- Group project presentation is on **January 18** 16:00 (for A*) and on **January 25** 16:00 (for B*); the order of presentation will be announced before (we will try to structure the presentations in a way that they fit together nicely) – please notify us as soon as possible in case there are any severe time constraints with your group (e.g. due to exams)
- Compulsory attendance of every group member in the **entire session** that your group presents in (A or B) (exceptions only for serious reasons)
- The group projects cover related topics that have not been discussed in the lecture in detail.
- If possible for your chosen topic, please prepare a live demo as a part of your presentation.
- Please send your files with your solutions/slides at least one day before the last meeting, that is 17th of January, resp. 24th of January.
- Please use the syntax „[regid]_project.[suffix]“ where regid is your group number; suffix depending on file type (zip, rar, tar.gz). In case you send an update please add a “U” at the end.

Procedure:

Topic Selection: Each group picks one project topic. Please send a mail to wdei@dbai.tuwien.ac.at with your topic preferences. In general, each topic will be distributed once only per unit (i.e. a group in B could choose the same topic as a group in A) and handled in a first come first served way. Please let us know your preferences until **December 14**.

Project/Talk: In the final session on *January 18/25,* one or more members of your group give the group talk, and you send the final paper to wdei@dbai.tuwien.ac.at too. Your papers and slides will be made available afterwards to the other participants of the course. Let me know if I need to install any particular software on the presentation notebook by **January 15/23**.

Preparation: The final exercise sheet (#6) handed out on **14th of December** will be about creating a mindmap for the group project that will be presented by each group on the **11th of January** and be used to give feedback for final preparations of the talk and the paper.

TOPIC LIST

Topic Area Wrapper/Deep Web Navigation Frameworks/Languages:

1 **Web Data Extraction using vision-based approach**. Give an overview about the paper “ViDE: A Vision-Based Approach for Deep Web Data Extraction” (http://www.cs.binghamton.edu/~meng/pub.d/ViDE-tkde.pdf), and the ideas and challenges of vision-based data extraction.

2 **Table Extraction.** Use the paper “Towards Domain-Independent Information Extraction from Web Tables” (http://www2007.org/papers/paper790.pdf) as a starting point to give an overview of approaches and methodologies for extracting tabular data from Web pages. Try to find other recent papers about Web table extraction and compare them (at least on a high level).

3 **Sikuli** is a visual technology to search and automate graphical user interfaces (GUI) using images (screenshots) (http://groups.csail.mit.edu/uid/sikuli/). Explain the approach, create sample macros recorded on the visual level, and compare the approach to common Web wrapper generation approaches. As one source of information about Sikuli use the paper “Sikuli: Using GUI Screenshots for Search and Automation” (UIST 2009) available on the Web page.

4. **DOM 3 Events and Ajax in Deep Web Navigation**. Use the paper “Web Navigation Automation in AJAX Websites” (from the group of Montoto/Pan/Bellas, available from us) as starting point to give an overview about challenges in Deep Web navigation in modern Web pages. Based on this paper, give an overview of the DOM3 event model and asynchronous interactions in general.

Topic Area HTML/XML:

5. Describe **SAXON CE** (http://www.saxonica.com/ce/doc/about/intro.xml), the Saxon Client Edition. Saxon CE is the first implementation of XSLT 2.0 running in browser environments and extends the language from just generating HTML to handle user interactions as well. Motivate and describe the approach, its JavaScript API, and give some examples (for instance based on sample applications and demonstrations on http://www.saxonica.com/ce/doc/samples/intro.xml and http://www.saxonica.com/ce/doc/demonstrations/intro.xml).
6. Give an overview and examples about the recently introduced **Web Workers, Web Sockets and Web Storage APIs** (http://www.w3.org/TR/workers/, http://www.w3.org/TR/websockets/, http://www.w3.org/TR/webstorage/). Web sockets enable Web pages for two-way communication with a remote host, Web workers allow for thread-like operations, and Web Storage extends data storage concepts in Web clients.

Topic Area: Web Mining and Web of Data

7. Give an overview on the area of **Web Opinion Mining / Sentiment Analysis**, e.g. based on material available on Bing Liu’s page (http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html.)
8. Give an overview about the **DBPedia** project (http://dbpedia.org). This project extracts data from Wikipedia and creates structured information as Linked Open Data knowledge base. Describe how data is identified and extracted and how DBPedia can be queried.
9. Give an overview about **plagiarims checkers** and their used (text extraction) techniques (use e.g. showing some example tool listed on http://webhugger.com/google/a-review-of-free-online-plagiarism-checkers/ and using this paper as starting point: http://jucs.org/jucs_12_8/plagiarism_a_survey/jucs_12_08_1050_1084_maurer.pdf)

Topic Area: The Crowd

10. **Crowd Sourcing**. Describe and give an overview about Web-based crowd-sourcing. As starting points, you can e.g. use the Amazon Mechanical Turk Service (http://mturk.com), look at articles/books of Jeff Howe, or crowd sourcing conferences such as CrowdConvention. In particular, elaborate use cases in which crowd sourcing can be useful in the Web data extraction and integration value chain.
11. Analyse some available **Web 2.0 Page Annotation Tools** that allow to annotate and comment on Web pages and share/publish/collaborate (e.g.: https://addons.mozilla.org/en-US/firefox/addon/reframe-it/, the (no longer available) Google SideWiki, http://en.wikipedia.org/wiki/Web_annotation, http://www.diigo.com/)
12. Give an overview of the **Microsoft Azure DataMarket** marketplace (https://datamarket.azure.com/), describe its concepts and give an overview of data sets, APIs and scenarios from developer and consumer perspectives.

Topic Area: Tools for Matching, Mashups and Accessibility

13. **COMA++** (Schema Matching Framework): Study and test the academic Schema Matching Framework Coma++ (http://dbs.uni-leipzig.de/Research/coma.html, use the Coma

Community Edition) and show some live examples (ideally based on Web data you extracted during the lecture).

14. Give an overview of the **IBM Mashup Center** (http://www-10.lotus.com/ldd/mashupswiki.nsf/) and its possibilities to compose mashup applications (refer to http://www-07.ibm.com/hk/e-business/events/archives/2008/downloads/why_mashups_matter.pdf for motivation and terminology).
15. Give an overview and experiment with the **Window Eyes** screen-reader (Windows) and *in particular* evaluate its usefulness for Web browsing and regarding understanding of elements (e.g. anchor/jump points, fill out forms, etc.) (http://www.windoweyes.de/).

Topic Area: Semantics

16. **Integration of Semantic Interaction in Web Applications**: Give an overview about Apache Stanbol (http://incubator.apache.org/stanbol/) and the Vienna IKS (http://www.iks-project.eu/) Editables on http://viejs.org/ (e.g. to query and embed information from DBPedia) and how to use it in Web applications.
17. **Navigate the Web of Data:** Give an overview about the swget tool (http://swget.wordpress.com/example1/) and the paper "Semantic Navigation on the Web of Data: Specification of Routes, Web Fragments and Actions" (http://www.dcc.uchile.cl/~cgutierr/papers/www2012.pdf).

Deliverables – Group Project Presentation and Handouts:

Please prepare a **paper** (e.g. in Word, OO, LaTeX, and convert it to PDF). Please use the Springer lecture notes format (templates are available on http://www.springer.com/computer/lncs?SGWID=0-164-6-793341-0).

The paper should contain an Abstract and literature/Web references as appropriate. The size of the paper should be approx. X pages in the above given format. The number X is the number of active group members times 2. It can contain (a reasonable amount of) screenshots/figures, summary tables, etc. Additionally, please prepare a **slide set** for a 10-15 minute presentation. Please concentrate on the main points there. Please use 5-10 slides. In case of a short demonstration of software please include some demo screenshots in the slides.

Note: The paper can be either in English or German language, and the presentation should be in English language.

IMPORTANT NOTE CONCERNING THE WRITING OF YOUR PAPER

Please stick to the general principles of writing papers - that is, you can of course use materials from external sources, as long as they are correctly cited. But please keep in mind, that you should not copy & paste entire passages. It is not the aim of a paper to glue foreign contents together. Citations should enrich or underline your own points, so basically you should use your own words and can back them up with some citations. Expressing with your words illustrates that you understood the topic.

Note: The paper can be either in English or German language, and the presentation should be in English language.