Collections, Capability, and Research

Leveraging the collections of cultural heritage institutions using HPC

Love Börjeson, PhD

Head of KBLab, National Library of Sweden

love.borjeson@kb.se, https://kb-labb.github.io/

What is the case for cultural heritage institutions using HPC?

The transformer transformation

- Pre-transformer AI-models require supervised training and a lot of annotated data
- Transformer AI-models require unsupervised training and massive unannotated data
- Cultural heritage institutions, like national libraries and archives, have exactly that: high quality, massive humanistic data (text, sound, images and videos)

Sweden has legal deposit laws installed in 1661 for everything printed.

During the the twentieth century, the law was gradually extended to include all modalities (text, sound, images and videos) and all formats (physical and digital). As a result, the KB has vast and ever growing collections, closing in on 26 Petabyte of data.

KB has the the largest, broadest and deepest collection of humanistic data for the Swedish language.

The collections includes objects such as..:

- Books
- Commercial leaflets
- Pizza menus
- Computer games
- School photos
- Hand-written manuscripts
- TV and radio broadcasts
- Newspapers and magazines
- ...and etc.

What is the case for cultural heritage institutions using HPC?

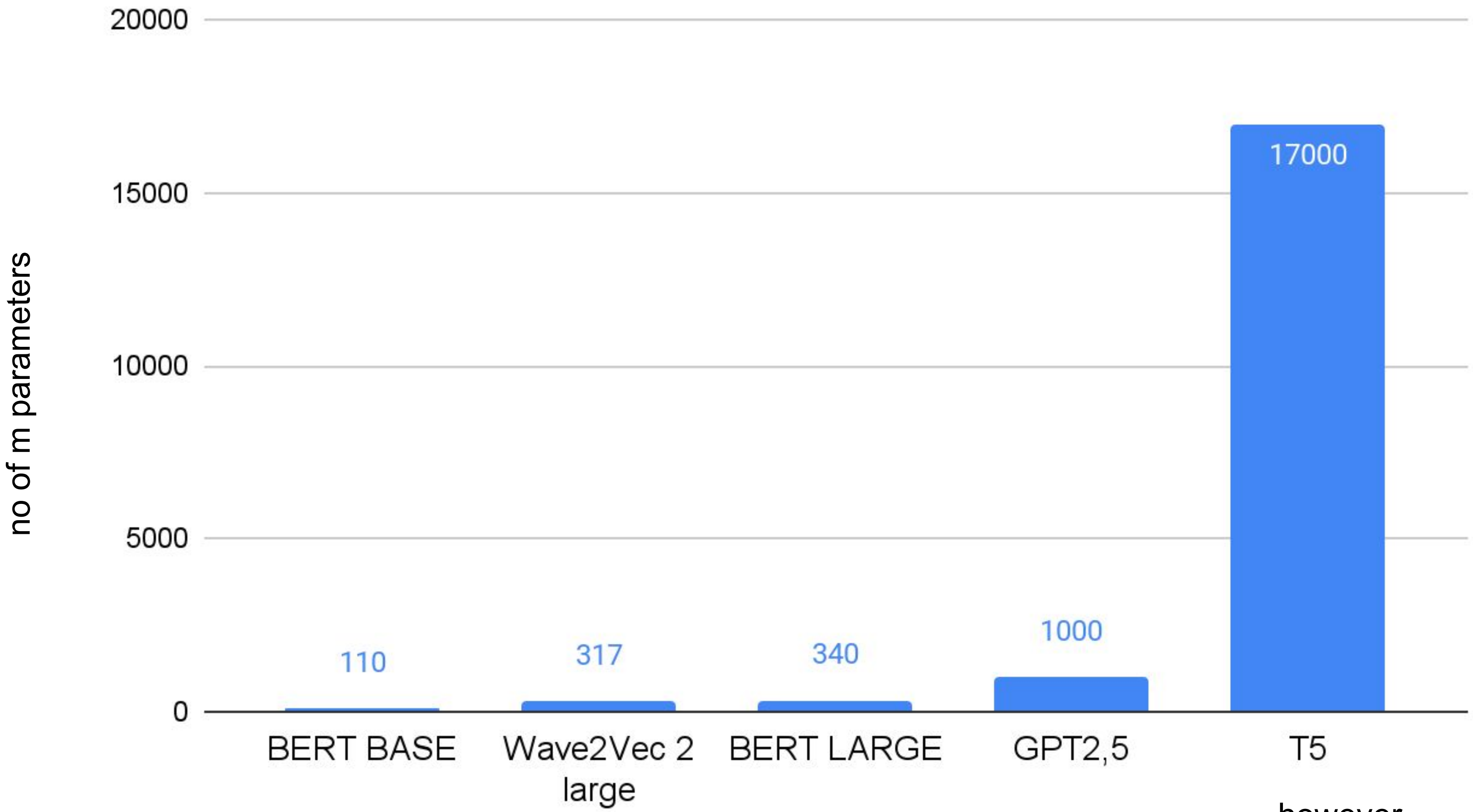
The paradigm of transferred learning

- Ground trained models have general capabilities (eg “understanding text”) that can be *transferred* via fine tuning and smaller annotated datasets to specific domains or capabilities (eg “understand Part-of-Speech”)
- Fine tuning is not data intensive and not computationally expensive - high usability downstream. Ground training however…

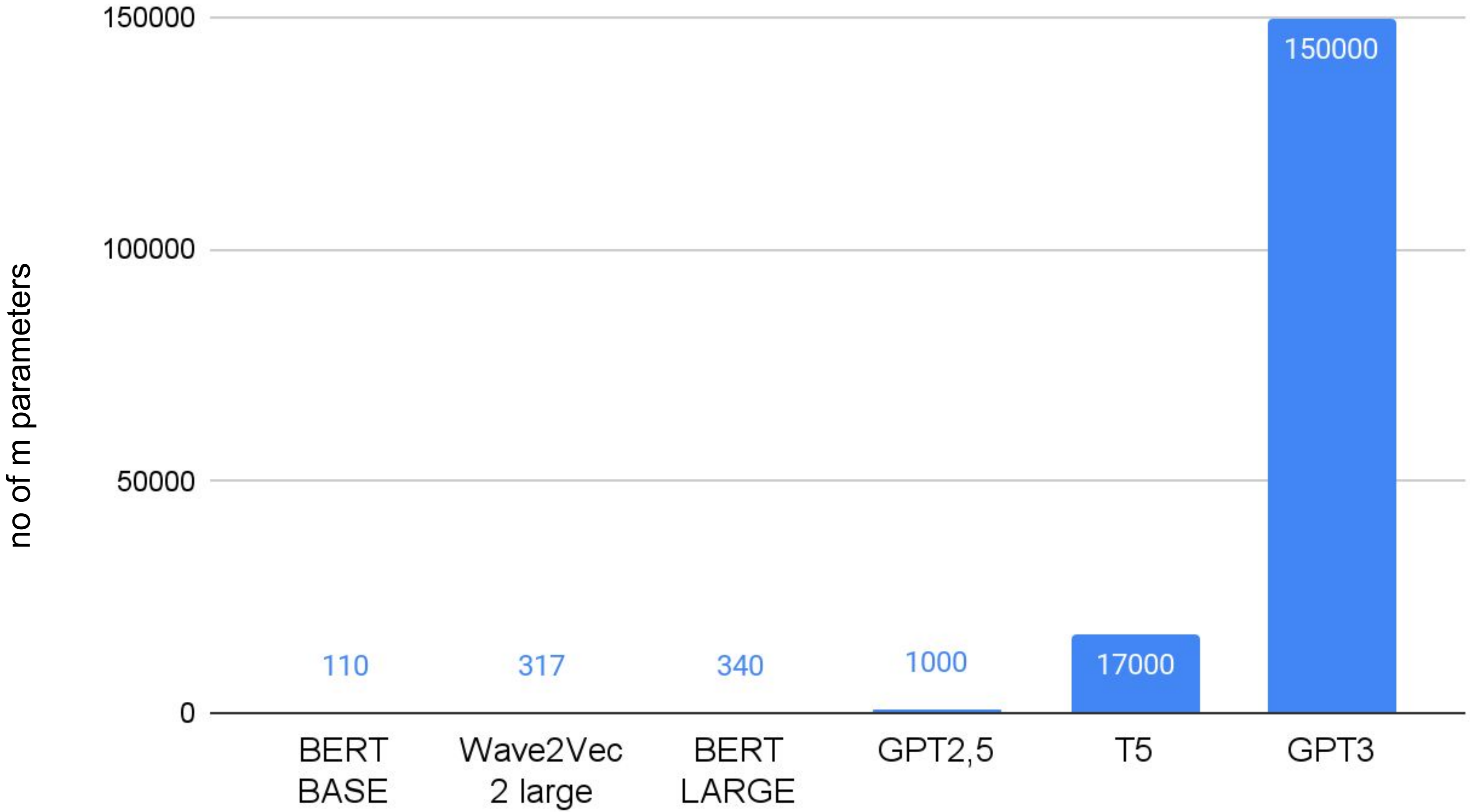
What is the case for cultural heritage institutions using HPC?

The never ending need for HPC

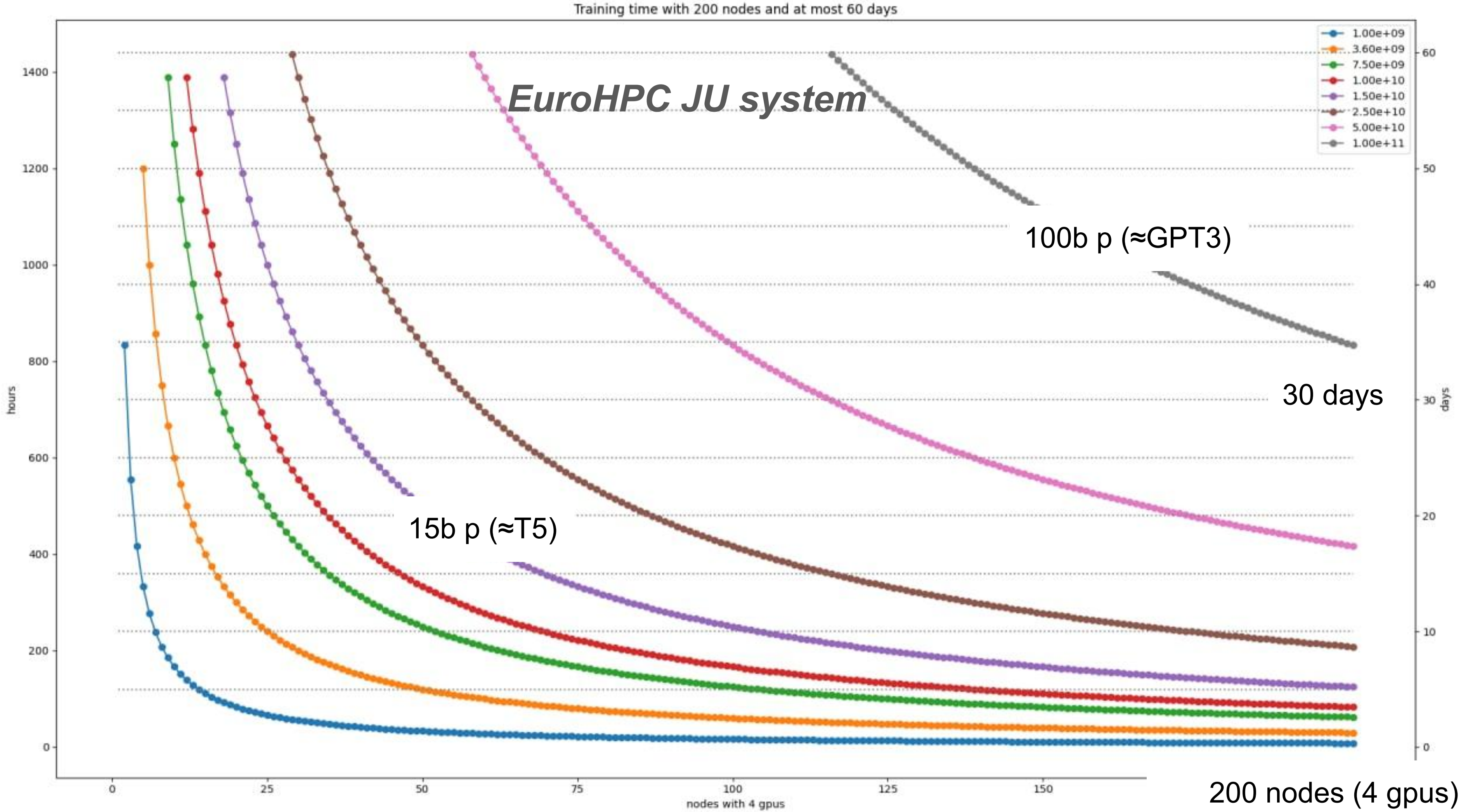
- For low- to mid-resource languages (eg Swedish), cultural heritage institutions have *the possibility* to train SOTA AI-models based on their collections of data on par with models for high-resource languages (eg English)
- The possibility to train SOTA models is also *an obligation* to train SOTA models, to support quality of Swedish research and ultimately the development of competitive Swedish-based AI-capabilities
- Balancing of data is democratically informed - KBLab has enough data to throw stuff away that would otherwise result in unbalanced and overly biased models
- However, to accomplish this and to make use of the collections, we need access to HPC-level computational resources, since ground training of transformer models is both data intensive and computationally expensive
- And models keep getting bigger…


no of m parameters
20000
15000
10000
5000
0
110
317
340
1000
17000
BERT BASE
Wave2Vec 2 large
BERT LARGE
GPT2,5
T5

however…



…and GPT3 is certainly not the end of it



EuroHPC JU system

100b p (≈GPT3)

15b p (≈T5)

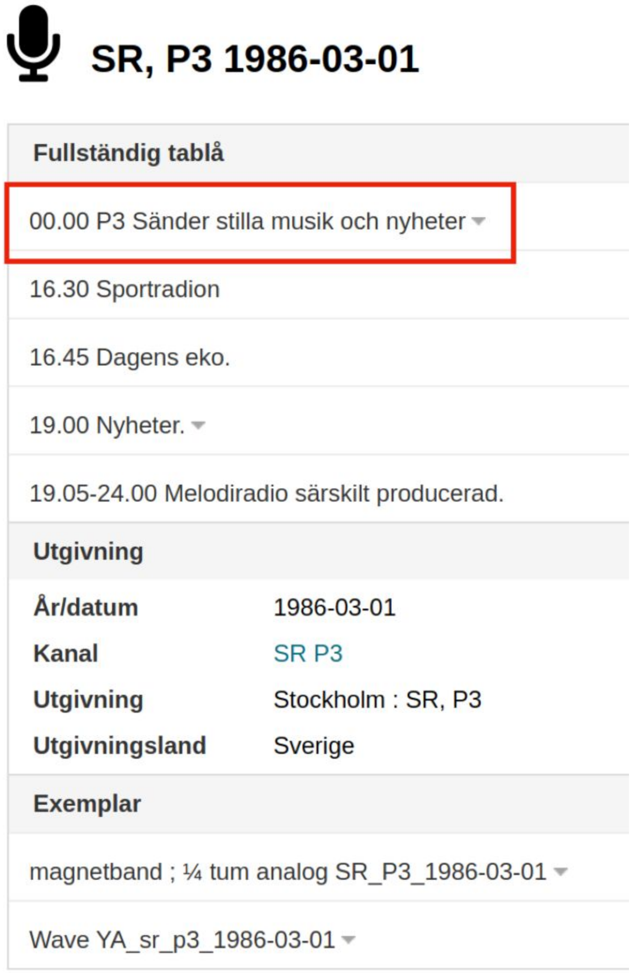
30 days

200 nodes (4 gpus)

… a user case

KBLab was the first European governmental agency to use the EuroHPC JU systems, specifically the Petascale podd VEGA in Slovenia (shoutout to the VEGA staff: we love you!).

How do we use our HPC-resources?



SR, P3 1986-03-01

Fullständig tablå

00.00 P3 Sänder stilla musik och nyheter

16.30 Sportradion

16.45 Dagens eko.

19.00 Nyheter.

19.05-24.00 Melodiradio särskilt producerad.

Utgivning

År/datum	1986-03-01
Kanal	SR P3
Utgivning	Stockholm : SR, P3
Utgivningsland	Sverige

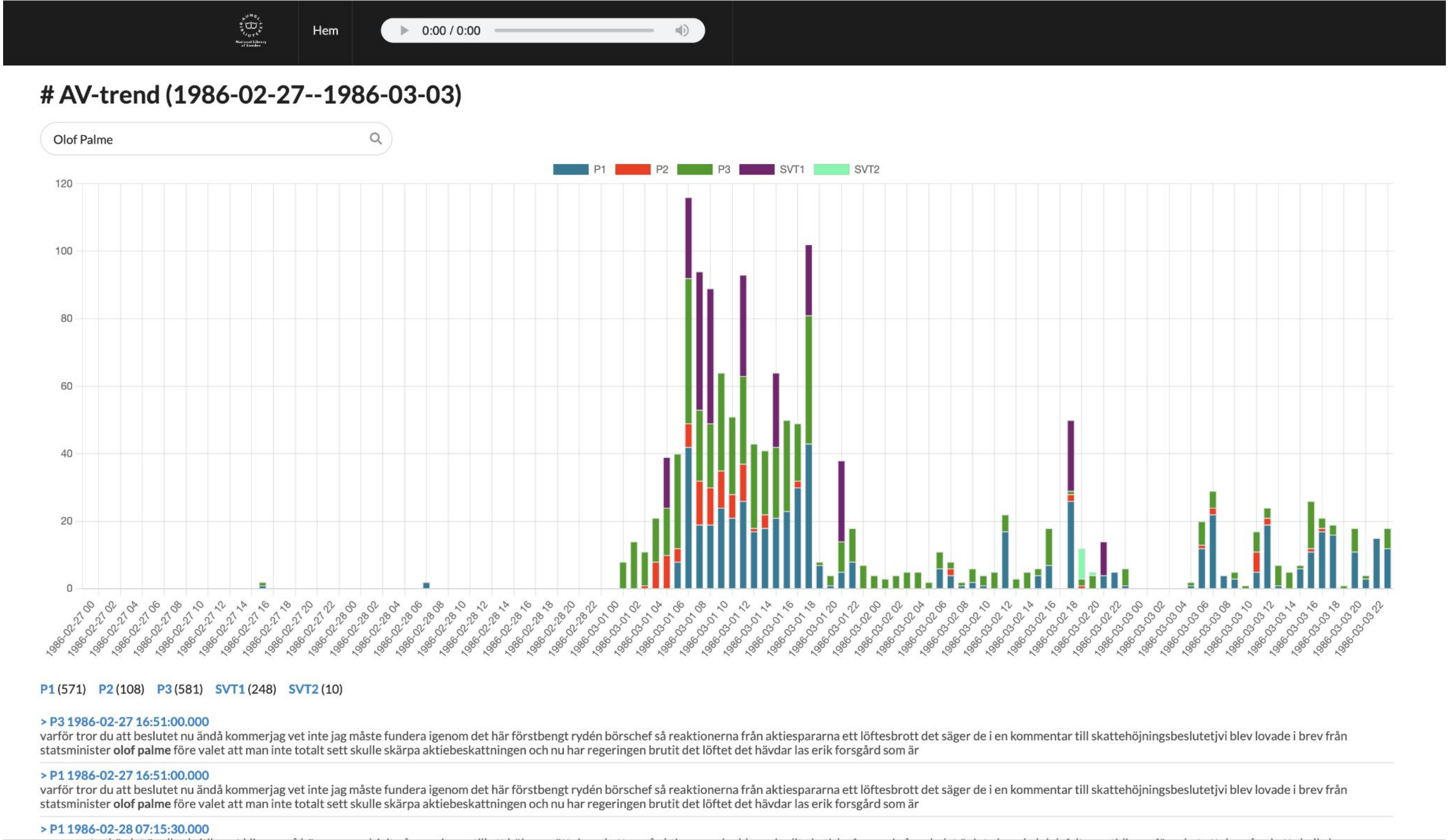
Exemplar

magnetband ; ¼ tum analog SR_P3_1986-03-01

Wave YA_sr_p3_1986-03-01

Information we have for the night of March 3rd, 1986:
Program 3 is broadcasting calm music and news.

But what happened that night?



AV-trend (1986-02-27--1986-03-03)

Olof Palme

P1 (571) P2 (108) P3 (581) SVT1 (248) SVT2 (10)

> P3 1986-02-27 16:51:00.000

varför tror du att beslutet nu ändå kommerjag vet inte jag måste fundera igenom det här förstbengt rydén börschef så reaktionerna från aktiespararna ett löftesbrott det säger de i en kommentar till skattehöjningsbeslutetjvi blev lovade i brev från statsminister **olof palme** före valet att man inte totalt sett skulle skärpa aktiebeskattningen och nu har regeringen brutit det löftet det hävdar las erik forsgård som är

> P1 1986-02-27 16:51:00.000

varför tror du att beslutet nu ändå kommerjag vet inte jag måste fundera igenom det här förstbengt rydén börschef så reaktionerna från aktiespararna ett löftesbrott det säger de i en kommentar till skattehöjningsbeslutetjvi blev lovade i brev från statsminister **olof palme** före valet att man inte totalt sett skulle skärpa aktiebeskattningen och nu har regeringen brutit det löftet det hävdar las erik forsgård som är

> P1 1986-02-28 07:15:30.000

The awakening - bringing back life to the dead collections

Using the KBLab’s sound model “VoxRex”, we can describe the content of the radio broadcasts. Digital objects that has been “dead” for practical purposes can thus be brought back to life.

The Vox Rex is SOTA because KBLab has access to almost unlimited amount of hours of sound with a complete representation of variations of spoken Swedish AND because we can use Euro HPC resources to train the models.

This is all but *one* application at the Nat. library. VoxRex and fine tuned versions of the model is however downloaded around 200 000 times per month for purposes/applications we do not really see.

Societal benefits from KBLab’s model is huge and largely unknown.