

Planning a shared stack and data repository

NeIC.NDHL Workshop 1

Kristoffer L Nielbo

`kln@cas.dk`

`knielbo.github.io`

Center for Humanities Computing|chcaa.io
Aarhus University, Denmark

Outline

Planning a shared
stack and data
repository

Kristoffer L Nielbo
kln@cas.dk
knielbo.github.io

Introduction

National State of the
Art

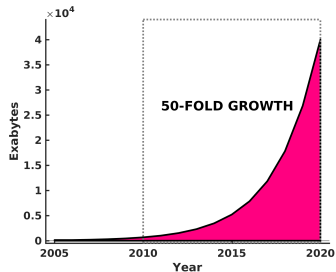
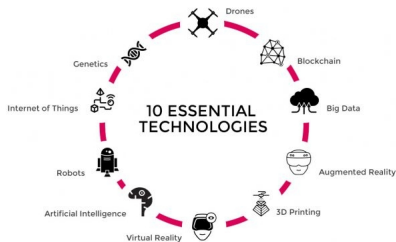
NDHL Wishlist

1 Introduction

2 National State of the Art

3 NDHL Wishlist

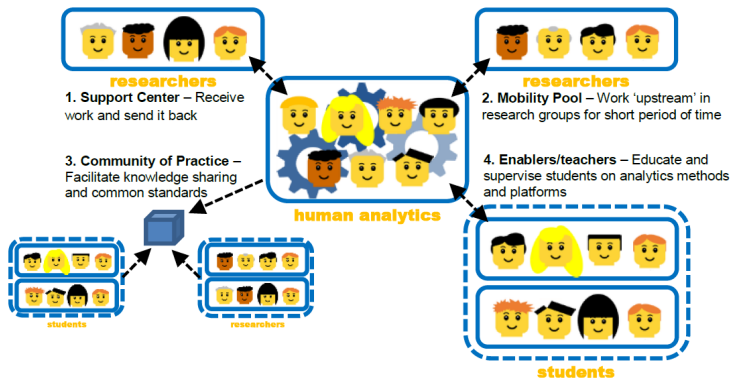




the data deluge is transforming knowledge discovery and understanding in every domain of human inquiry

a large part of these data are soft and unstructured \Rightarrow to get value from these data, humanities (and social sciences) must utilize automation

access to (high quality) data has become the biggest obstacle \Rightarrow NDHL will provide a shared software and application layer for modeling and analysis of cultural heritage* data across the Nordics.



Center for Humanities Computing Aarhus

Collaborators

- DelC, Danish Infrastructure Cooperation (HPC & RDM)
- Royal Danish Library
- DigHumLab
- CLARIN.DK & DARIAH-DK

Motivation

- Massive* national investments in supercomputing and “interactive HPC” is seen as the solution for onboarding SSH.
- Researchers are interested in newspapers, literature, and social media *at scale*

Challenges

- Increase in research-related cases of copyright infringement
- Currently “data mining” is a no go for cultural heritage data, unless a project has an agreement with the owner (case-by-case)

Options

- Access within the walls of the Royal Danish Library on the Cultural Heritage Cluster
- Derived data can be shared (e.g., summary statistics, neural embedding)
- Several projects are working on local solutions

We need an efficient solution for running applications on cultural data



Software and application layer

Planning a shared
stack and data
repository

Kristoffer L Nielbo
kln@cas.dk
knielbo.github.io

Introduction

National State of the
Art

NDHL Wishlist

Tools For CHCAA Python is the *lingua franca* and we rely heavily on the NumPy-SciPy ecosystem combined with TensorFlow with a CUDA backend; Numba switching to Dask.

Hardware Currently 40 active projects, 1/4 larger projects rely on accelerated HPC. A larger project uses 500-1000 node*hours on GPU nodes (2*v100) or (thin) CPU nodes (2*Intel CPUs w. 12 cores and 64 GB RAM).



Restricted national data resources

Planning a shared
stack and data
repository

Kristoffer L Nielbo
kln@cas.dk
knielbo.github.io

Introduction

National State of the
Art

NDHL Wishlist

DATA SET	LOCATION
Newspapers	KB
Internet	KB
Radio	KB
Literature (contemporary)	KB
TV	KB
FB	DataLab
Event-based Twitter	DataLab
Literature (pre 1920)	ADL
CoREST data	DSL
SMK collection	SMK



Resource sharing in the Nordics

“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, and ensure joint access to restricted and copyrighted cultural heritage data”

- Pan-Nordic access to raw data for modeling and analysis
- ... and meta data for efficient identification, access and evaluation of data with a FAIR mindset
- Easy sharing of compute resources
- Shared software and application stack to avoid parallel development
- Continued collaboration around data intensive research in the (digital) humanities

Model 1: Shared Virtual Laboratory

Planning a shared
stack and data
repository

Kristoffer L Nielbo
kln@cas.dk
knielbo.github.io

Introduction

National State of the
Art

NDHL Wishlist

“Inspired by NLPL, develop a DMZ that enables safe explorations of cultural heritage collections (restricted or otherwise) for research prototyping, piloting, and competency development.”



Model 2: Container Sharing

Planning a shared
stack and data
repository

Kristoffer L. Nielbo
kln@cas.dk
knielbo.github.io

Introduction

National State of the
Art

NDHL Wishlist

“tit for tat container sharing that gives access to restricted data through a research ring across the Nordics.”



THANKS

kln@au.dk
knielbo.github.io
chcaa.io

slides: http://knielbo.github.io/files/kln_ndhl_w1.pdf

