ABSTRACT:

ONLINE SERVICES OF HUGE STRUCTURED TEXT CORPORA

Király Péter, pkiraly@tesujionline.com
Tesuji Hungary Ltd. (http://www.tesujionline.com)

We are seeking answers for the following questions:
- how to present texts, mails, articles, bills, schedules, balance sheets etc. stored on DB, WedDav, mail- and webservers
- how to find at a glimpse the information you need, no matter where it is and which format the file is in that stores your data: Word, Excel, pdf, html, xml, rtf, odt, swx, txt...
- how to publishe on intranet or on Internet the whole archive of your organization, company, institution, digital library, collection of e-learning texts and books... any kind of structured or unstructured digital content, in a comprehensive, user friendly, easy to consult form...
- how to give your users variable and effective search opportunities
- how to integrate a sophisticated but user-friendly tool that always ensures up-to-date search results by automatically re-indexing the whole content into the existing portal or document-managing system

The lecture shows some Open Source subsystems, which can operate as modules of complex application, and one such application - namely Anacleto Digital Library - through an example of a realy huge text corpora (more than 7 billion character which equals 2-3,5 million book page).