WIKINGER
Wiki Next Generation Enhanced Repositories

Lars Bröcker, Stefan Paal
Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS)

Andreas Burtscheidt, Bernhard Frings
Kommission für Zeitgeschichte (KfZG)

Marc Rössler, Andreas Wagner, Wolfgang Hoeppner
Computerlinguistik, Universität Duisburg-Essen

Abstract

The regular indexing of text documents is based on the textual representation and does not evaluate the actual document content. In the semantic web approach, human-authored text documents are transformed into machine-readable content data which can be used to create semantic relations among documents. In this paper, we present ongoing work in the WIKINGER project which aims to build a web-based system for semantic indexing of text documents by evaluating manual and semi-automatic annotations. A particular feature is the continuous refinement of the automatically generated semantic network by considering community feedback. The feasibility of the approach will be validated in a pilot application.

1 Introduction

The Internet provides the technical infrastructure to access different data sources from around the world and to link documents from various sites into a global information space. For finding documents, web search engines like Google retrieve the document data and create a common document index which can be used to find documents matching a given query. The matching is performed by and large using string comparison. The related indexing procedure does not require any user input and can be instantly applied on any text document. However, regular search engines are not able to grasp the content of the documents but are limited to evaluate their textual representation. Thus, results from search engines lack topical context information that would allow the users to put them into perspective regarding their information needs. Another approach is the enrichment of human-authored text documents with machine-readable metadata, as proposed by the *Semantic Web* [1]. This metadata contain pieces of information regarding the semantics of the content it describes, thereby allowing machines to disambiguate different meanings of seemingly equal string representations. The crux at the moment lies in generating this metadata in a time- and cost-effective manner.

In this paper, we present the ongoing work towards the creation of the WIKINGER platform. It supports the semantic indexing of text documents based on manual and semi-automatic annotation of text fragments and the exploration of the resulting semantic network via a Wiki front end. A particular feature is the ability to continuously refine the semantic network by analysing new documents added later on. The system is implemented in Java and based on a service-oriented architecture which allows deploying the system in a heterogeneous and distributed environment like the Internet. New services can be added easily and distinct document sources may be connected via specific harvesters feeding the documents into the WIKINGER document repository. The feasibility of the approach is validated in a pilot application which particularly demonstrates the refinement of the semantic network using community feedback.

The paper is organized as follows. In section 2, we sketch the big picture of the WIKINGER project and introduce the process workflows and system components. This is followed by the detailed description of the named entity recognition in section 3, the semi-automatic generation of semantic networks in section 4 and the illustration of the first application in section 5. An overview of the results so far and the conclusions follow in section 6.

2 WIKINGER – The Big Picture

The overall goal of the project WIKINGER is the enrichment of textual data with semantic metadata. To this end, it supports the manual and semi-automatic semantic indexing of archived and community documents. Archived documents are read into a document repository and the related document structure is decomposed and transformed into an internal uniform document format for consistent processing across different document sources, e.g. by using a source-specific harvester. The subsequent workflows for semantic indexing are separated into manual annotation, community feedback and automatic annotation, as shown in fig. 1. Manual annotation is performed by using a client-side Java application called WALU (Wikinger Annotations- und Lern-Umgebung) which allows annotators to tag text portions according to a given set of semantic concepts (1), as further detailed in section 3. The annotations are fed back into the metadata repository and are used to generate a semantic network by connecting entities via semantic relations, as described in section 4. In the next step, the community can revise the annotations and relations by using a Wiki-frontend, e.g. XWiki (2). New documents may be added to the document repository as well and manually annotated and automatically linked into the semantic network as before. Since it is not feasible to tag every document by human annotators, we propose the use of a refined Named Entity Recognition (NER) service which automatically annotates new documents without particular user intervention (3), as described in section 3.

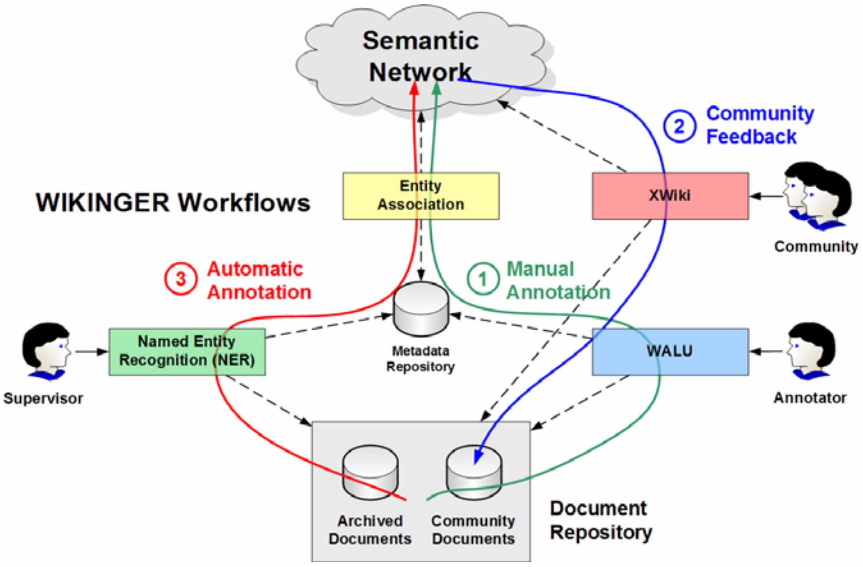


Figure 1: WIKINGER Workflows

The system design of WIKINGER is based on a service-oriented architecture. The decomposition of the implementation into distinct services and applications is shown in fig. 2.

The web service bus connects the system components and can be used to extend the original Java implementation with extra services and applications, e.g. additional annotation services [2]. The document service manages a versioned document repository in a relational database, e.g. MySQL, and stores documents harvested from different sources in a uniform document format. To this end, the object-relational mapping tool Apache OJB is employed which supports the transparent handling of Java objects stored in a relational database. In this context, the versioning feature allows creating persistent semantic networks which do not become invalid due to updates of document objects or links modified later on.

The named entity recognition (NER) service manages the annotations in the metadata repository which is also stored in a relational database. In addition to the automatic annotation processing, the client-side tool WALU can be used by annotators to add manual annotations. Supervisors may also monitor and adjust the annotation process. The entity service evaluates the annotations and creates a semantic network stored in the entity model repository. It uses Resource Description Framework (RDF) and HP Jena for modelling and storing the network in the database [3]. The servlet engine Jakarta Tomcat runs the web service framework Apache Axis and XWiki, which is used for web-based access. Finally, a regular web browser can be used by readers and editors to retrieve the documents, its annotations and the semantic relationships with other document entities.

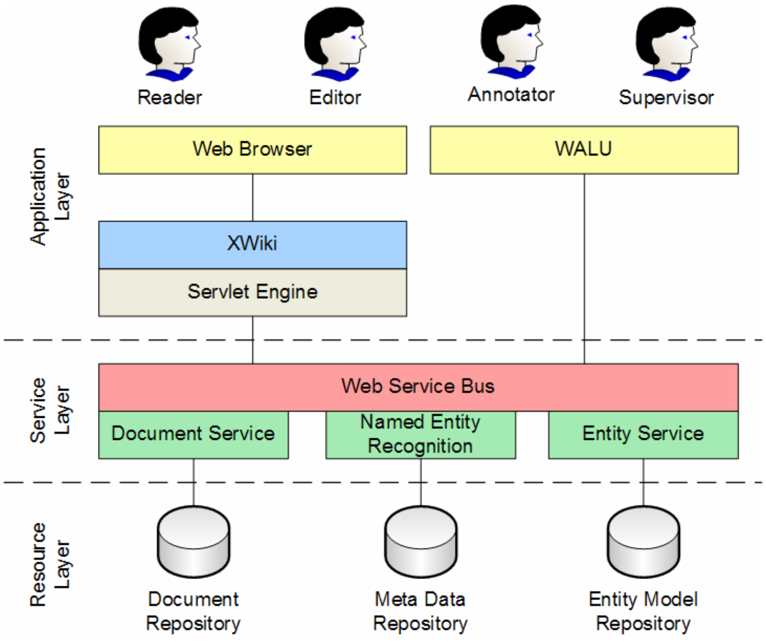


Figure 2: System Design

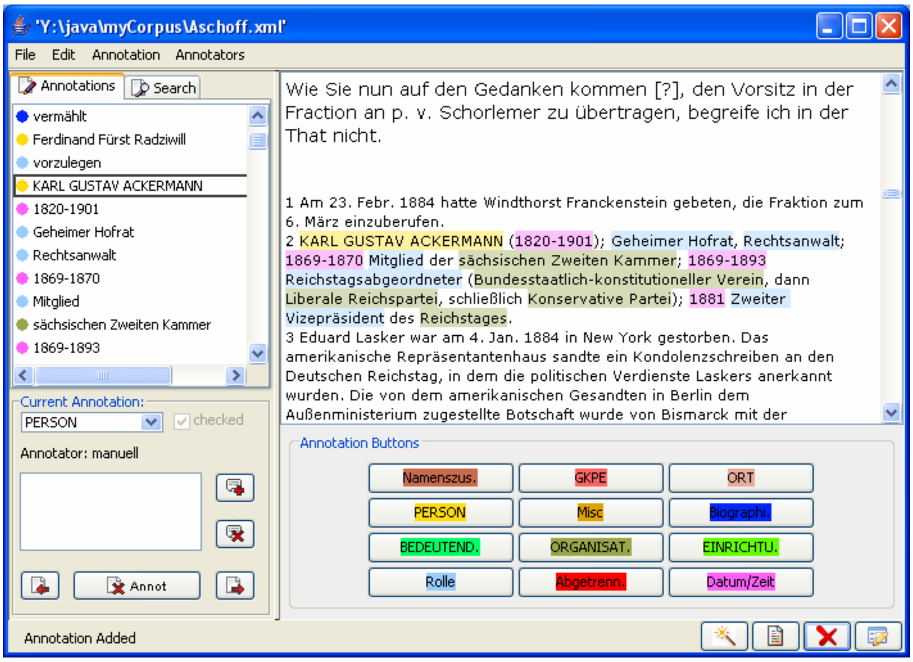
3 Named Entity Recognition

As described in Chapter 2, the nodes of the semantic network are provided by a Named Entity Recognition (NER) module that is applied to a large amount of textual data in order to extract the concepts, instances and topics relevant for the domain. NER is the task of identifying and classifying application-relevant names within written texts. The challenges of NER culminate in the problem of ambiguity and unknown words. Ambiguous word forms, i.e. "Washington" can be used to refer to an instance of the category **Person**, **Location**, or to the American government, which would belong to the category **Organisation**. Unknown words occur as it is impossible to enumerate all names in a dictionary. New companies for instance are founded each day and almost all of these get a new name. Both of these problems can only be tackled by taking into account certain characteristics of the words (or word sequences) to be identified, and their surrounding context.

NER requires the selection and definition of the relevant categories and a technique to recognise these items. For both aspects of the task, we opted for an example-based approach. This means that we try to avoid the need for the explicit *a priori* definition of domain-specific concepts and instead rely on an "empirical" determination of the categories based on manually labelled examples. In particular, we abstain from the linguistically motivated modelling of questions like "What is a name?" or "What linguistic structures usually indicate a name occurrence?" We rather induce a model for NER which is based on machine learning methods and a set of examples consisting of manually labelled text occurrences. With regard to the collaborative structure of WIKINGER, this example-based approach has an important advantage over rule-based NER approaches, which require the explicit modelling of knowledge: For domain experts it is straightforward to recognise and mark those entities occurring in a text which are relevant for the domain, and to provide the information which can be fed into machine learning approaches. On the contrary, it is a tedious task to communicate the domain expertise that a computational linguist needs to write sophisticated rules suitable for NER. Therefore, the example-based communication is a very efficient way to exchange expertise. In WIKINGER, this exchange is carried out via the annotation environment WALU (**W**ikinger **A**nnotations- und **L**ern-**U**mgebung [4]), a software which is being developed within the project.

WALU is a tool for finding relevant categories, specifying them in an example-based manner via manual annotations, applying machine learning models to automatically annotate texts, and inspecting and correcting the resulting annotations. Since all of these steps are to be executed primarily by the domain experts, it is essential that WALU provides all the required functionalities in an understandable and comfortable manner while encapsulating the complexity of the task. For the computational linguist, a fast and comfortable facility is crucial to browse through the annotated instances in order to identify the accuracy, specific problems, or pitfalls of the annotations. The current prototype of WALU shown in Figure 3 is optimised for manual annotations in order to detect and characterise the categories relevant for the domain. It supports this process with an easy to use GUI, a comfortable navigation through the annotations and simple but effective annotation support such as the automatic adjustment of markup-boundaries or a dynamic markup-dictionary. This dictionary is created during the annotation process and is used to instantaneously propose markup-labels for text passages corresponding to dictionary entries.

Using a context-sensitive menu, the annotator confirms or rejects these proposals and/or removes the entry from the dictionary. In our experience the immediate feedback of the dynamic markup-dictionary also helps the domain experts to clarify the task of string-based identification of domain-relevant concepts. Additionally, WALU also provides an automatic annotator for syntactically ‘easy’ strings referring to the category **Date** which is based on regular expressions. To be more concrete, this is a prototype of a series of automatic mechanisms that will be used to annotate all the available documents. Besides a few annotators based on regular expressions to classify entities with unique patterns (such as email addresses and URLs), most of these annotators are based on machine-learning algorithms that will be accessible via WALU.



- ● vermählt
- ● Ferdinand Fürst Radziwill
- ● vorzulegen
- ● KARL GUSTAV ACKERMANN
- ● 1820-1901
- ● Geheimer Hofrat
- ● Rechtsanwalt
- ● 1869-1870
- ● Mitglied
- ● sächsischen Zweiten Kammer
- ● 1869-1893

Current Annotation: PERSON ☑ checked

Annotator: manuell

Annotation Added

Wie Sie nun auf den Gedanken kommen [?], den Vorsitz in der Fraction an p. v. Schorlemer zu übertragen, begreife ich in der That nicht.

1 Am 23. Febr. 1884 hatte Windthorst Franckenstein gebeten, die Fraktion zum 6. März einzuberufen.
2 KARL GUSTAV ACKERMANN (1820-1901); Geheimer Hofrat, Rechtsanwalt; 1869-1870 Mitglied der sächsischen Zweiten Kammer; 1869-1893 Reichstagsabgeordneter (Bundesstaatlich-konstitutioneller Verein, dann Liberale Reichspartei, schließlich Konservative Partei); 1881 Zweiter Vizepräsident des Reichstages.
3 Eduard Lasker war am 4. Jan. 1884 in New York gestorben. Das amerikanische Repräsentantenhaus sandte ein Kondolenzschreiben an den Deutschen Reichstag, in dem die politischen Verdienste Laskers anerkannt wurden. Die von dem amerikanischen Gesandten in Berlin dem Außenministerium zugestellte Botschaft wurde von Bismarck mit der

Figure 3: WALU

In order to enable domain experts to set up new annotators, we focus on approaches to NER which consist of domain-independent features and of resources that can be easily adapted to a new domain or a new NER task (see [5]).

4 Semiautomatic Generation of Semantic Networks

The preceding chapter has shown how the NER module gathers the relevant concepts and their respective instances from the text corpus. At this stage in the workflow the nodes of the future semantic network representing the application domain are known. The last parts needed for a semantic network are the relations connecting the different nodes of the graph. This is the task of the next part in the workflow as depicted in fig 4: finding and extracting relations from the corpus in order to integrate the nodes into a semantic network. Relation Extraction is an active research topic within Computational Linguistics and Machine Learning communities, mainly driven by the quest for automatic translation tools or question answering systems. Accordingly, a lot of work has been performed and published on this topic. The approaches taken can be divided into two groups: rule-based and statistical relation extraction. Following the rule-based approach, rules are defined, that describe patterns for the target relations between different kinds of entities. This transforms the extraction task into one looking for occurrences in the data set fitting the patterns described by the rule set (see e.g. [7, 8]). The other group of approaches utilizes various statistical methods in order to train the computer to find relations in the corpus, thereby transforming it into a machine learning task (see e.g. [9, 10]).

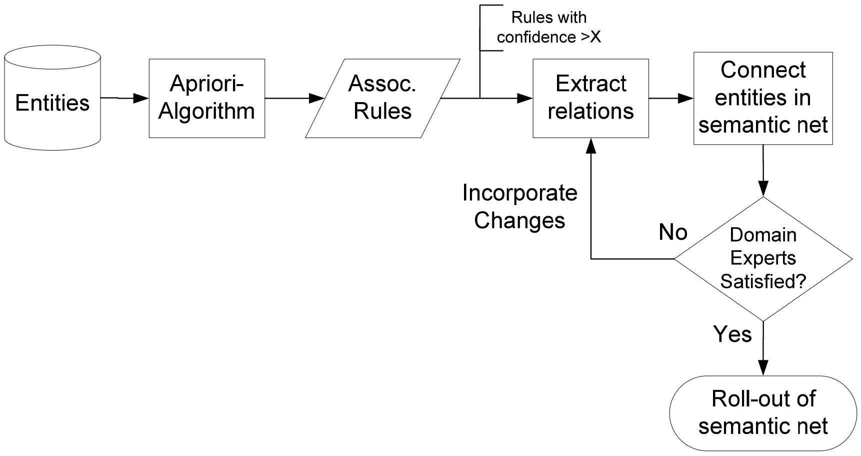


Fig 4: Workflow of the relation extraction module

Much work in this group uses previously agreed upon sets of relations, trading flexibility for easier training of the corresponding statistical models. This is only feasible with in-depth knowledge of both the target domain and the needs of the future users at design-time of the system. The relation extraction module of the WIKINGER platform falls into the second group of approaches as well. But since the platform is to be domain-independent, no fixed set of relations can be chosen. Figure 4 gives an overview of our proposed workflow for the relation extraction.

The first task is to determine a suitable set of target relations. This is done via a technique from Data Mining, i.e. association rule learning using the *apriori* algorithm. This algorithm first extracts all item sets from a list of items having a large support in the dataset. A rule-generation step follows that creates rules for individual items from these sets. In the context of our system, the item sets are built from entity classes, i.e. the co-occurrences of instances of the classes in a given context. The resulting association rules give evidence to the existence of statistically significant relations between participating entity classes. All rules having a confidence higher than a given threshold are subsequently used to automatically analyze the corpus. The algorithm searches for possible labels for the relations gained from each rule. Note that more than one relation type can correspond to a rule.

The next step connects the entities of the net using the labels gathered in the preceding step. To ascertain that the algorithm has learned the right relations, and furthermore has connected the right instances, the domain experts are involved at this stage. They get to see the current hypothesis of the semantic net, and are able to make changes or reinforce the decisions by the algorithm. The next iteration uses the feedback to get a better understanding of the relation to be learned, until the experts are satisfied. To ease the work of the experts, every relation gets learned separately.

5 First Application

The pilot application of the WIKINGER platform is the creation of a web-based "biographical-bibliographic compendium for Catholic Germany". Extensive bio-bibliographical information about German Catholicism in the 19th and 20th century is to be made available in a structured and semantically networked manner to allow the public in order to carry out research more efficiently.

The Commission for Contemporary History (in German KfZG) is in charge of developing and evaluating this system. It will organize and cultivate the contacts to the historical community, which will later have access to the articles via the front-end of the wiki-system.

This wiki will allow all registered participants and scholars of modern Catholic history to collaborate on the subsequent treatment of the extracted and networked data. Over time, it will expand the number and even improve the quality of the articles in the system. Those searching bio-bibliographical special data on Catholic personalities of the last 200 years have been forced to search laboriously through many specialized publications. The publications of the Commission for Contemporary History, the so-called "blue series", with numerous significant monographs on the social, political and cultural history of the German Catholicism in the 19th and 20th century, belong to those special studies. As of December 2006, 158 volumes with a total extent of more than 65,000 printed pages have appeared in this series. Series A features collections of primary sources, while Series B presents specialized monographs on German Catholicism in the 19th and 20th Century.

There are several external data sources which are being considered for inclusion in the course of the project. Among those are the databases of the Central Committee of German Catholics (ZdK) with regard to lay activities of the last 200 years, containing extensive data on the personalities in Catholic public life of that period. Additionally, the Catholic press agency (the German KNA) has offered use of its picture library, which contains approximately 100,000 pictures (usually for the time after 1945) in digital form, depicting many important personalities reported on in the articles of the wiki.

The Commission for Contemporary History serves as the nexus for research on modern Catholicism. As a center for research, it brings together European institutions, universities and individuals, who are researching German Catholicism. Several hundred scientists will belong to the research community, which will form the wiki-community. These include the following:

- Prominent individuals in the church, politics and society;

- Professors of modern history, church history and political science from German and foreign universities and their research assistants and graduate students;
- Authors of „the blue series “;
- Directors and staffs of the German catholic diocesan archives and monastic order archives
- Members of other institutes, such as the Protestant Working Group for Modern Church History;
- Editors of the religious press as well as broadcast and television journalists.

As a pilot user of the system, the Commission for Contemporary History will inform the research community about this new network and keep it updated about latest advances. The research community, which will already have been made aware of the project in its first stages, will be, in turn, informed about the web-based "biographic-bibliographical compendium for Catholic Germany" and introduced to the web-based manual. The presentation and practical introduction to WIKINGER will take place in the framework of meetings and conferences, organized by the Commission for Contemporary History and members of the larger research community.

6 Conclusion

In this paper, we have presented the goal of WIKINGER and illustrated the workflows for manual and semi-automatic annotation as well as the community feedback. The system design has been outlined and the named entity recognition and semi-automatic generation of semantic networks have been described in detail. Finally, the projected pilot application has been introduced.

The paper describes ongoing work and therefore we have not yet reached the evaluation phase. So far, the basic WIKINGER platform has been implemented which consists of a versioning document repository, a prototype of the WALU annotation system and an automatic entity association service for creating the semantic network. A working set of categories relevant for the pilot application scenario is defined using WALU and proved to be applicable. In this context, archived documents have been fed into the document repository and a set of manually annotated examples are available to test the entity association service. As an outcome, a first semantic net of the pilot domain is available, based on the annotated examples provided by the KfZG.

There are various issues for future work. First of all, we work on the completion of the proposed WIKINGER system and its evaluation within the projected pilot application. The transfer of the system to other application domains, e.g. providing harvester for standard web formats, and the integration and collaboration with related systems, e.g. from the eScience research program, are further issues.

Acknowledgements

The project WIKINGER is funded by the German Federal Ministry of Education and Research (BMBF) in the context of the programme “E-Science Wissensvernetzung”. It is jointly conducted by the Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS), the Department of Computer Linguistics at the University Duisburg-Essen, and the Commission for Contemporary History.

References

1. Berners-Lee, T. et al: The Semantic Web, Scientific American, 5/2001, pp. 216-218
2. Vaughan-Nichols, S. J. Web Services: Beyond the Hype. IEEE Computer. Vol. 35, Nr. 2. pp. 18-21. IEEE 2002.
3. Powers, S. Practical RDF. O'Reilly 2003.
4. Andreas Wagner & Marc Rössler: WALU – Eine Annotations- und Lern-Umgebung für semantisches Tagging. GLDV Frühjahrstagung 2007.
5. Marc Rössler: Korpus-adaptive Eigennamenerkennung. Dissertation. Duisburg 2007.
6. Hummel, Karl-Joseph/Burtscheidt, Andreas: Ein webbasiertes Handbuch für das katholische Deutschland. Das »Wikinger«-Projekt im Rahmen der »e-science«-Initiative der Bundesregierung, in: Tagungsband ».hist 2006. Geschichte im Netz – Praxis, Chancen, Visionen«, Berlin (im Druck).
7. Ahmed, S., Chidambaram, D. et al: Intex: A syntactic role driven protein-protein interaction extractor for bio-medical text. In: Proceedings ISMB/ACL Biolink 2005, pp. 54-61
8. Pustejovsky, J., Castano, J., and Zhang, J.: Robust relational parsing over biomedical literature: Extracting inhibit relations. In: Proceedings of the Pacific Symposium on Biocomputing 2002, pp. 362–373.
9. Palakal, M., Stephens, M., Mukhopadhyay, S. et al.: Multi-level text mining method to extract biological relationships. In Proceedings of the 2002 IEEE Computer Society Bioinformatics Conference, pp. 60–67
10. Hasegawa, T. et al.: Discovering Relations among Named Entities in large Corpora. Proceedings of the ACL’04, pp. 415-422.
11. Hummel, Karl-Joseph (Hg.), Zeitgeschichtliche Katholizismusforschung. Tatsachen, Deutungen, Fragen. Eine Zwischenbilanz, Paderborn 2004.
12. Abmeier, Karlies/Hummel, Karl-Joseph, Der Katholizismus in der Bundesrepublik Deutschland 1980-1993. Eine Bibliographie, Paderborn 1997.