Chapter 13

HYPERTEXT TYPES AND MARKUP LANGUAGES

The Relationship Between HTML and Web Genres

Georg Rehm
University of Tübingen, Germany
georg.rehm@uni-tuebingen.de

1. Introduction

It is vital to take a closer look at the role of the Hypertext Markup Language (HTML, Raggett et al., 1999) with regard to text technological applications that aim at processing web documents (for example, automatic summarisation, information extraction, or text classification). This article introduces the concept of hypertext types to highlight some of the most relevant aspects and applications. Hypertext types are very similar to traditional text types, i. e., actual hypertexts can be grouped into categories that share certain linguistic, textual, or pragmatic features such as communicative function, hypertextual structure, or content. Nowadays, the term hypertext (see, e. g., Kuhlen, 1991, Hammwöhner, 1997, Storrer, 2004) is used almost synonymously with the World Wide Web (WWW). Therefore, the term *web genre* seems to be more appropriate to refer to different types of websites.

Categorising web documents into web genres potentially comprises several advantages. The automatic identification of web genres (Rehm, 2002, 2007, Rehm and Santini, 2007) is an intuitive application: summarisation techniques can be tailored to the instances of specific web genres, collecting corpus data within the "web as corpus" approach can be restricted to a certain set of web genres, or information extraction engines can be implemented that are specialised to web genres such as *C. V.* or *List of Publications*.

The remainder of this chapter is structured as follows: first, section 2 describes the relationship between arbitrary markup languages and tra-

ditional text types. Section 3 details the concept of web genres and, among others, explains the evolutionary process that produces web genres. The contribution finishes with concluding remarks (section 4).

2. HTML as a Markup Language

HTML (Raggett et al., 1999) and XHTML (Pemberton, 2002) are applications of the metalanguages SGML (ISO 8879) and XML (Bray et al., 2004) respectively. Therefore, both HTML and XHTML can be considered SGML/XML-based markup languages. Before we can examine the specifics of HTML more closely, we need to concentrate on the features of a typical, generic markup language (assuming such a thing exists; cf. Lobin, 2001, p. 181).

Although markup languages can comprise more or less arbitrary pieces of information, a typical application scenario is to specify the structural patterns of a text type by means of a text grammar (or document type definition, DTD, in SGML/XML lingo) – popular examples often used in introductory classes or textbooks are the generic discourse structures of cooking recipes, poems, or books (Maler and Andaloussi, 1996, Lobin et al., in this volume). In an abbreviated, rewrite-rule-like notation, a DTD for non-fiction books often comprises element declarations such as (i) `book` → `contents`, `chapter+`, `index` (ii) `chapter` → `title`, `section+` (iii) `section` → `title`, `paragraph+`, `subsection+` etc.

Neither HTML nor the nearly equivalent XHTML contain nested declarations of this depth or semantic specifity. HTML 4.01 (Raggett et al., 1999) comprises declarations for 91 elements and 119 attributes but only very few of these rules are related to structural or informational units that themselves can be related to the constituents of the structural patterns of specific text types. In other words, the informational units provided by HTML do not correspond to the parts or building blocks of one type of text or hypertext. Rather, HTML specifies a set of generic elements and attributes that serve, for example, presentational and typographic (`i`, `b`, `font`), structural (`ol`, `ul`), semantic or logical (`em`, `strong`), referential or hypertextual (`a`), and functional (`object`, `embed`) means (see Walker, 1999, for a more detailed discussion).

HTML is not an ordinary, typical markup language. On the contrary: it is not aimed at one specific text type, it comprises several types of markup (see above) and must thus be considered a hybrid, heterogeneous markup language with a bias on presentational and functional aspects. Evidence for this can be found within HTML's sets of element and attribute names. Most of these are completely generic and universal and do not relate to text-structural or even genre-specific properties. In

addition, there are almost no restrictions with regard to the nesting or the combination of specific block-level elements (lists, headlines, tables, quotations etc.).[1] Further evidence can be found online: authors are constantly creating a multitude of different types of texts and hypertexts by employing only *one* single markup language – HTML.

3. Hypertext Types – Web Genres

The WWW comprises different types of hypertexts. Following the distinction of different types of texts as genres (*Textsorten* in German, cf. Brinker et al., 2000), these types can be conceptualised as individual web genres. In sharp contrast to the state of the art in traditional text linguistics, both theoretical and applied web genre-oriented research is, surprisingly, rather scarce (see, for example, Shepherd and Watters, 1999, Crowston and Williams, 2000, Rosso, 2005).

3.1 Genres and Web Genres

According to the framework sometimes referred to as North American Genre Theory and established by the works of Miller (1984), Bazerman (1994), and Swales (1990), among others, a genre "comprises a class of communicative events, the members of which share some set of communicative purposes. These purposes [...] constitute the rationale for the genre. This rationale shapes the schematic structure of the discourse and influences and constrains choice of content and style." (Swales, 1990, p. 58). Thus, genres are primarily function-oriented patterns of communication that aid in the realisation of specific communicative goals.

Yates and Orlikowski (1992, p. 299) apply genre theory in order to analyse genres of organisational communication and define genres as "typified communicative actions characterized by similar substance [social motives as well as content, G. R.] and form [the communicative action's physical manifestation, G. R.] and taken in response to recurrent situations." (see Erickson, 2000). In a later article, Orlikowski and Yates (1994, p. 543) emphasise the influence of the discourse community and define genre as "a distinctive type of communicative action, characterized by a socially recognized communicative purpose and common aspects of form [...]. The communicative purpose of a genre is not rooted in a single individual's motive for communicating, but in a purpose that

[1] The declaration of the element `body` in the HTML 4.01 DTD (Raggett et al., 1999) contains *arbitrary* sequences of the elements `p`, `h1`, `h2`, `h3`, `h4`, `h5`, `h6`, `ul`, `dl`, `pre`, `dl`, `div`, `noscript`, `blockquote`, `form`, `hr`, `table`, `fieldset`, `address`, and `script`. A related problem concerns the phenomenon known as "tag abuse" (Barnard et al., 1996): authors often use specific elements for their presentational features only and completely ignore their original semantics.

is constructed, recognized, and reinforced within a community". Yet another definition describes genres of organisational communication as "socially recognized types of communicative actions – such as memos, meetings, expense forms, training seminars – that are habitually enacted by members of a community to realize particular social purposes" (Orlikowski and Yates, 1994, p. 542). Therefore, a genre structures communication processes by means of shared expectations that refer to both content and structure. At the same time, the cost of producing or interpreting the instance of an established genre is reduced.

All of the abovementioned properties of traditional genres can be and indeed have to be applied to digital genres as well. For example, electronic mail is employed millions of times each day to carry instances of hundreds of distinct digital genres that can be described and analysed in similar ways as paper-based genres. In principle, the same is true for web genres, but the World Wide Web, as a means of communication, exhibits characteristics that necessitate revising and extending the traditional approaches and categories. The most important aspect refers to the web's roots in hypertext (see Storrer, in this volume): the foundation in the concept of hypertext is partly responsible for the often cited fragmented nature of HTML documents. Authors tend to bundle related informational units in lists of hyperlinks that connect individual nodes, therefore breaking up the traditional notion of a text as a coherent whole with an intrinsic beginning and end. Other aspects concern the influence of interactive features and the level of granularity that is needed for a proper analysis and description of web genres (see, for example, Shepherd and Watters, 1998, 1999, Haas and Grams, 1998, 2000, Rehm, 2007, and section 3.3).

3.2 The Evolution of Web Genres

It is necessary to have a closer look at the processes that are responsible for the gradual evolution of web genres in order to examine HTML's status as a markup language, especially with regard to operating on web pages by means of Natural Language Processing techniques. Generally speaking, we can distinguish between automatically converted and manually prepared HTML documents. Based on the respective tool or approach chosen by the author, additional stages with individual advantages and disadvantages can be identified (see figure 13.1).

The influence of the text type on the automatic conversion of existing documents is straightforward (see figure 13.2): a document created, for example, with a WYSIWYG text processor, usually belongs to a certain text type or genre such as *Scientific Article*, *List of Publications*, or

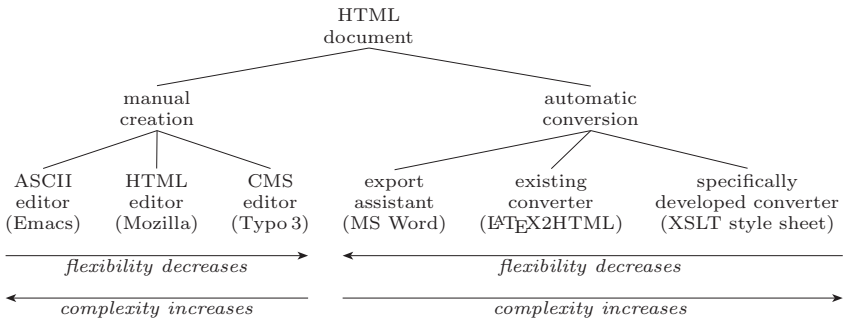

Figure 13.1. Facilities of creating HTML documents

Minutes. The conversion process does not significantly alter the text type, so that we can assume that a document's original genre is directly transferred into the web. Whereas slight modifications are often carried out in the generated HTML document (e. g., removal of navigation bars or footers), extensive changes are applied only to the source document so that yet another pass through the converter generates an up-to-date HTML version of the document. While the source document underlies a document life cycle (Lobin, 2000), the converter itself is subject to a software life cycle: new editions of a conversion tool introduce new features that might result in, for example, new layout templates, or novel navigation approaches in the machine-generated HTML documents.

The evolutionary processes that shape and form web genres with regard to the manual creation of websites (i. e., HTML documents) are

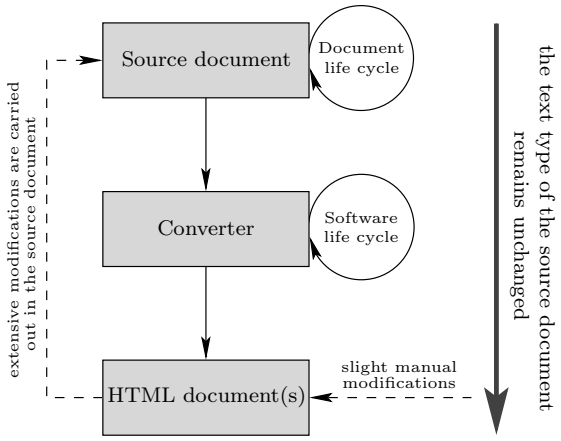

Figure 13.2. The influence of the text type on the conversion of documents into HTML

more complicated. Figure 13.3 depicts the most important influential factors, arranged in a cyclic and highly abstract model of web genre evolution that comprises four phases (Rehm, 2007, provides a more detailed discussion). The two phases on the right hand side (*hypertext production*, *modification*) can be associated with a generic author who intends to build a website. The significant influential factors are, for example, the purpose and communicative function of the website, its designated content, the software used to build the site, and the author's personal history and experience browsing and using the web as well as traditional genres (see, e. g., Furuta and Marshall, 1996, and Eckkrammer, 2001). The two phases on the left hand side (*change*, *reception*) represent the perpetual modification processes of other websites. These underlie rules and conventions whose spectrum ranges from very specific (e. g., obligatory functional elements that users expect to work the way they experienced on other sites) to very soft (rather optional features; see, e. g., Yates and Sumner, 1997). The cycle's four phases describe the slow-going process of emerging rules and conventions for specific web genres: authors of HTML documents are always readers of HTML documents. If an author wants to build a site with a certain purpose and content (i. e., a site with a specific web genre), he or she incorporates – consciously or unconsciously – elements of related websites. Over a period of time, this process generates web genre-specific conventions and rules that authors choose to apply, to extend or to break (cf. Yoshioka et al., 2001): "When establishing a new site that serves a purpose similar to existing sites, the genre characteristics are copied and refined to reflect resemblance to an existing genre" (Eriksen and Ihlström, 1999, p. 289). Empirical studies such as Ryan et al. (2003) or Emigh and Herring (2005), in which documents of specific web genres have been compared over long periods of time, yield further evidence for the cyclic model presented here.

The two abovementioned models complement and extend the taxonomy on the evolution of "cybergenres" (Shepherd and Watters, 1998) that can be characterised by specific features with regard to content, form and functionality. Although Shepherd and Watters do indeed use the term 'taxonomy', this concept is misleading (see Rehm, 2007, for critical remarks), as the model represents a kind of evolutionary continuum which comprises an "evolutionary path" that leads from "replicated genres" (*newspaper*) to "variant genres" (*electronic news*) and finally to "emerging genres" (*personalized news*). An additional node of the taxonomy is labeled "spontaneous cybergenres" (*homepage*).

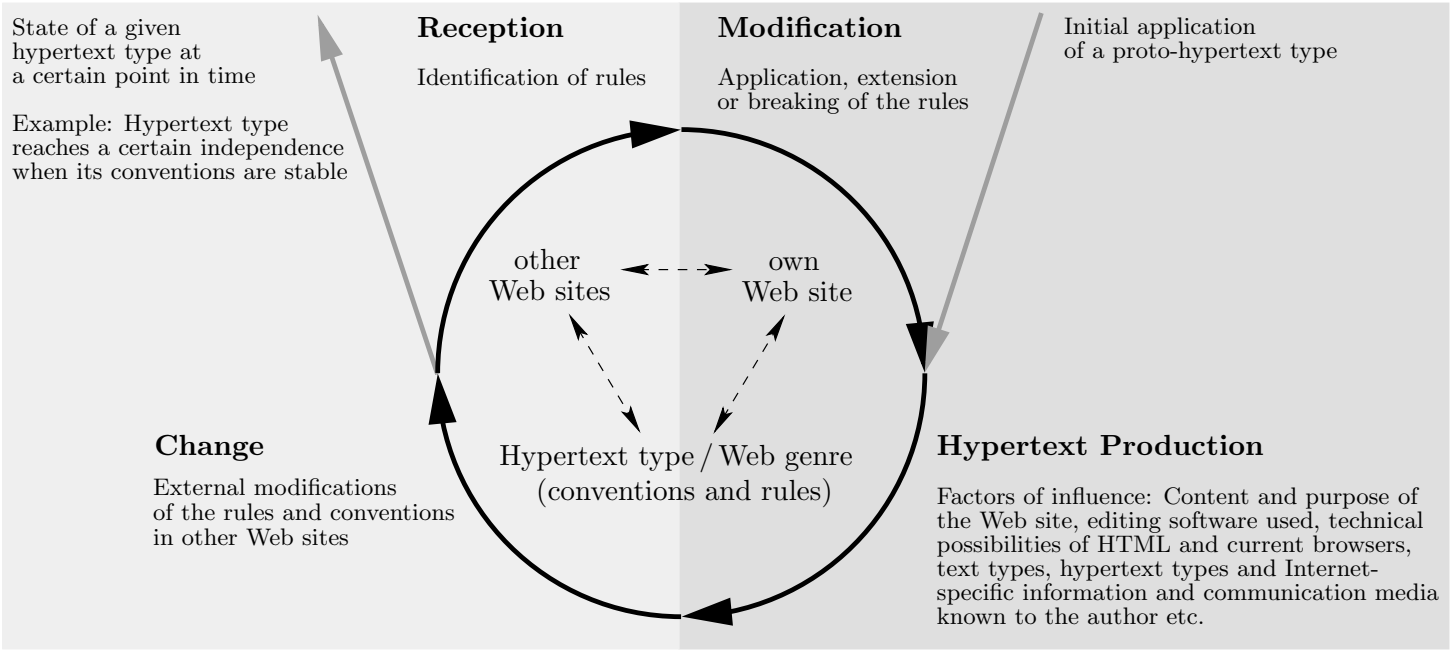

Figure 13.3. The phases and influential factors of the evolution of hypertext types with regard to documents created manually

3.3 The Web Genre Model

Though several specifics of digital genres and web genres have been addressed by previous research, an all-encompassing web genre theory is still missing. The model outlined in this section (see figure 13.4) aims to fill this gap, and is a revision and extension of the approach originally published in Rehm (2002). The roots of this model lie in text linguistics. It is primarily intended to aid linguistic analyses of websites as well as to guide approaches for their automatic processing. An exhaustive description can be found in Rehm (2007).

In principle, a web genre can be thought of as a generic text type, i. e., a conventionalised pattern of communication (see section 3.1) used within the World Wide Web, or, in other words, it is a hypertext type. Probably the most important feature of a text type is its communicative function which authors intend to fulfill by means of instantiating the (hyper)text type (Jakobs, 2003). Other features include contextual properties (such as the relationship between author and reader), specific hypertextual structuring conventions (for example, linear navigation from node to node used in online-payment sequences) or global web design features (decoration) that apply to all the nodes in a hypertext.

The model consists of three levels of granularity that represent the constituents of a web genre: the most fundamental of these is the web genre module that serves as a basic building block. As web genres are not monolithic but instead rather flexible (Haas and Grams, 1998), web genre modules act as basic logical-semantic entities that can be freely arranged in one or more HTML documents (see also Mehler, in this volume). It is necessary to assume this conceptual level of text structural entities beneath the individual document as, for example, the entry page of an instance of the web genre *Academic's Personal Homepage* may comprise the web genre modules *List of Publications*, *Contact Information*, *Current Courses* and *Current Projects* (Santini, 2007). While one author may arrange these web genre modules in a single HTML document, another author may choose to instantiate each of these web genre modules in HTML documents of their own. Based on their frequency, the fundamental constituents of a specific hypertext type can be divided into compulsory ($\geq 50\%$) and optional web genre modules ($< 50\%$), which form the peripheral boundaries of a web genre. Furthermore, we can distinguish between atomic and complex web genre modules with regard to their internal structure: for example, the atomic web genre modules *List of Publications*, *List of Presentations* and *Current Projects* constitute the complex web genre module *Scientific Profile* (with regard to the web genre *Academic's Personal Homepage*). In ad-

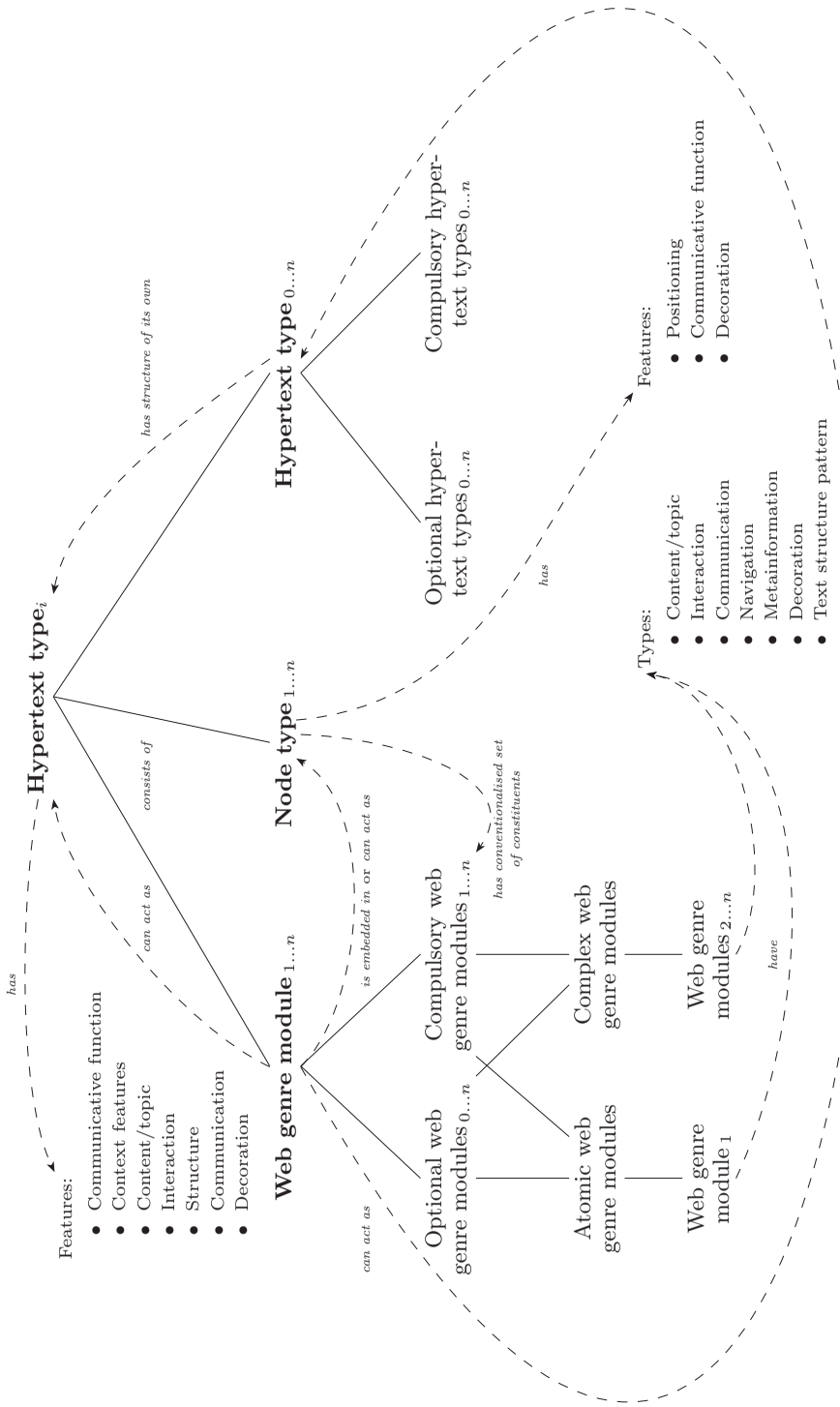

Figure 13.4. The generic constituent structure of a hypertext type/web genre

dition, every web genre module is reflected by at least one type: the *List of Publications* has a very significant *text structure pattern*, whereas the web genre module *E-Mail Address* is primarily reflected by the types *communication* and *interaction*.

The second group of constituents is called node type and relates to conventionalized configurations of web genre modules. In his information architecture handbook, Reiss mentions one such node type:

> More and more, [sic] companies are combining their site map, webmaster email address, disclaimer, and copyright information on a single page. Usually labeled "About this site", this combination page seems well on its way to becoming an established web convention. (Reiss, 2000, p. 135)

What Reiss calls a "combination page" is, according to the web genre model, a conventionalised node type that consists of web genre modules such as "site map" and "disclaimer". The conceptual level represented by node types is needed in order to be able to capture conventions with regard to the internal structure of individual nodes (i. e., HTML documents). Node types have features that relate to the positioning of web genre modules and a node type's individual web design (that might differ from the web genre's global decoration feature).

Finally, the instance of a web genre such as *University Website* can embed instances of subordinate web genres such as *Website of a University Department* or *Computer Centre Website*. This process of embedding complex constituents is, again, nothing but another set of rules or conventions. As trivial as this example might seem: every university website contains sub sites (embedded hypertexts) for every department within a university. Usually, members of the departments are responsible for creating and maintaining these websites. This circumstance (two groups of authors who, with their texts, pursue different objectives, i. e., communicative functions) justifies the distinction between two different hypertext types in this example.

The boundaries between web genre modules, node types and hypertext types is not as clear-cut as it might initially seem (Mehler et al., 2004). As neither binding norms nor straightforward standards for web genres exist, authors are rather free to structure their content as they see fit. Based on the analyses described in Rehm (2007), it is safe to state that conventions with regard to specific sets of web genre modules exist in most web genres, but the way these web genre modules are structured and sequenced within one or more hypertexts is less conventionalised. This is why the model specifies that web genre modules can act as node types as well as hypertext types. Consider the web genre *List of Publication*. With regard to most instances of the web genre *Academic's Personal Homepage*, this web genre module is either embedded

in a node together with other web genre modules, or it is the only web genre module in an HTML document (web genre module acts as node type). There are cases, however, in which authors chose to structure their lists of publications in more than one node, effectively creating a hypertext (for example, one node for monographs, another for refereed articles etc.). Therefore, a minimal web genre (or hypertext type) instance comprises at least one compulsory web genre module that can act as a node type, resulting in a single node.

Finally it should be noted that descriptions of web genres and node types rely on empirical evidence. For example, if one wants to construct a profile for the web genre *Academic's Personal Homepage*, a sample of corresponding hypertexts needs to be gathered first. The sample analysis includes inductively generalising the web genre modules, node types and embedded web genres contained within the sample's hypertexts and integrating the data as a web genre profile (see Rehm, 2007).

3.4 Analysing Web Genres

Analysing samples of HTML documents with regard to their web genres usually involves one of two goals. First, randomly collected instances of *one* web genre can be examined in order to construct empirically a web genre profile for this very web genre (see section 3.3). Second, if a sample contains randomly collected documents of *arbitrary* web genres (i. e., arbitrary documents), the aim is to identify all web genres, node types and possibly web genre modules contained in the sample, in order to get an idea about the variety of web genres in use within a particular domain or discourse community. Furthermore, annotating a document sample with web genre information results in a data set that can be used as training and testing data in a machine learning scenario that aims at the automatic identification of web genres.

As initially suggested in Rehm (2002), it is crucial to select an analysis domain first (see, e. g., Rosso, 2005), as the analysis of truly arbitrary documents such as the ones provided by randomly picking documents stored in the databases of search engines (an approach employed by Crowston and Williams, 2000, Shepherd and Watters, 1999, and Haas and Grams, 2000), leads to results that are too broad and too vague to be of any actual use. Furthermore, a snapshot, i. e., a stable corpus of HTML documents is needed that comprises the documents to be analysed. For this purpose, web crawlers and additional components such as language identification tools can be used in order to filter and preprocess the documents to be integrated into the corpus (Rehm, 2001).

A web-accessible corpus-database that assists the analysis of documents is described in Rehm (2007). Furthermore, the system provides an easy to handle interface for the generation and maintenance of document samples. The results of an analysis are automatically stored in a relational database. Several analyses have been carried out with the help of this tool. Gleim et al. (2007) present an integrated tool for the purpose of collecting, analysing, and annotating web corpora.

3.5 On the Variety of Web Genres

The analyses described in Rehm (2007) are based on a corpus of ca. four million HTML documents, written in German, that were crawled from the websites of 100 German universities. The domain of academia was chosen because it is assumed to be rather stable and highly structured, thus making it a prime candidate for a project related to the description and automatic identification of web genres. One of the analyses aimed to shed light on the number of web genres in use within the domain of academic websites. A random sample of 750 documents, generated by means of the corpus-database (see section 3.4), was analysed with regard to the node types, web genres (i. e., the superordinate web genre to which an individual node belongs) and the organisational unit that published a document on their website.

1. **Website of an organisational unit** (subtypes: 24) 28.4%; 2. **Website of an educational course** (subtypes: 4) 13.9%; 3. **Course program/directory** 6.0%; 4. **Software documentation** (subtypes: 4) 5.3%; 5. **Annual research report** 3.7%; 6. **Course materials (manuscript)** 3.7%; 7. **Photo gallery** (subtypes: 4) 3.5%; 8. **List of press releases** 3.2%; 9. **Organisational unit's publication** (subtypes: 8) 2.5%; 10. **Academic's personal homepage** 2.3%; 11. **Website of an organisation** (subtypes: 9) 1.9%; 12. **Student's homepage** 1.6%; 13. **School teaching materials** 1.5%; 14. **Study guide** 1.3%; 15. **Course of studies website**; 16. **Student presentations/theses** 1.2%; 17. **Employee directory**; 18. **Handbook** 1.1%; 19. **Virtual museum**; 20. **Instructions, manuals, documentations** 0.9%; 21. **Library catalogue** 0.8%; 22. **Textbook/textbook chapter**; 23. **Diploma thesis**; 24. **Digital library** 0.7%; 25. **Message/discussion board**; 26. **Student presentation/essay/thesis**; 27. **Conference website**; 28. **Medical diagnosis procedure** 0.5%; 29. **Lexicon**; 30. **Contest/event website**; 31. **Access statistics**; 32. **Tasks for student papers** 0.4%; 33. **Research projects of an organisational unit**; 34. **Medical diagnosis example**; 35. **Law, regulation, legal text**; 36. **Student statistics**; 37. **Final report (of a project)** 0.3%; 38. **Latest information, news**; 39. **Biography**; 40. **Digital map**; 41. **PhD thesis**; 42. **Subject-specific information portal**; 43. **FAQ document**; 44. **Graphical assistant for process development**; 45. **Internet journal (review forum)**; 46. **Mailing list archive**; 47. **Bibliography** 0.1%; 48. **Library classification scheme**; 49. **Historical building data**; 50. **Trip/excursion report**; 51. **Glossary**; 52. **Almanach**; 53. **Classifieds**; 54. **Cook book**; 55. **Art/cultural event**; 56. **Minutes archive**; 57. **Examination regulations**; 58. **Guidelines (for student papers)**; 59. **Materials on special reserve**; 60. **Study regulations**; 61. **Daily newspaper**; 62. **Betting game (on a sports event)**; 63. **Knowledge transfer catalogue**; 64. **Virtual library**; 65. **Scientific article**

Table 13.1. The 65 web genres found in a random sample of 750 documents

Table 13.1 shows the web genres found in the sample and illustrates the enormous variety in the set of hypertext types instantiated in academic websites. Web genres such as, for example, *Website of an Organisational Unit*, *Conference Website*, *Course Materials*, *PhD Thesis*, *Diploma Thesis*, and *Examination Regulations* can be expected in the domain. Furthermore traditional genres, such as *Cook Book*, *Classifieds*, and *Lexicon*, as well as a small number of highly specialised genres of sci-

entific communication (*Medical Diagnosis Procedure*, *Historical Building Data*) also occur. In addition, we find web genres primarily related to other internet services (*Mailing List Archive*, *FAQ Document*) or technical aspects of the World Wide Web (*Access Statistics*).

Table 13.2 shows the node types found in the sample of 750 randomly collected HTML documents. A rather salient example is *Page/paragraph* with occurrences in 119 documents. This category comprises 20 subtypes such as *Page/paragraph of a Software Documentation*, *Page/paragraph of a Handbook*, or *Page/paragraph of a Study Guide*. In printed manifestations of genres such as *Software Documentation*, *Handbook*, or *Study Guide*, logical constituents beneath the genre level or units corresponding to a physical page are usually not named (with the exception of constituents such as *Table of Contents*, *Chapter*, *Section*, *List of References* etc.). Therefore, all the node types which correspond to superordinate genres but that do not possess an identifying label are subsumed under the category *Page/paragraph*. As a natural consequence, the variety at the node level is greater than at the web genre level. The sample contains instances of node types as diverse as, for example, *Slide*, *Abstract*, *Press Release*, *Exercises*, *Announcement*, *Exhibit (of a Virtual Museum)*, *Minutes*, *Consent Form*, *City Map*, *Invitation*, *Review*, *"under construction" Notice*, *Classified Ad*, *Price List*, *Travel Diary*, and *Riddle*.

1. **Page/paragraph** (subtypes: 20) 15.9%; 2. **Slide** (subtypes: 6) 10.7%; 3. **Organisational data (course)** (subtypes: 4) 6.1%; 4. **Abstract** (subtypes: 6) 5.6%; 5. **Photo** 3.9%; 6. **Entry page** 3.2%; 7. **Press release**; 8. **Employee's professional homepage** (subtypes: 2) 2.4%; 9. **Article (organisational unit's publication)** (subtypes: 6) 2.1%; 10. **Primary navigation bar** 1.7%; 11. **Description of a work group** 1.6%; 12. **Instruction, manual, documentation** 1.3%; 13. **Hotlist**; 14. **Academic's personal homepage**; 15. **Exercises**; 16. **Course program/directory** (subtypes: 3); 17. **Schedule/program (of a course)** 1.2%; 18. **List of publications** (subtypes: 2); 19. ***Categorisation not possible***; 20. **Head line** 1.1%; 21. **Study guide** (subtypes: 3); 22. **School teaching materials**; 23. **Announcement** 0.9%; 24. **Photo gallery**; 25. **Exhibit (of a virtual museum)** 0.8%; 26. **Library catalogue (single record)**; 27. **E-mail**; 28. **Contact information**; 29. **Description of an organisational unit (functions and contact information)**; 30. **Description of technology transfer services** 0.7%; 31. **Lexicon entry**; 32. **Exercise solutions**; 33. **Employee directory**; 34. **Source code**; 35. **Student statistics**; 36. **Finished theses/possible topics for theses** 0.5%; 37. **Latest information, news**; 38. **Bibliography**; 39. **Invitation**; 40. **Table of contents** (subtypes: 3); 41. **Examination dates**; 42. **Medicial diagnosis procedure**; 43. **Statistical data (automatically generated)**; 44. **"under construction"**; 45. **Dispatcher**; 46. **Index (generated by web server)**; 47. **Possible topic for an essay/a thesis** 0.4%; 48. **Conference report**; 49. **Download list (multimedia resources)**; 50. **Course description**; 51. **Institution history**; 52. **Description of an organisational unit (profile/portrait)**; 53. **Minutes**; 54. **Question and answer**; 55. **Law, regulation, legal text**; 56. **Review**; 57. **Course of study description**; 58. **Study regulations**; 59. **Technical data/specification (hard-/software)**; 60. **Biography** 0.3%; 61. **Sports results**; 62. **Expose**; 63. **Footer**; 64. **Glossary entry**; 65. **Courses of studies list**; 66. **Medical diagnosis example**; 67. **Specification table**; 68. **Splash page**; 69. **Timetable**; 70. **Presentation manuscript**; 71. **Scientific article**; 72. **Newspaper article (scanned in)**; 73. **Access statistics**; 74. **Registration form** 0.1%; 75. **Directions**; 76. **Committee proposal**; 77. **Application form**; 78. **Medical information**; 79. **Library classification scheme (excerpt)**; 80. **Semester dates and deadlines**; 81. **Riddle**; 82. **Consent form**; 83. **Episode list (tv show)**; 84. **Errata**; 85. **Slides (thumbnails with interactive examples)**; 86. **Guestbook**; 87. **Calendar of memorial dates**; 88. **Glossary**; 89. **Image map**; 90. **City map**; 91. **Historical building data**; 92. **Plant data**; 93. **Cinema programme**; 94. **Test results**; 95. **Classified ad**; 96. **Recipe**; 97. **Club/society course directory**; 98. **Map/site plan**; 99. **List of course materials**; 100. **List of lexicon entries**; 101. **List of university projects**; 102. **List of doctoral projects within an organisational unit**; 103. **List of new or modified documents on a website**; 104. **Newsgroup (list of postings)**; 105. **Price list**; 106. **Press releases (list)**; 107. **Pupil's homepage**; 108. **Editors of a website**; 109. **Travel diary (single entry)**; 110. **Memo/circular**; 111. **Search form**; 112. **Participant list**; 113. **University newspaper (overview of one issue)**; 114. **Election results**

Table 13.2. The 114 node types found in a random sample of 750 documents

When we return to the role of HTML as a markup language, the results of this analysis are highly relevant for two reasons. First, the

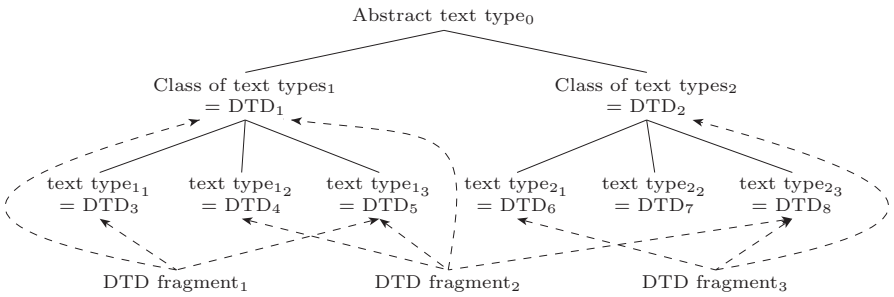

Figure 13.5. Representing a typology of text types by means of individual XML document type definitions and imported DTD fragments

study shows that hundreds of both traditional genres as well as novel web genres are used in the World Wide Web. Especially well-established genres, such as *Cook Book* and *Lexicon*, demonstrate that, in principle, one dedicated markup language is needed for every single genre to model the respective text structures' constituents. Second, the categories that have subtypes as well as closely related web genres reveal a problem that cannot be addressed with approaches such as DTDs or XML Schema descriptions. It is impossible formally to model typologies of genres, or, to be more precise, typologies of markup languages. Apart from representing identical parts of markup languages in DTD-fragments and importing these with the help of parameter entities, the family of XML standards does not provide a mechanism that is able adequately to represent a typology of document grammars (see Rehm, 2007). As figure 13.5 illustrates, the relationships that hold between similar or related text types or markup languages respectively, cannot be modelled by a set of isolated DTDs alone.

3.6 Modelling Web Genres with OWL

The two problems raised in section 3.5 can be overcome only with the help of an additional representation layer that encapsulates the relations that hold between (the constituents of) related text types. We have shown (see Rehm, 2007) that the Web Ontology Language (OWL) is an appropriate formalism for this text technological application (Rehm, 2004c). Usually, OWL is used to model traditional ontologies for knowledge representation and Semantic Web purposes (see Farrar and Langendoen, in this volume). OWL possesses several advantages such as class hierarchies, multiple inheritance, and different types of properties (features of classes or instances respectively).

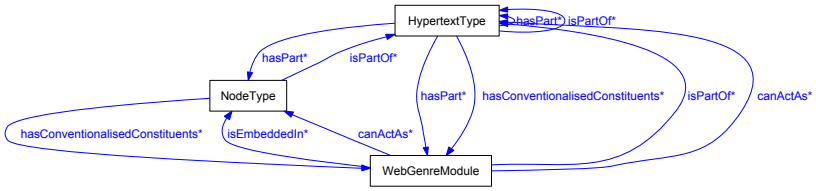

Figure 13.6. The three basic classes of the web genre ontology (see figure 13.4)

We developed a web genre ontology that is based as closely as possible on the internal structure of the web genre model (see section 3.3).[2] The ontology contains three classes named `HypertextType`, `NodeType` and `WebGenreModule` that correspond to the model's three main components. Figure 13.6 shows this basic framework along with some of the core relations that, again, match the relations included in the model.

The web genre ontology comprises several hundred classes as well as relations and it is tightly intervowen with a domain ontology that models the generic structure of a German university. A very small excerpt from the web genre ontology is depicted in figure 13.7. This figure presents the framework of the web genre typology governed by the abstract class *Homepage of a Person*. The typology contains, among others, the web genres *Academic's Personal Homepage* and *Student's Private Homepage* as subgeneric variants. As the analyses have documented, all web genres of the personal or private homepage type share common properties that can be conceptualised as a prototypical core. Due to OWL's inheritance mechanism, these prototypical features are defined in the abstract class *Homepage of a Person* and automatically propagated to all subclasses, so that exceptions and extensions can be handled individually. Furthermore, figure 13.7 shows references to specific subclasses of `NodeType`. These subclasses themselves reference subclasses of `WebGenreModule` (not shown in the figure) that have been identified and collected in empirical sample analyses (i. e., the web genre ontology reflects empirical data with regard to a particular analysis domain). Whereas the definitions of subclasses of `NodeType` mainly refer to specific configurations of web genre modules, the definitions of the subclasses of `WebGenreModule` can contain arbitrary information. The set of information associated with web genre modules can be comprised of, for example, DTD fragments, keywords, empirical data (e. g., word frequencies), comments or references to external resources. This flexibility

[2]The ontology was developed with the help of the Protégé-OWL ontology editor (see `http://protege.stanford.edu`). Figures 13.6 and 13.7 were generated using the Ontoviz plugin.

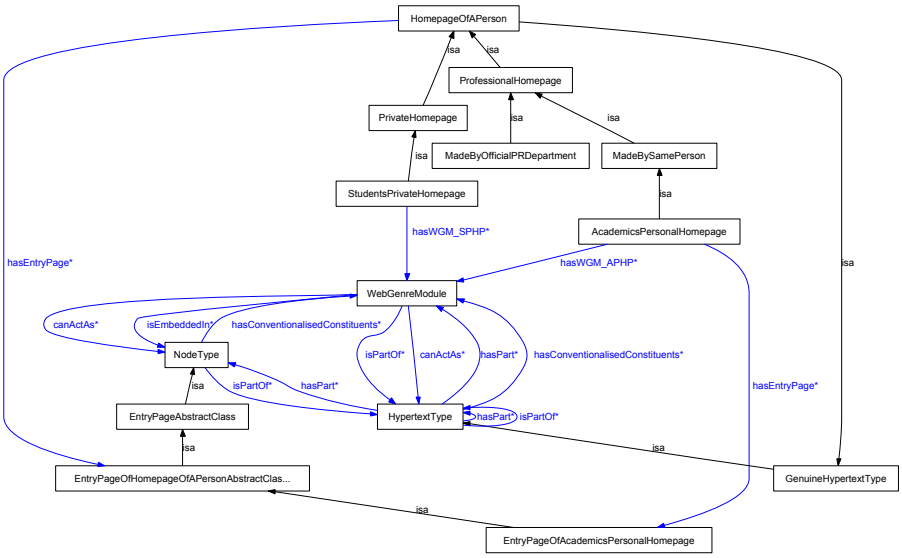

Figure 13.7. The web genres of the abstract hypertext type *Homepage of a Person*

is needed for the automatic processing of HTML documents, especially with regard to information extraction purposes or the categorization and identification of web genres. In addition, a lot of web genre modules are used in more than one web genre. To enable their re-use, definitions of web genre modules are simply referenced within the definitions of web genres and node types, so that the ontology itself is kept free of redundancies. Another challenge for the OWL-based modelling of web genres is the representation of conventions in hypertextual graph structures (see Mehler, in this volume). Currently, these conventions are modelled as an additional layer of object properties, i. e., relations between the instances of web genre modules such as *Navigation Bar* (link anchors) and corresponding web genre modules as well as node types (link targets).

The web genre ontology has several potential application scenarios (see Rehm, 2007, for a more detailed discussion). The automatic identification of web genres, for example, could be approached by categorising the macrostructural building blocks of HTML documents and mapping these objects onto web genre modules (Rehm, 2004b, also see Stede and Suriyawongkul, in this volume). Finally, the set of instantiated web genre modules could be used to compute the most probable web genre with the help of an inference engine. Another possible application involves the additional step of extracting information from instantiated

web genre modules (Rehm, 2004a): authors of web pages usually follow certain conventions that also apply to the basic informational units of web genre modules such as *Contact Information*. According to our analysis, this complex web genre module contains (in instances of the web genre *Academic's Personal Homepage*), the atomic components *E-Mail Address* (99%), *Street Address* (90%), *Phone Number* (86%), *Fax Number* (86%), *Room/Office Number* (30%), and *Office Hours* (27%). Assuming the majority of instances of a certain web genre module contains a specific set of information, it is possible to build specialised extraction engines that anticipate the presence of these sets of information and that represent the extracted data using tag sets such as `EMailAddress`, `StreetAddress`, `PhoneNumber` etc. These tag names are reflected in the abovementioned DTD fragments that can be referenced in web genre module definitions. If we assume now that the most basic structural information of every web genre module of specific web genres can be extracted by automatic means, we had a web genre-driven HTML-to-XML conversion as well as an information extraction system for a certain number of web genres.

Genres in the World Wide Web must not be viewed as restrictive and as standardised as some traditional genres, because, especially with regard to non-commercial websites, no one enforces web genres (i. e., certain rules and conventions) and no author has to fear any negative consequences should he or she breach the rules. Defining web genres by means of an OWL ontology makes allowance for the flexibility inherent in the web genre concept. Web genres consist of compulsory and optional web genre modules so that it is possible to compile the corresponding DTD fragments into an overall document grammar for a markup language that can be used to represent instances of the entire web genre (see Erdmann, 2001). Such a tool must be able to interpret the object properties employed to relate compulsory and optional web genre modules to a specific web genre and should be able to generate document grammars in several formats (DTD, XML Schema, Relax NG etc.).

4. Concluding Remarks

With regard to the very close relationship between genres (i. e., text types) and markup languages, a peculiar situation exists in the World Wide Web. Usually, the generalised discourse structure of a genre or text type is modelled in the form of one specific document grammar or markup language that authors apply for the process of writing a document belonging to this very genre. In the World Wide Web, however, we have the situation that *one* markup language, HTML, is used to

annotate and to design documents of hundreds of web genres. For this purpose, HTML and its companion standards (CSS etc.) are bent as far as possible in order to produce innovative and appealing designs, generally by employing visual tools (for example, WYSIWYG editors) instead of traditional SGML/XML editors. Therefore, in practice, HTML is used more like a page description language, such as Postscript, than as a genuine markup language that is rooted in the SGML tradition.

Hypertext itself is neither a genre nor a text type. Instead, it is – in the incarnation of the World Wide Web – a well-established means of communication that has spawned dozens of novel web genres. Furthermore, hundreds, if not thousands, of traditional genres are used throughout the World Wide Web, primarily by automatically converting existing text documents into HTML or by copying and pasting existing content into the editors of content management systems (Rehm, 2007). This observation has several consequences for current research trends in linguistics and computational linguistics. An important trend is the "web as corpus" approach (Kilgarriff, 2001, Mehler et al., 2008). In their overview and introduction, Kilgarriff and Grefenstette (2003, p. 342) complain about a "lack of theory of text types". Later, the authors mention one of the most central problems:

> "Text type" is an area in which our understanding is, as yet, very limited. Although further work is required irrespective of the web, the use of the web forces the issue. Where researchers use established corpora, such as Brown, the BNC, or the Penn Treebank, researchers and readers are willing to accept the corpus name as a label for the type of text occurring in it without asking critical questions. Once we move to the web as a source of data, and our corpora have names like "April03-sample77", the issue of how the text type(s) can be characterized demands attention. (Kilgarriff and Grefenstette, 2003, p. 343)

The situation can be approached from the opposite point of view as well: how can a web-based corpus of cook books, software manuals, guidelines for student papers or conference websites be constructed? These are but a tiny fraction of examples of established text types or genres in the World Wide Web. We need methods that enable us to identify genres automatically and to filter corpus collection processes based on that data. In other words: we do not want to label our corpus snapshot "April03-sample77" but, instead, we would like to use meaningful tags that, ideally, refer to a standardised inventory of web genre labels (Rehm, 2008).

The automatic identification of web genres requires very sophisticated and robust methods. Lim et al. (2005), for example, present a supervised machine-learning approach that uses a total of 329 features to classify documents into 16 genres (e. g., "public homepages", "commercial home-

pages", "simple tables/lists", "input pages", "official materials", "informative materials" and "others"; the genres are based upon the inventory prepared by Dewe et al., 1998, also see Levering et al., 2008, Kim and Ross, 2008). Although the approach by Lim et al. does operate with a precision of about 75%, it fails to address the key aspects of web genres. All the existing approaches to genre identification treat HTML documents as monolithic instantiations of web genres, i. e., one document is the manifestation of exactly one web genre (but see Mehler et al., 2004). Both the level of web genre modules as well as the superordinate level of web genres that consist of specific node types are ignored completely. Furthermore, all existing approaches lack a theoretical foundation (for example, neither the concept of web genres, nor the inventory of genres used are thoroughly discussed or motivated). What is needed is an approach that goes beyond the individual HTML document and analyses the hypertextual structure of component documents (Eiron and McCurley, 2003). In addition, the macrostructure of every HTML document needs to be processed in order to find instances of multiple web genre modules that might be used simultaneously in a single document. In Rehm (2005), the prototype of a parser for arbitrary HTML documents is presented. The system parses the HTML element tree and tries to identify the textual macrostructure's basic building blocks that can, in turn, be mapped onto web genre modules. First, the HTML document is converted to XHTML, so that it can be processed with XML tools (Myllymaki, 2001). A DOM and XPath engine is used to analyse recursively the element tree based on the "visual semantics" of all HTML elements and attributes. Their visual effect rather than the elements themselves need to be taken into account due to the problem of tag abuse (Barnard et al., 1996). The results of the analysis (purely structure-oriented tags that encapsulate individual HTML element subtrees with information about lists, headlines, separators, paragraphs of running text etc.) are stored within the scope of a new namespace directly in the converted XHTML document.

With regard to taking web genres into consideration from a theoretical as well as practical point of view, several important questions need to be addressed in the near future: genre rules and conventions stretch beyond the individual document, which is why approaches that treat single HTML documents fail in capturing the reality of everyday communication on the web. Furthermore, an extensible set of standardised web genre names needs to be established (see Rehm, 2007, for an initial proposal) and, ideally, accompanied by a reference corpus (Rehm, 2008, Rehm et al., 2008). An additional problem is introduced by the trend of bringing applications into the web. User interface programming

paradigms such as Ajax enable truly dynamic applications and operate by replacing individual subtrees of the DOM structure by means of JavaScript's XMLHttpRequest object instead of refreshing the whole (X)HTML document each time the user initiates an action. This technique severely challenges the concept of HTML *documents*, so that a distinction between HTML documents and HTML-based *applications* will be needed in the future.

Bibliography

Barnard, David T.; Burnard, Lou; DeRose, Steven J.; Durand, David G. and Sperberg-McQueen, C.M. (1996): "Lessons for the World Wide Web from the Text Encoding Initiative". *The World Wide Web Journal* 1 (1): pp. 349–357.

Bazerman, Charles (1994): "Systems of Genres and the Enactment of Social Intentions". In: *Genre and the New Rhetoric*, edited by Freedman, Aviva and Medway, Peter, London: Taylor and Francis, pp. 79–101.

Bray, Tim; Paoli, Jean; Sperberg-McQueen, C. M.; Maler, Eve; Yergeau, François and Cowan, John (2004): "Extensible Markup Language (XML) 1.1". Technical Specification, W3C. `http://www.w3.org/TR/2004/REC-xml11-20040204/`.

Brinker, Klaus; Antos, Gerd; Heinemann, Wolfgang and Sager, Sven F. (editors) (2000): *Text- und Gesprächslinguistik*, volume 16.1 of *Handbücher zur Sprach- und Kommunikationswissenschaft (HSK)*. Berlin, New York: de Gruyter.

Crowston, Kevin and Williams, Marie (2000): "Reproduced and Emergent Genres of Communication on the World Wide Web". *The Information Society* 16 (3): pp. 201–215.

Dewe, Johan; Karlgren, Jussi and Bretan, Ivan (1998): "Assembling a Balanced Corpus from the Internet". In: *Proceedings of the 11th Nordic Conference of Computational Linguistics*. Copenhagen, pp. 100–107.

Eckkrammer, Eva Martha (2001): "Textsortenkonventionen im Medienwechsel". In: *E-Text: Strategien und Kompetenzen – Elektronische Kommunikation in Wissenschaft, Bildung und Beruf*, edited by Handler, Peter, Frankfurt/Main, Berlin, Bern etc.: Peter Lang, volume 7 of *Textproduktion und Medium*, pp. 45–66.

Eiron, Nadav and McCurley, Kevin S. (2003): "Untangling Compound Documents on the Web". In: *Proceedings of the 14th ACM Conference on Hypertext and Hypermedia*. pp. 85–94. Nottingham.

Emigh, William and Herring, Susan C. (2005): "Collaborative Authoring on the Web: A Genre Analysis of Online Encyclopedias". In: *Proceedings of the 38th Hawaii International Conference on Systems Sciences (HICSS-38)*. Big Island, Hawaii.

Erdmann, Michael (2001): *Ontologien zur konzeptuellen Modellierung der Semantik von XML*. Ph. d. thesis, University of Karlsruhe, Karlsruhe.

Erickson, Thomas (2000): "Making Sense of Computer-Mediated Communication (CMC): Conversations as Genres, CMC Systems as Genre Ecologies". In: *Proceedings of the 33rd Hawaii International Conference on Systems Sciences (HICSS-33)*.

Eriksen, Lars Bo and Ihlström, Carina (1999): "In the Path of the Pioneers – Longitudinal Study of Web News Genre". In: *Proceedings of the 22nd Information Systems Research Seminar in Scandinavia (IRIS 22): "Enterprise Architectures for Virtual Organizations"*, edited by Käkölä, Timo K. University of Jyväskylä, Keuruu, pp. 289–304.

Furuta, Richard and Marshall, Catherine C. (1996): "Genre as Reflection of Technology in the World-Wide Web". In: *Hypermedia Design, Proceedings of the International Workshop on Hypermedia Design (IWHD 1995)*, edited by Fraïssé, Sylvain; Garzotto, Franca; Isakowitz,

Tomás; Nanard, Jocelyne and Nanard, Marc, Berlin, Heidelberg, New York etc.: Springer, Workshops in Computing, pp. 182–195.

Gleim, Rüdiger; Mehler, Alexander; Eikmeyer, Hans-Jürgen and Rieser, Hannes (2007): "Ein Ansatz zur Repräsentation und Verarbeitung großer Korpora". In: *Datenstrukturen für linguistische Ressourcen und ihre Anwendungen – Data Structures for Linguistic Resources and Applications: Proceedings of the Biennial GLDV Conference 2007*, edited by Rehm, Georg; Witt, Andreas and Lemnitzer, Lothar, Tübingen: Gunter Narr, pp. 275–284.

Haas, Stephanie W. and Grams, Erika S. (1998): "Page and Link Classifications: Connecting Diverse Resources". In: *Proceedings of Digital Libraries '98 – Third ACM Conference on Digital Libraries*, edited by Witten, I.; Akscyn, R. and Shipman, F. Pittsburgh, pp. 99–107.

Haas, Stephanie W. and Grams, Erika S. (2000): "Readers, Authors, and Page Structure – A Discussion of Four Questions Arising from a Content Analysis of Web Pages". *Journal of the American Society for Information Science* 51 (2): pp. 181–192.

Hammwöhner, Rainer (1997): *Offene Hypertextsysteme – Das Konstanzer Hypertextsystem (KHS) im wissenschaftlichen und technischen Kontext*. Number 32 in Schriften zur Informationswissenschaft. Konstanz: Universitätsverlag Konstanz.

ISO 8879 (1986): "Information Processing – Text and Office Information Systems – Standard Generalized Markup Language". Internationaler Standard, International Organization for Standardization, Genf.

Jakobs, Eva-Maria (2003): "Hypertextsorten". *Zeitschrift für germanistische Linguistik* 31 (2): pp. 232–252.

Kilgarriff, Adam (2001): "Web as Corpus". In: *Proceedings of the Corpus Linguistics 2001 Conference*, edited by Rayson, Paul; Wilson, Andrew; McEnery, Tony; Hardie, Andrew and Khoja, Shereen. Lancaster, pp. 342–344.

Kilgarriff, Adam and Grefenstette, Gregory (2003): "Introduction to the Special Issue on the Web as Corpus". *Computational Linguistics* 29 (3): pp. 333–348.

Kim, Yunhyong and Ross, Seamus (2008): "Examining Variations of Prominent Features in Classification". In: *Proceedings of the 41st Hawaii International Conference on Systems Sciences (HICSS-41)*. Big Island, Hawaii.

Kuhlen, Rainer (1991): *Hypertext – Ein nicht-lineares Medium zwischen Buch und Wissensbank*. Berlin, Heidelberg, New York etc.: Springer.

Levering, Ryan; Cutler, Michal and Yu, Lei (2008): "Using Visual Features for Fine-Grained Genre Classification of Web Pages". In: *Proceedings of the 41st Hawaii International Conference on Systems Sciences (HICSS-41)*. Big Island, Hawaii.

Lim, Chul Su; Lee, Kong Joo and Kim, Gil Chang (2005): "Multiple Sets of Features for Automatic Genre Classification of Web Documents". *Information Processing and Management* 41 (5): pp. 1263–1276.

Lobin, Henning (2000): "Service-Handbücher – Linguistische Aspekte im Document Lifecycle". In: *Raum, Zeit, Medium – Sprache und ihre Determinanten. Festschrift für Hans Ramge*, edited by Richter, Gerd; Riecke, Jörg and Schuster, Britt-Marie, Darmstadt: Hessische Historische Kommission, pp. 791–808.

Lobin, Henning (2001): *Informationsmodellierung in XML und SGML*. Berlin, Heidelberg, New York etc.: Springer.

Maler, Eve and Andaloussi, Jeanne El (1996): *Developing SGML DTDs – From Text to Model to Markup*. Upper Saddle River: Prentice Hall.

Mehler, Alexander; Dehmer, Matthias and Gleim, Rüdiger (2004): "Towards Logical Hypertext Structure — A Graph-Theoretic Perspective". In: *Proceedings of the Fourth International Workshop on Innovative Internet Computing Systems (I2CS '04)*, edited by Böhme, Thomas and Heyer, Gerhard. Berlin, New York: Springer, Lecture Notes in Computer Science.

Mehler, Alexander; Sharoff, Serge; Rehm, Georg and Santini, Marina (editors) (2008): *Genres on the Web: Computational Models and Empirical Studies*. In preparation.

Miller, Carolyn R. (1984): "Genre as Social Action". *Quarterly Journal of Speech* (70): pp. 151–167.

Myllymaki, Jussi (2001): "Effective Web Data Extraction with Standard XML Technologies". In: *Proceedings of the 10th International World Wide Web Conference (WWW-10)*. Hong Kong, pp. 689–696.

Orlikowski, Wanda J. and Yates, JoAnne (1994): "Genre Repertoire: The Structuring of Communicative Practices in Organizations". *Administrative Science Quarterly* (39): pp. 541–574.

Pemberton, Steven (2002): "XHTML 1.0: The Extensible Hypertext Markup Language (Second Edition)". Technical Specification, W3C. `http://www.w3.org/TR/xhtml1/`.

Raggett, Dave; Hors, Arnaud Le and Jacbos, Ian (1999): "HTML 4.01 Specification". Technical Specification, W3C. `http://www.w3.org/TR/html401/`.

Rehm, Georg (2001): "`korpus.html` – Zur Sammlung, Datenbank-basierten Erfassung, Annotation und Auswertung von HTML-Dokumenten". In: *Proceedings of the GLDV Spring Meeting 2001*, edited by Lobin, Henning. Gesellschaft für linguistische Datenverarbeitung (Society for Computational Linguistics and Language Technology), Giessen, Germany, pp. 93–103. `http://www.uni-giessen.de/fb09/ascl/gldv2001/`.

Rehm, Georg (2002): "Towards Automatic Web Genre Identification – A Corpus-Based Approach in the Domain of Academia by Example of the Academic's Personal Homepage". In: *Proceedings of the 35th Hawaii International Conference on System Sciences (HICSS-35)*. Big Island, Hawaii.

Rehm, Georg (2004a): "Hypertextsorten-Klassifikation als Grundlage generischer Informationsextraktion". In: *Automatische Textanalyse – Systeme und Methoden zur Annotation und Analyse natürlichsprachlicher Texte*, edited by Mehler, Alexander and Lobin, Henning, Wiesbaden: Verlag für Sozialwissenschaften, pp. 219–233.

Rehm, Georg (2004b): "Ontologie-basierte Hypertextsorten-Klassifikation". In: *Automatische Textanalyse – Systeme und Methoden zur Annotation und Analyse natürlichsprachlicher Texte*, edited by Mehler, Alexander and Lobin, Henning, Wiesbaden: Verlag für Sozialwissenschaften, pp. 121–137.

Rehm, Georg (2004c): "Texttechnologische Grundlagen". In: *Computerlinguistik und Sprachtechnologie – Eine Einführung*, edited by Carstensen, Kai-Uwe; Ebert, Christian; Endriss, Cornelia; Jekat, Susanne; Klabunde, Ralf and Langer, Hagen, Heidelberg: Spektrum, pp. 138–147. 2nd edition.

Rehm, Georg (2005): "Language-Independent Text Parsing of Arbitrary HTML-Documents – Towards A Foundation For Web Genre Identification". *LDV Forum* 20 (2): pp. 53–74.

Rehm, Georg (2007): *Hypertextsorten: Definition – Struktur – Klassifikation*. Norderstedt: Books on Demand. (Ph. D. thesis in Applied and Computational Linguistics, Giessen University, 2005).

Rehm, Georg (2008): "A Comparative Analysis of Genre Category Sets as a Prerequisite for a Reference Corpus of Web Genres". In: *Genres on the Web: Computational Models and Empirical Studies*, edited by Mehler, Alexander; Sharoff, Serge; Rehm, Georg and Santini, Marina. In preparation.

Rehm, Georg and Santini, Marina (editors) (2007): *Proceedings of the International Workshop Towards Genre-Enabled Search Engines: The Impact of Natural Language Processing*, Borovets, Bulgaria. Held in conjunction with RANLP 2007.

Rehm, Georg; Santini, Marina; Mehler, Alexander; Braslavski, Pavel; Gleim, Rüdiger; Stubbe, Andrea; Symonenko, Svetlana; Tavosanis, Mirko and Vidulin, Vedrana (2008): "Towards a Reference Corpus of Web Genres for the Evaluation of Genre Identification Systems". In: *Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008)*. Marrakech, Morocco.

Reiss, Eric L. (2000): *Practical Information Architecture – A Hands-On Approach to Structuring Successful Websites*. Harlow, London, New York etc.: Addison-Wesley.

Rosso, Mark A. (2005): *Using Genre to Improve Web Search*. Ph. D. thesis, School of Information and Library Science, University of North Carolina at Chapel Hill.

Ryan, Terry; Field, Richard H. G. and Olfman, Lorne (2003): "The evolution of US state government home pages from 1997 to 2002". *International Journal of Human-Computer Studies* 59 (4): pp. 403–430.

Santini, Marina (2007): "Characterizing Genres of Web Pages: Genre Hybridism and Individualization". In: *Proceedings of the 40th Hawaii International Conference on Systems Sciences (HICSS-40)*. Big Island, Hawaii.

Shepherd, Michael and Watters, Carolyn (1998): "The Evolution of Cybergenres". In: *Proceedings of the 31st Hawaii International Conference on Systems Sciences (HICSS-31)*. volume 2, pp. 97–109.

Shepherd, Michael and Watters, Carolyn (1999): "The Functionality Attribute of Cybergenres". In: *Proceedings of the 32nd Hawaii International Conference on Systems Sciences (HICSS-32)*.

Storrer, Angelika (2004): "Text und Hypertext". In: *Texttechnologie – Anwendungen und Perspektiven*, edited by Lobin, Henning and Lemnitzer, Lothar, Tübingen: Stauffenburg, Stauffenburg Handbücher, pp. 13–49.

Swales, John M. (1990): *Genre Analysis – English in academic and research settings*. The Cambridge Applied Linguistics Series. Cambridge: Cambridge University Press.

Walker, Derek (1999): "Taking Snapshots of the Web with a TEI Camera". *Computers and the Humanities* 33 (1–2): pp. 185–192.

Yates, Joanne and Orlikowski, Wanda J. (1992): "Genres of Organizational Communication: A Structurational Approach to Studying Communication and Media". *Academy of Management Review* 17 (2): pp. 299–326.

Yates, Simeon J. and Sumner, Tamara R. (1997): "Digital Genres and the New Burden of Fixity". In: *Proceedings of the 30th Hawaii International Conference on Systems Sciences (HICSS-30)*. volume 6, pp. 3–12.

Yoshioka, Takeshi; Herman, George; Yates, JoAnne and Orlikowski, Wanda (2001): "Genre Taxonomy". *ACM Transactions on Information Systems* 19 (4): pp. 431–456.