JRC Dataset

JRC-Names RDF: Person and organisation spelling variants as found in multilingual news articles

Description:
JRC-Names is a highly multilingual named entity resource for person and organisation names (called ‘entities’) developed by the European Commission's Joint Research Centre (JRC). JRC-Names consists of large lists of names and their many spelling variants (up to hundreds for a single person), including across scripts (Latin, Greek, Arabic, Cyrillic, Japanese, Chinese, etc.). The resource is the by-product of the Europe Media Monitor (EMM, see http://emm.newsbrief.eu/overview.html ) family of applications, which has been analysing up to 220,000 news reports per day, since 2004. EMM recognises names mentioned in the news in over twenty languages and decides automatically for each newly found name whether it belongs to a new entity or whether it is a spelling variant of a previously known entity. This resource allows EMM users to display news about people or organisations even if their names are spelt differently or if the news articles are written in different languages and scripts. JRC-Names has been available for download since September 2011, consisting of name variant lists and accompanying software. The new linked data edition, accessible through the European Union's Open Data Portal, offers more information compared to the previously released resource and tool, including: titles and function names that have been historically found next to the person mentions; information about the time period during which name variants and their titles were found; various frequency counts; as well as links to other linked datasets such as DBPedia. INFORMATION ON JRC- NAMES DATA MODEL The JRC-Names RDF representation is based on lemon (Lexicon Model for Ontologies, see http://lemon-model.net/ ), a model which allows the expression of lexical information relative to ontologies. JRC entities are modeled as instances of DBpedia classes (dbpedia:Person and dbpedia:Organisation) and the multilingual lexicalizations of their names and function names are represented as Lexical Entries of lemon Lexicons. Various other types of (linguistic) information and metadata are expressed using standardized vocabularies (LexInfo, OLIA, ISOCat, Lexvo, DCTerms, etc.). For cases where no already existing vocabulary could appropriately answer the needs, in-house classes and properties were defined ( see resource JRC data model for JRC names). The ‘JRC-Names schema’ gives an overview of how JRC-Names data is modeled. JRC-Names has links towards the following datasets: DBpedia ( http://dbpedia.org/ ), New York Times Open Data ( http://data.nytimes.com/ ) and Talk of Europe ( http://linkedpolitics.ops.few.vu.nl/ ). For further information on JRC-Names, see: https://ec.europa.eu/jrc/en/language-technologies/jrc-names.

Contributors:
- Steinberger, Ralf ralf.steinberger@ec.europa.eu
- Jacquet, Guillaume guillaume.jacquet@ec.europa.eu

How to cite:

Keywords:
JRC-Names, lemon, lexical variants, linked data, multilingual linguistic resource, named entities, natural language processing, news analysis, person and organisation names, person titles, variant and title usage time intervals

Related resources:
Data access
JRC-Names, RDF files
The compressed zip file contains an RDF file containing JRC names and spelling variants of JRC names.

SPARQL endpoint access
Access to the JRC-Names dataset via the Open Data Portal SPARQL endpoint. Some specific query examples are provided in the SPARQL endpoint webpage.

Publications
JRC-Names: A freely available, highly multilingual named entity resource
Other resources

JRC data model for JRC-Names
Definitions of classes and properties defined by JRC for JRC-names where no already existing vocabulary could appropriately answer the needs.

JRC-Names data statistics
Some statistics on JRC-Names

Additional information:
Last Modified: 2016-05-19
Issue date: 2015-05-19
Geographic area: European Union
Temporal coverage: From: 2004-01-01
Update frequency: None
Language: English
Data theme(s): Economy and finance; Science and technology
EuroVoc domain(s): 36 SCIENCE; 64 PRODUCTION, TECHNOLOGY AND RESEARCH
Identifier: http://data.europa.eu/89h/jrc-emm-jrc-names