# Semi-Structured Data
# Vorlesung mit Übung (VU 2.0)

Eleonora Laurenza     Emanuel Sallinger     Mantas Šimkus

Database and Artificial Intelligence Group
Institute of Logic and Computation
TU Wien

Winter Semester 2025

- Registration for the course in TISS not later than 09.10.2025 is required.
- Currently overbooked!

# Content

- Basics
    - XML
    - XML Namespaces
- Schema Languages
    - DTD
    - XML Schema
- Query Languages
    - XPath
    - XQuery
    - XSLT
- APIs
    - SAX
    - DOM

- Graph-structured Data
    - Property Graphs / Neo4J
    - Cypher QueryLanguage
- Other technologies
    - JSON (JavaScript Object Notation)
    - RDF (Resource Description Framework)

# Textbooks

- **XML in a Nutshell**

  Eliotte Rusty Harold, W. Scott Means
  O'Reilly, 3$^{rd}$ Edition, 2004

  - Free online access in the TU network

- **Java und XML** (in German)

  M. Scholz und S. Niedermeier
  Galileo Computing, 2$^{nd}$ Edition, 2009

- **XQuery** (English)

  P. Walmsley
  O'Reilly, 1$^{st}$ Edition, 2007

- **Graph Databases** (English) by Ian Robinson, Jim Webber
  & Emil Eifrem O'Reilly, 2nd Edition, 2015

- Free online access in the TU network

## Further sources

- W3C Recommendations
    - http://www.w3.org/TR/xml/
    - http://www.w3.org/TR/xml-names/
    - http://www.w3.org/TR/xmlschema-0/
    - http://www.w3.org/DOM/
    - http://www.w3.org/TR/xpath
    - http://www.w3.org/TR/xpath20/
    - http://www.w3.org/TR/xslt
    - http://www.w3.org/TR/xquery/
- Further links:
    - http://www.json.org/
    - http://www.saxproject.org/
    - http://w3schools.com/
    - http://www.cafeconleche.org/books/xmljava
      (Online version of the book "Processing XML with Java")
    - https://neo4j.com/docs/getting-started/

# Basic Information

- All key information and materials will be on TUWEL

- Lecture mode: live lectures
  Tuesdays, 10:00 - 12:00, FAV Hörsaal 1 Helmut Veith

- Lecture videos (pre-recorded) available on TUWEL

- Homework and work in groups:

  - There are 8 exercise sheets.

  - Students must present/discuss their solutions in groups of 3 students.

  - Students upload their solutions and discussion report to TUWEL.

  - The final grade is based on homework and group work, with no final exam.

# Lecture Plan

- October 7: Kick-off meeting. Introduction to Semi-structured Data and Towards XML
- October 14: XML Fundamentals  + Namespaces
- October 21: Document Type Definition (DTD)
- October 28: No lecture
- November 4: XML Schema Definition (XSD)
- November 11: XPath (Fundamentals, Further Examples, Usage in XSD)
- November 18: XSLT
- November 25: Simple API for XML (SAX) + Document Object Model (DOM)
- December 2: XQuery
- December 9: Other SSD Technologies
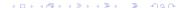- December 16: Graph Databases (incl. Property Graphs, Cypher query language)

## Exercises: General information

- You will receive 8 exercise sheets during the semester

- You are asked to solve them individually

- To count towards the final grade, each exercise sheet needs to be <span style="color:red">finalized</span>:

    - the student needs to present and discuss their solutions in a small group of students (3 people)

    - the student not only presents but also listens to the presentations of fellow students and provides feedback

    - this happens during dedicated "Work in Groups" sessions

    - approx. 45 minutes discussion per exercise sheet

# Exercise Sheets and Timeline

8 exercise sheets will be made available during the course:

1. XML+DTDs (to be provided by October 13, recommended finalization by November 13)
2. XML Schema (to be provided by November 3, recommended finalization by November 20)
3. XPath (to be provided by November 10, recommended finalization by November 27)
4. XSLT (to be provided by November 17, recommended finalization by November 27)
5. SAX (to be provided by November 24, recommended finalization by December 11)
6. DOM (to be provided by December 1, recommended finalization by December 16)
7. XQuery (to be provided by December 1, recommended finalization by January 15)
8. Property graphs (to be provided by December 15, finalization by January 22)

# Work in Group Sessions

During one concrete "Work in Group" session:

- A group of 3 students is formed at the beginning

- The students present and discuss their solutions in the group

  (ideally, at least one personal laptop per group is available)

- The students together fill in a discussion report (a paper form will be provided); the report is scanned at the end of the discussion

  (ideally, at least one smartphone per group is available)

# After a Work in Group Session...

After a discussion session, within 1 week, each student uploads to TUWEL:

- the scan of the discussion report

- their individual solution to the exercise sheet

- a short individual reflection on the exercise sheet and the student's solution to it, as well as a reflection on the work in the group (at most 300 words).

  You will use an XML file to store the individual reflection and other basic information about the submission.

# Work in Groups: Slots During the Semester

Sessions for group work (i.e. completion of exercise sheets) will be held during the whole semester, **starting from October 23  until January 22 (except October 28, October 30, December 18)** as follows:

1st slot: Tuesdays 08:00 - 09:00 in FAV Hörsaal 1 Helmut Veith - INF

2nd slot: Tuesdays 09:00 - 10:00 in FAV Hörsaal 1 Helmut Veith - INF

3rd slot: Thursdays 15:00 - 16:00 in EI 11 Geodäsie HS - INF

4th slot: Thursdays 16:00 - 17:00 in EI 11 Geodäsie HS - INF

**Place reservation will be required (you will be able to book your place for a slot here on TUWEL).**

# Reserving a slot for work in groups

- Exactly 1 exercise sheet can be finalized per session!
    - i.e., finalizing solutions to 8 exercise sheets requires presence in 8 sessions
- Schedule attendance in at most 1 session per week! (exception only for emergencies)
- With 4 available weekly slots during the semester, you have significant flexibility to plan

- However, it is critically important that you finalize your exercise sheets as soon as possible, because:
    - the number of students per session is limited
    - you need to reserve a spot for a discussion and the older the exercise sheet the lower the piority for getting a spot
    - if you are very late, you need to find two companions to deal with the same exercise sheet

## Some remarks

- We did not want to put deadlines for submitting/discussing exercises

- This was to support students with exceptional circumstances

- However, we need the vast majority of the students to work consistently during all of the semester

  - please do your best to follow the recommendations!

  - we are running this course with many students and very limited resources

- Constructive criticism is welcome (directly to us or on TISS at the end of the course)

# Communication

- TUWEL: please read carefully and visit regularly!
    - the general information about the course
    - the lecture videos, slides, and all supporting material
    - news in the "News Forum" and discussions in the "Discussion Forum"
    - booking a slot for work in groups

- TISS:
    - Announcements in TISS only for emergencies
    - Requirement: you need to be subscribed in order to receive the announcements

- Questions:
    - the TUWEL discussion forum
    - the Q&A sessions
    - ssd@dbai.tuwien.ac.at (Subject "SSD:")