

Wikimania 2009

WikiWord

Multilingual image search
and more

Daniel Kinzler



Image Search

1. Specify language, enter a word (term)
2. Find meanings for term
(topics/concepts/pages)
3. Find galleries and categories from Commons
4. List concepts, show images for each

Image Search

Term: Italian Images:
English

Note: this is a thesaurus lookup, not a full text search. Only exact matches are considered, matching is case-sensitive.

it: Roma

439 Roma (#9000283) (gallery)

Images:



3 Diocesi di Roma (#9000280) (gallery)

Images:



Wikipedia

- **Goal: *find and navigate topics***
- Wikipedia describes topics (concepts)
- Links define relations
- Links and titles define terms for concepts
- Interwiki-links provide cross-language concepts

Thesaurus

- Concepts + Terms + Relations = ***Thesaurus***
- Uses:
 - list meanings
 - navigate topics
 - relatedness (semantic proximity)
 - disambiguate in context

WikiWord: Relations

- Pages linked both ways: *related* topics (mutual relevance)
- Language-links overlap: *similar* topics
- Categories: *broader/narrower* topics
- Category main article: *same* topic
- Language-links both ways: *same* topic

WikiWord: Terms

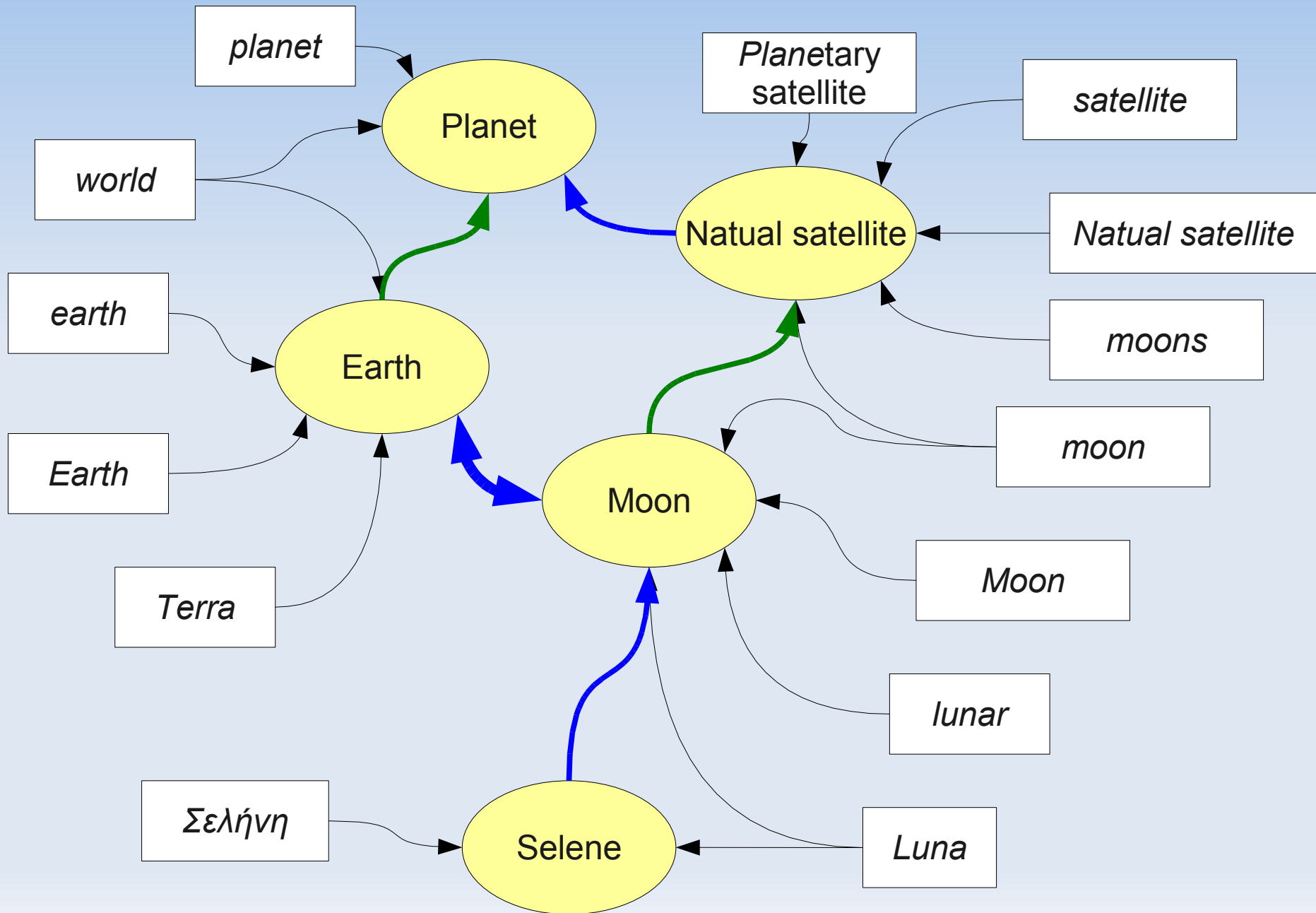
- Page titles
- Redirects
- Disambiguations
- **Link text!**
 - `[[Wikimedia Foundation|WMF]]`
 - *Frequency* of use
 - *Context* of use (not currently used)

WikiWord: Properties

- Infoboxes
- Biography templates
- Category names

- Rudimentary
- Use DBpedia

Structure

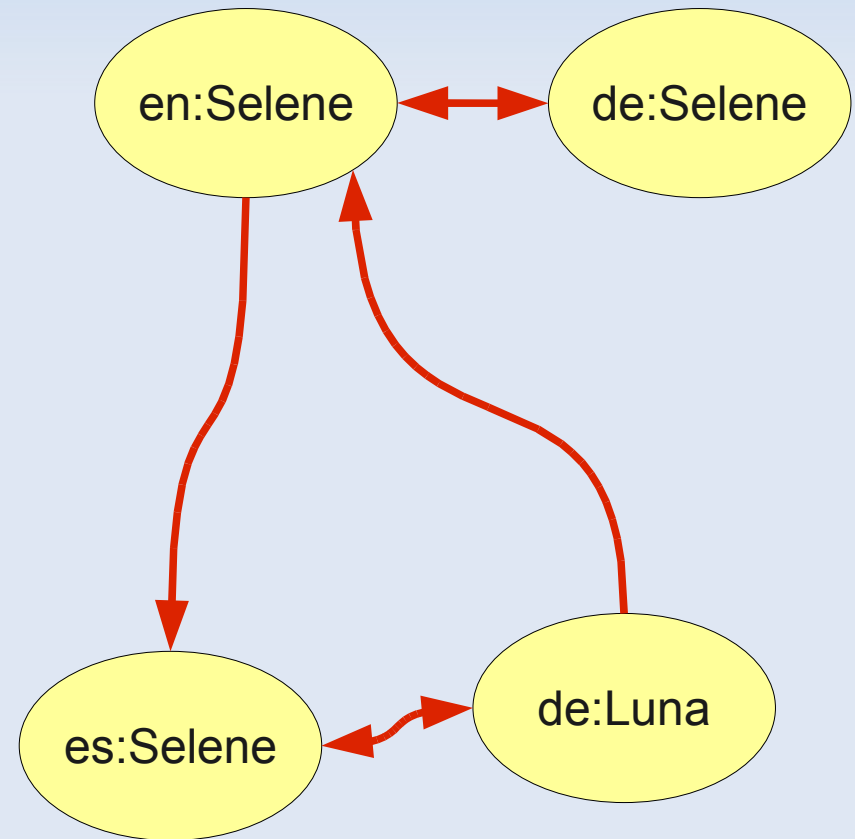
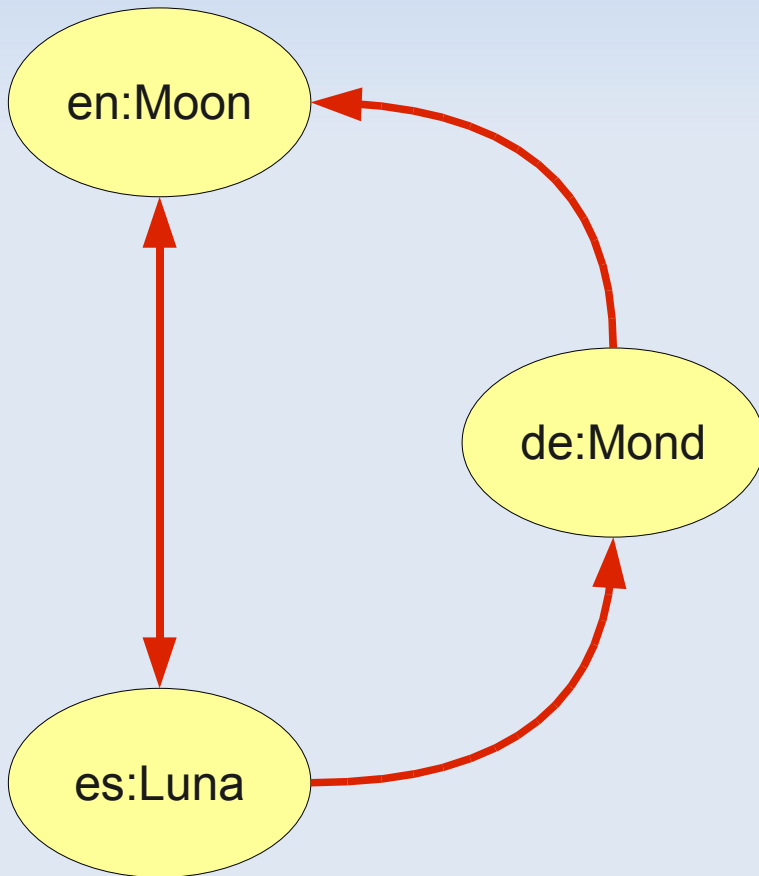


WikiWord: Merging

- Put concepts from all languages into one thesaurus
- Remember which language each concept came from
- If two concepts reference each other with language links, merge them
- Remember which languages are covered by a merged concept
- Each concept can cover each language only once

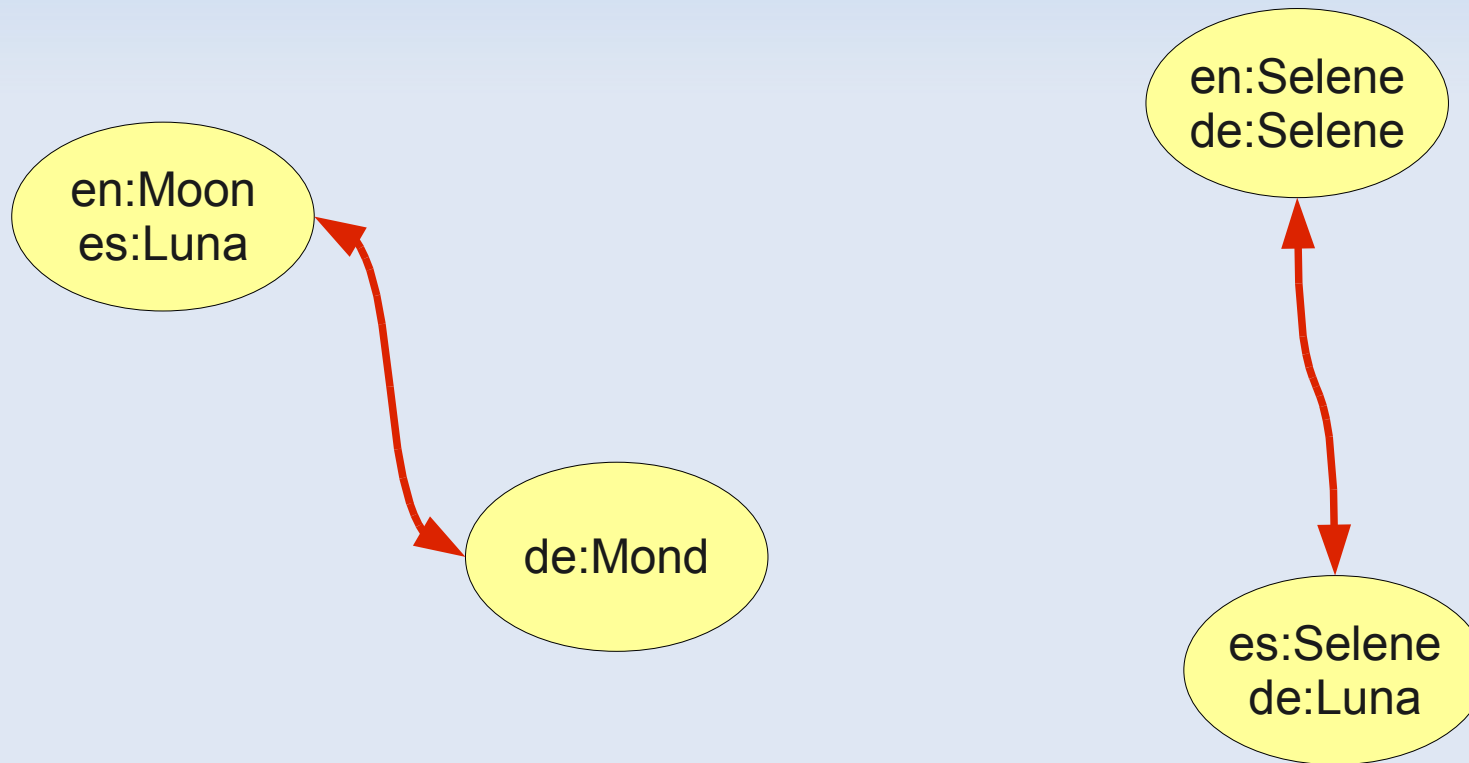
Merging

- Put concepts from all languages into one thesaurus
- Remember which language each concept came from



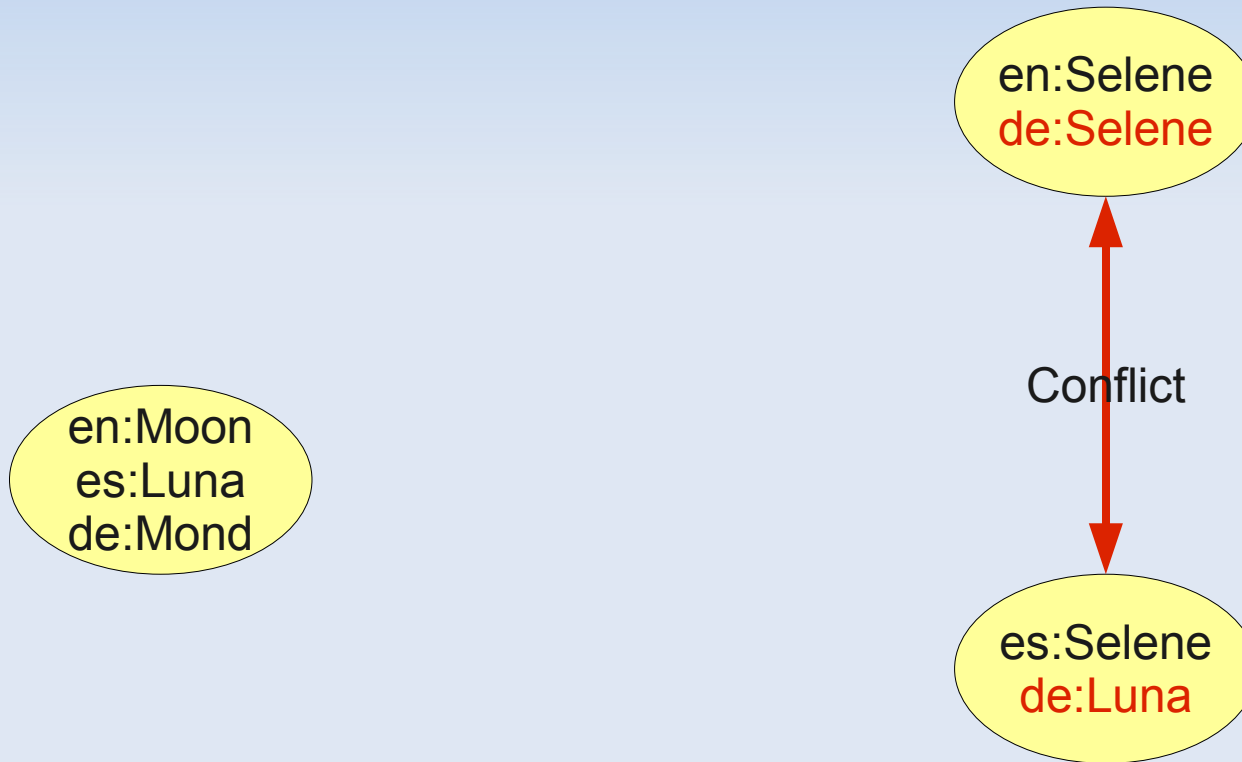
Merging

If two concepts reference each other with language links, merge them
Remember which languages are covered by a merged concept



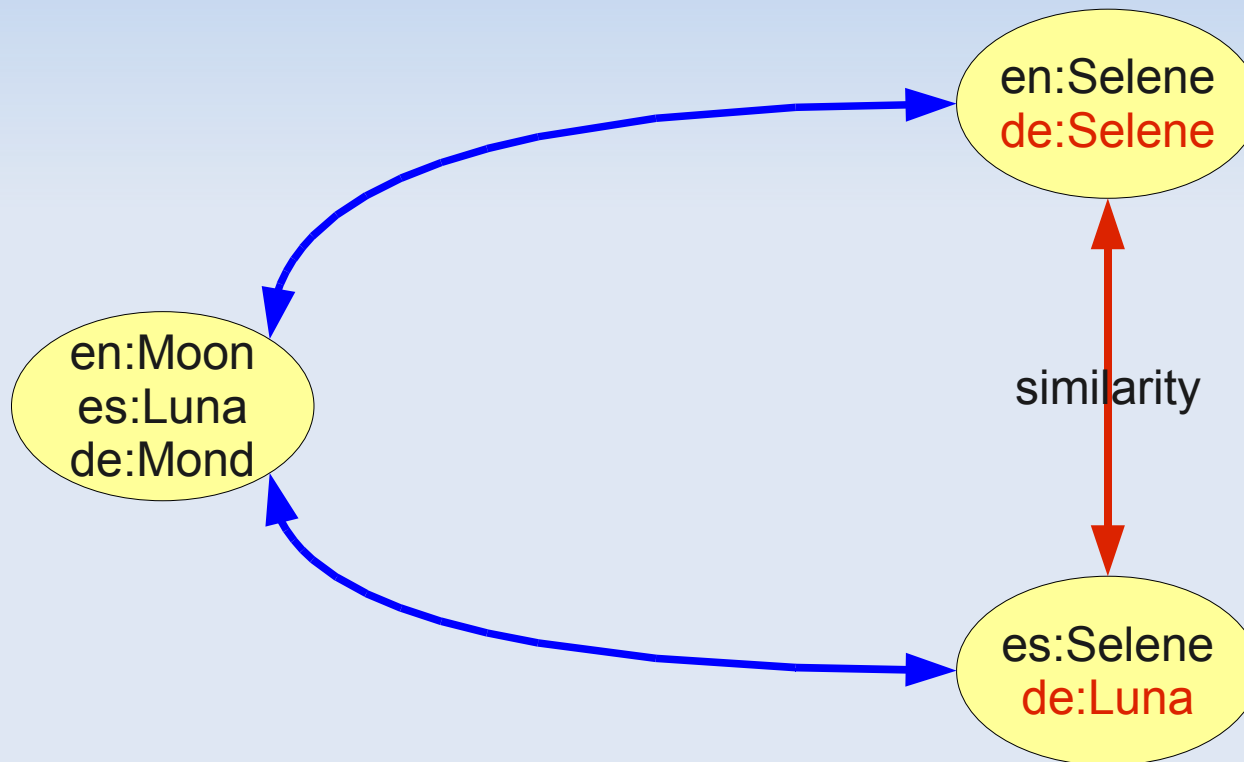
Merging

- Each concept can cover each language only once

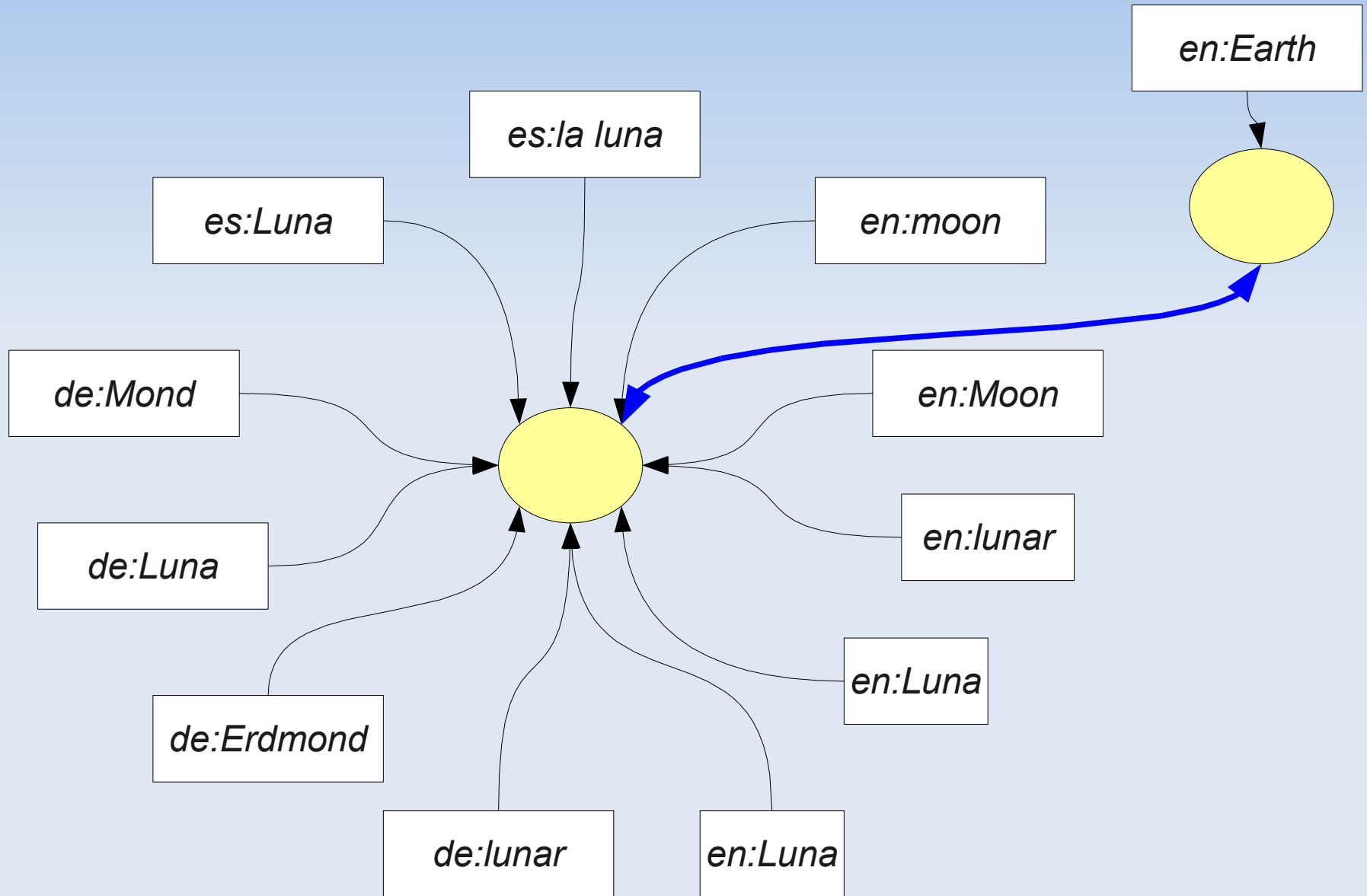


Merging

- Relations between concepts are accumulated



Multilingual Thesaurus



Using WikiWord

- Term → Concepts
- Concept → terms, properties, images
- Lookup, navigation, translation

- Commons: pretend...
 - Commons is a language
 - `{{Commons:Foo}}` is a language link

Searching with WikiWord

- List concepts for a term in a given language
- Allow navigation to related topics

Search

Term: English
English

Note: this is a thesaurus lookup, not a full text search. Only exact matches are considered, matching is case-sensitive.

en: moon

839 [Moon](#) (#25002396)

Definition: The Moon is Earth's only natural satellite, and is the fifth largest one in the Solar System.

154 [Natural satellite](#) (#24041541)

Definition: A natural satellite or moon is an celestial body that orbits another celestial body, planet or smaller, which is called the primary.

11 [Asteroid moon](#) (#21313621)

Definition: An asteroid moon is an asteroid that orbits another asteroid as its natural satellite.

11 [Mooning](#) (#22139153)

Definition: Mooning is the act of displaying one's bare buttocks by removing clothing, e.g. by lowering the back side of one's trousers and underpants, usually bending over, whether also exposing the genitals or not.

1 [Kalyke \(moon\)](#) (#25003625)

Definition: Kalyke, or Jupiter XXIII, is a retrograde irregular satellite of Jupiter.

Found 5 items.

The WikiWord Navigator is part of the [Wikimedia](#) project [WikiWord](#)

<http://toolserver.org/~daniel/wikiword/wikiword.php>

Search

Term: German ↕ go
English ↕ translate

Note: this is a thesaurus lookup, not a full text search. Only exact matches are considered, matching is case-sensitive.

de: Mond -> en

741 [Mond](#) (#25002396)

Definition: Der Mond ist der einzige natürliche Satellit der Erde.

61 [Satellit \(Astronomie\)](#) (#24041541)

Definition: Ein Satellit ist in der Astronomie ein natürlich entstandenes astronomisches Objekt, das ein Anderes – wie eine Galaxie, einen Planeten, einen Zwergplaneten oder auch einen Kleinkörper wie einen Asteroiden – umkreist.

3 [Mond \(Heraldik\)](#) (#1106347)

Definition: Der Mond oder auch Mondschein genannt, ist eine gemeine Figur in der Heraldik.

2 [Mond \(Mythologie\)](#) (#2214702)

2 [Gestirn#Mond](#) (#1270433)

2 [Mondgottheit](#) (#1286908)

Definition: Als Mondaottheit aelten in den Mythologien der unterschiedlichsten

[Moon](#) (#25002396)

Definition: The Moon is Earth's only natural satellite, and is the fifth largest one in the Solar System.

[Natural satellite](#) (#24041541)

Definition: A natural satellite or moon is an celestial body that orbits another celestial body, planet or smaller, which is called the primary.

Searching with WikiWord

- Not a text search
- Can't be used directly to categorize
- Still powerful and useful

Image Search with WikiWord

- Images from pages, galleries and categories
- Ranking: *used* images are best
- To do: favor featured/quality/valued images

- proof of concept implementation
 - For each concept, show images
 - Using old data, fake link to commons
 - Alpha-test with new data model: popes

Searching with WikiWord

Term: Italian Images:
English

Note: this is a thesaurus lookup, not a full text search. Only exact matches are considered, matching is case-sensitive.

it: Roma

439 [Roma](#) (#9000283) ([gallery](#))

Images:



3 [Diocesi di Roma](#) (#9000280) ([gallery](#))

Images:



<http://toolserver.org/~daniel/wikiword-trunk/wikiword.php>

Statistics

- Original WikiWord thesaurus (2007):
- en, de, fr, nl, no
- 8 million pages
- 12 million concepts (4 million "blue")
- 22 million term associations
- 100 million relations between concepts
- New thesaurus: nearly twice the size

Perspective

- Fresh complete thesaurus (really soon now)
 - This should be come a regular service
 - RDF/SKOS
- Usable image search (by the end of the year)
 - Alternative interface to commons
- Running on a toolserver machine
- Tricky because of huge size
- Development is ongoing

Thank You!

<https://wiki.toolserver.org/view/WikiWord>

<http://brightbyte.de/page/WikiWord>

<http://toolserver.org/~daniel/wikiword/wikiword.php>



CC-BY-SA 3.0

