

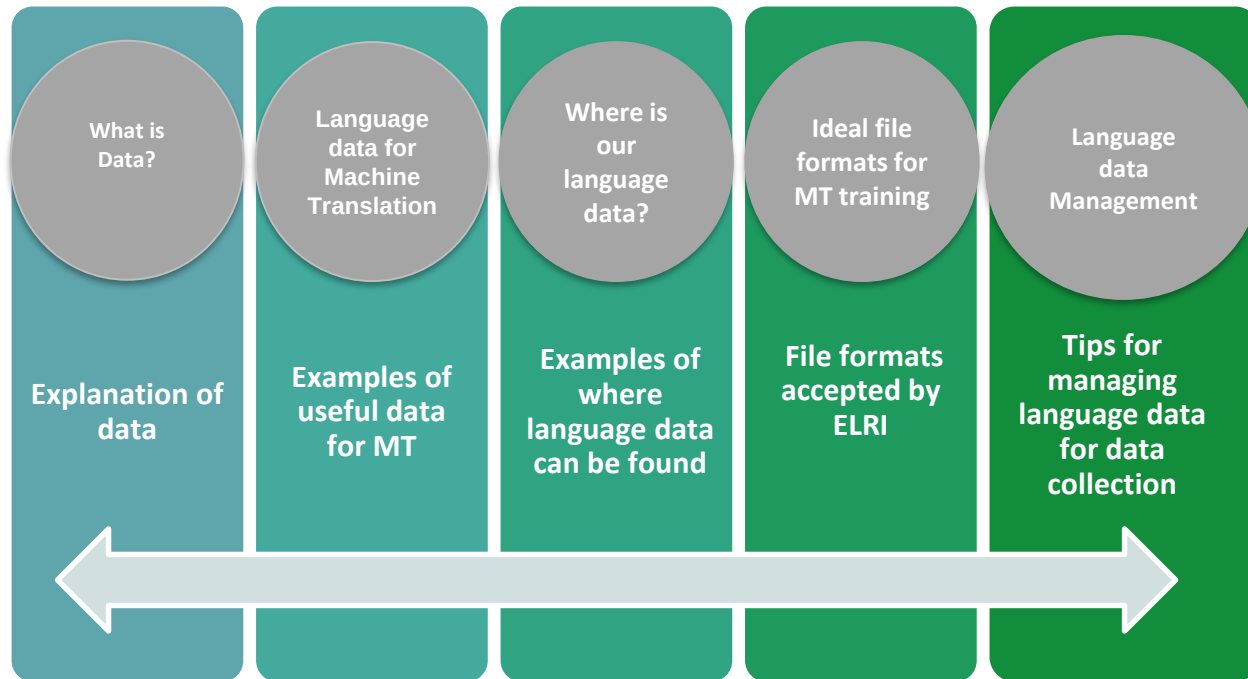


Engaging Content
Engaging People

To Share or Not to Share

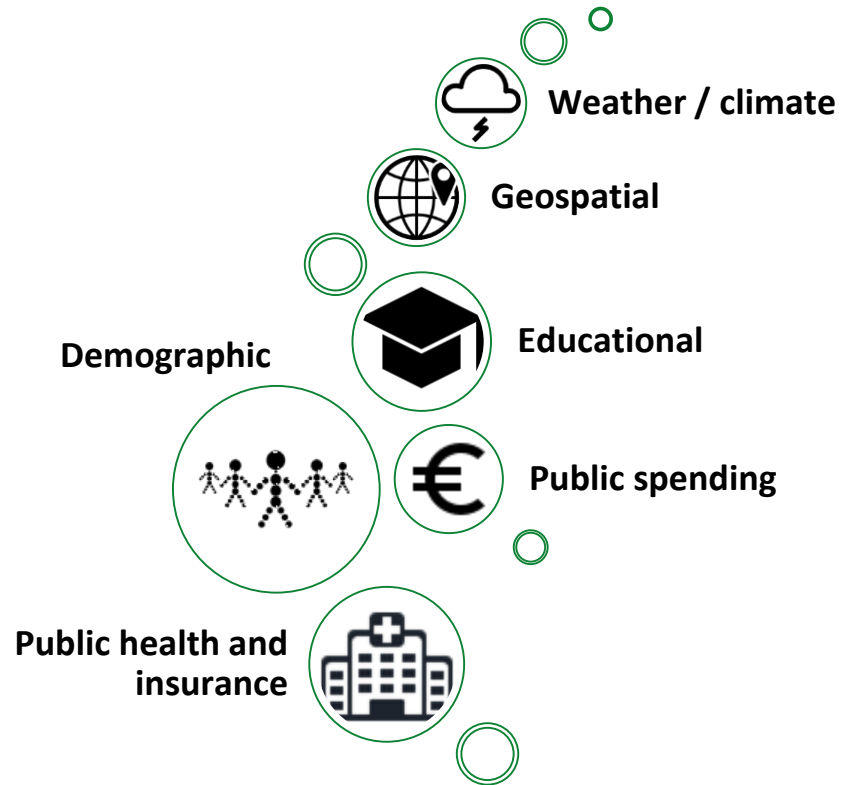
Practical guidance for Data Collection in ELRI

Jane Dunne (Dublin City University)



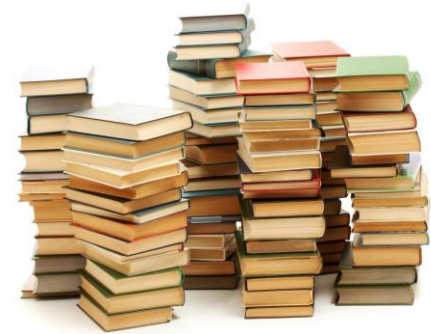
What is Data?



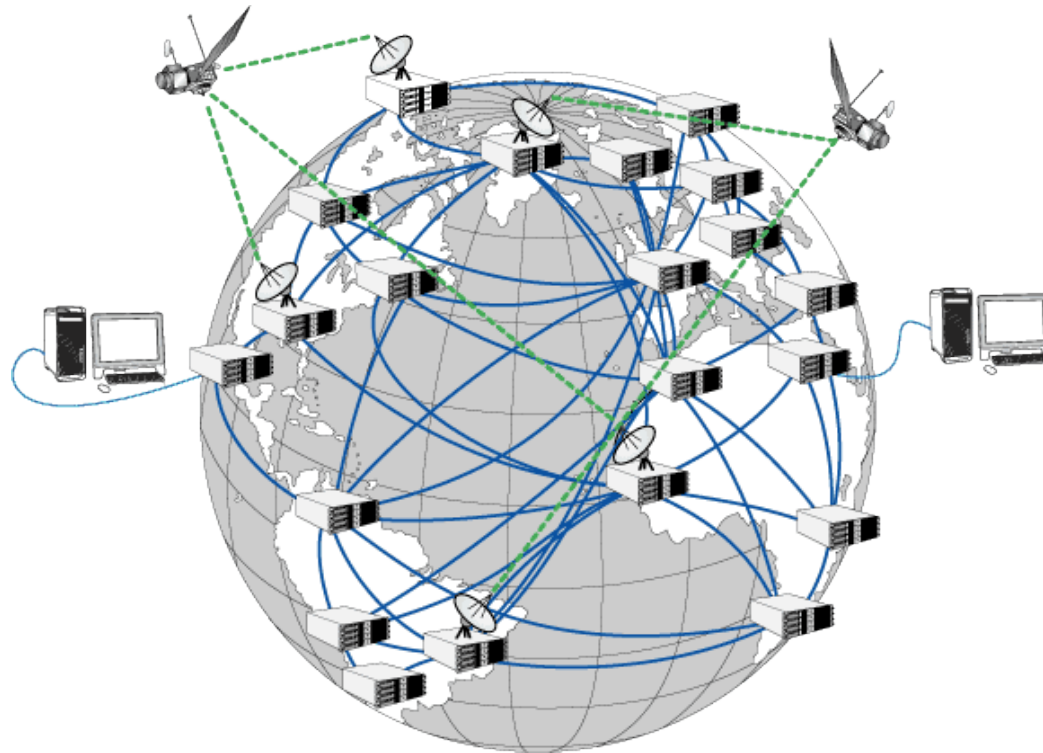


What is language data?

- Anything that contains “words”, preferences for “sentences”, even for sentences expressed in multiple languages
- Bags of “words”, “sentences”, multiple bags



wiseGEEK



<http://www.portices.fr/formation/Res/Internet/Res/InternetMonde.gif>

Data: any piece of electronically stored **content**



(Text) **Language Data:** any piece of electronically stored **text**

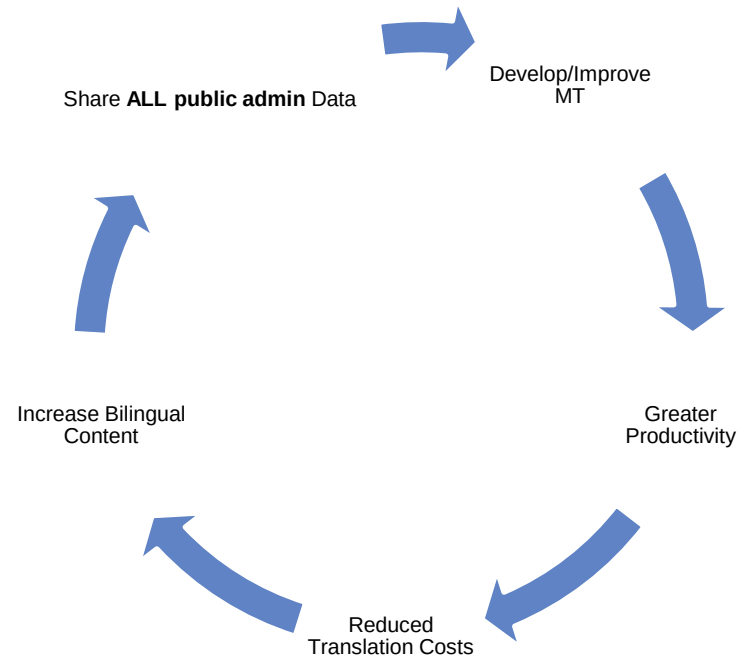
Machine Translation typically needs “aligned” Translation...



English

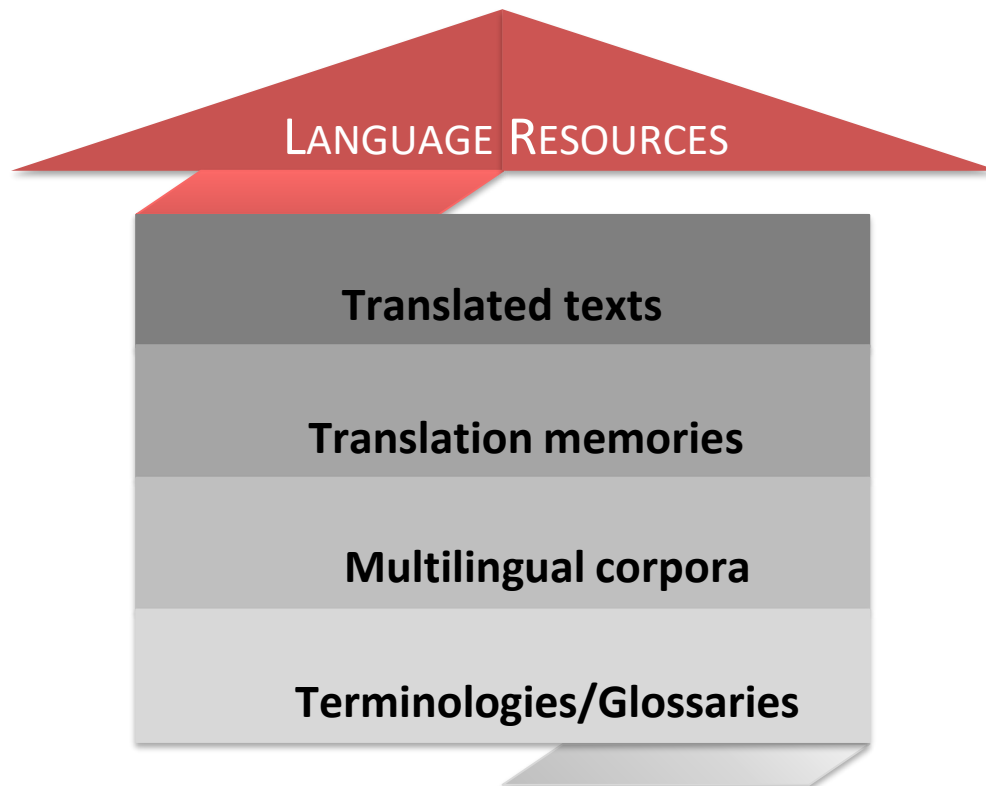


French



Language data for Machine Translation





Irish text

Dlúthchuid de chultúr, de thráchtáil, de thrádáil agus de chomhoibriú na hEorpa is ea na teangacha agus an éagsúlacht teanga. Chun cumarsáid ilteangach agus malartú doiciméad le húdaráis phoiblí thíortha na hEorpa, agus idir na húdaráis sin, a éascú, cuireann an Coimisiún Eorpach ardán aistriúcháin uathoibríoch ar fáil mar chuid den tSaoráid um Chónascadh na hEorpa (CEF AT).

Tá an t-ardán saor in aisce do gach institiúid phoiblí.

Faoi chlár na Saoráide um Chónascadh na hEorpa feabhsófar clúdach agus cáilíocht an chórais aistriúcháin, agus ullmhófar é le bheith comhtháthaithe ina iomláine le seirbhísí poiblí ar líne uile-Eorpacha, amhail na seirbhísí sin dá bhforáiltear sa chlár.

Translation in English

Languages and linguistic diversity are at the heart of European culture, commerce, trade and collaboration. In order to facilitate multilingual communication and the exchange of documents between and with national public administrations in Europe, the European Commission provides an Automated Translation platform within the Connecting Europe Facility (CEF AT).

The platform is free of charge for all public institutions

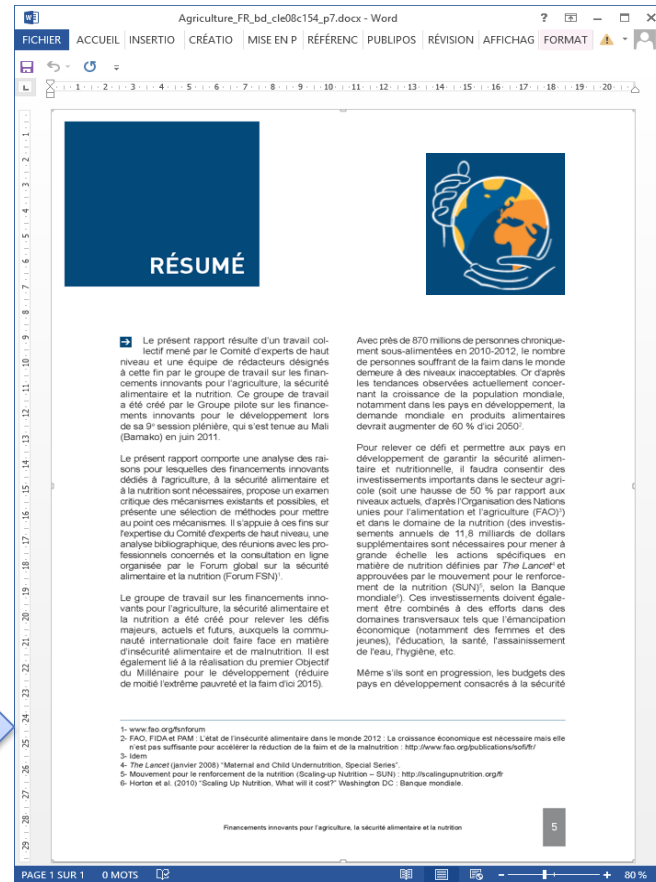
The CEF programme will improve the coverage and quality of the translation system, and will make it ready to be fully integrated to pan-European public online services, as those foreseen in the CEF programme.



English
version



French
version



Source: <http://www.diplomatie.gouv.fr/fr/photos-videos-publications/publications/enjeux-planetaires-cooperation/rapports/article/rapports-du-groupe-pilote>,
Financements innovants pour l'agriculture, la sécurité alimentaire et la nutrition, Ministère des Affaires étrangères et du Développement international

Translation Memory files are XML encoding of parallel texts

SDL Trados translation files are in **.TMX** format:

```
<tu reationdate="20100727T135602Z" creationid="AHGI\niarn" changedate="20100727T135602Z" changeid="AHGI\niarn"
lastusedate="20100727T135602Z">^M
  <prop type="x-Context">-1853973948243398551, -5048276635724717248</prop>^M
  <prop type="x-Origin">TM</prop>^M
  <prop type="x-ConfirmationLevel">Translated</prop>^M
  <tuv xml:lang="ga-IE">^M
    <seg>Is í an chuis oibiachtúil a bhfuil conradh do théarma socraithe eile á thabhairt duit go bhfuil an tÚdarás ag
fanacht go fóill ar chead chun post an phríomhfheidhmeannaigh a fhógairt agus príomhfheidhmeannach úr a cheapadh mar
chomharba ort.</seg>^M
  </tuv>^M
  <tuv xml:lang="en-GB">^M
    <seg>The objective for granting a contract for a further fixed term is that the Údarás is awaiting permission to
announce the position of chief executive and to appoint a new chief executive as your successor.</seg>^M
  </tuv>^M
</tu>^M
```

Translation Memory files are XML encoding of parallel texts

SDL Trados translation files are in **.TMX** format:

```
<tu reationdate="20100727T135602Z" creationid="AHGI\niarn" changedate="20100727T135602Z" changeid="AHGI\niarn"
lastusedate="20100727T135602Z" >^M
  <prop type="x-Context">-1853973948243398551, -5048276635724717248</prop>^M
  <prop type="x-Origin">TM</prop>^M
  <prop type="x-ConfirmationLevel">Translated</prop>^M
  <tuv xml:lang="ga-IE">^M
    <seg>Is í an chuis oibiachtúil a bhfuil conradh do théarma socraithe eile á thabhairt duit go bhfuil an tÚdarás ag
fanacht go fóill ar chead chun post an phríomhfheidhmeannaigh a fhógairt agus príomhfheidhmeannach úr a cheapadh mar
chomharba ort.</seg>^M
  </tuv>^M
  <tuv xml:lang="en-GB">^M
    <seg>The objective for granting a contract for a further fixed term is that the Údarás is awaiting permission to
announce the position of chief executive and to appoint a new chief executive as your successor.</seg>^M
  </tuv>^M
</tu>^M
```

Translation Memory files are XML encoding of parallel texts

SDL Trados translation files are in **.TMX** format:

```
<tu creationdate="20100727T135602Z" creationid="AHGI\niarn" changedate="20100727T135602Z" changeid="AHGI\niarn"
lastusedate="20100727T135602Z">^M
  <prop type="x-Context">-1853973948243398551, -5048276635724717248</prop>^M
  <prop type="x-Origin">TM</prop>^M
  <prop type="x-ConfirmationLevel">Translated</prop>^M
  <tuv xml:lang="ga-IE">^M
    <seg>Is í an chuis oibiachtúil a bhfuil conradh do théarma socraithe eile á thabhairt duit go bhfuil an tÚdarás ag
fanacht go fóill ar chead chun post an phríomhfheidhmeannaigh a fhógairt agus príomhfheidhmeannach úr a cheapadh mar
chomharba ort.</seg>^M
  </tuv>^M
  <tuv xml:lang="en-GB">^M
    <seg>The objective for granting a contract for a further fixed term is that the Údarás is awaiting permission to
announce the position of chief executive and to appoint a new chief executive as your successor.</seg>^M
  </tuv>^M
</tu>^M
```

Translation Memory files are XML encoding of parallel texts

SDL Trados translation files are in **.TMX** format:

```
<tu reationdate="20100727T135602Z" creationid="AHGI\niarn" changedate="20100727T135602Z" changeid="AHGI\niarn"
lastusedate="20100727T135602Z" >^M
  <prop type="x-Context">-1853973948243398551, -5048276635724717248</prop>^M
  <prop type="x-Origin">TM</prop>^M
  <prop type="x-ConfirmationLevel">Translated</prop>^M
  <tuv xml:lang="ga-IE">^M
    <seg>Is í an chuis oibiachtúil a bhfuil conradh do théarma socraithe eile á thabhairt duit go bhfuil an tÚdarás ag
    fanacht go fóill ar chead chun post an phríomhfheidhmeannaigh a fhógairt agus príomhfheidhmeannach úr a cheapadh mar
    chomharba ort.</seg>^M
  </tuv>^M
  <tuv xml:lang="en-GB">^M
    <seg>The objective for granting a contract for a further fixed term is that the Udarás is awaiting permission to
    announce the position of chief executive and to appoint a new chief executive as your successor.</seg>^M
  </tuv>^M
</tu>^M
```

Irish segment

English translation

The screenshot displays the Adapt Translation Memory tool interface. The top toolbar includes various functions such as Project Settings, Copy, Paste, Batch Tasks, File Actions, Formatting, QuickInsert, Concordance Search, Select Next Match, Apply Translation, Add New Term, Quick Add New Term, Copy Source to Target, Copy All Source to Target, Clear Target Segment, Segment Actions, Go To, and Navigation.

The main workspace is divided into several panes:

- Editor:** Shows the source text in Gaeilge (Irish) and its English translation. The source text includes: "Speaking at the announcement of the Peace Proms, Minister Heather Humphreys said: 1 Speaking at the launch announcement of the Peace Proms, Minister Heather Humphreys said: 2 Speaking at the announcement of the Peace Proms, Minister Heather Humphreys said: 3". The English translation includes: "The young performers, who range in age from 10 to 15, have been drawn from across the island of Ireland, and are all exceptional and award-winning artists and performers. 7 Speaking at the announcement of the Peace Proms, Minister Heather Humphreys said: 8 'This promises to be an extremely exciting and uplifting event; the ideal start to our year of commemorations in 2016. 9 The Peace Proms will be a wonderful occasion, bringing together talented and committed young musicians from across the island of Ireland. 10 I would like to thank all of the young performers who will be taking part; through their music and song they will set the perfect tone as kick off what is set to be a very special year.' 11".
- Translation Results:** Shows the translation progress and results. It includes a table with columns for the source text, the target text, and the translation status. The table shows that the translation is in progress, with a status of "Draft" and a score of 0%.
- Term Recognition:** A list of terms recognized by the tool, including "speaking", "announcement", "peace", "prom", "prom", "programmable read-only memory", and "Minister Humphreys".
- Project Settings:** A panel on the left side of the interface showing project details, including the project name "PR Peace Proms - Gaeilge", the project type "Translation", and the project status "Draft".
- Translation Details:** A panel on the right side of the interface showing the translation status, origin, and score. It indicates that the translation is in progress, with a status of "Draft", an origin of "Interactive", and a score of 0%.

The bottom status bar shows the current segment (All segments), the current language (INS), and the current progress (53.29%, 32.06%, 14.65%). It also displays the character count (Chars: 0 / 0/2630) and the current language (INS).

What is a terminology?

A list of special words or expressions used in relation to a particular subject or activity.

AutoSave OFF

bankingterminology.xml - Read-Only

Cut Copy Paste Format

Calibri (Body) 12

Wrap Text Merge & Center

General

Conditional Formatting Format as Table Cell Styles

Insert Delete

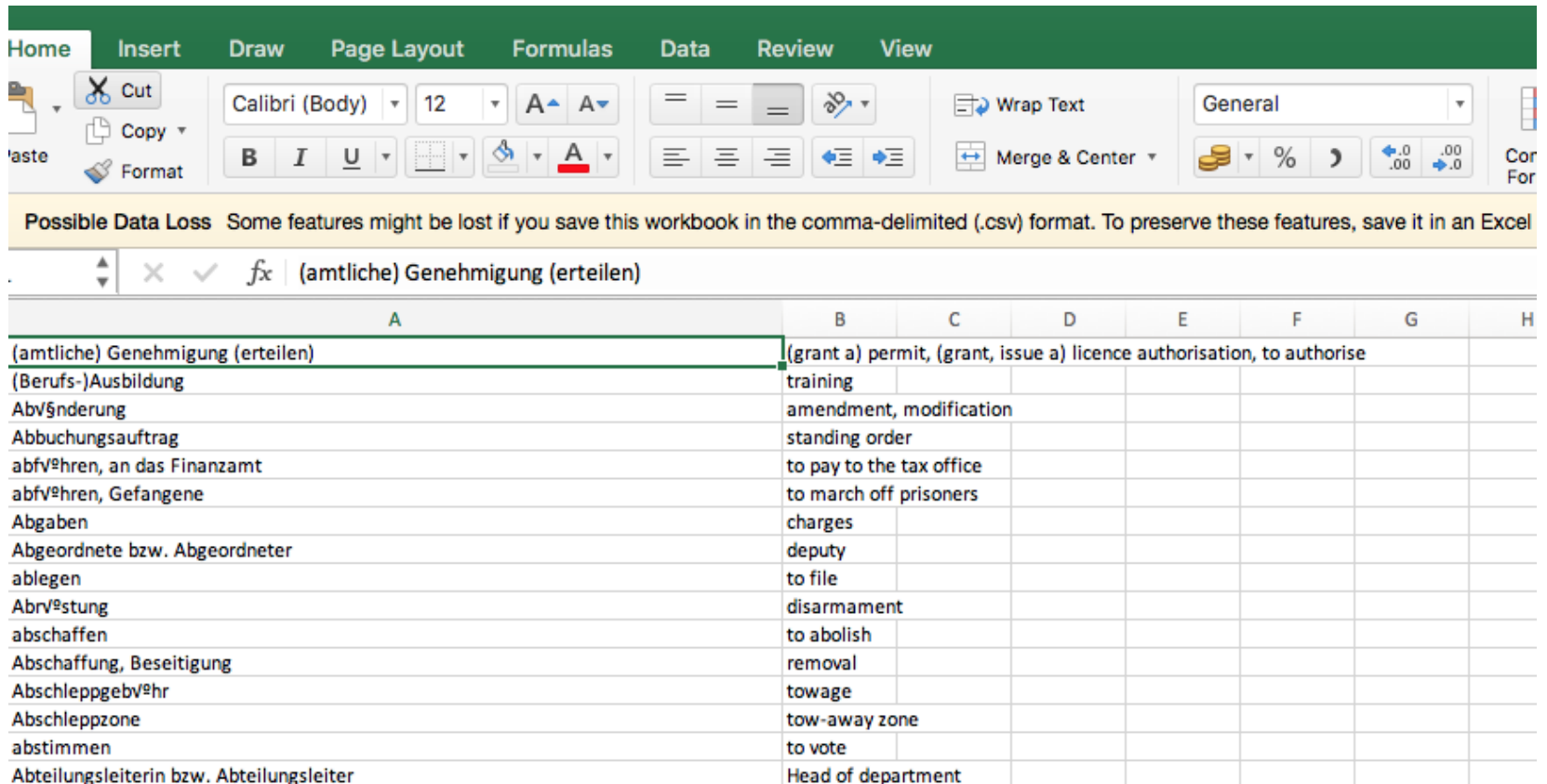
2019

fx Short exchange

A	B	C	D
/dict			
/term/cesdes	/term/cesterm	/term/engdes	/term/engterm
Daň uvalená na zisky společnosti.	Korporátní daň (daň z příjmů právnických osob)	Tax levied on the profits of a company.	Corporate tax
Dluhový papír vydaný společností za účelem navýšení jejího kapitálu.	Korporátní obligace	Debt issued by corporations as a means of raising capital	Corporate bond
Skupina akcií sestavená za účelem koupě nebo prodeje jako celku, obvykle	Koš (ve vztahu k derivátům)	Relates to derivative products. Group of stocks that is for	Basket
Opce umožňující držitelovi zachovat úrok z koše komodit/cenných papírů.	Košová opce	Option that allows the holder to maintain an interest in	Basket option
Obvykle dvoucestná cena, za níž je market maker (tvůrce trhu) připraven	Kotace	Normally a two way price at which a market maker is prepared	Quote
Elektronicky řízená tabule v makléřské firmě, která zobrazuje aktuální ceny	Kotovací tabule	An electronically controlled board at a brokerage firm that	Quotation board
Cena, za níž se uskutečnil poslední prodej a nákup určitého cenného papíru	Kotovaná cena	The price at which the last sale and purchase of a particular	Quoted price
Společnost uvedená na oficiální burze, jejíž akcie jsou "kotovány" a dostupné	Kotovaná společnost	One listed on an official stock exchange; its shares are "quoted"	Quoted company
Cena nabízená za koupi nebo prodej určitého cenného papíru.	Kotování	The price being offered or bid for a particular security.	Quotation
Objednávka na nákup při zahájení obchodního dne za cenu v otevřacím rozpisu	Koupě při otevření	An order to buy once the trading session has started at a particular	Buy on opening
Objednávka na nákup akcie na konci obchodního dne za cenu v uzavíracím rozpisu	Koupě při uzavření	An order to buy at the end of a trading session at a particular	Buy on close
Míra společného pohybu (regrese) dvou proměnných.	Kovariance	A measure of co-movement between two variables.	Covariance
Uplatnění práva bankovního zástupce. Označení "fiduciářská" se používá pro	Kovariance (fiduciářská) měření (standard)	Refers to banknotes and coins. These means of payment are	Fiduciary money (as standard)

What is a Glossary?

An alphabetical list, with meanings, of the words or phrases in a text that are difficult to understand.



A	B	C	D	E	F	G	H
(amtliche) Genehmigung (erteilen)	(grant a) permit, (grant, issue a) licence authorisation, to authorise						
(Berufs-)Ausbildung	training						
Abv�nderung	amendment, modification						
Abbuchungsauftrag	standing order						
abf�hren, an das Finanzamt	to pay to the tax office						
abf�hren, Gefangene	to march off prisoners						
Abgaben	charges						
Abgeordnete bzw. Abgeordneter	deputy						
ablegen	to file						
Abr�stung	disarmament						
abschaffen	to abolish						
Abschaffung, Beseitigung	removal						
Abschleppgeb�hr	towage						
Abschleppzone	tow-away zone						
abstimmen	to vote						
Abteilungsleiterin bzw. Abteilungsleiter	Head of department						

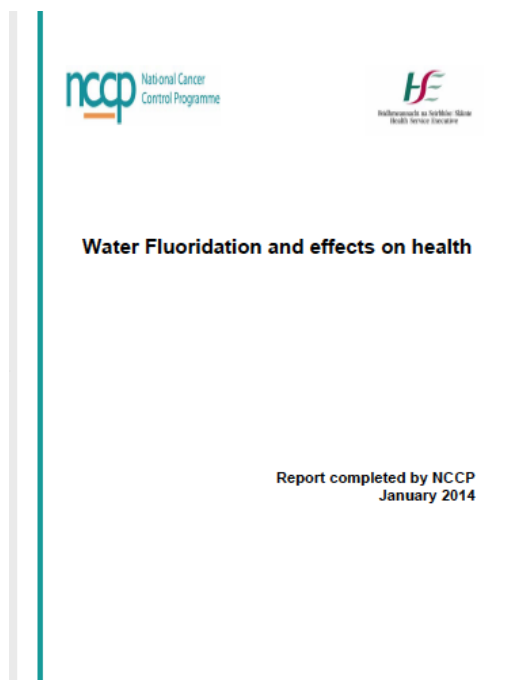
Glossary City of Vienna – downloaded from ELRC-SHARE. Provided by Executive Office, Verwaltungsakademie Vienna

Where to find language data?



Examples of digital content (data) in public organisations, produced in-house or outsourced:

- Reports
- Communication
- Policies



Examples of digital content (data) in public organisations, produced in-house or outsourced:

- Reports
- Communication
- Policies

Government Contributes to Fund to help Developing Countries Respond to Climate Change

The Minister for Communications, Climate Action and Environment, Minister Richard Bruton T.D. has secured government approval to make a contribution for Ireland to the replenishment of the Global Environment Facility Trust Fund (GEF). Ireland will contribute €5.73 million, paid in annual instalments, starting this year and continuing until 2022.

The GEF is a Fund managed by the United Nations and World Bank which provides grants to developing countries for a range of environmental projects, including to make the adjustments necessary to respond to climate change. Since its establishment in 1991, the GEF has provided over \$18 billion in grants and mobilised an additional \$93 billion in financing for more than 4,500 projects worldwide. The Fund is now being replenished for the 7th time.



Water Fluoridation and effects on health

Report completed by NCCP
January 2014



Examples of digital content (data) in public organisations, produced in-house or outsourced:

- Reports
- Communication
- Policies

Government Contributes to Fund to help Developing Countries Respond to Climate Change

The Minister for Communications, Climate Action and Environment, Minister Richard Bruton T.D. has secured government approval to make a contribution for Ireland to the replenishment of the Global Environment Facility Trust Fund (GEF). Ireland will contribute €5.73 million, paid in annual instalments, starting this year and continuing until 2022.

The GEF is a Fund managed by the United Nations and World Bank which provides grants to developing countries for a range of environmental projects, including to make the adjustments necessary to respond to climate change. Since its establishment in 1991, the GEF has provided over \$18 billion in grants and mobilised an additional \$93 billion in financing for more than 4,500 projects worldwide. The Fund is now being replenished for the 7th time.



Water Fluoridation and effects on health

Report completed by NCCP
January 2014

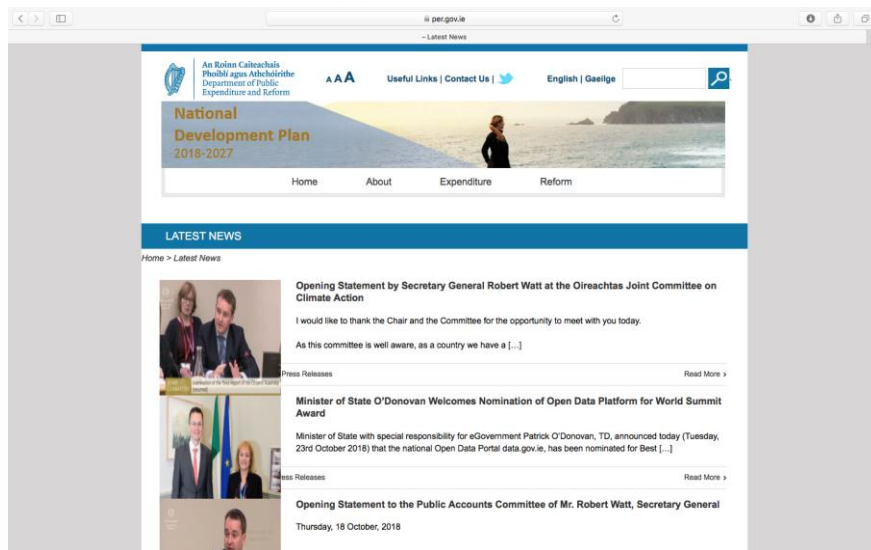


Contains Irish Public Sector Information licensed under a Creative Commons Attribution 4.0 International ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)) licence

Where to find useful language data for MT?

www.adaptcentre.ie

- News
- Web Content
- Archives



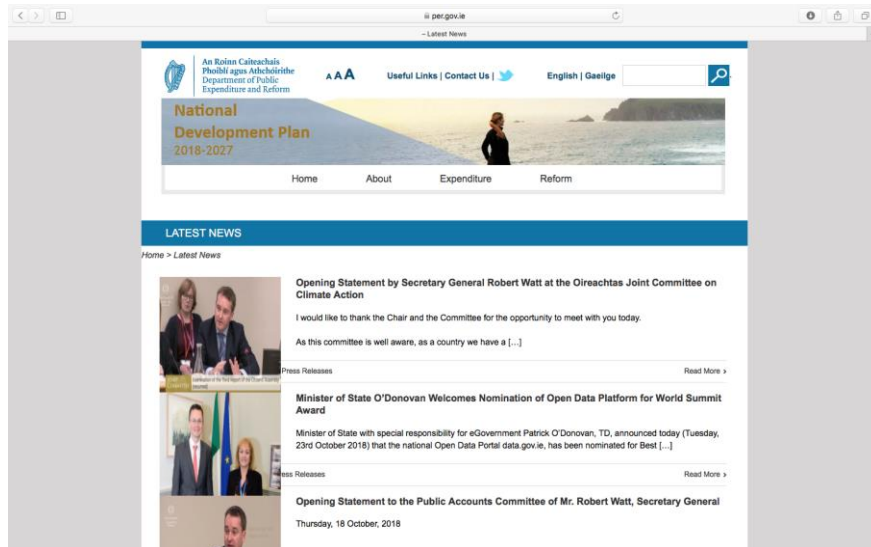
Contains Irish Public Sector Information licensed under a Creative Commons Attribution 4.0 International ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)) licence



Where to find useful language data for MT?

www.adaptcentre.ie

- News
- Web Content
- Archives

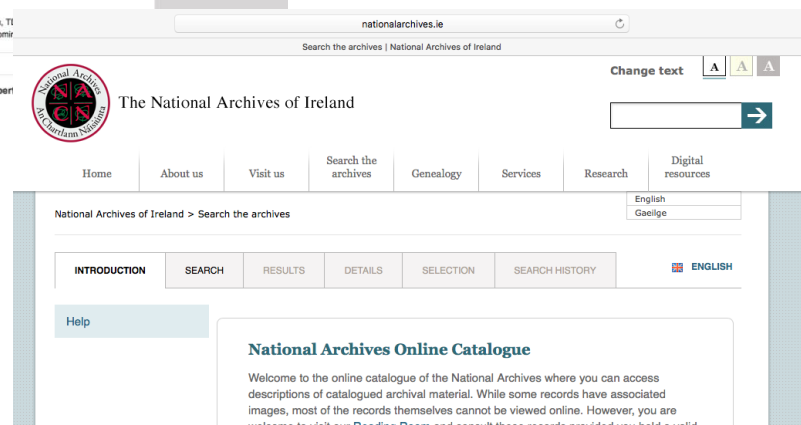
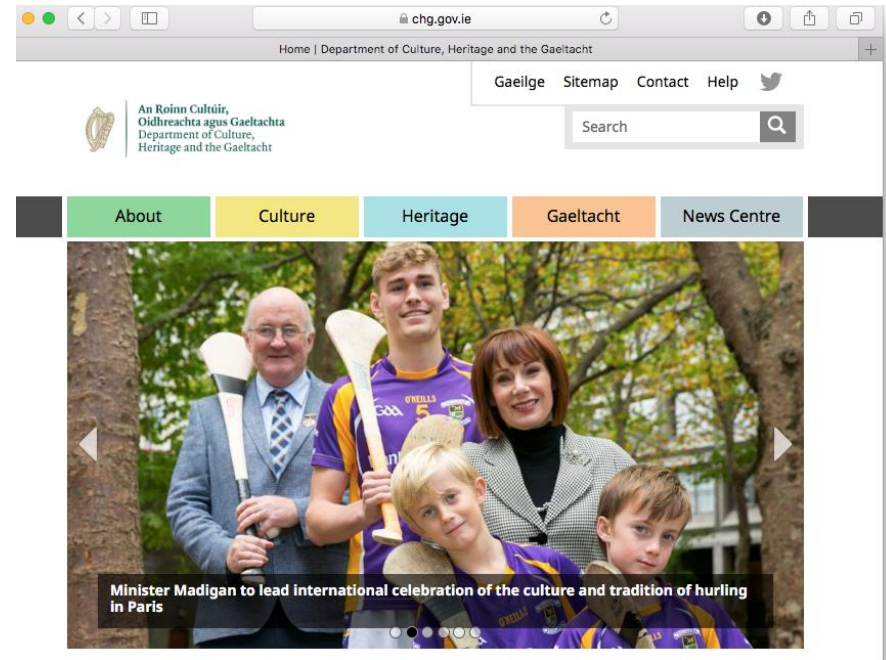
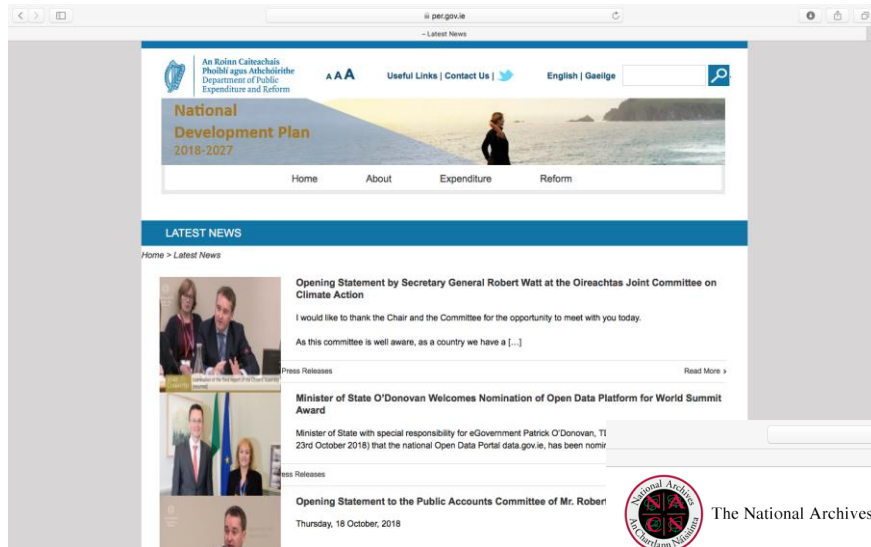


Contains Irish Public Sector Information licensed under a Creative Commons Attribution 4.0 International ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)) licence

Where to find useful language data for MT?

www.adaptcentre.ie

- News
- Web Content
- Archives



Contains Irish Public Sector Information licensed under a Creative Commons Attribution 4.0 International ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)) licence



Where to find useful language data for MT?

- Forms
- FAQs

Application Form for a Driving Licence D401



Please read accompanying guidance notes before completing this form. **Please complete this form in block capitals using a black ballpoint pen. Please place an X in the appropriate boxes e.g. ☒ Please do not photocopy this form as it may reduce its quality and result in your application being delayed or rejected.**

Part 1: Personal Details (See Part 1 of accompanying guidance notes)

*Mandatory field

- Have you previously held a learner permit and/or driving licence in Ireland?*
- Title
- First name(s)*
- Surname *

Yes ☐ No ☐

If yes, which one? Learner Permit ☐ Driving Licence ☐

Driver number* (if known)
(You will find this on Field 5 of the paper licence or Field 4d on a plastic card licence.)

Mr ☐ Mrs ☐ Miss ☐ Ms ☐ Other (please specify)

Name to appear on the licence. Acceptable photo ID must be provided in this name. Refer to list 1 on page 2 of the accompanying guidance notes.

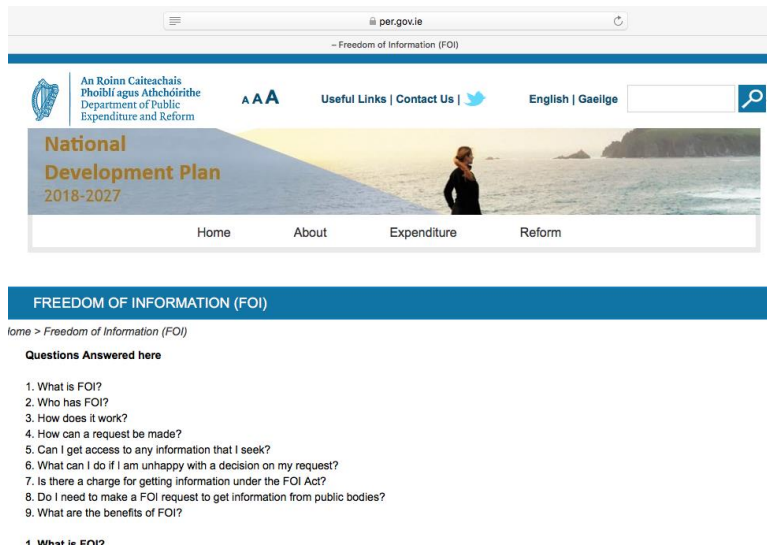
As it appears on your birth certificate



Where to find useful language data for MT?

www.adaptcentre.ie

- Forms
- FAQs



Application Form for a Driving Licence D401



Please read accompanying guidance notes before completing this form. **Please complete this form in block capitals using a black ballpoint pen. Please place an X in the appropriate boxes e.g. ☒ Please do not photocopy this form as it may reduce its quality and result in your application being delayed or rejected.**



Part 1: Personal Details (See Part 1 of accompanying guidance notes)

*Mandatory field

1. Have you previously held a learner permit and/or driving licence in Ireland?*

Yes ☐ No ☐

If yes, which one?

Learner Permit ☐

Driving Licence ☐

Driver number* (if known)

(You will find this on Field 5 of the paper licence or Field 4d on a plastic card licence.)

2. Title

Mr ☐

Mrs ☐

Miss ☐

Ms ☐

Other (please specify)

Name to appear on the licence. Acceptable photo ID must be provided in this name. Refer to list 1 on page 2 of the accompanying guidance notes.

3. First name(s)*

4. Surname *

As it appears on your birth certificate

Contains Irish Public Sector Information licensed under a Creative Commons Attribution 4.0 International ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)) licence

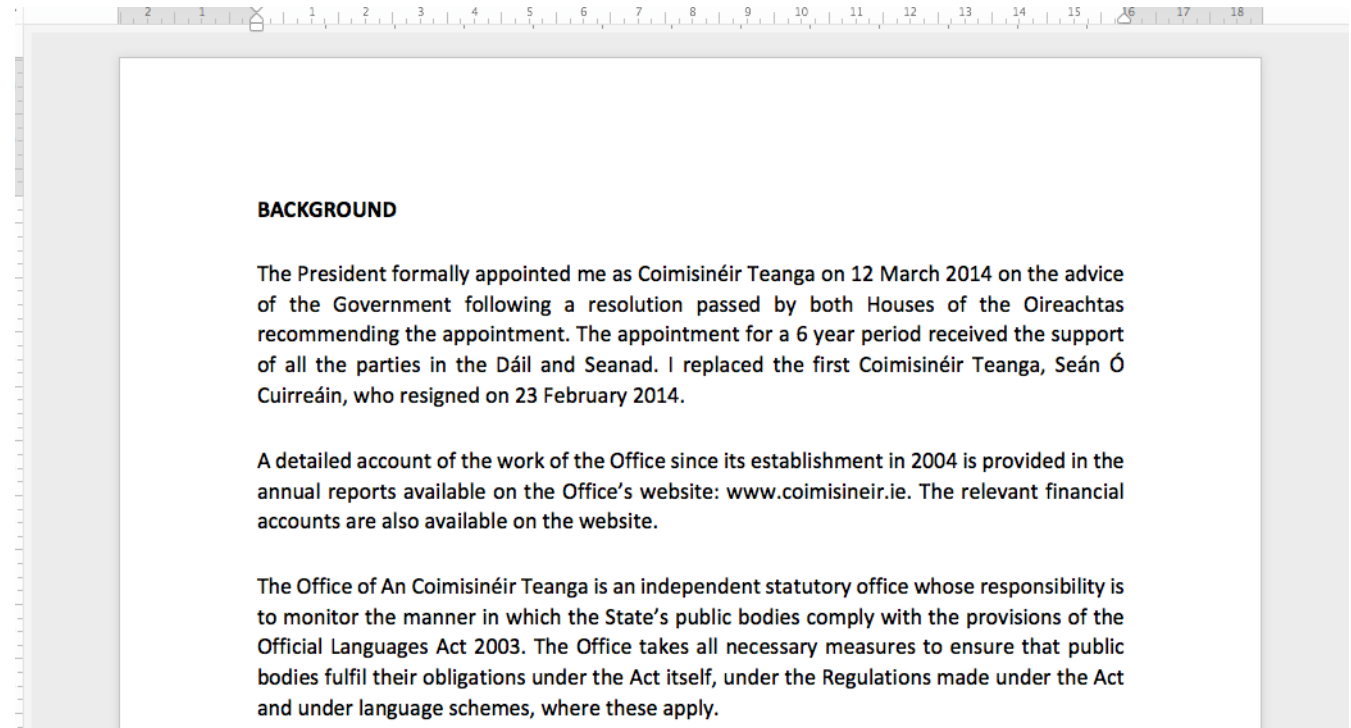


- Language Commissioner – Annual Reports, press releases, financial reports, complaints received, guidebook for Language Act
- Department of Justice – legislative acts
- Universities – signage, course content, application forms, flyers, web content
- Citizen's Information – website content e.g. social welfare, consumer affairs information, application forms
- Foras na Gaeilge (Irish language body) – dictionaries, monolingual national corpus
- Kings Inns (Law Society) – monolingual training material



Ideal File Content





Report from the Language Commissioner's Office



```
gaois.en.txt — Edited v
4 S.I. No. 171 / 1980 : ,
harbour RATES ( DUBLIN HARBOUR ) ORDER , 1980 . ,
( 1 ) This Order may be cited as the Harbour Rates ( Dublin Harbour ) Order , 1980 . ,
( 2 ) This Order shall come into operation on the 18th day of April , 1980 . ,
in this Order &quot; qu... the Board &quot; qu... means the Dublin port and Docks
Board . ,
the provisions set out in Part I of the Schedule to this Order shall apply in relation to
the rates chargeable under paragraph 6 of this Order . ,
Schedule . ,
rates based on weight may be converted to measurement rates . ,
unit or package rates . ,
size and type of packages . ,
rates which are chargeable by reference to type of package refer to the size and type of
package customary at the commencement of this Order . ,
average rates . ,
in the case of goods which are warehoused in diverse sizes or weights the Board may base
the charges on average figures or on averages within certain ranges of size or weight . ,
rent chargeable according to space . ,
in warehouse – 46.75p per square metre per week ,
in open yard – 28.20p per square metre per week ,
Rental periods . ,
the rate for the full rental period to apply to any fractional part thereof . ,
minimum charges . ,
Housing – £ 1.64 ,
Unhousing – £ 1.64 ,
rent – £ 26.13p per week or £ 1.05 per month . ,
( b ) In the absence of special circumstances the minimum rates shall not be applied where
average rates are charged or where occasional variations occur in traffic handled at the
standard rates . ,
increased rates on general goods in special circumstances . ,
```

Bilingual corpus of English-Irish legislation provided by the Department of Justice

1480

ΕΦΗΜΕΡΙΣ ΤΗΣ ΚΥΒΕΡΝΗΣΕΩΣ (ΤΕΥΧΟΣ ΠΡΩΤΟ)

United Nations Convention against Corruption

Preamble

The States Parties to this Convention,

Concerned about the seriousness of problems and threats posed by corruption to the stability and security of societies, to the integrity of institutions and values of democracy, ethical values and justice and to sustainable development and the rule of law,

Concerned also about the links between corruption and other forms of crime, in particular organized crime and economic crime, including money-laundering,

Concerned further about cases of corruption which involve vast quantities of assets, which may constitute a substantial proportion of the resources of States, and that threaten the political stability and sustainable development of those States,

Convinced that corruption is no longer a local matter but a transnational phenomenon that affects all societies and economies, making international cooperation to prevent and control it essential,

Convinced also that a comprehensive and multidisciplinary approach is required to prevent and combat corruption effectively



Don'ts



This is an English paragraph. This is an English paragraph. This is an English paragraph. This is an English paragraph. This is an English paragraph. This is an English paragraph. This is an English paragraph. This is an English paragraph.

Αυτή είναι η Ελληνική μετάφραση της παραπάνω παραγράφου. Αυτή είναι η Ελληνική μετάφραση της παραπάνω παραγράφου. Αυτή είναι η Ελληνική μετάφραση της παραπάνω παραγράφου. Αυτή είναι η Ελληνική μετάφραση της παραπάνω παραγράφου. Αυτή είναι η Ελληνική μετάφραση της παραπάνω παραγράφου. Αυτή είναι η Ελληνική μετάφραση της παραπάνω παραγράφου. Αυτή είναι η Ελληνική μετάφραση της παραπάνω παραγράφου. Αυτή είναι η Ελληνική μετάφραση της παραπάνω παραγράφου.

This is another English paragraph. This is an English paragraph. This is an English paragraph. This is an English paragraph. This is an English paragraph. This is an English paragraph. This is an English paragraph. This is an English paragraph.

Αυτή είναι η Ελληνική μετάφραση της παραπάνω παραγράφου. Αυτή είναι η Ελληνική μετάφραση της παραπάνω παραγράφου. Αυτή είναι η Ελληνική μετάφραση της παραπάνω παραγράφου. Αυτή είναι η Ελληνική μετάφραση της παραπάνω παραγράφου. Αυτή είναι η Ελληνική μετάφραση της παραπάνω παραγράφου. Αυτή είναι η Ελληνική μετάφραση της παραπάνω παραγράφου. Αυτή είναι η Ελληνική μετάφραση της παραπάνω παραγράφου. Αυτή είναι η Ελληνική μετάφραση της παραπάνω παραγράφου.

This is an English sentence.

Αυτή είναι η μετάφραση της παραπάνω πρότασης.



This-is-an-English-paragraph.This-is-an-
English-paragraph.This-is-an-English-
paragraph.This-is-an-English-paragraph.
This-is-an-English-paragraph.This-is-an-
English-paragraph.This-is-an-English-
paragraph.→ ¶

१

9

This-is-another-English-paragraph..This-is-an-English-paragraph..This-is-an-English-paragraph..This-is-an-English-paragraph..This-is-an-English-paragraph..This-is-an-English-paragraph..¶

9

This is an English sentence.¶

Αυτή·είναι·η·Ελληνική·μετάφραση·της·
 παραπάνω·παραγράφου·Αυτή·είναι·η·
 Ελληνική·μετάφραση·της·παραπάνω·
 παραγράφου·Αυτή·είναι·η·Ελληνική·
 μετάφραση·της·παραπάνω·παραγράφου·
 Αυτή·είναι·η·Ελληνική·μετάφραση·της·
 παραπάνω·παραγράφου·Αυτή·είναι·η·
 Ελληνική·μετάφραση·της·παραπάνω·
 παραγράφου·¶

Αυτή·είναι·η·Ελληνική·μετάφραση·της·
 παραπάνω·παραγράφου·Αυτή·είναι·η·
 Ελληνική·μετάφραση·της·παραπάνω·
 παραγράφου·Αυτή·είναι·η·Ελληνική·
 μετάφραση·της·παραπάνω·παραγράφου·
 Αυτή·είναι·η·Ελληνική·μετάφραση·της·
 παραπάνω·παραγράφου·Αυτή·είναι·η·
 Ελληνική·μετάφραση·της·παραπάνω·
 παραγράφου·¶

Αυτή·είναι·η·μετάφραση·της·παραπάνω·
πρότασης.¶

Don'ts



English	Greek
This is the content of an English document. This is the content of an English document. This is the content of an English document. This is the content of an English document. This is the content of an English document. This is the content of an English document. This is the content of an English document. This is the content of an English document. This is the content of an English document. This is the content of an English document. This is the content of an English document. This is the content of an English document. This is the content of an English document.	Αυτή είναι η Ελληνική μετάφραση του κειμένου. Αυτή είναι η Ελληνική μετάφραση του κειμένου. Αυτή είναι η Ελληνική μετάφραση του κειμένου. Αυτή είναι η Ελληνική μετάφραση του κειμένου. Αυτή είναι η Ελληνική μετάφραση του κειμένου. Αυτή είναι η Ελληνική μετάφραση του κειμένου. Αυτή είναι η Ελληνική μετάφραση του κειμένου. Αυτή είναι η Ελληνική μετάφραση του κειμένου. Αυτή είναι η Ελληνική μετάφραση του κειμένου. Αυτή είναι η Ελληνική μετάφραση του κειμένου.

Ideal File Formats for MT training



- In principle, any text in machine readable format
- But, some formats are more “MT-ready” than others, i.e. they require less manual or automatic processing (e.g. Word Doc vs PDF)
- More processing introduces more errors in the final output, making it less useful to train MT systems

- Word Docs (.doc, .docx, .odt)
- PDF
- Terminology database (.tbx)
- Translation Memories (.tmx, sdltm)
- XML files (.xml)
- Plain text (.txt)
- Spreadsheets (.xlsx, .xls)
- ZIP files



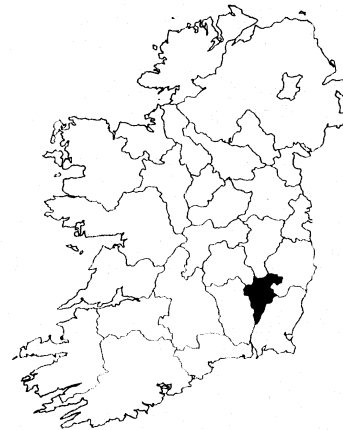
Don'ts



Recorded Monuments

Protected under Section 12 of the
National Monuments (Amendment) Act, 1994

County Carlow



Archaeological Survey of Ireland
The Office of Public Works

Don'ts



970MM X 650MM

COLLAISTE ALBERT
DCU
ALBERT COLLEGE

Urrár na Talún
Ground Floor

2	Oifig an Phríomh-Oifigigh Oibríochán Chief Operations Officer	←
2	Leas-Uachtarán um Ghnóthas Acadúil Vice-President Academic Affairs	←
2	Seomra Cruinnithe (A204) Meeting Room (A204)	←
1	Oifig an Uachtaráin President's Office	←
1	Oifig an Leas-Uachtaráin Deputy-President's Office	←
1		
G	Club 1838 1838 Club	↑
G	Seomra Comhairle (AG01) Boardroom (AG01)	→
G	Seomra Cruinnithe an Uachtaráin (AG04) President's Meeting Room (AG04)	←
G		
G	Shíneadh le Collaiste Albert Albert College Extension	→

NEW NOV 2017
420MM X 450MM

COLLAISTE ALBERT
DCU
ALBERT COLLEGE

An Chéad Lár
First Floor

Oifig an Uachtaráin
President's Office

Oifig an Leas-Uachtaráin
Deputy-President's Office

Lárnaí
Lobbies

COLLAISTE ALBERT
DCU
ALBERT COLLEGE

An Dara Talún
Second Floor

Phríomh-Oifigigh Oibríochán
Chief Operations Officer

Seomra Cruinnithe M204
Meeting Room (M204)

Leas-Uachtarán um Ghnóthas Acadúil
Vice-President Academic Affairs

Lárnaí
Lobbies

An tríú lár
Located on First Floor

3NO 6MM GLASS/LOOK PERSPEX PANELS WITH PRINTED BANDS TO FACE
MOUNTED ON ANODISED ALUMINIUM STUD FIXINGS

ACADEMY
SIGNS

DRAWING NAME	ALBERTCOLLEGE.PLT	REVISIONS			LED TYPE	SIGNATURE FOR APPROVAL OR EMAIL CONFIRMATION REQUIRED:	
CLIENT	DCU ALBERT COLLEGE	NO.	DATE	DETAILS			
DRAWN BY	KEITH HYNES		01.11.17		PHOTOREF.		
		2.0	24.01.18	JAMES WHELAN			
ACADEMY SIGNS, RARE 98 NUTGRAVE AVENUE, RATHFARNHAM, DUBLIN 14. T. 01-4935188, F. 01-4937337 W. www.academydesigns.com							

PLEASE NOTE:
PROOFREAD THIS DOCUMENT CAREFULLY AS
A SIGNED COPY OF THIS DOCUMENT IS REQUIRED BY ACADEMY SIGNS
ALL DRAWINGS TO BE VERIFIED BY THE CLIENT
VISUAL REPRESENTATION: NONE CANNOT BE GUARANTEED
ALL PRICES AND SPECIFICATIONS LISTED
BY ACADEMY SIGNS ARE THE PROPERTY
OF ACADEMY SIGNS AND PROTECTED
BY COPYRIGHT ©

PROOF READ THIS PDF CAREFULLY AS
RESPONSIBILITY FOR ANY ERRORS OR OMISSIONS
WILL BE WITH THE CLIENT UPON SIGN OFF.
ANY PRINTS/COPIES WILL BE REPRODUCED (COPY
AND PASTE) AND REPRESENT THE FINAL PRINTED OR
PHOTOGRAPHIC QUALITY OF PRINT.
A.P.S. IS ONLY A COLOUR AND DRAWING
REPRESENTATION AND WILL NOT REPRODUCE
DETAILS, LINES, PATTERN OR SHADY SHAPES (TAPER
EFFECTS)

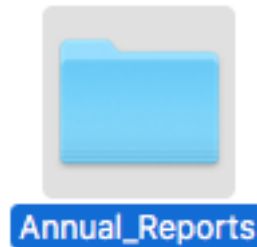
Campus signage courtesy of DCU

Language data management tips – going forward





Create a folder structure to easily manage paired translation files



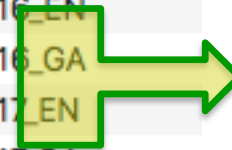


Name	
	Annual_Report_2015_EN
	Annual_Report_2015_GA
	Annual_Report_2016_EN
	Annual_Report_2016_GA
	Annual_Report_2017_EN
	Annual_Report_2017_GA
	Annual_Report_2018_EN
	Annual_Report_2018_GA

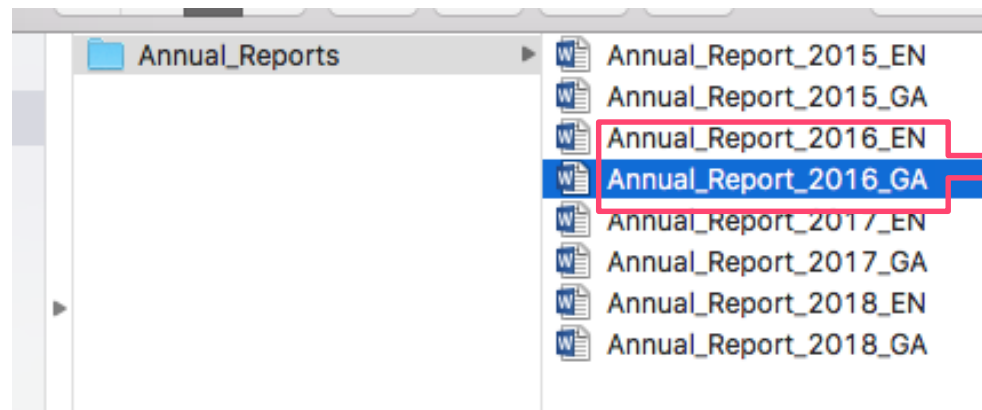
Use an identical filename for each document pair



Name	
	Annual_Report_2015_EN
	Annual_Report_2015_GA
	Annual_Report_2016_EN
	Annual_Report_2016_GA
	Annual_Report_2017_EN
	Annual_Report_2017_GA
	Annual_Report_2018_EN
	Annual_Report_2018_GA



Include language
identifiers in the
filename



Each file is monolingual with its translation in a different file.

So at the end, you have a folder with document pairs that are easy to find and understand.

- A dataset is a collection of data **grouped according to certain criteria**
 - **Language(s)**: each collection is defined by the language or language pairs of its data, e.g.
 - *Parallel* (e.g. English – Irish)
 - *Multilingual* (e.g. English – Portuguese – Spanish – Irish)
 - **Domain**: each collection ideally belongs to a single domain, e.g.
 - *Culture* (e.g. 1916 centenary celebration)
 - *Social security*
 - *Legal Text*
 - *Finance*
 - *Foreign Affairs*
 - *Agriculture*
 - ...

- Now that I know the value of data, what should my plans be?
- What are the best ways to collect, maintain, archive and re-use my data?
- In particular how can I use it for improving MT performances?



“Where computer aided translation tooling is used to produce translations then, in addition to the finished target language translation, any and all translation memory data, or similar, derived from the translation process should be returned to the department contracting for translation services as part of the final end product delivery.”

Go raibh maith agat

Gracias

Obrigado

Merci

Thank you!

