

B R U
G G E

ERFGOED
BRUGGE

Brugge Grenzeloos Digitaal: Ondersteuning meertaligheid

Eindrapport v0.1

./DATABASE

Titel	Brugge Grenzeloos Digitaal: Ondersteuning meertaligheid: eindrapport
Auteur(s)	Alexander Soetaert, Luc Truyens, Henk Vanstappen
Publisher	Datable BV, Antwerpen www.datable.be

Document history

Versie	Datum	Auteur(s)	Wijzigingen
0.1	2021-05-31	Henk Vanstappen	document setup

Inhoud

1. Management samenvatting	5
2. Inleiding	6
3. Context en doelstellingen	7
3.1. Project Grenzeloos Digitaal	7
3.2. Doelstelling	7
3.3. As is situatie	8
3.3.1. Huidige architectuur	8
3.3.2. Datasets	8
4. Strategieën	10
4.1. Manuele vertaling	10
4.1.1. Methodologie	10
4.1.2. Evaluatie	10
4.2. Meertalige Controlled Vocabularies	11
4.2.1. Use case: vertalen archeologische termen	11
4.2.2. Evaluatie	12
4.3. Machine translation (MT)	12
4.3.1. Machine translation services	13
4.3.2. Methodologie en globale beoordeling	14
4.3.3. Vergelijking vertaalservices	16
4.3.4. Impact van de lengte en nood aan context	16
4.3.5. Vergelijking Engels/Frans	16
4.3.6. Evaluatie MT services	18
4.4. Query translation met MT	18

4.4.1. Onderzoeksvraag	18
4.4.2. Methodologie	19
4.4.3. Vertaling zoektermen	20
4.4.4. Zoekopdrachten	20
4.4.5. Conclusies	21
4.4.6. Evaluatie	21
4.5. Gebruik van Glossaries	21
4.6. Named entity recognition (NER)	22
4.7. Combinatie	22
5. Besluit vertaalstrategieën	23
6. Implementatie	24
6.1. Strategieën	24
6.1.1. Ondersteuning MT door CAT	24
6.1.2. Manuele evaluatie MT	24
6.2. Technische implementatie	24
6.2.1. Opslag vertaalde data	24
6.2.1.1. Geen opslag	24
6.2.1.2. Abstract datamodel	24
6.2.1.3. In bronapplicatie	24
6.2.1.4. In aggregatielaag / Atlantis	24
6.2.1.5. Hergebruik van vertalingen over deeloplossingen heen	24
6.2.2. Ontsluiting	24
6.2.2.1. Meertalige website	24
6.2.2.2. Zoeken (in toepassing)	25
6.2.2.3. Zoekmachines (google)	25
6.2.2.4. SEO	25
7. Addenda	26
7.1. Overzicht datasets	26
7.1.1. 20210303_Archeologie_Sites_exportAtlantisErfgoedBrugge.xlsx	26
7.1.2. 20210303_Archeologie_VondstenExportAtlantisRaakvlak.xlsx	29
7.1.3. CollectieArcheologie_Wikimedia.docx	31
7.1.4. 20210303_Beeldbank_exportMemorixMaior.csv	32
7.1.5. 20210308_Beeldbank_exportAtlantis.csv	33
7.1.6. 20210302_Groeningemuseum_exportAtlantis.csv	34
7.1.7. 20210317exportkatrien_Prentenkabinet.xml	34
7.1.8. Kranten	36
7.2. Richtlijnen en kwaliteitscriteria vertaling	36
7.3. CAT en localization tools	40

7.3.1. Crowdin	40
7.3.2. OmegaT	40
7.3.3. Smartcat	40
7.3.4. SDL Trados Studio	41
7.3.5. Zanata	41
7.5. Architectuurschema (huidige situatie)	42
7.6. Technische implementatievereisten	43

Glossarium

bronapplicatie	toepassing waarin de gegevens (bv. collectiedata) worden beheerd. Voorbeelden van bronapplicaties zijn Adlib/Axiell, Memorix Maior)
brondata	data zoals bewaard in de bronapplicatie
machine translation	gebruik van software voor het vertalen van tekst of spraak
MT	→ machine translation
Named Entity Recognition	TODO
NER	→ Named Entity Recognition
query translation	vertaling van de zoekvraag (in plaats van de brontekst)

1. Management samenvatting

2. Inleiding

probleem

oorspronkelijke doelstelling: geautomatiseerd vertaling ; onderzoek leid naar combinatie van MT en beheerstools die validatie ondersteunen.

3. Context en doelstellingen

3.1. Project Grenzeloos Digitaal

Het project Grenzeloos Digitaal¹ vertrekt vanuit de vaststelling dat Erfgoed Brugge een online schatkamer is, maar moeilijk toegankelijk blijft voor wie geen Nederlands spreekt. Dankzij een subsidie van de Vlaamse Overheid wordt onderzocht hoe de Brugse erfgoedcollecties meertalig doorzoekbaar en internationaal toegankelijker kunnen worden gemaakt.

Het project is opgebouwd uit drie fasen. In de eerste fase stond het verwerven van inzicht in het bestaande en potentiële doelpubliek en hun verwachtingen centraal. Er werd een gebruikersonderzoek uitgevoerd op ErfgoedBrugge.be om de huidige gebruikers van het portaal in beeld te brengen en hun ervaringen te evalueren. Dat gebeurde aan de hand van interviews, een enquête, analyse van de gebruikersstatistieken en het zoekgedrag en live gebruikerstesten. Om het potentiële bereik van meertalige data in kaart te brengen, werden daarnaast twee tests opgezet met LearnMore.be. Met deze online-applicatie kunnen bezoekers van het Groeningemuseum en het Burgplein informatie opzoeken over schilderijen en historische gebouwen.

De huidige tweede fase (nov. 2020 – 2021) is een onderzoeks- en ontwikkelingsfase die vertrekt van de inzichten rond doelgroepen uit fase 1. Aan de hand van enkele subsets van data met internationale relevantie wordt een workflow uitgewerkt om Brugse collecties te ontsluiten via diverse internationale platformen (Wikimedia, Europeana, maar ook specifieke platformen voor de verschillende erfgoedtypes). Daarnaast wordt op basis van dezelfde datasets een methodologie uitgewerkt om meertalige doorzoekbaarheid uit te bouwen op het eigen platform. Tot slot ontwikkelt deze fase ook een concept voor een niet-Nederlandstalige en minder taalafhankelijke interface met aandacht voor de specifieke verwachtingen en achtergrond van internationale doelgroepen.

Fase 3 is de implementatiefase. De nadruk ligt in deze fase vooral op het de implementatie van een meertalige interface en meertalige doorzoekbaarheid op ErfgoedBrugge.be.

Projectpartners zijn de Musea Brugge, het Stadsarchief Brugge, Raakvlak en de Openbare Bibliotheek Brugge. Ze worden bijgestaan door de externe partners meemoo, Vlaamse Kunstcollectie en FARO. Datable ondersteunt het project in de uitwerking van methodologieën om meertalige doorzoekbaarheid uit te bouwen op het eigen platform.

3.2. Doelstelling

Om meertalige toegankelijkheid uit te bouwen op het platform van Brugge (zie verder) zijn verschillende methodologieën mogelijk om de vertaling te realiseren. Deze worden verder toegelicht in het hoofdstuk strategieën. Uitgangspunten hierbij zijn dat:

1. de vertaling kostenefficiënt gebeurt en daarom waar mogelijk wordt geautomatiseerd
2. de kwaliteit van de vertaling binnen een aanvaardbare marge moet liggen
3. de implementatie van de vertaalprocessen onderhoudsvriendelijk moet zijn
4. het bewaren en beheren van vertalingen zoveel mogelijk moet aansluiten bij de bestaande infrastructuur.

De meertalige toegankelijkheid kan toegepast worden om verschillende use cases te realiseren:

¹ <https://erfgoedbrugge.be/grenzeloos-digitaal/>

1. doorzoekbaar maken van de gegevens in de taal van de gebruiker (waarbij de resultaten niet per se in de taal van de gebruiker worden gepresenteerd)
2. presentatie van gegevens in de taal van de gebruiker

Geen van beide use cases vereist dat gegevens volledig en voorafgaand worden vertaald. Bij implementatie kan er ook voor worden gekozen om de zoekopdracht en/of de brontekst ad hoc te vertalen. De verschillende mogelijkheden worden verder besproken in het hoofdstuk Implementatie.

De doelstelling van dit onderzoek is dus vast te stellen hoe kwaliteitsvolle meertalige toegang op een kostenefficiënte en onderhoudsvriendelijke manier kan worden gerealiseerd en geïmplementeerd. Daartoe worden op basis van geselecteerde datasets verschillende vertaalstrategieën vastgesteld en onderzocht.

Als doeltaal wordt in de eerste plaats het Engels gekozen, omdat deze taal ook door vele niet-Engelssprekenden wordt begrepen. Daarnaast werd ook Frans als doeltaal onderzocht.

3.3. As is situatie

3.3.1. Huidige architectuur

De huidige architectuur bestaat uit meerdere componenten en lagen die elk instaan voor een bepaald gegevenstype en/of functie. Afhankelijk van het contenttype (kranten, foto's, objecten) worden verschillende applicaties ingezet die één of meerdere functies (digital asset management en collectiebeheer) uitvoeren.

1. Memorix Maior (Picturae) staat in voor het beheer van foto's ('Beeldbank'). Memorix bevat het beeldmateriaal zelf, met alle technische en inhoudelijke metadata die bij het beeld horen. In Memorix maior kunnen SKOS entities meertalig worden ingevoerd. SKOS entities zijn: thema's, (deel)gemeente/straat en trefwoorden.² In de huidige configuratie van Erfgoed Brugge is enkel Nederlands voorzien. Ook voor de overige velden ondersteunt deze toepassing geen meertaligheid.
2. Adlib (Axiell) wordt gebruikt voor het beheer van museale objecten. Adlib voorziet de mogelijkheid om velden meertalig te maken.
3. Atlantis (Deventit) bevat de gedigitaliseerde kranten en bijbehorende metadata. Ook de archeologische collecties ('Raakvlak') worden beheerd in een Atlantis toepassing. Atlantis voorziet eveneens in een integratielaag voor de data van de verschillende brontoepassingen. Vanuit de Atlantis integratielaag worden de verschillende content management systemen (WordPress, SceneMaker) en applicaties (Xplore app) van data voorzien. Atlantis ondersteunt momenteel nog geen meertaligheid. Dit wordt voorzien in een volgende versie van Atlantis. De verwachting is, dat deze versie in het laatste kwartaal van dit jaar beschikbaar komt.³

Een gedetailleerd schema wordt gegeven in bijlage.

² <https://maior.memorix.nl/skos/schema/index/locale/en>

³ Mededeling via e-mail van Devent-it, 19 maart 2021

3.3.2. Datasets

Onderstaand schema geeft een overzicht van de datasets die voor dit onderzoek werden gebruikt. Een gedetailleerd overzicht van de aangeleverde datasets wordt gegeven in bijlage.

naam	bestand	inhoud	aantal	opmerkingen
Groeninge	20210317exportkatrien_Prentenkabinet.xml	objectbeschrijvingen van museale collectie	9674	volgens Adlib XML schema
Beeldbank	Beeldbank exportAtlantis.csv	fotocollectie	87514	
Archeologie	20210407_Archeologie_Vondsten_Brugge	archeologische vondsten	1336	
Kranten	JournalDeBruges 1838_1878 JournalDeBruges 1898 JournalDeBruges 1919_1938_1952	historische kranten	5202	op basis van OCR, met groot aantal fouten

Met uitzondering van de kranten bevat elke dataset zowel gestructureerde als ongestructureerde data. Gestructureerde metadata betreft content die een specifiek informatietype bevat, zoals een datum, getal, coördinaat of een 'entity'. Entities zijn informatie-eenheden die betrekking hebben op een gedefinieerd concept, persoon, plaats etc. Voorbeelden zijn:

- bord
- Brugge
- Belfort (Brugge)
- Tweede Wereldoorlog

In het bronsysteem worden deze entities vaak beheerd aan de hand van een thesaurus of gecontroleerd woordenlijst.

Ongestructureerde data is afkomstig uit vrije tekstvelden zoals titel en beschrijving en bevatten meestal unieke waarden. Ongestructureerde data zijn doorgaans in 'natuurlijke taal' opgesteld, al komen er ook codes en afkortingen in voor. Enkele voorbeelden:

- "Waarschijnlijk een oord (liard) ofwel dubbele oord. Oostenrijkse periode:1711-1794. Maria Theresia: 1740-1780. Voorzijde: Buste naar rechts van Maria-Theresia. Keerzijde: AD/VSVM/BELGII/AVSTR./17.."
- "Portret van een deken van het Brugse Sint-Michielsgilde"
- "We zien de heer en mevrouw Rooseboom."
- "Drainagesysteem voor het standbeeld van Lodewijk XIV, Fonte des Statues Equestres"

Samengevat zijn de brondata verdeeld over de volgende applicaties:

	Groeninge	Beeldbank	Kranten	Raakvlak
Atlantis			x	x
Adlib	x			

Memorix		x		
Atlantis integratielaag	x	x	x	x

4. Strategieën

4.1. Manuele vertaling

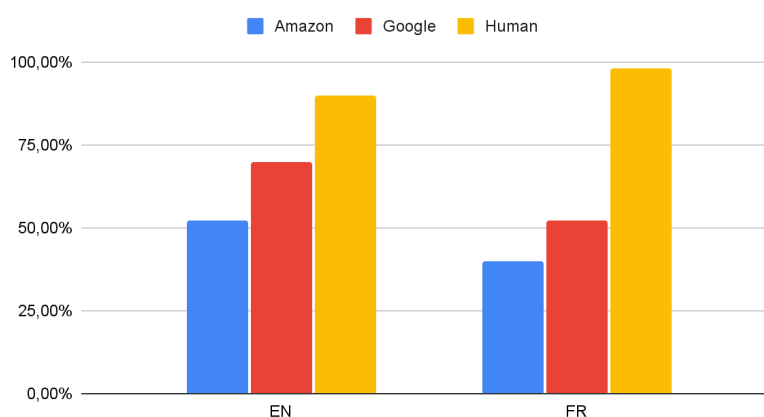
Manueel vertalen is de traditionele manier van vertalen en wordt vaak door professionele vertalers met de nodige domeinkennis uitgevoerd. Een belangrijk voordeel is de relatief hoge kwaliteit, maar daar tegenover staat een relatief hoge kostprijs.⁴ Die kost kan eventueel gedrukt worden door het inzetten van vrijwilligers (eventueel met crowdsourcing), maar zeker dan blijft een doorgedreven kwaliteitscontrole absoluut noodzakelijk. Professionele vertalers maken vaak gebruik van CAT-software (computer assisted translation, zie verder), die helpt om repetitieve vertalingen sneller te verwerken en specifieke woorden consequent te vertalen. Toch is het manueel vertalen een moeilijk te automatiseren proces.

Het vertalen van de ca. 4000 unieke titels uit de Groeninge dataset (170 000 woorden) zou 20 000 euro per taal kosten. Bovendien moet tijd worden geïnvesteerd in feedback aan de vertalers en controle. Voor manuele vertaling van de beschrijvingen (2 miljoen woorden) zou de kost van de vertaling helemaal onhaalbaar zijn.

4.1.1. Methodologie

Om de kwaliteit van manuele vertaling te meten werden de manuele Engelse en Franse vertaling van 50 titels geëvalueerd. Dezelfde titels werden ook door Google Translate en Amazon Translate vertaald en geëvalueerd. Elke vertaling werd door meerdere personen (3) beoordeeld, waarbij de beoordelaar niet wist wat de oorsprong van de vertaling was. Het resultaat van deze 'objectieve' beoordeling wordt weergegeven in onderstaande grafiek. Hieruit blijkt dat ook manuele vertalingen niet steeds correct worden uitgevoerd: voor vertaling van de Engelse titels bleek in 90% van de gevallen niet correct. Anderzijds zijn manuele vertalingen qua grammatica en zinsbouw betere vertalingen. De vertaling is vaak vrijer en daarom vlotter. Zinnen worden bijvoorbeeld samengevoegd/gesplitst waar dat leidt tot betere vertalingen. De vrijere vertalingen leiden ook wel tot de vaststelling dat de vertaling niet correct is. Ze bevatten ook tikfouten (bv. "vergetable" i.p.v. "vegetable").

MT vs. Human



Grafiek Vergelijking tussen kwaliteit menselijke vertaling en machine Translation (n=50)

⁴ Marktprijzen variëren tussen 6 en 15 cent per woord. Voor specialistische vertalingen worden zelfs bedragen tot 25 cent per woord aangerekend

4.1.2. Evaluatie

Samenvattend kunnen we stellen dat manuele vertaling de beste kwaliteit oplevert, maar tegen een kostprijs die niet haalbaar is binnen de voorziene budgetten van het project Grenzeloos Digitaal. Ook bij deze vorm van vertalen moet rekening worden gehouden met validatie/nazicht en feedback voor de vertalers over specifieke (lokale) terminologie. Omdat de aangeleverde vertalingen correct zijn, is de tijdsinvestering hier wel kleiner dan wanneer de automatische vertalingen moeten worden nagelezen en gecorrigeerd.

4.2. Meertalige Controlled Vocabularies

Meertalige controlled vocabularies zoals de Art & Architecture Thesaurus (AAT)⁵ beschrijven concepten in een gestructureerde vorm. Aan elk concept worden termen gekoppeld in verschillende talen, met onderscheid tussen voorkeurs- en niet voorkeursvormen. In principe kunnen deze vocabularies net als glossaries gebruikt worden om domeinspecifieke termen te vertalen. Daarbij wordt de te vertalen woord(combinatie) opgezocht in de authority, waarna de string in de doeltaal wordt opgehaald. Deze aanpak lijkt dus in de eerste plaats geschikt voor het vertalen van erfgoedconcepten als objectnamen, materialen en technieken. Ook plaatsnamen en namen van personen kunnen in een bepaalde taalvariant worden vertaald.⁶ Autoriteiten zoals Wikidata of CONA bevatten vertalingen van titels van kunstwerken, waarop eventueel een beroep kan worden gedaan. Over het algemeen betreft het goed onderhouden en kwaliteitsvolle bronnen en vertalingen, waardoor werd verwacht dat de betrouwbaarheid van het resultaat toeneemt ten opzicht van niet-domeinspecifieke services. Aangezien veel controlled vocabularies binnen het erfgoedveld via een API aanspreekbaar zijn, is dit proces relatief eenvoudig te automatiseren.

	erfgoedconcepten	geografie	kunstwerken
Art & Architecture Thesaurus (AAT)	x		
Thesaurus of Geographic Names (TGN)		x	
Wikidata	x	x	x
Cultural Objects Names Authority (CONA)			x

Tabel: Controlled Vocabularies voor vertaling van erfgoedconcepten, geografische termen, titels van kunstwerken

4.2.1. Use case: vertalen archeologische termen

Voor dit onderzoek maakten we gebruik van drie sets archeologische termen, waarvoor een gevalideerde Engelse vertaling beschikbaar was.⁷ Het betrof 110 objectnamen, 45 materiaaltermen en 7 termen die een periode aanduiden. De lijst bevatte weinig gespecialiseerde terminologie (hamer, kanonbal, kom) waarvan voor de meeste termen een overeenkomstig concept in de AAT te vinden is.

Elk van de termen werd geautomatiseerd gematched met een term uit de AAT. Er werd steeds gezocht op een exact match van de één van de Nederlandstalige termvarianten die aan het AAT-concept zijn gekoppeld. Meerdere matches werden niet weerhouden. Vervolgens werd de Engelse term opgehaald en vergeleken met de gevalideerde Engelse vertaling. Deze beoordeling gebeurde manueel, om aanvaardbare afwijkingen (bv.

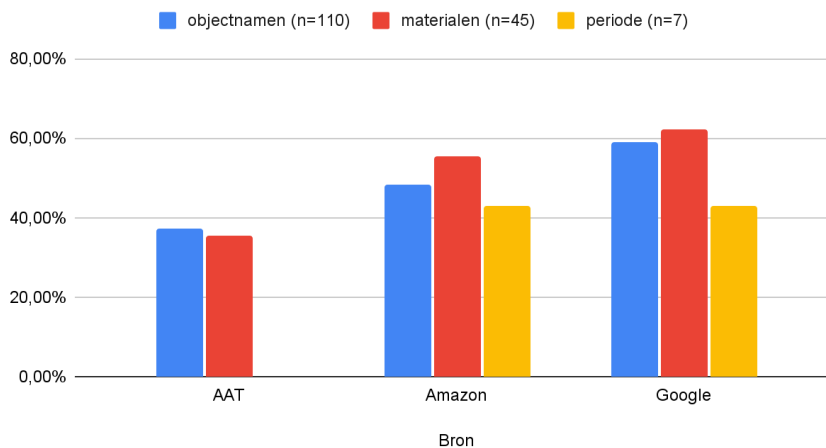
⁵ <https://www.getty.edu/research/tools/vocabularies/aat/>

⁶ Aangezien het voorkomen van taalvarianten van persoonsnamen zeer beperkt is, werd dit type entity niet verder onderzocht.

⁷ Achtergrond: bij upload van vondsten naar Wikidata (een onderdeel van WP1 van het project Grenzeloos digitaal) werd een lijst van Engelse vertalingen van voorwerpen, periodes en materialen samengesteld.

de toevoeging van een qualifier 'materiaal') als geldige vertaling aan te merken. Ter vergelijking lieten we dezelfde termen ook vertalen door Google Translate en Amazon. Het resultaat wordt weergegeven in onderstaande grafiek.

Controlled Vocabularies vs. Machine Translation



Grafiek Vergelijking vertaling door controlled vocabulary en machine translation

Uit deze test blijkt dat AAT onverwacht slecht scoort: minder dan 40% van de objectnamen werden correct herkend en vertaald. De 7 termen die een periode aangeven werden door AAT niet herkend. De resultaten zouden verbeterd kunnen worden door de zoekopdrachten te verfijnen (bv. door materiaaltermen enkel op te zoeken in het Materials facet van de AAT). Het impliciete voordeel dat gebruik wordt gemaakt van een domeinspecifieke bron, blijkt tegelijk een nadeel: voor elk gegevenstype moet een andere bron worden geselecteerd en geconfigureerd. Het gebruik van controlled vocabularies vereist dus een gedegen voorkennis van deze bronnen zelf én van de manieren om de verschillende API's te bevragen. Naarmate de zoekvraag verder wordt verfijnd, verbeteren de resultaten weliswaar, maar verhogen ook de inspanningen om die resultaten te verkrijgen.

Een tweede set termen bestond uit concepten uit de Groeninge dataset. Hiervoor was geen gevalideerde Engelse vertaling beschikbaar. Gegevens moesten bewerkt worden voor ze aan AAT konden worden aangeboden, bv. het splitsen van termen, verwijderen van qualifiers, etc. De kwaliteit van deze vertaling bleek na een korte evaluatie eveneens ondermaats en werd niet verder geanalyseerd.

4.2.2. Evaluatie

Het gebruik van controlled vocabularies als AAT in functie van een geautomatiseerde vertaling is weinig efficiënt. In het geval gegevens al eerder manueel met een controlled vocabulary gereconcilieerd werden, is een meertalige controlled vocabulary natuurlijk wel een zeer waardevol hulpmiddel.

4.3. Machine translation (MT)

Machinevertaling (MT) maakt gebruik van software voor het vertalen van tekst of spraak van de ene taal naar de andere. Daarbij kan gebruik worden gemaakt van verschillende methodologieën.

- Statistische machinevertaling (SMT) maakt gebruik van statistische modellen die zijn gebaseerd op de analyse van grote hoeveelheden tweetalige tekst. Hierbij worden overeenkomsten tussen

woorden uit de bron- en doelttekst bepaald. Het grootste nadeel is dat deze methode weinig of geen rekening houdt met de context.⁸

- Rule-based machinevertaling (RBMT) daarentegen vertaalt op basis van grammaticale regels. Het voert een grammaticale analyse uit van de brontaal en de doeltaal om de vertaalde zin te genereren. Vaak wordt RBMT gecombineerd met SMT (Hybride machinevertaling of HMT).
- De meeste MT systemen maken tegenwoordig echter gebruik van Neurale machinevertaling (NMT). NMT is een type machinevertaling dat afhankelijk is van neurale netwerkmodellen (gebaseerd op het menselijk brein) om statistische modellen voor vertaling te ontwikkelen. Bekende voorbeelden zijn Google Translate en Deepl.⁹

4.3.1. Machine translation services

Voor dit onderzoek maakten we gebruik van twee services, namelijk Google Translate en Amazon Translate. Wellicht de meest bekende service is Google Translate.¹⁰ Google Translate is een vertaalservice die in april 2006 door Google is ontwikkeld. Oorspronkelijk werd Google Translate uitgebracht als een statistische machinevertaalservice. In november 2016 heeft Google zijn vertaalmethode overgezet naar een systeem dat gebruik maakt van van deep learningtechnieken, waardoor de kwaliteit van de vertalingen significant is toegenomen.

De service van Google ondersteunt meerdere applicaties, waaronder een grafische interface,¹¹ een Android app etc. De Translation API is relatief eenvoudig op te zetten en te implementeren in eigen applicaties. Deze service kost \$ 20 per miljoen tekens, maar de eerste 500.000 tekens per maand worden niet aangerekend. De AutoML Translation laat ook toe eigen modellen te ontwikkelen. Deze modellen worden getraind met language pairs, waarvan de kostprijs per uur wordt verrekend (\$ 45 per uur). De vereiste tijd hangt af van het aantal training pairs en complexiteit van de data. Voor 10.000 pairs is naar schatting 3 tot 4 uur vereist. Het vertalen zelf verloopt aan een vergelijkbare kost als de Translation API.

Omdat de beschikbare data onvoldoende trainingsmateriaal bevatten en er nog geen glossaries beschikbaar zijn, werd enkel de Google Translate API getest.

De tweede service in de test is Amazon Translate. Amazon Translate is sinds 2018 beschikbaar en sterk vergelijkbaar met Google Translate: de service is eveneens gebaseerd op deep learning en beschikt over grafische interface en een API waarmee verschillende toepassingen kunnen worden ondersteund, zoals het in real time vertalen van websites. Het belangrijkste verschil is dat de services van Amazon enkel beschikbaar zijn voor wie een Amazon AWS account heeft.

De prijszetting van Amazon Translate is vergelijkbaar met die van Google: maandelijks kunnen 2 miljoen karakters gratis vertaald worden, vervolgens wordt \$ 15 per miljoen karakters aangerekend.

Andere services zijn bv. Bing,¹² IBM Watson Language Translator,¹³ Deepl¹⁴ en Lingea¹⁵ Ook deze services zijn vergelijkbaar in de zin dat ze gelijkaardige services aanbieden en beperkt gratis te gebruiken zijn.¹⁶

⁸ Een voorbeeld hiervan is Moses (<http://statmt.org/moses/>)

⁹ <https://techpulse.be/premium/7629/hoe-werkt-google-translate-vertaalmachine/>

¹⁰ <https://cloud.google.com/translate/docs>

¹¹ <https://translate.google.com/>

¹² <https://www.bing.com/translator>

¹³ <https://www.ibm.com/watson/services/language-translator/>

¹⁴ <https://www.deepl.com/translator>

¹⁵ <https://translator.lingea.com/>

¹⁶ Voor een uitgebreide vergelijking, zie o.a. https://en.wikipedia.org/wiki/Comparison_of_machine_translation_applications

Een uitzondering in de reeks vertaalservices is eTranslation (EC). EC is kosteloos een online vertaaldienst van de Europese Commissie (EC). De dienst werd in de eerste plaats ontwikkeld voor teksten over EU-gerelateerde zaken, maar er wordt ook gewerkt aan een model dat kan worden gebruikt voor het vertalen van data met betrekking tot erfgoedcollecties.¹⁷ Tot nu toe is het domein cultuur nog enkel beschikbaar voor vertaling Engels - Duits. De uitbreiding van deze service voor erfgoedcollecties is één van de doelstellingen van het project Europeana Translate (2021-2023).¹⁸

De services werden getest met een set 'corner cases': Nederlandstalige fragmenten met moeilijk te vertalen onderdelen, zoals:

- eigennamen die mogelijk foutief vertaald kunnen worden ('Frans Francken de Jongere en Simon de Vos')
- grammaticale fouten in de brontekst ('Op de achtergrond een het lot van de zoon:')
- meertalige brontekst ('Figuratieve kaart voor de afwerking van het Schipdonkkanaal, Plan figuratif des deux projets présentés au gouvernement pour l'achèvement du canal de Schipdonck')
- Bijbelse namen en concepten ('Geboorte van Maria en de ontmoeting van Anna en Joachim bij de Gouden Poort')
- woorden met onduidelijke/verschillende betekenissen ('Zittende man die een pleister van zijn rechterhand trekt')

Op basis van deze test werden de best scorende services geselecteerd, nl. Google Translate en Amazon Translate.

4.3.2. Methodologie en globale beoordeling

Elke service werd geëvalueerd aan de hand van een steekproef (n=50) van elke dataset. Voor de evaluatie zijn in principe meerdere methodes mogelijk¹⁹:

1. In een Round-trip translation methode wordt de tekst in doeltaal weer vertaald naar de brontaal, en vergeleken met de oorspronkelijke tekst. Het nadeel is dat er daarbij twee vertaalservices worden beoordeeld: van brontaal naar doeltaal en omgekeerd, terwijl alleen de eerste vertaling relevant is.
2. Wanneer meerdere services worden gebruikt kan worden verondersteld dat het resultaat correct is wanneer alle services tot een gelijk of gelijkaardige resultaten komen. De mate van gelijkheid kan worden bepaald door de similarity tussen de resultaten van een aantal services (op basis van 5-gram overlap). De methode bleek echter onvoldoende betrouwbaar bij het relatief laag aantal services, aantal bronteksten en hun lengte.
3. Automatische evaluatie maakt gebruik van diverse statistische methodes om het verschil tussen een vertaling door een mens en die van een machine te meten.²⁰ Deze statistische methodes vergelijken de vertaalde tekst met een referentietekst waarvan wordt verondersteld dat het een correcte vertaling is ('ground truth'). Een nadeel van deze methode is dat er geen rekening wordt gehouden met de mogelijkheid dat een zin op meerdere geldige manieren kan worden vertaald. Bovendien was er slechts in beperkte mate een gevalideerde vertaling beschikbaar.

¹⁷ <https://pro.europeana.eu/post/approaches-for-handling-multilingualism-for-digital-cultural-heritage>

¹⁸ <https://pro.europeana.eu/project/europeana-translate>

¹⁹ https://en.wikipedia.org/wiki/Evaluation_of_machine_translation

²⁰ https://en.wikipedia.org/wiki/Evaluation_of_machine_translation#Automatic_evaluation

4. De vertalingen werden telkens als geheel manueel beoordeeld: alle vertalingen die minstens één fout bevatten, werden in hun geheel als fout aangemerkt. Daarbij worden verschillende parameters beoordeeld:
 - a. grammaticale correctheid van de vertaalde tekst
 - b. inhoudelijk correctheid en volledigheid van de vertaling
 - c. consistentie tussen verschillende vertalingen

We selecteerden uiteindelijk methode 4 om de resultaten te evalueren. Een nadeel van deze methode is dat deze de meest strikte is: enkel een volledig correcte vertaling wordt weerhouden, of met andere woorden: de kleinste vergissing wordt genadeloos afgestraft.

Ondanks deze strenge beoordelingsmethode werden de resultaten van de Engelse vertalingen werden algemeen als redelijk goed beoordeeld. De problemen waren van uiteenlopende aard:

- Bij aanduidingen van personen op een foto (bv. "van links naar rechts", "v.l.n.r.", "op de achterste rij") worden veel fouten gemaakt die vertalingen niet acceptabel maken.
- Een aantal termen komt in de dataset heel vaak voor: bijvoorbeeld "jubilaris", "opname" (van een foto), "aanlegsteiger" worden altijd fout vertaald. Als deze 3 termen consequent correct zouden worden vertaald, zou de score voor Google aanzienlijk hoger liggen. Dit wijst er op dat een vertaalservice gebaat zou zijn met het gebruik van een glossary.
- Inconsequenties: "in de Sint-Clarastraat" wordt door Amazon vertaald als "à Sint-Clarastraat", maar "in de Steenstraat" als "dans la Steenstraat." "Gistfabriek" wordt soms vertaald als "yeast factory", soms niet vertaald. Ook hier zou een Glossary kunnen helpen om inconsequenties te vermijden.
- Beide services hebben het moeilijk met aanhalingstekens (vooral bij Franse vertalingen): enkele en dubbele aanhalingstekens worden door elkaar gebruikt binnen een vertaling, soms weggelaten of verplaatst.
- Sommige vertalingen blijken correct, maar niet optimaal: "hoofdman" wordt vertaald als "headman", maar in de specifieke context van de beschrijving zou "dean" correcter zijn.
- Bij meertalige bronteksten vertalen de services slechts één taal en behouden de rest van de brontekst in de oorspronkelijke taal. Voorbeeld: "Figuratieve kaart voor de afwerking van het Schipdonkkanaal, Plan figuratif des deux projets présentés au gouvernement pour l'achèvement du canal de Schipdonck"
- Dialectwoorden en -uitdrukkingen zorgen voor misinterpretatie, bv. "doorgaan" (in de betekenis van "plaatsvinden") wordt vertaald als "continue".
- In een aantal titels komen stukken oud Nederlands voor. Deze worden door de services gedeeltelijk vertaald, maar zouden onvertaald moeten blijven.
- Eigennamen worden niet vertaald, ook wanneer dit wel vereist is, bv. "Geboorte van Maria" wordt vertaald als "Birth of Maria" (niet "Mary"). Andere namen worden dan weer wél vertaald, wanneer dat niet vereist is, zoals in "Achiël of Ackerplein."
- Wanneer de brontekst zelf ambigu is, voegt de vertaalservice vaak een (gewenste of ongewenste correctie toe: "Vier vissen: Acipencer, Acipenser Zeelandicus, Fundulus, Carpio, Vissen, Piscium Vivae Icones" wordt vertaald als "Four fishes: Acipencer, Acipenser Zeelandicus, Fundulus, Carpio, Pisces, Piscium Vivae Icones." ("Vissen" wordt hier dus naar het Latijnse "Pisces" vertaald.)

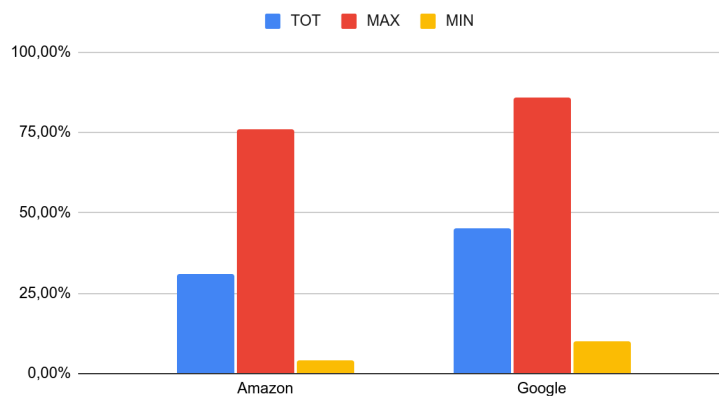
4.3.3. Vergelijking vertaalservices

Tussen de twee vertaalservices die in deze test werden opgenomen, bleken aanzienlijke kwaliteitsverschillen te bestaan. Google scoorde aanmerkelijk beter. Bovendien merkten de uitvoerders van de evaluatie op dat de fouten van Google meestal minder groot waren. Amazon vaker problemen met interpretatiefouten, zoals de

functie van "deken" die foutief werd vertaald als "blanket", of de vertaling van "een voornaam man" als "a first name man."

De matige score van Amazon werd nog vergroot omdat deze service regelinden interpreteert als het eind evan een zin, waardoor de brontekst vaak fout werd geïnterpreteerd. Google gaat blijkbaar intelligenter om met regelinden en negeert deze om tot een zinvolle brontekst te komen.

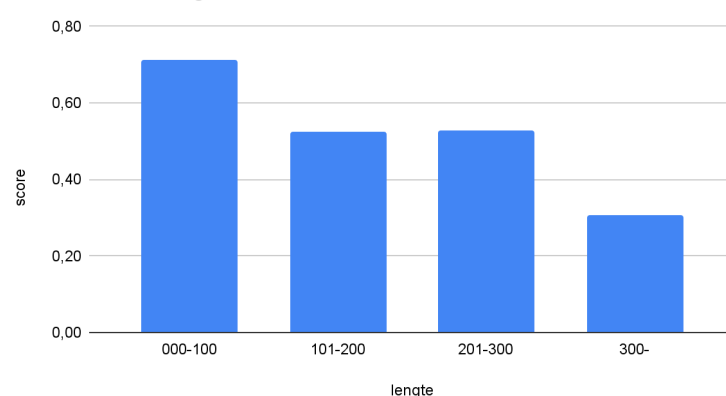
Amazon vs. Google



4.3.4. Impact van de lengte en nood aan context

Lengte van de brontekst kan op twee manieren impact op het resultaat hebben: enerzijds vergroot de kans op fouten evident wanneer de tekst langer is, maar anderzijds zorgt een langere tekst mogelijk voor meer context en daardoor voor een meer accurate vertaling. Onderstaande diagram toont het verband tussen de lengte van alle unieke Engelstalige titels en en beschrijvingen en hun gemiddelde score. Korte tekstfragmenten blijken uiteindelijk minder foutgevoelig dan langere. Machine translation zou volgens deze vaststelling dus meer geschikt zijn voor (doorgaans kortere) titels dan voor beschrijvingen.

score versus lengte



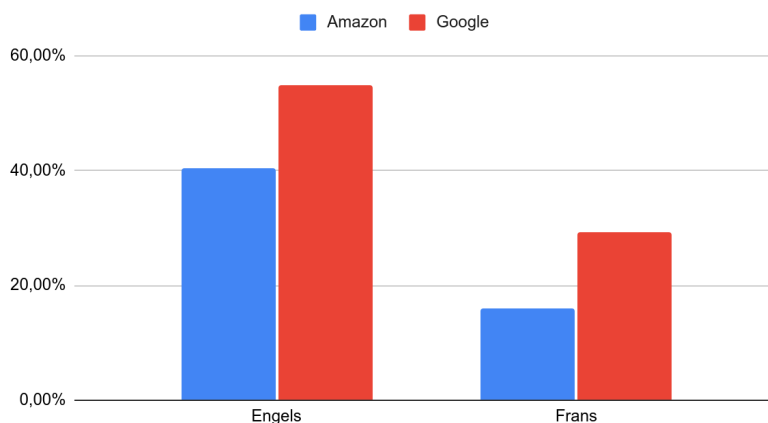
4.3.5. Vergelijking Engels/Frans

De kwaliteit van de vertalingen van Nederlands naar Frans bleek bij zowel Google als Amazon van veel mindere kwaliteit dan de vertalingen naar het Engels. De reden daarvoor is dat er gebruik wordt gemaakt van indirecte vertaling: de brontekst wordt naar het Engels, vervolgens uit het Engels naar het Frans vertaald. De meeste vertaalservices hebben immers slechts een beperkt aantal taalparen: om directe vertaling van elke

brontaal naar elke doeltaal mogelijk te maken, zou immers een enorm aantal vertaalcombinaties moeten worden gecreëerd en onderhouden. Het ondersteunen van directe vertaling tussen 100^2 talen zou 10.000 combinaties vereisen. Bovendien zijn er voor veel taalcombinaties onvoldoende tekstparen beschikbaar om een model afdoende te trainen.

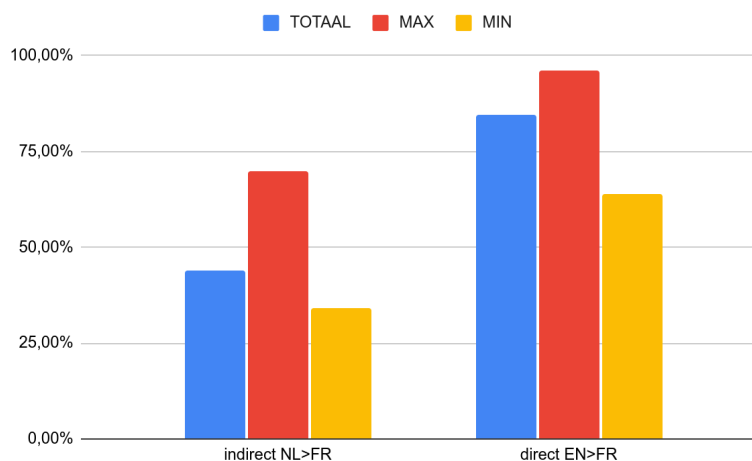
De matige kwaliteit van Nederlands naar Franse vertaling uit zich op verschillende manieren: enerzijds zijn er interpretatiefouten die leiden tot betekenisveranderingen, zoals "gezelschap" dat via de Engelse tussenvertaling "company" werd vertaald als "entreprise". Betere of meer correcte vertalingen worden niet gegeven, bv. "titelprent" wordt via "title print" vertaald als "impression de titre" in plaats van "frontispice". In veel gevallen bleven er Engelse woorden onvertaald in de Franse vertaling, bv. "Achter de schermen van het Heilig-Bloedspel" werd: "Dans les coulisses du Holy Blood Game."

Engels vs. Frans



Globale kwaliteit vertaling naar Engels, resp. Frans

In een tweede fase werd daarom een vertaling gemaakt van gecorrigeerde Engelse vertalingen naar het Frans (enkel met de Google vertaalservice). Voor een aantal datasets (titels en beschrijvingen van de Beeldbank en titels van Groeningemuseum) leidde dit tot significant betere en aanvaardbare resultaten. In andere gevallen (beschrijvingen van het Groeningemuseum) was de kwaliteit ook merkbaar beter, maar bleef de foutmarge via deze methode nog te groot (35%).



Vergelijk indirecte en directe vertaling naar Frans

4.3.6. Evaluatie MT services

- goed, niet goed genoeg
- snel, goedkoop
- verschillen in taal, service
- relatief eenvoudige implementatie (as a service)
- fine tuning door creëren van eigen modellen relatief complex en arbeidsintensief (glossary samenstellen, trainen)

Kwaliteit Engels doorgaans beperkt

Franse vertaling onvoldoende

Waardenlijsten:

- manueel vertalen
- manueel reconciliëren met multilingual authority

Pre- en postprocessing vereist:

- regeleinden
- codering diakriten

glossarium

Ondersteunen manuele controle

- glossarium
- NER?
-

Zoeken vs representatie?

Presentatie:

"Over het algemeen is de kwaliteit te slecht om de automatische vertalingen publiek te maken op een collectieplatform."

Zoeken:

"[Overwegen] om de automatische vertalingen te gebruiken als basis voor meertalige doorzoekbaarheid. Daarbij moet worden afgewogen welke meerwaarde deze functionaliteit heeft t.o.v. een wellicht hoger aantal foutieve zoekresultaten en het feit dat de gebruiker niet alle doorzoekbare metadata zal kunnen zien."

"75% van de zoekopdrachten hebben betrekking op eigennamen, onderwerpen, gebeurtenissen of plaatsnamen binnen Brugge"

Fouten in vertaling named entities

Duke of Alva (Duke of Alba)

Ahasverus (Ahasuerus)

Achiel of Ackerplein

4.4. Query translation met MT

Bij query translation wordt niet de brontekst vertaald, maar de zoekopdracht van een eindgebruiker. Omdat deze vertaling onmiddellijk moet gebeuren en de zoekopdracht vooraf niet bekend is, kan dit enkel worden gerealiseerd met machine translation. Query translation is vooral geschikt wanneer het voorafgaand vertalen van de brontekst niet haalbaar is, bijvoorbeeld vanwege de omvang van de brontekst.

De haalbaarheid van query translation werd uitgetest op het corpus van Franstalige kranten. Deze krantencollectie was via OCR full text doorzoekbaar gemaakt, maar omdat de kranten gedeeltelijk Franstalig zijn, leveren Nederlandstalige zoektermen geen of onvoldoende resultaat. Met dit onderzoek wilden we te weten komen of zoekresultaten verbeterd kunnen worden door de zoektermen geautomatiseerd te vertalen. De vraag was dus enerzijds of query translation meer resultaten oplevert (grotere 'vangst'), anderzijds of dit een impact heeft op de kwaliteit in de vorm van ongewenste zoekresultaten (ruis).

4.4.1. Methodologie

Voor dit onderzoek werd gebruik gemaakt van een corpus gedigitaliseerde en ge-OCRde kranten.²¹ In totaal bevat dit corpus ruim 5000 pagina's met 17 miljoen woorden. Het betreft Franstalige kranten, maar een gedeelte van de niet-redactionele inhoud is Nederlandstalig (o.a. advertenties en aankondigingen).

Voor dit onderzoek werd gebruik gemaakt van een corpus zoektermen dat werd verzameld in de loop van november en december 2020. Deze zoektermen werden getagged naar type (NAAM, PLAATS, ONDERWERP) en taal (NL, FR, EN, ...). Daaruit bleek dat bij minder dan de helft van de zoekopdrachten gezocht wordt op onderwerp. Enkel deze onderwerpst termen werden in het onderzoek gebruikt.²² Hieruit werd een random sample van 200 unieke termen gehaald.

De zoektermen werden vertaald d.m.v. de Google Translate service. Van de 200 zoektermen werden er 139 als correcte vertaling beoordeeld. Alle foute vertalingen (bv. "peerdenrennen", vertaald als "course de poire"), gedeeltelijk vertaalde en onvertaalde termen (bv. "ketsekot") werden als incorrect beoordeeld.

De oorspronkelijke en vertaalde zoektermen werden opgezocht in het corpus. We gingen er hierbij van uit dat een correct vertaalde zoekterm een valide zoekresultaat oplevert en omgekeerd. De bijna 70.000 zoekresultaten zelf werden als zodanig niet gevalideerd.

- 24 (12%) van de Nederlandse zoektermen leiden tot een resultaat, met 2.646 hits. Het merendeel van deze resultaten was niet relevant.
- 49 (24,5%) van de Franstalige zoektermen leverde een resultaat op, met bijna 70.000 hits. Ook hier mag verwacht worden dat een deel van de zoekresultaten niet relevant is, omdat deze zoektermen ook incorrect vertaalde termen bevat.
- 35 (17.5%) van de correct vertaalde zoektermen leverde een resultaat op, met in totaal 28840 hits.

	NL	FR (valid)	FR
zoekvragen	200	139	200
aantal beantwoorde zoekvragen	24	35	49
aantal matches	2.646	28.840	68.591

Tabel: samenvatting respons zoekvragen

²¹ Journal de Bruges jg. 1838, 1878, 1898, 1919, 1938, 1952

²² Zoeken op eigennamen is in de meeste gevallen taalonafhankelijk. Dit geldt in mindere mate voor plaatsnamen (cf Diksmuide vs. Dixmude).

4.4.2. Gebruikersbeoordeling

Tijdens dit onderzoek werden twee gebruikers van de historische krantendatabase gevraagd naar hun verwachtingen met betrekking tot oplossingen om Franstalige kranten beter doorzoekbaar te maken. De eerste gebruiker was al gewend aan het opzoeken in meerdere talen. Hij verwachtte dat dit mogelijk een voordeel kon zijn, maar wilde graag controle houden: een gebruiker zou de vertaaloptie moeten kunnen aan- en uitschakelen, of de vertaling zou als zoeksuggestie moeten worden aangeboden. De tweede gebruiker had niet eerder stilgestaan bij het feit dat er veel Franstalige pagina's zijn. Hij verwachtte dat dit een meerwaarde zou bieden bij het doorzoeken van krantenpagina's. De fouten in zoekresultaten wegen voor hem niet op tegen dit het voordeel om een ruimere zoekactie te kunnen uitvoeren. Net als gebruiker 1 verkoos ook deze gebruiker dat zoektermen worden voorgesteld door de zoekfunctie.

4.4.3. Evaluatie

Hoewel de kwaliteit van de zoekresultaten als zodanig niet werd geëvalueerd, kunnen we veronderstellen dat query translation op basis van machine translation een aanzienlijk hogere vangst oplevert. Keerzijde is dat ook de ruis aanzienlijk groter is.

Ondanks de hoge 'ruis' biedt een geautomatiseerde vertaling van de zoektermen een meerwaarde bij het zoeken door een corpus Franstalige kranten.

Het is belangrijk om de eindgebruiker te informeren over het feit dat de zoekopdracht geautomatiseerd werd vertaald, zodat de - soms verrassende - resultaten geduid kunnen worden. Om de eindgebruiker ook enige controle te geven op de zoekopdracht, kan in de zoekomgeving de keuze gegeven worden om de zoektermen naar het Frans te vertalen (en eventueel manueel aan te passen).

Bij implementatie moet ook rekening gehouden worden met de visie dat alle content van het Erfgoed Brugge platform integraal doorzoekbaar moet zijn.

4.5. Glossaries

Een glossary (woordenlijst) is een aangepast woordenboek dat wordt gebruikt om de domeinspecifieke terminologie consistent te vertalen. Glossaries ondersteunen een betere vertaling van:

1. eigennamen (plaatsnamen, namen van gebouwen, gebeurtenissen en personen): "hertog van Alva" wordt bv. vertaald als "duke of Alba" (niet "duke of Alva")
2. niet te vertalen woorden: "Heilig Bloedspel" wordt niet vertaald als "Holy Blood game"
3. dubbelzinnige woorden: het woord "glas" kan bijvoorbeeld een materiaal of een drinkglas betekenen. Een domeinspecifieke glossary kan er voor zorgen dat de juiste keuze wordt gemaakt.

Het creëren van een glossary is een manueel proces, maar het resultaat kan worden gebruikt om gestructureerde data snel te vertalen. Ze kunnen als zelfstandig instrument worden gebruikt voor het vertalen van termen. Binnen het project Grenzeloos Digitaal werd bijvoorbeeld een glossary opgesteld om archeologische termen te vertalen. De glossary werd ook opgeladen in Wikidata, waardoor archeologische vondsten nu meertalig beschikbaar zijn.²³

Glossaries kunnen ook worden gebruikt als onderdeel van andere vertaalstrategieën. Zo worden ze ingezet in CAT systemen (zie verder), waarmee de consistentie van de vertaling wordt ondersteund. Ook machine translation (zie verder) maakt soms gebruik van glossaries.

²³ <https://erfgoedbrugge.be/grenzeloos-digitaal/>

4.5.1. Named Entity Recognition

Het opstellen van een glossary is een manueel proces. Het kan worden ondersteund door het gebruik van named entity recognition (NER). NER is automatisch herkennen van entities (persoonsnamen, plaatsnamen e.d.) in ongestructureerde tekst. Op die manier worden onderdelen van vrije tekst verwerkt tot gestructureerde tekst en kunnen ze gericht worden vertaald.

Een voorbeeld: "de kerk van het Engels Klooster." moet worden vertaald als "the church of the Engels Klooster", niet "English Monastery". "Engels Klooster" kan d.m.v. NER worden geïdentificeerd waarna de correcte vertaling aan een glossary wordt toegevoegd.

Als test maakten we gebruik van Frog,²⁴ een open source NER tool die geschikt is voor Nederlandstalige teksten. In een testtekst van 569 woorden werden handmatig 91 entities geïdentificeerd. Frog identificeerde daarvan 71 (88%) entities. Frog klasseerde bovendien 81 van de herkende entities correct (als persoon, organisatie, gebeurtenis, locatie, plaatsnaam).

4.5.2. Evaluatie

Het onderzoek voor de dataset archeologische vondsten (zie strategie Meertalige Controlled Vocabularies) toont aan dat het manueel opstellen van een glossary voor specifieke termen (bv. voorwerp, materiaal, periodes) meestal efficiënter is en een betere kwaliteit oplevert. Anders gesteld: voor vertaling van gecontroleerde woordenlijsten binnen een bronsysteem (bijvoorbeeld objectnamen) is een manuele aanpak doorgaans de beste.

4.6. Combinatie van strategieën

Een laatste strategie die we hier willen voorstellen is een combinatie van bovenstaande strategieën, waarbij gebruik wordt gemaakt van machine translation, die door een mens wordt gevalideerd en waar nodig gecontroleerd. De vertaling wordt daarbij ondersteund door één of meerdere glossaries. Deze glossaries kunnen handmatig worden samengesteld of gecreëerd op basis van NER.

Deze *Human in the Loop* aanpak heeft als voordeel dat machine translation en glossaries het menselijke werk grotendeels automatiseren en versnellen. Menselijke controle blijft echter mogelijk om de kwaliteit van de vertaling op het gewenste niveau te brengen. Het samenbrengen van deze verschillende strategieën in één workflow is complex, maar kan worden gecoördineerd door een geschikt Computer-aided translation systeem.

4.6.1. Computer-aided translation tools

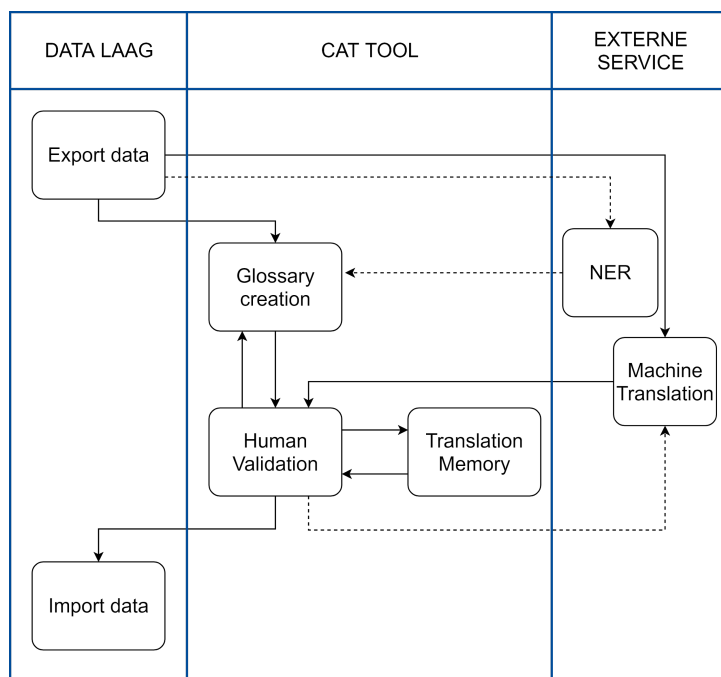
Computer-aided translation (CAT) is het gebruik van software om een menselijke vertaler te helpen bij het vertaalproces. De vertaling wordt gemaakt door een mens en bepaalde aspecten van het proces worden gefaciliteerd door software.

Onder CAT-tools worden doorgaans programma's verstaan die specifiek het eigenlijke vertaalproces vergemakkelijken. Belangrijke functionaliteiten zijn onder andere:

- een geheugen, waarin eerdere vertalingen worden bewaard en hergebruikt kunnen worden.
- integratie van glossaries, waarmee een vertaling kan worden aangestuurd.
- de mogelijkheid te integreren met externe services, zoals machine translation services die via de API worden gekoppeld.

²⁴ <http://ilk.uvt.nl/frog/>

- CAT wordt vaak gebruikt om *localization files* van software te vertalen.²⁵ Localization files zijn de bestanden waarin labels, helpteksten e.d. voor elke taalversie van een software worden bewaard. Om die reden ondersteunen CAT tools een groot aantal bestandsformaten.
- samenwerking en coördinatie wordt mogelijk gemaakt door een webbased interface, met mogelijkheid om de vorderingen te monitoren en vertalingen te valideren door domeinspecialisten en verschillende doeltalen te ondersteunen. Dit laat bijvoorbeeld ook toe meerdere vrijwilligers aan het werk te zetten en op te volgen



Bovenstaand diagram geeft het proces vereenvoudigd weer:

1. Gegevens worden geëxporteerd uit het de data laag van het bronsysteem. Dit kan een applicatie voor collectiebeheer zijn, of een data laag die gegevens uit verschillende applicaties aggregeert (zie hiervoor het onderdeel Implementatie).
2. Uit de gegevens worden waar mogelijk vooraf Glossaries gecreëerd, eventueel met ondersteuning van NER service.
3. De gegevens worden vertaald door Machine Translation service.
4. De machinevertalingen worden gevalideerd en gecorrigeerd door een gebruiker, hierbij ondersteund door de glossary.
5. Gevalideerde vertalingen worden opgeslagen in het vertaalgeheugen, dat daarmee de validatie verder ondersteunt. In een geavanceerde vorm kunnen gevalideerde vertalingen gebruikt worden om de modellen die de MT-service gebruikt verder te trainen.
6. Het resultaat wordt opgeslagen in de data laag.

4.6.2. Evaluatie

Voor dit project maakten we een korte marktverkenning van CAT tools die de nodige functionaliteiten hebben. Een meer uitgebreide test van de CAT tool Crowdin toonde de werkbaarheid van deze oplossing aan:

²⁵ In dit geval wordt gesproken van Localization tools, die specifiek geschikt zijn om formaten van localization files te lezen en schrijven.

hoewel de efficiënt moeilijk gemeten kan worden zonder uitgebreide analyse en vergelijking met een manueel proces, was de winst in tijd (tegenover manueel vertalen) en kwaliteit (tegenover MT) overduidelijk. In het addendum geven we een korte bespreking van Crowding en enkele andere vertaaltools.

5. Besluit vertaalstrategieën

manueel te duur, auto niet goed genoeg → hitl met cat tools

8. Addenda

8.1. Overzicht datasets

8.1.1. 20210303_Archeologie_Sites_exportAtlantisErfgoedBrugge.xlsx

Inhoud: Opgravingen ('projecten') met adres, type opgraving e.d

Aantal:

Types: keywords, vrije tekst, codes, datering

Opmerkingen: Beschrijvende tekst bevat html tags.

Veldnaam	sample	publiek	type	Named Entities
Projectnaam	Brugge Leeuwstraat	TRUE	text	adres
Identificatienummer	BR78/LE	TRUE	id	--
Beschrijvingpubliek	In&nbsp;de Leeuwstraat, aan de noord-oosthoek&nbsp;van de Leeuwebrug,&nbsp;is bij een werfcontrole een trap in Doornikse kalksteen waargenomen. De trap is tegen de gevel van een huis langs de Reie geconstrueerd. De trap, die&nbsp;afdaalt in de Reie, is te zien op het plan van Marcus Gerards.
 (foto: Fragment uit Kaart van Brugge uitgegeven door Marcus Gerards, Uitgever van de copie: Karel Reyaert (Storie), 1562-1900.)
 &nbsp;</p>	TRUE	text	agent, geo, concept, time/period
Beschrijvingwo	Deze vondsten bevinden zich in De Pakhuizen en hebben, ondanks hun benaming St. Trudo en de afkorting TR, niks te maken met de het gebouw van de Sint Trudoabdij te Steenbrugge (Assebroek).	FALSE	text	all

	De vondsten in ons depot met benaming St. Trudo uit 1955, komen niet uit Assebroek. In 1954 verhuisden de zusters, uit de Abdij van Steenbrugge, naar het voormalig grafelijk slot in Male dat door hun toedoen volledig gerenoveerd werd. Bij deze renovatiewerken werden enkele vondsten gedaan en vond een klein archeologisch onderzoek plaats o.a. op de binnenkoer. Deze vondsten zitten in ons depot en hebben (ondanks hun benaming St. Trudo) dus niks te maken met de Abdij van Steenbrugge. (ze kopen een pand op de) Oude Oostendse Steenweg te Sint-Pieters Brugge. In 2014 zouden de zusters er hun intrek kunnen nemen. Het grafelijk slot van Male werd aangekocht door de familie Hein Deprez uit Belsele, met de belofte dit Vlaamse erfgoed te conserveren. uit: https://www.odis.be/hercules/toonOrg.php?taalcode=nl&id=15116 De congregatie van zusters heet Sint-Trudoabdij ongeacht hun woonplaats. + info uit: Devliegheer, L. 1959: Sint-Kruis (West-Vlaanderen). Kasteel van Male, Archeologie 1959.2, 307. ---- PUBLICATIE TE ZOEKEN			
Begindatum_Onderzoek	14/02/2014	TRUE	date	--
Einddatum_Onderzoek	25/02/2014	TRUE	date	--

Termijn_Onderzoek	jul/78	FALSE	text	--
Opmerking	Vast punt: 0 cm	FALSE	text	--
Assessment	0	FALSE	binair	--
Adres	Leeuwstraat	FALSE	text	adres
Huisnummer	96a	FALSE	text	--
Kadasternummer	afd 1, sectie c, nr 342 c	FALSE	text	--
Vindplaats_Cairecordnummer	206033 en 154949	TRUE	identifie r	--
Vindplaats_Opmerking	Dudzele/Ramskapelle/Oostkerke	FALSE	text	geo
TypeOnderzoek_Naam	werfcontrole	TRUE	list	concept
Rechthouder_Naam	[leeg]	FALSE	--	--
Rechthouder_Email	[leeg]	FALSE	--	--
Rechthouder_Postcode	[leeg]	FALSE	--	--
Rechthouder_Plaats	[leeg]	FALSE	--	--
Rechthouder_Land	[leeg]	FALSE	--	--
Rechthouder_Telefoon	[leeg]	FALSE	--	--
Beheerder_Naam	[leeg]	FALSE	--	--
Beheerder_Email	[leeg]	FALSE	--	--
Beheerder_Postcode	[leeg]	FALSE	--	--
Beheerder_Plaats	[leeg]	FALSE	--	--
Beheerder_Land	[leeg]	FALSE	--	--
Beheerder_Telefoon	[leeg]	FALSE	--	--
OndernemingActor_Naam	Stedelijke Archeologische Dienst Brugge	TRUE	list	agent
Eigendomsstatus_Naam	[leeg]	FALSE	--	--
Deelgemeente_Naam	Brugge	TRUE	list	geo
Gemeente_Naam	Brugge	TRUE	list	geo
ActorPerSite_Actornaam	Hubert De Witte	FALSE	text	agent
ActorSpecialisme_Naam	Archeoloog	FALSE	list	concept
MateriaalTypePerSite_Materiaaltype_Naam	aardewerk	FALSE	list	concept
MateriaalTypePerSite_Materiaaltype_Opmerking	gebruiksaardewerk, vaatwerk	FALSE	list	concept
PeriodePerSite_Periode	volle middeleeuwen / late middeleeuwen	FALSE	list	time/period

PublicatiesPerSite_Publicatie_Titel	DE WITTE Hubert 1978: Archeologisch onderzoek in Brugge, overdruk uit: Het Brugs Ommeland, jg. 18, 1-7. / http://www.cartesius.be/CartesiusPortal/	TRUE	text	--
PublicatiesPerSite_Publicatie_Jaartal	2009 /	TRUE	date	--
PublicatiesPerSite_Publicatie_Isbnnummer	9789076297439 /	TRUE	id	--
PublicatiesPerSite_Publicatie_AuteursPerPublicatie_Auteur_Naam	[leeg]	FALSE	--	--
PublicatiesPerSite_Publicatie_AuteursPerPublicatie_Auteur_Voornaam	[leeg]	FALSE	--	--
PublicatiesPerSite_Publicatie_AuteursPerPublicatie_Auteur_Opmerking	[leeg]	FALSE	--	--
TrefwoordenPerSite_Trefwoord_Naam	gracht / kustpolders	FALSE	list	concept
Vergunning_Vergunningsnummer	1998/53	TRUE	id	--
VondstOmstandighedenPerSite_VondstOmstandigheid_Naam	werfbegeleiding	FALSE	list	concept

8.1.2. 20210303_Archeologie_VondstenExportAtlantisRaakvlak.xlsx

Zie <https://docs.google.com/spreadsheets/d/14TXIH9aFKQPk8GrAongwXBjqKLzdfxPRHvWVukKOYb8>

Inhoud: Vondsten uit archeologische opgravingen

Aantal: 1655

Types: keywords, vrije tekst, codes, datering, getal, pad

Opmerkingen:

- geen relatie met opgravingen?
- gepubliceerd op Raakvlak.be ?

Veldnaam	sample	publiek	type	Named Entities
Import	Onbekend	WAAR	list	--
Extrainformatie	Een koperen ketel met hengel. Op de ketel zijn duidelijk verschillende reparaties zichtbaar.	WAAR	text	concept, geo, agent

Titel	kraal uit amber	WAAR	text	concept
Materiaal_Naam	koperlegering	WAAR	list	concept
Materiaal_Materiaalgroep_Naam	metaal	WAAR	list	concept
ZekerheidMateriaal_Naam	zeker	WAAR	list	--
Behandeling_Naam	tinglazuur	WAAR	list	concept
Functie_Naam	voedselbereiding	WAAR	list	concept
ZekerheidFunctie_Naam	zeker	WAAR	list	--
Stad_Naam	Utrecht	WAAR	list	geo
Regio_Naam	Vlaanderen	WAAR	list	geo
Land_Naam	Frankrijk	WAAR	list	geo
ZekerheidStadRegioLand_Naam	zeker	WAAR	list	agent
Type_Opmerking	cent komt van het Latijnse "centum" honderd. De benaming cent werd voor het eerst gebruikt in de V.S. voor de Washington-cent in 1781	WAAR	text	concept, geo, agent
Type_Voorwerp_Naam	munt	WAAR	list	concept
Type_Voorwerp_Opmerking	Een zalfpot is cilindrisch, heeft een opening aan de bovenzijde en opgaand werk dat recht of uitstaand is. Additieven zijn zeldzaam; meestal heeft een zalfpot een standvlak of standvoet, in een enkel geval een standing of ziel. Incidenteel zijn zalfpotten gedecoreerd met bijvoorbeeld een aanduiding van de inhoud.	WAAR	text	concept, geo, agent
Type_Voorwerp_Maat1	randdiameter	WAAR	list	concept
Type_Voorwerp_Maateenheid1	cm	WAAR	list	concept
Type_Materiaal_Naam	faience	WAAR	list	concept
Type_Materiaal_Materiaalgroep_Naam	keramiek	WAAR	list	concept
Baksel_Naam	Roodbeschilderd wit geel	WAAR	text	concept
Baksel_Beschrijving	roodbeschilderd wit geel baksel orangerode verf	WAAR	text	concept
Baksel_Materiaal_Naam	roodbeschilderd aardewerk	WAAR	list	concept
Baksel_Materiaal_Materiaalgroep_Naam	keramiek	WAAR	list	concept
Randtype_Naam	LK2A	WAAR	list	concept

Randtype_Randtypegroep_Naam	pot	WAAR	list	concept
Volledigheid_Naam	bijna archeologisch volledig	WAAR	list	--
ConservatiesPerVondst_Conservatie_Naam	reiniging droog met borstel	WAAR	list	--
GebruiksporenPerVondst_Gebruiksspoor_Naam	slijtage, rand / slijtage, binnenzijde / krassen / slijtage buitenzijde	WAAR	list	--
KenmerkenPerVondst_Kenmerk_Naam	bodem, standring, glad	WAAR	list	--
PeriodePerVondst_Periode	nieuwe tijd	WAAR	list	time/period
PublicatiesPerVondst_Publicatie_Titel	CARMIGGELT Arnold 1991: Een beeld van een vondst. Haagse archeologische vondsten in particulier bezit. in: VOM REEKS 1991-4.	WAAR	text	--
VersieringenPerVondst_Naam	handbeschilderd / figuratief	WAAR	list	concept
Voorwerpmeting_Voorwerp_Naam	pot	WAAR	list	--
Voorwerpmeting_Voorwerp_Afkorting	pot	WAAR	list	--
Voorwerpmeting_Voorwerp_Opmmerking	Een pot is een gesloten vorm en heeft een primaire opslagfunctie. De formaten, baksels en additieven doen de specifieke functie wijzigen. Potten hebben vaak twee oren en soms een schenklip, tuit of hengel.	WAAR	text	concept
Voorwerpmeting_Voorwerp_Maat1	randdiameter	WAAR	list	concept
Voorwerpmeting_Voorwerp_Maat eenheid1	cm	WAAR	list	concept
Voorwerpmeting_Zekerheid_Naam	zeker	WAAR	list	concept

8.1.3. CollectieArcheologie_Wikimedia.docx

- URL naar dataset Vondsten uit Raakvlak, opgeladen in Wikidata: <https://w.wiki/33c5> (525)
- URL naar beelden op Wikimedia Commons [https://commons.wikimedia.org/wiki/Category:Media_contributed_by_Raakvlak_\(Brugge\)](https://commons.wikimedia.org/wiki/Category:Media_contributed_by_Raakvlak_(Brugge)) (525)

8.1.4. 20210303_Beeldbank_exportMemorixMaior.csv

Veldnaam	sample	publiek	type	Named Entities
uuid	00002ecc-e848-4b91-b67a-9882b23213a3	FALSE	id	--
Fotonummer	COU001000146	TRUE	id	
Bewaarplaats	Privécollectie	TRUE	list	agent
Inventarisnummer	FO/A30482	TRUE	id	--
Plaats- en straatnaam-uuid	052243d6-0eb4-11e1-98c8-cda787f35be8	FALSE	--	--
Plaats- en straatnaam-altLabel	[leeg]	FALSE	--	--
Plaats- en straatnaam-hiddenLabel	[leeg]	FALSE	--	--
Plaats- en straatnaam-prefLabel	Brugge, Centrum, Gouden-Handstraat	TRUE	list	geo
Huisnummer		TRUE	text	--
Exacte datering (dd-mm-jjjj)	1972-08-00	TRUE	date	--
Beginjaar	1984	TRUE	date	--
Eindjaar	1984	TRUE	date	--
Titel	Een piraatboot	TRUE	text	concept, agent, geo
Beschrijving	Deze piraatboot van Pietje 'Puppe' Hendryckx voer op de reien rond met mensen die niet voor de vaartocht betaalden. Dit beweerde hij toch. Hij was van meni	TRUE	text	concept, agent, geo
Thema's-uuid	f334632a-0eb3-11e1-ad8a-a81b87f35be8	FALSE	--	--
Thema's-altLabel		FALSE	--	--
Thema's-hiddenLabel		FALSE	--	--
Thema's-prefLabel	Stadsbeeld, Water en bruggen, Waterwegen	TRUE	list	concept
Trefwoorden-uuid	0f6c1fd9-228f-05cd-6934-a53d36d4f26d	FALSE	--	--
Trefwoorden-altLabel		FALSE	--	--

Trefwoorden-hiddenLabel		FALSE	--	--
Trefwoorden-prefLabel	rei	TRUE	list	concept
Vervaardiger	Edmond Coucke	TRUE	list	agent
Uitgever		TRUE	list	agent
Soort document	foto	TRUE	list	concept
Kleurmodel	Kleur	FALSE	list	concept
Verzameling	Edmond Coucke	TRUE	list	agent
Werver / schenker	Edmond Coucke	FALSE	list	agent
Rechthebbende	Stadsarchief Brugge	FALSE	list	agent

8.1.5. 20210308_Beeldbank_exportAtlantis.csv

Veldnaam	sample	publiek	type	Named Entities	service
Dekking	Brugge, Centrum, Garenmarkt	WAAR	list	geo	
Plaats	Brugge	WAAR	list	geo	Geonames
Wijk	Centrum	WAAR	list	geo	Geonames
Adres	Garenmarkt	WAAR	list	geo	Geonames
Auteur of maker	Zaza Bertrand	WAAR	text	agent	VIAF
Omschrijving	De Aarde' van Karel Aubroeck uit Temse (1894-1986), volgens de beeldhouwer zijn beste werk. Een bronzen afgietsel bevindt zich in Verenigde-Natieslaan te Gent. Aubroeck werkte vooral in expressionistische stijl, gekenmerkt door een gesloten karakter. Van zijn hand zijn o.a. het Koning Albertmonument te Nieuwpoort, beelden op de Paxpoort te Diksmuide en de gedenksteen voor Lt. De Winde te Westrozebeke.	WAAR	text	agent, concept, geo	ML translation
Bron	Stadsarchief Brugge / Fotoverzameling	WAAR	list	agent	manual

	Stadsarchief Brugge				
Onderwerp en trefwoorden	Stadsbeeld, Standbeelden en straatmeubilair, Standbeelden	WAAR	list	concept	manual
Titel	Annunciatie en de aanbidding van het Jezuskind' (1452) van Petrus Christus, Groeningemuseum	WAAR	text	agent, concept, geo	ML translation
Bestandstype	foto	WAAR	list	concept	AAT

8.1.6. 20210302_Groeningemuseum_exportAtlantis.csv

Veldnaam	sample	publiek	type	Named Entities	service
Titel	Heilige Cecilia van Raphael - NB aan te vullen!	WAAR	text	concept, agent, geo	ML translation
Description	Bouwtekeningen voor de pier en de kaai van Zeebrugge.	WAAR	text	concept, agent, geo	ML translation
Object_Name	lithografie (prent)	WAAR	list	concept	AAT
Production_Period	19de eeuw	WAAR	text	time/period	manual
Content_Motif_General	architectuur	WAAR	list	concept	
Dimensions	(geheel) hoogte 53 mm, (geheel) breedte 35 mm	WAAR	text	concept	
Creator	17de eeuw Anonieme meester, Jan van Eyck	WAAR	list	agent	
Collection	Prentenkabinet	WAAR	list	agent	
Material	inkt (drager), papier (drager)	WAAR	list	concept	
Copyright	Lukas - Art in Flanders vzw (www.lukasweb.be) - Musea Brugge	WAAR	list	concept	

8.1.7. 20210317exportkatrien_Prentenkabinet.xml

AdlibXML export

Veldnaam	sample	publiek	type	Named Entities
acquisition.method	schenking	WAAR	list	concept
acquisition.source	Brangwyn, Frank	WAAR	list	agent

collection	Brangwyn	WAAR	list	--
content.motif.general	stadsgezichten	WAAR	list	concept
creator	Brangwyn, Frank	WAAR	list	agent
creator.history		WAAR	text	--
current_owner	Stad Brugge	WAAR	list	agent
description	Gezicht op twee pittoreske bruggen te Ventimiglia (Noord-Italië). Een kermisnummer wordt opgevoerd voor een woonwagen. Zie ook 0.2.GRO0002.II.	WAAR	text	concept, agent, geo
dimension.notes/0	blad	WAAR	text	concept
dimension.part/0	geheel	WAAR	list	concept
dimension.type/0	hoogte	WAAR	list	concept
dimension.unit	mm	WAAR	list	concept
documentation.author	Marechal, Dominique	WAAR	list	agent
documentation.page_reference	p. 238	WAAR	text	--
documentation.title	Collectie Frank Brangwyn. Catalogus	WAAR	text	--
documentation.title.lead_word	De	WAAR	text	--
inscription.content/0	FB	WAAR	text	--
inscription.position/1	links onderaan	WAAR	list	concept
inscription.type/0	monogram	WAAR	list	concept
institution.name	Musea Brugge	WAAR	list	agent
institution.place	Brugge	WAAR	list	geo
location.history.executor/0		WAAR	list	--
material.part/0	medium	WAAR	list	concept
material/0	aquarelverf	WAAR	list	concept
object_category	tekeningen	WAAR	list	concept
object_name	aquarel	WAAR	list	concept
owner_hist.acquired_from	Brangwyn, Frank	WAAR	list	agent
owner_hist.acquisition.method	schenking	WAAR	list	concept
physical_description	aquarel en potlood op lichtgrijs papier	WAAR	text	concept
production.period	20ste eeuw	WAAR	text	time/period
reproduction.creator	Provost, Dominique	WAAR	list	agent

title	Brug te Ventimiglia	WAAR	text	concept, agent, geo
title.translation/0	Bridge, Ventimiglia	WAAR	text	concept, agent, geo
title.type	toegekende titel	WAAR	list	concept

8.1.8. Kranten

ALTO/XML

Veldnaam	sample	publiek	type	Named Entities
Titel krant	Journal de Bruges	WAAR	list	--
[tekst]	Samedi, Dimanche et Lnndi N° 1-2-3 « 62° Année 1', 2 et 3 Janvier lb98 ABONNEMENTS : (PAYABLES D'AVANCE) On an fr. iq Six moiii. a q Troi« moi . g Pour l'étranger, le port en aui. Un numéro : 10 oentimeï. OM S'ABONNE AU SUHEAU OU JOURNAL PLACE MEMLINC, N» 1, ainsi qna ohes tous les librairea et directeurs de poste Le Bureau eet [...]	WAAR	text	agent, geo, concept

TITEL	JAARTAL	files	woorden
JournalDeBruges	1838_1878	2.456	
JournalDeBruges	1.898	1.232	
JournalDeBruges	1919_1938_1952	1.514	
JournalDuDépartementDeLaLys	1811-1814	--	--
TOTAAL		5.202	17.013.893

8.3. Richtlijnen en kwaliteitscriteria vertaling

1. Richtlijnen voor de vertaling van eigennamen en plaatsnamen

- Namen van Brugse straten, pleinen, parken enz. worden niet vertaald.
- Straatnamen worden in het Engels voorafgegaan door het lidwoord *the*, tenzij de straatnaam op het einde van een titel staat en volgt op een komma (bv. *Arrival at the Koninklijke en Prinselijke Gilde Sint-Joris Stalen Boog, Stijn Streuvelsstraat.*)
- Namen van huizen, gebouwen, instellingen, bedrijven en evenementen worden niet vertaald (bv. Middenstandshuis, Hogere Polytechnische School, Gistfabriek, Praalstoet van de Gouden Boom, Heilig Bloedspel, ...).
- Gebouwen waarvoor een goede en courante vertaling bestaat, vormen een uitzondering:

Stadhuis	<i>City Hall</i>
Belfort	<i>Belfry</i>
Gruuthusepaleis	<i>Gruuthuse Palace</i>
- Namen van religieuze gebouwen en organisaties waarin de naam van een heilige voorkomt, kunnen wel worden vertaald:

Confrérie van Sint-Jozef	<i>Confraternity of St Joseph</i>
Onze-Lieve-Vrouwekerk	<i>Church of Our Lady</i>
Sint-Salvatorskathedraal	<i>St Salvator's Cathedral</i>
Sint-Jorisgilde	<i>Guild of St George</i>
- Namen van vorsten, Bijbelse figuren en historische personages met een courante vertaling, worden vertaald:

Willem III van Oranje	<i>William III of Orange</i>
de hertog van Alba	<i>the Duke of Alba</i>
Anna en Joachim	<i>Anne and Joachim</i>

2. Vaste vertaling terminologie

Voor aantal termen die heel vaak voorkomen (vaak specifieke terminologie) kunnen specifieke vertalingen worden afgesproken. Dit kan het aantal fouten reduceren, maar vereist wel manueel nazicht van de vertalingen.

Een lijst samengesteld o.b.v. de twee gebruikte datasets:

Nederlandse term	Context/betekenis	Vertaling Google	Correcte vertaling
reactie	bij een beeld op Erfgoed Brugge	<i>response</i>	<i>comment</i>
opname	opname van een foto	<i>recording</i>	<i>image</i>
voornaam bezoeker		<i>distinguished visitor</i>	<i>prominent visitor</i>
dienstjubileum	m.b.t. de Gistfabriek	<i>jubilee</i>	<i>work anniversary</i>
jubilair	bij dienstjubileum	<i>jubilee</i>	<i>employee celebrating his work anniversary</i>
bootexploitant	op de reien	<i>boat operator</i>	<i>boat company</i>
v.l.n.r.	van links naar rechts		<i>from left to right</i>
in de achterste / middenste / voorste rij	positie op groepsfoto		<i>in the back / middle / front row</i>
in de voorgrond / het middenplan / de achtergrond	positie op een foto, schilderij, ...		<i>in the foreground / middle ground / background</i>
het blad	blad papier	<i>the magazine</i>	<i>the sheet</i>

broederschap	religie	<i>brotherhood</i>	<i>confraternity</i>
linker- / rechterluik	diptiek, triptiek		<i>left / right panel</i>
luik(en)	diptiek, triptiek		<i>panel(s)</i>
hoofdman	van een gilde	<i>headman / captain</i>	<i>dean</i>
pronkzaal	van de Academie	<i>pageantry / showroom</i>	<i>pageantry room</i>
memorietafel			<i>memorial coat of arms</i>

3. Aanvaardbare fouten

Een aantal fouten maken een vertaling niet meteen onaanvaardbaar. Voor deze fouten kan een foutmarge van 20% worden gehanteerd (fouten in max. 2 op 10 vertalingen).

Hieronder kunnen vallen:

- De betekenis blijft dezelfde of is duidelijk, maar grammaticaal is de vertaling niet vlot of niet helemaal correct. De oorzaak hiervan is meestal de specifieke formulering in het Nederlands.

Nederlandse tekst	Automatische vertaling	Vlotter alternatief
Van links naar rechts zien we op de achterste rij ...	<i>From left to right we see in the back row ...</i>	<i>In the back row we see, from left to right ...</i>
Op 22 februari 2020 deelt de heer Vanderschaeghe ons het volgende mee ...	<i>On February 22, 2020, Mr. Vanderschaeghe informs us of the following ...</i>	<i>On February 22, 2020, Mr. Vanderschaeghe informs us of the following details ...</i>
... Garemijn die uitmunt in het schilderen van ...	<i>... Garemijn who is excellent in painting ...</i>	<i>... Garemijn who excelled in painting ...</i>

- Inconsequent gebruik van lidwoorden bij straatnamen: bv. *the Boeveriestraat* is de regel, maar blijft ook zonder *the* acceptabel.
- De naam van een gebouw of organisatie blijft onvertaald, waar dat volgens de richtlijnen wel mogelijk is: bv. Sint-Salvatorskathedraal, Stadhuis.
- Een naam van een vorst, Bijbelse figuur of historisch personage blijft onvertaald: bv. *Albrecht, Archduke of Austria* (i.p.v. *Albert*); *the Duke of Alba* (i.p.v. *Alba*)
- De vertaling is minder specifiek dan het Nederlandse origineel, waardoor betekenis verloren gaat: bv. *coat of arms* als vertaling voor alliantiewapen, i.p.v. *arms of alliance*.

Bij voorkeur wordt een melding voorzien die duidelijk maakt dat het om een automatische vertaling gaat die fouten kan bevatten. De gebruiker zou ook de mogelijkheid moeten krijgen om te de originele Nederlandse tekst weer te geven (zie bv. *European Multilingual Strategy*, p. 14).

4. Niet-aanvaardbare fouten:

Voor deze fouten wordt een foutmarge van 5% gehanteerd (fouten in max. in 1 op de 20).

Hieronder vallen:

- Fouten die zorgen voor een verandering van betekenis:

Nederlandse origineel	Context	Automatische vertaling	Correcte vertaling
opname	van een foto	<i>recording</i>	<i>image</i>
het blad	blad papier	<i>the magazine</i>	<i>the sheet</i>
voorstelling	op foto, schilderij, prent, ...	<i>performance</i>	<i>depiction, image</i>

bukschietters of kolveniers	schuttersgilde	<i>bus shooters or flasks</i>	<i>culverin shooters</i>
was	materiaal	<i>laundry</i>	<i>wax</i>
luiken	diptiek, triptiek	<i>shutters</i>	<i>panels</i>
dekenen	van een gilde	<i>blankets</i>	<i>deans</i>

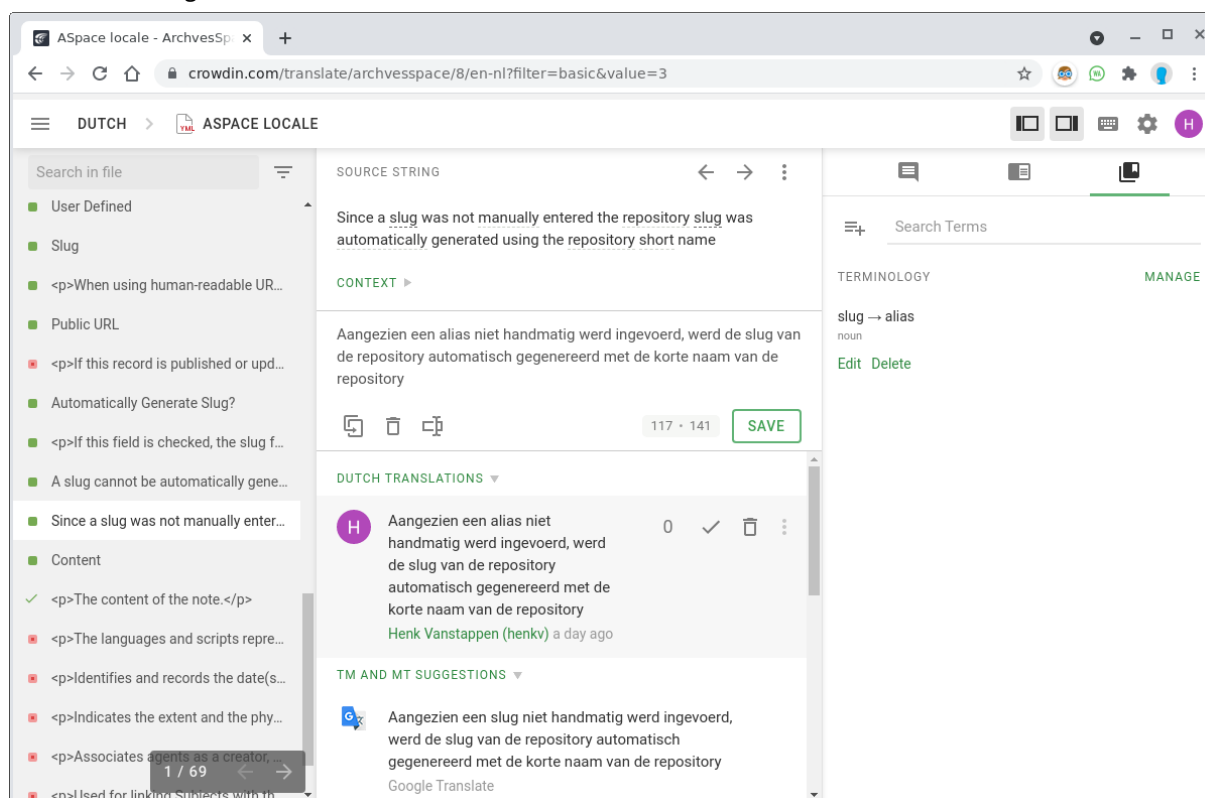
- Grammaticale fouten die leiden tot een betekenisverandering:
it i.p.v. *he*
de militair laat zich ... portretteren als ... > *the soldier can be portrayed as ...*
- Gedeeltelijke vertaling van eigennamen:
Edele Confrerie van het Heilig Bloed > *Noble Confrerie van het Heilig Blood*
Sint-Janshospitaal > *Saint Jans hospital (St John's Hospital of origineel behouden)*
Antonius en Paulus > *Antonius and Paul* (beter: *Anthony and Paul*)

8.4. CAT en localization tools

8.4.1. Crowdin

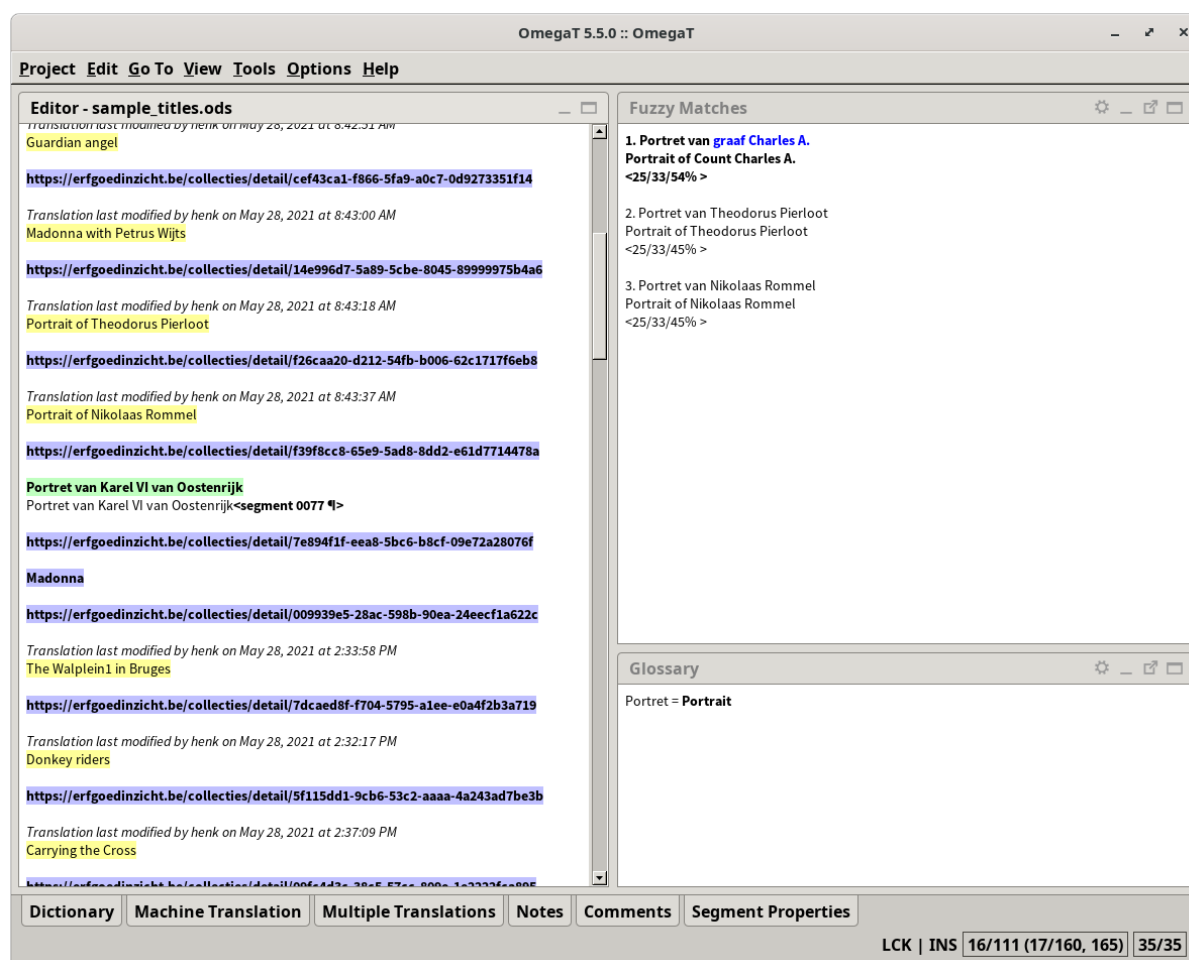
Crowdin is een cloudgebaseerde oplossing voor het beheer van localization files. De tool heeft een online vertaaleditor met vertaalgeheugen en glossary, waar teksten kunnen worden vertaald en nagelezen door domeinspecialisten. Crowdin kan gebruikt worden door een eigen vertaalteam, crowdsourcing, maar heeft ook een marktplaats waar vertaalbureaus rechtstreeks kunnen gecontacteerd worden..

Crowdin heeft machinevertaling geïntegreerd in de vertaalworkflow. Het ondersteunt momenteel Microsoft Translator, Yandex.Translate, Google Translate, Amazon Translate, Watson (IBM) Translator, DeepL Translator. Machinevertalingen kunnen achteraf worden bewerkt.



8.4.2. OmegaT

OmegaT is een open-sourcesoftware (Java) waarmee vertalers gebruik kunnen maken van een vertaalgeheugen dat gebruik maakt van fuzzy matching, waarmee automatisch vertalingen worden voorgesteld. Deze tool gebruikt meerdere vertaalgeheugens tegelijk en stelt gebruikers in staat om meerdere projecten tegelijkertijd te verwerken. OmegaT ondersteunt Unicode (UTF-8), spellingcontrole en kan geïntegreerd worden met MT services.



8.4.3. Smartcat

<https://smartcat.com/>

A collaborative CAT tool, SmartCAT streamlines the translation process by allowing the translator, editor, and other contributors to work and collaborate in real time. This cloud-based software leverages the concept of translation memories (TM) and allows translators to create glossaries for consistent translations. In addition, SmartCAT now supports multilingual translation memories, which makes creating and managing TMs easier than with some desktop applications. This feature can also be beneficial for translators working in mixed language pairs.

8.4.4. SDL Trados Studio

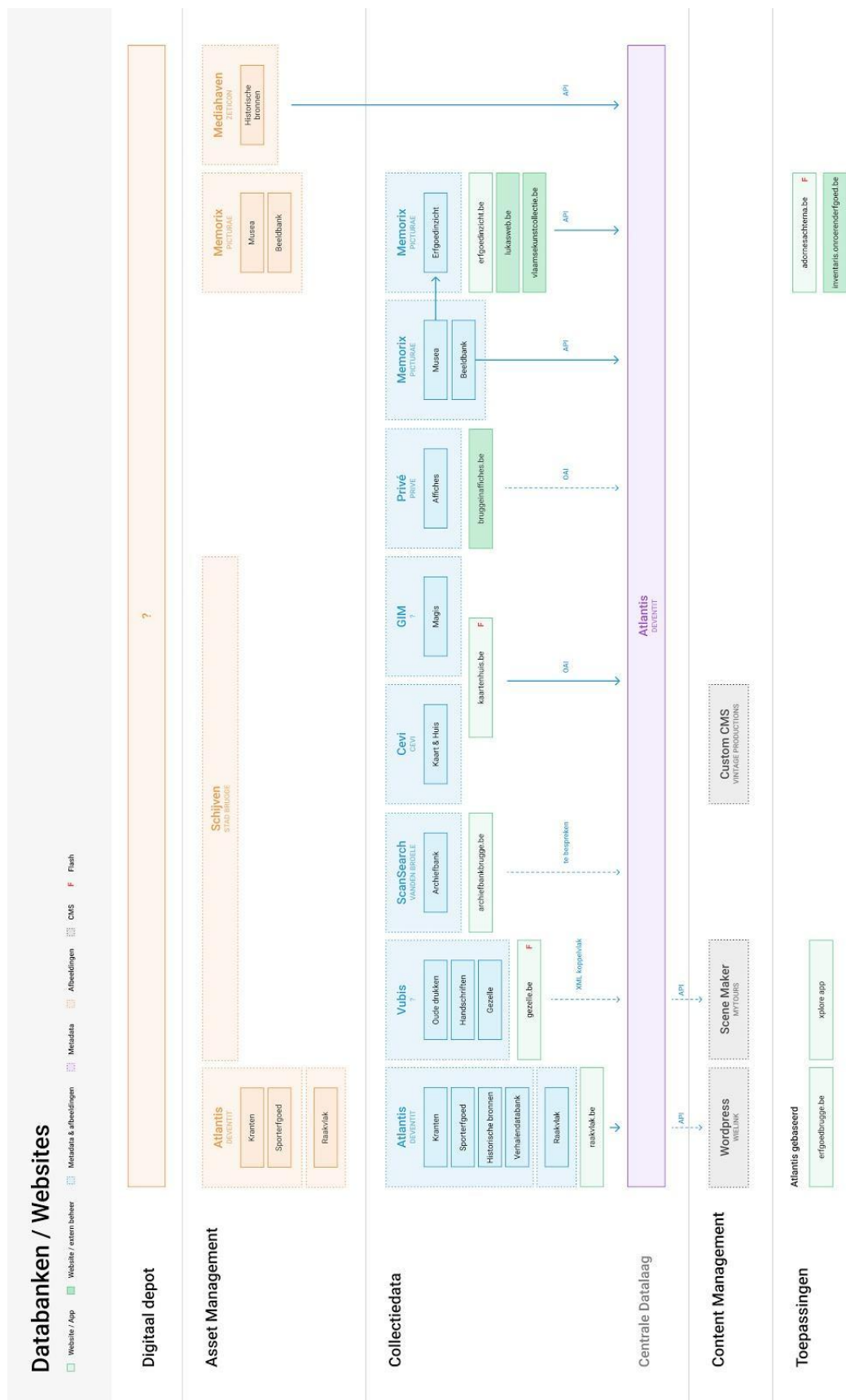
<https://www.sdltrados.com/products/trados-studio/>

The most recommended computer-assisted translation (CAT) tool by Gengo Wordsmiths, SDL Trados is a wise investment for full-time translators. This software features TM, terminology, machine translation and software localization. Most large agencies require translators to work with one of the established CAT tools, so using SDL Trados could also increase your client base and broaden your horizons. If you need time to decide before purchase, try a free demo version for 30 days.

8.4.5. Zanata

A web-based system for translators, content creators and developers to manage localization projects, Zanata handles the entire translation workflow and allows translators to focus on translations, not tools and formats. Its TM also finds and suggests the best translation matches in the entire system. Zanata's Editor works on any web browser with no installation necessary. Multiple translators can also work on the Editor, with a chat room for real-time communication.

8.6. Architectuurschema (huidige situatie)



8.7. Technische implementatievereisten

8.7.1. Gebruik erkende datastandaarden

Het gebruik van een aantal internationale standaarden voor data registratie en uitwisseling zal interpretatie en integratie van de betrokken data sterk vereenvoudigen. In het bijzonder zijn ISO-639²⁶ met codes voor de meest courante talen en ISO-8601²⁷ voor het registreren van datums aangewezen.

8.7.2. Language tags

Voor XML of RDF gebaseerde systemen wordt het gebruik van de language tag `xml:lang=` ten stelligste aangeraden.

8.7.3. Meertalige interne vocabularies

Bij het gebruik van waardenlijsten wordt bij voorkeur gebruik gemaakt van een model dat gebruik maakt van concepten, waaraan meerdere vertalingen kunnen gekoppeld worden. Een breed gedragen model hiervoor (dat ook (hiërarchische) relaties tussen termen en concepten toelaat is vastgelegd in ISO-25964.²⁸ SKOS, het met RDF werkende model is hierop gebaseerd.²⁹

8.7.4. Relatieve databases

Er zijn verschillende opties mogelijk die we hier samenvatten:

Methode	Hoe werkt dit	Voordelen	Nadelen
Extra kolommen	Voor elke taal wordt een extra kolom voorzien in een record, bijv. Tabel Object: id=1, naam=Jan, titel_nl=Man met hoed, titel_fr=Homme au chapeau	b	Moeilijk onderhoudbaar Moeilijk uitbreidbaar Intelligentie meer in code (waar moet info gezocht worden)
Extra rijen	Voor elke taal wordt een extra record voorzien in dezelfde tabel bijv. Tabel Object(record1): id=1, id2=1, naam=Jan, titel_nl=Man met hoed Tabel Object(record2): id=1, id2=2 naam=Jan, titel_fr=Homme au chapeau	Snelle implementatie Eenvoudige queries	Moeilijk onderhoudbaar Moeilijk uitbreidbaar Veel dubbele data
Unieke vertaal-tabel	In aparte tabel worden alle vertalingen voorzien via id en vertalingen. Bijv. Tabel Object: id=1, naam =Jan, titel= 11 Tabel Vertaalid: id=11 Tabel Vertaling (record1): id = 111,	Makkelijk uitbreidbaar Vertaling op 1 plaats, helder	Moeilijker queries Moeilijker onderhoud (verschillende tabellen, niet meteen duidelijk waar vertalingen (kunnen) worden gebruikt)

²⁶ [Overzicht van de taalcodes](#) via de Library of Congress

²⁷ [Samenvatting en links naar de officiële documentatie](#) via Wikipedia

²⁸ <http://www.niso.org/schemas/iso25964>

²⁹ <https://www.w3.org/2004/02/skos/intro>

	Vertaalid = 11, taal=nl, tekst: Man met hoed Tabel Vertaling (record2): id = 112, Vertaalid = 11, taal=fr, tekst: Homme au chapeau		
Bijkomende vertaaltabel	Per tabel wordt een aparte vertaaltabel voorzien, te vertalen velden worden in de vertaaltabel gezet. Bijv. Tabel Object: id=1, naam =Jan Tabel ObjectVertaald(record1): id=1, id2=11, taal=nl, titel= Man met hoed Tabel ObjectVertaald(record2): id=1, id2=12, taal=fr, titel=Homme au chapeau	Eenvoudige queries Makkelijk uitbreidbaar	Complexer datamodel

8.7.5. 7.7.5. Vertalingen in complexe architectuur met meerdere toepassingen

Een aantal parameters hebben invloed op de ideale plaats waar vertalingen (semi-)automatisch gebeuren en de vertalingen worden bijgehouden.

8.7.5.1. Doel van de vertalingen

In tegenstelling tot internationale organisaties/bedrijven wordt in Brugge meertaligheid niet gezien als een vereiste voor de back-end, maar als onderdeel van de publieksfunctie. Dit houdt in dat voor het pure beheer van kennis en de uitvoering van interne bedrijfsprocessen vertalingen weinig relevant zijn, tenzij wanneer brongegevens naar het Nederlands moeten worden vertaald.

Het is daarom logisch dat vertalingen best zo dicht mogelijk bij de publieke ontsluitingsfaciliteiten worden uitgevoerd en bewaard.

8.7.5.2. Vertalingen als apart proces

Het doel van vertalen in Brugge heeft dan ook als gevolg dat bij het registreren van kennis of documenteren van processen vertalen niet tot deze kernprocessen behoort: het is pas wanneer gegevens publiek worden gemaakt dat vertalen wordt overwogen. Het vertalen is dus een proces dat autonoom kan uitgevoerd, aangestuurd en opgevolgd worden.

Gevolg: vertalen hoeft niet uitgevoerd te worden bij de oorspronkelijke registratie van data, en hoeft dus niet noodzakelijk georganiseerd binnen applicaties die specifieke processen ondersteunen.

8.7.5.3. Hergebruik

Het principe om data slechts op één plaats te bewaren en op meerdere plaatsen te gebruiken is een best practice die leidt tot meer kwaliteitsvolle data en kostenefficiënter beheer van zowel data als applicaties.

Gevolg: aangezien in de applicatiearchitectuur van Brugge verschillende applicaties met gelijkaardige data (bijv. thesauri) werken zal een gecentraliseerde 'vertaalservice' leiden tot homogenere data, en dient niet in elke applicatie apart geïnvesteerd te worden wanneer deze niet ingericht is op meertaligheid.

8.7.5.4. Vereisten voor CAT en localisation tools

In een ideaal scenario integreren CAT-software en localisation tools vlot met een centrale data-laag, die instaat voor het leveren van data aan website en fungeert als data-provider. De nood om vlotte integratie tussen CAT-software, localisation tools en de reeds gebruikte software vlot mogelijk te maken wordt enkel maar hoger omdat wellicht op korte termijn niet enkel met de centrale data-laag dient geïntegreerd te worden. Een goed gedocumenteerde API, bij voorkeur via REST of GraphQL, is dan ook een must.