TAUS Presents:

TAUS Data Cloud

Capitalizing on Translation Data

www.taus.net/data

TAUS

The Data Cloud is …

- A neutral and secure platform to
  - Search for general or domain-specific translation data
  - Share and pool translation data, based on a reciprocal model
  - Acquire credits for data, if there is no or not enough translation data to share
- A very large industry-shared repository of translation data
  - Accessible through the Web or API
- A super cloud for the global translation industry helping to
  - Improve translation quality – its primary use case being MT engines' training
  - Fuel translation automation and business innovation

TAUS

Founding Members

The Data Cloud contains …

- Translation data of
  - Tens of billions of words in multiple translation directions across 62 language locales, 17 industry domains and 9 content types
- Data sources
  - Industry-shared TMs contributed by TAUS members (providers, buyers, LSPs)
  - Individual TMs contributed by TAUS users (freelance translators/professionals)
  - Public data made available to the community as such by large organizations and institutions
  - Public data generated and/or curated by TAUS members and users
- See also relevant FAQ: What types of data does the Data Cloud contain?

TAUS

Features

- Search
- Upload
- Discover & Download
- Account History
- My Data
- Feedback

- API for integration to other technologies
- Matrix function

TAUS

Attributes

- Translation Direction
- Industry Domain
- Content Type
- Data Owner
- (Data Provider)

TAUS

Data Set Acquisition

- Data Cloud data sets can be acquired in 3 ways:
  - Sign up for a TAUS membership and receive free bonus credits
  - Earn credits by sharing own data
  - Buy credits if there is no or not enough data to share

- TAUS ensures worry-free use of the data in Data Cloud through:
  - Easy upload & download of standard TMX files
  - The TAUS legal framework addressing concerns over ownership and IP

TAUS

Legal Framework

- Main points included in TAUS Legal Framework:
  - Free to use translations
  - Free to develop derivative work
  - Copyright remains with data owner
  - Agreed by all users and members
  - No re-distribution of the translation data
  - TAUS Data Cloud owns the IP to the infrastructure

- For more information, read TAUS Terms of Use:
  - https://www.taus.net/taus-terms-of-use

TAUS

Data Quality

- Built-in automated quality checks
  - See FAQ: Does the Data Cloud monitor data quality?
- Monitoring of data uploads
  - Corrupted or bad data is archived. Used credits on such data can be refunded
- Community-driven monitoring of data quality
  - Through the Feedback page
  - By viewing data set samples, users can assess quality prior to data pooling

TAUS

Highlights of Data Cloud v2.1.0 (October 2016)

- Enhancement of data discoverability:
  - Select the translation direction and other attributes
  - Get the list of available data sets with metadata
  - Sort the list by attribute or volume
  - View the relation of required credits to the available data sets
- Data relevance for your projects:
  - View random samples of the data sets

TAUS

Highlights of Data Cloud v2.2.0 (June 2017)

- Faster search and indexing
  - Through modernized Data Cloud Infrastructure
- Users subscribed to the Free Tier
  - Can now discover and browse data sets
- Bonus Data Cloud credits
  - For new or renewed memberships

Some Use Scenarios

- Discover Translation Data in a Flash
  - Identify relevant translation data for your projects through the associated metadata and by browsing data samples
- Download Translation Data Sets for Machine Translation
  - Use Data Cloud resources to train, evaluate & improve performance of MT engines
- Assess Data Quality at a Glance
  - Make a first-hand data quality assessment by browsing random samples of retrieved data sets
- Download Translation Data Sets for Terminology Extraction
  - Extract terminology from the industry-shared TMs contributed by TAUS members

TAUS

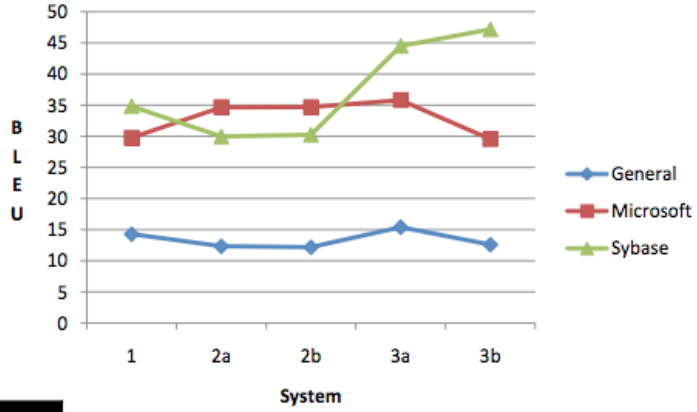
Use Cases

Microsoft, ModernMT, KantanMT

TAUS

Microsoft – Use Case

- In-domain (IT) data of multiple providers (EMC, Intel, Dell, Sybase, Adobe, Microsoft, etc.) were pooled from the Data Cloud
- The system trained with the combined in-domain data had a gain of 8 BLEU points



Chinese

			Test Set		
System	Size	System Description	General	Microsoft	Sybase
1	8.3M	General domain	14.26	29.74	34.81
2a	2.6M	Microsoft	12.32	34.65	29.95
2b	2.8M	Microsoft with Sybase	12.16	34.66	30.24
3a	11.5M	General and Microsoft and TAUS	15.38	35.80	44.49
3b	11.5M	System 3a with Sybase lambda	12.57	29.51	47.16

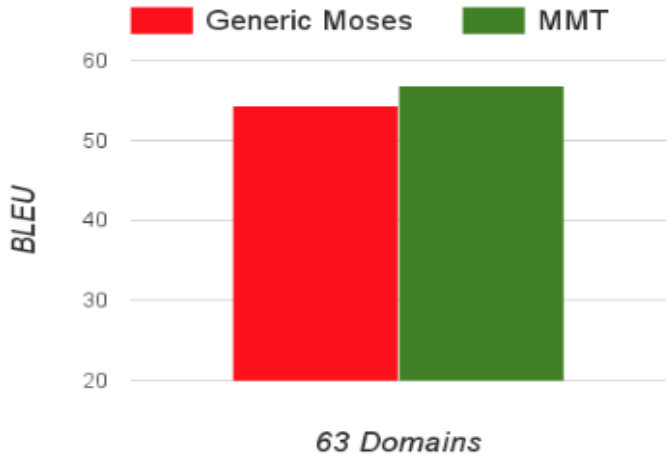
Source: Improving the quality of a customized SMT system using shared training data, C.Wendt, W.Lewis,MT Summit 29/8/2009

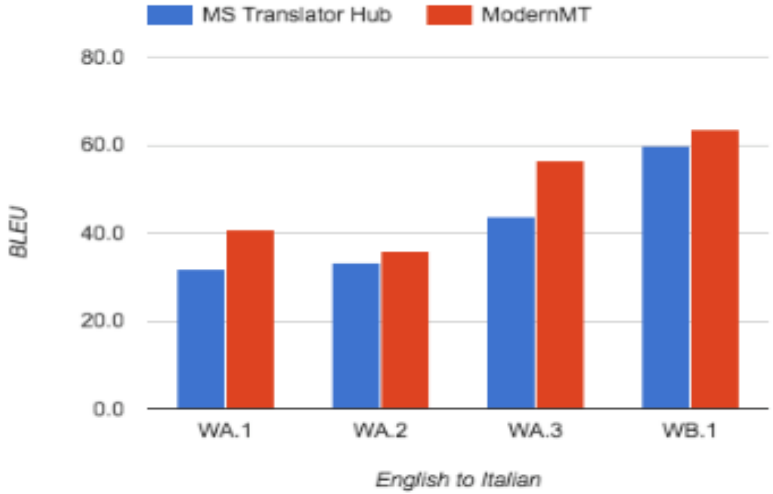
TAUS

ModernMT – Use Case

ModernMT or MMT is context adaptive and incremental learning MT technology

- Evaluation & comparison of two engines trained with combined data from MyMemory, Data Cloud, web data, public data and user-provided TMs.
- The high volume and quality of Data Cloud data is useful and valuable.





Source: MT adaptation from TMs in ModernMT, Marcello Federico (FBK), AMTA, Austin, Texas, 29/10/2016

KantanMT – Use Case

The addition of TAUS Data Cloud data relevant to the translation project increased the **BLEU score** of the engine by 12% improving therefore the quality and performance of the MT output. **F-measure** and **TER** were also improved.

https://www.taus.net/data/kantanmt-use-case-taus-data-cloud-data-used-to-boost-up-overall-quality, 2017

TAUS

Data Cloud Background

Resources & Roadmap

TAUS

Resources

- Browse the Data Cloud pages
- Watch the Data Cloud Tutorial Videos
- Find answers in the Data Cloud FAQ section
- Request a live, customized Data Cloud Demo
- Download for free and read the following two reports
  - TAUS Translation Data Landscape Report (December 2015)
  - TAUS Data Market White Paper (June 2017)

TAUS

Roadmap

- Matching Scores
  - For discovery of translation data most similar to users' projects
- Transaction interface and backend
  - For payments, credits and bids on unavailable/new translation data
- Feedback and tracking
  - Number of sample views and usages/downloads
- Premium API
  - For power users and MT developers to provide Data Cloud services from their own platforms to their clients
- Data Market
  - Market to buy, sell and bid on translation data
- Expand with Speech Data

TAUS

Get Involved!

- Sign up for the Data Cloud Newsletter to stay informed on the latest news: http://bit.ly/2cXS7Ga
- If you don't have an account yet, get started with the Data Cloud: https://www.taus.net/data/get-started-with-data-cloud
- Contribute your translation data sets to the Data Cloud and earn credits to download valuable translation data relevant to your projects!
- Follow the TAUS MT User Group and TAUS Data Summit 2017

TAUS

Data Cloud Screenshots

TAUS

Search

with respect to this

English (United Kingd > Dutch (Netherlands) Search

Less options

Industry: Computer Software

Content Type: Select...

Data Owner: Select...

Source Segment	Target Segment
Invert **with respect to this** circle	Inverteer ten opzichte van deze cirkel
Patients should be monitored **with respect to this**.	De patiënt dient hierop gecontroleerd te worden.
Device is performing no action **with respect to this** property	Apparaat doet niets wat deze eigenschap betreft

Previous 1 Next

Show 10 entries

Showing 1 to 3 of 3 entries

Link to this search

TAUS

Hi **TAUS Member** | Logout

Search | Upload | Discover & Download | Account History | My Data | Feedback | Data Cloud Guide

Upload

Data Owner: TAUS Provider

By uploading data, the uploader certifies that they own the data and license the data to Data Could under the Terms of Use.

Product Line: Default

Industry: Professional and Business Services

Content Type: Support Content

Zipped TMX File: TM_Upload_en-GB-nl-NL.tmx.zip Choose file

Upload

	Job	Created	Translation Direction			
[+]	7585 TM_Upload_en-...	2017-04-25 16:42:28	en-GB	>	nl-NL	Importing...

Data Cloud API | FAQs | Terms of Use

TAUS Data Cloud Release Version: 2.2.0
For any enquiries, please contact us: data@taus.net.

TAUS

Discover & Download

Translation Direction* :
English (United Kingdom)
Dutch (Netherlands)

Industry:
Select...

Content Type:
Select...

Data Owner:
Select...

Product Line:
Select...

Include Matrix translations (?)

Exclude data uploaded by my organization

▶ **Available Words:** 107,473,485 (?)
Required Credits: 107,473,485 (?)

▶ **Already Downloaded:** 0 (?)
Current Credits: 2,363,513,840 (?)

Balance: 2,256,040,355 (?)

Export

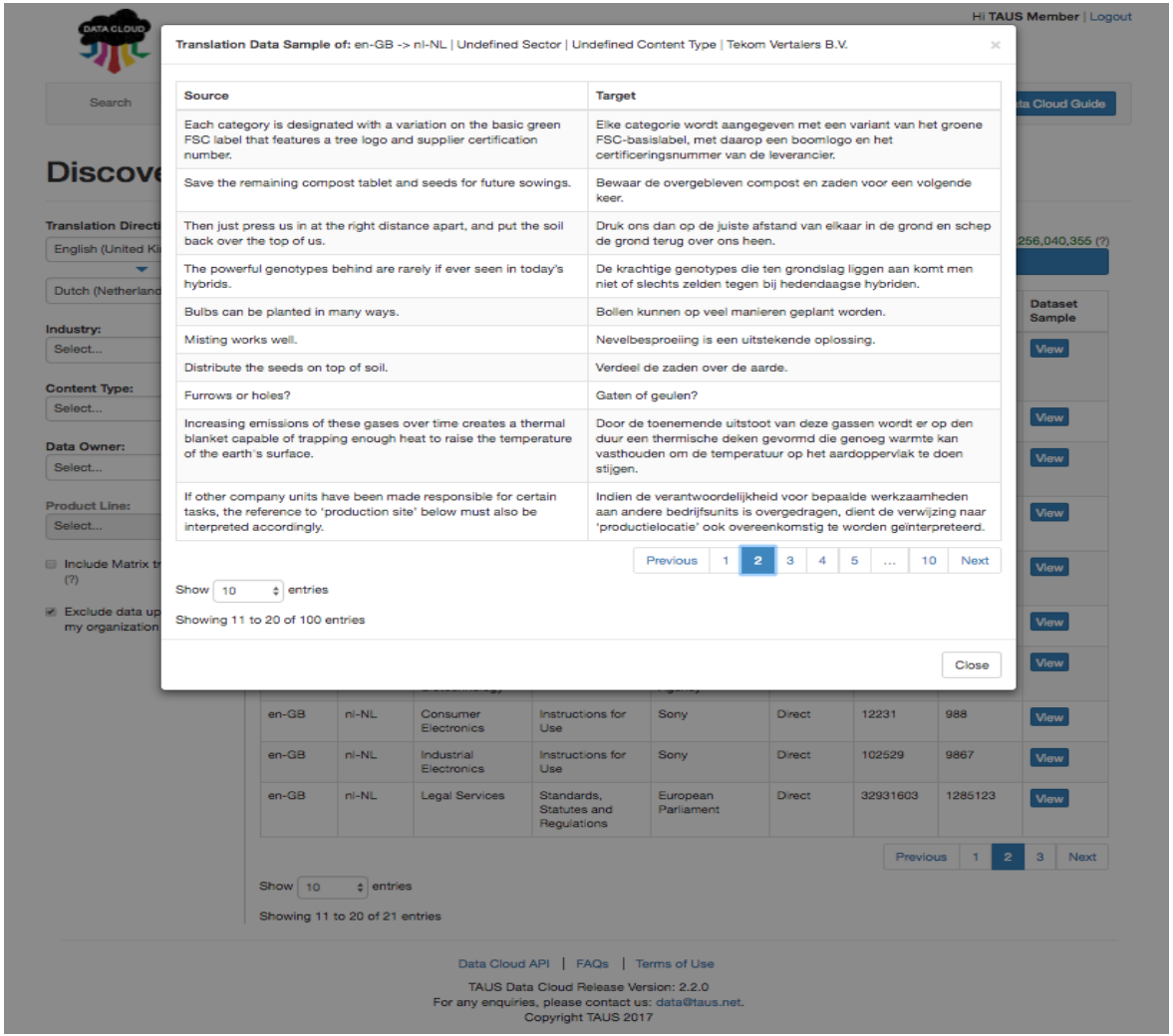
Source Language	Target Language	Industry	Content Type	Data Owner	Direct / Matrix	Word Count	Segment Count	Dataset Sample
en-GB	nl-NL	Financials	News Announcements, Reports and Research	Sebastiaan Vandenbore (8814)	Direct	3217508	108840	View
en-GB	nl-NL	Undefined Sector	Undefined Content Type	Tekom Vertalers B.V.	Direct	29618	2409	View
en-GB	nl-NL	Legal Services	Standards, Statutes and Regulations	European Parliament	Direct	25005596	1205972	View
en-GB	nl-NL	Computer Software	Software Strings and Documentation	Sebastiaan Vandenbore (8814)	Direct	882	61	View
en-GB	nl-NL	Undefined Sector	Undefined Content Type	Sebastiaan Vandenbore (8814)	Direct	355	16	View
en-GB	nl-NL	Undefined Sector	Undefined Content Type	Translavic B.V.	Direct	3104	261	View
en-GB	nl-NL	Pharmaceuticals and Biotechnology	Instructions for Use	European Medicines Agency	Direct	6248470	362882	View
en-GB	nl-NL	Consumer Electronics	Instructions for Use	Sony	Direct	12231	988	View
en-GB	nl-NL	Industrial Electronics	Instructions for Use	Sony	Direct	102529	9867	View
en-GB	nl-NL	Legal Services	Standards, Statutes and Regulations	European Parliament	Direct	32931603	1285123	View

Previous 1 2 3 Next

Show 10 entries

Showing 11 to 20 of 21 entries

TAUS


Hi TAUS Member | Logout
Search
Translation Data Sample of: en-GB -> nl-NL | Undefined Sector | Undefined Content Type | Tekom Vertalers B.V.
Source
Target
Each category is designated with a variation on the basic green FSC label that features a tree logo and supplier certification number.
Elke categorie wordt aangegeven met een variant van het groene FSC-basislabel, met daarop een boomlogo en het certificeringsnummer van de leverancier.
Save the remaining compost tablet and seeds for future sowings.
Bewaar de overgebleven compost en zaden voor een volgende keer.
Then just press us in at the right distance apart, and put the soil back over the top of us.
Druk ons dan op de juiste afstand van elkaar in de grond en schep de grond terug over ons heen.
The powerful genotypes behind are rarely if ever seen in today's hybrids.
De krachtige genotypes die ten grondslag liggen aan komt men niet of slechts zelden tegen bij hedendaagse hybriden.
Bulbs can be planted in many ways.
Bollen kunnen op veel manieren geplant worden.
Misting works well.
Nevelbesproeiing is een uitstekende oplossing.
Distribute the seeds on top of soil.
Verdeel de zaden over de aarde.
Furrows or holes?
Gaten of geulen?
Increasing emissions of these gases over time creates a thermal blanket capable of trapping enough heat to raise the temperature of the earth's surface.
Door de toenemende uitstoot van deze gassen wordt er op den duur een thermische deken gevormd die genoeg warmte kan vasthouden om de temperatuur op het aardoppervlak te doen stijgen.
If other company units have been made responsible for certain tasks, the reference to 'production site' below must also be interpreted accordingly.
Indien de verantwoordelijkheid voor bepaalde werkzaamheden aan andere bedrijfsunits is overgedragen, dient de verwijzing naar 'productielocatie' ook overeenkomstig te worden geïnterpreteerd.
Previous 1 2 3 4 5 ... 10 Next
Show 10 entries
Showing 11 to 20 of 100 entries
Close
Translation Direct
English (United Ki
Dutch (Netherland
Industry:
Select...
Content Type:
Select...
Data Owner:
Select...
Product Line:
Select...
Include Matrix tr
(?)
Exclude data up
my organization
Discove
Data Cloud Guide
256,040,355 (?)
Dataset Sample
View
en-GB nl-NL Consumer Electronics Instructions for Use Sony Direct 12231 988 View
en-GB nl-NL Industrial Electronics Instructions for Use Sony Direct 102529 9867 View
en-GB nl-NL Legal Services Standards, Statutes and Regulations European Parliament Direct 32931603 1285123 View
Previous 1 2 3 Next
Show 10 entries
Showing 11 to 20 of 21 entries
Data Cloud API | FAQs | Terms of Use
TAUS Data Cloud Release Version: 2.2.0
For any enquiries, please contact us: data@taus.net.
Copyright TAUS 2017

TAUS

Hi **TAUS Member** | Logout

Search | Upload | Discover & Download | Account History | My Data | Feedback | Data Cloud Guide

Account History

Membership Type:	Member
Words uploaded to TAUS Data Public Sharing:	500,000,055
Words downloaded from TAUS Data Public Sharing:	136,486,435
Current Credits:	2,363,513,840 Buy credits

Uploads | Downloads

Refresh

	Job	Created	Translation Direction			
[-]	13411	2016-11-30 00:15:18	en-US	>	es-MX	Download 350716 words
	Created by: TAUS Member Industry: Healthcare Data Owner/Source: Molina Healthcare					27338 segments
[-]	13410	2016-11-29 23:51:13	en-US	>	es-XL	Download 61248 words
	Created by: TAUS Member Industry: Industrial Manufacturing Data Owner/Source: OMAX Corporation					8777 segments
[+]	13409	2016-11-29 23:48:43	en-US	>	es-XL	Download 3196309 words
[+]	13408	2016-11-29 23:48:02	en-US	>	es-XL	Download 3324757 words

TAUS

Hi **TAUS Member** | Logout

Search | Upload | Discover & Download | Account History | My Data | Feedback | Data Cloud Guide

My Data

Translation Direction : English (United Kingdom) to Dutch (Netherlands)

Search : Dear Sir

Search

Archive | Activate

	Id	Source	Target	Product	Status
☐	6605576	Dear Sir	Geachte heer	Default	Active

1/1

Data Cloud API | FAQs | Terms of Use

TAUS Data Cloud Release Version: 2.2.0
For any enquiries, please contact us: data@taus.net.

TAUS

Hi **TAUS Member** | Logout

Search | Upload | Discover & Download | Account History | My Data | Feedback | Data Cloud Guide

Feedback

Data Owner : Select...

Product Line : Select...

Translation Direction : Select...

Rating : ☆☆☆☆☆

Comment :

Submit

Data Cloud API | FAQs | Terms of Use

TAUS Data Cloud Release Version: 2.2.0
For any enquiries, please contact us: data@taus.net.

TAUS

TAUS

TAUS