4IZ441: Semestral project (up to 35 points; minimum 20)

Of this, at most 2 points are provided for the oral presentation of the project, and at most 33 points for the written / implemented part of the project.

The semestral project is an individual, original project providing a research or practical contribution to the field/s of graph data and knowledge. There is a number of options of how to tune such a project, explained below.

Note: The options are partly a legacy of the pre-cursor course, thus their linkage to the 4IZ441 is not perfect, and some links might be slightly outdated. Any other topic related to 4IZ441 would however fall under the final variant and is open for negotiation.

Schedule:

- **December 6, until 23:59:** submission of a project 1-pager (outline) to InSIS
  - This is a condition for getting points for the subsequent presentation
- **December 9, from 14:15:** oral presentations of project proposals (ideally, with some early results) – within a "workshop" held during the labs.
  - Duration of one presentation: 2-5 minutes
  - Awarded by up to 2 points
- **January 17, 2025, until 23:59:** submission of a complete project
  - The project can still be improved afterwards if more points are needed, according to the lecturer
- **Week of January 20, 2025:** brief presentation (Q&A) of the projects, individually, to the lecturer.
  - Both the physical and online modes are possible.
  - Registration for the individual slots will start in early January at the latest

For students with **special deadlines** (final defence; end of exchange; etc.) the schedules can be adapted appropriately.

Semestral project content: general rules

- All variants of project that are not by themselves of "textual" nature – e.g., programming or vocabulary development – imply that at least a short **document** (such as 3-5 pages in PDF) is submitted to InSIS in the role of "project" anyway. The document should contain a **link** to the actual project achievement (e.g., stored in a github repo), the description of the **problem** addressed, and (in sufficient detail) the approach used in **solving it**.
- Only briefly describe pre-existing technologies you are using. Rather focus on the **original part** of your solution, on the **new findings**, on your original **experimental protocol**, etc.

Tentative, non-exclusive variants of the project

1. Topical projects associated with the **Department's research**. In this setting, the project can correspond to an early phase of work on a **diploma thesis**. However, it is possible to undertake such a kind of project even within a link to own diploma thesis. (Your accomplished 4IZ441 project can then be potentially reused by the actual person preparing the corresponding diploma thesis, as an aid in his/her onboarding.)
2. Testing the usability of a linked data tool using some test methodology that you can find in the software testing literature. The testing must include aspects of the tool that are relevant

to working with RDF data (or other data specifically relevant to the subject), i.e. not just common features shared by other classes of software.

3. **Extraction/transformation/linking** performed on real data, with usable data output. The output must be evaluated in some way, e.g. using RDFunit, or specifically qSKOS for resources in SKOS.
    - Especially welcome are activities over Czech public data, either currently exposed in the National Catalogue of Open Data (NKOD, one should look at RDF formats such as RDF TriG, etc.), or prospectively leading to enrichment of these resources with linked data. SPARQL queries to them can rely on e.g. the Handbook of Data Journalism, or process some other dataset not yet processed.
    - Similarly, activities in the direction of the Czech DBpedia are welcome.

    Specific technological elements (can be combined):
    - Cleaning the selected RDF dataset using SPARQL UPDATE operations
        - One possibility may be to extend the cleaning set for the Public Procurement Bulletin of the Czech Republic (cooperation with J. Mynarz)
    - Extraction from tabular data using LinkedPipes ETL, tarql (for CSV), XSLT (for XML), ev. OpenRefine for HTML (with subsequent tarql application) npod.
    - SPARQL UPDATE operation performing linking (deduplication) of entities within one or different datasets
    - A set of mapping rules for Czech DBpedia.
4. Interesting custom applications **consuming** linked data, e.g.
    - A visualization application, including source code, preferably (but not necessarily) primarily built on top of datasets from the Czech Republic (e.g. those from NKOD).
        - In addition to a description of the application from a technological and user perspective, the accompanying document will also include a proposal for the range of people for whom the application could be useful and a scenario for its use to meet a realistic information need.
    - A "smart" client for the Triple Pattern Fragments interface.
    - A sample application for the **decentralized web** built on top of the Solid platform.
5. Design of a custom, non-trivial, **data vocabulary**, e.g., a product ontology (RDFS or simple OWL) corresponding to schema.org and applicable for describing public contracts. This may be a different vocabulary depending on interest. The vocabulary should, as far as relevant, reuse other existing vocabularies and must be complemented with:
    - **sample data** (separately in Turtle, not just as part of the vocabulary file) described by this vocabulary and other relevant vocabularies.
    - at least 2 sample **SPARQL queries** that will return results over this sample.
6. Design a **SHACL shape graph** for entities from some existing vocabulary, with verification of functionality over some real dataset.
7. A completely **custom** topic in consultation with the lecturer (again, this may include e.g. a specification of application that is to be later realized within a master thesis).