

OntoMedia - Creating an Ontology for Marking Up the Contents of Fiction and Other Media

K. Faith Lawrence, dr monica schraefel

Image, Speech and Intelligent Systems Group
School of Electronics and Computer Science
University of Southampton

1 Introduction

The OntoMedia project began over a cup of tea. If this was a work of fiction instead of an academic paper then the tea might be a metaphor, a literary conceit, an important part of the plot or just a cup of tea. In both cases the statement would be inaccurate. In reality the project began as a convergence of three interests - the creation of metadata to describe the content of online amateur fiction, the identification of events in multimedia objects to seed external applications such as sound effects or automatic music generation and, finally, as a potential way of generating narrative. This paper will focus on the first of these issues.

The Internet has opened up many opportunities in electronic publishing. While the commercial world is weighed down by formats and digital rights management the amateur world has embraced the new means of distribution. There are two types of amateur writing available online - original fiction and fan fiction. Fan fiction, in its simplest form, is amateur fiction written about characters or set in a world that has been previously created by someone else. Both fan fiction and original fiction have been around since the oral tradition held sway and telling the difference between the two isn't always obvious. In its most current form fan fiction is generally agreed to have been around since the 1960s with credit (or possibly blame) most often being given to Star Trek as the catalyst.

The proportion of original to fan fiction on the Internet is impossible to estimate but, while it still remains widely distributed, amateur fiction probably represents one of the largest electronic libraries currently in existence. Unfortunately, one which doesn't come with a catalogue. While the larger archives are fairly fixed in their position, the smaller archives and personal pages are frequently changing address, going down temporarily for maintenance or due to bandwidth limitations or just vanishing. "Can anyone tell me where to find..?" is a frequent question on many lists as is the popular "Can anyone recommend..?" or "I am trying to find a story that contains...". A large part of this ever-changing nature is attributable to the subject matter and the very amateur nature of the enterprise. Complaints, witch-hunts and legal threats can force relocation while lack of resources or loss of interest can remove sites entirely. All of which add additional complication to an already diverse system of differing standards and expectations and vocabularies.

1.1 Related Work

The idea of creating top level conceptual model to act as a bridge for media ontologies and standards is not new. The ABC ontology designed by Lagoze and Hunter was intended to “integrate information from multiple genres of multimedia content within digital libraries and archives[3]. This ontology focused on factual information both within the document and at the level of document creation, provenience and rights management. The strength and weakness of this approach was that it tried to describe everything using the same basic structures and had its roots in factual events outside the document. While the ontology can be used to describe the events that occurred within the document it was “designed to model physical, digital and analogue objects held in libraries, archives, and museums and on the Internet” [5]. Where it goes beyond the physical it focuses on “abstract concepts such as intellectual content and temporal entities such as performance or lifecycle events that happen to an object”[5].

With the OntoMedia model we were concerned not only with the facts but also the possibilities and concepts that existed behind those facts. The ABC model includes the ‘abstraction category’ as a way of expressing ideas which do not exist in the context of a situation but sees this primarily as a way of binding together various manifestations of the same intellectual expression. Conversely the OntoMedia structure deals primarily with abstractions that exist behind the content. While, like the ABC model, our structure was based on the concepts of entities and events we specifically focus on the needs of best describing content of both analogue and multimedia. In this way the OntoMedia ontology can be integrated with the ABC model if required but otherwise provides augmentation and an alternative method of interfacing with the expression of that content while allowing other models to describe the relationship between that expression and the physical media that contains it as well as the bibliographic information that relates to either the expression or the physical media.

2 OntoMedia: Media Content and Reusability

2.1 User Requirements for an Ontology for Fiction and Other Media Content

Due to the amount and nature of the material produced by the amateur writing community and its hobbyist origins the current systems of storage, access and retrieval is immensity variable and frequently contentious. A discussion group was created to allow interaction with some of those involved in this community and to look into the possibility of creating a definitive taxonomy of terminology and meaning. This taxonomy would have acted as the knowledge base from which a tailored ontology could be created. It was quickly discovered that not only were many of the definitions vague but most were fluid and gained specific significance or meaning within one community that was lost or transformed in another. Only interaction within any of the communities allowed the user to pick up the most common meaning for the term within that group. This could be very confusing for those people coming into the group either from another community or from outside. One of the most striking examples of this was the use of the word “gen to describe a work of fiction. Depending on the context it could either be intended to mean a story with no romantic interests, one in which the romantic interests

are of a level with and the same as the source material which inspired the piece or very occasionally a story which may contain any non-explicit heterosexual relationship. Thus if the source material included a specific relationship or a high level of sexual explicitness it could still be marked as “gen - a potentially unpleasant surprise for a reader who was expecting only platonic feelings and actions between the characters.

Because of the shifting vocabulary it became clear that creating meta data to a level of detail that would engender a semantic application would require that data to exist at the level beneath that of the community specific concepts. This would also solve the problem of different vocabularies existing within the fan fiction section of the amateur writing community and between that section of the domain and the rest of it. While this had the potential of requiring more detail than would otherwise be needed it also offers the possibility of user-defined and personalized lexicons of terms thus increasing the usability of the system and lowering the level of expertise that would be required of a new user. Creating an event driven system proved an obvious method to implement this.

2.2 The OntoMedia Ontology

The OntoMedia model is event based. Since events have to happen to something the top level of our ontology is divided into Events and Entities (see fig 1). We define an Entity as an object or concept. An Event describes an interaction between one or more Entities during which zero or more attributes of those entities are modified or a new entity is created. A number of Events subclasses are defined based on the most common types of events found in literature. The subclasses of the Entity construct fall into three different types. Those related to objects both physical and abstract (`ontomedia:Characters` and `ontomedia:Items`), those related to spatial models (`ontomedia:Space`) and those relating to time (`ontomedia:Timeline` and `ontomedia:Occurrence`).

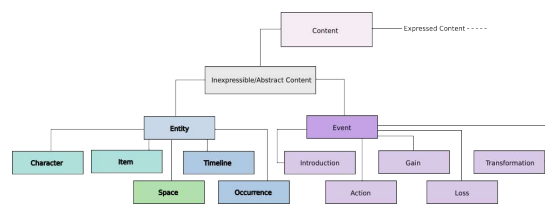


Fig. 1. The Top Levels of the OntoMedia Ontology Structure

Entities of the `ontomedia:character` are defined as having a personality. For example a toaster would be a type of `ontomedia:Physical_Item`, a subclass of `ontomedia:Item`, whereas the talkie toaster from the television show *Red Dwarf* and the related books would be classed as an `ontomedia:Character` despite having the physical appearance of a toaster. The class `ontomedia:Item` is

subclassed into physical and abstract items. An `ontomedia:Occurrence` is a specific instance of an event which occurs within a single `ontomedia:Timeline`. The reason for this is that any `ontomedia:Inexpressible_Content` can have a `ontomedia:Timeline` associated with it. It is therefore likely that more than one time line instance will need to be defined for any given work. One event will therefore occur on multiple time lines but the relationship between the events on any given time line will not necessarily be the same as on any other time line. For example if we define one time line to describe the events which occur in the narrative and another to describe the events chronologically these two will differ whenever we encounter a 'flashback' or one of the characters mentions either a historical event or something from their past. This flexibility regarding time is especially important since stories may involve time travel and related paradoxes. While a 1:1 mapping exists between an instance of `ontomedia:Timeline` and an instance of `ontomedia:Occurrence`, a 1:Many relationship exists between any instance of an `ontomedia:Event` and the instances of `ontomedia:Occurrence` that contextualise it. This can even allow more than one occurrence of the same event to exist on the same time line for example if a character meets their future self on their personal time line this event will occur twice, once when then are the younger version of themselves and once when they are the older version.

Time travel was not the only occasion when characters have been known to meet themselves. The idea of multiple universes or realities is a staple in science fiction and fantasy. The `ontomedia:Context` class, a subclass of `ontomedia:Abstract_Item`, was created to separate the many different versions of the same entity that may exist. It allows us to differentiate both between different representations of the same fictional character, different version of the same character and between real people, fictionalisations of real people and fictional characters. This is a particular issue when considering the contents of fiction, especially when those works have been reinterpreted across media, within the same work or after a period of time. Since these different interpretations can be physically distinctive, for example when a character is portrayed by different actors, or given different personality traits or back history it becomes necessary to recognise that while they may be supposed to be to same entity there are occasions when there differences are as important as their similarities. Examples of this can be seen in almost every movie adaptation of a book. For example the recent Lord of the Rings movie or the transformation of 'We Can Remember It For You Wholesale' to the movie 'Total Recall'. In the first case the character of Faramir as portrayed in the movie by David Wenham was both physically and emotionally different to the character described in the book much to the disappointment of many fans. In the second the hero of the short story 'Douglas Quail' becomes 'Douglas Quaid'. An even more extreme case can be seen in the new Battlestar Galactica mini-series in which two of the characters (Starbuck and Boomer) have changed gender since the original series aired. Interaction with users showed that they frequently distinguished between these different representations as well as creating their own. Since the separation was important to them it was also important to be able to model that distinction and allow the meta date to reflect it for the purposes of reducing ambiguity in search and retrieval.

With the aim of modeling the contents of fiction, whatever media format it is presented in, we faced the potentiality of having to deal with anything that the human mind can come up with. The only way to deal with this lack of limitations is to plan for it. By making the ontology modular and expendable we leave the option open for those situations that arise which we had not envisioned. Beyond acknowledging that we are not going to be able to cover every situation, supporting extensibility allows us to reuse existing ontologies where they already exist. For example we extended the location ontology created for the Signage project [6] to provide a basic spacial model. We chose this ontology because it had a level of detail that matched what we envisioned for the rest of OntoMedia. This choice worked well in the examples that we created. However should the need have arisen for a different type of spacial representation then a different ontology could have been used instead.

Work is currently under way on modularising the ontology even further to allow greater integration with other ontologies. We are also in the process of creating a series of examples of different types. Basic queries have been run on these examples to demonstrate such things as identifying main characters by looking at who is involved in the most actions or finding which characters have secrets by comparing observed, portrayed and standard traits. The creation of examples is allowing us to check the validity of our model while at the same time we are returning back to the previously identified user needs to confirm that we can meet them. This includes comparing the results from the amateur writing access and usability questionnaire which relate to the types of information that readers wish to be told and checking that against what we can currently describe.

2.3 Examples

To test the robustness of the ontology a number of examples have been created. The examples are of two types, what we consider 'top level' fandom information and story narrative. For the second of these we chose to create the meta data for 'We Can Remember it for you Wholesale' by Philip K. Dick [1]. This short story was selected because it contained a number of interesting constructions such as false memories while at the same time being surprisingly simple in it's basic narrative path. Through this exercise we were able to show that we could carry out a number of detailed queries on the meta data such as identifying the main characters through analysis of which characters were involved in the most events and we could display the events according to both chronological and narrative order. The other type of example that was created was not concerned with the structure of any particular story but instead modelled the basic elements that existed in the domain of a specific intellectual creation. These examples included the domains of the J.R.R. Tolkien's Middle Earth, MASH, Highlander and an original universe. This was done to test the more complex aspects of entity creation and the use of contexts since each of the mentioned examples contains overlapping but distinct versions of the entities and events that they contain. As well as confirming the the ontology could handle this sort of continuity issues the examples also laid the groundwork for the defined entities and events to be reused if an in depth mark up of one or more of the works or works containing one or more of the elements was desired. By allowing a definitive version of commonly used elements to be created referential

integrity can be assisted since all the versions of a specific object will point back to the same definition. From an usability perspective we intended to investigate if such a system will aid users by reducing the time needed to create the meta data for a given work.

3 Future Work

3.1 Integratation with the FicNet Project

The FicNet project aims to create an integrated semantic system with a wide area of appeal in a non-academic field and to make that system usable by the average Internet user so that the advantages of the system outweigh any effort required to participate. The OntoMedia ontology is one of three related projects that support this creation. The other required aspects are a way of describing the users and a way of creating and using a web of trust to integrate standing in the community with information access. The Fan Online Persona (FOP) ontology has been created to support these two requirements and future work will include combining the the two types of meta data to allow inferences to be made across them.

Having created a workable ontology research will investigate ways to facilitate the addition of meta-data to amateur fiction. While personal data is comparatively trivial the marking up of the stories, illustrations and videos that are created is anything but. Applications such as SMORE from the University of Maryland assist in the association of meta data to on line content [4]. The intention is to go beyond this type of tool so that the average user does not need to deal with the base ontology. Due to the reuse of entities and event types it is postulated that one way to aid meta-data creation within our case study is with the generation of reusable objects which can then be connected with a story or through a simple drag and drop interface. While this requires the initial time spent creating the objects, subsequent time will be saved since only the differences from the generic form or other individualities will be needed to be expressed. Investigation will need to be undertaken on the relative merits of when and how much of this information should be distributed for optimum performance.

It is expected that a series of interface designs will be created to facilitate the creation of both FOP files and OntoMedia based meta data files. Examples of both types of files will also be created to confirm the expected linkages between the meta data and how they could be used. This will result in a test system which will clarify where, if anywhere, the technology is not yet ready to support such a system and where further research needs to be done.

3.2 Beyond FicNet

One of the interesting aspects of this project is how it can be taken beyond its initial objectives. The OntoMedia ontology was designed by a multidisciplinary team who were interested in using the resulting meta data for different purposes. While the development of the ontology was driven by the needs of those involved there was an ongoing acknowledgement that the type of information that would be made available could be used in a number of other ways.

Narrative Study and Automatic Classification One of the first possibility raised for consideration was for automatic classification. Biographic records often include some indication of genre. Research is currently ongoing into automatic classification through natural language systems. A similar system to be implemented using the events that occur during a work of fiction. The experiment would be twofold. Firstly to see whether this is a valid method of sorting and secondly to consider what is actually described by a genre such as 'drama' or 'science fiction'. Such a use would not be limited to fiction since it could be applied to anything that could be modeled using the OntoMedia Ontology. It is also possible to take the ontology beyond fiction into history with the possibility of using it to describe news events and analyse these.

Myths, Legends and Cultural Heritage Further work is also intended to map the OntoMedia structure to other related semantic models. The CIDOC Conceptual Reference Model (CRM) was created as a semantic approach to integrated access[2, p.2] for cultural heritage data. By providing a conceptual basis that can be used for automated mapping the CIDOC CRM acts as a bridging technology between existing data structures and a guide to creating new structures. The similarity between cultural heritage and fictional content is one of underling components. Both are concerned with people and events, the only difference being the type of evidence that exists of those manifestations and where they were believed to have taken place. The fictional aspect of the historical narrative has been commented on by people as diverse as Plato and Churchill. The historical aspect of fictional narrative is more often the aegis of fanatical devotees of a particular work. However, ignoring the fact that it is an imaginary history, works of fiction still tell a history.

The possibility of mapping the OntoMedia ontology to the CIDOC CRM system opens up a number of possibilities in the heritage field. Work is already being done to harmonize the CIDOC CRM and the Functional Requirements for Bibliographic Record (FRBR) conceptual model that was created by the International Federation of Library Associations and Institutions [7] however this mapping still exists at the bibliographic object level rather than the content level. One can search within a specific genre or for a known author but not on a type of action that is contained or a particular character. The addition of this type of meta data to cultural heritage texts, especially when linked by the CIDOC CRM to non-literary artefacts or information could expose previously unnoticed links and provide a new tool for researchers.

Since beyond the defined context of the 'real world', the OntoMedia model does not distinguish between real and fictional it can be used to annotate and describe the mythological histories that are exist in both written and oral tradition. Comparative mythology would just be the first discipline to benefit from being able search for myths of a particular type or ones in which certain events occurred. This information can then be combined with external sources such as historical events and trade patterns to provide an integrated view of the conditions that existed in the period under investigation and the influences that would have existed from and between the social, economic and religious elements.

Bibliography

- [1] Philip K. Dick. *The Preserving Machine*, chapter We Can Remember It for You Wholesale. Ace Books, 1969.
- [2] Martin Doerr, Jane Hunter, and Carl Lagoze. Towards a core ontology for information integration. *Journal of Digital Information*, 4(1), April 2003.
- [3] Jane Hunter. Enhancing the semantic interoperability of multimedia through a core ontology. *IEEE Transactions on Circuits and Systems for Video Technology*, January 2003.
- [4] Aditya Kalyanpur, Bijan Parsia, James Hendler, and Jennifer Golbeck. SMORE - semantic markup, ontology, and RDF editor.
- [5] Carl Lagoze and Jane Hunter. The abc ontology and model. *Journal of Digital Information*, 2(2), November 2001.
- [6] Ian Millard, David De Roure, and Nigel Shadbolt. The use of ontologies in contextually aware environments. In Jadwiga Indulska and David De Roure, editors, *Proceedings of First International Workshop on Advanced Context Modelling, Reasoning And Management*, pages 42–47, 2004.
- [7] K. G. Saur. Functional requirements of bibliographic records: final report. Technical report, IFLA Study Group on the Functional Requirements of Bibliographic Records, München, 1998.