relationship between these two arguments (Nowak, 2003).

3.2.2 Selection Strategy Example

The selection strategy is essentially of a similar form to the discourse strategy but each discourse segment is now accompanied by media items. Using the *What is Melbourne?* example, the selection strategy might contain Virtual Advisor media along with Virtual Battlespace media showing the location of Melbourne, in order to illustrate the first discourse segment, "`Melbourne is a city.`" The selection strategy then might consist of an Image or video of the Melbourne CBD along with Virtual Advisor media to illustrate the elaboration discourse segment, "`It has a population of over 3 million`."

3.3 Presentation Layer Agent

The purpose of the Presentation Layer Agent (PLA) is to create a presentation plan. The presentation plan is a result of the discourse and selection strategies that are passed up from the CLA and MSLA respectively. The PLA is instructed what media items to present and needs to decide on the spatial and temporal relations between these media items. It makes these decisions based on factors such as how many media objects are to be displayed, the physical size of the display medium (e.g. screen size), complementary media types and the temporal ordering properties of the presentation.

3.3.1 Implementation

The PLA is divided into two main parts. The first part sets the temporal constraints and the second part calculates the spatial constraints of the presentation.

The PLA first reformats the selection strategy according to the linear form of the presentation plan. This is done by a recursive process that interprets the nucleus and satellite RST relationships and produces a temporal ordering as sequence and parallel elements, similar to functionality within SMIL (SYMM Working Group, 2001). Next, each element is given temporal constraints that are based on the type of media items, the order in which they are to appear in the presentation and their relationship with other media items to be played, in parallel or in sequence with the element. The constraints specify duration times (which can be explicit or relative to another element), a *replace* value that determines whether the element is to stay on the screen or be removed upon completion, and a *timeout* value that acts as a default time when the image can be removed from the presentation if necessary.

Once the temporal constraints and the ordering sequence of the presentation are determined, the spatial constraints are set. The first step in this process is to determine the environment for which the constraints of the presentation are going to be designed for. This involves asking the Local Rendering Agent (described in the next section) about the properties of the screen environment, including the number of screens and their size. Next, the PLA will make decisions on screen placement according to the display environment available. For example, the spatial constraints for showing a particular presentation in the FOCAL environment will be considerably different from those for showing the same presentation on a small tablet PC.

The PLA uses a combination of dynamic and template based approaches for screen placement of media. An example of screen layout is shown in Figure 3, which demonstrates the use of predefined grid layers of various resolutions to aid in the placement of media items. The layers are selected on the basis of how many and what types of visual media items are to be displayed at once.

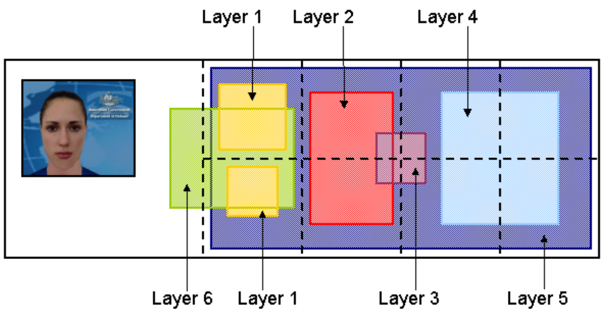

Figure 3. Screen layout using a combination of grid layers

The placement of media items is also dependant on what other media items are already on the screen, or are due to be displayed. An advantage of using a combination of a grid based and dynamic approach to screen placement is that a choice can be made whether or not use a cell if there is a need for it to remain free. Consequently a decision can then be dynamically made to choose another cell to display the media or even to choose an entirely different grid layer.

3.3.2 Presentation Plan Example

The presentation plan of the *What is Melbourne* example contains the nucleus – satellite relationship of the two discourse segments from the selection strategy as a sequence of two elements, using the RST relation to decide the order. Each of these elements are parallel elements containing the two media items to display and also the details of any temporal and spatial constraints. For example, the second `par` element of the sequence contains details about the Virtual Advisor media element and the image media element. It specifes for the Virtual advisor to say, "`It has a population of over 3 million.`" and to place the image media item to the right of the screen. The media elements also specifies any environment constraints, such as which screen or screens to display the media items on.

3.4 Distribution Layer Agent

The Distribution Layer Agent (DLA) takes the output from the PLA and identifies the media items to display. It uses the timing constraints specified in the presentation strategy to send a message to the Local Rendering Agent (LRA), specifying which media items to render. The messages that the DLA sends to the LRA need to be high level and non-specific in order to preserve flexibility and portability of the MPP Agents. For example, the DLA might send a string to the LRA which only specifies the media type, file and some temporal and spatial constraints. The LRA will then transform this data into a format that can be used to communicate with the local rendering application that is used to render a specific media type.

3.4.1 Implementation

The DLA iterates through the `seq` and `par` elements of the Presentation Plan and handles the temporal execution and ordering of the Multimedia Presentation. The DLA tells the LRA to play, update, remove, move, minimize or maximize media items. It has rules that are specific to a media type, that determine whether to play or update a media item and also decide whether remove or minimize it after it has finished playing. When playing a `seq` element, the DLA simply iterates through each element in a loop, waiting for the element to have finished before it iterates onto the next in the sequence. Once the last element has finished, the agent will know that the whole sequence has been executed. When playing a `par` element, all elements are executed at the same time. The DLA will determine that a `par` element has finished playing in one of three ways. If there are media items to be played that have an explicit duration time, for example a video, then the `par` element is finished once the media item with the longest explicit duration has completed playing. If there are no media items to be played in parallel that have explicit duration times as a constraint, for example an image, then the `par` element will be considered as finished once the accompanying utterance has finished or after a specified timeout value.

3.4.2 Media Output

The output of the Distribution layer is a simple, high level command that specifies whether to play, update, move, remove, minimize or maximize a media item. It contains details of the media type, file name (if there is one) and any other essential details such as screen position, size, and audio channel.

3.5 Local Rendering Agent

The LRA is the only MPP agent which is specific to the environment in which the presentation is to be displayed. This encapsulation ensures portability across platforms and operating systems. The LRA knows which tools are available in its local environment and what types of media it can handle. Thus, its primary purpose is to wait to receive input from the DLA, for example "display *this* image in *this* position", and then forward that information to the appropriate rendering device. The LRA can also be queried by other layers of the MPP in order to provide information about the display environment or media availability. For example, the MSLA might want to know if a specific media type is available to play on the system it is running on, or the PLA might need to know the screen size and number of screens to display the presentation on.

3.5.1 Implementation

The LRA can play, update, move, remove or minimize media once the DLA asks it to do so. Once any of these requests are executed, the LRA will use a set of commands specific to the media type which will know how to play, update, move, remove or minimize the media item. The LRA also has the ability to service requests made by the MSLA and PLA, and can provide details of the screens and media types that are available for use in the multimedia presentation.

3.5.2 Media Rendering

The output of the LRA is a formatted string that is sent to the desired proxy agent which will parse the string and execute commands in order to play the media item.

4. More complex example

A more complex example of this implementation is taken from the scenario serving as the testbed for FOCAL development. It was developed in the Annex of the CSIRO report (Colineau and Paris, 2003) to illustrate the proposed framework. We show how the discourse strategy and the selection strategy constructed in that example would be implemented and displayed as a multimedia presentation using the MPP Agent. This example is taken from the Geopolitical Summary in an Induction Brief*:*

"The Admiralty Islands belong to Papua New Guinea. Because of its international agreements with Papua New Guinea, Australia is inclined to protect Papua New Guinean sovereignty. In addition, a decade of hostility has existed between Australia and Kamaria, manifested by poor diplomatic relations and at times minor maritime and air confrontations."

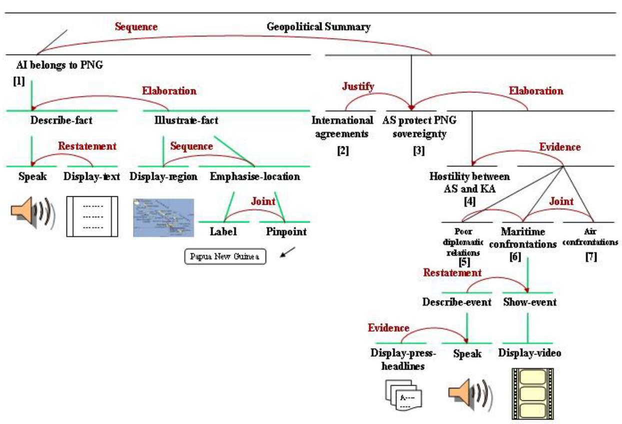

Figure 4. RST implementation of FOCAL scenario example from CSIRO report (Colineau & Paris, 2003)

In the implementation described in this paper, the discourse strategy is produced by the CLA. It uses binary relations to illustrate the RST relationships between discourse segments and content expressions to represent information.

```
(complex_rst
  :relation sequence
  :nucleus (simple_rst
    :content (belongs_to AI PNG))
  :satellite (complex_rst
    :relation elaboration
            ...
        :relation joint
        :nucleus (simple_rst
          :content (rel_diplomatic AUS KM))
                ...
```

We take the discourse strategy chosen by Colineau and Paris and send it as input to the MSLA which produces the selection strategy using the media items as shown in Figure 4. Each discourse segment is accompanied by an utterance by the Virtual Advisor. There are also visual media elements associated with the first and sixth discourse elements. The discourse segments from the selection strategy are converted by PLA into six sequential elements. Of these, two are parallel elements that contain Virtual Advisor media along with text, video or Edge (Virtual Battlespace) media elements. An example of the presentation plan structure is provided below.

```
(seq
  :element ((par
             :element ((element
                          :media (media :type text))
                       (element
                          :media (media :type edge))
                       (element
                          :media (media :type VA))))
           (element
             :media (media :type VA))
                  ...
           (par
             :element ((element
                         :media (media :type text))
                       (element
                         :media (media :type video))
                       (element
                         :media (media :type VA))))))
```

Each element has an associated media item description and a set of constraints that are set by the PL. For example, for the first discourse segment and its accompanying media, each element has had temporal and spatial constraints assigned to it. In this case, the Edge map has been situated in the background and takes up the right two thirds of the screen. The text has been located on top of Edge in the centre of the display, and the virtual advisor will play from its usual place on the left of the screen. Figure 5 illustrates this layout arrangement.

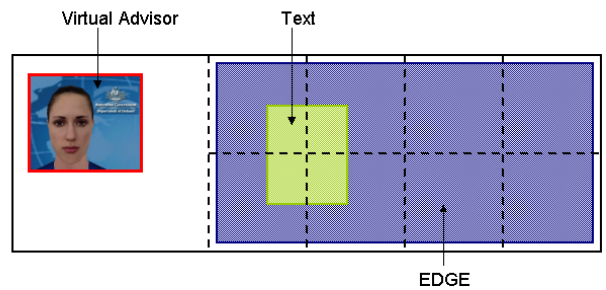

Figure 5. Example of screen layout showing multiple media items.

```
(par
 :element ((element
          :media (media
                 :type text
                 :text "The Admiralty Islands ...")
          :constraints (constraints
                      :dur default_time
                      :replace leave
                         ...
                      :x-pos 1420 :y-pos 272
                      :width 320 :height 480))
         (element
          :media (media
                 :type edge
                 :location (location
                      :geographic-identifier S-W-Pacific
                      :location-type region )
                 :highlight ((location
                      :geographic-identifier PNG
                      :location-type country)
                             (location
                      :geographic-identifier A-I
                      :location-type state)))
          :constraints (constraints
                      :screen FOCAL_main
                         ...
                      :x-pos 1257 :y-pos 37
                      :width 2178 :height 950))
         (element
          :media (media
                 :type VA
                 :advisor Bob
                 :say "The Admiralty Islands ..."))))
```

In this example, temporal cues can be taken from the Virtual Advisor's speech utterances as to when this parallel set of elements has finished playing. This is because the text and Edge elements do not have an associated length or duration with their media type. Once the utterance has stopped, the presentation will continue to the next element in the sequence, as shown in Figure 6. The temporal cues are the same for the next four elements in the sequence. However, the temporal cue for the parallel set of media elements at the end of the presentation is set to the end of the video, as it does have an explicit duration, unlike the text and image media types. The text and video media elements that appear at the end of the presentation will have spatial constraints set by the PLA which place them side by side in the centre of the screen, whilst still having Edge in the background.

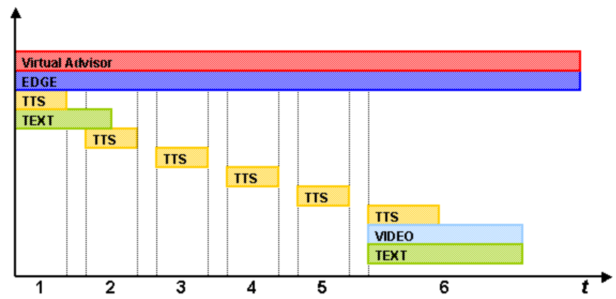

Figure 6. Example of temporal ordering.

5. Conclusion

This paper has outlined the design and implementation of an agent based intelligent MPP within a framework based on Rhetorical Structure Theory and extending the standard Reference Model. We have demonstrated how the MPP has been customized for FOCAL's display and novel media types. This MPP has been integrated in an existing spoken dialogue system already implemented with an agent based architecture and written in ATTITUDE. Current work is aimed at expanding the current range of presentations that can be designed. Finally, implementations of the LRA will be developed to enable utilization of the MPP in other large multi-screen environments such as LiveSpace. Similarly, a LRA will also be developed for single screen applications and other environments that access more conventional media types. This work has yet to be evaluated; subjective and objective evaluations will be conducted after the media coverage has been increased to allow for greater richness and variety in the presented information.

6. ACKNOWLEDGMENTS

Our thanks to the HSI group and the FOCAL team, in particular Steve Wark and Andrew Zschorn.

7. REFERENCES

[1] André, E. and Rist, T. The design of illustrated documents as a planning task. In M. Maybury (Ed.), Intelligent Multimedia Interfaces, 1993, chapter 4, 94-116. AAAI Press / The MIT Press.

[2] André, E., Müller, J. and Rist, T. WIP/PPP Automatic Generation of Personalized Multimedia Presentations. In *Proceedings of Multimedia 96, 4th ACM International Multimedia Conference*, Boston, MA, 1996, 407-408.

[3] Bateman, J., Kamps, T., Kleinz, J. and Reichenberger, K. Communicative goal-driven NL generation and data-driven graphics generation: an architectural synthesis for multimedia page generation. In *Proceedings of the 9th International Workshop on Natural Language Generation, Niagara-on-the-Lake, Canada*, 1998.

[4] Bordegoni, M., Faconti, G., Maybury, M.T., Rist, T., Ruggieri, S., Trahanias, P. and Wilson, M. A standard reference model for intelligent multimedia presentation systems. *Computer Standards and Interfaces: the International Journal on the Development and Application of standards for computers, Data Communications and Interfaces*, 1997, 18,6-7.

[5] Colineau, N. and Paris, C. *Framework for the Design of Intelligent Multimedia Presentation Systems: An architecture proposal for FOCAL*. CMIS Technical Report 03/92, CSIRO, May 2003.

[6] Endsley, M. R. Toward a Theory of Situation Awareness in Dynamic Systems. Human Factors, 37,1, (1995), 32-64.

[7] Estival, D., Broughton, M., Zschorn, A. and Pronger, E. Spoken Dialogue for Virtual Advisers in a semi-immersive Command and Control environment. *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*, Sapporo, Japan, 2003.

[8] Feiner, S. K. and McKeown, K. R. An analysis of COMET and MAGIC using the standard reference model for intelligent multimedia presentation systems. *Computer Standards and Interfaces: the International Journal on the Development and Application of standards for computers, Data Communications and Interfaces*, 1997, 18(6-7). 565-574.

[9] Global InfoTek Inc.: CoABS website. http://coabs.globalinfotek.com. Accessed 4 Aug 2005

[10] Lambert, D. A. Advisers With ATTITUDE For Situation Awareness. In *Proceedings of the 1999 Workshop on Defense Applications of Signal Processing*, La Salle, Illinois, 1999, 113-118.

[11] Mann, W. and Thompson S. Rhetorical Structure Theory: Towards a Functional Theory of Text Organization. *Text*, 8,3 (1988), 243-281.

[12] Maybury, M. *Planning multimedia explanations using communicative acts*. In M. Marbury (Ed.), Intelligent Multimedia Interfaces, 1993, chapter 2, 59-74. AAAI Press / The MIT Press.

[13] Nowak, C., Estival, D., and Zschorn, A. Towards Ontology-Based Natural Language Processing. *In Proceedings of RDF/RFDS and OWL in Language Technology: 4th Workshop on NLP and XML*, Barcelona, Spain, 2004, 59-66.

[14] Nowak, C. On ontologies for high-level information fusion. In *Proceedings of the Sixth International Conference on Information Fusion*. Cairns, Australia, 2003.

[15] Paris, C., Colineau, N., and Estival, D. Intelligent Multi Media Presentation of information in a semi-immersive Command and Control environment. *Australasian Language Technology Workshop*, Sydney. 2004, ASSTA, 85-92.

[16] SYMM Working Group. Synchronized Multimedia Integration Language Version 2.0, World Wide Web Consortium.http://www.w3.org/TR/SMIL20 , 2001.

[17] Wark, S., A. Zschorn, M. Broughton and D. Lambert. FOCAL: A Collaborative Multimodal Multimedia Display Environment. *SimTecT 2004*, Canberra, Australia, 298-303.

[18] Wark, S., Broughton, M., Nowina-Krowicki, M., Zschorn, A., Coleman, M., Taplin, P. and Estival, D. The FOCAL Point - Multimodal Dialogue with Virtual Geospatial Displays. In *Proceedings of SimTecT 2005*, Sydney, Australia.