

## Document Processing (E018921)

**Course size** (nominal values; actual values may depend on programme)

**Credits** 4.0      **Study time** 120 h      **Contact hrs** 30.0 h

**Course offerings and teaching methods in academic year 2017-2018**

|                |            |        |
|----------------|------------|--------|
| A (semester 1) | group work | 15.0 h |
|                | lecture    | 15.0 h |

**Lecturers in academic year 2017-2018**

|                    |      |                    |
|--------------------|------|--------------------|
| Bronselaer, Antoon | TW07 | lecturer-in-charge |
|--------------------|------|--------------------|

**Offered in the following programmes in 2017-2018**

|   | crdts | offering |
|---|-------|----------|
| <a href="#">Master of Science in Computer Science Engineering</a> | 4     | A        |
| <a href="#">Master of Science in Computer Science Engineering</a> | 4     | A        |

**Teaching languages**

English

**Keywords**

Document, document processing, information retrieval, PDF, XML, search engine

**Position of the course**

Today, a very significant fraction of an organization's knowledge and information is stored in documents with little structure. Document processing and document management aim at the introduction of methods to structure documents, and to manage efficiently collections or ensembles of document that may be very large. To achieve this, it is necessary to acquire an elementary knowledge, not only of text and document processing, but also the technologies used in storing and retrieving documents.

**Contents**

- Models and transformations of documents: Document formats, transitions from logical structure to physical representation, transitions from physical representation to logical structure.
- Text processing inside documents: Typefaces and fonts, line-breaking algorithms, page description languages and the portable document format.
- XML technology: XML (eXtensible Markup Language), the schema languages DTD, XML schema and RelaxNG, formal document models, XML transformations.
- Information retrieval: Boolean retrieval, Index construction, Vector-space model, probabilistic document model.
- Text mining: Document clustering, document classification.
- Web search: Web crawlers, link analysis, XML retrieval, XPath.

**Initial competences**

Elementary knowledge of programming

**Final competences**

- 1 Knowledge of the various formats for document storage.
- 2 Knowledge of aspects of document layout.
- 3 Knowledge of aspects of XML technology and the ability to apply them.
- 4 Knowledge of the principles of information retrieval and the ability to apply them.
- 5 Knowledge of the basic methods for classification and clustering of documents.

**Conditions for credit contract**

Access to this course unit via a credit contract is determined after successful competences assessment

**Conditions for exam contract**

This course unit cannot be taken via an exam contract

**Teaching methods**

Lecture, self-reliant study activities

**Learning materials and price**

The course notes are made available on Minerva, as the course progresses throughout the semester.

**References**

- Digital Typography, Donald Knuth, CSLI Publications, 1999
- Digital Typography, An Introduction to Type and Composition for Computer System design, Richard Rubinstein, Addison-Wesley, 1988
- The Concise SGML Companion, Neil Bradley, Addison-Wesley, 1996
- The XML Companion, Neil Bradley, Addison-Wesley, 1998
- The XML Schema Companion, Neil Bradley, Addison-Wesley, 2003
- XSL Formatting Objects, Sharon Adler Ed., Sams Publishing, 2003
- Document Warehousing and text Mining, Dan Sullivan, Wiley, 2001
- Introduction to Information Retrieval, C. D. Manning, P. Raghavan, H. Schuetze, Cambridge, 2008.
- Understanding Search Engines, Michael Berry and Murray Browne, SIAM, 2005

**Course content-related study coaching**

Interactive support and coaching through Minerva (a course forum; students may open up new threads themselves); appointments, upon request by e-mail, for personal issues.

**Evaluation methods**

end-of-term evaluation

**Examination methods in case of periodic evaluation during the first examination period**

Written examination, open book examination

**Examination methods in case of periodic evaluation during the second examination period**

Written examination, open book examination

**Examination methods in case of permanent evaluation****Possibilities of retake in case of permanent evaluation**

not applicable

**Extra information on the examination methods**

During the examination period, there will be an exam consisting of a part in which theoretical aspects are tested (written closed-book exam) and a part in which practical aspects are tested (written open-book exam)

**Calculation of the examination mark**

Weighing:

- 1/2 of the end score is determined by the evaluation of the answers to the questions on the theoretical part of the exam
- 1/2 of the end score is determined by the evaluation of the answers to the questions on the exercise part of the exam