

COVID-19 MLIA EVAL

ROUND 1: DATA ACQUISITION

Vassilis Papavassiliou, Stelios Piperidis
{vpapa, spip}@athenarc.gr
Virtual Meeting, 12-14 January 2021



DATA COLLECTION: 1ST ROUND OF THE MT TASK

- From the aspect of data collection (during the 1st round),
- how to simulate a very quick response of the MT community in an emergency situation, like the current pandemic
 - Generate an initial collection of parallel corpora
 - in health and medicine domains
 - from well-known web sources
 - Enrich with identified COVID-19 parallel data.

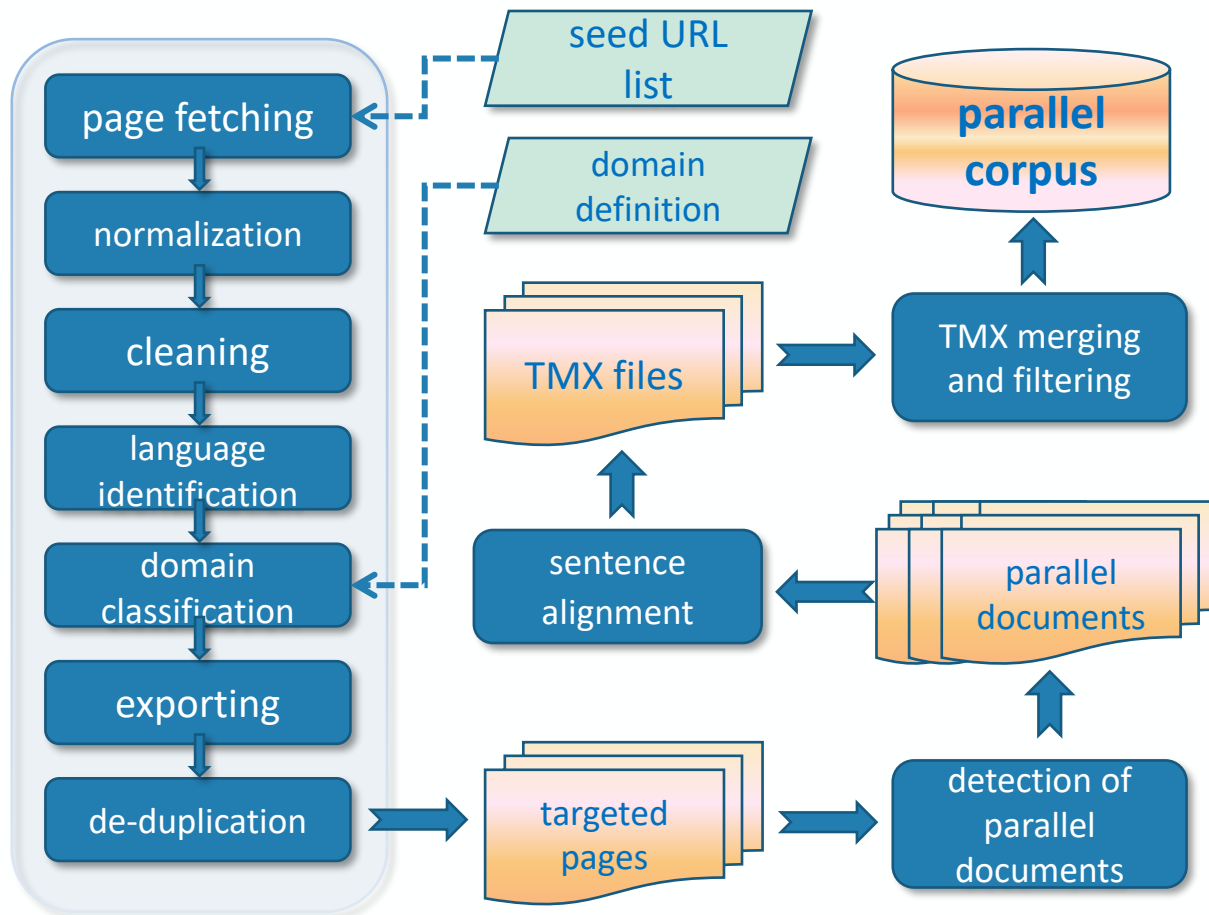
DATA COLLECTION: “GENERAL” SUBSET OF TRAINING DATA

- An updated version of the EMEA corpus by
 - harvesting the website of the European Medicines Agency
 - applying new (more robust and efficient) methods for
 - text extraction from PDF files (enhanced version of the PDFBox library),
 - identification of sentence pairs (LASER toolkit)
 - parallel corpus filtering.
- Medical/Health-related multilingual collections
 - offered by the Publications Office of EU
 - processed in a similar manner.

DATA COLLECTION: “COVID-19” SUBSET OF TRAINING DATA

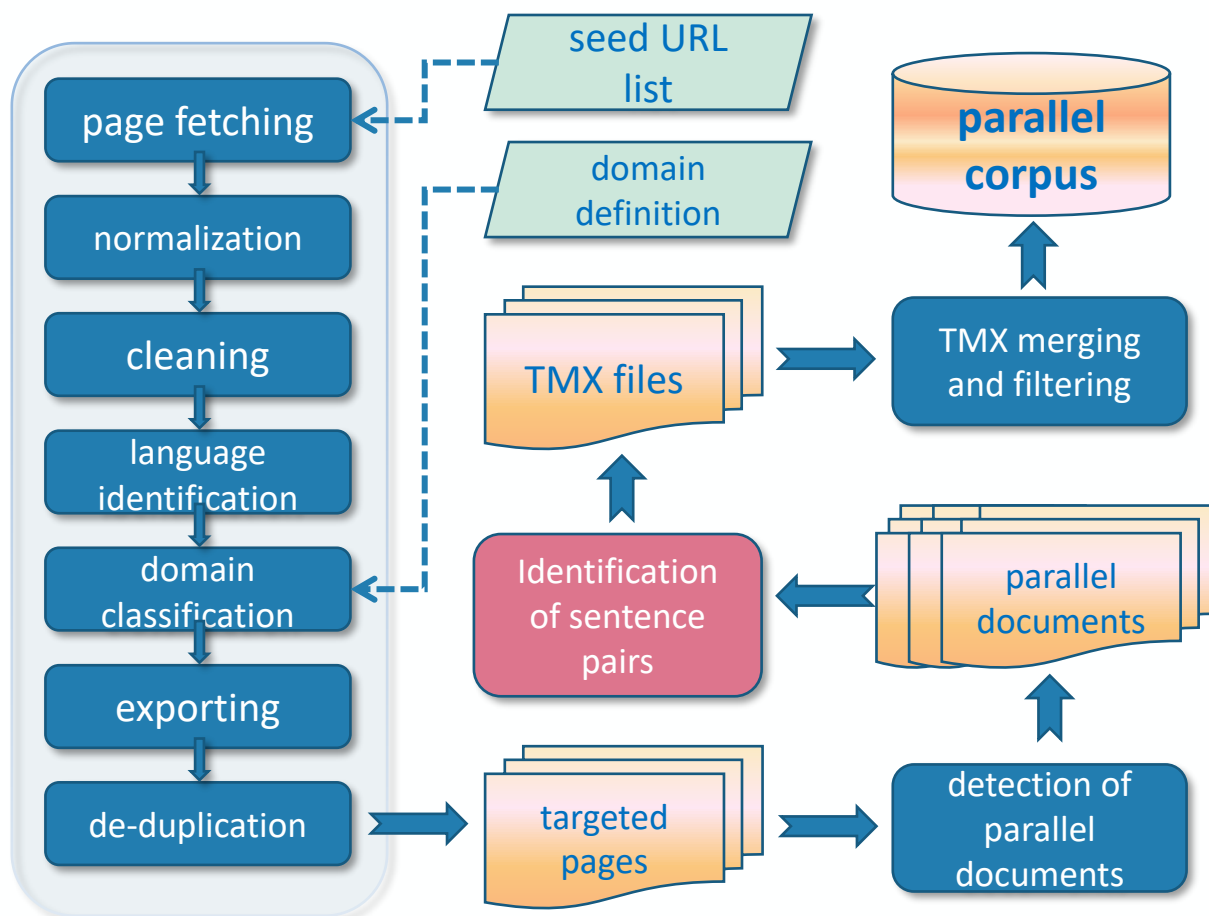
- First step in acquiring COVID-19-related data:
 - Identification of multi/bi-lingual websites with such content.
- With the aim of constructing publicly available data sets,
 - we targeted websites of
 - national authorities and public health agencies (<https://www.ecdc.europa.eu/en/COVID-19/national-sources>),
 - EU agencies and
 - specific broadcast websites (e.g., Voxeurop, GlobalVoices, etc.)
- In the next rounds we plan to also include relevant data from several international organizations and outcomes of broader crawls.

WORKFLOW FOR ACQUIRING COVID-19-RELATED DATA



- A recent version of ILSP-FC, a modular toolkit that integrates modules for
 - text normalization,
 - language identification,
 - document clean-up,
 - text classification,
 - de-duplication,
 - bilingual document alignment and
 - sentence alignment.
- In this emergency situation,
 - a “rapid” approach based on keywords was adopted for text classification

WORKFLOW FOR ACQUIRING COVID-19-RELATED DATA



- For the identification of sentence pairs, the LASER toolkit was used instead of the integrated aligner.
- A battery of criteria was applied on aligned sentences to automatically filter out sentence pairs:
 - with potential alignment or translation issues
 - of limited use for training MT systems
- and, thus, generate precision-high language resources.

DATA COLLECTION: DATASETS FOR 1ST ROUND OF THE MT TASK

- The size of the constructed datasets varies from 0.81 to 1.1M sentence pairs.
- “general” subset covers ~80%
- “covid-19” subset covers ~20%
- The organizers of the MT Task split them into Training, Validation and Test Sets per language pair.

Language Pair	Number of TUs
EN-FR	1 103 974
EN-DE	938 506
EN-ES	1 086 007
EN-IT	909 548
EN-EL	840 943
EN-SV	816 698

DATA COLLECTION: 1ST ROUND OF IE AND MSS TASKS

DATA COLLECTION: 1ST ROUND OF IE AND MSS TASKS

- Description of the 2020 EMM-MediSys COVID19 Dataset
- Postprocessing steps for the IE and MSS tasks

Data Acquisition: Europe Media Monitoring (EMM) and the 2020 EMM-MediSys COVID19 Dataset

Guillaume Jacquet
JRC.I.3 – Joint Research Centre – European Commission

Covid-19 MLIA Virtual Meeting, 12/01/2021

Data Acquisition:

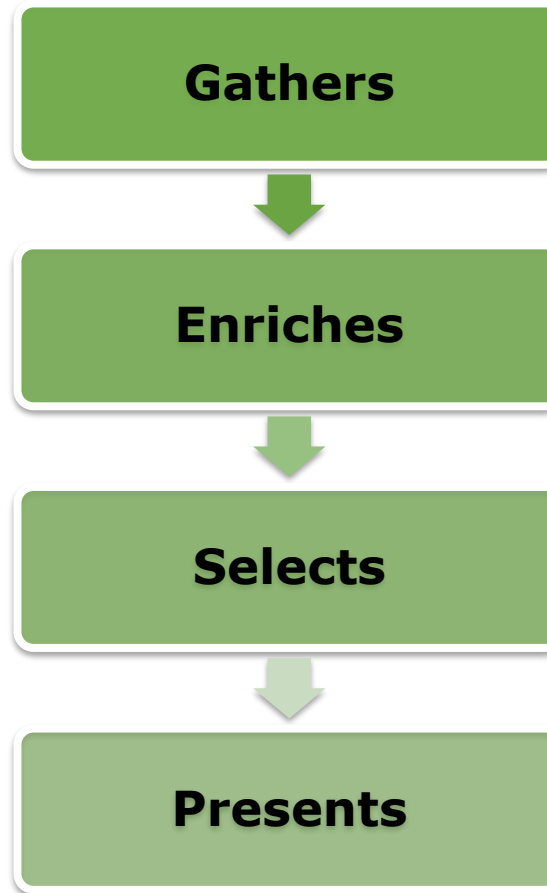
Europe Media Monitoring (EMM)

Guillaume Jacquet
JRC.I.3 – Joint Research Centre – European Commission

Covid-19 MLIA Virtual Meeting, 12/01/2021

EMM: INFORMATION FROM TEXT


Objective: Find knowledge that is hidden in text



Results: Find perhaps 100 relevant articles a day from 300 000

EMM Architecture

Custom Domain (Sources and Topics)



Gathers
~300000
new news articles
per day

70^{In}
languages
~8000
news portals
world-wide

Classifies all news according
to hundreds of subjects and
countries

24
hours
per day

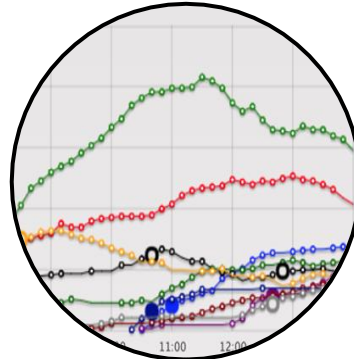
7
days
a week

Operational since
2002

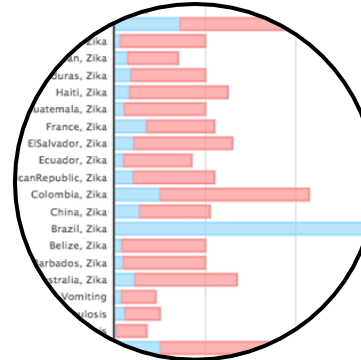
EMM Open Source Monitoring Engine



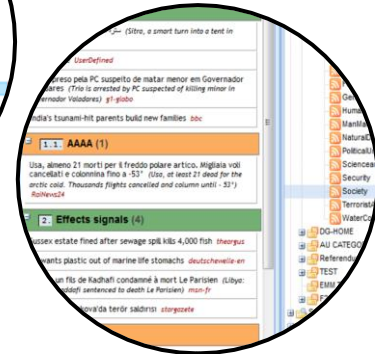
Newsbrief



Medisys



NewsDesk



Other
Uses



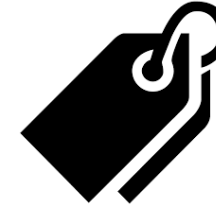
End Use Applications

Structured data: logical view



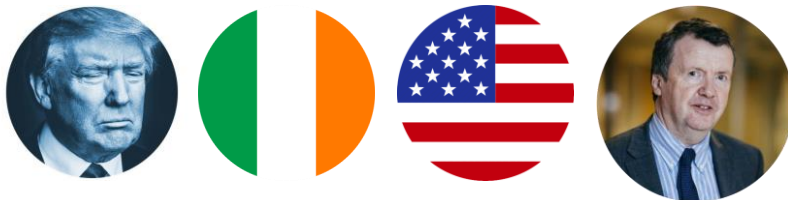
Language:
English

US president Donald Trump intention to cut the US corporate tax rate to 15 per cent may not be achievable with the border adjustment tax, which might have funded the reduction, now off the table. Irish tax experts remain convinced that US tax reform, whatever shape it takes, will not adversely damage Ireland's economic prospects, suggesting incentives for US multinationals to base operations here will continue. *"All in all, I wouldn't be unduly worried at this juncture,"* Alan McQuaid said.



Categorization:
Taxation
Economy
EU-US trade

Entities:



Quotes:

"All in all, I wouldn't be unduly worried at this juncture"

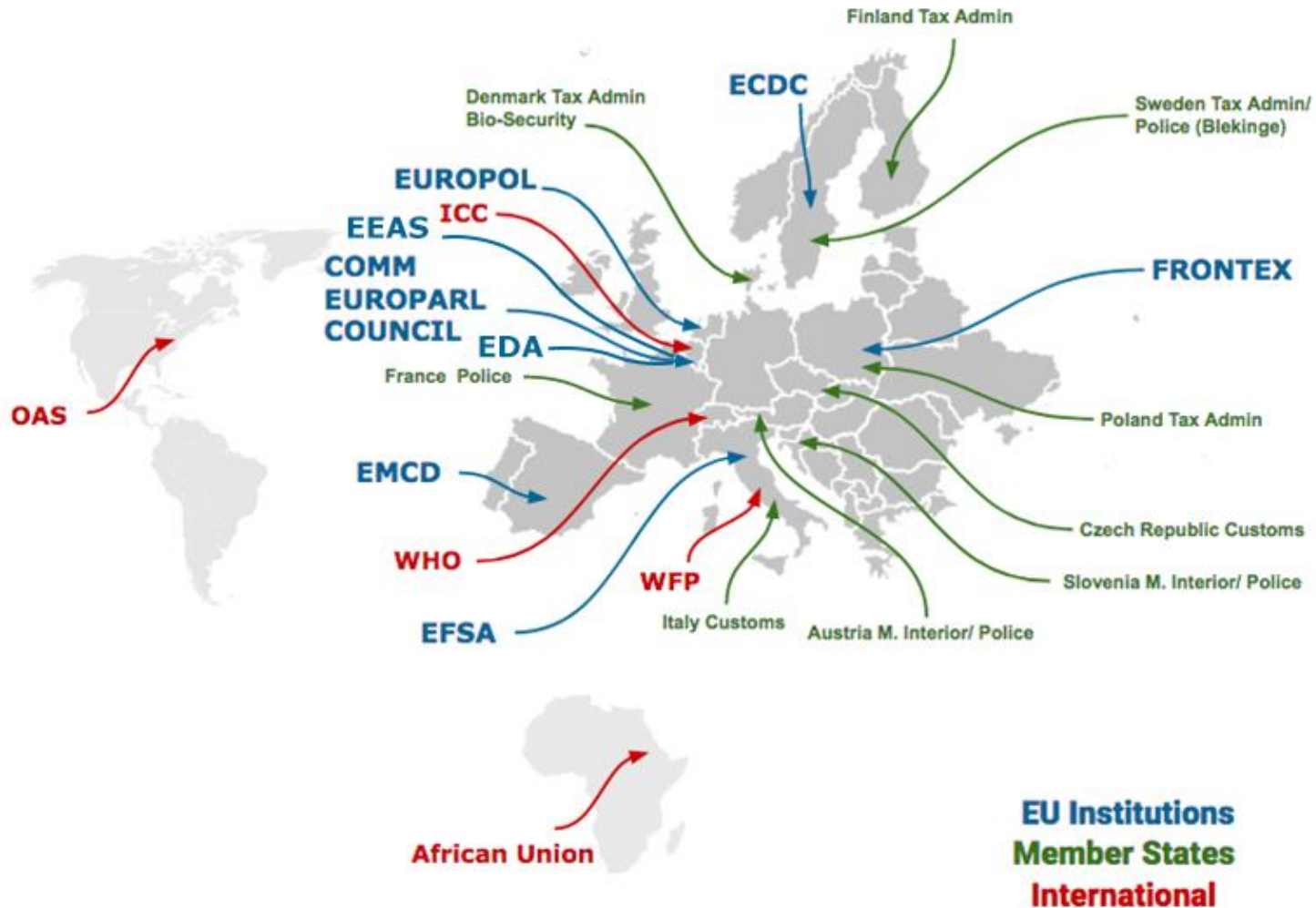
Geo:



Sentiment:
+ POSITIVE



SOME EXISTING CLIENTS



**EU Institutions,
EU Agencies,
UN Agencies,
African Union,
Organization of
American States,
Member States
Tax Agencies and
Customs**

from Member States have
already started working with
EMM and OSINT tools.

Top Stories





6.25am Coronavirus likely of animal origin, not made in a lab: WHO



Articles: 85 | Last update: 2 minutes ago | Start: Apr 21, 2020 12:31 PM | Sources: 0 | Peak: New | Current rank: 1

6.25am Coronavirus likely of animal origin, not made in a lab: WHO

 theage Apr 21, 2020 11:05 PM | Info 

The World Health Organisation says that all available evidence suggests the novel coronavirus originated in animals in China late last year and was not manipulated or produced in a laboratory. US President Donald Trump said last week that his government was trying to determine whether the virus....



More deaths, no benefit from malaria drug in VA virus study



Articles: 40 | Last update: 2 minutes ago | Start: Apr 21, 2020 6:16 PM | Sources: 0 | Peak: 2 | Current rank: 2

More deaths, no benefit from malaria drug in VA virus study

Articles (570)



« 1 2 3 ... 55 56 57 »

EMM News sources by country



Map of sources of 570 articles



Data Acquisition:

and the 2020 EMM-MediSys COVID19 Dataset

Guillaume Jacquet
JRC.I.3 – Joint Research Centre – European Commission

Covid-19 MLIA Virtual Meeting, 12/01/2021

Media data access

Goal:

- Give access to the metadata created in EMM-MediSys from news articles
- Focus on Covid-19
- From Dec 2019 to April 2020

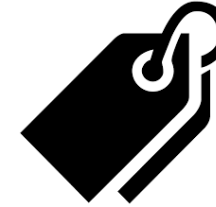


Structured data: logical view



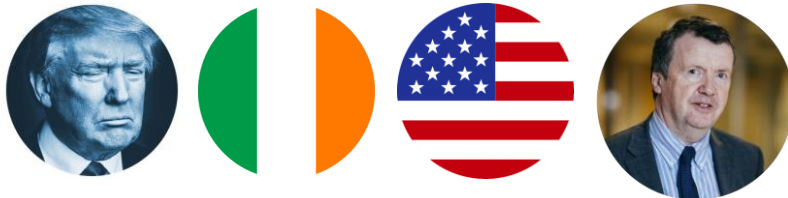
Language:
English

US president Donald Trump intention to cut the US corporate tax rate to 15 per cent may not be achievable with the border adjustment tax, which might have funded the reduction, now off the table. Irish tax experts remain convinced that US tax reform, whatever shape it takes, will not adversely damage Ireland's economic prospects, suggesting incentives for US multinationals to base operations here will continue. *"All in all, I wouldn't be unduly worried at this juncture,"* Alan McQuaid said.



Categorization:
Taxation
Economy
EU-US trade

Entities:



Quotes:

"All in all, I wouldn't be unduly worried at this juncture"

Geo:



Sentiment:
+ POSITIVE



Media data access

- Limitations:
 - IPR compliant
 - Not initially made for distribution
 - Information extraction based on algorithms
→ Incomplete and possibly noisy data.



Media data access

- Some numbers:
- One month →
 - 4.1 M news articles
 - 76 languages
 - 37.6 M entities (per, org, loc, ...)
 - 15 M dates
 - 0.8 M quotes



Media data access

- Where?
 - ODP portal
- What?
 - data from Dec 2019 to April 2020

<https://data.europa.eu/euodp/en/data/dataset?q=covid+jrc>

COVID-19 news monitoring with Medical Information System (MediSys)

Publisher

[Joint Research Centre »](#)

Description

Dataset of metadata created with Europe Media Monitor (EMM)/Medical Information System (MediSys) processing chain from news articles.

MEDISYS is a media monitoring system providing event-based surveillance to rapidly identify potential public health threats using information from media reports. The system displays only those articles with interest to public health (e. g. diseases, plant pests, psychoactive substances), analyses news reports and warns users with automatically generated alerts.

This dataset has a focus on Covid-19. It provides a large set of metadata automatically extracted from news articles related to Covid -19, stored as rss/xml format. It is publicly available, and anyone can build applications on top of that. The current version contains 4 months of news articles, from December 2019 to April 2020, which corresponds to more than 6 Million news articles. There is one zip file per month, containing the whole metadata information. As a example, the biggest month is March 2020, it contains 4.1 million news articles, from 76 different languages, 36 million entity occurrences (person names, organization names, location names, ...), 15 million dates, 0.8 million quotations.

The information processed by MediSys is derived from the Europe Media Monitor (EMM). The freely accessible Europe Media Monitor (EMM) is a fully automatic system that analyses on-line media. It gathers and aggregates about 300,000 news articles per day from news portals world-wide in up to 80 languages

DATA COLLECTION: 1ST ROUND OF IE AND MSS TASKS

- Based on the 2020 EMM-MediSys COVID19 Dataset
 - Dataset of metadata created with Europe Media Monitor (EMM)/Medical Information System (MediSys) processing chain from news articles.
- Processing steps:
 - Parse the metadata
 - Download web pages
 - Clean-up acquired web pages
 - Construct monolingual collections of XML documents including:
 - Main content segmented into paragraphs
 - Basic metadata (e.g. url, title, etc.)



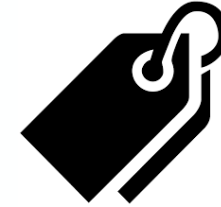
Any questions?

Structured data: logical view



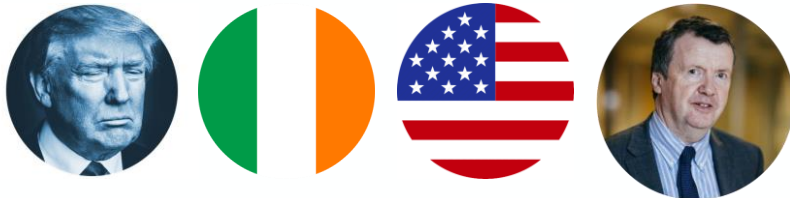
Language:
English

US president Donald Trump intention to cut the US corporate tax rate to 15 per cent may not be achievable with the border adjustment tax, which might have funded the reduction, now off the table. Irish tax experts remain convinced that US tax reform, whatever shape it takes, will not adversely damage Ireland's economic prospects, suggesting incentives for US multinationals to base operations here will continue. *"All in all, I wouldn't be unduly worried at this juncture,"* Alan McQuaid said.



Categorization:
Taxation
Economy
EU-US trade

Entities:



Quotes:

"All in all, I wouldn't be unduly worried at this juncture"

Geo:



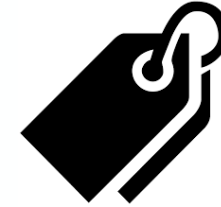
Sentiment:
+ POSITIVE



Structured data: logical view

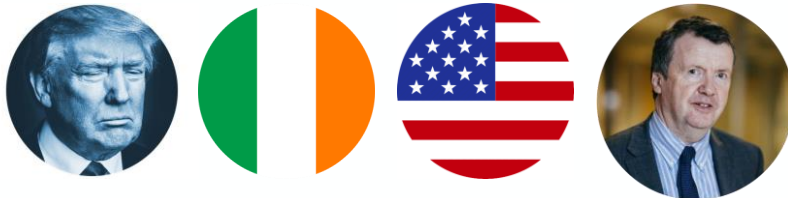


Language:
English



Categorization:
Taxation
Economy
EU-US trade

Entities:



Quotes:

*"All in all, I wouldn't be
unduly worried at this
juncture"*

Geo:



Sentiment:
+ POSITIVE



Structured data: logical view

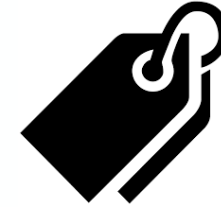


Language:
English

Original URL

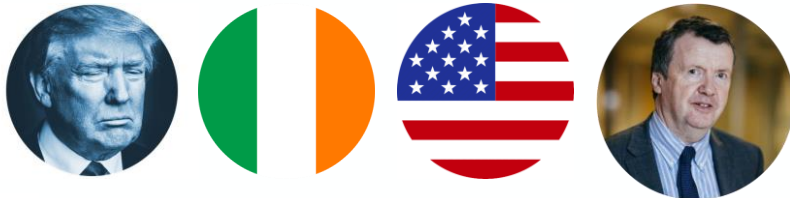
50% of full article in alphabetic order:

15 ; achievable ; adjustment ; be ; border ; cent ;
convinced ; corporate ; cut ; donald ; experts ; funded ;
have ; intention ; irish ; may ; might ; not ; now ; off ;
per ; president ; rate ; reduction ; reform ; remain ;
table ; tax ; tax ; tax ; tax ; that ; the ; the ; the ; the ; to
; to ; trump ; us ; us ; us ; which ; with



Categorization:
Taxation
Economy
EU-US trade

Entities:



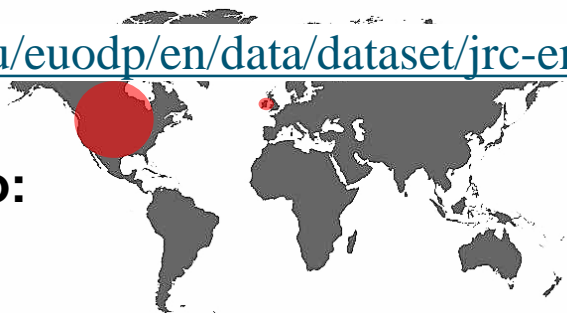
Quotes:

*"All in all, I wouldn't be
unduly worried at this
juncture"*

JRC-Names RDF

<https://data.europa.eu/euodp/en/data/dataset/jrc-emm-jrc-names>

Geo:



Sentiment:
+ POSITIVE



Structured data

```
<item emm:id="DerBund-f0b47a38fb5ad0208bf5dfe9ab9f293d" duplicate="baz-f0b47a38fb5ad0208bf5dfe9ab9f293d">
  <title>Angst vor der Rezession - jetzt soll der Staat helfen Panik an den Börsen verstärkt die Ängste vor schweren wirtschaftlich
  <link>https://www.derbund.ch/wirtschaft/angst-vor-der-rezession-jetzt-soll-der-staat-helfen/story/29691971</link>
  <description/>
  <emm:contentType>text/html</emm:contentType>
  <pubDate>2020-03-01T01:04+0100</pubDate>
  <source country="CH" url="https://www.derbund.ch/wirtschaft/">DerBund</source>
  <iso:language>de</iso:language>
  <guid>DerBund-f0b47a38fb5ad0208bf5dfe9ab9f293d</guid>
  <category>ContaminationNew</category>
  <category>USA</category>
  <category>MedicalAlertPublicAll</category>
  <category>CoronavirusInfection</category>
  <emm:favicon>https://www.derbund.ch/webapp/img/favicon.png?almostappearneveragain</emm:favicon>
  <emm:entity id="253715" name="Hans Jürgen Maurus" pos="163" count="1" subtype="PER" type="p">Hans-Jürgen Maurus</emm:entity>
  <emm:entity id="2079243" name="Armin Müller" pos="183" count="1" subtype="PER" type="p">Armin Müller</emm:entity>
  <emm:georss iso="US" id="18956535" name="de New York:Kings:New York:U S" pos="316" count="1" wordlen="8" charpos="316" class="2">
  <emm:timex pos="745" count="1" type="date" value="2008">2008</emm:timex>
  <emm:ifs pos="736" count="1" subtype="TIM-DA" type="x" day="null" year="2008" month="10">Oktober 2008</emm:ifs>
  <emm:sentiment mode="svm-v1">neutral</emm:sentiment>
  <emm:emotion mode="svm-v1">fear</emm:emotion>
  <emm:tonality>-11</emm:tonality>
  <emm:text wordCount="111">an an angesichts ausbreitung börsen börsenhändler coronavirus das den den des die ein eine eine ergeben
</item>
```