Title:
Nordic Digital Humanities Laboratory|NDHL - A Virtual Laboratory for Compute- and Data-intensive Humanities Research

Partners:
Kristoffer L. Nielbo (Center for Humanities Computing Aarhus, Aarhus University, Denmark), *Eetu Mäkelä* (Helsinki Centre for Digital Humanities, Helsinki University, Finland), *Lars Johnsen* (National Library of Norway, Norway), *Lars Borin* (SWE-Clarin, The Swedish Language Bank, Sweden), *Katrine H. Gasser* and *Per Møldrup-Dalum* (DeIC Cultural Heritage Cluster, Royal Danish Library, Denmark), *Birte Christensen-Dalsgaard* (DigHumLab, DARIAH, Denmark), *Aleksi Kallio* (CSC, Finland), *Nina Tahmasebi* (Centre for Digital Humanities, University of Gothenburg, Sweden), *Mats Malm* (University of Gothenburg, Sweden).

Introduction

A recent expert survey on digital competencies for humanities and arts in the Nordic countries supported by NeIC[1], showed a strong interest in shared compute and data infrastructure in order to facilitate code and data sharing (henceforth e-Infrastructure) for highly heterogeneous and unstructured cultural heritage data (i.e., text, images, and audio accessed through national libraries). This interest however was particular to researchers who already had rich code repositories, technical competencies and were well-established in Digital Humanities (DH) (henceforth *expert users*). Because the humanities and arts have been more challenged by the "digital revolution" than other research areas (see below), development of e-Infrastructure has favored domain experts in the core humanities and arts areas (e.g., cultural studies, history, literature and languages) that have very limited knowledge of automation, software-driven innovation, and e-Infrastructure. This strategy has resulted in several interesting research projects, but it has had the unfortunate consequence of neglecting and siloing expert users, because they have had to rely on local and, often, divergent e-Infrastructure. Our vision therefore is to *create convergence in Nordic humanities and arts e-Infrastructures for expert users through a Nordic Digital Humanities Laboratory (NDHL), a participant-driven virtual laboratory for compute- and data-intensive research*[2]. NDHL's goals are to 1) create new ways to enable compute- and data-intensive research by implementing a common data, software and service stack at royal libraries and HPC centres across the Nordics; 2) ensure joint access to restricted and copyrighted cultural heritage data; and 3) develop a sandbox environment that enables safe explorations of cultural heritage collections (restricted or otherwise) for research prototyping, piloting, and competency development. To initiate this ambitious infrastructure, we propose an NDHL community-forming pre-study under NeIC in order to develop prototypes, map synergies, identify stakeholder, and build an inclusive community.

The rapid growth of digital data and the development of computational technologies (the above mentioned digital revolution) are transforming knowledge discovery and understanding in every domain of human inquiry. Humanities and arts however are challenged by this transformation threefold due to the i) domain-specific requirements to and restrictions on data; ii) heterogeneity, historical origin and lack of documentation of data; and iii) lack of e-Infrastructure for developing and supporting exploratory compute- and data-intensive research. The NDHL partners consist of expert users in the humanities that have extensive experience with navigating the problem space of i-iii. They come together in order to connect humanities and arts research that rely on compute- and data-intensive applications in a stronger research community where access to and sharing of data and compute resources are made faster and more efficient through a Nordic collaboration. It is important to emphasize that NDHL is fundamentally about facilitating research collaboration and that its utilization will not be exclusive to its partners, on the contrary. When the virtual laboratory is developed, its codebase, documentation, and data will be available to all researchers in the Nordics following OS (Open Source) and FAIR (Findable, Assessible, Interoperable, Re-usable) principles. The development and management however is exclusive to the partners in order to ensure an efficient and high quality project life cycle that minimizes risk and maximizes research benefit (i.e., NDHL's goals).

Several infrastructures which support humanities already exist at European, national and local levels, but they do not directly target compute- and data-intensive humanities and arts research. To ensure maximal knowl-

[1] Final report was submitted to NeIC by emphBirte Christensen-Dalsgaard (DigHumLab, Denmark), bcd@cc.au.dk.

[2] Compute- and data-intensive research designates any kind of research that use CPU/GPU- and/or data-driven high performance computing as an essential part of their research.

edge transfer, the relevant infrastructures are either directly included as partners and stakeholders (see Partners), or indirectly through joined activities (seeActivities). The main inspiration for NDHL comes from our computer science colleagues' NeIC project Nordic Language Processing Laboratory (NLPL). While great care will be given to utilize NLPL's "knowledge base" and resources, NDHL deviates considerably due to the diversity of and requirements for cultural heritage data. On the European level CLARIN develops language technology and DARIAH focuses more broadly on digital humanities communities. To minimize reduplication, relevant resources from both infrastructures will be included in NDHL and national representatives are included as partners. National infrastructure providers are either represented among the initial partners (e.g., DigHumLab, DeIC's Cultural Heritage Cluster and CSC) or targeted as stakeholders during the pre-study (SNIC, UNINETT Sigma2 and RH Net). Finally, several of the NDHL partners represent existing laboratories and centers (e.g., Center for Humanities Computing Aarhus, Helsinki Centre for Digital Humanities, Centre for Digital Humanities Gothenburg) who are responsible for the majority of compute- and data-intensive research in the humanities.

NDHL will advance research in humanities and arts significantly by creating convergence between leading research groups in order to enable and empower internationally competitive compute- and data-intensive research across the Nordics through development and sharing of participant-driven e-Infrastructure.

1 Activities

The NDHL pre-study activities have three components: 1) pre-study collaboratory infrastructure; 2) NDHL hack days; and 3) Participant-driven workshops. Combined, the components ensure effective planning and scoping while facilitating participant involvement in an open science environment that promote community trust and collaboration.

1.1 Collaboratory Infrastructure

Communication, project planning and source code management is during the pre-study managed with Gitter and the GitLab instance for Nordic research software under NeIC and CodeRefinery. Center for Humanities Computing Aarhus (CHCAA) that is already contributing to CodeRefinery will be responsible for NDHL collaboratory infrastructure.

1.2 NDHL Hack Days

Hack days are spontaneously organized events where one or more partners commit to developing a prototype for specific parts of the NDHL's software stack and data repositories. Hack days only require one partner and will typically be budgeted for and run by the partner. Hack days and sprints are already common research and development practice among the partners and NDHL hack days will therefore piggyback on existing infrastructure. The purpose of hack days are to de-risk NDHL during the pre-study by rapid prototyping of risky parts on relatively small and accessible data samples.

1.3 Particpant-Driven Workshops

1.3.1 W1: Planning a shared stack and data depositories

The initial workshop brings together partners in Denmark to define, design, and prioritize the requirements to NDHL's software stack and data repositories. The workshop purpose is to set specific goals, sub-goals and anti-goals for the execution phase beyond the pre-study, separate must-haves from nice-to-haves, and prepare for prototyping in hack days in order to de-risk NDHL. The workshop will be held in Helsinki and HELDIG and CHCAA will provide on-site facilitators and developers for the workshop.

1.3.2 W2: *Nanos gigantum humeris insidentes*

Acknowledging that NDHL will be *standing on the shoulders of infrastructure giants*, we organize en e-Infrastructure workshop with national infrastructure providers and stakeholders in order to increase knowledge sharing and community-building and draw on previous discoveries. The purpose of the workshop is to modify goals and sub-goals in the NDHL project plan to avoid rewriting systems from scratch and present NDHL to stakeholders and potential partners. The workshop will be hosted at Gothenburg University during February 2020 and is a two-day event with Data infrastructure as the topic of the first day and Code and Compute Infrastructures as the topic of the second day. Planned invites are representatives from (day 1) CLARIN and DARIAH, the Nordic Royal Libraries, and the Danish DigHumLab, and (day 2) national e-Infrastructure providers (CSC, DeIC, SNIC, UNINETT Sigma2 and RH Net), NLPL and CodeRefinery.

1.3.3 W3: Hilbert problems of digital humanities

During the pre-study the NDHL partners also want to target and involve leading researchers and growing talent from the humanities in general and more specifically digital humanities. NDHL will therefore organize a BarCamp session on the topic *Hilbert problems of digital humanities* at the 5th Digital Humanities in the Nordic Countries conference in Latvia March 2020. To ensure a productive outcome, each partner functions as facilitator and proposes an important unsolved (Hilbert) problem in humanities that can benefit from the use of NDHL. Participants are then to define how and in what way NDHL can contribute.

1.3.4 W4: Scoping NDHL

The last workshop of the NDHL pre-study is dedicated to finalizing the goals and scoping NDHL's execution in order to formulate and submit infrastructure applications for NDHL to NeIC and national infrastructure funding across the Nordics. The purpose of the workshop is to divide NDHL into smaller modular components that can be distributed across the partners and define measurable milestones and time buffers to get to the NDHL goals. Given the experience from NLPL, we will invite Stephan Oepen and Bjørn Lindi to take advantage of "historical data" from NLPL's project life cycle.

1.4 Activity and Coordination Budget

Activity	transport	accommodation	meeting	other	total
Infrastructure & Coordination	NA	NA	NA	40,000	40,000
NDHL Hack Days	NA	NA	NA	15,000	15,000
W1: Planning	20,000	10,000	5,000	NA	35,000
W2: Nanos gigantum	40,000	20,000	10,000	NA	70,000
W3: Hilbert problems	NA	NA	5,000	NA	5000
W4: Scoping	20,000	10,000	5,000	NA	35,000
Total	80,000	40,000	25,000	55,000	200,000

Table 1: Estimated activity cost in Norwegian Kroner (NOK) based on Carlson Wagonlit Travel.

2 Partners of the Collaboration

Kristoffer L. Nielbo (applicant and Danish coordinator) is associate professor of humanities computing and director of Center for Humanities Computing Aarhus University and has extensive experience with eScience for the humanities from DeIC's University of Southern Denmark's eScience Center, membership of CodeRefinery's SG and several commisions of trust in DeIC. *Katrine H. Gasser* and *Per Møldrup-Dalum* are development managers at DeIC's Cultural Heritage Cluster for Humanities and Social Sciences at the Royal Danish Library. *Birte Christensen-Dalsgaard* is director of the Danish infrastructure DigHumLab and national representative in DARIAH. *Eetu Mäkelä* (Finish coordinator) is Professor in Human Sciences-Computing Interaction at University of Helsinki and Helsinki Centre for Digital Humanities where he leads a research track that develops computational tools to support humanities and social science research based on complex unstructured and structured data. *Aleksi Kallio* is Development manager at CSC - IT Center for Science. *Lars Johnson* (Norwegian Coordinator) is researcher and developer at the National Library of Norway and has extensive experience from developing their research infrastructure for, among other things, Bokhylla. *Lars Borin* (Swedish coordinator) director of the Swedish Language Bank, National coordinator of SWE-CLARIN, and professor of NLP at Gothenburg University. *Nina Tahmasebi* is an NLP researcher at the University of Gothenburg where she works with text mining for the digital humanities at the Centre for Digital Humanities and the Swedish Language Bank. *Mats Malm* is professor of Comparative literature and head of the Swedish Literature Bank.

3 Additional Information

Center for Humanities Computing Aarhus at Aarhus University, see CHCAA, and Helsinki Centre for Digital Humanities at Helsinki University, see HELDIG, are the principal organizers behind the NDHL proposal. Both centers are leading in applications of eScience in the humanities and arts across the Nordics and show how high performance computing and compute and data-intensive approaches can go hand in hand with strong domain expertise in the humanities and arts. Both centers combine a strong background in computational and formal sciences with extensive experience in the humanities and arts, two necessary conditions for creating infrastructural convergence and promoting adoption of technology at the faculties of humanities and arts in the Nordic countries.