European Commission

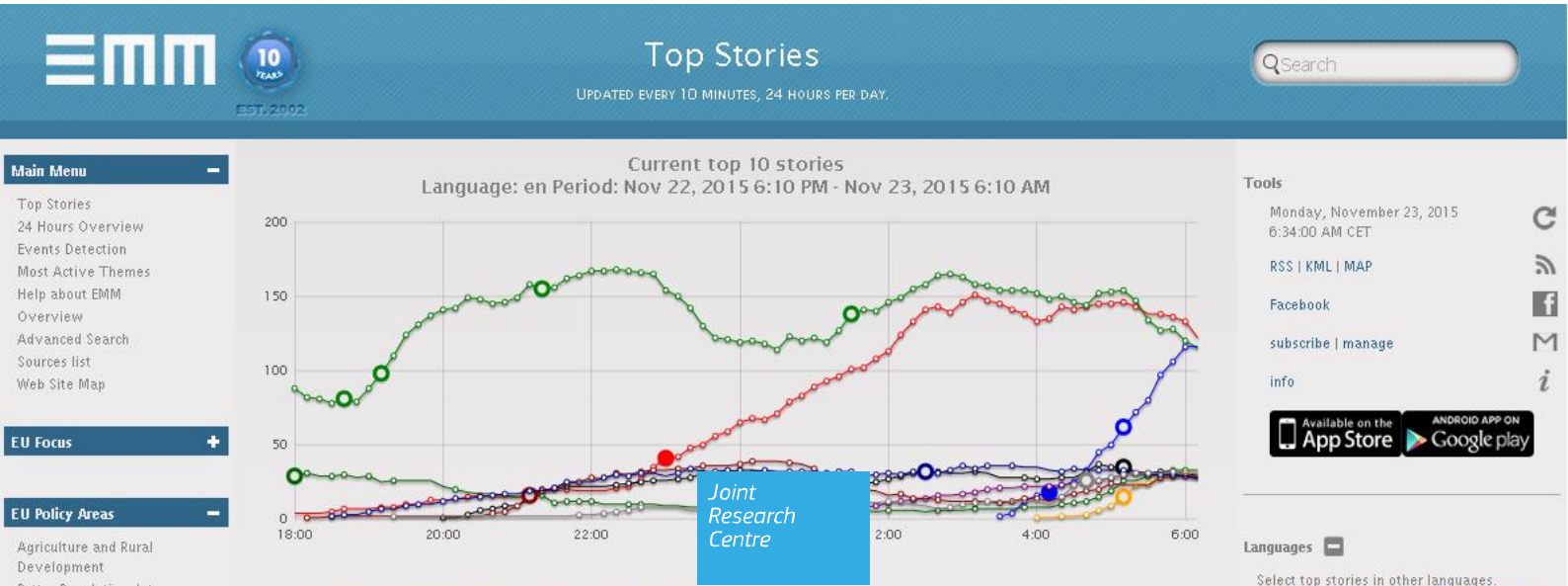
JRC SCIENTIFIC INFORMATION SYSTEMS AND DATABASES REPORT

Europe Media Monitor

Metadata format V. 1.0

JRC.I.3

2020



This publication is a Scientific Information Systems and Databases report by the Joint Research Centre (JRC), the European Commission's science and knowledge service. It aims to provide evidence-based scientific support to the European policymaking process. The scientific output expressed does not imply a policy position of the European Commission. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use that might be made of this publication. For information on the methodology and quality underlying the data used in this publication for which the source is neither Eurostat nor other Commission services, users should contact the referenced source. The designations employed and the presentation of material on the maps do not imply the expression of any opinion whatsoever on the part of the European Union concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

Contact information
Name: Marco Verile
Email: marco.verile@ec.europa.eu

EU Science Hub
https://ec.europa.eu/jrc

Ispra: European Commission, 2020

Contents

Acknowledgements

The Europe Media Monitor technology is the result of a collective passionate effort over more than 15 years. Many people have contributed to address text mining challenges as well as source curation, data collection, processing, and visualization tools design. We would like to thank all of them, and the many more that, we hope, will contribute to EMM in the future.

1. Introduction

EMM is the Europe Media Monitor[1], a system for monitoring open source news information. EMM is developed and maintained by the Text & Data Mining Unit, in the Directorate for Competences of the European Commission's Joint Research Centre (JRC). EMM started in 2002 as a project to support the Commission with its media monitoring activities. The main purpose of EMM is to provide monitoring of a large (but selected) set of electronic media, reducing the information flow to manageable proportions by: clustering related news, categorising articles, and applying language technology tools to derive further metadata, such as recognising and disambiguating entities in the text, extracting quotes by and about people, applying sentiment/tonality analysis and more. The system continuously monitors almost 11 000 sources to find new articles published on the Internet (~300 000 articles daily).

The results of EMM processing are captured in metadata describing, for each analysed document, the associated topics, mentioned named entities (people, organisations, and locations), quotes extracted, etc. This document describes the format of the EMM metadata.

1.1 Copyright Notice

[1] Public web site: https://emm.newsbrief.eu .

2. Format

EMM metadata is stored in RSS format[2] with some customisations. Each file contains a variable number of RSS items.

```
<?xml version="1.0" encoding="UTF-8"?>
<rss version="2.0" xmlns:emm="http://emm.jrc.it" xmlns:iso="http://www.iso.org/3166">
  <channel>

    <title>my selection</title>
    <pubDate>Tue, 21 Apr 2020 02:38:27 CEST</pubDate>

    <item emm:id="news-9147bf1262c6d838918051ac68da64e5">
          [...]
    </item>

    [... more items here...]

    <item emm:id="news-AFFF4556678c6d838918051ac68da6342">
          [...]
    </item>

  </channel>
</rss>
```

2.2 Channel element

The channel element contains the RSS creation date and title.

Table 1. "channel" element.

Field	Short Description	Example
pubDate	RSS file creation date.	Tue, 21 Apr 2020 02:38:27 CEST
Title	Title of the selection.	my selection

[2] See RSS 2.0 specification: http://www.rssboard.org/rss-specification .

2.3 Item element

The item element contains the metadata about one document processed by EMM.

Table 2. "item" element.

Attribute	Short Description	Example
link	Link to the news article on the publisher web site (the same used in the web browser).	https://www.lemonde.fr/planete/article/2020/04/07/coronavirus-les-francais-appeles-a-ne-pas-relacher-leurs-efforts-de-confinement_6035777_3244.html
Title	Title of the news article[3].	Coronavirus : plus de 10 000 morts en France, un salarié sur quatre en activité partielle
Guid	EMM Internal unique identifier.	middleeastmonitor-93b4e24dd35a4d910ce30dff47c7c898
source	a) Publisher's name b) Publisher's web site c) Publisher's country	a) LeMonde b) url=https://www.lemonde.fr/rss/une.xml c) country="FR"
pubdate	Date and time when the article was discovered.	2020-04-07T21:15+0200
iso:language	Language of the text.	fr
emm:entity	Known people or organizations mentioned in the text (already part of the entity repository). a) EMM Internal unique identifier b) Type c) Subtype (refer to chapter 3 for possible values)	 a) id="2844" b) type="p"

[3] See §1.1 for copyright restrictions.

Attribute	Short Description	Example
	d) Occurrences in the text e) Position in the text f) Display name	c) subtype="PER" d) count="2" e) pos="3684,4540" f) name="Anne Hidalgo"
emm:guess	New people or organization names discovered in the text (not yet in the entity repository). a) Type b) Subtype (refer to chapter 3 for possible values) c) Occurrences in the text d) Position in the text e) Display name f) Matching rule	 a) type="o" b) subtype="ORG-CO" c) count="1" d) pos="1539" e) name="BFMTV" f) rules="ho0_alone_0"
emm:georss	Known location mentioned in the text. a) Full name b) Internal unique identifier c) Latitude d) Longitude e) Occurrences in the text f) Position in the text g) Class h) ISO country code i) Position in the text j) Word length k) Display name	 a) name="Paris:Ile-de-France:France" b) id="16912462" c) lat="48.8521" d) lon="2.34899" e) count="3" f) pos="3460,3678,4566" g) class="1" h) iso="FR" i) charpos="3460,3678,4566" j) wordlen="5,5,5" k) Paris
emm:fullgeo	Known location mentioned in the text. a) Name b) Internal unique identifier	 a) name="France" b) id="78"

Attribute	Short Description	Example
	c) Adjective d) Internal ranking e) Latitude f) Longitude g) Occurrences in the text h) Position in the text i) Class j) ISO country code k) Position in the text l) Word length m) Display name	c) adjective="false" d) rank="5" e) lat="48.8521" f) lon="2.34899" g) count="5" h) pos="38,2344,2477,2944,4126" i) class="-1" j) iso="FR" k) charpos="38,2344,2477,2944,4126" l) wordlen="6,6,6,6,6,6" m) France
category	Topic/subject mentioned in the text. a) Name. b) Internal ranking (position in the text). c) Keywords matching score. d) Relevant words found in the text.	 a) CommunicableDiseases b) emm:rank="0" c) emm:score="80" d) emm:trigger="Coronavirus[1]; coronavirus[1];"
emm:link	Link to other web pages referenced in the text.	https://www.interieur.gouv.fr/Actualites/L-actu-du-Ministere/Attestation-de-deplacement-derogatoire-et-justificatif-de-deplacement-professionnel
emm:sentiment	Sentiment of the overall text.	negative
emm:emotion	Emotion of the overall text.	anger
emm:tonality	Tonality score of the overall text.	-4
emm:timex	Numerical expression. a) Type b) Position in the text c) Occurrences in the text d) Exact value extracted.	 a) type="date" b) pos="332" c) count="1" d) value="2020-04-07"

Attribute	Short Description	Example
emm:ifs	Other entities, usually numerical or composed of specific textual patterns like email addresses, currency, … a) Type (in this case always with the value "x") b) Subtype (refer to 3for possible values) c) Position in the text d) Occurrences in the text e) Exact value extracted.	a) type="date" b) subtype=" IDT-EM" c) pos="332" d) count="1" e) value="JRC-EMM-SUPPORT@ec.europa.eu"
emm:text	50% of the processed words sorted in alphabetic order.	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

3. EMM Named Entity Hierarchy

The **main rationale for creating this Named-Entity hierarchy** is to enumerate the **Named-Entity TYPES** to be used **CONSISTENTLY** when encoding any kind of linguistic resources (e.g., lexical resources or grammars) for the purpose of **Named Entity Recognition** and make easier the sharing of information between the different modules involved in this process. This doesn't mean that the description of an entity would be limited to these types and subtypes only. More fine-grained or application-specific information could be declared by adding arbitrary attribute-value pairs to each entity, like for instance a profession for a person entity.

Since names can be ambiguous and some entities could have multiple sub-types or even types, a given entity can be potentially assigned to **more than ONE type or subtype**.

PERSON		
Subtype	**Example/Explanation**	**Encoding**
-	*"John Smith", "George W. Bush", "James Bond"*	PER[4]
ORGANISATION		
Subtype	**Example/Explanation**	**Encoding**
POLITICAL-PUBLIC[5]	political parties, e.g., "Democratic Party", "CDU", political organisations, e.g. *"Palestine Liberation Organisation", "JSIS",* military organisations, e.g. *"US Air Force"*, government institutions, e.g., "Ministry of Interior of Italy", *"Thatcher's Cabinet", "Embassy of USA", public institution, e.g., "European Commission", "European Patent Office", "New York Public Library"*	ORG-PP
COMMERCIAL	*"Toyota", "Apple", "Microsoft", "Bank of Scotland"*	ORG-CO
RELIGIOUS	*"Anglican Church of Canada", "Islamic Forum of Europe"*	ORG-RE
SPORT	*sport clubs and organisations, e.g., "FC Barcelona", "Serie A", "Bundesliga",*	ORG-SP
EDUCATION-RESEARCH	*"University of Lugano", "European School of Varese"*	ORG-ER
OTHER	any organisation that do not fit in the categories above	ORG-OT
LOCATION		
Subtype	***Example/Explanation***	**Encoding**
CITY	*"London"*	LOC-CI
COUNTY**Erreur ! Signet non défini.**	*"West Chester County"*	LOC-CN
PROVINCE	*"Province of Varese"*	LOC-PR
COUNTRY	*"Italy"*	LOC-CT

[4] There was a proposal to introduce subtypes for PER, e.g., to distinguish between profession of the person and/or whether it is a historical person or from 20/21 century, etc. We concluded that this type of information could be stored in the appropriate attribute for PER mainly due to the fact that we are talking here of some characteristics that change over time (e.g. profession).

[5] There was a proposal to introduce subtypes for this category, namely: (a) parties, (b) administrative, (c) civil protection and disaster management services like fire brigade, (d) juridical - courts, procuratura, (e) political movements, recognized and not recognized, (f) criminal - Ndrangeta, Mafia, Sacra corona unita, (g) terrorist organizations. Since the borders between some of these subtypes are to some extent blurred for now there will not be any distinction between them. However, additional information can be stored in attribute appropriate for that type.

REGION	part of a city, e.g., *"Bronx"*, special economic zone, geographical region, e.g., *"Provence"*	LOC-RE
FACILITY[6]	sport facilities, e.g., *" "Yankee Stadium"*, *recreation facilities, e.g., "Central Park", "Berlin Zoo", "Disneyland", etc., cultural facilities, e.g., "British Museum", "Louvre", "Royal Opera House", "La Scala", hotels, tourist sites, e.g., "Archeological Ruins at Moenjodaro", "Forum Romanum", hospitals, cemeteries, all kind of transportation hubs (ports, railway stations, airports), e.g., "Schipol Airport", "165th Street Bus Terminal", "Berlin Hauptbahnhof", "Port of Honk Kong", "Victoria Harbour", churches, urban/non-urban facilities such as roundabouts,* railroads, roads, tunnels, e.g. *"St. Gothard Tunnel"*, bridges, *etc.*	LOC-FA
OTHER	any mentions of locations (e.g. landforms) that do not fit in the categories above, e.g., mountains *"Mount Everest"*, islands, water bodies *"Baltic"*, rivers, valleys, islands, etc.	LOC-OT
IDENTIFIER		
Subtype	**Example/Explanation**	**Encoding**
STREET-NAME	*"Via Fermi 27"*	IDT-SN
POSTAL-CODE	*"NY 10202"*	IDT-PC
POSTAL-ADDRESS	*"Via Fermi 27, 21200 Ispra, Italy"*	IDT-PA
GEO-COORDINATES	*"51° 28' 38" N"*	IDT-CO
PHONE-NUMBER	*"+49 60 1234576"*	IDT-PN
EMAIL	*"abc@derf.com"*	IDT-EM
URL	*"http://www.google.com"*	IDT-UR
IP-ADDRESS	*"255.255.123.212"*	IDT-IP
VAT-NUMBER	*"ATU99999999"*	IDT-VA
BANKING-IDENTITY	*"DE44 5001 0517 5407 3249 31"*	IDT-BI
CREDIT-CARD-NUMBER	*2345 3523 2453 3453*	IDT-CR
SOCIAL-MEDIA-ID	*"@jakubP"*	IDT-SM
PRODUCT		
Subtype	**Example/Explanation**	**Encoding**
ELECTRONICS	*"Commodore 64"*	PRO-EL
DRUG-MEDICINE	*"Aspirin C"*	PRO-DM
WEAPON	*"AGM-1 Carbine"*	PRO-WE
VEHICLE	*"Mitsubishi Pajero"*	PRO-VE
FOOD	*"Snickers"*	PRO-FO
ART	*"Star Wars"*	PRO-AR
SERVICE	*"Google Search Engine"*	PRO-SE
OTHER	any product mention that does not fall under the above categories	PRO-OT
EVENT[7]		
Subtype	**Example/Explanation**	**Encoding**

[6] Note that this category has been inspired by the FACILITY category in Sekine NE.
[7] Please note that classification of the events is done from the perspective of NAMED MENTIONS of events in text.

INCIDENT[8]	*"Chernobyl Disaster", "John Kennedy assassination", "World war II"*	EVT-IN
NATURAL	*"Great Alaska Earthquake", "Hurricane Katrina"*	EVT-NA
OCCASION[9]	conferences, e.g., *"LREC 2016", "Yalta Conference",* religious holiday. e.g., *"Christmas",* sport events, e.g., *"Football World Cup 2014",* ceremonies, e.g. *"Nobel Prize Awards"*, etc.	EVT-OC
OTHER	*"Kuril Island dispute"*	EVT-OT
TIMEX		
Subtype	**Example/Explanation**	**Encoding**
TIME	*"2PM", "18:42"*	TIM-TM
DATE	*"1 April 2016", "12.01.2016"*	TIM-DA
PERIOD	*"4 hours", "20 years"*	TIM-PE
OTHER	*"Victorian Age"*	TIM-OT
NUMEX		
Subtype	**Example/Explanation**	**Encoding**
NUMERICAL-EXPRESSION	*"12", "12.00", "12 million", "22%", "2/3",*	NUM-EX
CURRENCY-EXPRESSION	*"100 USD", "3.50 EUR"*	NUM-CU
AGE	*"12 years old"*	NUM-AG
MEASUREMENT	*"30 kg", "100 gallons", "36° C"*	NUM-ME
COUNTX	*"10 people"*	MUM-CT
OTHER	any numerical expressions that do not fall under the above categories	NUM-OT
OTHER		
Subtype	**Example/Explanation**	**Encoding**
-	everything else that does not fall under any of the above main categories	OTH

[8] Man-made incidents

[9] There was a proposal to make distinction between the different subtypes, e.g., political, sport, conferences …. Such sub-classification would not be consistent with the current breakdown into the main categories, etc., e.g., conferences can be related to both politics and sports. Waiting for more input in this regard and a refined proposal.

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: https://europa.eu/european-union/contact_en

On the phone or by email

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),

- at the following standard number: +32 22999696, or

- by electronic mail via: https://europa.eu/european-union/contact_en

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

EU publications

You can download or order free and priced EU publications from EU Bookshop at: https://publications.europa.eu/en/publications. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/european-union/contact_en).

The European Commission's science and knowledge service

Joint Research Centre

JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.

EU Science Hub

ec.europa.eu/jrc

@EU_ScienceHub

EU Science Hub - Joint Research Centre

EU Science, Research and Innovation

EU Science Hub