Malcolm Clark, Ian Ruthven and Patrik O'Brian Holt

The Evolution of Genre in Wikipedia

This paper presents an overview of the ways in which genres, or structural forms, develop in a community of practice, in this case, Wikipedia. Firstly, we collected data by performing a small search task in the Wikipedia search engine (powered by Lucene) to locate articles related to global car manufacturers, for example, British Leyland, Ferrari and General Motors. We also searched for typical biographical articles about notable people, such as Spike Milligan, Alex Ferguson, Nelson Mandela and Karl Marx. An examination of the data thus obtained revealed that these articles have particular forms and that some genres connect to each other and evolve, merge and overlap. We then looked at the ways in which the purpose and form of a biographical article have evolved over six years within this community. We concluded the work with a discussion on the usefulness of Wikipedia as a vehicle for such genre investigations. This small analysis has allowed us to start generating a number of detailed research questions as to how forms may act as descriptors of genre and to discuss plans for experimental work aimed at answering these questions.

1 Introduction

The research reported and discussed in this paper combines information retrieval (IR), cognitive science and genre, merging and utilizing these for one particular purpose: to analyze how texts are used in different contexts with the final goal of retrieving structured texts. The main goals of effective IR are the identification of users' information needs and the evaluation of the results by creating IR applications that can discern better matches between users' information needs and the available documents (Clark, 2005). According to Ingwersen and Jarvelin (2005), IR is divided into computer science lab experiments versus 'user-orientated' social studies. Our approach is concerned with the latter and forms part of a wider human context to examine the ways in which the framework of a community of practice (CoP) (Wenger, 2000) gives rise to standardized information forms. The evolution of genre is an important part of this research and this paper describes the results of a preliminary study on genre development in Wikipedia In the recent past, the IR community, such as the text retrieval conference (TREC), and more recently, the initiative for the evaluation of XML retrieval (INEX) (Lalmas et al., 2004) have started to understand the importance of (technologically) structured text retrieval but up to now have largely overlooked two important concepts: naturally occurring structures called *genres* and the human perceptual processes which are used to identify and employ them. Genre has been discussed for centuries, most notably by Plato and, of course, by Aristotle, in his work on substance and form. Of course,

there is much more substance to Aristotle's doctrines but for this work, we plan to look at the ways in which humans extract the form which determines the nature of the observed object, in this case, Wikipedia articles or Electronic mail (Clark et al., 2008).

Genre (or kind), is used to differentiate between differing types of texts (especially in classical literature studies) such as reports, novels, poems, memoranda and so on. Although the form and function issue is central to genre theory, some theorists focus on the style, function, form and/or content of genre to distinguish between the 'kinds'. The aim of our discussion and research is to investigate genres and, particularly, how they evolve within Wikipedia. The Wikipedia Encyclopaedia, which first appeared in 2001, is growing and evolving day by day and has articles in more than 250 languages. Currently, the English version alone consists of more than 2.5 million articles and has more than 8 million registered editors (Ehmann et al., 2008). Only a small amount of genre analysis research utilising Wikipedia has as yet been carried out, but as Emigh and Herring (2005) pointed out, Wikipedia can offer an extraordinary insight into how a community can democratically participate in creating forms or genres to show the meaning of an article. Further to this, the work carried out by Collins et al. (2001) showed how there tend to be socially constructed communicative behaviours, namely genres, which emerge to improve the efficiency of the activities in a CoP. The purpose of this paper is to describe an initial study of these behaviours and the evolution of some articles in Wikipedia (English version only), in which classic forms of genres are found, such as Biographies. Some other types of 'new' structured genres, mainly defined by form and content, are also continuously evolving in the Wikipedia community. The question also arises, however, as to whether Wikipedia editors interact, discuss, debate and jointly learn? Does the community consist of the vital characteristics of a CoP, namely, "The Domain, The Community and The Practice," described by Wenger (see Section 2.2)? Our questions for this initial feasibility study were:

1. Is Wikipedia, as a CoP, a suitable vehicle for demonstrating the evolution/development of genre?

2. Are Wikipedia articles consistently composed of a combination of purpose and form?

3. What are the constituent parts of the CoP in the Wikipedia domain?

4. How does a classic genre, such as Biography, evolve in this community? Are there any possible new genres?

Section 2 begins with an introduction to genre, ecologies and CoPs. The third section examines the methodology for this study, the presence of Wikipedia genres by showing the results of a small search of genres and by mapping the genres. In part 3.3 there is a case study to take a closer look at the ways in which a biographical article has evolved since 2001. The conclusions drawn from the research and the plans for future work are presented in section four.

2 Genre, Ecologies and Communities of Practice

2.1 Genre

Genres have been around as an idea for thousands of years. Early examples can be found in the context of Plato's "ideas, forms or reality" and Aristotle's "rhetoric and poetics" (Aristotle, 1984). Aristotle disregarded Plato's musings on 'reality': he considered that whatever was perceivable by the individual was reality. He believed that the entire visual array was made up of *substance* and, most importantly for this research, *form*. Form was knowable, "which specified the individual and which could be abstracted from the objects in a process of perception. External objects impinged upon the senses, and due to the power of reason, the mind was able to extract the essence (or form), which determined the nature of the observed thing." (Breure, 2001). Contemporary authors writing on genre have continued with this theme, for example, Dewdney et al. (2001) refer to Substance and Form in their work. In the seminal book, 'Genre and the New Rhetoric',describe two prominent schools of thought based in different hemispheres: The North American School (heavily influenced by Miller 1984) and The Sydney School (heavily influenced by Halliday 1973, Halliday 1978, Kress and Threadgold 1988 and Martin 1999). In spite of the intrinsic differences between the two schools, some similarities can also be observed: they both acknowledge the superiority of the social in understanding genres and the role of context; in addition, they highlight the value of community or social factors. However, they do differ in other respects. The Sydney School focuses on the textual features in terms of linguistic analysis that stresses the static characteristics and rigid qualities. In contrast, the North American School emphasises the dynamic nature of genres, with the cornerstone of the theory based on interplay and interaction, and in particular, on the intricate associations between context and text. Both of these schools have implications for this work: not only are the textual features vital, but also the interaction and interplay of genres. There are also many genres that are of a static or dynamic nature.

Any thorough literature review of works on genre will reveal a general lack of consensus on the question of finding an appropriate definition for genre because so many questions remain unanswered as to how genres function, overlap and interact with each other, which rules and patterns constitute a genre and how these characteristics are perceived. We argue that the backgrounds of researchers influence the way they define genre, as Kwasnic and Crowston (2005) point out, the researcher chooses the definition appropriate to the current investigation. That said, there are significant similarities between scholars: compare, for example, Berkenkotter and Huckin (1995) *Situatedness* and *Duality of structure* with Yates and Orlikowski's *Genres of Organizational Communication*(1992). As Kwasnic and Crowston (2005) explain, the many definitions of genre and lack of agreement are not due to slipshod attitudes or lack of effort, but are rather indicative of the diversity of genre.

As Breure (2001) states: in most contemporary genre analysis, content and form are supplemented by purpose and function. In the context of this paper, it is the set

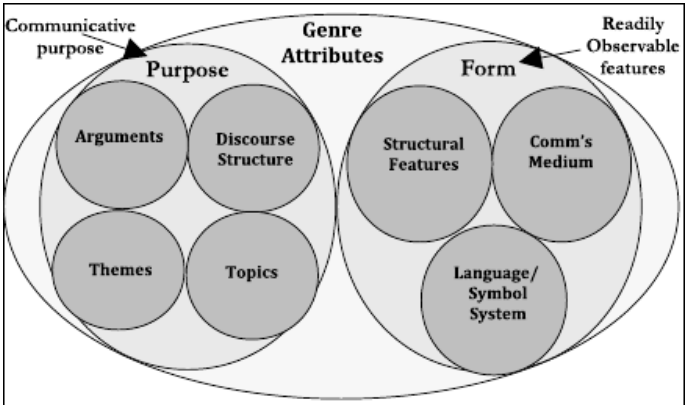

Figure 1: Orlikowski and Yates (1994) devised from the definition and attributes of purpose and form being used for this paper.

of structures and layout that show the user the documents' purpose (substance) and form through its structure, regardless of the topical nature of the writing. Figure 1 illustrates the definitions of purpose and form provided by Yates and Orlikowski (1992, 301-303), Orlikowski and Yates (1994, 544-545) and Yates et al. (1997, 550-551) that was influenced by Giddens (1986) structuration theory. Form, in the context of this project, simply refers to the easily perceptible features of the communication, such as those found in calls for papers, which include:

1. **Structural Features**: text formatting devices such as lists and headings, and devices for structuring interactions at meetings, such as agenda and chairpersons.

2. **Communication Medium**: pen and paper, telephone, or face to face.

3. **Language or Symbol System:** linguistic characteristics, such as the level of formality and the specialized vocabulary of corporate or professional jargon.

The purpose of the genre refers to the communicative purpose, in particular the social motives, themes and topical nature assembled and perceived in the communicative genre, for example, the purpose of a shareholders' meeting is to present the company's past accomplishments and future outlook to stockholders, or the purpose of a curriculum vitae is to summarise an individual's educational and employment history for a potential employer[1]. This particular technique defined by Orlikowski and Yates, and used by Emigh and Herring (2005) analysed genre by looking at the common and shared purpose to typical aspects of substance and form that are particularly useful for this small

[1] A full overview of genre key issues and definitions can be found in Boudouride's excellent and thorough literature review (Boudourides, 2001)

feasibility study of the Wikipedia domain. First of all, however, it would be helpful to look at Ecologies, which perfectly describe how these texts evolve and are modified in this domain.

2.2 Ecologies

Duff (2000, 4) pointed out that due to the existence of biological metaphors in genre theory (Erickson, 2000; Kwasnic and Crowston, 2005), it was only natural that the evolutionary paradigms found in Darwin's "Origin of Species" (Darwin, 1859) would be used to model the ways in which literary forms change over time by evolving, being modified and being replaced. Duff (2000, xii) also suggested that some genre theorists also extend the biological metaphor in "quasi-Darwinian terms" by describing some of the mechanisms of literary evolution as "the competition of genre", genres struggling for survival, their "fitness" for an environment and the "possibility of extinction" but this could be criticised for extending the metaphor too far. Kwasnic and Crowston (2005, 87) gave an impressive description of how the genres behave when they extended Erickson's genre "ecology" metaphor (Erickson, 2000). They compared a genre to an organism in an ecological community: they all rely on other organisms for their effectiveness, have an effect on each other, evolve over an unspecified course of time at different paces, and can even replace each other, i.e. memo-genre. They declared that these ecological habitats are CoPs (see Section 2.2), Wikipedia, in the context of this paper. As is the case in most areas of research, however, there are issues with Web genres that have to be considered when studying digital media such as Wikipedia. Kwasnic and Crowston (2005, 87) described these issues and how the problems arise in a genre ecology by explaining two phenomena which occur more or less concurrently: firstly, traditional genres appearing on the Web and, secondly, the appearance of new unique genres appearing on the WWW. Both of these phenomena have genres that divide, merge, transform and evolve. This is an important implementation issue that has to be taken into consideration because the genres have to be identifiable by all systems and perceptible to all users.

2.3 Communities of Practice

Wenger (2000) stated that CoPs are social institutions or sites where human agents draw on genre rules to engage in organizational communication which operate by producing, reproducing, or modifying whatever they are producing (in this case, genres). (Yates and Orlikowski, 1992, 305) stated: "In structurational terms, genres are social institutions that are producing, reproducing, or modifying when human agents draw on genre rules to engage in organizational communication". If the behaviour of the community could be comprehended, this could be exploited in the implementation of skimming and categorization tools that would provide search and retrieval of important community objects. Further to this, Collins et al. (2001) explained that what the community sees as important will be reflected in the implicit structures found in the objects they create and share and as Watt (2009) has observed "convergence on a set of standardised

document structures is both natural and helpful". These objects are genres that occur in the web; CoPs are utilised, but we need to look at the ways in which these web pages are structured in Wikipedia and the types of features of which they consist. Wenger (2000) described what he considered to be the characteristics of a CoP as:

The community: In pursuing their interest in their domain, members engage in joint activities and discussions, help each other, and share information. They build relationships that enable them to learn from each other. A website in itself is not a CoP. Having the same job or the same title does not make for a CoP unless members interact and learn together.

The Domain: A CoP is not merely a club of friends or a network of connections between people. It has an identity defined by a shared domain of interest. Membership therefore implies a commitment to the domain, and therefore a shared competence that distinguishes members from other people.

The Practice: They develop a shared repertoire of resources: experiences, stories, tools, ways of addressing recurring problems-in short a shared practice. This takes time and sustained interaction.

3 Evolution of Wikipedia Genres

Wikipedia is an important and popular domain for accessing information about a huge range of information. Not only do individuals use it for reference, but many organisations, such as the BBC News, use it for information. However, Wikipedia does have its detractors, who criticise it for being inaccurate; it suffers from vandalism, of course, which is carried out sometimes with malicious intent, but also sometimes just to raise a laugh. The Now Show (British comedy program) on BBC Radio 4 has even used Wikipedia for some of its sketch material.At a higher level there are many types of offshoots of Wikipedia such as WikiBooks (Cookbooks, StudyGuide etc,), Wikizine, Portals etc. However, this study concentrates on the evolving types in 'Wikipedia The Free Encyclopaedia'. This Wikipedia operates in an editorial hierarchy of: all, users, Autoconfirmed users, Bots, Administrators, Bureaucrats, Checkusers, Stewards and Board Vote Administrators with least permissible editing powers being assigned to "all" and "users" and the most 'power' to "Stewards" and "Board Vote Administrators". For example, once an edit is submitted 'live' by a least empowered editor, a modification is accepted/rejected by Stewards et al. The full hierarchy and list of responsibilities is published in Wikipedia but will not be listed here. Wikipedia contributors are allowed to edit each page and are given a toolbox of HTML functions to use for text formatting, linking files, adding photographs, inserting tables and so on. Much like Kwasnic and Crowston (2005) describe traditional genres are appearing on the web. The Wikipedia community, we believe, contains a wide array of such types, such as FAQ, lists for example list of films etc, Reviews, Guides, News Articles, Events and so on. Not only that, new unique genres also appear, transform and evolve, much like

Kwasnic and Crowston (2005) pointed out. Section's 3.1-3.3 will be used to examine how some of these structural forms (or genres) evolve. It could be argued that Wikipedia (encyclopaedic) itself could be called a genre in its own right but for this study, we look at the articles (maybe sub-genres?) of which the content and form are constantly evolving as a result of editors employing certain devices or tools, such as formatting text, lists, tables and photographs and also studying multiple sources, such as biographical books, for amending and adding factual content. Underlying each article in Wikipedia there is also a discussion area (or aka Talk Pages) between users that re-enforces our potential understanding of the whole CoP aspect of this domain. For example, much of the current discussion about General Motors Corporation (see Figure 3) is the likelihood of its demise in the current financial crisis and debate about what content to include. The Wikipedia site says the purpose of the talk pages is to provide areas for editors to discuss changes to the linked article or project page. Also provided is a history from when the article was first created to the present day as each amendment no matter how big or small is recorded. This small study is overall being used to examine the suitability of Wikipedia for our study into structural forms and how structure is perceived and used by purpose and form. Our overall goals, at this stage, are to examine the suitability of Wikipedia and its constituent parts (discussion etc) as a vehicle for demonstrating the CoP and evolutionary paradigm in this context in which we have devised a methodology (3.1 below). We have chosen to look at the evolution of several possible new and old types of structured articles (see Figures 2, 3, 4 and 5) such as discographies, lists musical groups/bands, footballers etc as well as conduct a small case study of how a Wikipedia biographical article such as Spike Milligan evolves.

3.1 Methodology

The methodology for this study consists of several parts which tie in with the Ecology, CoP and the Orlikowski and Yates (1994, 544-545) definition of purpose and form.

1. Search: REM, Margaret Thatcher, General Motors and Alex Ferguson of Manchester United Football Club etc
2. Examine the potential genres by purpose and form.
3. Look at how the articles are constructed and note if they lead to any other types of structure (Kwasnic and Crowston, 2005) such as discography, FAQ, Biography, List and so on. Look at the articles, noting in particular whether:
    a) They are traditional types of genre such as Biography.
    b) The article is a NEW style of genre.
    c) Examine the underlying CoP to see whether the discussions (in articles) indicate the expected characteristics indicated (Wenger, 2000).

3.2 Search and Record Genres

The Wikipedia articles were first perceived for their potential usefulness during the relevance judgements ('paper' exercise) for the INEX in 2006 (Huang et al., 2006). While examining the topical relevance of organisations' submissions during the relevance judgements' phase of INEX 2006, it was noticed that particular structures or genres were starting to appear throughout. This showed that Wikipedia would be a potentially suitable vehicle for studying the evolution or development of genre in a CoP and also for studying highly visual types of text with perceivable purpose and form. After the search by subject most of the important types of articles linked to the main articles were mapped, recorded and analysed. As all the genres could not be mapped out due to space issues, they have been narrowed down to internal categories such as: biographies, lists, football clubs, motor vehicle manufacturers, and political parties which have their own particular purpose and form. Conducting the search enabled the recording of the relevant statistics, purpose and form attributes that are shown in Table 2.

A popular area in modern culture is, of course, music such as rock and pop. While searching for rock music, it was noticed that there was a hierarchy of genres which are connected to musical groups such as REM, Muse, etc (Figure 2) which link to other types of genres such as discography, biography, musical group, several types of lists list of bands under the same record label, chronological list of Rock and Roll Hall of Fame inductees which is in two forms. One list [2] has a large table with the band information containing the year order, name, image of artist and year inducted and the second list type is in alphabetical order (Table 2 has more information).

Figure 2 shows that there are some already existing web genres in Wikipedia such as list and index but also new ways of structuring information. The Musical Group, Band member and Discography contains a layout consisting of lists and tables but some titles also show up consistently throughout different examples of Musical Groups (U2, Muse, etc), Discographies and Band Members. It is also clear that, similarly to the evolutionary paradigms in section 2.1, some of these literary forms are evolving, being modified and being replaced. Some of the existing genres are actually evolving and outliving their usefulness and in some circumstances leading to a new type, for example, the histories of the articles for rock bands REM and U2. Three years after the original articles had appeared, they seemed to have become too big and thus seemed to have outlived their usefulness. The editors created other articles, such as discographies[3], to help contain the information, leaving the textual information laid out helpfully for the readers who were then able to filter to the content they would most need. This is particularly helpful in an information search task. As can be seen in Figure 3 and Table 2 Automobile Manufacturers, such as General Motors and Ford, had several different types of articles linked to the main result.

At the top of the hierarchy, the Automobile Manufacturers could be categorised as an Organisation (for example, British Petroleum and GM Corporation nearly have similar

[2] http://en.wikipedia.org/wiki/List_of_Rock_and_Roll_Hall_of_Fame_inductees

[3] http://en.wikipedia.org/w/index.php?title=Talk:R.E.M._discography\&oldid=94780788

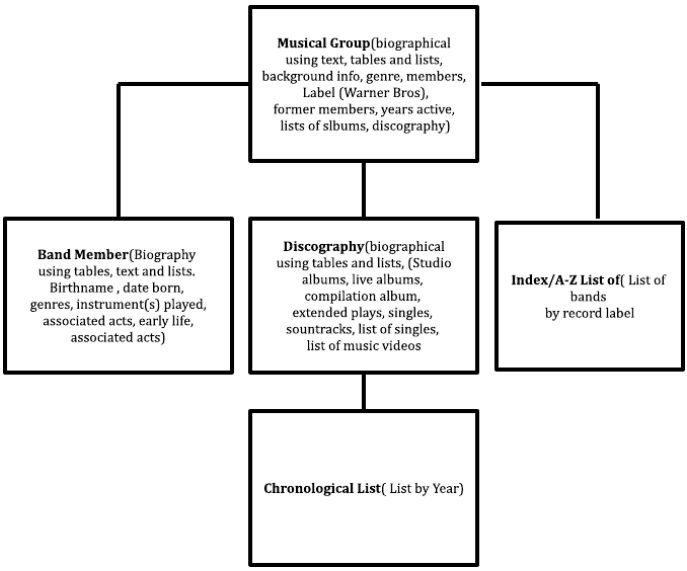

Figure 2: Band/Music Group example (see REM for a good example): visual format from Wikipedia.

structures) and display a particular structural form that allows the perceiver/reader to understand and find quickly the appropriate content pertaining to the organisation or Automobile Manufacturer. By emphasising the most important information related to each article, for example, in figure 3 (also Table 2) the community of editors has decided that the most important information defining an Organisation (such as General Motors) are what you see in the boxes above (as well as an image of the Organisation logo). This information is heavily formatted due to its prominence in the article whereas the rest of the article is composed only of text and citations of a biographical nature that elaborate on this information. Wikipedia has many articles on particular organisations in the automobile industry, such as GM Corporation, British Leyland and Ferrari (Figure 3). In the next level of the hierarchy, the first two organisations are more famed for producing consumer or family cars whilst the latter, Ferrari, produces Formula 1 or SuperCars (Figure 3). The SuperCar and Family Car have their own individual forms, but occasionally overlapping, attributes, such as, an engine. During the analysis of the biography genre, it was noticed that several types of biography exist along with links to their genres.

There was another type of biographical sub-genre or, arguably, mixed genre found: Football Manager. This structured article also naturally led to Football-Player, Team and Ground, which also linked to County and Country. The football team/club article seemed to outgrow its purpose and lead to new genres such as manager, ground and

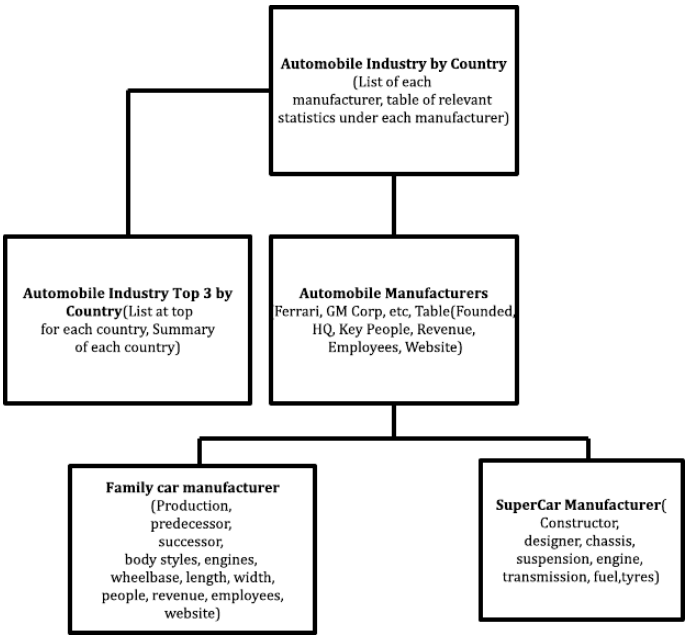

Figure 3: Small search for Automobile related Wikipedia articles and analysis of how they are structured and linked.

player. As Figure 4 (table 2 for more information) shows, each genre type is defined by certain forms that have been created in this particular community.

Search for football related Wikipedia articles and analysis of how they are structured and linked (see also Table 2). Several different types of biographical genre exist in Wikipedia with many different sets of characteristics for some notable figures in history, such as, Spike Milligan, Nelson Mandela, Alex Ferguson (Manchester United manager), Pol Pot, John Howard, Karl Marx and Margaret Thatcher. Other than the sole biographical structures for Spike Milligan, Karl Marx et al., a different form existed for ex-prime Minister John Howard, ex-president Nelson Mandela and ex-prime Minister Margaret Thatcher which, as can be seen in Figure 5, contains particular layout titles along with a biographical 'substance' in chronological order – this genre could be called: Leader. Many kinds of genres that are represented by several types of structure and meaning have been recorded in figures 2-5. Table 2 lists most of these recorded types and shows the attributes according to which we would contend they qualify to be categorised by form and purpose. An examination of the related interactions on the discussions pages and edits of the articles mentioned above showed that Wikipedia can qualify as a CoP because it contains the three characteristics outlined by Wenger (2000):

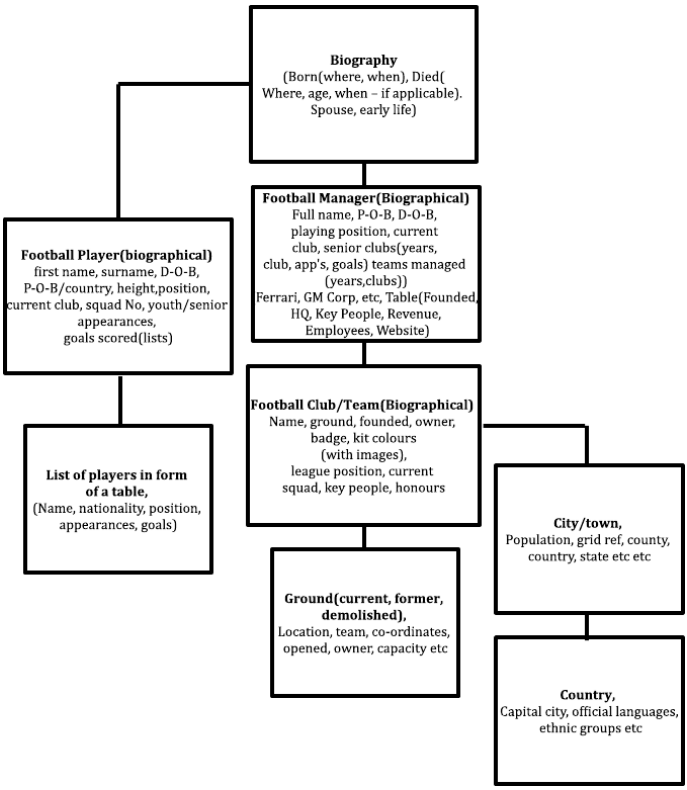

Figure 4: Search for football related Wikipedia articles and analysis of how they are structured and linked (see also Table 2).

The Practice, The Community and The Domain. The editors involved demonstrate a commitment to the domain and also seem to value their collective competence and the chance to learn from each other. The members engage in joint activities, such as voting, interaction and discussion. The editors develop a large and shared repertoire of resources, such as stories, tools and ways of addressing recurring problems, a mechanism for this being that the editors actually practice democracy by initiating voting cycles to discuss the merits of carrying out an alteration to an article[4].

[4]Vote Proposal: `http://en.wikipedia.org/w/index.php?title=Talk:R.E.M.\_discography\&oldid=94780788`

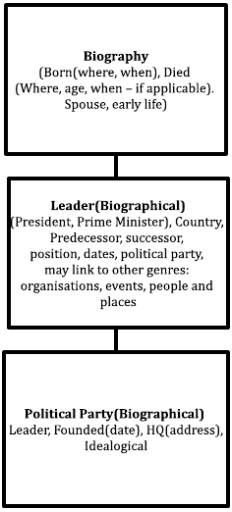

Figure 5: Search for notable people in history and analysis of how the main and related articles are structured and linked.

3.3 A Closer Look at Genres

An analysis of the literature on ecologies, CoPs and Yates and Orlikowski (2002, 15) work will help us to further identify the ways in which articles are created in domains such as Wikipedia. Also, by referring to the history of articles being made available, we can find out the details of how and when the particular articles (or genres!) are/were produced, reproduced and modified. Although it could be argued that carrying out an analysis of the edit histories, discussion/talk pages would, instead of demonstrating genre evolution, simply suggest the supplementing of previous knowledge or thoroughness, this would be a narrow-minded view of the genre evolution. The analysis of the edit histories and discussion clearly indicate a CoP implementing the division, merging, transformation and evolution of the article genres in this very complex domain. We looked closely at an example of a 'classical' genre, the biographical article, in this case about Spike Milligan, the celebrated and highly influential comedian and author who died in 2002. The purpose of the article is, obviously, to provide biographical information to the reader about Spike Milligan. As can be seen, the form of this web page article is continuously evolving and being transformed, much as Kwasnic and Crowston (2005) described in their ecological metaphor. The original article was first created on November 5, 2001; note the sparse and poorly organised information it contains (Figure 6).

After seven years, approximately 487 different users have submitted edits to the Spike Milligan page with only ten editors submitting more than 10 edits per person. The community for this particular article is evidently quite large and, as can be seen in

Terence Alan "Spike" Milligan (1918-) 'Irish' comedian, novelist, poet, and member of the Goons. Spike Milligan has suffered from Bipolar Disorder for most of his life.

Comedy shows:

* The Goon Show
* Q8

Books:

* Puckoon
* Adolf Hitler, My Part in his Downfall

Resources

* `http://www.fireflycafe.org/spike/`
* `http://www.google.com/Top/Arts/People/M/Milligan,_Spike/`

Figure 6: Spike Milligan Wikipedia article containing no formatting or notable structure dated 5 November 2001.

Table 1, the article has grown considerably. Over this time period, images were placed within the article. Eventually, the portrait picture in Figure 7 (after being in many different positions) ended up at the top right as nearly all pictures now do. On the 26th November 2006 a table with the title *Spike Milligan* was created by a contributor.

Since the screenshot was captured in early 2008, the biographical form in Figure 7 has yet again been transformed after much discussion by the editors involved. Not only has the contents table on the left been extended, but the table that encapsulates the name has also changed. The Birth name, Born, Died and Children information has now changed to Born, Died, Nationality, Influences and those people on whom he arguably exerted an influence. The focus is now concentrated on the career instead of on the person, much as with the Football Player or Leader, and is thus maybe moving towards forming another unique kind of genre which we could rename Artist or maybe Comedian. There could also possibly be overlaps with one of the new genres with classical forms, such as Obituary (as Milligan has died) and Biography. Another possible issue which could be linked with the merging and overlapping of genres is the reaching of a consensus on what constitutes a type of genre in a community, in this case a biography. Recently, the Spike Milligan article has evolved to contain more professional information than biographical (a human life in its course). The main elements in Figure 8 (below), "Children," has been amended to show professional influences and those whom he influenced instead of children, spouses (some time ago). The available 'histories' and underlying discussion area (Talk Page) do, of course suggest this but the information is not conclusive. It is obvious that by operating as a community, the contributors have added and enhanced information that they deem important (in a

Figure 7: Table 'feature' located in top-right of each biographical article.

hierarchy of importance) and have placed extra structural emphasis on the elements which are deemed most important about each article genre even if they do not always agree on these details. We noticed, by examining the history and discussion areas, that the Wikipedia editors have utilised a toolbox of HTML functions for formatting and embedding various media links, such as, video and photographs. The editors also seem to access unlikely sources to obtain information as indicated by one editor in the Talk page discussion: an un-named 'source' in the Daily Telegraph is cited as possessing a photograph of Spike Milligan's gravestone (for inclusion in the article) which is famous for the Gaelic inscription: "I told you I was ill".

4 Conclusion and Future Work

This paper constitutes the first steps in research on the Wikipedia domain, in particular, how the structures evolve in this "organic" and "biological" type of community. Wikipedia seems to be a suitably large and hierarchically structured CoP to demonstrate how genres evolve over a scale of time which will allow us to look closely at the evolution of a biography even though not all articles featured in Wikipedia are as formed as others. The viewed articles also contain a good fusion of form and purpose although some of the less formed articles contain a very small amount of form. The next step in this research is to formulate a study on genre and perception in this new area, that is, Wikipedia, which has the same aims and objectives as described in the earlier research paper by Clark (2007), and in the electronic mail study of Clark et al. (2008). A particular user search study will be set up to complement further research by looking into how the Wikipedia articles are used and perceived when a user extracts the form and recognises the purpose of the documents during an information search. The plan is to study how human beings cognitively interact and use genres of documents, which features or attributes they perceive and whether their perceptive processes can be explained or understood. Users are typically asked to read and categorise material from different genres and with different structures and forms. Measuring user categorisation according to genre, structure and form is further enhanced by recording eye movements during the tasks. Detailed data can thus be obtained regarding the attention paid to

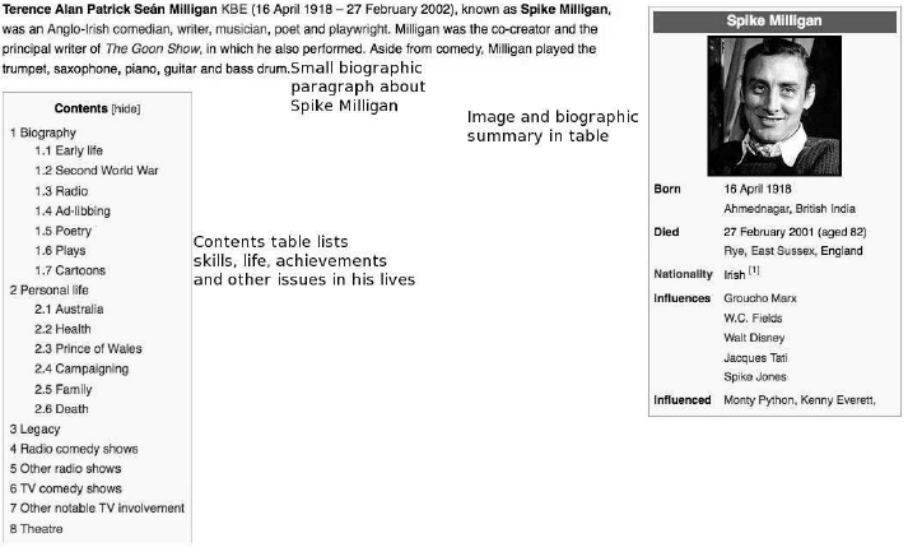

Figure 8: Biography example: visual format from Wikipedia containing tables, lists and image dated early 2008.

structures and forms by users when recognising, judging and determining genre. This research has the potential to show how human categorisation behaviour can be emulated computationally by a machine that actually 'understands' the meaning of a text for automatic retrieval. In some contexts, in particular, it is important to find out which of the two predominant processes – ecological (perceiving for action and affordances – cf. Gibson 1986) and constructivist (perceiving for recognition – cf. Gregory 1966) – are present in the subjects' genre recognition tasks.

References

Aristotle (1984). *The Rhetoric and the Poetics of Aristotle*. Modern Library College Editions Series. McGraw-Hill Higher Education, 1st edition.

Berkenkotter, C. and Huckin, T. N. (1995). *Genre knowledge in disciplinary communication: cognition, culture, power*. L. Erlbaum Associates, Hillsdale.

Boudourides, M. A. (2001). Commorg topics of genre literature review. Unpublished Article.

Breure, L. (2001). Development of the genre concept. *(last checked = 2009-05-22)*. `http://people.cs.uu.nl/leen/GenreDev/GenreDevelopment.htm`.

Clark, M., Ruthven, I., and Holt, P. O. (2008). Genre analysis of structured emails for corpus profiling. In *Proceedings of the Workshop on Corpus Profiling for Information Retrieval and Natural Language Processing*. EWICS.

Clark, M. J. (2005). Classifying xml documents by genre vol. 1. Master's thesis, The School of Computing.

Clark, M. J. (2007). Structured text retrieval by means of affordances and genre. In *BCS IRSG Symposium: Future Directions in Information Access BCS IRSG Symposium: Future Directions in Information Access BCS IRSG Symposium: Future Directions in Information Access BCS IRSG Symposium: Future Directions in Information Access*. British Computer Society.

Collins, T. D., Mulholland, P., and Watt, S. N. K. (2001). Using genre to support active participation in learning communities. In *In: The European Conference on Computer Supported Collaborative Learning (Euro CSCL 2001)*, Maastricht.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, London.

Dewdney, N., VanEss-Dykema, C., and MacMillan, R. (2001). The form is the substance: classification of genres in text. In *Proceedings of the workshop on Human Language Technology and Knowledge Management*, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.

Duff, D. (2000). *Modern genre theory*. Longman Publishing Group, London, 1st edition.

Ehmann, K., Large, A., and Beheshti, J. (2008). Collaboration in context: Comparing article evolution among subject disciplines in wikipedia. *First Monday: Peer-Reviewed Journal on the Internet*, 13(10).

Emigh, W. and Herring, S. C. (2005). Collaborative authoring on the web: A genre analysis of online encyclopedias. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05)*, volume 4, page 99a. IEEE.

Erickson, T. (2000). Making sense of computer-mediated communication: Conversations as genres, cmc systems as genre ecologies. In *Proceedings of the 33rd Hawaii International Conference on System Sciences-Volume 3*. IEEE Computer Society.

Gibson, J. J. (1986). *The ecological approach to visual perception*. LEA, New Jersey, 2nd edition.

Giddens, A. (1986). *The Constitution of Society: Outline of Theory of Structuration*. University of California Press.

Gregory, R. L. (1966). *Eye and Brain the psychology of seeing*. World University Library, London, 1st edition.

Halliday, M. A. K. (1973). *Explorations in the Functions of Language*. Edward Arnold (Publishers) Ltd.

Halliday, M. A. K. (1978). *Language as Social Semiotic: The Social Interpretation of Language and Meaning*. Edward Arnold, London.

Huang, F., Watt, S., Harper, D. J., and Clark, M. (2006). Robert gordon university at inex 2006: Adhoc track. In *Overview of INEX 2006*, volume 4518/2007, pages 64–72. Springer-Verlag.

Ingwersen, P. and Jarvelin, K. (2005). *The Turn: Integration of Information Seeking and Retrieval in Context*, volume 18 of *The Information Retrieval Series*. Springer, 1st edition.

Kress, G. and Threadgold, T. (1988). Towards a social theory of genre. *Southern Review*, 21(3):215–243.

Kwasnic, B. and Crowston, K. (2005). Introduction to the special issue genres of digital documents. *Information Technology and People*, 18(2):76–87.

Lalmas, M., Rolleke, T., Szlavik, Z., and Tombros, T. (2004). Accessing xml documents: the inex initiative. In Agosti, M. and Fuhr, N., editors, *DELOS WP7 Workshop on the Evaluation of Digital Libraries*, University of Padua, Italy.

Martin, J. R. (1999). Mentoring semogenesis: 'genre-based' literacy pedagogy. In Christie, F., editor, *Pedagogy and the Shaping of Consciousness: Linguistic and social processes*, Open Linguistics Series, pages 123–155. Cassell, London.

Miller, C. R. (1984). Genre as social action. *Quarterly Journal of Speech*, 70(2):151–67.

Orlikowski, W. J. and Yates, J. A. (1994). Genre repertoire: norms and forms for work interaction. *Administrative Science Quarterly*, 39:541–574.

Watt, S. (2009). *Text categorisation and genre in information retrieval.* (In Press), chapter In Press. John Wiley and Sons.

Wenger, E. (2000). Communities of practice and social learning systems. *Organization*, 7(2):225–246.

Yates, J., Orlikowski, W. J., and Rennecker, J. (1997). Collaborative genres for collaboration: Genre systems in digital media. In *Proceedings of the 30th Hawaii International Conference on System Sciences: Digital Documents - Volume 6*, volume 6, pages 50–59. IEEE.

Yates, J. A. and Orlikowski, W. (2002). Genre systems: Structuring interaction through communicative norms. *Journal of Business Communication*, 39(1):13–35.

Yates, J. A. and Orlikowski, W. J. (1992). Genres of organizational communication: a structurational approach to studying communication and media. *Academy of Management Review*, 17(2):299–326.

Table 1: Article Structure by Contents Table (positioned top-left of each article see Figure 4 Left Hand Side) Evolving bi-annually

Nov 2001	Nov 2004	Nov 2006	Nov 2009
Comedy Shows: ∗The Goon Show ∗Q8 Books: ∗Adolf Hitler, My Part in his Downfall ∗Puckoon	∗1 Biography ∗2 Radio Comedy Shows ∗3 TV Comedy ∗4 Theatre ∗5 Movies ∗6 Books ∗7 Quotations ∗8 External Links	∗1 Biography ∗2 Posthumously ∗3 Trivia ∗4 Radio Comedy ∗5 Other radio shows ∗6 TV comedy shows ∗7 Theatre ∗8 Films ∗9 Books	∗1 Biography o 1.1 Early Life o 1.2 WW II o 1.3 Radio o 1.4 Ad-libbing o 1.5 Poetry o 1.6 Plays o 1.7 Cartoons ∗2 Personal life o 2.1 Australia o 2.2 Health o 2.3 Prince of Wales o 2.4 Campaigning o 2.5 Family o 2.6 Death ∗3 Legacy ∗4 Radio comedy shows ∗5 Other radio shows ∗6 TV Comedy Shows ∗7 Other TV ∗8 Theatre ∗9 Films ∗10 Books ∗11 Quotations ∗12 External links ∗13 References

Table 2: Article, Genre, Purpose and Form

Genre	Stats (Date Created/Amount of Editors/Edits)	Attributes of Purpose (Themes, topics, discourse structure)	Attributes of Form(Structural features e.g titles, lists etc)

Band/Musical Group (Query REM)	1 February 2002, 1067 editors ,1564 edits	To biographically present the past and present members of the group, show their work output and list their achievements.	∗TABLE TITLES, HEADINGS: Background information, Origin Genre(s) ,Years active Label(s), Associated acts, Website(URL), Former members. MAIN TEXT HEADINGS Chronological History, URL(s)to listen/download radioone or more song samples, Summary of the Discography.
War (query Napoleonic Wars)	22 March 02, 991 editors, 2361 edits		TABLE TITLES/HEADINGS date, location, result. List of Beligerants, names of sides, List of commanders on each side, casualities and losses on each side in numerics. MAIN TEXT HEADINGS (title and years of stage) Lead up to start of war, major phases of war(battles etc), outcome, legacy and effects.
Discography(query REM)	17 December 2005/ 145 editors /410 edits	To present and list the output produced by an entity such as musical artists TABLE TITLES, HEADINGS Small summary table with type of release and amount e.g. album 5.	MAIN TEXT HEADINGS: Tables. Each table by title such as Studio Albums, Singles etc with sub-titles such as Year, Album and Single Details, chart positions.
A to Z index List of Bands (by genre query alternative rock bands)	27 March 2004, 832 editors, 4520 edits	To present a comprehensive alphabetically structured index of alternative musical groups throughout the world.	TABLE TITLES, HEADINGS Contents table 0 to 9 A B C D E F G H I J K L M N O P Q R S T U V W X Y Z MAIN TEXT HEADINGS Small summary. Index of alphabetical sections with list of bands name beginning with 0 to 9 to Z.
Leader query Nelson Mandela	7 June 2005, 317 editors, 1053 edits	To present the biographic details of how and when a person became a leader in a political party etc	TABLE TITLES HEADINGS Title of office held, dates held position, Vice president, succeeded and/or proceeded by, born where and when, political party MAIN TEXT HEADINGS Early life, key moments in life and leadership

City query Aberdeen	5 February 2002, 28 editors, 2234 edits	To detail the geographic, population and historical information pertaining to a particular city.	TABLE TITLES HEADINGS Name of city, map with location, population, density, language spoken, location Council area, Lieutenancy area, Constituent country, Sovereign state, Post town, Postcode district Dialling code, Police or fire ambulance(name of service, European Parliament, UK Parliament Scottish Parliament MAIN TEXT HEADINGS Geography, demography, climate, Landmarks, transport, culture
Football Club query Scarborough Athletic	25 June 2007,68 editors,451 edits	To present current and historical information, including achievements, regarding a football and/or soccer team.	TABLE TITLES HEADINGS image of coat of arms, Full name, Nicknames, Founded, Ground (Capacity), Owner, Managing Director, League, Premier League. Images of club strip(shorts, socks and top). MAIN TEXT HEADINGS: Stadiums, Supporters, Table of honours, records, Table with list of current aquad players. Tables (with lists by name and years) coaching staff, key people, manager history, chairman history.
List of Football Players query List of Newcastle United F.C. players	11 February 2006, 110 editors,357 edits	To present current(still playing) and historical information(now retired), including achievements, regarding a football or soccer player.	TABLE TITLES/ HEADINGS None MAIN TEXT HEADINGS Large table with headings Name Nationality, Position, Club Name career, appearances, Goal Table with list of first team captains (year and name)
List of Lists query list of bands by genre etc	10 December 2003/195 editors/340 edits	A comprehensive list of lists sorted by certain categories.	TABLE TITLES None. MAIN TEXT Title(By Genre, By Instrument) then list under each
Political Party query socialist party of Ireland	12 February 2004/92 editors/553 edits	Presents biographic information regarding a political party in a particular country or region in the world.	TABLE TITLES/ HEADINGS Name, Logo, Founded, Leader, Headquarters, Political ideology, International Affiliation, European Affiliation European Parliament Group, Colours , Website. MAIN TEXT HEADINGS: Electoral history, Key policies, List of elected members(name, position, district)

Automobile Manufacturer (query General Motors)	25 February 2002/1772 editors/5233 edits	Presents information to the public regarding the general business structure and financial performance.	TABLE TITLES/HEADINGS Type , Founded, Founder(s) Headquarters, Area served, Key people, Industry , Products, Services , Revenue ▽ currency (year), Operating income ▽ currency(year), Net income ▽ currency(year), Total assets ▽ currency (year), Total equity ▽ currency (year), Employees, (number)(year), Divisions, Subsidiaries, Website (url) MAIN TEXT HEADINGS: History, Company Overview, Corporate Structure, Table listing open manufacturing plants, Table of Yearly Sales, List of brands/defunct brands, subsidiaries