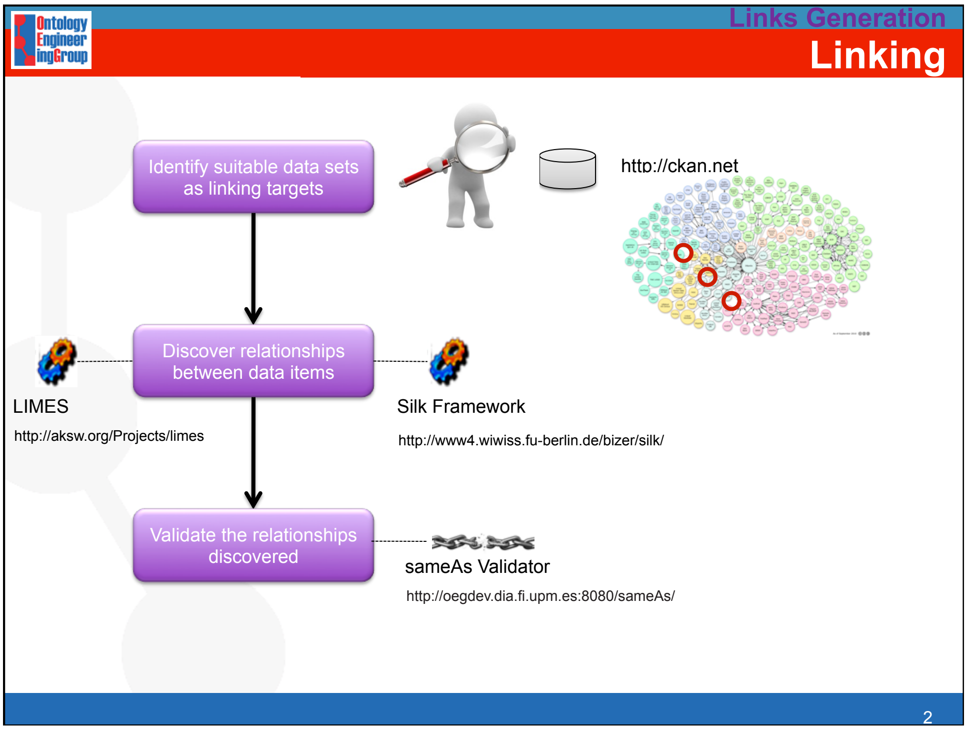
Silk: Discovering Links

Daniel Vila Suero, Boris Villazón-Terrazas
dvila@fi.upm.es

Ontology Engineering Group, Universidad Politécnica de Madrid

Agradecimientos: A los miembros del OEG que han participado en la elaboración de estas transparencias



Links Generation

Linking

Identify suitable data sets as linking targets

http://ckan.net

Discover relationships between data items

LIMES
http://aksw.org/Projects/limes

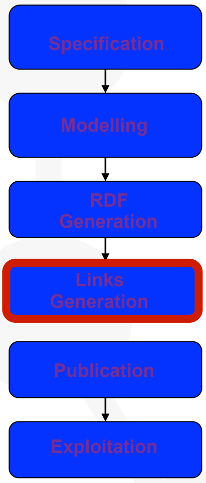
Silk Framework
http://www4.wiwiss.fu-berlin.de/bizer/silk/

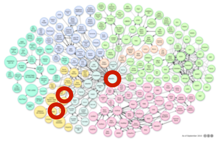
Validate the relationships discovered

sameAs Validator
http://oegdev.dia.fi.upm.es:8080/sameAs/

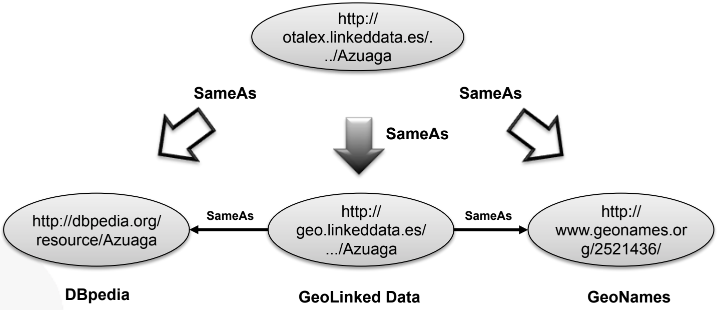
Links Generation

Linking





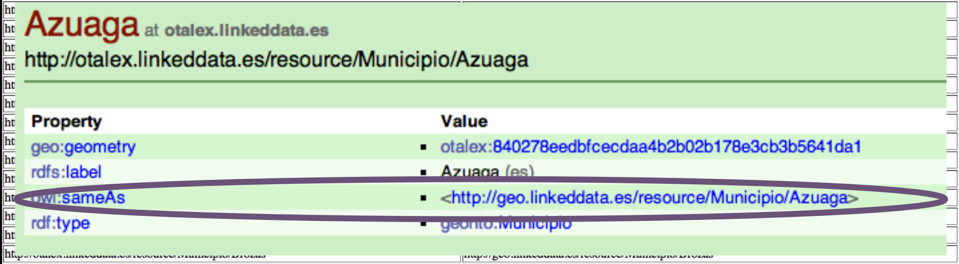
Silk framework


http://otalex.linkeddata.es/../Azuaga
SameAs
SameAs
SameAs
http://dbpedia.org/resource/Azuaga
SameAs
http://geo.linkeddata.es/.../Azuaga
SameAs
http://www.geonames.org/2521436/
DBpedia
GeoLinked Data
GeoNames

Links Generation

Linking

s	o
http://otalex.linkeddata.es/resource/Municipio/Don%20Benito	http://geo.linkeddata.es/resource/Municipio/Don%20Benito
http://otalex.linkeddata.es/resource/Municipio/Almendral	http://geo.linkeddata.es/resource/Municipio/Almendralejo
http://otalex.linkeddata.es/resource/Municipio/Almendralejo	http://geo.linkeddata.es/resource/Municipio/Almendralejo
http://otalex.linkeddata.es/resource/Municipio/Majadas	http://geo.linkeddata.es/resource/Municipio/Miajadas
http://otalex.linkeddata.es/resource/Municipio/Miajadas	http://geo.linkeddata.es/resource/Municipio/Miajadas
http://otalex.linkeddata.es/resource/Municipio/Villafranca%20De%20Los%20Barros	http://geo.linkeddata.es/resource/Municipio/Villafranca%20de%20los%20Barros
http://otalex.linkeddata.es/resource/Municipio/Villalba%20De%20Los%20Barros	http://geo.linkeddata.es/resource/Municipio/Villafranca%20de%20los%20Barros
http://otalex.linkeddata.es/resource/Municipio/Badajoz	http://geo.linkeddata.es/resource/Municipio/Badajoz
http://otalex.linkeddata.es/resource/Municipio/Talayuela	http://geo.linkeddata.es/resource/Municipio/Talayuela
ht	[illegible]
ht	[illegible]
ht	[illegible]
ht	[illegible]
ht	[illegible]
ht	[illegible]
ht	[illegible]
ht	[illegible]
ht	[illegible]
ht	[illegible]
ht	[illegible]
ht	[illegible]
ht	[illegible]
http://otalex.linkeddata.es/resource/Municipio/Brozas	http://geo.linkeddata.es/resource/Municipio/Brozas
http://otalex.linkeddata.es/resource/Municipio/Brotas	http://geo.linkeddata.es/resource/Municipio/Brozas



Azuaga at otalex.linkeddata.es

http://otalex.linkeddata.es/resource/Municipio/Azuaga

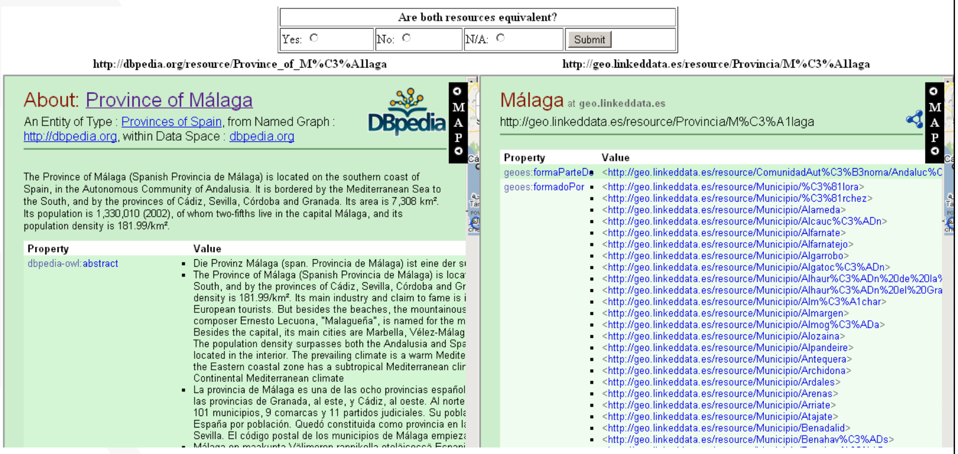
Property	Value
geo:geometry	▪ otalex:840278eedbfcecdaa4b2b02b178e3cb3b5641da1
rdfs:label	▪ Azuaga (es)
owl:sameAs	▪ <http://geo.linkeddata.es/resource/Municipio/Azuaga>
rdf:type	▪ geonto:Municipio

Con quién enlazamos:

- **GeoLinked Data**
- **DBpedia**

Links Generation

Ejemplo- Linking



Are both resources equivalent?			
Yes: ○	No: ○	N/A: ○	Submit

http://dbpedia.org/resource/Province_of_M%C3%Allaga

About: Province of Málaga

An Entity of Type : Provinces of Spain, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

The Province of Málaga (Spanish Provincia de Málaga) is located on the southern coast of Spain, in the Autonomous Community of Andalusia. It is bordered by the Mediterranean Sea to the South, and by the provinces of Cádiz, Sevilla, Córdoba and Granada. Its area is 7,308 km². Its population is 1,330,010 (2002), of whom two-fifths live in the capital Málaga, and its population density is 181.99/km².

Property	Value
dbpedia-owl:abstract	▪ Die Provinz Málaga (span. Provincia de Málaga) ist eine der s ▪ The Province of Málaga (Spanish Provincia de Málaga) is loca South, and by the provinces of Cádiz, Sevilla, Córdoba and Gr density is 181.99/km². Its main industry and claim to fame is i European tourists. But besides the beaches, the mountainous composer Ernesto Lecuona, "Malagueña", is named for the m Besides the capital, its main cities are Marbella, Vélez-Málag The population density surpasses both the Andalusia and Spa located in the interior. The prevailing climate is a warm Medite the Eastern coastal zone has a subtropical Mediterranean clir Continental Mediterranean climate ▪ La provincia de Málaga es una de las ocho provincias español las provincias de Granada, al este, y Cádiz, al oeste. Al norte 101 municipios, 9 comarcas y 11 partidos judiciales. Su pobla España por población. Quedó constituida como provincia en la Sevilla. El código postal de los municipios de Málaga empieza ▪ Málaga on maakunta Välimeren rannikolla eteläisessä Espanj

http://geo.linkeddata.es/resource/Provincia/M%C3%Allaga

Málaga at geo.linkeddata.es

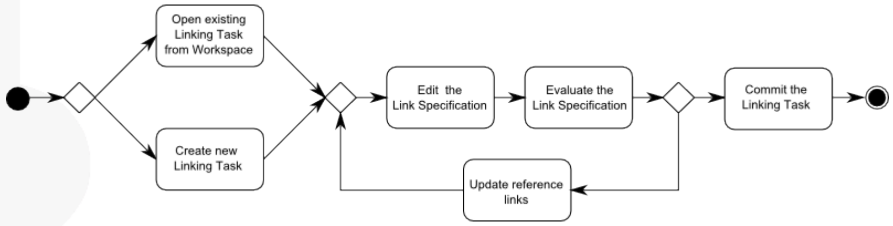
http://geo.linkeddata.es/resource/Provincia/M%C3%A1laga

Property	Value
geoes:formaParteDe	▪ <http://geo.linkeddata.es/resource/ComunidadAut%C3%B3noma/Andaluc%C
geoes:formadoPor	▪ <http://geo.linkeddata.es/resource/Municipio/%C3%81lora> ▪ <http://geo.linkeddata.es/resource/Municipio/%C3%81rchez> ▪ <http://geo.linkeddata.es/resource/Municipio/Alameda> ▪ <http://geo.linkeddata.es/resource/Municipio/Alcauc%C3%ADn> ▪ <http://geo.linkeddata.es/resource/Municipio/Alfarnate> ▪ <http://geo.linkeddata.es/resource/Municipio/Alfarnatejo> ▪ <http://geo.linkeddata.es/resource/Municipio/Algarrobo> ▪ <http://geo.linkeddata.es/resource/Municipio/Algatoc%C3%ADn> ▪ <http://geo.linkeddata.es/resource/Municipio/Alhaur%C3%ADn%20de%20la% ▪ <http://geo.linkeddata.es/resource/Municipio/Alhaur%C3%ADn%20el%20Gra ▪ <http://geo.linkeddata.es/resource/Municipio/Alm%C3%A1char> ▪ <http://geo.linkeddata.es/resource/Municipio/Almargen> ▪ <http://geo.linkeddata.es/resource/Municipio/Almog%C3%ADa> ▪ <http://geo.linkeddata.es/resource/Municipio/Alozaina> ▪ <http://geo.linkeddata.es/resource/Municipio/Alpandeire> ▪ <http://geo.linkeddata.es/resource/Municipio/Antequera> ▪ <http://geo.linkeddata.es/resource/Municipio/Archidona> ▪ <http://geo.linkeddata.es/resource/Municipio/Ardales> ▪ <http://geo.linkeddata.es/resource/Municipio/Arenas> ▪ <http://geo.linkeddata.es/resource/Municipio/Arriate> ▪ <http://geo.linkeddata.es/resource/Municipio/Atajate> ▪ <http://geo.linkeddata.es/resource/Municipio/Benadalid> ▪ <http://geo.linkeddata.es/resource/Municipio/Benahav%C3%ADs>

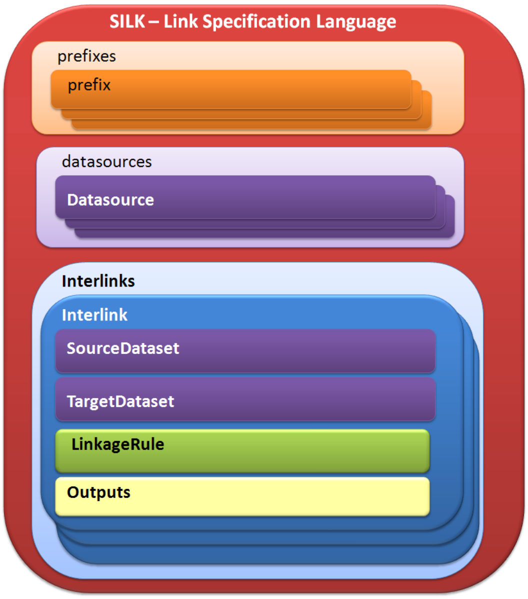
http://oegdev.dia.fi.upm.es:8080/sameAs/

SILK Intro

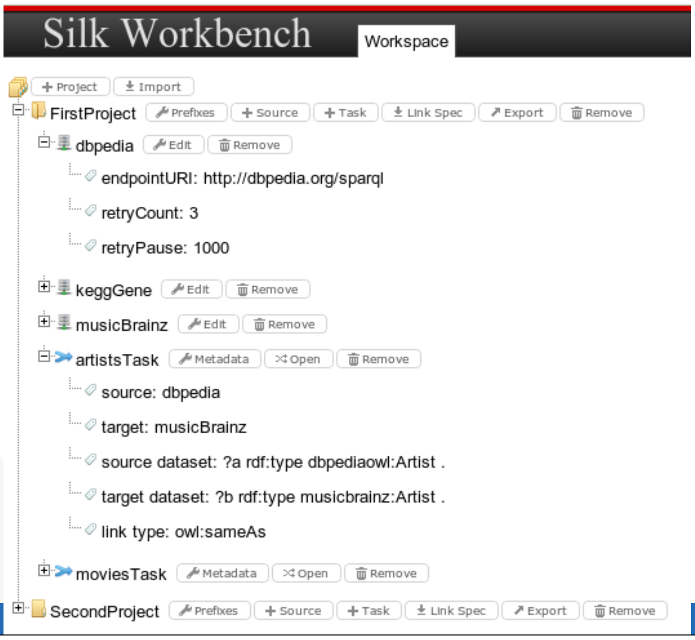
- Silk Workbench:


Open existing Linking Task from Workspace
Create new Linking Task
Edit the Link Specification
Evaluate the Link Specification
Commit the Linking Task
Update reference links

Link Specification Language



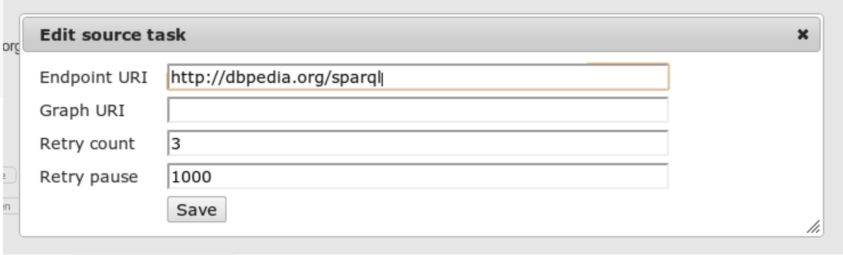
Workspace



```
FirstProject
  dbpedia
    endpointURI: http://dbpedia.org/sparql
    retryCount: 3
    retryPause: 1000
  keggGene
  musicBrainz
  artistsTask
    source: dbpedia
    target: musicBrainz
    source dataset: ?a rdf:type dbpediaowl:Artist .
    target dataset: ?b rdf:type musicbrainz:Artist .
    link type: owl:sameAs
  moviesTask
SecondProject
```

Workspace components

- A **project** holds the following information:
  - All URI prefixes which are used in the project.
  - A list of data sources
  - A list of linking tasks
- A **data source** holds all information that is needed by Silk to retrieve entities from it:

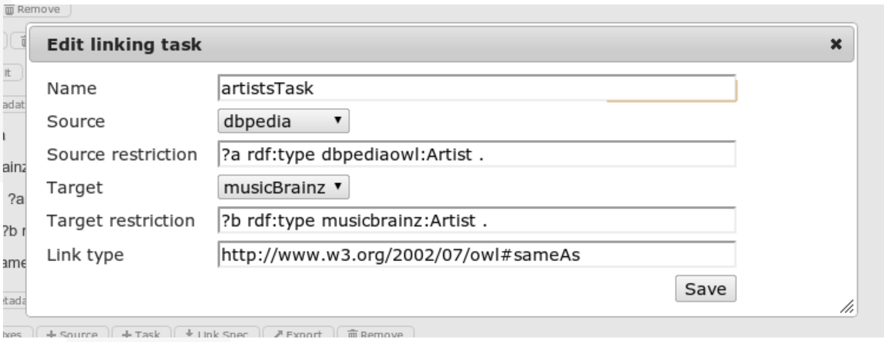


Edit source task

Endpoint URI	http://dbpedia.org/sparql
Graph URI	
Retry count	3
Retry pause	1000

Workspace components: Linking Tasks

- A **linking task** consists of the following elements:
  - Metadata
  - A link specification
  - Positive and negative reference links

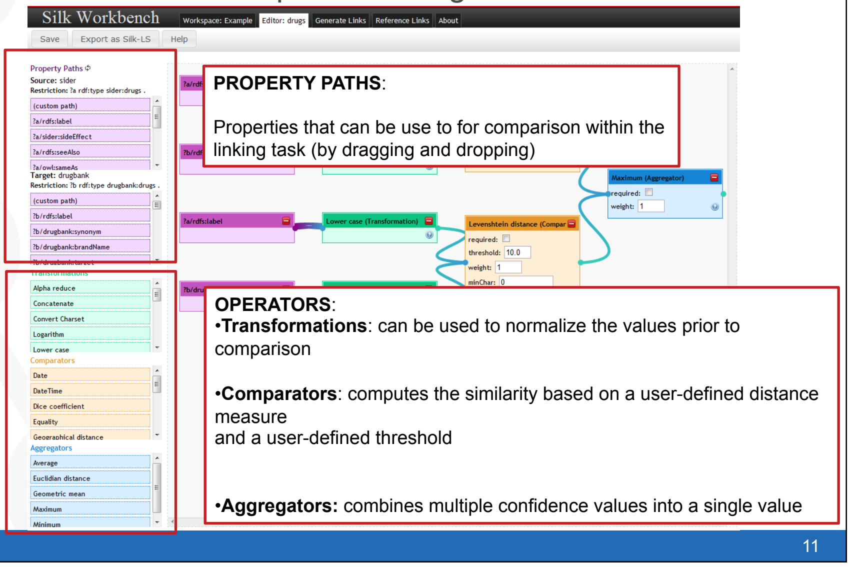


Edit linking task

Name	artistsTask
Source	dbpedia
Source restriction	?a rdf:type dbpediaowl:Artist .
Target	musicBrainz
Target restriction	?b rdf:type musicbrainz:Artist .
Link type	http://www.w3.org/2002/07/owl#sameAs

Linkage rules editor

- Clicking on the ***OPEN*** button opens the *Linkage Rules Editor* for a specific linking task



PROPERTY PATHS:

Properties that can be use to for comparison within the linking task (by dragging and dropping)

OPERATORS:
•**Transformations**: can be used to normalize the values prior to comparison

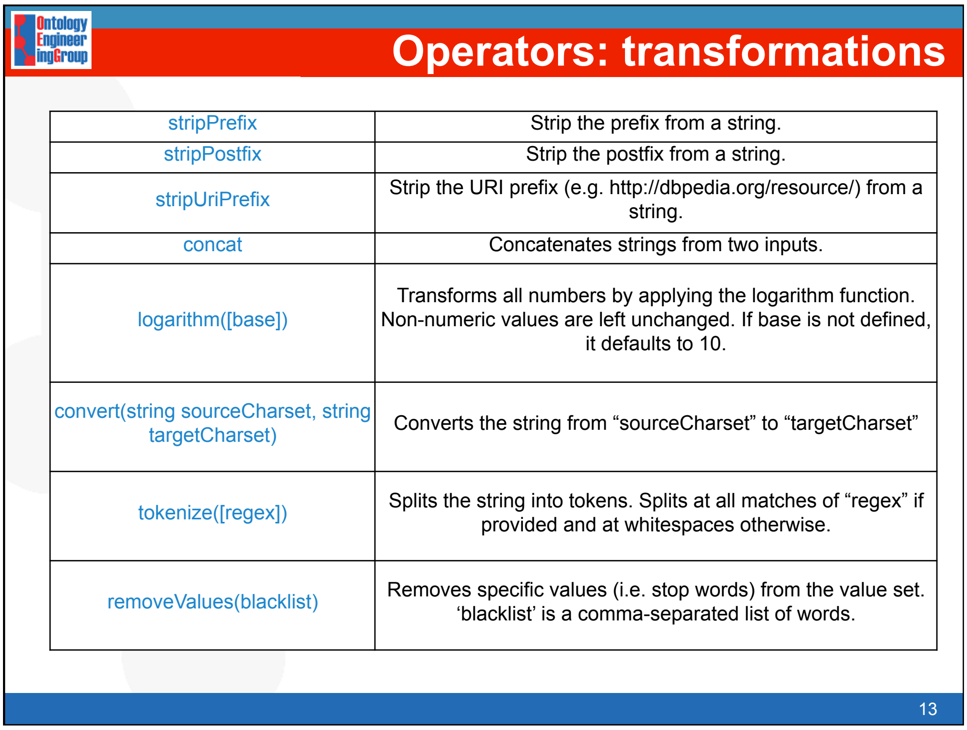
•**Comparators**: computes the similarity based on a user-defined distance measure
and a user-defined threshold

•**Aggregators:** combines multiple confidence values into a single value

Operators: transformations

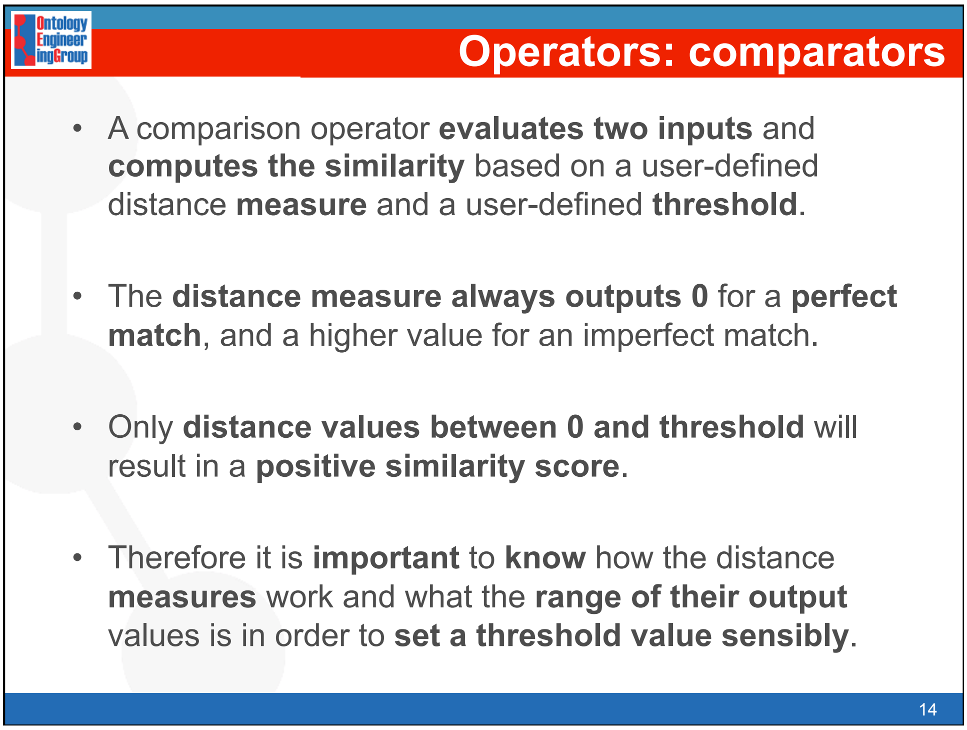
*A **transformation** can be used to normalize the values prior to comparison.*

Function and parameters	Description
removeBlanks	Remove whitespace from a string.
removeSpecialChars	Remove special characters (including punctuation) from a string.
lowerCase	Convert a string to lower case.
upperCase	Convert a string to upper case.
capitalize(allWords)	Capitalizes the string i.e. converts the first character to upper case. If ‘allWords’ is set to true, all words are capitalized and not only the first character. By default ‘allWords’ is set to false.
stem	Apply word stemming to the string.
alphaReduce	Strip all non-alphabetic characters from a string.
numReduce	Strip all non-numeric characters from a string.
replace(string search, string replace)	Replace all occurrences of “search” with “replace” in a string.
regexReplace(string regex, string replace)	Replace all occurrences of a regex “regex” with “replace” in a string.



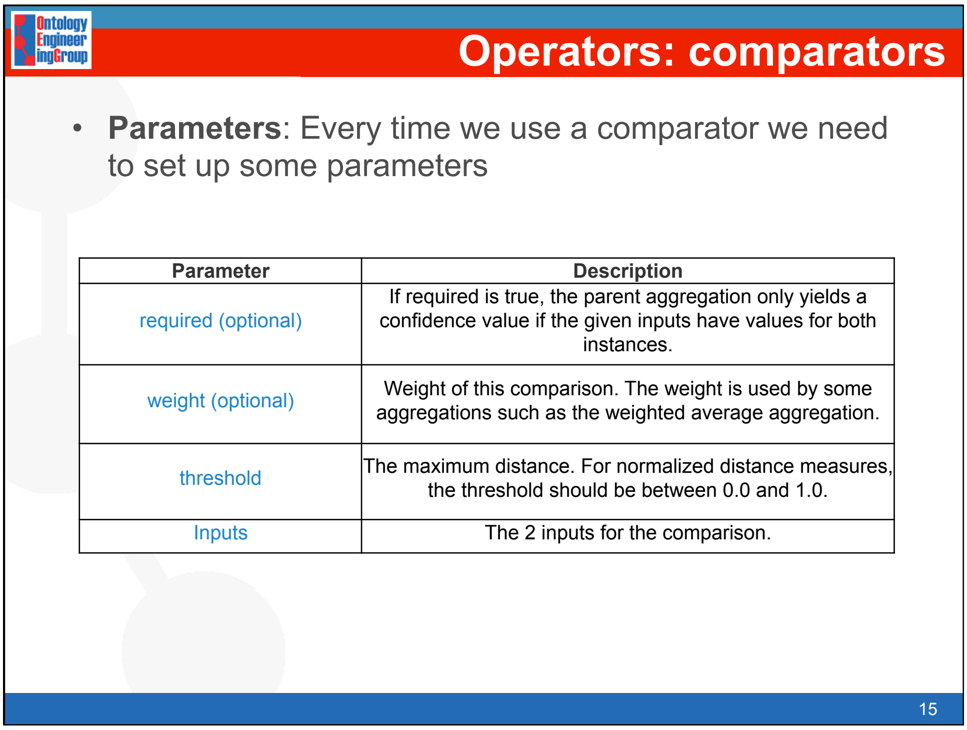
Operators: transformations

stripPrefix	Strip the prefix from a string.
stripPostfix	Strip the postfix from a string.
stripUriPrefix	Strip the URI prefix (e.g. http://dbpedia.org/resource/) from a string.
concat	Concatenates strings from two inputs.
logarithm([base])	Transforms all numbers by applying the logarithm function. Non-numeric values are left unchanged. If base is not defined, it defaults to 10.
convert(string sourceCharset, string targetCharset)	Converts the string from "sourceCharset" to "targetCharset"
tokenize([regex])	Splits the string into tokens. Splits at all matches of "regex" if provided and at whitespaces otherwise.
removeValues(blacklist)	Removes specific values (i.e. stop words) from the value set. 'blacklist' is a comma-separated list of words.



Operators: comparators

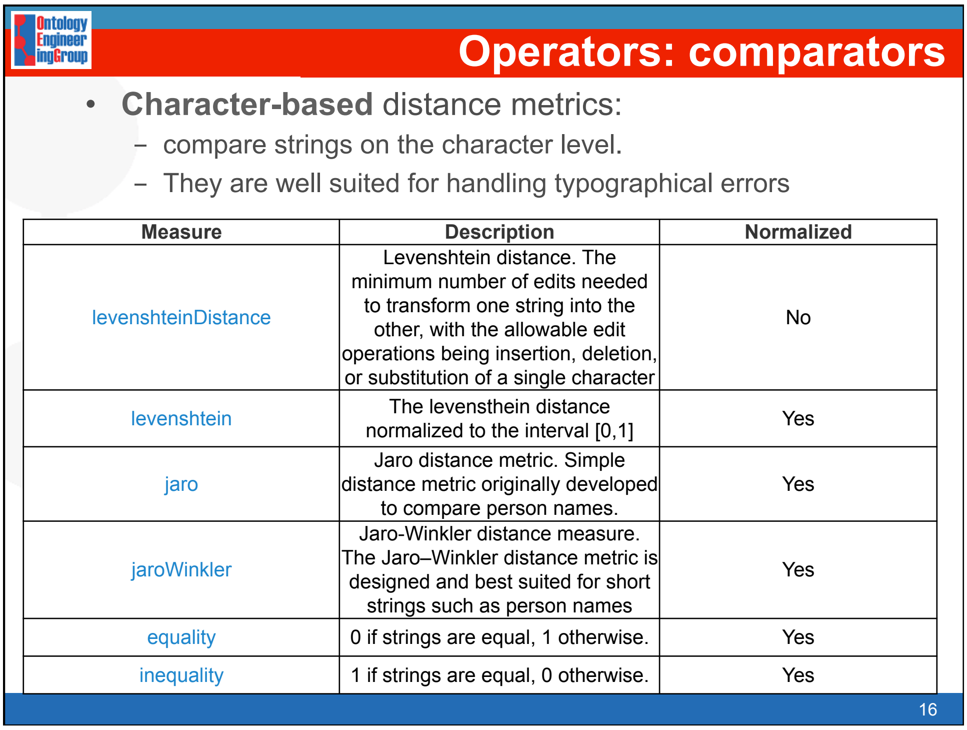
- A comparison operator **evaluates two inputs** and **computes the similarity** based on a user-defined distance **measure** and a user-defined **threshold**.
- The **distance measure always outputs 0** for a **perfect match**, and a higher value for an imperfect match.
- Only **distance values between 0 and threshold** will result in a **positive similarity score**.
- Therefore it is **important** to **know** how the distance **measures** work and what the **range of their output** values is in order to **set a threshold value sensibly**.



Operators: comparators

- **Parameters**: Every time we use a comparator we need to set up some parameters

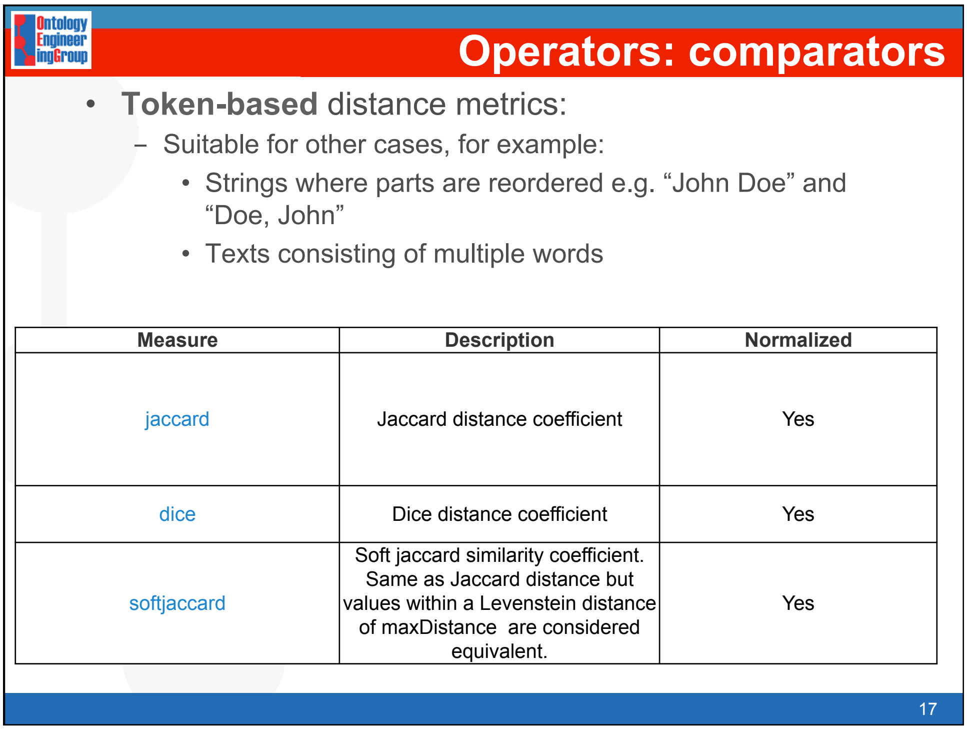
Parameter	Description
required (optional)	If required is true, the parent aggregation only yields a confidence value if the given inputs have values for both instances.
weight (optional)	Weight of this comparison. The weight is used by some aggregations such as the weighted average aggregation.
threshold	The maximum distance. For normalized distance measures, the threshold should be between 0.0 and 1.0.
Inputs	The 2 inputs for the comparison.



Operators: comparators

- **Character-based** distance metrics:
  - compare strings on the character level.
  - They are well suited for handling typographical errors

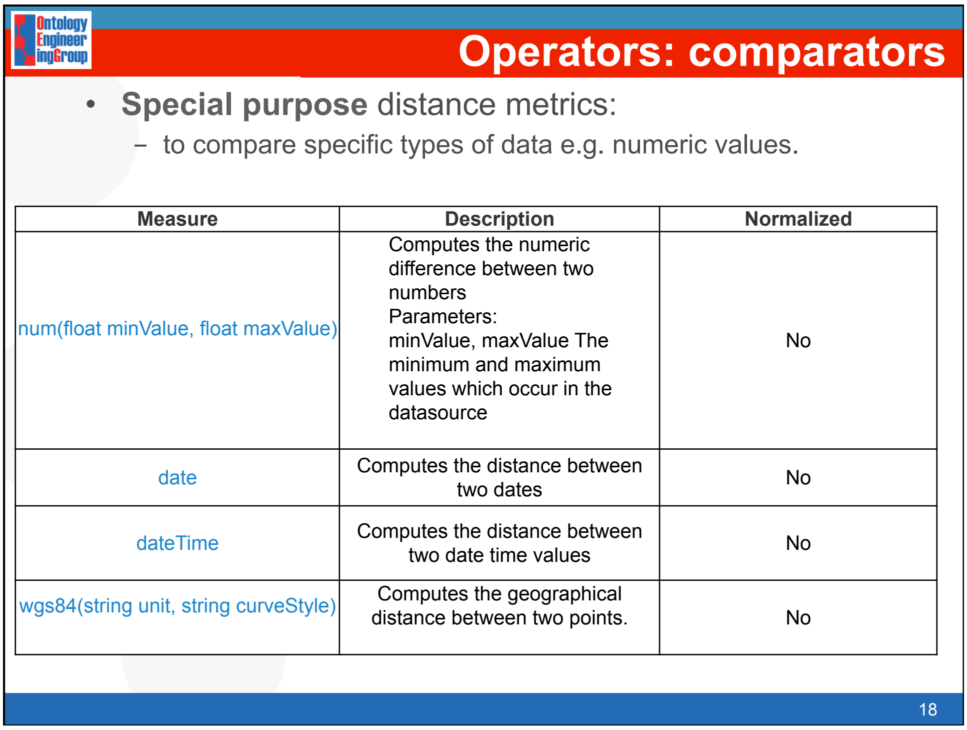
Measure	Description	Normalized
levenshteinDistance	Levenshtein distance. The minimum number of edits needed to transform one string into the other, with the allowable edit operations being insertion, deletion, or substitution of a single character	No
levenshtein	The levensthein distance normalized to the interval [0,1]	Yes
jaro	Jaro distance metric. Simple distance metric originally developed to compare person names.	Yes
jaroWinkler	Jaro-Winkler distance measure. The Jaro–Winkler distance metric is designed and best suited for short strings such as person names	Yes
equality	0 if strings are equal, 1 otherwise.	Yes
inequality	1 if strings are equal, 0 otherwise.	Yes



Operators: comparators

- **Token-based** distance metrics:
  - Suitable for other cases, for example:
    - Strings where parts are reordered e.g. “John Doe” and “Doe, John”
    - Texts consisting of multiple words

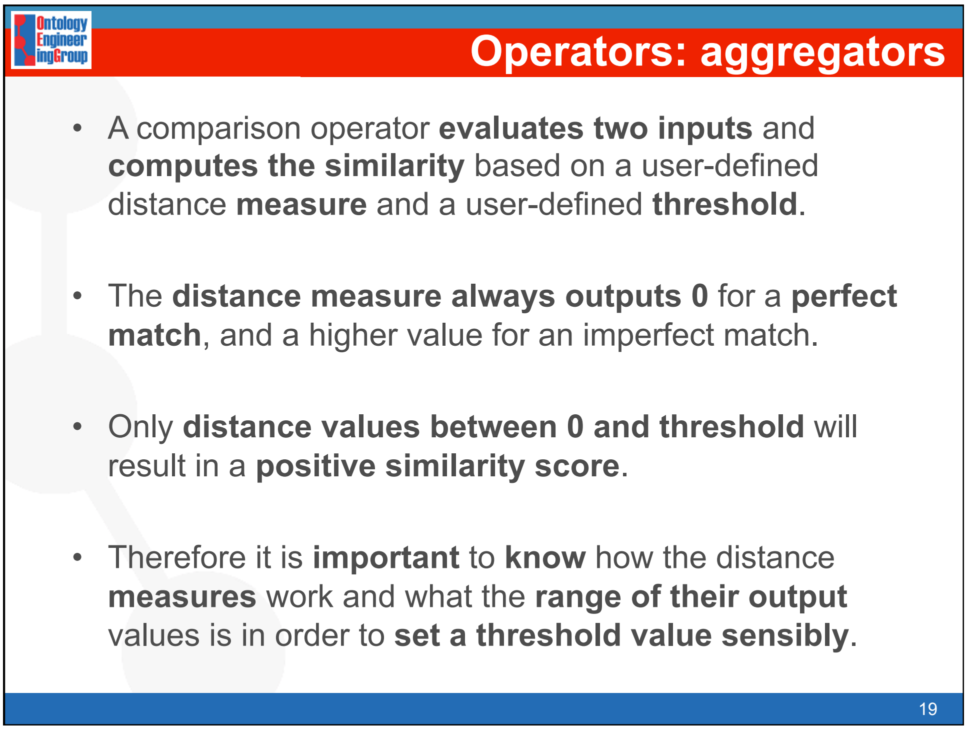
Measure	Description	Normalized
jaccard	Jaccard distance coefficient	Yes
dice	Dice distance coefficient	Yes
softjaccard	Soft jaccard similarity coefficient. Same as Jaccard distance but values within a Levenstein distance of maxDistance are considered equivalent.	Yes



Operators: comparators

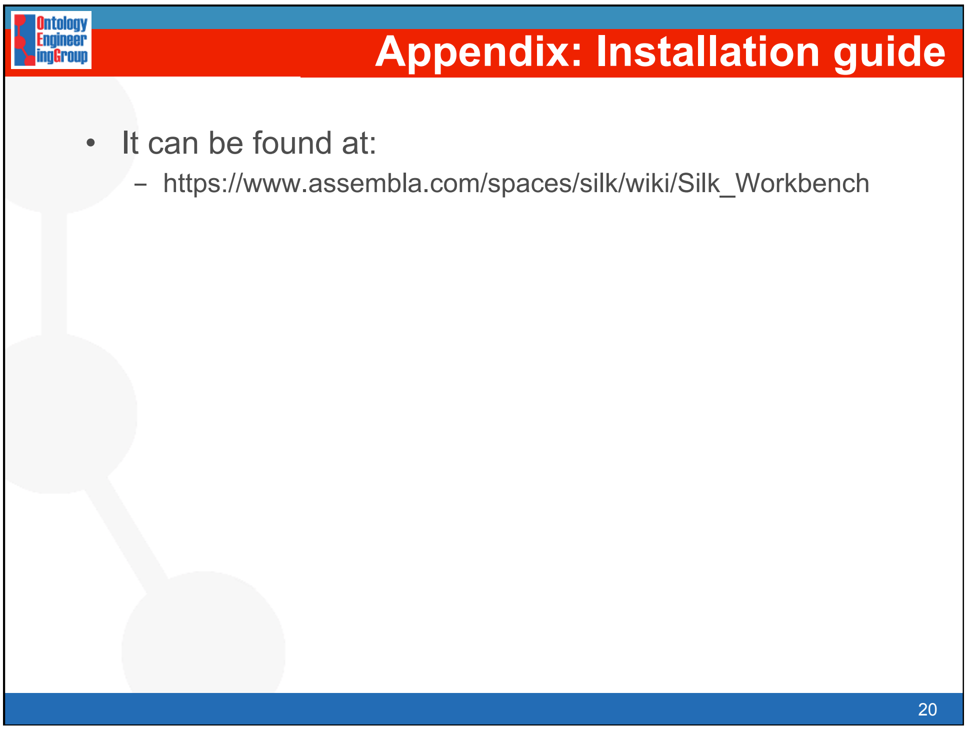
- **Special purpose** distance metrics:
  - to compare specific types of data e.g. numeric values.

Measure	Description	Normalized
num(float minValue, float maxValue)	Computes the numeric difference between two numbers Parameters: minValue, maxValue The minimum and maximum values which occur in the datasource	No
date	Computes the distance between two dates	No
dateTime	Computes the distance between two date time values	No
wgs84(string unit, string curveStyle)	Computes the geographical distance between two points.	No



Operators: aggregators

- A comparison operator **evaluates two inputs** and **computes the similarity** based on a user-defined distance **measure** and a user-defined **threshold**.
- The **distance measure always outputs 0** for a **perfect match**, and a higher value for an imperfect match.
- Only **distance values between 0 and threshold** will result in a **positive similarity score**.
- Therefore it is **important** to **know** how the distance **measures** work and what the **range of their output** values is in order to **set a threshold value sensibly**.



Appendix: Installation guide

- It can be found at:
  - https://www.assembla.com/spaces/silk/wiki/Silk_Workbench