Towards Automatic Web Genre Identification

A Corpus-Based Approach in the Domain of Academia by Example of the Academic's Personal Homepage

Georg Rehm
Research Unit for Applied and Computational Linguistics
Otto-Behaghel-Str. 10 D, Justus-Liebig-Universität, 35394 Giessen, Germany
Georg.Rehm@uni-giessen.de

Abstract

We argue for a systematic analysis of one particular, well structured domain—academic Web pages—with regard to a special class of digital genres: *Web genres*. For this purpose, we have developed a database-driven system that will ultimately consist of more than 3 000 000 HTML documents, written in German, which are the empirical basis for our research. We introduce the notions of *Web genre type* which constitutes the basic framework for a certain Web genre, and *compulsory* and *optional* Web genre *modules*. These act as building blocks which go together to make up the structure characterised by the Web genre type and furthermore, operate as modifiers for the default <content, form, function> assignment involved. The analysis of a 200 document sample illustrates our notion of *Web genre hierarchy*, into which Web genre types and modules are embedded. The analysis of four different documents of the Web genre *Academic's Personal Homepage*, not only illustrates our approach, but also our long-term goal of automatically extracting the contents of Web genre modules in order to build up *structured* XML documents of groups of *unstructured* HTML documents.

1 Introduction

Nowadays, there seems to be a consensus emerging in favor of the evolution of new digital genre systems on the World Wide Web. Traditional as well as digital genres, have been studied by scholars concentrating on a specific form of genre theory [24, 37, 27] which emphasizes the impact of recurring communicative situations within discourse communities, and characterizes genres by means of the triple <content, form, function>.

Automatic Web genre identification (AWGI) is one of the key factors in improving the often inadequate results of search engines, as the user would be able to specify the desired *Web genre* along with a set of keywords.[1] Several prerequisites have been partially approached: one genre in particular, the personal homepage, has been analysed with regard to several key features [10]. In other studies, small samples of literally all types of documents (from commercial, private, academic and other domains) have been randomly selected with the help of search engines, whereupon the documents were classified into broad sets of genres [18]. These all-encompassing approaches are—with regard to the heterogenous diversity of Web genres—inevitably rather coarse and incomplete concerning the set of distinct features that constitute a certain genre resp. group of genres.

Our approach concentrates on a domain restricted enough to exclude a lot of problematic "genres", yet broad enough to precisely identify a Web genre hierarchy: the relatively stable domain of academic Web pages. Currently, a corpus of 3 000 000 Web pages from German universities is being constructed. From this corpus, four sample-documents of the Web genre *Academic's Personal Homepage* were selected to illustrate our feature-based AWGI-approach which relies on the novel notion of *Web genre types*, which are composed of *Web genre modules*. An additional goal of our project is the automatic extraction of information likewise based on the notions of Web genre types and modules, formally specified by XML Schema definitions within a Web genre hierarchy framework, illustrated by the analysis of a 200 document sample.

2 The State of the Art

Most studies presented within the Digital Genre community deal with specific genres. Crowston and Williams [7] examine different uses of hyperlinking in FAQ documents. Eriksen and Ihlström [12] studied three digital newspapers over a period of three years and found that these differ from their paper cousins in several respects. Fortanet et al. [14] identifiy computer-related target ads as a subgeneric variation of the "netvertising" genre. A very thoroughly studied digital genre is the personal homepage: Walters [39] conducted a survey in which she analysed 100 students' homepages. Although she did find distinct categories (*professional* vs. *interest page*), these could not be considered as belonging to genres: "in practice, few homepages actually have a specific purpose." Furuta and Marshall [15] regard "representation and construction of self on the Internet" as a primary communicative purpose: homepages often contain personal information, a portrait of the

[1] Applications of genre detection systems in a Computational Linguistics context, e. g., parsing or part-of-speech tagging, are listed in [21].

author, as well as biographical details. They conclude that the "accessible document structure allows authors to assimilate conventions quickly", Erickson [11] supports this view of homepages "being used to construct identity." de Saint-Georges [9] analyses 38 students' homepages with regard to deictic linguistic elements and gives a detailed definition of the "personal homepage". Roberts [32] examines narrative clauses in 41 students' homepages, and considers hyperlinks being narrative clauses that "allow the author to maintain discourse cohesion". Amitay [1] analyzes 1 000 personal homepages with regard to general features of hyperlinks and word frequency: "it is not surprising that the words *I*, *my* and *me* would be very high in the list. However the word *you* is also placed very high thus indicating a tendency to use direct and informal language – *from me* [the author] *to you* [the reader]." Dillon and Gushrowski [10] collected more than 100 documents from homepage repositories and analysed these regarding title, e-mail address, etc. From the most and least frequent elements, new pages were created and presented to 57 subjects who had to name elements crucial to good homepages: there "is a correlation between features selected [. . .] and the frequency of features that appear on existing pages", which is an indication that the personal homepage has established itself as a unique digital genre: "personal home pages [. . .] seem to have evolved very quickly into a standard form [. . .]. Added to this, users' preferences [. . .] correlate positively with the presence or absence of these key common elements." [10]. Crowston and Williams [6] studied to what extent the acceptance of the Web results in the adaptation of existing genres or the emergence of novel ones. 100 English documents were randomly selected by means of AltaVista's *Surprise* function. The sample contains documents from 12 different countries—from the commercial, scientific, and governmental area—that were classified into 48 genres. The results are manifold: first, Crowston and Williams state that 82 of the 100 documents are "more or less faithfully reproduced genres [. . .] familiar in traditional media". In contrast, the "hotlist", "home page", "Web server statistics" and "letter column", were identified as being novel. Second, a lot of documents were deep nodes of larger hypertext networks. Third, the genres of three documents could not be named as their purposes could not be determined. In a follow-up study, the authors extensively analysed 837 documents [8]. At least 64 genres were identified (the total number is not given) which were partly grouped into a hierarchy based on the structure of the *Art and Architecture Thesaurus*. Shepherd and Watters [35] discuss cybergenres, specified by <content, form, functionality>. A fuzzy taxonomy is proposed that helps in characterizing the evolution of certain cybergenres, e. g., whether it is either extant or novel. "News" and "math dictionary" are discussed and integrated into the taxonomy at different evolutionary stages. In their follow-up work [36], the authors concentrate on understanding and defining the functionality attribute. 96 Web pages, randomly chosen by `random.yahoo.com`, were examined and classified into home page (40% of the documents), brochure (17%), resource (35%), catalogue (5%), search engine (figures n. a.), and game (3%). The authors "recognize that there are more specific categories [. . .], for example, the personal home page and the corporate home page." These six abstract classes contain subgeneric variations, described by very abstract values, e. g., "information about person", "subject-specific information", "challenge to user" (content), "hierarchical", "video", "query box", "scenes" (form), "browsing", "e-mail", "interaction" (functionality). Shepherd and Watters [36] are of the opinion that "there are actually relatively few classes of cybergenres on the Web". Haas and Grams [18, 17, 19] propose a classification system for documents and links, based on an analysis of 75 randomly selected English documents (by means of AltaVista's *Surprise* function) and their 1 500 links. The authors regard the following characteristics of a page as central regarding a "classification system for Web page types": function, intended audience, content or format, types of links it contains, and relationship to the pages to which it provides links. The classification system contains seven major categories, each comprising subcategories: Organizational, Documentation, Text, Home Page, Multimedia, Tools, Database Entry. Haas and Grams [19] identified four major groups of links: Navigation, Expansion, Resource, and Miscellaneous. Furthermore, correlations were identified with regard to certain page types and the link types these documents contain: "table of contents" pages often contain within-document "navigation" links, "index" pages often contain "resource" links. Roussinov et al. [33] emphasize on using Web page genres to improve navigation. They present a user study, conducted by interviewing 184 students while browsing the Web. 1 234 Web pages were collected, 1 076 were classified, resulting in a total of 116 genres. The most often stated purpose for searching the Web was "scholarly research" (22.95%). Furthermore, five genre groups were identified: Topics, Publications, Products, Educational Material, and FAQ. Additionally, sets of intuitively refined "recognition indicators" were assigned to these groups. Toms and Campbell [38] hypothesize that a document genre can be described by a "parsimonious set of attributes." They suggest that a document instance can be characterised on the levels *function* (semantic content), *form* (visual appearance), and *interface* (means of access). A study was conducted to detect central features associated to specific genres, both paper-based and digital (journal article, reading list, memo, dictionary, minutes, course calendar). To test for the recognition of the form, the text was masked (substituted by x, X and 9). To test for function, the layout was masked (the text was transformed into a flat sequence of words). The authors conclude that "the visual cues [. . .] act in tandem with the semantic content to influence the user during the crucial seconds of initial exposure", and "document structure can be used as a means of identifying documents [. . . Those] same cues that make a document immediately identifiable in the paper world are readily transferrable to the digital world." Karlgren and Cutting [20] present an approach for "text genre recognition" based on descriptive statistics. Target documents are (originally non-digital) texts from the Brown corpus. Experiments were conducted by slicing pre-selected documents into partitions by means of genre analysis. For this purpose, a set of discriminant functions using parameters of pre-categorized sets was utilized. These can be used to classify new documents once their parameters are extracted. In the first study, 500 texts were divided into two categories ("informative", "imaginative") with a total of 22 incorrect allocations. In the second study, 500 texts were divided into four categories ("press", "non-fiction", "fiction", "misc.") with a total of 134 errors. Kessler et al. [21] present another approach, likewise using 499 texts of the Brown corpus. The authors work with features which are assigned to sets of extractable parameters (generic cues). By means of logistic regression methods, predictor functions and neural nets were developed. The results are very promising with about 90% precision for the reportage and fiction genre, whereas editorial and legal texts are hard to categorize. Roussinov et al. [33] describe the only approach to automatic genre detection in the Web context we currently know of, but this approach has not yet been implemented. Their objective is to identify five to six "major groups of Web genre".

3 Automatic Web Genre Identification

3.1 Web Genre Types and Web Genre Modules

Crowston and Williams [6] illustrate their thesis that genres form a hierarchy: the *social science paper* belongs to the *research paper* genre, which is a type of *paper*. Other types within *research paper* combine identical features, e. g., name of author(s), title, and bibliography, whereas other features are variations, e. g., expected section headings. Toms and Campbell [38] note that a "concept of taxonomic families of documents based on document structure" exists. Certain masked paragraph structures in their study were identified as a series of bibliographic citations, whereas other participants directly identified the whole document as an instance of a reading list that *contains* a bibliography. The difference between these two notions is that Crowston and Williams [6] assume an IS-A relationship between different genre levels (*social science paper* IS-A *research paper* IS-A *paper*), whereas the example in [38] implies a partitive CONTAINS taxonomy. The complex issue of *multiple* genre hierarchies is closely related to the question of whether Web pages are discrete monolithic entities, or whether they consist of modular "building blocks" [19].

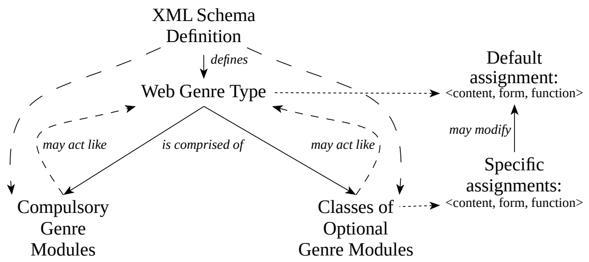

XML Schema Definition
defines
Web Genre Type
Default assignment: <content, form, function>
may modify
Specific assignments: <content, form, function>
may act like
is comprised of
may act like
Compulsory Genre Modules
Classes of Optional Genre Modules

Figure 1. Web genre types and modules

In several of the studies cited in sect. 2, the authors noted that assigning a digital genre to a specific document was often very difficult, due to its diverse constituent elements [19]. Thus, the assumption that all HTML documents are monolithic entities with regard to the notion of genre are, in our opinion, ungrounded. Building on the work of Haas and Grams, we assume that primitive, *generalized Web genre types* exist which constitute the basic framework of a certain Web genre at its most abstract level. This framework comprises one or more *compulsory genre modules*. Additionally, it can be extended by various other modules whose concrete occurrence in the *Web genre instance* (i. e., in a document that belongs to this Web genre), marked by the framework is *optional*.

The status of a genre module can vary on several levels: first, it can be obligatory or optional with regard to the respective Web genre type. Second, it can be central to the function of a document (e. g., the author's name on a personal homepage), it can be a navigational element, or of negligible importance (e. g., the *last modified* date). Thus, we propose that the triple of <content, form, function> can, and should, be assigned to generalized *classes of optional genre modules* too, in order to lay the foundation for the overall, i. e., document-wide, assignment of these features which we can compute, based on the attribute-values of the involved instances of optional genre modules.[2] Using this approach, we can modify the default <content, form, function> assignment of the underlying Web genre type to allow for variations. Third, certain genre modules can act as generalized frameworks themselves, e. g., the *list of publications*, which exists as a compulsory module in the Web genre type *Academic's Personal Homepage* (cf. fig. 1).

[2] This is based on Frege's Principle of Compositionality: a complex expression's meaning is a function of the meaning of its constituent parts.

An instance of this module can either occur directly in the physical document that consitutes the academic's personal homepage, or it exists as an HTML document on its own, linked from the personal homepage. Thus, the issue of cross-document linking is a central one within our concept of genre types and genre modules.

3.2 A Corpus of Academic Web Pages

As previous approaches have shown (cf. sect. 2), multiple Web genres do exist. The number of identified genres so far, is impressive, but, in our opinion, the Internet-wide random sample-generation performed in all-encompassing approaches (including academic, commercial and governmental pages etc.) with virtually no restrictions whatsoever, inherently leads to results which are too broad, and rather vague. Consequently, suggested recognition indicators are too shallow to be implementable. Therefore, we propose a significantly restricted approach to sampling and analyzing data by means of focusing on one certain thematic domain: academic Web servers. These are a favorite research object in AI studies [5, 25], due to their extreme structuring in terms of content and linking: universities consist of administrative units and departments or faculties, which are divided into institutes resp. research units, where, in turn, different professors, research fellows, student assistants and administrative staff, work. Thematic topics focus on research, education, and administration. Almost every university possesses this structure, which is reflected in the structuring of academic Web servers and which leads to a certain ubiquitious Web genre system [3] in academia. We explore this thesis in the domain of German academic Web pages for which we are currently building up a corpus.[3]

The initial step was to set up a database-driven system for automatic data collection. At first, we shuffled a list of the central entry points of all German universities to create a random sequence. Then, we started collecting data with the help of Open Source software [31] and custom tools: a highly configurable spider program is given one single starting point (e. g., `http://www.uni-giessen.de`), whereby several restrictions apply: only textual documents of a small set of MIME types (most importantly, `text/html`) and a size of less than 500 kb are to be incorporated into the corpus. For all the other files (images, audio/video data etc.), only the HTTP response header information is stored in the MySQL-database for statistical reasons. A further requirement is that exclusively documents written in German are collected. For this purpose, we have implemented a language identification [26] tool that is able to distinguish between documents written in *German* resp. *Unknown* with a precision of about 97%. Further technical details are discussed in [29]. At the time of writing, the corpus contains data from 28 different German universities with a total of 4 204 unique HTTP server/port combinations, 5 415 131 records and 1 300 510 locally stored HTML documents. We expect the final version of the corpus to contain about 3 000 000 documents, from some 50 different universities In order to enable intuitive and distributed methods of access, automatic sample generation, and feature annotation, a Web-based corpus-frontend is being developed in PHP. The user can navigate by means of several modes: one may randomly pick a document from the current server or the full corpus, one can select a university, request a list of its servers, select one, and then retrieve document lists. Another option is to search for substrings contained in the server name or in the path-component of URLs. Then, the user can trigger software modules to analyze a document.

As we have to manually perform extensive analyses on an empirical basis, we implemented a method of generating and saving random samples of arbitrary size, applying several restrictions (length of URL, sub-strings in URL and server name, directory-depth etc.). A sample can be analyzed inside the frontend with the help of HTML forms, results can be stored in the database for further examination. An initial sample of 200 documents has been analyzed in order to formulate an initial web genre taxonomy (see sect. 3.6). Moreover, two larger samples will be examined semi-automatically: The first one will contain 1 000 'lower level' documents, the second one will comprise the 'upper level' entry points of all the universities contained in the corpus, as well as the documents directly accessible from these. Thus, we can characterize the academic Web genre system in a combined top-down-bottom-up-manner.

3.3 The "Academic's Personal Homepage"

Four documents of the Web genre *Academic's Personal Homepage* (cf. the screenshots[4] depicted in fig. 5 to 8) comprise the sample we use to illustrate our goals. We defined requirements for a document to be included in the sample and collected these from the corpus by means of the frontend (see sect. 3.2): (a) we wanted to include four *Academics' Personal Homepages* by intuitively assessing whether a document belongs to this genre or not, (b) the documents should be available in both German and English, (c) not use frame sets, (d) the respective academics should work in different areas and (e) in different cities.[5]

[3] This approach is backed by Roberts [32]: "A study [of the entire WWW] would involve creating and maintaining an electronic corpus [. . .] and developing tools to efficiently analyze these data."

[4] These have been taken using Communicator 4.73 on NT 4.0 with a tool that is capable of auto-scrolling long windows. To decrease the height of fig. 8, we removed small amounts of superfluous whitespace.

[5] We did *not* look for documents as similar as possible, rather we used the first four documents we found by iteratively exploring the corpus.

The four sample-documents fulfil multiple *functions*: all the authors would first like to introduce themselves by noting their name, accompanied by a portrait photograph, and to *establish an individual scientific profile* by specifying professional achievements and scientific as well as administrative activities, by listing their publications[6], C. V., research projects, a list of talks and presentations, and current as well as past educational courses taught. As all documents are available in both German and English, we can assume the authors would like to target all interested parties world-wide. Examining the separate course-related documents on Andreas Neumann's homepage, we find course-descriptions, assignments, solutions, slides etc., which means that one of the functions of his homepage is to *distribute course-related material to students*.[7] His page works as a digital replacement for materials on special reserve, which, for the author, are faster to create and easier to maintain, and for the student, easier to access. The other three documents only list titles of courses, lectures and dates, sometimes descriptions. An additional purpose of the documents is to *make contact information available* to students, guests, or visitors: all the documents present an e-mail address, office phone and fax number (accompanied in two documents by the secretaries' phone numbers), and postal address. S. Baumgärtner and A. Neumann provide their room numbers. Furthermore, Neumann offers source code for computer programs that he developed in a research project (cf. [31]).

Fig. 2 specifies the framework which resulted from the analysis of the sample. Especially regarding the automatic extraction of Web genre modules, it should be noted, that a special *form* feature is observable in each of the four documents: the spatial seperation of individual genre module instances by means of horizontal rules (realized by either using the <HR> tag or inline images acting as visual replacements). If applicable, the list presented in fig. 2 has been generalized with regard to the top-down sequence as well as to the embedded nature of some genre modules (with exceptions, see *Postal address*, so the list can still be presented in its entirety). Names of genre modules are noted in *italics*, their status in SMALL CAPS. Some of these values have not simply been determined based on the sample, but rather on our intuition of which genre modules this Web genre type usually contains. Of course, several additional genre modules of optional status can exist in concrete instances, e. g., a search box, or sections about someone's family, friends, hobbies etc. Furthermore, generic modules which apply to *every* Web genre document instance (and that can be inherited into the representation of this very hierarchy, the XML Schema definition, see sect. 3.5) are not presented in fig. 2, e. g., *metadata* (the URLs of all involved documents, the HTTP header etc.), and *unclassifiable* (reserved for sections that cannot be extracted automatically).

- *Affiliation* (COMPULSORY; logo graphics of resp. university, department/institute; maybe accompanied/substituted by plain text carrying the same information; can alternatively appear at the bottom of a document)
- *Alternative version* of document in a different lanugage, here: English; maybe accompanied by flag (OPTIONAL)
- Homepage owner's *name* (COMPULSORY), maybe accompanied by title (e. g., "Dr.") and phrase (e. g., "Homepage of *Firstname Lastname*")
- *Portrait photo* of author, spatially close to his/her name (thus building a module of its own), usually in upper third of document (OPTIONAL)
- *Contact information* (COMPULSORY)
  - *Postal address* (OPTIONAL; author's name, institute, university, street, postbox, zip code, city, country)
  - *Phone number* (OPTIONAL)
  - *Secretary's phone number* (OPTIONAL)
  - *Fax number* (OPTIONAL)
  - *Electronic mail address* (COMPULSORY)
  - *Room number* (OPTIONAL)
  - *Office hour* (OPTIONAL)
- *C. V.* or general biographical information (COMPULSORY)
- *Information about educational courses* (COMPULSORY)
- *Research interests* and/or *Projects* (COMPULSORY)
- *List of publications* (COMPULSORY)
- *List of talks or presentations* (OPTIONAL)
- *Related links* (OPTIONAL)
  - *Link to university's homepage* (COMPULSORY)
  - *Link to own department* (COMPULSORY)
  - *Link to own institute/research group* (COMPULSORY)
- *Last update information* (OPTIONAL)

Figure 2. Web genre type "Academic's Personal Homepage" and genre modules' status

Adapting de Saint-Georges's [9] definition of the *personal homepage*, we can, based on the analysis, define the Web genre *Academic's Personal Homepage: presentation of the self in digital, hypertextual form, authored by one individual working at a university or similar institution, and which (i) emphasizes this person (by a name and possibly a picture) and clearly states his/her affiliation to the university (including hyperlinks); and (ii) a person's current (and possibly past) research activities; and (iii) professional experience (by means of listing past and current educational courses, a C. V., and a list of publications); and (iv) displays a person's research interests (in the body of the text and/or through a list of hyperlinks to other sites); and (v) presents contact information (e. g., phone- and fax-numbers, postal- and e-mail-addresses). Purposes of this Web genre include: (i) establishing an individual scientific profile; (ii) distributing course-related material; (iii) making publications, current research activities and contact information easily available.*

[6] In all four pages, these were available on seperate documents. Publications can be downloaded in abstract, draft or final form, thus enabling instant access, as well as prompting the readers to give feedback.

[7] Making course notes and additional material available online can serve auxiliary purposes, e. g., advertising (by presenting the own research) or to offer the material to the outside world, so that others can both peer-review a lecture script or incorporate it into course notes.

When we compare this definition with others, two issues become clear: first, the notion of “personal homepage”, or “homepage”, has been used much too vaguely in the related literature up to now. With this term, some refer to specifically defined individual’s pages [9], others refer to a vast group of digital genres. Roussinov et al. [8] define this concept as “personal or organizational information plus links to other pages reflecting the subject’s interests that are intended to introduce the person or organization to the world and to facilitate further contact.” Thus, we need more precise definitions of the different Web genres reported as being “homepages”. As this is not in the scope of this paper, we hereby propose to use this term only in connection with individuals, not with organizations (*entry point* could be used as a more appropriate alternative). Second, the *Academic’s Personal Homepage* has established itself as a unique and novel Web genre that is one specific genre of the general hierarchy of *personal homepages*.

3.4 Extracting Content from Web Genre Instances

The overall goal is not only the *identification* of Web genres, but also the *extraction* of the content contained in genre modules and its *integration* into a structured XML document (Extensible Markup Language, [4]), which itself is based on the abstract form defined by the involved Web genre type’s XML Schema definition. The XML standard defines a formal meta-language. Its purpose is to enable the definition of specialized markup languages.

Fig. 3 shows an example representation by making the information implicitly contained in fig. 5, explicit. Several issues are of interest: first, this manually generated example only contains the information of the original HTML document that we think can be extracted automatically. That is why, e. g., three list items have been put into the `<unclassifiable>` Web genre module, as we think these three items will pose serious problems for Web genre-driven methods of information extraction. Second, the XML-representation has been simplified to keep it concise: A large volume of detailed information was cut out (“`...`”, lines 6, 33, etc.), e. g., `ID`-attributes which uniquely specifiy informational objects, the content of the `<phone>` and `<fax>` tags were further divided into `<areacode>` and `<phonenumber>`, and the `<metadata>` section contained more information. Third, lines 31–43 (“Teaching”) illustrate a central point of our approach: the breaking of physical document boundaries. These lines contain detailed information regarding Neumann’s educational courses. On his homepage, only a *hyperlink* exists, leading to the document containing the lists of courses and additional material (cf. sect. 3.3). An automated system requires knowledge of content, form, and structure of a Web genre type like the *Academic’s Personal Homepage*, whereupon structure implies knowledge about the typical *hypertextual* structure, too. Consequently, optional and compulsory Web genre modules need not exist as data in the homepage’s physical file. It would be legitimate for them to occur only virtually as hyperlinks to pages that need to fulfil a set of constraints (e. g., they have to physically exist in the document-space editable by the author: `.../~neumann/` → `.../~neumann/Papers/`). Thus, we need to further specify which compulsory genre modules *must* occur on the personal homepage itself. In the case of the Web genre type *Academic’s Personal Homepage*, this is only the name of the individual. The other compulsory genre modules (contact information, list of publications, information about courses, etc.) can be made available by means of independent documents, linked to from the homepage.

```
1  <?xml version="1.0" encoding="UTF-8"?>
2  <homepage type="Academic’s Personal Homepage" sys="aph"
3            xsi:noNamespaceSchemaLocation="aph.xsd">
4    <metadata>
5      <documentaddress>
6         <url lang="en" type="alterna">http://.../~neumann/index-e.html</url>
7         <url lang="de" type="default">http://.../~neumann/</url>
8      </documentaddress>
9      <date type="incorporation_into_corpus">02-06-2001</date>
10     <date type="last_modified">05-12-2000</date>
11     <name title="Dr." photo="http://www.informatik.../andreas-large.jpg">
12       <firstname>Andreas</firstname><lastname>Neumann</lastname>
13     </name>
14     <affiliation>
15       <university url="http://www.uni-trier.de">Trier</university>
16       <dept url="http://www.informatik.uni-trier.de">Computer Science</dept>
17       <research-unit>Programming Languages and Compilers</research-unit>
18     </affiliation>
19   </metadata>
20   <contact>
21     <address type="postal_address">
22        <name title="Dr.">
23           <firstname>Andreas</firstname><lastname>Neumann</lastname>
24        </name>
25        <dept>Computer Science</dept><university>University of Trier</university>
26        <country>Germany</country><zipcode>54286</zipcode><city>Trier</city>
27     </address>
28     <room>Building V, Room 216a)</room><phone>++49 651 201 2823</phone>
29     <fax>++49 651 201 3822</fax><email>neumann@PSI.Uni-Trier.DE</email>
30   </contact>
31   <teaching url="http://www.informatik.uni-trier.de/~neumann/Teach/">
32     <term term="summer" year="2000">
33        <course url="http://www...Lehre/dokver00.html" type="Vorlesung">
34           <title>Grundlagen der Dokumentenverarbeitung</title><day>FRI</day>
35           <time type="start">10</time><time type="end">12</time><room>V301</room>
36        </course>
37     </term>
38     <term term="summer" year="1999">
39        <course url="http://www.../PSI/fs_ss99.html" type="Uebung">
40          <title>Formale Semantik</title><day>WED</day>[...]
41        </course>
42     </term>
43   </teaching>
44   <listofpublications url="http://.../~neumann/Papers/index.html">
45     <entry type="Ph.D." university="Universitaet Trier" date="Dec. 1999">
46       <title>Parsing and Querying XML Documents in SML</title>
47       <abstract>[...]</abstract>
48       <online type="Postscript" url="http://.../~neumann/Papers/thesis.ps.gz"/>
49       <online type="PDF" url="http://.../~neumann/Papers/thesis.pdf.gz"/>
50     </entry>[...]
51   </listofpublications>
52   <unclassifiable>
53     <list>
54       <item url=".../~neumann/Fxp/">fxp - the functional XML parser</item>
55       <item url=".../~neumann/Fxgrep/">fxgrep - the functional ...</item>
56       <item url=".../~seidl">My boss</item>
57     </list>
58   </unclassifiable>
59 </homepage>
```

Figure 3. XML-representation of figure 5

3.5 XML Schema for Web Genre Modelling

Detailed Web genre analyses enable us to formulate XML Schema [13] definitions for Web genre types, which make the generalized information about compulsory and optional genre modules, their structuring, datatypes etc., explicit.[8] The purpose of such a bank of XML Schema definitions (see fig. 1) is to provide a document-grammar-like formal definition and framework for the overall structuring of generalized Web genre types and participating modules. Furthermore, we can improve these sets of XML Schema definitions in several aspects. First, we can define general *element types* at a very high level in our hierarchy of Schema definitions, so that we can reuse certain module or datatype definitions in more specialized definitions by means of inheritance. Second, we can enrich the definitions of the basic Web genre types, as well as the compulsory and optional modules that can occur in certain Web genres' instances, with information about <content, form, function>: a "search-box" module can be defined as being optional for the Web genre *Academic's Personal Homepage*. The default assignment regarding the values of <content, form, function> of this Web genre should roughly resemble the Web genre definition presented in sect. 3.3. If an optional genre module of the type "search box" existed on A. Neumann's homepage, we could compute the level of interactivity: we take the default assignment of the generalized Web genre type, extract the <content, form, function> assignments of all participating optional genre modules and allocate the specific assignment for this document. As a "search box" exists on this page, and as the genre module type "search box" is (among others) defined as, e. g., `function.interactivity="high"`, this would improve the assignment for the whole document as far as this feature is concerned.

3.6 Recognition Features – Web Genre Hierarchy

Automatic genre identification is possible, as shown in [20] and [21]. Extracting significant features and basing genre computation on the concrete values of these features should be feasible in the Web domain, too. The following section introduces several sets of features as the first prerequisite for our AWGI-system. The final section illustrates the second requirement: a comprehensive Web genre hierarchy.

Documents available on the Web are, due to the characteristics of HTML, a lot 'richer' than the plain texts of the Brown Corpus [20, 21]. This is why large parts of AWGI can be carried out by not examining the actual text at all. Therefore, we propose an approach that is predominantly founded on the feature-based analysis of the HTML structure of a document (or group of documents). Due to space restrictions, we only list the particular feature-categories and most of the features resp. short descriptions:

- *Metadata* – URL of a document, HTTP header, size, title, content of `<meta>` tags, reference to the utilized HTML-DTD
- *HTML Structure* – Overall structure of the HTML element-tree
  - *Hyperlinks in Document/Document Group* – Link number, internal vs. external, hypertext structure, target's file-type, target's Web genre, anchor, link function, link position
  - *Inline Graphics* – Dimensions of an inline image, file- and directory-names, content, alternative text, format, overall number of graphics
  - *Interactive Elements* – HTML forms, JavaScript, Plug-Ins, Applets
- *Document-Spanning Features* – Recurrence of genre modules, position of document in terms of the hypertext structure of document group
- *Linguistic and Structural Cues* – Certain linguistic expressions at Web genre-specific positions
- *Language Issues* – Spoken [30] vs. written nature of language, spelling

The methods of application of these features within a concrete implementation of an AWGI-system depend heavily upon the results of the future empirical analyses. However, it is safe to assume that our two main goals—automatic Web genre identification and information extraction—will, in turn, require two main software components. Extrapolating the promising classification results reported in [20] and [21], we will most probably concentrate our efforts on the information extraction task. Up to now, there has been a lot of research upon which we can build within the field of Information Extraction with regard to the Web; most importantly, on different approaches to the manual or automatic construction of Wrappers [22, 16, 34, 2, 28, 23]. Wrappers are highly specialised software modules that are able to parse HTML documents belonging to a tightly defined thematic domain (e. g., car- or real-estate-advertisements) in order to extract their information. The Web genre notion could be optimally used to generalize these up to now extremely restricted wrapper-based approaches.

As we have shown in sect. 3.1, a hierarchy of Web genres should at least reflect the CONTAINS and the IS-A aspects of Web genre types and modules. Fig. 4 shows a small excerpt of our comprehensive hierarchy into which these are incorporated (solid lines indicate compulsory, dashed lines indicate optional Web genre modules, dashed arrows identify mandatory hyperlinks, arcs mark the IS-A relationship). Hierarchies of this kind are the basic foundation for our definitions of Web genre types and modules. In the upper part, fig. 4 illustrates the structure of a typical institute's resp. research unit's entry point. A crucial point is that the staff listing contains (among others) lists of personal homepages that may belong to academics, administrative staff, or student assistants. At this position in the hierarchy, two characeristics coincide: (i), the *Staff Listing* Web genre may optionally contain a *list of personal homepages* Web genre module which itself may contain one or more lists of hyperlinks[9] to instances of the three aforementioned Web genre types. (ii), there exists an abstract IS-A relationship between the Web genre types *Student Assis-*

[8] Due to space restrictions, we are not able to present the extensive XML Schema definition for the *Academic's Personal Homepage* in this paper.

[9] These are *mandatory* hyperlinks (dashed arrows), i. e., the content referenced by the links *must not* be contained in the source document.

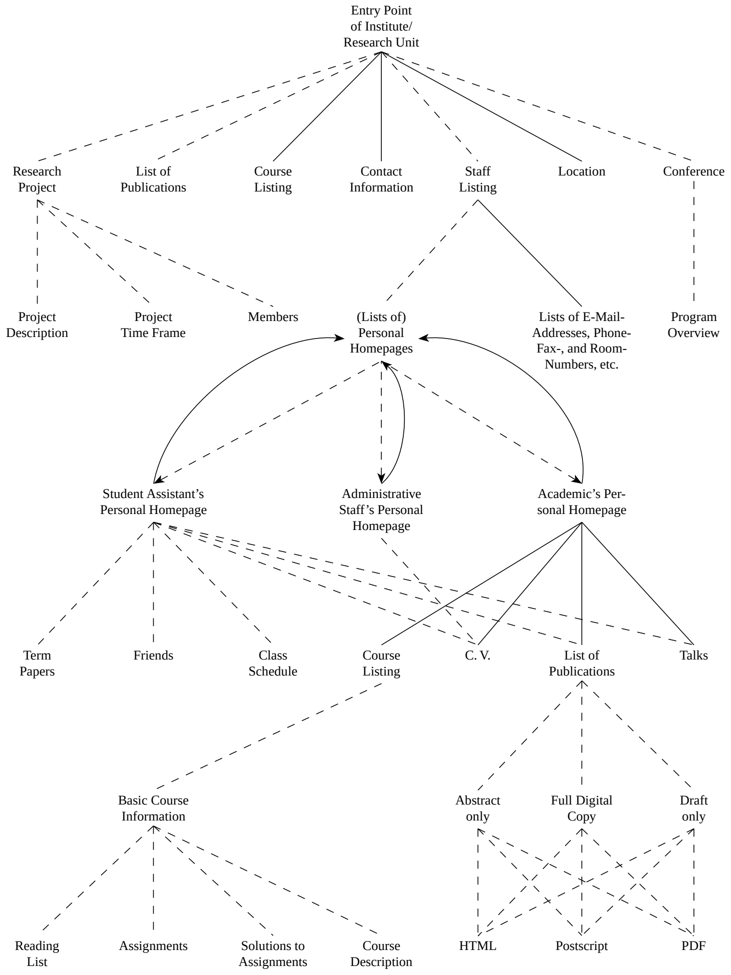


Figure 4. Excerpt of our Web genre hierarchy.

tant's Personal Homepage, *Administrative Staff's Personal Homepage*, *Academic's Personal Homepage* and the upper level, more generic Web genre type *Personal Homepage*. Furthermore, fig. 4 illustrates our approach towards inheriting general Web genre modules: using XML Schema definitions, we can separately define genre modules such as *List of Publications* or *C. V.* and can reference the definition of these specific modules in the formal definition of a Web genre type such as, e. g., *Student Assistant's Personal Homepage*, which contains the above-mentioned modules as optional constituents, as well as other optional modules that only apply to this Web genre type (among others, *Term Papers*, *List of Friends*, *Class Schedule*). Additionally, the aforementioned optional Web genre modules are good examples in order to illustrate our approach of modifying the default <content, form, function> assignment of the Web genre type *Student Assistant's Personal Homepage*: we can imagine two instances, one containing a C. V., a list of term papers and a class schedule, the other primarily containing a list of hyperlinks to friends, and a photo gallery. By means of modifying the default assignment of the genre type by the concrete instantiations of Web genre modules, we can deduce that the former page is rather formal/official, whereas the latter is of a more informal/private nature.

The aforementioned analysis of a random sample containing 200 documents resulted in an initial version of an academic Web genre hierarchy. The following excerpt only shows the major thematic Web genre groups, along with the most frequently found Web genres. 84 Web genres and groups were found altogether, 11 documents could not be allocated for technical reasons, 4 Web genres could not be reasonably incorporated into the Web genre hierarchy (Association Chronicle, Clause of Legislative Regulation, Sports Ranking List, Travelogue). The number of documents in the Web genre group resp. Web genre are given in brackets. As we excluded universities' main entry points and upper level documents in the sample generation, these are inherently not included. Moreover, due to spatial restrictions, we cannot provide the full graph with all interconnections, but emulate these by indenting the respective Web genre group based on the number of Web genres found and their allocations into upper groups.

- *Administrative Information* (14)
  - Study Regulations (2)
  - Course-related Information (2)
    - Basic Course Information (9)
    - Course Description (7)
    - Assignments (5)
    - Course Listing (2)
    - Course-related Material (2)
    - Solutions to Assignments (1)
  - Information about a Scholarship (1)
  - Vacancy Advertisement (1)
- *Information about the University* (5)
  - Location/Directions/Floor Plan (3)
  - Description of an Informational Service (1)
- *Institute/Research Unit* (15)
  - Entry Point (4)
    - *Conference* (5)
      - Description of a Working Group (2)
      - Program Overview (1)
      - Registration Form (1)
  - Staff Listing (4)
    - *Personal Homepage* (14)
      - Academic's Personal Homepage (4)
      - Student Assistant's Personal Homepage (3)
      - Virtual Business Card of Staff Member (1)
  - Description of Main Research Focus (3)
    - *Bibliography/List of Publications* (9)
      - Author Bibliography (3)
      - List of Institution Publications (3)
      - Thematically categorised Bibliography (2)
    - *Research Project* (9)
      - Project Description (7)
      - Project Time Frame (1)
  - Short Description of Institute/Research Unit (1)
  - Organizational Plan (1)

4 Future Research

The next step towards developing a Web genre identification system will comprise an extensive empirical analysis of about 2 000 documents. Thereafter, we will conceptualize a set of Web genre type generalizations by means of XML Schema definitions. The last phase of our project will be the implementation of a robust multi-agent AWGI system for the academic domain. As the methods to be utilized depend significantly on the results of the empirical studies, we cannot provide any specific information yet. We estimate these methods to be a combination of both symbolic—utilizing the features listed in sect. 3.6—and machine learning-methods, using the 2 000+ manually categorized documents as training data for the classification algorithm. At a later stage, we will equip the system with capabilities for extracting information from Web genre modules in order to instantiate XML documents that will conform to the above-mentioned XML Schema definitions and will contain structured information that was only implicitly included in the original, unstructured Web pages.

References

[1] E. Amitay. Anchors in Context: A Corpus Analysis of Web Pages Authoring Conventions. In L. Pemberton and S. Shurville, editors, *Words on the Web*, pages 25–35. Intellect Books, Bristol, 2000. `http://www.mri.mq.edu.au/~einat/`. The printed version of this article has been shortened.

[2] N. Ashish and C. A. Knoblock. Semi-Automatic Wrapper Generation for Internet Information Sources. In *Conference on Cooperative Information Systems*, pages 160–169, 1997.

[3] C. Bazerman. Systems of Genres and the Enactment of Social Intentions. In A. Freedman and P. Medway, editors, *Genre and the New Rhetoric*, pages 79–101. Taylor and Francis, London, 1995.

[4] T. Bray, J. P. Paoli, C. M. Sperberg-McQueen, and E. Maler. Extensible Markup Language (XML) 1.0 (Second Edition). Technical report, World Wide Web Consortium, 2000. `http://www.w3.org/TR/2000/REC-xml-20001006`.

[5] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery. Learning to Extract Symbolic Knowledge from the World Wide Web. In *Proceedings of AAAI-98*. American Association for Artificial Intelligence, 1998.

[6] K. Crowston and M. Williams. Reproduced and Emergent Genres of Communication on the World-Wide Web. In *Proceedings of the 30th Hawaii International Conference on Systems Sciences (HICSS-30)*. IEEE, 1997.

[7] K. Crowston and M. Williams. The Effects of Linking on Genres of Web Documents. In *Proceedings of HICSS-32*. IEEE, 1999.

[8] K. Crowston and M. Williams. Reproduced and Emergent Genres of Communication on the World Wide Web. *The Information Society*, 16(3):201–215, 2000.

[9] I. de Saint-Georges. Click Here if You Want to Know Who I Am. Deixis in Personal Homepages. In *Proceedings of HICSS-31*. IEEE, 1998.

[10] A. Dillon and B. A. Gushrowski. Genres and the Web: Is the Personal Home Page the First Uniquely Digital Genre? *Journal of the American Society for Information Science*, 51(2):202–205, 2000.

[11] T. Erickson. The World Wide Web as Social Hypertext. *Communications of the ACM*, 39(1):15–17, 1996.

[12] L. B. Eriksen and C. Ihlström. Evolution of the Web News Genre – The Slow Move Beyond the Print Metaphor. In *Proceedings of HICSS-33*. IEEE, 2000.

[13] D. C. Fallside, H. S. Thompson, D. Beech, M. Maloney, N. Mendelsohn, P. V. Biron, and A. Malhotra. XML Schema. Technical report, World Wide Web Consortium, 2001. W3C Recommendation. `http://www.w3.org/XML/Schema`.

[14] I. Fortanet, J. C. Palmer, and S. Posteguillo. Netvertising: Content-Based Subgeneric Variations in a Digital Genre. In *Proceedings of HICSS-31*. IEEE, 1998.

[15] R. Furuta and C. C. Marshall. Genre as Reflection of Technology in the World-Wide Web. Technical report, Hypermedia Research Lab, Department of Computer Science, Texas A&M University, 1996.

[16] X. Gao and L. Sterling. AutoWrapper: Automatic Wrapper Generation for Multiple Online Services. In *Proceedings of Asia Pacific Web Conference 1999 (APWeb99)*, pages 61–70, September 1999.

[17] S. W. Haas and E. S. Grams. A Link Taxonomy for Web Pages. In C. Preston, editor, *Proceedings of the 61st Annual Meeting of the American Society for Information Science*, pages 485–495, 1998.

[18] S. W. Haas and E. S. Grams. Page and Link Classifications: Connecting Diverse Resources. In I. Witten, R. Akscyn, and F. Shipman, editors, *Proceedings of Digital Libraries '98 – Third ACM Conference on Digital Libraries*, pages 99–107, Pittsburgh, 1998.

[19] S. W. Haas and E. S. Grams. 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):181–192, 2000.

[20] J. Karlgren and D. Cutting. Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. In *COLING 94 – The 15th International Conference on Computational Linguistics*, volume 2, pages 1071–1075, Kyoto, Japan, August 1994. Association for Computational Linguistics.

[21] B. Kessler, G. Nunberg, and H. Schütze. Automatic Detection of Text Genre. In *Proceedings of the 35th Annual Meeting of the ACL and the 8th Meeting of the European Chapter of the ACL*, pages 32–38, San Francisco, 1997. Morgan Kaufmann.

[22] L. Liu, C. Pu, and W. Han. XWRAP: An XML-Enabled Wrapper Construction System for Web Information Sources. In *International Conference on Data Engineering (ICDE)*, pages 611–621, 2000.

[23] W. May and G. Lausen. Information Extraction from the Web. Technical Report 136, Computer Science Institute, Freiburg University, March 2000. `http://www.informatik.uni-freiburg.de/~may/Publics/`.

[24] C. R. Miller. Genre as Social Action. *Quarterly Journal of Speech*, (70):151–167, 1984. Reprinted in: *Genre and the New Rhetoric* (1995), edited by Freedman, A. and Medway, P. London: Taylor and Francis, pp. 23–42.

[25] M. Müller. Inducing Conceptual User Models. In S. Wrobel, editor, *Proceedings of ABIS-99 (Adaptivität und Benutzermodellierung in interaktiven Softwaresystemen)*, Magdeburg, September 1999. Gesellschaft für Informatik. `http://www-mmt.inf.tu-dresden.de/joerding/abis99/proceedings.html`.

[26] Y. K. Muthusamy and L. Spitz. Automatic Language Identification. In R. Cole, J. Mariani, H. Uszkoreit, G. B. Varile, A. Zaenen, and A. Zampolli, editors, *Survey of the State of the Art in Human Language Technology*, pages 314–317. Cambridge University Press, Cambridge, 1998. `http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html`.

[27] W. J. Orlikowski and J. Yates. Genre Repertoire: The Structuring of Communicative Practices in Organizations. *Administrative Science Quarterly*, (39):541–574, 1994.

[28] F. Paradis. Information Extraction and Gathering for Search Engines: The Taylor Approach. RIAO (Recherche d'Informations Assiste par Ordinateur), Paris, France, 2000.

[29] G. Rehm. `korpus.html` – Zur Sammlung, Datenbank-basierten Erfassung, Annotation und Auswertung von HTML-Dokumenten. In H. Lobin, editor, *Proceedings of the GLDV Spring Meeting 2001*, pages 93–103, Giessen, Germany, 2001. Society for Computational Linguistics and Language Technology. `http://www.uni-giessen.de/fb09/ascl/gldv2001/`.

[30] G. Rehm. Schriftliche Mündlichkeit in der Sprache des World Wide Web. In A. Ziegler and C. Dürscheid, editors, *Kommunikationsform E-Mail*. Stauffenburg, 2002. In press.

[31] G. Rehm and H. Lobin. From Open Source to Open Information – Collaborative Methods in Creating XML-based Markup Languages. In *Proceedings of Electronic Publishing 2000*, Kaliningrad, Svetlogorsk, 2000. International Federation for Information Processing and International Council for Computer Communication. `http://www.uni-giessen.de/~g91063/pdf/open-information.pdf`.

[32] G. F. Roberts. The Home Page as Genre: A Narrative Approach. In *Proceedings of HICSS-31*. IEEE, 1998.

[33] D. Roussinov, K. Crowston, M. Nilan, B. Kwasnik, J. Cai, and X. Liu. Genre Based Navigation on the Web. In *Proceedings of HICSS-34*. IEEE, 2001.

[34] A. Sahuguet and F. Azavant. Building Intelligent Web Applications Using Lightweight Wrappers. *Data and Knowledge Engineering*, 36(3):283–316, 2001.

[35] M. Shepherd and C. Watters. The Evolution of Cybergenres. In *Proceedings of HICSS-31*. IEEE, 1998.

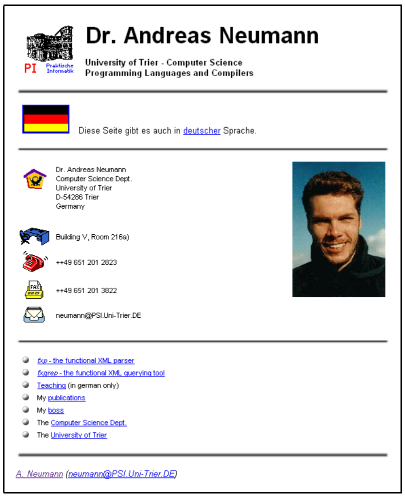
[36] M. Shepherd and C. Watters. The Functionality Attribute of Cybergenres. In *Proceedings of HICSS-32*. IEEE, 1999.

[37] J. M. Swales. *Genre Analysis – English in academic and research settings*. Cambridge University Press, Cambridge, 1990.

[38] E. G. Toms and D. G. Campbell. Genre as Interface Metaphor: Exploiting Form and Function in Digital Environments. In *Proceedings of HICSS-32*. IEEE, 1999.

[39] A. Walters. An Analysis of Purposes and Forms of Personal Homepages on the World Wide Web. Thesis, Sloan School of Management, Massachusetts Institute of Technology, 1996.

A Screenshots of the Sample-Documents



Dr. Andreas Neumann

University of Trier - Computer Science
Programming Languages and Compilers

Diese Seite gibt es auch in deutscher Sprache.

Dr. Andreas Neumann
Computer Science Dept.
University of Trier
D-54286 Trier
Germany

Building V, Room 216a)

++49 651 201 2823

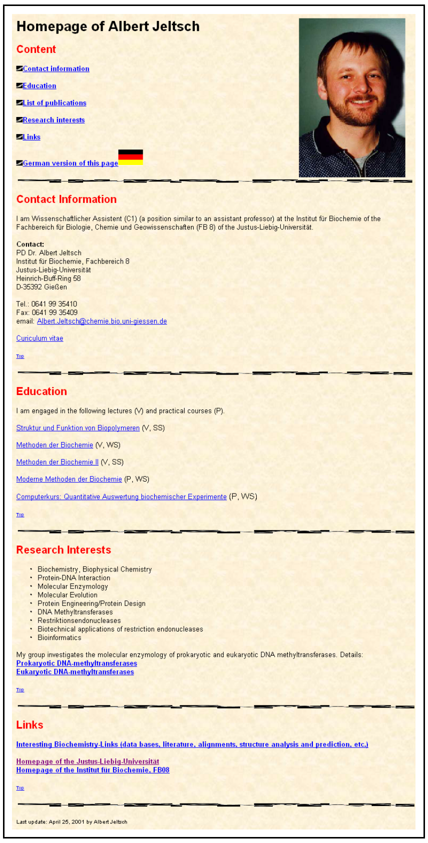
++49 651 201 3822

neumann@PSI.Uni-Trier.DE

- *fxp* - the functional XML parser
- *fxgrep* - the functional XML querying tool
- Teaching (in german only)
- My publications
- My boss
- The Computer Science Dept.
- The University of Trier

A. Neumann (neumann@PSI.Uni-Trier.DE)

Figure 5. Homepage of A. Neumann, CS Dept., Trier Univ. (`http://www.informatik.uni-trier.de/˜neumann/`)



Homepage of Albert Jeltsch

Content

Contact information

Education

List of publications

Research interests

Links

German version of this page

Contact Information

I am Wissenschaftlicher Assistent (C1) (a position similar to an assistant professor) at the Institut für Biochemie of the Fachbereich für Biologie, Chemie und Geowissenschaften (FB 8) of the Justus-Liebig-Universität.

Contact:
PD Dr. Albert Jeltsch
Institut für Biochemie, Fachbereich 8
Justus-Liebig-Universität
Heinrich-Buff-Ring 58
D-35392 Gießen

Tel.: 0641 99 35410
Fax: 0641 99 35409
email: Albert.Jeltsch@chemie.bio.uni-giessen.de

Curiculum vitae

Top

Education

I am engaged in the following lectures (V) and practical courses (P).

Struktur und Funktion von Biopolymeren (V, SS)

Methoden der Biochemie (V, WS)

Methoden der Biochemie II (V, SS)

Moderne Methoden der Biochemie (P, WS)

Computerkurs: Quantitative Auswertung biochemischer Experimente (P, WS)

Top

Research Interests

- Biochemistry, Biophysical Chemistry
- Protein-DNA Interaction
- Molecular Enzymology
- Molecular Evolution
- Protein Engineering/Protein Design
- DNA Methyltransferases
- Restriktionsendonucleases
- Biotechnical applications of restriction endonucleases
- Bioinformatics

My group investigates the molecular enzymology of prokaryotic and eukaryotic DNA methyltransferases. Details:
Prokaryotic DNA-methyltransferases
Eukaryotic DNA-methyltransferases

Top

Links

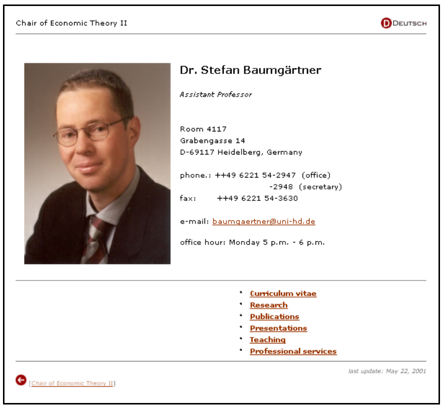
Interesting Biochemistry-Links (data bases, literature, alignments, structure analysis and prediction, etc.)

Homepage of the Justus-Liebig-Universität
Homepage of the Institut für Biochemie, FB08

Top

Last update: April 25, 2001 by Albert Jeltsch

Figure 6. Homepage of A. Jeltsch, Institute of Biochemistry, Giessen Univ. (`http://www.uni-giessen.de/˜gf1020/`)



Chair of Economic Theory II

Dr. Stefan Baumgärtner

Assistant Professor

Room 4117
Grabengasse 14
D-69117 Heidelberg, Germany

phone.: ++49 6221 54-2947 (office)
-2948 (secretary)
fax: ++49 6221 54-3630

e-mail: baumgaertner@uni-hd.de

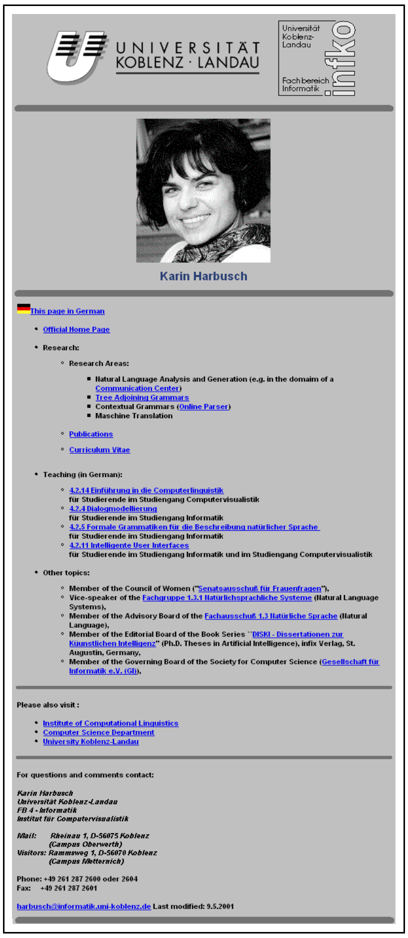
office hour: Monday 5 p.m. - 6 p.m.

- **Curriculum vitae**
- **Research**
- **Publications**
- **Presentations**
- **Teaching**
- **Professional services**

last update: May 22, 2001

[Chair of Economic Theory II]

Figure 7. Homepage of S. Baumgärtner, Economic Dept., Heidelberg Univ. (`http://www.rzuser.uni-heidelberg.de/˜mw3/baumgaertner/homed.html`)



Karin Harbusch

This page in German

- **Official Home Page**
- **Research:**
    - **Research Areas:**
        - **Natural Language Analysis and Generation (e.g. in the domaim of a Communication Center)**
        - **Tree Adjoining Grammars**
        - **Contextual Grammars (Online Parser)**
        - **Maschine Translation**
    - **Publications**
    - **Curriculum Vitae**
- **Teaching (in German):**
    - **4.2.14 Einführung in die Computerlinguistik**
      für Studierende im Studiengang Computervisualistik
    - **4.2.4 Dialogmodellierung**
      für Studierende im Studiengang Informatik
    - **4.2.5 Formale Grammatiken für die Beschreibung natürlicher Sprache**
      für Studierende im Studiengang Informatik
    - **4.2.11 Intelligente User Interfaces**
      für Studierende im Studiengang Informatik und im Studiengang Computervisualistik
- **Other topics:**
    - **Member of the Council of Women ("Senatsausschuß für Frauenfragen"),**
    - **Vice-speaker of the Fachgruppe 1.3.1 Natürlichsprachliche Systeme (Natural Language Systems),**
    - **Member of the Advisory Board of the Fachausschuß 1.3 Natürliche Sprache (Natural Language),**
    - **Member of the Editorial Board of the Book Series ``DISKI - Dissertationen zur Küunstlichen Intelligenz" (Ph.D. Theses in Artificial Intelligence), infix Verlag, St. Augustin, Germany,**
    - **Member of the Governing Board of the Society for Computer Science (Gesellschaft für Informatik e.V. (GI)),**

Please also visit :

- **Institute of Computational Linguistics**
- **Computer Science Department**
- **University Koblenz-Landau**

For questions and comments contact:

Karin Harbusch
Universität Koblenz-Landau
FB 4 - Informatik
Institut für Computervisualistik

Mail: Rheinau 1, D-56075 Koblenz
(Campus Oberwerth)
Visitors: Rammsweg 1, D-56070 Koblenz
(Campus Metternich)

Phone: +49 261 287 2600 oder 2604
Fax: +49 261 287 2601

harbusch@informatik.uni-koblenz.de Last modified: 9.5.2001

Figure 8. Homepage of K. Harbusch, Institute of Computational Visualistics, Koblenz-Landau Univ. (`http://www.uni-koblenz.de/˜harbusch/`)