GHENT
UNIVERSITY

# Course Specifications

From the academic year 2021-2022 up to and including the academic year

## Document Processing (E018921)

**Due to Covid 19, the education and assessment methods may vary from the information displayed in the schedules and course details. Any changes will be communicated on Ufora.**

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

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

**Course offerings and teaching methods in academic year 2021-2022**

| | | | | |
|---|---|---|---|---|
| A (semester 1) | English | Gent | lecture | 15.0h |
| | | | self-reliant study activities | 15.0h |

**Lecturers in academic year 2021-2022**

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

| **Offered in the following programmes in 2021-2022** | **crdts** | **offering** |
|---|---|---|
| Master of Science in Computer Science | 4 | A |
| Master of Science in Computer Science Engineering | 4 | A |
| Exchange Programme in Computer Science (master's level) | 4 | A |

### Teaching languages

English

### Keywords

Document, document processing, information retrieval, Knuth's algorithm, XML, XPath

### 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, Knuth's line-breaking algorithm, 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.
- Document search: regular expressions and XPath.
- Web search: Web crawlers, link analysis, XML retrieval.

### 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 elementary methods for searching in 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 the electronic learning platform, 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 the electronic learning platform (a course forum; students may open up new threads themselves); appointments, upon request by e-mail, for personal issues.

### Assessment moments

end-of-term assessment

### Examination methods in case of periodic assessment during the first examination period

Written examination, Open book examination

### Examination methods in case of periodic assessment during the second examination period

Written examination, Open book examination

### Examination methods in case of permanent assessment

### Possibilities of retake in case of permanent assessment

not applicable

### Extra information on the examination methods

During the examination period, there will be a written open-book exam

### Calculation of the examination mark