

Programme

RESAW Technical Meetup, December 16-17, 2015, in Aarhus, Denmark

Theme: Corpus creation in web archives

The event is organised by [NetLab, Aarhus University](#).

Day 0 - Tuesday, December 15

For attendees who will be travelling to Aarhus:

- Arrive in Aarhus C (Central Station)
- Walk to the Radisson Blu Scandinavia Hotel <URL: <https://www.radissonblu.com/en/hotel-aarhus>>, it is only a 500m walk
- Check in
- Optional: Experience the Christmas feel of Aarhus C <URL: <http://www.visitaarhus.dk/danmark/jul-i-aarhus-2015>>
- Sleep well and see you tomorrow

Day 1 - Wednesday, December 16

Theme: A web corpus - what is it?

Location: State & University Library (morning) and Aarhus University (afternoon)

9:30-10:00	Arrival at State & University Library. Stairs/Elevator to 3rd floor, Meeting room "Skibet" (Danish: the (mother)ship)
9:45	A warm welcome by Ulrich & Niels, NetLab
10:00-11:00	"The concept of a corpus. How do researchers perceive of a web corpus? Why do we need it?" Presentations by RESAW members Ulrich K. Have & Niels Brügger of NetLab
11:00-11:20	"Netarkivet: data & infrastructure" by Tue Larsen & Jonas Lindberg Frellesen, Netarkivet
11:20-11:40	"Access Tools: Wayback Machine and Netarchive Search" by Tue Larsen & Jonas Lindberg Frellesen, Netarkivet
11:40-12:00	"Building a SolrCloud cluster for search on top of the WARC files" by Thomas Egense, State & University Library
12:00-12:45	Lunch
12:45-13:00	500m walk to the Nygaard-295, the IT Park, Aarhus University, umbrella if needed
13:00-13:20	"Creating corpus collections for the CLARIN center at the University of Copenhagen" by Mitchell Seaton, the University of Copenhagen
13:20-13:40	"Beyond Search: Exploring Corpus Creation Support within the WebART project" by Hugo Huurdeman, University of Amsterdam
13:40-14:00	"ArchiveSpark, framework to facilitate access to Web archives and enabling easy data extraction" by Helge Holzmann, Leibniz University
14:00-14:30	Discussion

14:30-1500	Coffee break
	For Fun (and inspiration):
15:00-15:30	"Image Search" by Thomas Egense, State & University Library
15:30-16:00	"Mass-generated screenshots" by Colin Rosenthal, Netarkivet
16:00-16:30	Summary of today's theme. Intro of Day 2's theme
	Dinner:
18:00	We meet at the entrance to ARoS Art Museum
18:10-20:00	Dinner at ARoS Food Hall
20:00-22:00	Exhibitions
22:00-	Good night/Cheers (pubs nearby)

Day 2 - Thursday, December 17

Theme: A web corpus - how do we make it?

Location: Aarhus University

9:30	Arrival & welcome at Nygaard-295, the IT Park, Aarhus University
9:30-9:45	Coffee, tea and fruit
9:45-10:00	Intro to today's theme by Ulrich K. Have, NetLab
10:00-10:20	"An algorithm for extracting a web corpus" by Ulrich K. Have, NetLab
10:20-10:40	"The Corpus as a View" by Tøke Eskildsen, State & University Library
10:40-11:00	"How can we analyse a corpus as big data?" by Per Møldrup-Dalum, State & University Library
11:00-11:10	Coffee Break
11:10-11:20	"Preserving the corpus (Data Management)" by Ulrich K. Have, NetLab
11:20-11:40	"WebCorpus, a prototype application to identify and extract corpora out of a web archive" by Sara Aubrey, Bibliothèque de France
11:40-12:00	Summary and discussion
12:00-13:00	Lunch & relax with a game of pool/table football
13:00-14:00	Tech discussion at the whiteboard
14:00-14:30	Researcher Q&A w. Niels Brügger, NetLab
14:30-1500	Coffee Break
15:00-16:00	Tech discussion at the whiteboard
16:00-16:30	Summary. Action points on the whiteboard.
16:30	Good bye (see you next time)

RESAW — A Research infrastructure for the Study of Archived Web materials

www.resaw.eu

Version: 4

Date: December 10, 2015