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[1]
ORGANISATION		
Subtype	**Example/Explanation**	**Encoding**
POLITICAL-PUBLIC[2]	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

[1] 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).

[2] 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.

OTHER	any organisation that do not fit in the categories above	ORG-OT
LOCATION		
Subtype	***Example/Explanation***	**Encoding**
CITY	*"London"*	LOC-CI
COUNTY	*"West Chester County"*	LOC-CN
PROVINCE	*"Province of Varese"*	LOC-PR
COUNTRY	*"Italy"*	LOC-CT
REGION	part of a city, e.g., *"Bronx"*, special economic zone, geographical region, e.g., *"Provence"*	LOC-RE
FACILITY[3]	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

[3] Note that this category has been inspired by the FACILITY category in Sekine NE.

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[4]		
Subtype	**Example/Explanation**	**Encoding**
INCIDENT[5]	*"Chernobyl Disaster", "John Kennedy assassination", "World war II"*	EVT-IN
NATURAL	*"Great Alaska Earthquake", "Hurricane Katrina"*	EVT-NA
OCCASION[6]	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

[4] Please note that classification of the events is done from the perspective of NAMED MENTIONS of events in text.

[5] Man-made incidents

[6] 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.