Protocol: Software Development meeting

Date: 30.09.2008, 13-15 pm
Participants: Malcolm Hyman, Wolfgang Schmidle, Peter Damerow, Robert Casties, Stefan Trzeciok, Josef Willenborg, Dirk Wintergrün
Protocol writer: Josef Willenborg

1. Presentation of Lucene (by Dirk Wintergrün)
Indexing: Separate fields for the content and the normalized content (with donatus) and language specific indexing.
One problem: Lucene query results does not deliver the position in an indexed document
Harvester: read files from specified directories and use them for indexing
Language analyzers: for each language a special analyzer, also a special donatus analyzer
Filter: LowCaseFilter, DonatusFilter
While analyzing documents the Donatus-Server is called (by a special XML-file) and delivers for each language and document all tokens (base and inflectional form) as an DonatusResultDocument.
While analyzing also each document could be separated into some documents (e.g. for each sentence) which are then each indexed seperately (Language analyzer: addDocument(...)).
Querying: langugae specific querying (FulltextSearch.java)
OCR: Ocropus documents are analyzed and indexed: see vlp.mpiwg-berlin.mpg.de (→ fulltext search)
Performance of OCR indexing: 500 documents (each with 300 pages) needs 8 hours
Sources see: https://itgroup.mpiwg-berlin.mpg.de:8080/svn/fulltextsearch/trunk/

2. Next dates:
- Presentation ESciDoc 16.10.-17.10.2008 in Munich
- Presentation eXist: in 2 weeks by Dirk Wintergrün
- Presentation Digilib: 6.10.2008, 14-15 pm by Robert Casties
- Meeting of the Schema group: 6.10.2008, 15-16 pm