Hva er språkdata og hvordan kan vi få tak i dem?

Magnus Breder Birkenes
Språkbanken, Nasjonalbiblioteket

Nasjonalbiblioteket

Hva er språkdata?

- “Alt er språk”, “Alt er tekst”
- Tekst
  - Hele tekstdokumenter, korpus (annotert, ikke-annotert) eller reduserte representasjoner av disse
  - Oversettelser (parallelltekst)
  - Ordbøker
  - Termlister
- Tale
  - (Store mengder) lydopptak av talt språk
  - Transkripsjoner av den samme talen: gjengivelse av lydene i skriftlig form
  - Uttaleleksikon
- Representativitet viktig (og problematisk)

Utvalgte bruksområder

- tekstdata:
  - automatisk oversettelse mellom språk og målformer
  - skrivestøtte: stavekontroll, grammatikkontroll, stilistiske tips
  - samtaleroboter
  - data mining: hente ut trekk i tekster
    - emneklassifisering/modellering
    - sentimentanalyse
- taledata:
  - talegjenkjenning
  - talesyntese

Hvor finner jeg språkdata for norsk?

- Språkbankens ressurskatalog: https://www.nb.no/sprakbanken/ressurskatalog/
- Inneholder:
  - åpne språkdata fra Språkbanken ved Nasjonalbiblioteket (fortrinnsvis med svært liberale bruksbetingelser, CC-0)
  - data fra andre norske institusjoner som er medlem i CLARIN-samarbeidet (bruksbetingelser kan være strengere)

To eksempler

- Doffin: Parallellstilling og generering av oversettelsesminner
- Målfrid-korpuset: Massenedlasting og -prosessering av tekst fra statlige domener

Doffin

Doffin

- Doffin = nasjonal kunngjøringsdatabase for offentlige anskaffelser (hos DFØ): https://www.doffin.no/
- Oversettelser til engelsk i TED Tenders Electronic Daily (TED): https://ted.europa.eu
- XML-dump av databasene
- Originaler og oversettelse forbundet gjennom dokument-identifikator

Direktoratet for forvaltning og økonomistyring

Bokmål | Nynorsk | English |

DOFFIN

Database for offentlige innkjøp

Om Doffin | Kontakt | Logg inn

Søk kunngjøring med filter | Søk kjøperprofil | Oppdragsgivere | Leverandører | Brukerstøtte

Vis dokumenter | Se Ted kunngjøring

Kunngjøring

Anskaffelsesforskriften

Alminnelig kunngjøring av konkurranse

EØS-kunngjøring

Del I: Oppdragsgiver

I.1) Navn og adresser

Nasjonalbiblioteket
976029100
Henrik Ibsensgate 110
OSLO
0255
NO
Kontaktperson: Cyrille Nolin
E-post: cyrille.nolin@nb.no
NUTS-kode: NO - NORGE
Sted: hele Norge
Internettadresse(r):
Nettsted oppdragsgiver:
https://permalink.mercell.com/74690480.aspx
Nettsted Kjøperprofil: http://www.nb.no/

I.3) Kommunikasjon

Konkurransegrunnlaget er elektronisk tilgjengelig med gratis, direkte og ubegrenset tilgang på:
https://permalink.mercell.com/74690480.aspx
Tilleggsinformasjon finnes på en annen adresse:
Mercell Norge AS
Karihaugveien 89
Oslo
1086

Ted

English (en)

Search TED for public tenders

Advanced search / Expert search

EUROPA > TED home > Display TED notice in current language

TED | TED SIMAP | TED eNotices | TED eTendering

041
2021

OJ S current issue

Release calendar

Check out our COVID-19 dedicated page for tenders related to medical equipment needs.

Brexit

Supplies - 401940-2017

Original language | Data | Document family

Share

11/10/2017 S195

I. II. III. IV. VI.

Norway-Oslo: Cinematographic film

2017/S 195-401940

Contract notice

Supplies

Legal Basis:
Directive 2014/24/EU

Section I: Contracting authority

I.1) **Name and addresses**

2017-677443_orginal.xml

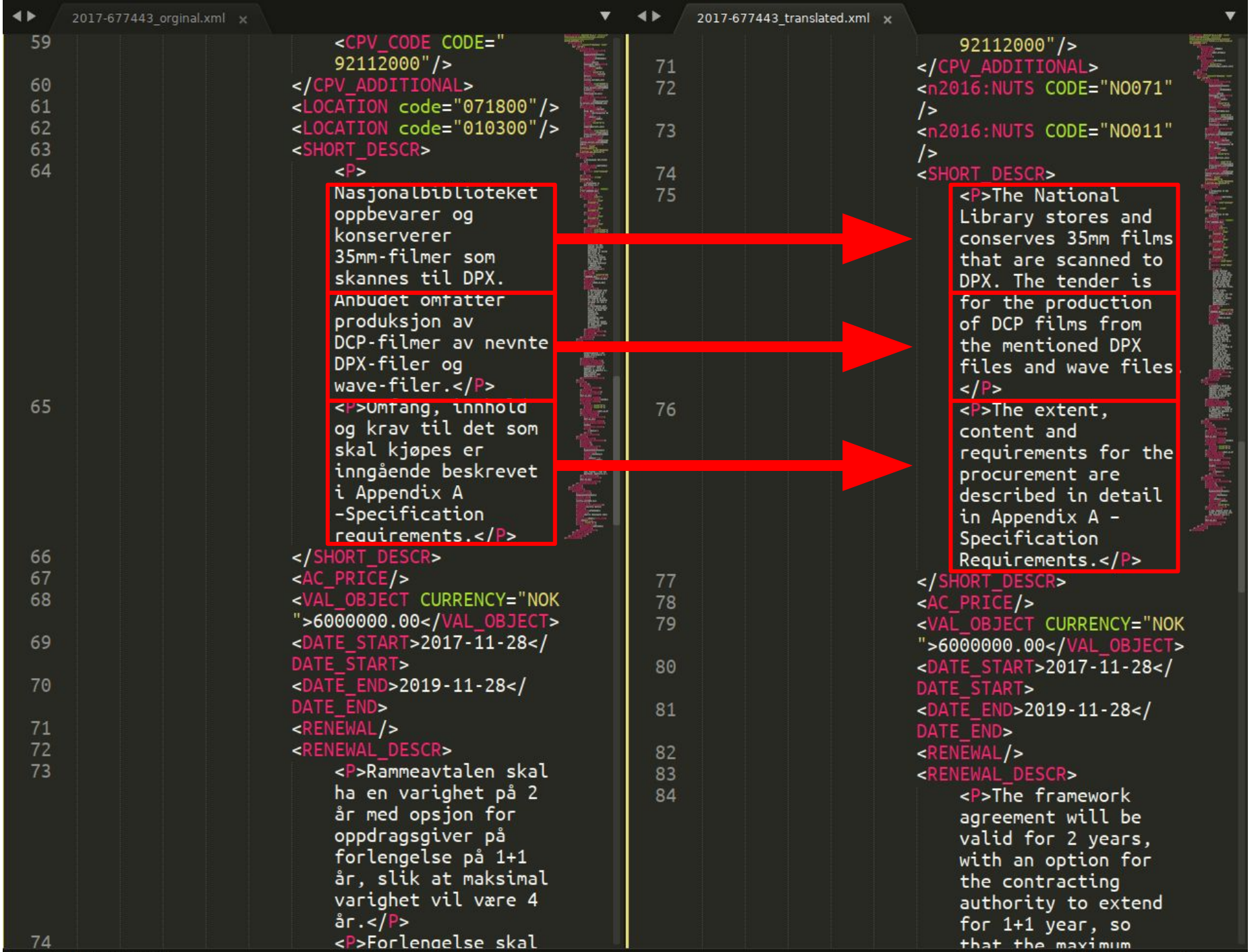
```
59          <CPV_CODE CODE="
            92112000"/>
60      </CPV_ADDITIONAL>
61      <LOCATION code="071800"/>
62      <LOCATION code="010300"/>
63      <SHORT_DESCR>
64          <P>
            Nasjonalbiblioteket
            oppbevarer og
            konserverer
            35mm-filmer som
            skannes til DPX.
            Anbudet omfatter
            produksjon av
            DCP-filmer av nevnte
            DPX-filer og
            wave-filer.</P>
65          <P>Omfang, innhold
            og krav til det som
            skal kjøpes er
            inngående beskrevet
            i Appendix A
            -Specification
            requirements.</P>
66      </SHORT_DESCR>
67      <AC_PRICE/>
68      <VAL_OBJECT CURRENCY="NOK
        ">6000000.00</VAL_OBJECT>
69      <DATE_START>2017-11-28</
        DATE_START>
70      <DATE_END>2019-11-28</
        DATE_END>
71      <RENEWAL/>
72      <RENEWAL_DESCR>
73          <P>Rammeavtalen skal
            ha en varighet på 2
            år med opsjon for
            oppdragsgiver på
            forlengelse på 1+1
            år, slik at maksimal
            varighet vil være 4
            år.</P>
74          <P>Forlengelse skal
```

2017-677443_translated.xml

```
            92112000"/>
71      </CPV_ADDITIONAL>
72      <n2016:NUTS CODE="NO071"
        />
73      <n2016:NUTS CODE="NO011"
        />
74      <SHORT_DESCR>
75          <P>The National
            Library stores and
            conserves 35mm films
            that are scanned to
            DPX. The tender is
            for the production
            of DCP films from
            the mentioned DPX
            files and wave files.
            </P>
76          <P>The extent,
            content and
            requirements for the
            procurement are
            described in detail
            in Appendix A -
            Specification
            Requirements.</P>
77      </SHORT_DESCR>
78      <AC_PRICE/>
79      <VAL_OBJECT CURRENCY="NOK
        ">6000000.00</VAL_OBJECT>
80      <DATE_START>2017-11-28</
        DATE_START>
81      <DATE_END>2019-11-28</
        DATE_END>
82      <RENEWAL/>
83      <RENEWAL_DESCR>
84          <P>The framework
            agreement will be
            valid for 2 years,
            with an option for
            the contracting
            authority to extend
            for 1+1 year, so
            that the maximum
```

Setningstokenisering

- Oppdeling av tekst i setninger
- Bruker ulike heuristikker:
  - Skilletegn: . ! ?
  - Stor/liten forbokstav
  - Forkortelser: *f.eks.*
  - Typiske ord setninger begynner med: *Jeg, det*
  - Kollokasjoner: *Johann S. Bach*
- Her: “Punkt Tokenizer” med norsk modell
- Trent på et større materiale av “gode setninger”


2017-677443_orginal.xml
2017-677443_translated.xml
<CPV_CODE CODE="
92112000"/>
</CPV_ADDITIONAL>
<LOCATION code="071800"/>
<LOCATION code="010300"/>
<SHORT_DESCR>
<P>
Nasjonalbiblioteket oppbevarer og konserverer 35mm-filmer som skannes til DPX.
Anbudet omfatter produksjon av DCP-filmer av nevnte DPX-filer og wave-filer.</P>
<P>Omfang, innhold og krav til det som skal kjøpes er inngående beskrevet i Appendix A -Specification requirements.</P>
</SHORT_DESCR>
<AC_PRICE/>
<VAL_OBJECT CURRENCY="NOK
">6000000.00</VAL_OBJECT>
<DATE_START>2017-11-28</
DATE_START>
<DATE_END>2019-11-28</
DATE_END>
<RENEWAL/>
<RENEWAL_DESCR>
<P>Rammeavtalen skal ha en varighet på 2 år med opsjon for oppdragsgiver på forlengelse på 1+1 år, slik at maksimal varighet vil være 4 år.</P>
<P>Forlengelse skal
92112000"/>
</CPV_ADDITIONAL>
<n2016:NUTS CODE="NO071"
/>
<n2016:NUTS CODE="NO011"
/>
<SHORT_DESCR>
<P>The National Library stores and conserves 35mm films that are scanned to DPX. The tender is for the production of DCP films from the mentioned DPX files and wave files.</P>
<P>The extent, content and requirements for the procurement are described in detail in Appendix A - Specification Requirements.</P>
</SHORT_DESCR>
<AC_PRICE/>
<VAL_OBJECT CURRENCY="NOK
">6000000.00</VAL_OBJECT>
<DATE_START>2017-11-28</
DATE_START>
<DATE_END>2019-11-28</
DATE_END>
<RENEWAL/>
<RENEWAL_DESCR>
<P>The framework agreement will be valid for 2 years, with an option for the contracting authority to extend for 1+1 year, so
that the maximum
ok, Line 5, Column 30
Tab Size: 4
XML

2017-677443_nob.txt

```
Produksjon DCP-filmer
Produksjon av DCP-filmer.
Nasjonalbiblioteket oppbevarer og konserver
Anbudet omfatter produksjon av DCP-filmer a
Omfang, innhold og krav til det som skal kj
Rammeavtalen skal ha en varighet på 2 år me
Forlengelse skal gis til leverandøren senes
Forlengelse skal meddeles til leverandøren
Leverandøren skal være registrert i et fore
Vilkårene angitt i Appendix 1 - Terms of Co
Oslo
Eventuelle klager skal fremmes i tråd med g
```

2017-677443_eng.txt

```
Production of DCP films
Production of DCP films.
The National Library stores and conserves 3
The tender is for the production of DCP fil
The extent, content and requirements for th
The framework agreement will be valid for 2
The tenderer shall be notified of the exter
The tenderer shall be notified of the exter
Tenderers shall be registered in a company
The terms mentioned in Appendix 1 - Terms o
Oslo
Any appeals shall be filed in accordance wi
```

doffin-nob.tmx

```
196597      <tuv xml:lang="nob">
196598        <seg>Produksjon av DCP-filmer.</seg>
196599      </tuv>
196600      <tuv xml:lang="eng">
196601        <seg>Production of DCP films.</seg>
196602      </tuv>
196603    </tu>
196604    <tu>
196605      <tuv xml:lang="nob">
196606        <seg>Nasjonalbiblioteket oppbevarer og konserverer 35mm-filmer som skannes til
              DPX.</seg>
196607      </tuv>
196608      <tuv xml:lang="eng">
196609        <seg>The National Library stores and conserves 35mm films that are scanned to
              DPX.</seg>
196610      </tuv>
196611    </tu>
196612    <tu>
196613      <tuv xml:lang="nob">
196614        <seg>Anbudet omfatter produksjon av DCP-filmer av nevnte DPX-filer og
              wave-filer.</seg>
196615      </tuv>
196616      <tuv xml:lang="eng">
196617        <seg>The tender is for the production of DCP films from the mentioned DPX
              files and wave files.</seg>
196618      </tuv>
196619    </tu>
196620    <tu>
196621      <tuv xml:lang="nob">
196622        <seg>Omfang, innhold og krav til det som skal kj&#248;pes er inng&#229;ende
              beskrevet i Appendix A &#150;Specification requirements.</seg>
196623      </tuv>
196624      <tuv xml:lang="eng">
196625        <seg>The extent, content and requirements for the procurement are described in
              detail in Appendix A &#150; Specification Requirements.</seg>
```

Aligning

- en forholdsvis enkel jobb i dette tilfellet
  - alignert på dokumentnivå allerede
  - avsnitt i lik rekkefølge
- noen steder tilsvarer én setning på norsk flere setninger på engelsk og omvendt
- tilordning gjennom verktøyet HunAlign

```
<tu>
<tuv xml:lang="nob">
   <seg>Utstillingen "Ville vakre Vega" vil være
   hovedutstillingen ved verdensarvsenteret som
   skal formidle og fortolke områdets verdier som
   har gitt Vegaøyan en plass på Unescos
   verdensarvliste.</seg>
 </tuv>
 <tuv xml:lang="eng">
   <seg>The exhibit "Wild and beautiful Vega" shall
   be the main exhibit at the World Heritage
   Centre. ~~~ The exhibit shall communicate and
   interpret the area's values which have led to a
   place on the UNESCO's World Heritage list.</seg>
 </tuv>
</tu>
```

Resultat

- 40.000 dokumentpar (original og oversettelse)
- 300.000 oversettelsespar (translation units) etter deduplisering
- datasettet er fritt tilgjengelig under: https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-63/

Målfrid-korpuset

Målfrid

- samarbeidsprosjekt mellom Språkrådet og Nasjonalbiblioteket på oppdrag fra Kulturdepartementet
- mål: automatisere målformsrapportering (kravet om minst 25% nynorsk/bokmål i statlig informasjonsmateriale)
- middel: høsting av statlige nettsider, prosessering og automatisk språkdeteksjon av materiale

Høsting

- verktøy: *wget*
- rekursiv høsting ned til nivå 12 (det dypeste nivået vi fant under en prøveinnhøsting)
- kun tekstdokumenter (HTML, PDF, DOC)
- alle statlige domener med visse unntak
- alle undersider med unntak (f.eks. ikke metadataregistre)

Prosessering

- HTML: “Boilerplate removal”, Justext
  - navigasjonselementer og repeterende elementer, lenkelister fjernes
  - heuristikker. semantikken i HTML-elementer, stoppordtetthet, lenketetthet osv.
- PDF: full OCR med Google Vision API
- WORD/ODT: Python-biblioteket Textract
- alle dokumenter dedupliseres på domenenivå (sjekksum av redusert representasjon)

- Forretningsplan
- Om Bane NOR
- Nasjonal transportplan
- Organisasjon og ledelse
- Eierskap og styring
- Investor Relations
- Jernbanereformen
- Miljø
- Etiske retningslinjer
- Varslingskanal
- Bane NORs historie
- Arkiv Jernbaneverket

Om Bane NOR

Skriv ut

Bane NOR er et statlig foretak med ansvar for den nasjonale jernbaneinfrastrukturen.

Bane NORs formål er å sørge for tilgjengelig jernbaneinfrastruktur og effektive og brukervennlige tjenester, inkludert knutepunkts- og godsterminalutvikling.

Bane NOR har ansvaret for planlegging, utbygging, forvaltning, drift og vedlikehold av det nasjonale jernbanenettet, trafikkstyring og forvaltning og utvikling av jernbaneeiendom. Bane NOR har det operative koordineringsansvaret for sikkerhetsarbeidet og operativt ansvar for samordning av beredskap og krisehåndtering.

Bane NOR har om lag 3 400 ansatte og har hovedkontor i Oslo.

Bane NOR SF er 100 prosent eid av staten og er underlagt Samferdselsdepartementet.

Last ned dokumenter

- Brosjyre Bane NOR - Vi skaper fremtidens jernbane.pdf
- Brochure Bane NOR - We create the railway of the future.pdf

Publisert av: Bane NOR 12.06.2018

Del saken

- Forretningsplan
- **Om Bane NOR**
- Nasjonal transportplan
- Organisasjon og ledelse
- Eierskap og styring
- Investor Relations
- Jernbanereformen
- Miljø
- Etiske retningslinjer
- Varslingskanal
- Bane NORs historie
- Arkiv Jernbaneverket

Om Bane NOR

Bane NOR er et statlig foretak med ansvar for den nasjonale jernbaneinfrastrukturen.

Bane NORs formål er å sørge for tilgjengelig jernbaneinfrastruktur og effektive og brukervennlige tjenester, inkludert knutepunkts- og godsterminalutvikling.

Bane NOR har ansvaret for planlegging, utbygging, forvaltning, drift og vedlikehold av det nasjonale jernbanenettet, trafikkstyring og forvaltning og utvikling av jernbaneeiendom. Bane NOR har det operative koordineringsansvaret for sikkerhetsarbeidet og operativt ansvar for samordning av beredskap og krisehåndtering.

Bane NOR har om lag 3 400 ansatte og har hovedkontor i Oslo.

Bane NOR SF er 100 prosent eid av staten og er underlagt Samferdselsdepartementet.

Skriv ut

Last ned dokumenter

- Brosjyre Bane NOR - Vi skaper fremtidens jernbane.pdf
- Brochure Bane NOR - We create the railway of the future.pdf

Publisert av: Bane NOR 12.06.2018

Del saken

Bane NOR SF

Postadresse
Postboks 4350
2308 Hamar

Hovedkontor (besøksadresse)
Schweigaards gate 33
0191 Oslo

Bane NOR kundesenter

Åpningstider

Mandag - fredag	08.00 - 19.00
Lørdag - søndag/helligdag	09.00 - 15.00

Chat med oss

Telefon

Om nettstedet

- Nettstedskart
- Cookies (informasjonskapsler) på Bane NOR
- Personvernerklæring

Ansvarlig redaktør
Konserndirektør Torild Lid Uribarri

- Forretningsplan
- **Om Bane NOR**
- Nasjonal transportplan
- Organisasjon og ledelse
- Eierskap og styring
- Investor Relations
- Jernbanereformen
- Miljø
- Etiske retningslinjer
- Varslingskanal
- Bane NORs historie
- Arkiv Jernbaneverket

Om Bane NOR

Bane NOR er et statlig foretak med ansvar for den nasjonale jernbaneinfrastrukturen.

Bane NORs formål er å sørge for tilgjengelig jernbaneinfrastruktur og effektive og brukervennlige tjenester, inkludert knutepunkts- og godsterminalutvikling.

Bane NOR har ansvaret for planlegging, utbygging, forvaltning, drift og vedlikehold av det nasjonale jernbanenettet, trafikkstyring og forvaltning og utvikling av jernbaneeiendom. Bane NOR har det operative koordineringsansvaret for sikkerhetsarbeidet og operativt ansvar for samordning av beredskap og krisehåndtering.

Bane NOR har om lag 3 400 ansatte og har hovedkontor i Oslo.

Bane NOR SF er 100 prosent eid av staten og er underlagt Samferdselsdepartementet.

Skriv ut

Last ned dokumenter

- Brosjyre Bane NOR - Vi skaper fremtidens jernbane.pdf
- Brochure Bane NOR - We create the railway of the future.pdf

Publisert av: Bane NOR 12.06.2018

Del saken

Bane NOR SF

Postadresse
Postboks 4350
2308 Hamar

Hovedkontor (besøksadresse)
Schweigaards gate 33
0191 Oslo

Bane NOR kundesenter

Åpningstider
Mandag - fredag 08.00 - 19.00
Lørdag - søndag/helligdag 09.00 - 15.00

Chat med oss

Telefon

Søk herSøk
Meny
Forside
Om oss
Om Bane NOR
Om Bane NOR
Bane NOR er et statlig foretak med ansvar for den nasjonale jernbaneinfrastrukturen.
Bane NORs formål er å sørge for tilgjengelig jernbaneinfrastruktur og effektive og brukervennlige tjenester, inkludert knutepunkts- og godsterminalutvikling.
Bane NOR har ansvaret for planlegging, utbygging, forvaltning, drift og vedlikehold av det nasjonale jernbanenettet, trafikkstyring og forvaltning og utvikling av jernbaneeiendom. Bane NOR har det operative koordineringsansvaret for sikkerhetsarbeidet og operativt ansvar for samordning av beredskap og krisehåndtering.
Bane NOR har om lag 3 400 ansatte og har hovedkontor i Oslo.
Bane NOR SF er 100 prosent eid av staten og er underlagt Samferdselsdepartementet.

final class	good
cotext-free class	good
heading	False
length (in characters)	157
number of characters within links	0
link density	0.000
number of words	18
number of stopwords	6
stopword density	0.333
html.body.main.div.div.article.div.p	

VEDTEKTER FOR STATENS LÅNEKASSE FOR UTDANNING

Fastsatt av Kunnskapsdepartementet den 15.januar 2016 i medhold av lov 3. juni 2005 nr. 37 om utdanningsstøtte.

1 Formål

Statens lånekasse for utdanning (Lånekassen) skal forvalte utdanningsstøtten i samsvar med bestemmelsene gitt i eller i medhold av utdanningsstøtteloven. Rammene for virksomheten til Lånekassen fastsettes for øvrig av overordnet myndighet.

Utfyllende bestemmelser om virksomhetens formål framgår av Hovedinstruks for Lånekassen.

2 Lånekassens ledelse og oppgaver

Lånekassen er et ordinært forvaltningsorgan som ledes av eget styre og tilsatt administrerende direktør.

Styret er Lånekassens øverste organ og er ansvarlig for den samlede virksomheten. Styret skal blant annet påse at Lånekassen drives i samsvar med regelverket, samt styringssignaler og retningslinjer gitt av overordnede myndigheter. Styret skal sikre at fastsatte mål oppnås og at ressursbruken er effektiv. Styret har også ansvar for at det foreligger vurderinger som identifiserer de viktigste risikofaktorene knyttet til virksomheten og at det foreligger planer for oppfølging av disse faktorene.


```
    <div class='ocr_carea' id='block_2' title='bbox 100 153 169 161'>
    <p class='ocr_par' dir='ltr' id='par_2_0' title='bbox 100 153 169 161'>
<span class='ocr_line' id='line_2_0_0' title='bbox 100 153 169 161; baseline 0 0'>
    <span class='ocrx_word' id='word_2_0_0' title='bbox 100 153 108 161'>1</span>
    <span class='ocrx_word' id='word_2_0_1' title='bbox 114 153 169 161'>Formål</span>
</span>
<span class='ocr_line' id='line_2_0_1' title='bbox 114 153 169 161; baseline 0 0'>
</span></p></div>
```

```
    <div class='ocr_carea' id='block_3' title='bbox 100 174 735 201'>
    <p class='ocr_par' dir='ltr' id='par_3_0' title='bbox 100 174 735 201'>
<span class='ocr_line' id='line_3_0_0' title='bbox 100 174 735 201; baseline 0 0'>
    <span class='ocrx_word' id='word_3_0_0' title='bbox 100 174 148 182'>Statens</span>
    <span class='ocrx_word' id='word_3_0_1' title='bbox 154 174 216 182'>lånekasse</span>
    <span class='ocrx_word' id='word_3_0_2' title='bbox 222 174 241 182'>for</span>
    <span class='ocrx_word' id='word_3_0_3' title='bbox 247 174 312 182'>utdanning</span>
    <span class='ocrx_word' id='word_3_0_4' title='bbox 318 174 407 182'>(Lånekassen)</span>
    <span class='ocrx_word' id='word_3_0_5' title='bbox 413 174 438 182'>skal</span>
    <span class='ocrx_word' id='word_3_0_6' title='bbox 445 174 496 182'>forvalte</span>
    <span class='ocrx_word' id='word_3_0_7' title='bbox 502 174 621 182'>utdanningsstøtten</span>
    <span class='ocrx_word' id='word_3_0_8' title='bbox 626 174 629 182'>i</span>
    <span class='ocrx_word' id='word_3_0_9' title='bbox 635 174 689 182'>samsvar</span>
    <span class='ocrx_word' id='word_3_0_10' title='bbox 694 174 723 182'>med</span>
</span>
<span class='ocr_line' id='line_3_0_1' title='bbox 694 174 723 182; baseline 0 0'>
    <span class='ocrx_word' id='word_3_0_11' title='bbox 100 184 202 192'>bestemmelsene</span>
    <span class='ocrx_word' id='word_3_0_12' title='bbox 208 184 229 192'>gitt</span>
    <span class='ocrx_word' id='word_3_0_13' title='bbox 234 184 237 192'>i</span>
    <span class='ocrx_word' id='word_3_0_14' title='bbox 244 184 273 192'>eller</span>
    <span class='ocrx_word' id='word_3_0_15' title='bbox 278 184 281 192'>i</span>
    <span class='ocrx_word' id='word_3_0_16' title='bbox 287 184 345 192'>medhold</span>
    <span class='ocrx_word' id='word_3_0_17' title='bbox 350 184 365 192'>av</span>
    <span class='ocrx_word' id='word_3_0_18' title='bbox 370 184 523 192'>utdanningsstøtteloven.</span>
    <span class='ocrx_word' id='word_3_0_19' title='bbox 529 184 595 192'>Rammene</span>
    <span class='ocrx_word' id='word_3_0_20' title='bbox 601 184 619 192'>for</span>
    <span class='ocrx_word' id='word_3_0_21' title='bbox 625 184 717 192'>virksomheten</span>
    <span class='ocrx_word' id='word_3_0_22' title='bbox 723 184 735 192'>til</span>
</span>
<span class='ocr_line' id='line_3_0_2' title='bbox 723 184 735 192; baseline 0 0'>
    <span class='ocrx_word' id='word_3_0_23' title='bbox 100 193 178 201'>Lånekassen</span>
    <span class='ocrx_word' id='word_3_0_24' title='bbox 183 193 244 201'>fastsettes</span>
    <span class='ocrx_word' id='word_3_0_25' title='bbox 250 193 268 201'>for</span>
    <span class='ocrx_word' id='word_3_0_26' title='bbox 274 193 308 201'>øvrig</span>
    <span class='ocrx_word' id='word_3_0_27' title='bbox 314 193 328 201'>av</span>
    <span class='ocrx_word' id='word_3_0_28' title='bbox 333 193 406 201'>overordnet</span>
    <span class='ocrx_word' id='word_3_0_29' title='bbox 411 193 488 201'>myndighet.</span>
</span>
<span class='ocr_line' id='line_3_0_3' title='bbox 411 193 488 201; baseline 0 0'>
</span></p></div>
```

Språkdeteksjon

- Modeller for bokmål, nynorsk, samiske språk og andre språk (utviklet av UIT/Giellatekno)
- Frekvens av bokstavsekvenser og enkeltord
- Algoritme: Out-of-place-ranking (Cavnar/Trenkle 1994)
- Implementasjon: Textcat
- Klassifikasjon på dokument- og avsnittnivå

Språkdeteksjon på avsnitt

	paragraph_nr	line	lang	tokens
0	0	VEDTEKTER FOR STATENS LÅNEKASSE FOR UTDANNING		6
1	1	Fastsatt av Kunnskapsdepartementet den 15.januar 2016 i medhold av lov 3. juni 2005 nr. 37 om utdanningsstøtte.	nob	17
2	2	1 Formål		2
3	3	Statens lånekasse for utdanning (Lånekassen) skal forvalte utdanningsstøtten i samsvar med bestemmelsene gitt i eller i medhold av utdanningsstøtteloven. Rammene for virksomheten til Lånekassen fastsettes for øvrig av overordnet myndighet.	nob	30
4	4	Utfyllende bestemmelser om virksomhetens formål framgår av Hovedinstruks for Lånekassen.		10
5	5	2 Lånekassens ledelse og oppgaver		5
6	6	Lånekassen er et ordinært forvaltningsorgan som ledes av eget styre og tilsatt administrerende direktør.	nob	14
7	7	Styret er Lånekassens øverste organ og er ansvarlig for den samlede virksomheten. Styret skal blant annet påse at Lånekassen drives i samsvar med regelverket, samt styringssignaler og retningslinjer gitt av overordnede myndigheter. Styret skal sikre at fastsatte mål oppnås og at ressursbruken er effektiv. Styret har også ansvar for at det foreligger vurderinger som identifiserer de viktigste risikofaktorene knyttet til virksomheten og at det foreligger planer for oppfølging av disse faktorene.	nob	71
8	8	Styret skal fastsette Lånekassens strategi og sørge for at denne blir realisert. Styret skal behandle forslag fra Lånekassen om vesentlige endringer i utdanningsstøtteordningen, samt øvrige viktige saker for Lånekassens virksomhet.	nob	30
9	9	Styret ansetter administrerende direktør på åremål. Betingelsene for ansettelsesforholdet fastsettes av Kunnskapsdepartementet.	nob	12

Korpus

- 4,3 milliarder løpende ord (deduplisert)
  - 3,2 milliarder på bokmål
  - 765 millioner på engelsk
  - 289 millioner på nynorsk
  - 5,8 millioner på nordsamisk
- Muligens den største frie tekstressursen for norsk
  - Common Crawl (OSCAR):
    - Bokmål: 804 millioner
    - Nynorsk: 9,4 millioner

Emnemodellering: Helsedir.

Topic 0: kommune, pasient, helse-, pårørend, besøk, omsorgstjeneste, tjeneste, behov, ansette, helsepersonell, råd, anbefaling, sykehjem, kommunal, kompetanse, beboer, nødvendig, omsorgsinstitusjone, vurdere, informasjon, institusjon, smitte, sikre, fastlege, oppfølging

Topic 1: prosent, pasient, helsevern, psykisk, antall, tertial, poliklinisk, aktivitet, nedgang, konsultasjon, periode, endring, sykehus, kontakt, ventetid, døgnopphold, andel, datum, sammenligne, voksen, fastlege, måned, talle, somatisk, behandling

Topic 2: metode, behandling, metodevurdering, pasient, legemiddel, legemiddelverk, drøfting, innspill, metodevurderinge, beslutning, kostnad, effekt, klinisk, studium, vurdering, forslag, al, metodevarsel, rapport, hurtig, fagdirektør, dokumentasjon, oppdrag, indikasjon, fullstendig

Topic 3: barn, skole, barnehage, elev, ung, veileder, ansette, råd, tiltak, skolehelsetjeneste, smitte, sykdom, tjeneste, avstand, gravid, behov, ungdom, kohort, møte, anbefaling, foresette, faglig, helsestasjons-, drift, ivareta

Topic 4: person, tiltak, kommune, smitte, karantene, reise, arrangement, avstand, land, test, oppdrag, covid, teste, testing, utbrudd, uke, nærkontakt, vurdering, unntak, hurtigtest, symptom, innreisekarante, område, antall, tilfelle

Emnemodellering: Helsedir.

01. https://www.helsedirektoratet.no/rapporter/tiltak-pa-skole-og-barnehageomradet-under-koronautbruddet-varen-2020/Tiltak%20p%C3%A5%20skole-%20og%20barnehageomr%C3%A5det%20under%20koronautbruddet%20v%C3%A5ren%202020.pdf/_/attachment/inline/1811d255-4e8d-4c8f-8433-c97c002df1c0:85099ff1d203787e87babbacfda6ee4bddfc9658/Tiltak%20p%C3%A5%20skole-%20og%20barnehageomr%C3%A5det%20under%20koronautbruddet%20v%C3%A5ren%202020.pdf
02. https://www.helsedirektoratet.no/veiledere/smittevern-for-skoletrinn-1-7-covid-19/bakgrunn
03. https://www.helsedirektoratet.no/veiledere/smittevern-i-ungdomsskole-og-videregaende-skole-covid-19/bakgrunn
04. https://www.helsedirektoratet.no/tema/beredskap-og-krisehandtering/koronavirus/anbefalinger-og-beslutninger/Presentasjon%20rapport%20ekspertgruppe%203.%20april%202020.pdf/_/attachment/inline/c4f51efd-ac8c-46ea-8a89-91e086f3e693:c1d87b7cc09d4c11612355f9b2978729297f61a6/Presentasjon%20rapport%20ekspertgruppe%203.%20april%202020.pdf
05. https://www.helsedirektoratet.no/veiledere/covid-19-smittevern-i-barnehager/bakgrunn
06. https://www.helsedirektoratet.no/veiledere/smittevern-for-skoletrinn-1-7-covid-19/er-det-noen-barn-og-ansatte-det-ma-tas-spesielle-hensyn-til
07. https://www.helsedirektoratet.no/veiledere/covid-19-smittevern-i-barnehager/er-det-barn-og-ansatte-det-ma-tas-spesielle-hensyn-til
08. https://www.helsedirektoratet.no/nyheter/kunnskap-om-barn-og-koronasmitte
09. https://www.helsedirektoratet.no/veiledere/smittevern-i-ungdomsskole-og-videregaende-skole-covid-19/er-det-noen-ungdom-og-ansatte-det-ma-tas-spesielle-hensyn-til
10. https://www.helsedirektoratet.no/nyheter/oppdaterer-veiledere-anbefalinger-og-rad-om-barn-og-unge

Veien videre

Avleveringsløsning for språkdata

- Språkbanken ønsker å etablere en digital avleveringsløsning for språkdata
- Skal gjøre det mulig for statlige institusjoner og private virksomheter å laste opp (anonymisert) materiale som kan publiseres i Språkbankens ressurskatalog

Takk for oppmerksomheten!

Nasjonalbiblioteket