1.**JRC Names 实体专有名词数据集**

**-下载地址：**[**http://dataju.cn/Dataju/web/datasetInstanceDetail/78**](http://dataju.cn/Dataju/web/datasetInstanceDetail/78)

**- 简介：**

**-数据集概述：**JRC Names 是一个跨语言实体名称语料库，该语料库包含了大量跨语言人名和组织名称（称之为“实体”），包括不同语言的实体名称，包括汉语、英语、拉丁语、希腊语、阿拉伯语、斯拉夫语、日本语等。2016年之后，JRC Names还关联了其它附加信息，如每个实体在每种语言中出现的频率和时间段等。

**-数据来源：**[**https://ec.europa.eu/jrc/en/language-technologies/jrc-names**](https://ec.europa.eu/jrc/en/language-technologies/jrc-names)**。**JRC Names 实体专有名词是欧洲媒体监控（EMM）通过每天分析约22万条新闻报道所产生的。从2004年以来，分析了数以百万计的新闻文章，最多可以使用21种语言，识别实体（主要是人员，组织，事件名称等）的名称，并检测其中的哪一种新发现的名字是彼此的变体拼写。因此，JRC Names中的大多数命名变体都是在现实生活中采集的（包括频繁的拼写错误）。此外，对于实体集合的一个子集，软件会自动从维基百科的跨语言链接中提取许多其他语言（例如中文，泰语，日语，...）的拼写变体。对于非常频繁或重要的名称，另外手动验证命名实体资源。由于JRC名称主要是自动生成的，所以会包含一些错误。

**-文件类型：**多个文件

**-文件大小：**756.00Mb

**-用处:**JRC实体专有名词，可用于查找命名实体，即使它们拼写不同，但它在计算机处理文本信息过程中也能识别，可用于例如，用于文本挖掘。该工具具有多种用途并解决了各种问题，其中包括：

1.当搜索数据库，互联网和其他存储库时，正确的名称是一个问题，因为通常找不到搜索名称的变体。这导致对文档，图像和视听内容的存储库的非最佳使用和利用。 JRC-Names允许标准化名称，从而改进检索;

2.机器翻译在翻译实体专有名词时也存在问题，因为它们不应像其他单词一样翻译;利用JRC-Names可以在翻译过程之前提取名称，并且可以用目标语言重新插入外语变体来解决这个问题;

3.两个不同语言中的实体列表通常用于学习音译规则;

4.实体可以在文本中识别和标记，以便在训练机器学习命名实体识别系统中直接利用;
5.数据来源于多国国家，可以减少不同国家观点的偏见;
6.命名实体识别对于意见挖掘的计算语言学任务，共同参考解析，总结，主题检测和跟踪，跨语言的相关文档的跨语言链接等都是有用的。

**-统计信息:**

JRC Names包含EMM实体数据库中一些重要的实体，如频繁搜索或手动验证或在维基百科上发现的实体。JRC名称的第一个版本（2011年9月）包含约205,000个不同的已知实体的名称，以及这些实体的大致相同数量的变体拼写。此外，它包含这些名称的许多形态上变形的变体。截至2016年3月，该资源已增至307,000个不同实体，另有333,000个变体。EMM每天都会识别新的名称，还可以从JRC的网页上每天下载包含最近发现的名称和名称拼写的文件。截至2011年7月，数据库包括27个不同脚本中拼写的名称。最常用的是拉丁语（包括英语和大多数其他欧洲语言），西里尔语（例如俄语和保加利亚语），阿拉伯语（包括波斯语），日语（汉语，平假名和片假名）和汉语汉语（简体中文）。JRC名称中的64％的名称没有额外的拼写变体。对于28％的名字，JRC名称有两到三个拼写。有3760个实体有10个拼写或更多，37个实体有超过100个拼写变体。具有最多拼写变体的名字是Muammar Gaddafi（413个拼写），Mikhail Saakashvili（256）和Mahmoud Ahmadinejad（246）。

**-相关论文:**

1**.**Steinberger Ralf,  Bruno Pouliquen, Mijail Kabadjov, Jenya Belyaeva & Erik van der Goot (2011).

[**JRC-Names: A freely available, highly multilingual named entity resource**](https://ec.europa.eu/jrc/sites/jrcsh/files/11_RANLP_JRC-Names_Final.pdf). Proceedings of the 8th International Conference Recent Advances in Natural Language Processing (RANLP). Hissar, Bulgaria, 12-14 September 2011.

2.Ehrmann Maud, Guillaume Jacquet & Ralf Steinberger (2016). [**JRC-Names: Multilingual Entity Name Variants and Titles as Linked Data**](http://www.semantic-web-journal.net/content/jrc-names-multilingual-entity-name-variants-and-titles-linked-data-1). Semantic Web Journal (March 2016).

3.[STEINBERGER Ralf](https://ec.europa.eu/jrc/en/person/ralf-steinberger)，ATKINSON Martin，GARCIA DOMINGO Teofilo，VAN DER GOOT Erik

LINGE Jens，MACMILLAN Charles，TANEV Hristo，VERILE Marco，WAGNER Gerhard（2017）

[EMM: Supporting the Analyst by Turning Multilingual Text into Structured Data](https://ec.europa.eu/jrc/en/publication/emm-supporting-analyst-turning-multilingual-text-structured-data)

4.[JACQUET Guillaume](https://ec.europa.eu/jrc/en/person/guillaume-jacquet)，EHRMANN Maud，[STEINBERGER Ralf](https://ec.europa.eu/jrc/en/person/ralf-steinberger)，VAEYRYNEN Jaakko（2016）.[Cross-lingual linking of Multi-word Entities and their corresponding Acronyms](https://ec.europa.eu/jrc/en/publication/cross-lingual-linking-multi-word-entities-and-their-corresponding-acronyms).

5.王志娟, 李福现. 跨语言命名实体翻译对抽取的研究综述[J]. 计算机科学, 2017, 44(s1):14-18.

6.胡亚楠, 舒佳根, 钱龙华,等. 基于机器翻译的跨语言关系抽取[J]. 中文信息学报, 2013, 27(5):191-198.

7.吴丹, 何大庆, 陆伟. 跨语言信息检索中的命名实体识别与翻译[J]. 图书情报知识, 2012(3):13-19.