Towards a Unified Knowledge-Based Approach to Modality Choice

Yulia Bachvarova, Betsy van Dijk, Anton Nijholt
Human Media Interaction Group, University of Twente,
PO BOX 217, 7500 AE Enschede, The Netherlands
{y.s.bachvarova, e.m.a.g.vandijk, anijholt}@cs.utwente.nl

Abstract

This paper advances a unified knowledge-based approach to the process of choosing the most appropriate modality or combination of modalities in multimodal output generation. We propose a Modality Ontology (MO) that models the knowledge needed to support the two most fundamental processes determining modality choice – modality allocation (choosing the modality or set of modalities that can best support a particular type of information) and modality combination (selecting an optimal final combination of modalities). In the proposed ontology we model the main levels which collectively determine the characteristics of each modality and the specific relationships between different modalities that are important for multi-modal meaning making. This ontology aims to support the automatic selection of modalities and combinations of modalities that are suitable to convey the meaning of the intended message.

Keywords: Modality Ontology, Modality Choice, Modality Allocation, Modality Combination.

1 INTRODUCTION

The process of choosing and combining modalities to best convey the intended message is central for multimodal output generation. It is also a complex and highly knowledge-intensive process that depends on the type of the information that has to be represented and the specifics of the context, the user and the particular goal of the multimodal presentation on the one hand and the proper understanding and modelling of the nature of each modality and of multimodal meaning making on the other hand. Research on all these different aspects has been conducted by different communities. A lot of the research results gained, though relevant for multimodal output generation, remain scattered and not really employed to their potential. A unified framework, capturing the aforementioned aspects in their array of dependencies can properly address and formalize the complexity of the problem of modality choice.

The work that we present in this paper attempts to start addressing the issues related to modality choice in a unified and systematized manner. The two most fundamental processes related to modality choice are modality allocation and modality combination. Modality allocation assigns the most appropriate modalities that can best represent the types of information that have to be represented. Modality combination is the process where modalities are integrated into a coherent final multimodal message.

We start with the assumption that there is a formal representation, for example a domain ontology, of what has to be represented. We look at the types of information that have to be represented and the existing relationships between them and map this to the specific features of modalities describing their strengths and weaknesses in representing such information types and relationships. We further apply principles for optimal cognitive information processing or exploit the interdependencies between different

modalities that determine multimodal meaning making in order to generate the most optimal modality combination(s).

Central in the design of the ontology is the idea that there are two main aspects that properly describe each modality – the content it represents, and its nature. While modelling the content of some modalities has recently received significant attention, research on the nature of a modality has not been properly systematized. Therefore, we address the issue of what describes the nature of a modality void of its content. Moreover, the focus on the relation between modality content and modality form will be shown to have important implications for multimodal meaning making.

We start by describing the main levels of the Modality Ontology providing examples on how the knowledge modelled in these levels can support modality choice. We then provide an example of the relation between modality content and modality profile. Finally, we conclude by outlining our future research directions.

2 MODALITY ONTOLOGY

The main purpose of the ontology we propose is to be able to support the automatic selection of modalities and combination(s) of modalities, hence the processes of modality allocation and modality combination. To be able to support these two processes, the Modality Ontology (MO) has to model the following main types of knowledge about modalities - knowledge about the capacity of each modality to represent different types of information, knowledge about the cognitive and perception related aspects of each modality's nature, and knowledge about the structural dependencies that exist between the different modalities and that determine the syntax of a given modality combination.

We demonstrate, but not in detail, how each of these aspects of knowledge about modalities is modelled by the ontology and provide simple examples how the ontology can support modality allocation and combination.

2.1 THE UPPER LEVEL OF THE MODALITY ONTOLOGY

The central idea of the approach we advance in this paper is that the meaning that each modality carries is determined by its content (the particular information it represents), its nature per se, that is its content-independent characteristics, and the relations existing between these two main aspects. In the MO the nature of a modality is modelled by the profile level. Further in this subsection we describe this level in more detail. Figure 1 shows the upper level of MO where the Modality class represents the operational concept of the ontology. Modality presents ModalityContent and is described by ModalityProfile.

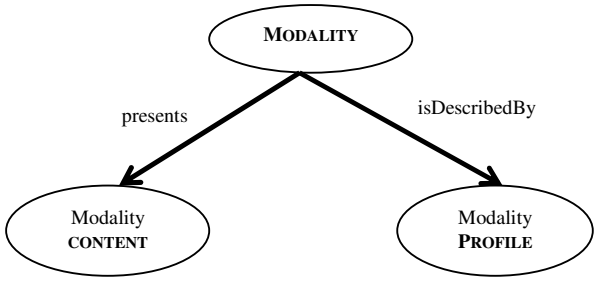

Figure 1: Upper level of the Modality Ontology

The ModalityProfile class describes knowledge about modalities at three different main levels – the information presentation level, the perception level, and the structural level. In MO these three levels are presented by the classes InformatonPresentationProfile, PerceptionProfile and StructuralProfile respectively (see Figure 2).

2.1.1 Information presentation level

The *information presentation level* models those modality characteristics that describe the strengths and weaknesses of each modality in representing particular types of information.. At the upper level of the InformationPresentationProfile we distinguish between *linguistic* and *analogue* modalities. The characteristics *linguistic* and *analogue* have been chosen based on their argued generality and robustness in profoundly distinguishing the different capabilities of modalities in representing information (Bernsen, 1994; Stockl, 2004). Linguistic representations, such as text and speech, are based on existing syntactic-semantic-pragmatic systems of meaning (Bernsen, 1994). An important feature of linguistic representations is that they *lack specificity* (Stenning & Oberlander, 1991); that is, they cannot specify precisely how things, situations or events look, sound, feel, smell or taste. Instead, linguistic representations are *abstract* and *focused* – they focus at some level of abstraction on the subject matter to be communicated. Those characteristics of linguistic representations determine their strength in representing abstract concepts, states of affairs and relationships. Analogue representations, such as images, represent through aspects of similarity between the representation and what they represent (Bernsen, 1994; Stockl, 2004). This determines the strong capacity of analogue representations to portray essentially visual or spatial information (Tversky, Morrison & Betrancourt, 2002). Analogue representations lack focus and can only to a limited extent represent abstract information. Knowing which modality feature is responsible for representing which information type allows mapping between what has to be represented and the modalities which can actually do that, i.e., MO supports the automatization of the modality allocation process. The information presentation level of the ontology can also support the modality combination process. The features of linguistic and analogue modalities we have chosen to describe here are complementary. The complementarity of features of analogue and linguistic modalities determines their frequent use together. In Section 2.1.2 we provide a concrete example of how modality combinations based on complementarity can be calculated.

The features of analogue and linguistic modalities that determine their capacity in representing different types of information are members of the class AnalogueModalityFeatures and LinguisticModalityFeatures respectively (see Figure 2).

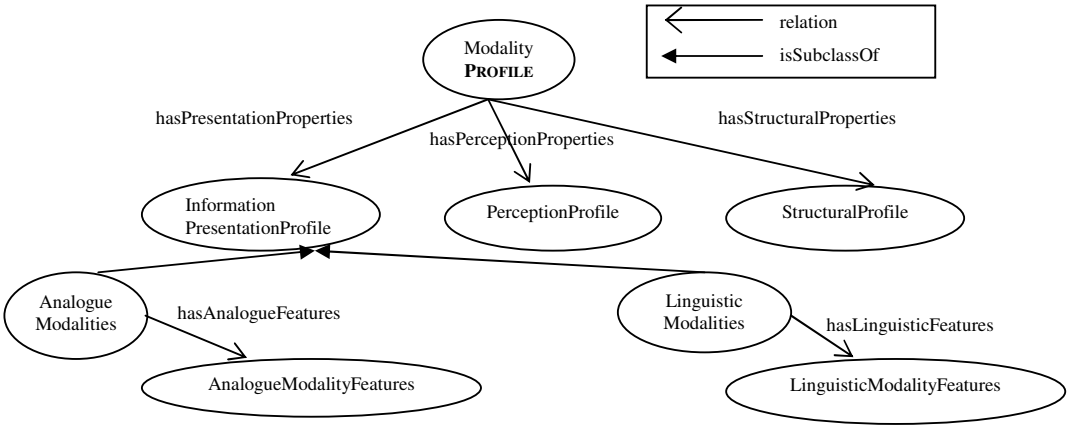

Figure 2: Information presentation level of the Modality Ontology

2.1.2 Perception level

The perception level models those modality characteristics which determine how a particular modality is perceived and processed by the human perceptual-sensory system (see Figure 3). At this level we distinguish between visual, auditory and haptic modalities. Visual are the modalities that are perceived

through the visual sensory channel, for example written text or images. In the ontology visual modalities are represented by the class VisualModalities. Auditory modalities are perceived through the auditory sensory channel, for example speech or music and are represented in the ontology by the AuditoryModalities class. Haptic modalities are related to the sensory system of touch. This modality class falls out of the scope of interest of this paper.

An important dimension in the way a particular modality is processed is the time allowed for its processing. Static modalities, for example pictures or static text, allow unlimited time for inspection and processing. In contrast, dynamic modalities (animation, video) are transient and do not allow freedom of perceptual inspection. In MO static modalities are represented by the class StaticModalities and dynamic modalities are represented by the class DynamicModalities.

We further describe an example of how the knowledge modelled by the perception profile can support the process of modality combination by generating multimodal output in accordance with well established principles for cognitive information processing. More concretely, our example demonstrates how to generate multimodal combinations that comply with the cognitive Modality Principle postulated and empirically tested in (Moreno & Mayer, 1999). This principle states that when giving multimedia explanations words should be presented as auditory narration rather than as visual on-screen text. The Modality Principle is based on two important themes from theories of human cognitive processing (Baddley 1992; Chandler & Sweller, 1991; Pavio 1986): (i) the processing capacity (or working memory capacity) of the visual and auditory information-processing channels is limited and (ii) active processing involves selecting relevant visual and verbal information, organizing the material into coherent mental models and integrating between visual and verbal representations as well as existing knowledge from the long-term memory. In accordance with (ii) the combination between visual and verbal information (in MO between linguistic and analogue modalities) is realized based on the complementarity of the features specificity, abstractness and focus (see description of information presentation level). Avoiding cognitive overload (i) will require that the above generated combination is also a combination between visual and auditory modalities (in MO modelled by the perception level). Thus the final combination is calculated to be between a modality belonging to the linguistic and auditory classes, that is, speech, together with analogue visual modality, for example animation. The choice of which analogue modality to use can be subject to applying additional principles or design rules. Generating multimodal output based on the modality cognitive principle makes use jointly of the information presentation and perception levels of the proposed Modality Ontology by applying modality combination rules.

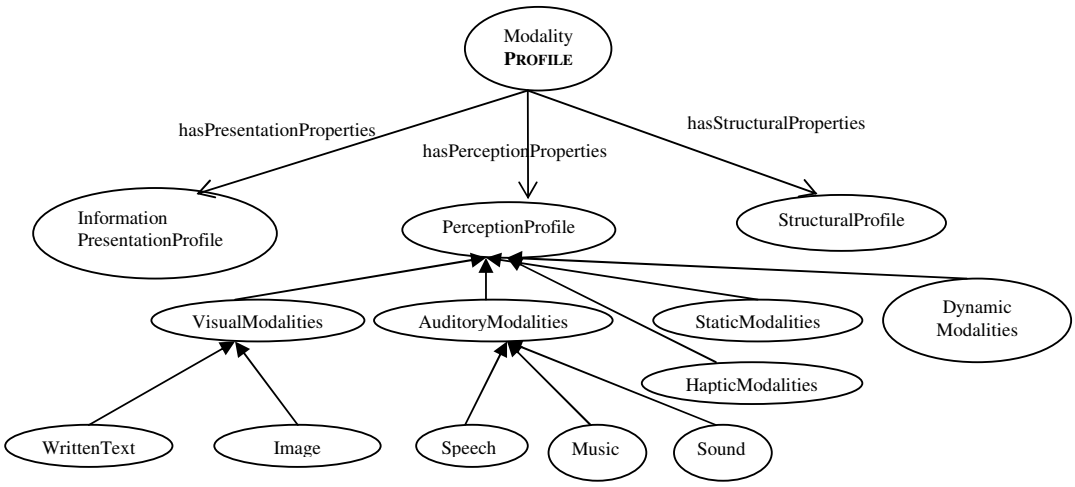

Figure 3: Perception level of the Modality Ontology

2.1.3 Structural level

The structural level models the structural dependencies that can exist between the composite modalities of a multimodal presentation. Structural dependencies form the syntax of multimodal presentations and as such have a direct bearing on the way multimodal messages construct and convey meaning. For an illustration consider the structural dependence of a substrate (background) and the information carried by the modality situated on that substrate. By virtue of being a substrate one modality can determine the interpretation scope and provide the semantic context of the modality which is situated within the substrate. A more concrete example is the combination of an icon on a map substrate. The map used to describe a region of the world possesses an internal structure – points on it correspond to points in the region it charts. When used as a background of an icon, one may indicate the location of the object represented by the icon by placing it in the corresponding location on the map substrate (Arens, Hovy & Vossers, 1993).

Pertinent to the structural level is the distinction between dependent and independent modalities made by (Bernsen, 1994). Independent modalities can do much of their representational work on their own; for example text alone can express almost everything. In contrast, dependent modalities need other modalities to serve representational purposes. Graphs are examples of dependent modalities as they almost always require clear and detailed linguistic annotation for their interpretation. Structural dependencies are important for calculating modality combinations. We have chosen to model these dependencies as properties relating the classes of modalities forming the dependency and not necessarily as part of the structural level. For example, in the ontology the classes Graphs and Labels (see Section 2.2.2 for more in-depth explanation) are related by the inverse properties *annotates* and *areAnnotatedBy*.

2.2 ANALOGUE AND LINGUISTIC MODALITIES

At this level of the ontology we describe which more specific differentiations can be made between modalities in terms of their capacity to represent different types of information. The members of each modality class at this level are characterized not only by the set of features related to that particular level by also, through inheritance, by the set of features characterizing the upper level.

2.2.1 Analogue modalities

Zooming in on the AnalogueModalities class, it comprises of the disjoint classes of Images, Maps, Graphs and Diagrams (see Figure 4). This classification is based on Bernsen's taxonomy of output modalities and Lohse's classification of visual representations (Lohse et al., 1994). The specific characteristics describing the way each of these modalities represents information are members of the classes ImageFeatures, MapFeatures, GraphFeatures and DiagramFeatures respectively. Table 1 presents some of the features characterizing images, graphs, maps and the three types of diagrams – structural, process and conceptual. The features have been selected from existing literature describing the different characteristics and aspects related to the nature of different modalities (Bernsen, 1994; Lohse et al., 1994; Tufte, 1983; Twyman, 1979). This set of features is by no means exhaustive. It is not the aim of this paper to describe such exhaustive set, but just to illustrate the approach we adopt in modeling the knowledge about modalities.

MO represents modalities at levels deeper than the specific attention of this subsection. For example, graphs can be scatterplot, categorical, line, stacked bar, bar, pie, box, fan, response surface, histogram etc. Each of these graph types has specific characteristics which distinguish it as a type on its own. In a fashion similar to the one applied for the aforementioned ontology levels, each graph type is a modality class which is related to the class of properties describing this modality.

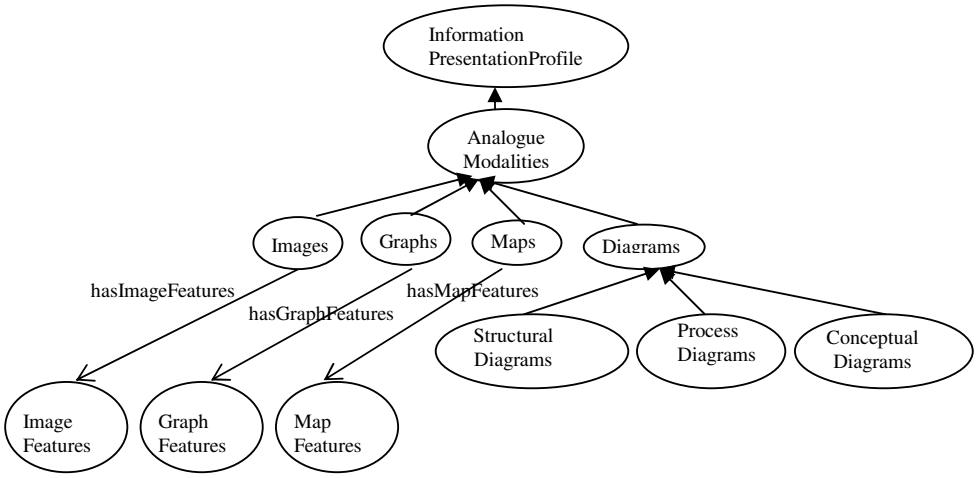

Figure 4: Analogue modalities

Modality	Information presentation related features
Image	- high specificity - full correspondence with the represented object - preserves distance properties of real world space - preserves interval properties of real world space
Map	- represents physical geography - represents location - represents relational structure of objects and events
Graph	- encodes quantitative information - emphasizes the whole display - symbolic (no recognizable similarity to the subject matter or domain of representation) - supports analysis of data information - supports reasoning about data information
Conceptual Diagram	- presents analytical decomposition of an abstract entity - facilitates the perception of structure and relationship
Structural Diagram	- describes a physical object - conveys spatial, nonnumeric, concrete information
Process Diagram	- describes the interrelationship and processes associated with physical objects - the spatial data expresses dynamic, continuous or temporal relationships among the objects

Table 1: Information presentation related features of analogue modalities

2.2.2 Linguistic modalities

At the *linguistic modalities* level the main distinction is between *text, discourse, label* and *notation linguistic modalities* (see Figure 5). The distinction between *text* and *discourse* modalities stems from the different behaviour of written language and spontaneous spoken language. While written language is situation independent, i.e., the recipient and the author of the communication do not need to share the same space, time and situation, spoken language has evolved to serve situated communication. Label and notation modalities are brief expressions of focused information. These features make labels well suited in combinations with modalities that require short textual annotation, for example graphs or conceptual diagrams. Relationships of that kind are directly encoded in the ontology (see Figure 5) and can be used for a straightforward calculation of certain modality combination. In the particular example with graph and label modalities the properties *annotate* and *areAnnotatedBy* are inverse. It is possible to specify that in OWL using owl:inverseOf.[1] *Notations* are for specialist users and their most prominent feature is limited expressiveness.

Similarly to the depicted relations between modalities and their features at the different already described levels of the ontology, all the classes of linguistic modalities are described by their corresponding features. We did not choose to show all the feature classes of linguistic modalities in Figure 5 as our attention is mainly on depicting the new aspects of knowledge about modalities that each level introduces.

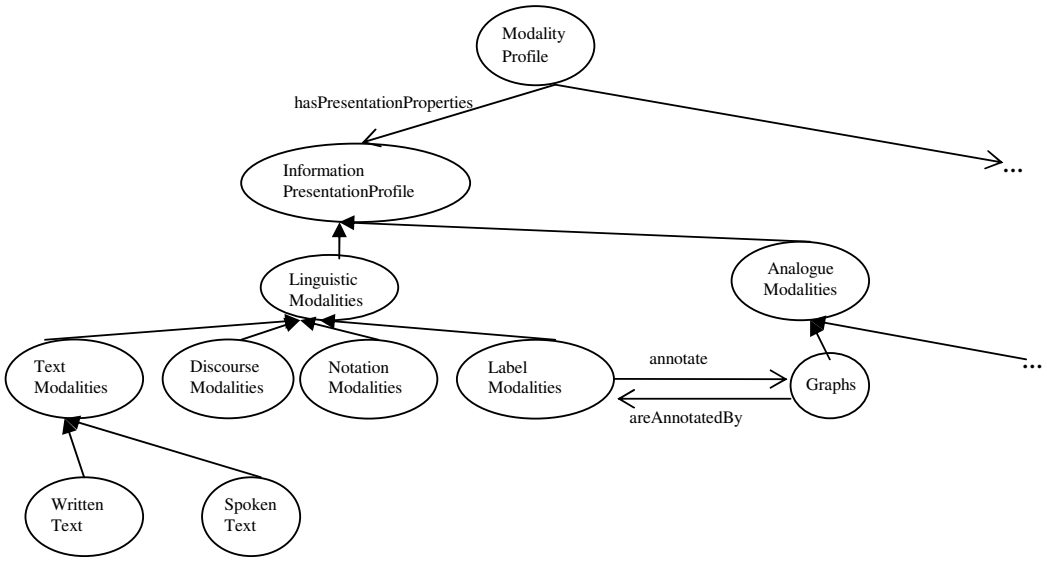

Figure 5: Linguistic Modalities

2.2.3 Information channels

Information channels are an important aspect determining the way modalities convey information. Information channel has been defined as a perceptual aspect (an aspect accessible through human perception) of some medium which can be used to carry information in context (Bernsen, 1994) or an independent dimension of variation of a particular information carrier in a particular substrate (Arens, Hovy & Vossers, 1993). An example of the latter definition would be an icon that can convey information by its shape, color and position and orientation in relation to a substrate map. What Bernsen (2004) and

[1] MO has been implemented in OWL.

Arens et al., (2003) call information channels, Stockl (2004) calls sub-modes, defining them as the building blocks of a core mode's grammar (core modes correspond to the level describing linguistic and analogue modalities in MO). In what follows we describe the way the information channels of typography and colour are modelled by MO. The approach applied for these two information channels can be generalized for the remaining information channels.

Typography is an important aspect in representing written text and can contribute to its meaning beyond the linguistics. We have chosen to model typography at the profile level (see Figure 6) because it is related to the modality form, i.e., one and the same typography can accommodate different contents. In the following subsection we will use this ontological distinction to demonstrate how MO can capture important meaning making relations between content and profile. The class Typography contains all the main constructs that describe typography, such as font type and size, spacing, paragraphing, margins, etc. In the ontology they are modelled as subclasses (Paragraphing, Font, Colour) of the class Typography.

Colour is an information channel that describes not only typography but also images. In order to properly capture all the important features that describe colour we align MO at this level with the MPEG-7 ontology (Hunter, 2005) and more specifically with the MPEG-7 colour visual descriptor. In the MPEG-7 ontology some widely used visual and audio features or properties are represented by a choice of descriptors. The visual properties described are colour, texture, shape and motion while the audio properties are silence, timbre, speech, musical structure and sound effects. The property colour is described by the descriptors (classes in the MPEG-7 ontology) DominantColour, ScalableColour, ColourLayout, ColourStructure, GoFGoPColour (see Figure 6).

We follow the same approach of aligning with MPEG-7 ontology feature descriptors when describing the remaining information channels.

3 RELATIONSHIP BETWEEN CONTENT AND PROFILE

To model the relationship between content and profile we need a proper representation of modality content in addition to the modality profile representation we describe in this paper. Modeling content is not our focus and for that reason we try to make use of already existing frameworks. At the content level we align with the MPEG-7 ontology and more specifically the part that concerns content representation (see Figure 7).

To illustrate the capacity of MO to capture and model meaning that is derived from the relationship between modality content and profile we use an example described in (Stockl, 2004). We have chosen this example because of the necessity it poses on modelling content and profile separately and establishing a connection between the two.

The example is that of an advertisement of the RSPCA (the Royal Society for the Prevention of Cruelty to Animals) for free range eggs where the verbal text is typographically designed to yield the visual form and appearance of a supermarket receipt. The language contained in the receipt is not what we would normally expect to read on a receipt (the bought items and their prices) but the textual message of the advertisement (the appeal to people not to buy battery eggs). In this example the exported typographical repertoire has a semantic impact. The receipt form of the text makes the pivotal point that it is in the supermarket where farming policies are shaped via the price of the eggs and consumer behaviour.

In order to be able to capture or generate such sophisticated interplay between content and form we need to have the proper frameworks to model the two aspects separately as well as their relations.

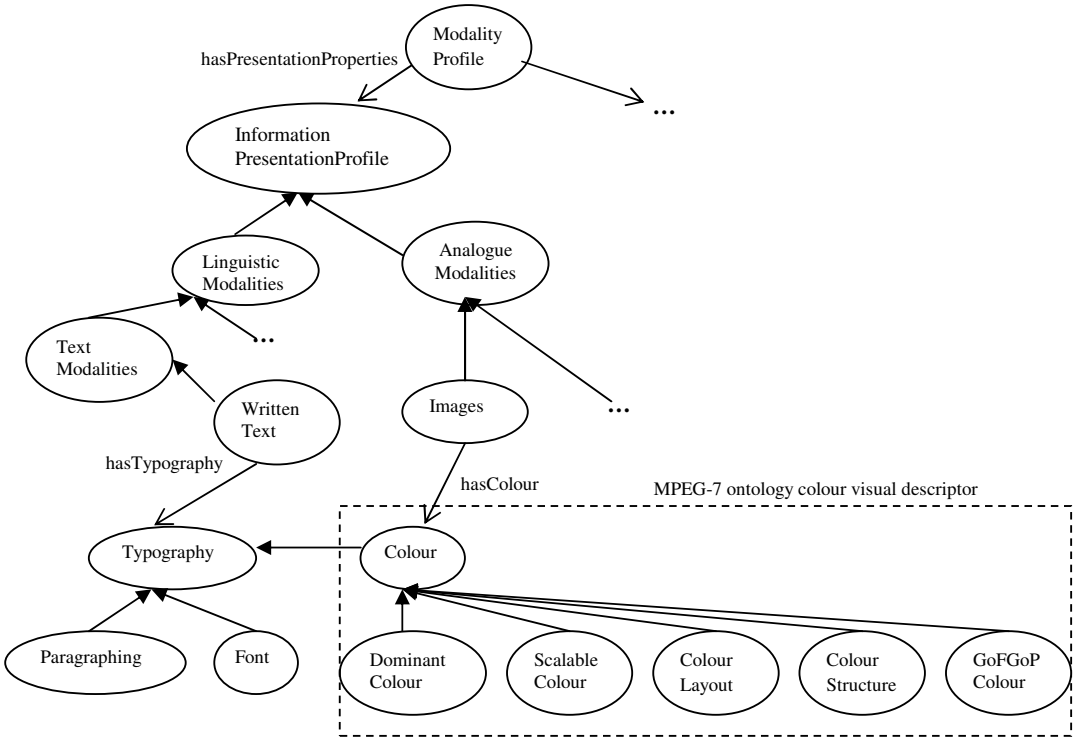

Figure 6: Information channels

Using MO the representation provided by the RSPCA advertisement can be modelled on the content and profile levels. On the profile level we describe the specific features of the typography of a supermarket receipt (the specific type and size of the font, paragraphing, etc.) and relate it to the concept of a supermarket receipt. On the content level the representation of the pair - item and its corresponding price - is also related to the concept of supermarket receipt as this is the information that you normally find on a receipt. The instantiations of the specific content of a receipt and its typography are related by the hasTypography relation. In other words, text which says which items have been bought and in what price has a specific typography – narrow margins marked by lines of three stars each, dotted font typical for cash-desk printer, etc. The text and the typography are both characteristic for supermarket receipts. Now when in this relationship between content and form only the content is changed, in our particular example the content of the receipt is substituted with the advertisement text, the advertisement text appears in the form of a supermarket receipt (the hasTypography relation to the receipt typography stays unchanged). The meaning derived from the new representation is a combination of the meaning derived from the content level (what the text of the advertisement says) and the meaning associated with the specific instantiation of the Typography class, that is a supermarket receipt. In other words content and form (profile) shift and blend and users translate or transpose meaning from one of those two aspects to the other.

.

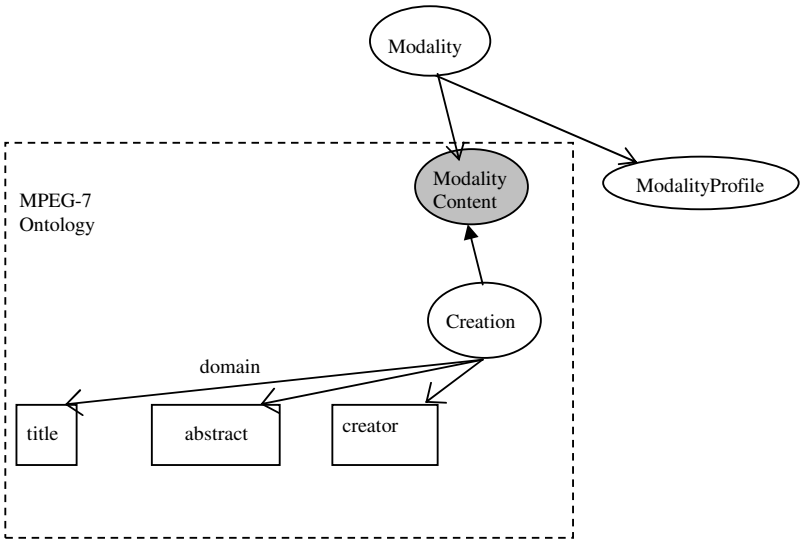

Figure 7: Aligning with the MPEG-7 ontology on the content level

4 CONCLUSION AND FUTURE WORK

The processes of modality allocation and modality combination are knowledge intensive and require proper representation of the knowledge that supports them. The Modality Ontology we propose models that knowledge at three different levels – properties of modalities that determine their capacities to represent different types of information, properties that determine the way each modality is perceived and processed by human cognitive systems and structural dependencies between different modalities. The knowledge described on the first level supports mainly the modality allocation process while the second and the third level are used for calculating modality combinations. MO has the capacity to serve as a unified framework that captures different aspects of knowledge about modalities that have already been modelled for different purposes by different research communities. We have demonstrated a possible alignment with the MPEG-7 ontology.

We are currently developing more robust and generalized methods for modality allocation.

ACKNOWLEDGEMENTS

This work was supported by the ICIS program (http://www.decis.nl/html/icis.html). ICIS is sponsored by the Dutch government under contract BSIK 03024.

REFERENCES

Arens, Y., Hovy, E. & Vossers, M. (1993) On the knowledge underlying multimedia presentations. In *Intelligent Multimedia Interfaces.* Mark Maybury, editor. AAAI Press.

Baddeley, A. (1992). Working memory. *Science, 255.*

Bernsen, N. O. (1994). Foundations of multimodal representations: A taxonomy of representational modalities. *Interacting with Computers*, 6(4):347–371.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8*.

Hunter, J. (2005). Adding multimedia to the Semantic Web - Building and applying an MPEG-7 ontology. *Multimedia Content and the Semantic Web: Standards, Methods and Tools*, Giorgos Stamou and Stefanos Kollias (Editors), Wiley.

Lohse, G., Biolsi, K., Walker, N., & Rueter, H. (1994). A classification of visual representations. *Communications of the ACM*, v.37.

Moreno, R. & Mayer, R.E. (1999). Cognitive principles of multimedia learning: The role of modality and contiguity. *Journal of Educational Psychology*, 91, 358-368.

Paivio, A. (1986). *Mental Representations: A Dual Coding Approach*. Oxford, England: Oxford University Press.

Stenning, K. & Oberlander, J. (1991). Reasoning with words, pictures and calculi: Computation versus justification. In Barwise, J., Gawron, J.M., Plotkin, G., and Tutiya, S. (Eds.). *Situation Theory and Its Applications.* Stanford, CA:CSLI, Vol.2.

Stockl, H. (2004). In between modes. Language and image in printed media. In E.Ventola, C. Charles, and M. Kaltenbacher (Eds.). *Perspectives on Multimodality*. John Benjamins Publishing Company.

Tversky, B., Morrison, J.B., & Betrancourt, M. (2002). Animation: Can it facilitate? *Int. J. Hum.-Comput. Stud.* 57, 4.

Tufte, E.R. (1983). *The Visual Display of Quantitative Information*, Graphics Press, Cheshire, Connecticut.

Twyman, M. (1979). A schema for the study of graphic language. In Kolers, P., Wrolstad, M., and Bourna, H. (Eds.). *Processing of Visible Language* Vol. 1. New York: Plenum Press.