Technical Report

The Laurin Inferface Suite: A software package for newspaper clipping archives

Gregor Retti
The Laurin Project
Department of German Language, Literature and Literary Criticism
University of Innsbruck
A-6020 Innsbruck
Austria
Email : gregor.retti@uibk.ac.at

Abstract : This paper is a technical report about the Laurin Interface Suite. It covers architecture and platforms as well as the most important tools of the software package.

1. Introduction

The Laurin Interface Suite was developed during the R&D-project Laurin (1998-2000) (Mühlberger, 2000) and enhanced and completed in the follow-up project Laurin+ (2000-2002). As the major part of the Laurin System (Calvanese *et al.*, 2001) the Laurin Interface Suite is used by the electronic clipping archive of the *Innsbrucker Zeitungsarchiv* since 1999 (cf. http://iza.uibk.ac.at). The aim of the Laurin project was to create a software package for clipping archives which would allow them to entirely digitise the clipping, indexing, storing, and retrieval of the archived material. The Laurin System comprises several components: an acquisition tool scanning newspaper pages and clipping articles, the Laurin Database, which holds all textual data of the system, the Laurin Thesaurus, a multilingual, standards compliant thesaurus (Retti & Stehno, 2003), and the Laurin Interface Suite, which is made up of several tools for managing the data, for indexing and retrieval. This paper focuses on the tools the Laurin Interface Suite.

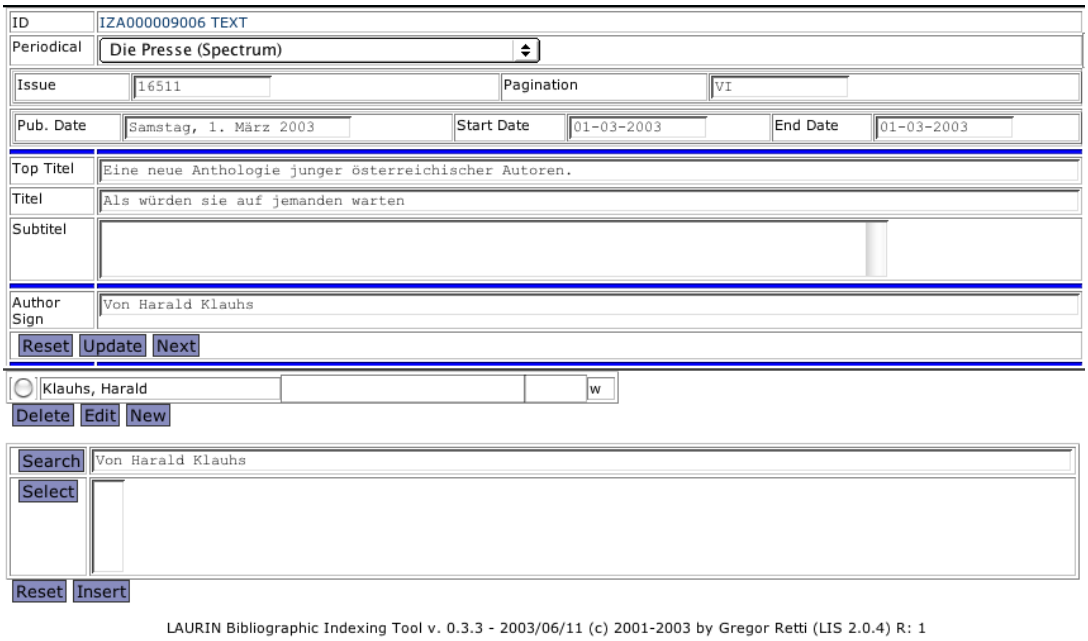
2. Architecture and Platforms

From a technical point of view the Laurin System can be regarded as consisting of a data repository, i.e. a relational database (Retti 2003), an image repository, i.e. a hierarchical file system holding files with image data and OCR text, a scanning and clipping station for data acquisition, and an application server, providing all necessary tools to work with the clipping archive's data. The database has been implemented in Oracle, although it should be rather easy to migrate to an Open Source equivalent like PostgreSQL. Due to the architecture the database server can be separated from the image repository as well as from the application server. The application server is web-based, therefore any current web-browser on a client computer may be used with the system regardless of the operating system of the client computer. The Laurin Interface Suite is written in Perl (Wall *et al.*, 1996) and runs under the Apache HTTP Server (cf. http://httpd.apache.org). Linux is used as the development environment, the production environment at the University of Innsbruck runs partly under Sun Solaris.

3. Indexing Workflow and Indexing Tools

During the analysis conducted in the LAURIN project (Habitzel & Retti, 1999) two major steps in the indexing process have been isolated and described in detail: *bibliographic indexing* and *content indexing*. Bibliographic indexing refers to the exact description of the object in question, i.e. the clipped article, and the required meta-data can usually be found on the object itself. The only exception may be the author's name, which – although depending on the newspaper or journal as well as on national traditions – may appear in an abbreviated form or as a pseudonym. Content indexing, on the other hand, should be understood as a description of the content of the article by determining its *text type*, e.g. "news: breaking news" or "opinion: editorial", and by associating the article to keywords from the Laurin Thesaurus (Retti, *et al.* 2000; Retti & Stehno 2003).

Obviously the skills required to perform these two different tasks are not the same. Therefore, two different tools have been designed and implemented to complete the two different steps of indexing: LBIX, the "Laurin Bibliographic Indexing Tool", is a front-end to display, check, change, and commit the articles' data one record after the other.



ID	IZA000009006 TEXT
Periodical	Die Presse (Spectrum)
Issue	16511
Pagination	VI
Pub. Date	Samstag, 1. März 2003
Start Date	01-03-2003
End Date	01-03-2003
Top Titel	Eine neue Anthologie junger österreichischer Autoren.
Titel	Als würden sie auf jemanden warten
Subtitel	
Author Sign	Von Harald Klauhs

○ Klauhs, Harald | w

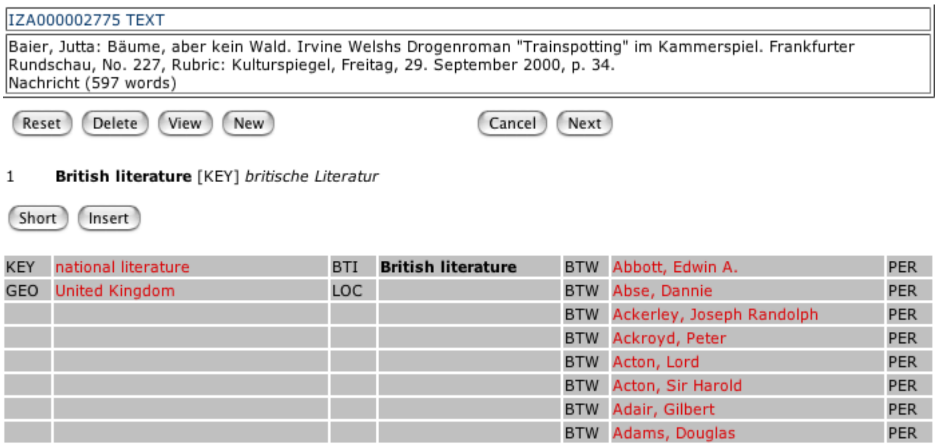
Search: Von Harald Klauhs

LAURIN Bibliographic Indexing Tool v. 0.3.3 - 2003/06/11 (c) 2001-2003 by Gregor Retti (LIS 2.0.4) R: 1

Figure 1.

LBIX, article and author editor

LCIX, the “Laurin Content Indexing Tool”, includes a search facility to retrieve keywords from the Laurin Thesaurus and a thesaurus browser to select and associate the correct thesaurus entries for indexing. Both tools provide a revision component for corrections at a later step.



IZA000002775 TEXT

Baier, Jutta: Bäume, aber kein Wald. Irvine Welshs Drogenroman "Trainspotting" im Kammerspiel. Frankfurter Rundschau, No. 227, Rubric: Kulturspiegel, Freitag, 29. September 2000, p. 34.
Nachricht (597 words)

1 **British literature** [KEY] *britische Literatur*

KEY	national literature	BTI	**British literature**	BTW	Abbott, Edwin A.	PER
GEO	United Kingdom	LOC		BTW	Abse, Dannie	PER
				BTW	Ackerley, Joseph Randolph	PER
				BTW	Ackroyd, Peter	PER
				BTW	Acton, Lord	PER
				BTW	Acton, Sir Harold	PER
				BTW	Adair, Gilbert	PER
				BTW	Adams, Douglas	PER

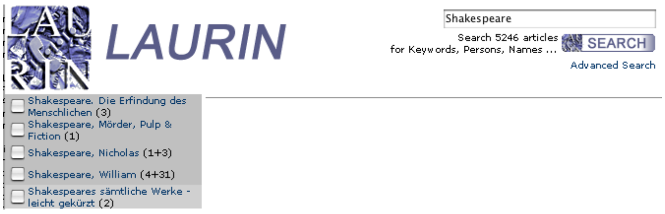
Figure 2.

LCIX, thesaurus browser

New indexing terms may be added to the thesaurus during the process of content indexing. Those new entries are to be controlled, corrected, accepted, or rejected later on during the thesaurus maintenance workflow. It should be noted, that from within LBIX and LCIX the scanned images of the current article are always accessible as a one- or multi-page PDF-file, which is created on the fly from the images stored in TIFF-format.

4. Information Retrieval

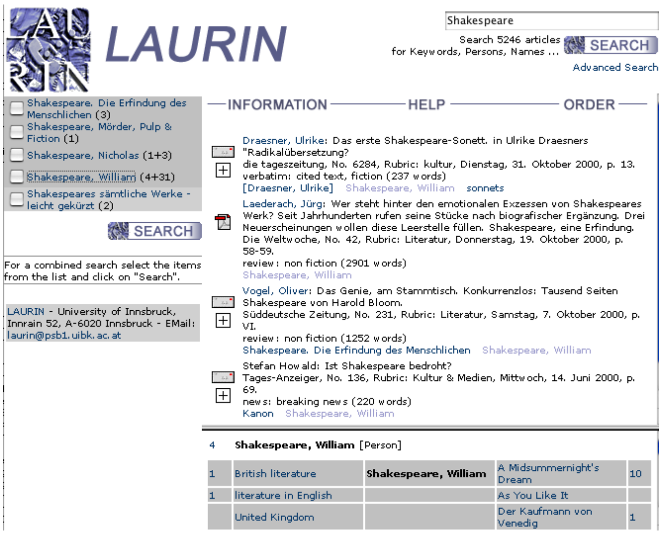
The “LAURIN Search Interface” (LAUS) is used to search for articles in the digital archive. The basic search feature uses a word index, which is build up from the thesaurus entries associated with the articles. This approach fits to the behaviour of the casual web-user, who is used to enter a term into one field and hit the search button or the enter key. Then a scored list of thesaurus entries is presented to the user and a click on one of the list items will display the articles referenced by the item.



□ Shakespeare. Die Erfindung des Menschlichen (3)
□ Shakespeare, Mörder, Pulp & Fiction (1)
□ Shakespeare, Nicholas (1+3)
□ Shakespeare, William (4+31)
□ Shakespeares sämtliche Werke - leicht gekürzt (2)

Figure 3.
LAUS, search result with scored list

LAUS uses frames to display the basic search form, the list of thesaurus entries and the list of articles. Furthermore, an additional frame at the bottom of the browser windows shows a thesaurus browser for the user to navigate through the data. The clipped articles can be downloaded as PDF-files, if the copyright allows the archive to deliver them, or they may be ordered as print-outs through a simple interface resembling an online-shop. The question of copyright has been addressed during the LAURIN Project, but it was impossible to find a general solution. Therefore, the *Innsbrucker Zeitungsarchiv* took a different approach and asked the newspaper publishers for permission. Fortunately as a non-profit research institutions it was granted the right to deliver the articles electronically without charge by almost half of the publishers asked.



□ Shakespeare. Die Erfindung des Menschlichen (3)
□ Shakespeare, Mörder, Pulp & Fiction (1)
□ Shakespeare, Nicholas (1+3)
□ Shakespeare, William (4+31)
□ Shakespeares sämtliche Werke - leicht gekürzt (2)

For a combined search select the items from the list and click on "Search".

LAURIN - University of Innsbruck,
Innrain 52, A-6020 Innsbruck - EMail:
laurin@psb1.uibk.ac.at

Draesner, Ulrike: Das erste Shakespeare-Sonett. in Ulrike Draesners "Radikalübersetzung?
die tageszeitung, No. 6284, Rubric: kultur, Dienstag, 31. Oktober 2000, p. 13.
verbatim: cited text, fiction (237 words)
[Draesner, Ulrike] Shakespeare, William sonnets

Laederach, Jürg: Wer steht hinter den emotionalen Exzessen von Shakespeares Werk? Seit Jahrhunderten rufen seine Stücke nach biografischer Ergänzung. Drei Neuerscheinungen wollen diese Leerstelle füllen. Shakespeare, eine Erfindung.
Die Weltwoche, No. 42, Rubric: Literatur, Donnerstag, 19. Oktober 2000, p. 58-59.
review: non fiction (2901 words)
Shakespeare, William

Vogel, Oliver: Das Genie, am Stammtisch. Konkurrenzlos: Tausend Seiten Shakespeare von Harold Bloom.
Süddeutsche Zeitung, No. 231, Rubric: Literatur, Samstag, 7. Oktober 2000, p. VI.
review: non fiction (1252 words)
Shakespeare. Die Erfindung des Menschlichen Shakespeare, William

Stefan Howald: Ist Shakespeare bedroht?
Tages-Anzeiger, No. 136, Rubric: Kultur & Medien, Mittwoch, 14. Juni 2000, p. 69.
news: breaking news (220 words)
Kanon Shakespeare, William

4 **Shakespeare, William** [Person]

1	British literature	**Shakespeare, William**	A Midsummernight's Dream	10
1	literature in English		As You Like It	
	United Kingdom		Der Kaufmann von Venedig	1

Figure 4.
LAUS, search result with articles and thesaurus browser

Searches may also be performed through an advanced search form, which includes popup-lists of the journals and newspapers covered by the archive, rubrics in these periodicals, text types, and time ranges in month. These lists may be combined with a request for a specific author or a generic search term. The latter will be looked up in the word index mentioned above as well as in an additional word index, which is build from the data in the title, sub-title, and abstract field of the article, thus providing a restricted full text retrieval. Full text search on the OCR-text of the articles is currently on the agenda for future developments.

5. Thesaurus Maintenance

The LAURIN Thesaurus complies with the relevant standards for mono- and multilingual thesauri (ISO 2788, 1986; ISO 5964, 1985). Thesaurus entries are moulded after the well-known linguistic sign model (Saussure, 1967), differentiating expression and meaning as *concept* and *name* (Retti & Stehno, 2003). The LAURIN Thesaurus has been filled with data from the Getty Thesaurus of Geographic Names™ (Harpring 1998) and the Nomenclature of Territorial Units for Statistics for the European Community (NUTS 1995). A top level structure according to the IPTC Subject Reference System was added (IPTC) and most of the OECD Macrothesaurus was merged into the LAURIN Thesaurus (OECD 1997). Furthermore, a database maintained at the *Innsbruck Zeitungsarchiv* holding about 35.000 records about writers and poets was imported. Thus, the LAURIN Thesaurus started from the early beginning with an inventory of approximately 200.000 concepts, i.e. thesaurus entries, and 250.000 names.

Due to the fact that the LAURIN System was originally designed as a network of online archives with a thesaurus distributed throughout that network on local nodes grouped around a central node, the issue of quality control regarding the maintenance of the LAURIN Thesaurus had to be addressed from the very beginning of the project. The LAURIN Interface Suite provides a tool for the *local* level, LNTM, the “Local Node Thesaurus Manager”, and one for the *central* level, CNTM, the “Central Node Thesaurus Manager”. The following schema gives an overview of this workflow:

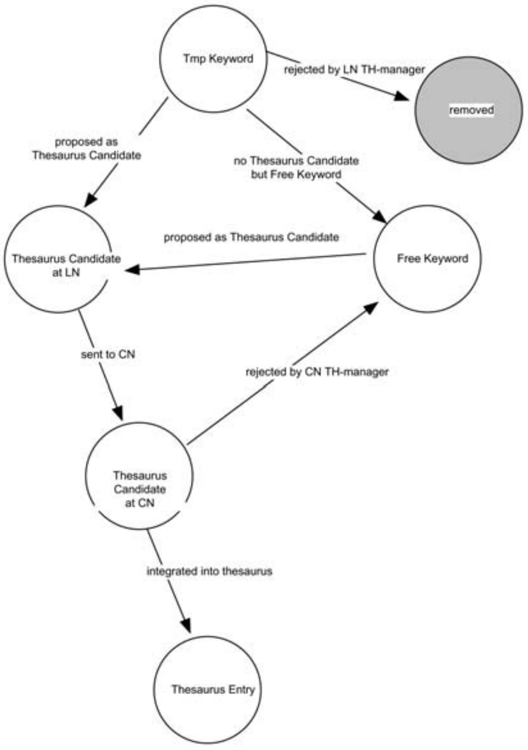


Figure 5.

State diagram for the thesaurus entries

When a new item is added to the thesaurus, it is marked as a *temporary keyword*. It has then to be decided whether this temporary keyword will be removed, whether it will we be promoted to be a *thesaurus candidate*, or whether it will be turned into a *free keyword*. Free keywords are not really integrated into the thesaurus, as the do not maintain relations to other thesaurus entries. On the other hand, free keywords are very useful for new or upcoming terms often to be found in newspapers. Some of those terms may be turned into regular thesaurus entries later, but other are just fashion and disappear from the media soon. When such free keywords appear first, it is often difficult to obtain an exact definition and, therefore, to determine the correct place for them within the thesaurus. Thesaurus candidates are forwarded to be examined on the central level, where they may be accepted or rejected. If the model of the thesaurus entries as made up of concepts and names is taking into account, the picture of the workflow gets a little more complicated:

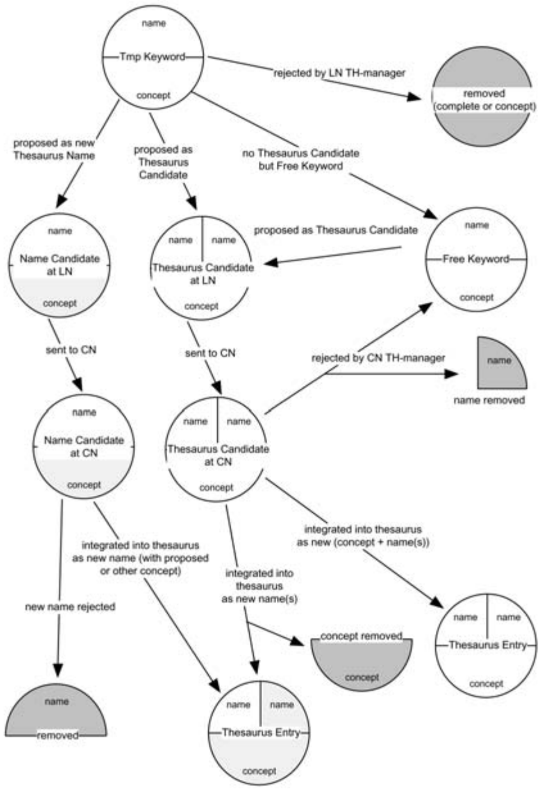


Figure 6.

State diagram for the thesaurus entries including *concept* and *name*

CNTM, the "Central Node Thesaurus Manager", is used to review thesaurus candidates, to apply structural changes to the thesaurus by merging or moving entries, and to add new thesaurus entries if required.

6. Conclusion

The Laurin System, though designed and planed to be run by a network of clipping archives, is used in a production environment only be the *Innsbrucker Zeitungsarchiv* at the University of Innsbruck. As a matter of fact the software was hardly usable after the first phase of the project was finished. Therefore, the other partners from the EU-funded project did not deploy the system. In the second phase of the project, which was only conducted on a national level, the Laurin Interface Suite was enhanced and completed. Besides the tools mentioned above there are interfaces to edit users of the system, to edit periodical data, to display statistics on acquired and indexed data etc. Together with the Laurin Database it is available for free download at http://laurin.uibk.ac.at/.

Journal of Digital Information Management Special issue

ITCC 2004 Track on

Methodologies, technologies and applications in distributed and Grid systems

To be held in conjunction with the **5th International Conference on**:
Information Technology: Coding and Computing
The Orleans, Las Vegas, Nevada, USA
April 5-7, 2004

CONTEXT and GOALS:

Computational Grids, initially used for the sharing of distributed computation resources in scientific applications, start to be used in different application domains offering basic services for application definition and execution in heterogeneous distributed systems.

In health systems, the Grid offers the power and ubiquity needed to the acquisition of biomedical data, processing and delivering of biomedical images (CT, MRI, PET, SPECT, etc) located in different hospitals, within a wide area. So, the Grid acts as a *Collaborative Working Environment*: doctors often want to aggregate not only medical data, but also human expertise and might want colleague around the world to visualize the examinations in the same way and at the same time so that the group can discuss the diagnosis in real time.

The Grid offers a dynamic infrastructure for retrieving and on-demand processing of remote sensing data, for instance, retrieving of SAR metadata related to terabyte of SAR data, starting on-demand processing on raw data, starting on-demand post-processing on focalized data and creating a complex application composing simple tasks.

For atmospheric and climate modeling, a Grid offers tools for simulate and forecasting meteorological phenomena, simulate emission and dispersion of pollutants for air quality studies and simulate complex phenomena about the impact of global climate changes.

Grid Computing techniques can be used in the motor industry, reducing the optimization process time for improvement of diesel engine emission performance using, for instance, micro-genetic algorithms for engine chamber geometry optimization and Kiva3 code to calculate chamber geometry fitness.

In the computer aided medicine, a new research area involves the use of the Grid technologies for surgical simulations. Some simulations could be performed in a distributed system to allow surgeons to practise executing of particular surgical procedures. Analysis of the problems relevant to the use of GRID in medical virtual environments will be appreciated.

Finally, bioinformatic applications call for the ability to read large datasets (e.g. protein databases) and to create new datasets (e.g. mass spectrometry proteomic data). They can require the ability to change (updating) existing datasets; consequently a Data Grid, i.e. a distributed infrastructure for storing large datasets, is needed. In the bioinformatic field, a Data Grid could reveal useful to build Electronic Patient Record systems (EPRs) for the management of patient information (data, metadata and images), to support data replication, allowing the integration and sharing of biological databases and, generally, for the developement of efficient bioinformatics (in particular proteomic) applications.

The main goal of the Conference Track is to discuss well-known and emerging data-intensive applications in the context of distributed systems and Grid systems, and to analyze technologies and methodologies useful to develop such applications in such environments.

In particular, this Conference Track aims at offering a forum of discussion where young researchers and PhD students could present their research activities, either at an early or mature phase.

TOPICS:

The topics of interest include (but are not limited to) the following:

- Data intensive applications in distributed and Grid systems:
- Grid for biomedical imaging;
- Grid for remote sensing and GIS application;
- Grid for Atmospheric and Climate Modeling;
- Grid for motor industry (diesel engine simulation);

- Grid for surgery simulations;
- Bioinformatic for:
- Biomedical Imaging;
- Proteomics and genomics;
- Electronic Patient Records;
- Medical images, data and metadata management; Image Recognition,
- Processing and Analysis.
- Technologies and methodologies in distributed and grid-based applications:
- Grid technologies (Grid portals, Web & Grid services, portlets);
- Grid Information and Monitoring services and related (OO,Relational,XML) data models;
- Grid Security;
- Grid Workload and Data management services;
- Grid Resource management;
- Parallel and Distributed application (cluster and grid based);
- Simulation and Applications of Modeling.

PUBLICATION:

The selected papers after review and extension will be published in a special issue of the Journal of Digital Information Management.

TRACK CHAIR

Maria Mirto

Center for Advanced Computational Technologies
(CACT/ISUFI),
University of Lecce,
Via per Monteroni, 73100 Lecce, Italy
Email: maria.mirto@unile.it

Guest Editors of the special issue:
Giovanni Aloisio.