# Datasets de carga de bioentidades y/o relaciones entre ellas

Una vez revisado el estado del arte de los procesos NER para bioentidades, el alcance del proyecto se circunscribe a las mutaciones genéticas. Para ello a partir de papers relacionados con la notación en texto de mutaciones, se procederá a entrenar un modelo NER especializado en la extracción de estas bioentidades. Para ello se parte de varias fuentes de datos:

## Nomenclatura genética

* Ncbi for NCBI gene IDs eg. 673
* Hgnc for HGNC gene IDs eg. HGNC:1097
* Ensembl for ensembl gene IDs eg. ENSG0\*0000157764
* Symbol for HGNC approved symbol eg. TP53

http://ibi.imim.es/befree/

### tmVar <https://www.ncbi.nlm.nih.gov/research/bionlp/Tools/tmvar/>

Base de datos de mutaciones con 500 artículos tageados

### ClinVar

### Mutation finder <http://mutationfinder.sourceforge.net/>

Sistema de identificación de bioentidades relacionadas con mutaciones genéticas

### Bronco <https://www.ncbi.nlm.nih.gov/pubmed/27074804>

Sistema de identificación de bioentidades relacionadas con mutaciones genéticas

### Oisris <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2277400/>

### Thomas

### NCBI <https://www.ncbi.nlm.nih.gov/gene>

Base de datos genética del centro nacional de información biotecnológica. Se trata de un fichero de texto de medio TB que incluye información de todos los genes. La información relevante para el TFM será:

* GenId: Identificador único del gen
* Symbol: Símbolo del gen estándar
* Synonyms: Símbolos no oficiales para el gen separados por “|”

A pesar de tener enlazados los genes con los artículos científicos, el hecho de que este dataset no contenga nombres hace que nos tengamos que plantear su uso dentro del TFM.

### HGNC (<https://www.genenames.org/>)

hgnc\_complete\_set.txt contiene una nomenclatura de genes estandarizada con información textual, identificadores estándar, identificadores en otros dataset y nomenclaturas anteriores. El dataset es un fichero de texto de unas 40000 líneas, separado por tabuladores con el nombre de las columnas en el encabezado. De todas las columnas, las interesantes de acuerdo al proyecto son:

* Hgnc\_id: Identificador según este estándar del gen
* Symbol: Símbolo del gen estándar
* Name: Nombre del gen
* Alias\_symbol: Otros símbolos del gen. En este caso separados por la barra vertical “|”
* Alias\_name: Otros nombres del gen, también separados por “|”
* Prev\_symbol: Símbolo anterior, separado por “|”
* Prev\_name: Nombre anterior separado por “|”
* Entrez\_id: nomenclatura NCBI
* Ensembl\_gene\_id: Identificador del ensamblado
* Vega\_id: Identificador en el estándar Vega
* Usc\_id: Identificador en el estándar UCSD
* Pubmed\_id: Identificador de las citas de este gen separados por “|”

### UniProt

## Nomenclatura de enfermedades

UMLS (https://www.nlm.nih.gov/research/umls/)

### MESH (<https://www.nlm.nih.gov/databases/download/mesh.html>)

MESH es un esfuerzo de Biblioteca Nacional de Medicina de los Estados Unidos para proporcionar una terminología médica que permite categorizar información biomédica actualizada diariamente.