Proceedings of the TERM21 Workshop , pages 13–18
Language Resources and Evaluation Conference (LREC 2022), Marseille, 20-25 June 2022

Converting from the Nordic Terminological Record Format to the TBX Format

Maria Skeppstedt, Marie Mattson, Magnus Ahltorp, Rickard Domeij
Institute for Language and Folklore
Stockholm, Sweden
firstname.lastname@isof.se

Abstract

Rikstermbanken (Sweden's National Term Bank), which was launched in 2009, uses the Nordic Terminological Record Format (NTRF) for organising its terminological data. Since then, new terminology formats have been established as standards, e.g., the Termbase eXchange format (TBX). We here describe work carried out by the Institute for Language and Folklore within the Federated eTranslation TermBank Network Action. This network develops a technical infrastructure for facilitating sharing of terminology resources throughout Europe. To be able to share some of the term collections of Rikstermbanken within this network and export them to Eurotermbank, we have implemented a conversion from the Nordic Terminological Record Format, as used in Rikstermbanken, to the TBX format.

Keywords: Term banks, Nordic Terminological Record Format, TBX

1. Introduction

Rikstermbanken,[1] (Sweden's National Term Bank) was originally developed by "Terminologicentrum TNC" (The Swedish Centre for Terminology, TNC), which in 2006 was commissioned by the Swedish government to develop a national termbank (Nilsson, 2009; Bucher, 2009). The first technical implementation of Rikstermbanken was launched in 2009, and the product has since then been available through a search interface on a public website, which has been used by translators and terminologists at public agencies and other organisations in Sweden. The termbank hosts externally developed term collections, both from the public and private sector, as well as collections developed by TNC. Rikstermbanken contains both small and large term collections, with a total of 130,000 term entries, many of them multi-lingual.

When TNC closed down in the end of 2018, the responsibility of maintaining Rikstermbanken was handed over to ISOF (the Swedish Institute for Language and Folklore), i.e., the responsibility of maintaining the terminological content as well as the technical product. ISOF replaced the original Java and SQL-based implementation of Rikstermbanken in 2021 by a new technical implementation based on Python, Flask and the document database MongoDB.

According to the Language Act (Språklag (2009:600), 2009)[2] , the Swedish government agencies have the responsibility to ensure that terminology in their various areas of expertise is accessible, used and developed. ISOF provides the other agencies with support for implementing the Language Act, and Rikstermbanken forms one part of this work.

[1] https://www.rikstermbanken.se

[2] https://www.riksdagen.se/sv/dokument-lagar/dokument/svensk-forfattningssamling/spraklag-2009600_sfs-2009-600

2. NTRF and TBX

The format chosen for storing the terminological data when developing the original version of Rikstermbanken was NTRF, the Nordic Terminological Record Format (Rådet for teknisk terminologi, 1999). This is a terminology format developed by central terminology institutions of Finland, Norway and Sweden. The standard NTRF version was adapted to requirements specific for the terminology data stored in Rikstermbanken to a local version of NTRF, which uses the fields shown in the first column of Table 2 for organising the data.

Since then, new terminology formats have been established as standards, e.g., the Termbase eXchange format (TBX) (Localization Industry Standards Association, 2008). TBX is an international standard for representation of structured terminological resources, and it defines an XML format for the exchange of terminology data. It is, for instance, used in terminology software and CAT tools (computer-assisted translation tools) to support the functionality of importing and exporting term lists.

There are also other possible formats, such as the SKOS format. However, as there are no hierarchical relations in the term collections in Rikstermbanken nor any linked data relations to entities outside of each term collection, we considered TBX to be the format most suitable to the data in Rikstermbanken.

Since Rikstermbanken was originally developed, it has also become more common that term collections are released as open data, and some of the term collections in Rikstermbanken are possible to release with an open license. These term collections are more useful to the third party user if they are made available in a standard

format. They can then, for instance, easily be imported into a CAT tool or into other TBX-based termbanks.
Therefore, to be able to share some of the data from Rikstermbanken in a standardised format, we have implemented a conversion from NTRF, as used in Rikstermbanken, to the TBX format.

2.1. An NTRF example

NTRF is a row-based format. Thereby, its structure is very different from the XML-based, hierarchical, TBX format. Table 1 shows the NTRF representation for the concept "biologisk mångfald" (biological diversity) from a collection of sustainability terms.[3]
The format also allows for expressing formatting of terms and texts, e.g., to express italics with HTML-like markup: {i}biologisk mångfald {/i}.

sv**TE** biologisk mångfald
sv**SYTE** biodiversitet
sv**UPTE** artrikedom
en**TE** biological diversity
en**SYTE** biodiversity
sv**DF** rikedom av arter, av genetisk variation inom arter samt av de ekosystem som arterna ingår i
sv**AN** Begreppet biologisk mångfald betonar betydelsen av variationsrikedom bland alla levande organismer, exempelvis bakterier, växter, svampar och djur, samt de ekosystem och livsmiljöer de ingår i, allt från landskap med många olika naturtyper till städer med parker och grönområden.
I FN-fördraget "Konventionen om biologisk mångfald" definieras {i}biologisk mångfald {/i} som 'variationsrikedom bland levande organismer av alla ursprung, inklusive från bland annat landbaserade, marina och andra akvatiska ekosystem och de ekologiska komplex i vilka dessa organismer ingår; detta innefattar mångfald inom arter, mellan arter och av ekosystem'. Den definitionen används bland annat i juridiska sammanhang.
Benämningen {i}biologisk mångfald {/i} skiljer sig från det tidigare använda uttrycket {i}artrikedom{/i}, genom att det även avser den genetiska variationen inom en art
sv**EX** Ett exempel på biologisk mångfald är när det finns olika arter av bin och humlor: jordhumlor, snäckmurarbin, tapetserarbin, honungsbin, fjällhumlor osv. De har olika kroppsform och längd på tunga, och olika preferenser när det gäller pollen och nektar, vilket innebär att de kan pollinera olika arter av växter.

Table 1: The NTRF representation for the concept "biologisk mångfald" (biological diversity) in Rikstermbanken.

Figure 1 shows how the term-post above is presented in the user interface of Rikstermbanken.

[3]The collection of sustainability terms, developed by the Institute for Language and Folklore, is licensed under a Creative Commons Attribution 4.0 International License, CC-BY 4.0 (`http://creativecommons.org/licenses/by/4.0/`)

3. The Federated eTranslation TermBank Network Action

The work of implementing a conversion from NTRF to TBX has been carried out within the Federated eTranslation TermBank Network Action, which is a network that develops a technical infrastructure for facilitating sharing of terminology resources throughout Europe. Among other initiatives, the network has implemented an API for pushing terminology resources in the TBX format from other termbanks into Eurotermbank. We will use this API and the TBX conversion described here for exporting some of Rikstermbanken's term collections to Eurotermbank[4].
In addition to the API for pushing terminology resources, the network has also implemented the Eurotermbank toolkit. This is a toolkit for managing terminology resources, i.e., for creating, editing, importing and exporting terminology in various formats. The toolkit is set up as a local web-based application, which functions as a local node that can export the terminology data to Eurotermbank. There are thus two main methods for exporting data into Eurotermbank, (i) either to use the API (i.e., the method ISOF uses), or (ii) to create or import terminology lists into a local node of the Eurotermbank toolkit.[5] The term lists exported to Eurotermbank are then exported further into the repository of the European Language Resource Coordination initiative, ELRC-SHARE[6]. This repository is used for training eTranslation[7], the machine translation system developed by the European Commission.
Figure 2 illustrates the conversion from NTRF to TBX and the export to Eurotermbank, and Figure 3 shows the term-post for "biologisk mångfald" (biological diversity) when it has been imported into Eurotermbank. Eurotermbank uses TBX 2.0, and this version was therefore chosen for the TBX export[8].

3.1. Term lists that are shared within the Federated eTranslation TermBank Network Action

The focus of the Federated eTranslation TermBank Network Action has been to construct and evaluate the technical infrastructure for sharing terminology resources, rather than to actually carry out the collection of resources to share. However, in order to practically evaluate the infrastructure, we have decided to start by exporting the following four resources from Rikstermbanken, with either a CC0 or a CC-BY license.

[4]`https://www.eurotermbank.com`
[5]Information on the network is available here: `https://www.eurotermbank.com/participants-network`
[6]`https://elrc-share.eu`
[7]`https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation`
[8]With the following specification: `https://eurotermbank.com/tbx-0.5.1.xcs`

Termpost

SVENSKA TERMER:	**biologisk mångfald** biodiversitet
DEFINITION:	rikedom av arter, av genetisk variation inom arter samt av de ekosystem som arterna ingår i
ANMÄRKNING:	Begreppet biologisk mångfald betonar betydelsen av variationsrikedom bland alla levande organismer, exempelvis bakterier, växter, svampar och djur, samt de ekosystem och livsmiljöer de ingår i, allt från landskap med många olika naturtyper till städer med parker och grönområden. I FN-fördraget ”Konventionen om biologisk mångfald” definieras *biologisk mångfald* som 'variationsrikedom bland levande organismer av alla ursprung, inklusive från bland annat landbaserade, marina och andra akvatiska ekosystem och de ekologiska komplex i vilka dessa organismer ingår; detta innefattar mångfald inom arter, mellan arter och av ekosystem'. Den definitionen används bland annat i juridiska sammanhang. Benämningen *biologisk mångfald* skiljer sig från det tidigare använda uttrycket *artrikedom*, genom att det även avser den genetiska variationen inom en art
EXEMPEL:	Ett exempel på biologisk mångfald är när det finns olika arter av bin och humlor: jordhumlor, snäckmurarbin, tapetserarbin, honungsbin, fjällhumlor osv. De har olika kroppsform och längd på tunga, och olika preferenser när det gäller pollen och nektar, vilket innebär att de kan pollinera olika arter av växter.
ENGELSKA TERMER:	biological diversity biodiversity
KÄLLA:	Hållbarhetstermgruppen: Termlista | 2021

Figure 1: The concept “biologisk mångfald” (biological diversity), as shown in Rikstermbanken (www.rikstermbanken.se).

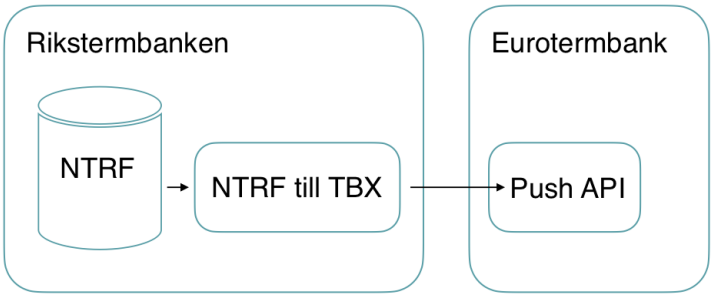

Figure 2: An illustration of the conversion from NTRF to TBX and the export to Eurotermbank

3.1.1. A collection of sustainability terms

ISOF organises the Sustainability Terminology Group, which aims to standardise and clarify Swedish terminology and concepts related to sustainable development. The group includes experts from a number of different knowledge areas, as well as actors who work to communicate such expert knowledge to a wider audience. The group members represent various scientific disciplines, government agencies, the media as well as NGOs. The linguistic expertise is provided by two terminologists and a discourse analyst, all from ISOF, as well as a translator from English to Swedish.

3.1.2. Terms from the Swedish authority terminology group

ISOF has also been organising a terminology group with the long-time goal of creating a more standardised terminology within the public sector. The terminology group included five different public agencies: the Swedish Tax Agency, the Swedish Public Employment Service, the Swedish Social Insurance Agency, the Swedish Police Authority and the National Board of Health and Welfare. The group members were terminologists, language experts, translators and business architects, and the group was led by a project manager, a project assistant and a terminologist from ISOF. The

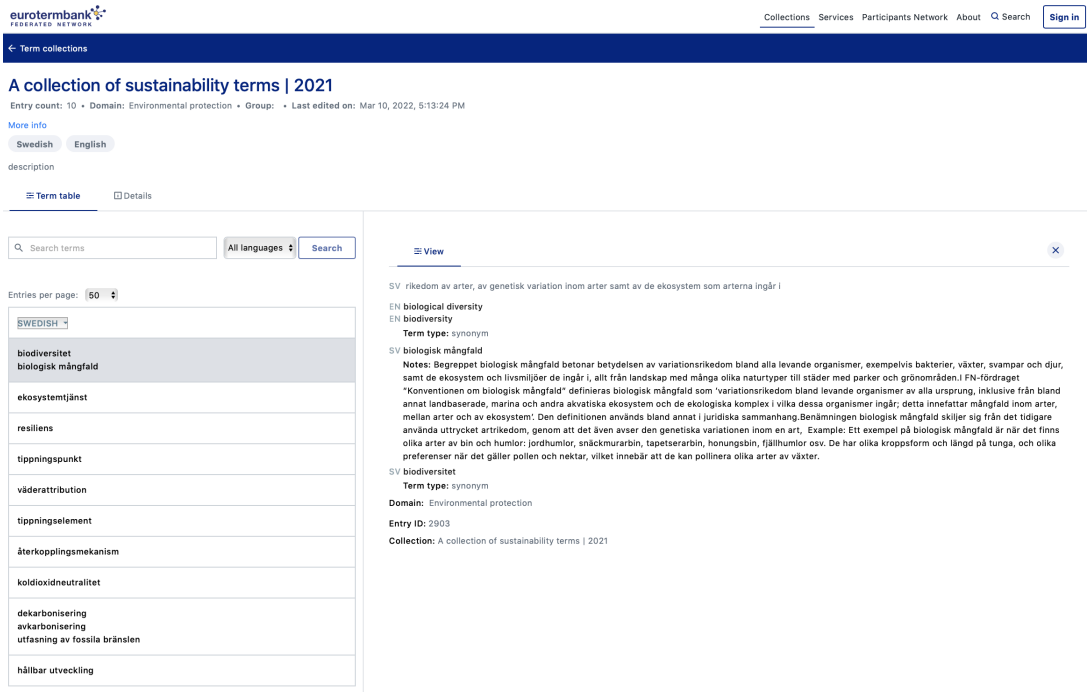

Figure 3: The concept "biological diversity" as shown when imported into Eurotermbank.

term list developed in the project consists of 27 term-posts, which are translated into five languages: English, Arabic, Finnish, Romani Arli and Romani Kelderash. The work began in August 2021 as a pilot project, but the group will hopefully continue to expand the terminology collection.

3.1.3. Statistics Sweden's term list

Statistics Sweden (SCB) is a public agency responsible for official statistics. Their term list, which contains terms in Swedish and English, consists of terms related to statistics, society, and other topics that can be useful when communicating statistics. The most recent update of the list was created by two statistics experts, two language experts and a terminologist consultant.

3.1.4. Swedish Council for Higher Education's term list

The Swedish Council for Higher Education (UHR) is responsible for supporting the higher education sector by providing admission services, IT systems, and the Swedish Scholastic Aptitude Test, among other things. Their term list consists of over 2,000 terms and synonyms related to higher education. The list is updated yearly with the help of institutions in the higher education sector.

4. The conversion

It could be concluded that the correspondence between NTRF and TBX in general was very good. There were only three pieces of information in NTRF that could not be directly expressed in TBX. These were (i) the Swedish common gender, (ii) domain on a language level, and (iii) translation equivalence comment on a language level. Table 2 shows a simplified mapping between the two formats, i.e., simplified in the way that the hierarchy of the TBX format has been left out. The pieces of information that can not be expressed in TBX are shown in boldface. The table is divided into the following five sections:

(2.1) TBX uses the hierarchical structure of XML to divide the term-post into language segments, whereas NTRF specifies the language for each term- and text row.

(2.2) NTRF allows the user to specify a number of different types of terms, e.g., standard term, synonym, deprecated term, whereas TBX uses the $<term>$ tag for all kinds of terms, and lets the user add additional tags to specify different kinds of term attributes, e.g., if it is a synonym or a deprecated term.

(2.3) For some features of the term, NTRF too uses attributes for specifying them, e.g., grammatical information, geographical usage, abbreviation/full form and homograph information. For all the attributes, there is a corresponding TBX tag. As stated above, we were, however, not able to find any standard in TBX for expressing that a noun has the common gender, which is one of the gender categories of Swedish. Terms having another gender category than "feminine", "masculine" and "neuter" were therefore given the gender category "othergender".

(2.4) There is also information on a language level, both for NTRF and TBX, e.g., a definition or explanation of the concept, or a note. These pieces of information can optionally have a reference. On the lan-

NTRF	TBX
1. The language:	
la	<langSet xml:lang=*la*> … (where *la* is a variable containing the language)
2. The term:	
*la*TE *word*	<term>*word*</term> …
*la*AVTE *word*	<term>*word*</term> … <termNote type="normativeAuthorization">deprecatedTerm</termNote> <termNote type="administrativeStatus">deprecatedTerm-admn-sts</termNote>
*la*BT *word*	<term>*word*</term> … <termNote type="termType">formula</termNote>
*la*PH *word*	<term>*word*</term> … <termNote type="termType">phraseologicalUnit</termNote>
*la*SYPH *word*	<term>*word*</term> … <termNote type="termType">synonym</termNote> <termNote type="termType">synonymousPhrase</termNote> <termNote type="termType">phraseologicalUnit</termNote>
*la*SYTE *word*	<term>*word*</term> … <termNote type="termType">synonym</termNote>
*la*INTE *word*	<term>*word*</term> … <termNote type="termType">fullForm</termNote>
Also for: *la*TE and *la*PH	(If the setting 'troligenUppdelat' ('probably split') is true for the language:) <termNote type="normativeAuthorization">preferredTerm</termNote> <termNote type="administrativeStatus">preferredTerm-admn-sts</termNote>
Also for: *la*SYPH and *la*SYTE	(If the setting 'troligenUppdelat' ('probably split') is true for the language:) <termNote type="normativeAuthorization">admittedTerm</termNote> <termNote type="administrativeStatus">admittedTerm-admn-sts</termNote>
3. Information associated with the term:	
GNGR f	<termNote type="grammaticalGender">feminine</termNote>
GNGR m	<termNote type="grammaticalGender">masculine</termNote>
GNGR t	<termNote type="grammaticalGender">neuter</termNote>
GNGR **other**	<termNote type="grammaticalGender">**otherGender**</termNote> (Can, e.g., be used for Swedish common gender, which is not expressed in TBX.)
GR pl	<termNote type="grammaticalNumber">plural</termNote>
GR sing	<termNote type="grammaticalNumber">singular</termNote>
GR koll	<termNote type="grammaticalNumber">mass</termNote>
another GR	<termNote type="grammaticalNumber">otherNumber</termNote>
OKGR subst	<termNote type="partOfSpeech">noun</termNote>
OKGR adj	<termNote type="partOfSpeech">adjective</termNote>
OKGR verb	<termNote type="partOfSpeech">verb</termNote>
OKGR adv	<termNote type="partOfSpeech">adverb</termNote>
OKGR itr	<termNote type="partOfSpeech">verb</termNote> <termNote type="grammaticalValency">monovalent</termNote>
OKGR tr	<termNote type="partOfSpeech">verb</termNote> <termNote type="grammaticalValency">divalent or more</termNote>
another OKGR	<termNote type="partOfSpeech">other</termNote>
FRKT F	<termNote type="termType">abbreviation</termNote>
FRKT OF	<termNote type="termType">fullForm</termNote>
HONR *nr*	<termNote type="homograph">*nr*</termNote>
UT *text*	<termNote type="pronunciation">*text*</termNote>
RF *text*	<xref type="xSource" target=" *text* " />
SA *text*	<termNote type="usageNote">*text*</termNote>
GE *text*	<termNote type="geographicalUsage">*text*</termNote>
EKVI *text*	<termNote type="transferComment">*text*</termNote> (EKVI is currently not used in Rikstermbanken on a term level.)
4. Information associated with the language:	
*la*DF *text*	<descrip type="definition">*text*</descrip>
*la*EX *text*	<descrip type="example">*text*</descrip>
*la*FK *text*	<descrip type="explanation">*text*</descrip>
*la*KT *text*	<descrip type="context">*text*</descrip>
la{DF/EX/FK/KT} *text* RF *text*	(The four preceding text attributes can also be given an optional reference) <descrip type="{definition/example/explanation/context}">*text*</descrip> <admin type="sourceIdentifier">*text*</admin>
*la*AN *text*	<note>*text*</note>
*la*AN *text.1* RF *text.2*	(In addition, if there is a reference associated with the note) <admin type="annotatedNote">*text.1*</admin> <adminNote type="noteSource">*text.2*</adminNote>
***la*SA** *text*	<note>**Domain:** *text*</note>
***la*EKVI** *text*	<note>**Equivalence:** *text*</note>
*la*UPTE *w.1, w.2*	<admin type="searchTerm">*w.1*</admin> <admin type="searchTerm">*w.2*</admin>
5. References to other terms, which are associated with the language in NTRF and with the term-post in TBX:	
*la*RETE *word* (<HONR *nr*>)	<ref type="crossReference" target="*word*(-*nr*)">*word*</ref> (A homograph number is needed when referencing to homographs)
*la*SU *word* (<HONR *nr*>)	<ref type="see" target="*word*(-*nr*)">*word*</ref> (A homograph number is be needed when referencing to homographs)

Table 2: A simplified mapping table between Rikstermbanken NTRF and TBX, not showing the hierarchical structure of the TBX. *la* is variable containing the language, *word* (and *w.1*/*w.2*) contains a word, *text* contains a text, and *nr* contains a number. The three pieces of information in NTRF that could not be expressed in TBX is shown in boldface.

guage level, both NTRF and TBX also lets the user specify search words, i.e., words that should lead to this term-post being retrieved when used in a search query. NTRF also allows the user to specify a "translation equivalence comment" as well as a "domain" on the language level. We have not been able to find support for adding this information on the language level in TBX, and have therefore instead added a standard TBX note that starts with the text "Domain:" and "Equivalence:", respectively. This is shown in boldface in the table.

(2.5) Finally, related terms and see-under terms are expressed on a language level in NTRF, whereas they are expressed with XML tags on a term-post level in TBX. We moved the information to the term-post level when carrying out the conversion.

The typographic formatting of the text and terms, i.e., the HTML-like markup, is not exported in the current implementation of the conversion.

The conversion is implemented in Python. The conversion procedure consists of first retrieving the NTRF formatted files from their representation in the document database, and thereafter converting them into TBX.

5. Future work

In the future, we will continue to select term lists from Rikstermbanken to export to Eurotermbank, as well as to develop and collect new term lists to include in Rikstermbanken.

We plan to continue to support NTRF for Rikstermbanken, as there is knowledge within ISOF on how to use this format. We have, however, also implemented the first stage of a conversion in the other direction, i.e., a conversion *from* TBX. Such a conversion would make it possible to import data available in a TBX format into the MongoDB database of Rikstermbanken, i.e., making the TBX format one of the formats supported for importing data into Rikstermbanken.

6. Acknowledgements

The Federated eTranslation TermBank Network Action is co-financed by the Connecting Europe Facility of the European Union.

The contents of this paper are the sole responsibility of the authors and do not necessarily reflect the opinion of the European Union.

7. Bibliographical References

Bucher, A.-L. (2009). Terminologisamordning inom svenska myndigheter ny språklag på väg. In *NORDTERM16: Ontologier og taksonomier*.

Localization Industry Standards Association. (2008). Systems to manage terminology, knowledge, and content - termbase exchange (TBX). https://www.gala-global.org/sites/default/files/migrated-pages/docs/tbx_oscar_0.pdf.

Nilsson, H. (2009). The realisation of a national term bank – how and why? In *ELETO – 7th Conference Hellenic Language and Terminology*.

Rådet for teknisk terminologi. (1999). Nordic terminological record format (NTRF).

Språklag (2009:600). (2009). Kulturdepartementet.