

Getting geographical answers from Wikipedia: the GikiP pilot at CLEF

Diana Santos, Nuno Cardoso

Other organizers: **Paula Carvalho, Yvonne Skalban**

Participants: **Nuno Cardoso, Iustin Dornescu,
Johannes Leveling, Sven Hartrumpf**

Acknowledgements

- The organization work was done in the scope of Linguateca, contract no. 339/1.3/C/NAC, project jointly funded by the Portuguese Government and the European Union, and administratively led by FCCN.
- This presentation was also partially funded by SINTEF ICT in the scope of GikiP follow-up that was submitted to CLEF by Nuno Cardoso (Univ. of Lisbon, Linguateca, and SINTEF ICT)



Purpose of this presentation

- Present the general pilot and its outcome
- Give an idea of plans for next year
- The participants will present their work at 15:30 in the *Hornung* room at the GeoCLEF parallel session (14:00-16:00)

Never heard about *Linguateca*?

- It is a (Portuguese-)government funded initiative to significantly raise the quality and availability of resources for the **computational processing of Portuguese**
- After an initial plan for discussion by the community (white paper, in 1999) a network was launched, headed by a small group (Linguateca's Oslo node) at SINTEF ICT, having as main goal to guarantee that
 - Information was provided and gathered at one place on the Web
 - Resources were made public, maintained, and further developed in connection with the scientific community
 - Evaluation initiatives were launched: Morfolimpíadas, HAREM
 - ... and **with CLEF since 2004!**

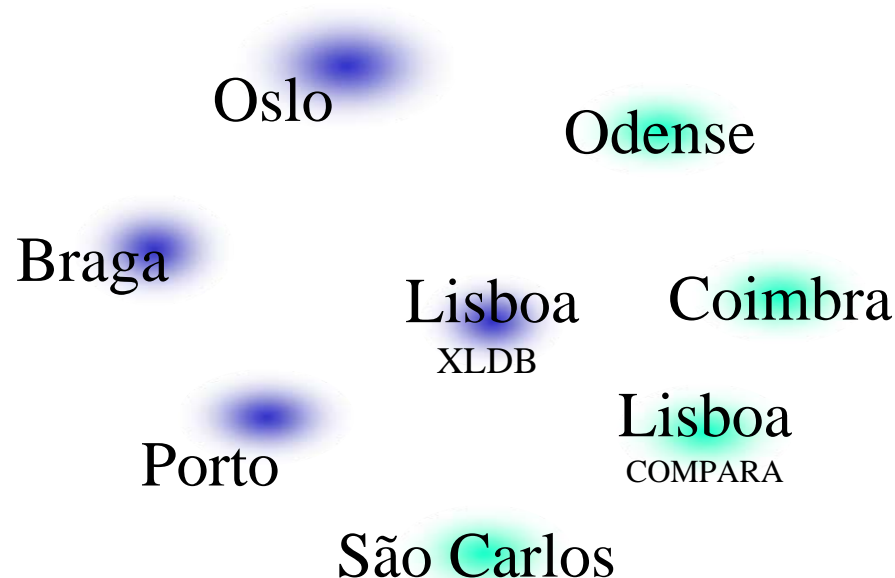
Linguateca, a project for Portuguese

- A distributed resource center for Portuguese language technology

IRE model

- Information
- Resources
- Evaluation

www.linguateca.pt



Language engineering at SINTEF

- Question answering
- Ontologies
- Geographical reasoning
- Contrastive studies
- Information extraction (NER, etc.)
- Corpus search
- Evaluation
- Crossmedia applications



Publication management
Log analysis

This is the group that inherited and hosted Linguateca experience in SINTEF and most probably will back up the next edition of GikiP

What is GikiP?

- GikiP is a pilot evaluation task run under the GeoCLEF umbrella
- Task: *Find Wikipedia entries (i.e. articles) that answer a particular information need which requires geographical reasoning of some sort*
- Scientific goal: Create synergies between the geographic information retrieval (GIR) and the question answering (QA) “disciplines”.
- Practical goal: Wouldn't it be good if we had systems that could mediate between us & Wikipedia, answering our complex questions, no matter the language?

In 2007, we had German, Portuguese and English

Topic titles in GikiP 2008

| ID | English topic title |
|------|--|
| GP1 | Which waterfalls are used in the film “The Last of the Mohicans”? |
| GP2 | Which Vienna circle members or visitors were born outside the Austria-Hungarian empire or Germany? |
| GP3 | Portuguese rivers that flow through cities with more than 150,000 inhabitants |
| GP4 | Which Swiss cantons border Germany? |
| GP5 | Name all wars that occurred on Greek soil. |
| GP6 | Which Australian mountains are higher than 2000 m? |
| GP7 | African capitals with a population of two million inhabitants or more |
| GP8 | Suspension bridges in Brazil |
| GP9 | Composers of Renaissance music born in Germany |
| GP10 | Polynesian islands with more than 5,000 inhabitants |
| GP11 | Which plays of Shakespeare take place in an Italian setting? |
| GP12 | Places where Goethe lived |
| GP13 | Which navigable rivers in Afghanistan are longer than 1000 km? |
| GP14 | Brazilian architects who designed buildings in Europe |
| GP15 | French bridges which were in construction between 1980 and 1990 |

Topic titles in GikiP 2008

| ID | English topic title |
|------|--|
| GP1 | Which waterfalls are used in the film “The Last of the Mohicans”? |
| GP2 | Which Vienna circle members or visitors were born outside the Austria-Hungarian empire or Germany? |
| GP3 | Portuguese rivers that flow through cities with more than 150,000 inhabitants |
| GP4 | Which Swiss cantons border Germany? |
| GP5 | Name all wars that occurred on Greek soil. |
| GP6 | Which Australian mountains are higher than 2000 m? |
| GP7 | African capitals with a population of two million inhabitants or more |
| GP8 | Suspension bridges in Brazil |
| GP9 | Composers of Renaissance music born in Germany |
| GP10 | Polynesian islands with more than 5,000 inhabitants |
| GP11 | Which plays of Shakespeare take place in an Italian setting? |
| GP12 | Places where Goethe lived |
| GP13 | Which navigable rivers in Afghanistan are longer than 1000 km? |
| GP14 | Brazilian architects who designed buildings in Europe |
| GP15 | French bridges which were in construction between 1980 and 1990 |

Which Spanish writers lived in America in the XIX century?

- Answers in a lot of Wikipedia languages
- Kind of answers: NE (names)
- Assessment relatively easy
- Promotes multilinguality and crosslinguality

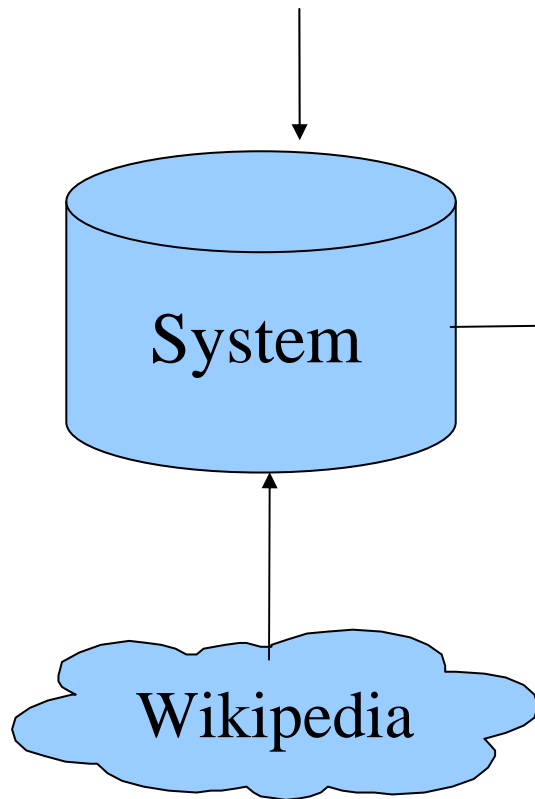
GikiP's collection: Wikipedia

Wikipedia is a great collection to work on:

- Available
- Truly multilingual (dozens of languages)
- Spans several subjects, and their users/contributors strive for consistency
- According to some, documents are well written, constantly reviewed and their content validated
- Rich content, structure and metadata that can be explored (categories, infoboxes, links)
- Multimedia resource
- Widely used!!!! A lot of users with a lot of different information needs

GikiP: the simplest example

Topic: “Which Swiss cantons border Germany?”



Returned answers:

de/k/a/n/Kanton Aargau.html
de/k/a/n/Kanton Basel-Landschaft.html
de/k/a/n/Kanton Basel-Stadt.html
de/k/a/n/Kanton Zürich.html



en/a/a/r/Aargau.html
en/b/a/s/Basel-Land.html
en/c/a/n/Canton of Zurich.html
en/t/h/u/Thurgau.html



pt/a/r/g/Argóvia (cantão).html
pt/b/a/s/Basiléia-Campo.html
pt/b/a/s/Basiléia-Cidade.html
pt/c/a/n/Cantão de Zurique.html



The system should...

- ...understand what the topic really wants (a list of cities, rivers or mountains), and its restrictions (a given population/length/height threshold)
- ...reason over the Wikipedia collection and over the geographic domain (i.e., “does this river flows to the Atlantic Ocean?”)
- ...return Wikipedia pages for the answers: not lists, not overview pages, just the answers.

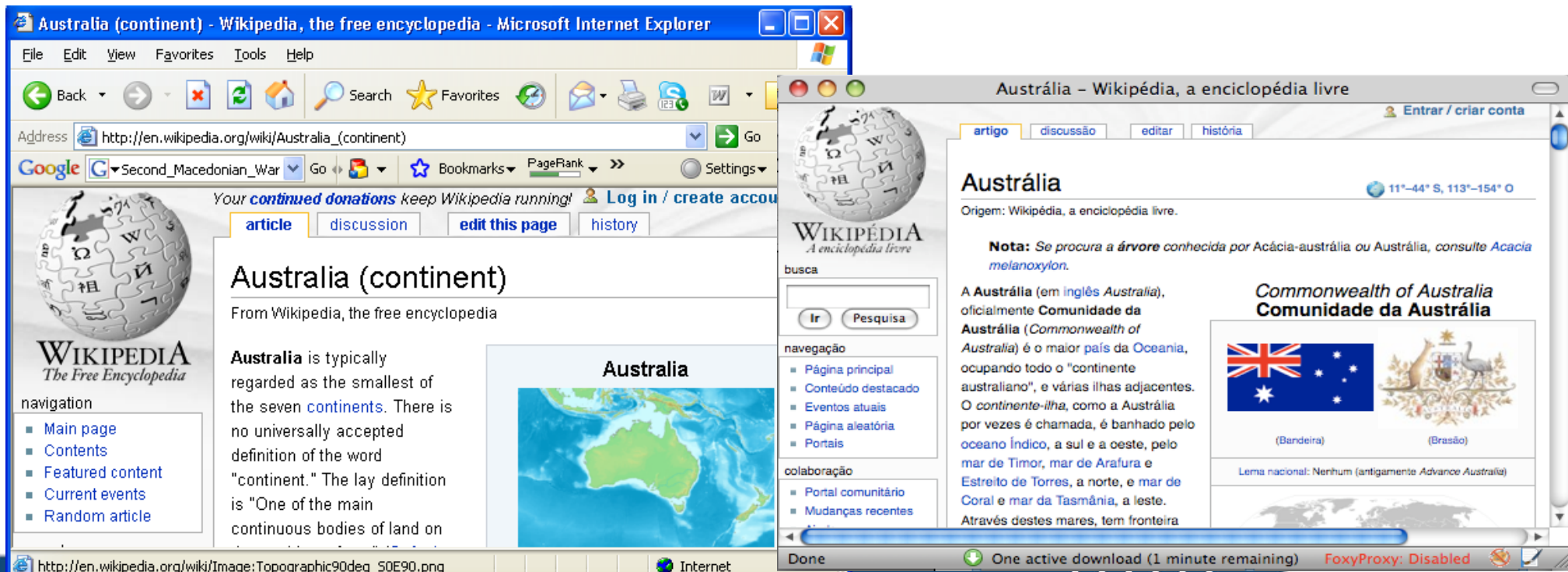
Interesting issues (1)

- Names change, roles change!
- Topic: *African capitals...*



Interesting issues (2)

- Different languages, different meanings of geographic scope
- *Australia*: both a continent and a country in EN, but only a country in PT (continent: *Oceânia*)
- Topic: *The highest mountains of Australia...*



Interesting issues (3)

- Different languages, different information sources, different data
- Ex: *African capitals with more than x habitants*



Wikipedia EN on “Harare”:

| | |
|--------------------------|------------------------|
| Country | Zimbabwe |
| Province | Harare |
| Founded | 1890 |
| Incorporated (city) | 1935 |
| Government | |
| - Mayor | Muchadeyi Masunda |
| Elevation ^[1] | 1,490 m (4,888 ft) |
| Population (2006) | |
| - City | 1,600,000 |
| - Urban | 2,800,111 estimated |

Wikipedia PT
on “Harare”:

| Harare | |
|--|----------------------|
| Capital | Harare |
| População | 1.903.510 habitantes |
| Censo | 2002 |
| Área | 872 km² |
| Densidade | 2.182,92 hab/km² |
| Mapa | |
|  | |

Wikipedia DE on “Harare”:

| Wappen | Karte |
|---|---|
|  |  |
| Basisdaten | |
| Geografische Lage: | 17° 51′ 50″ S, 31° 1′ 47″ O |
| Höhe: | 1.490 m ü. NN |
| Fläche: | 872 km² |
| Einwohner: | 1.903.510 (2006) |
| Bevölkerungsdichte: | 2.183 Einwohner/km² |

Interesting issues (4)

- Not all questions can be answered easily by a person!
 - Topics GP2 and GP15 had zero hits
- For example: “Name all wars that occurred on Greek soil”
 - There is no straightforward category in Wikipedia to start with.
 - Even if there were a “Greek War” category, would it include only wars fought on Greek soil, or all wars involving Greece?
 - Temporal issues: How was the Greek soil back then? Narrower or longer than today's boundaries?

See the topic typology initially presented at GIR06 and adopted by GeoCLEF in Gey et al. (2006)

Interesting issues (5)

- Reasoning over the geographic domain
- Topic GP11: “Which plays of Shakespeare take place in an Italian setting?”

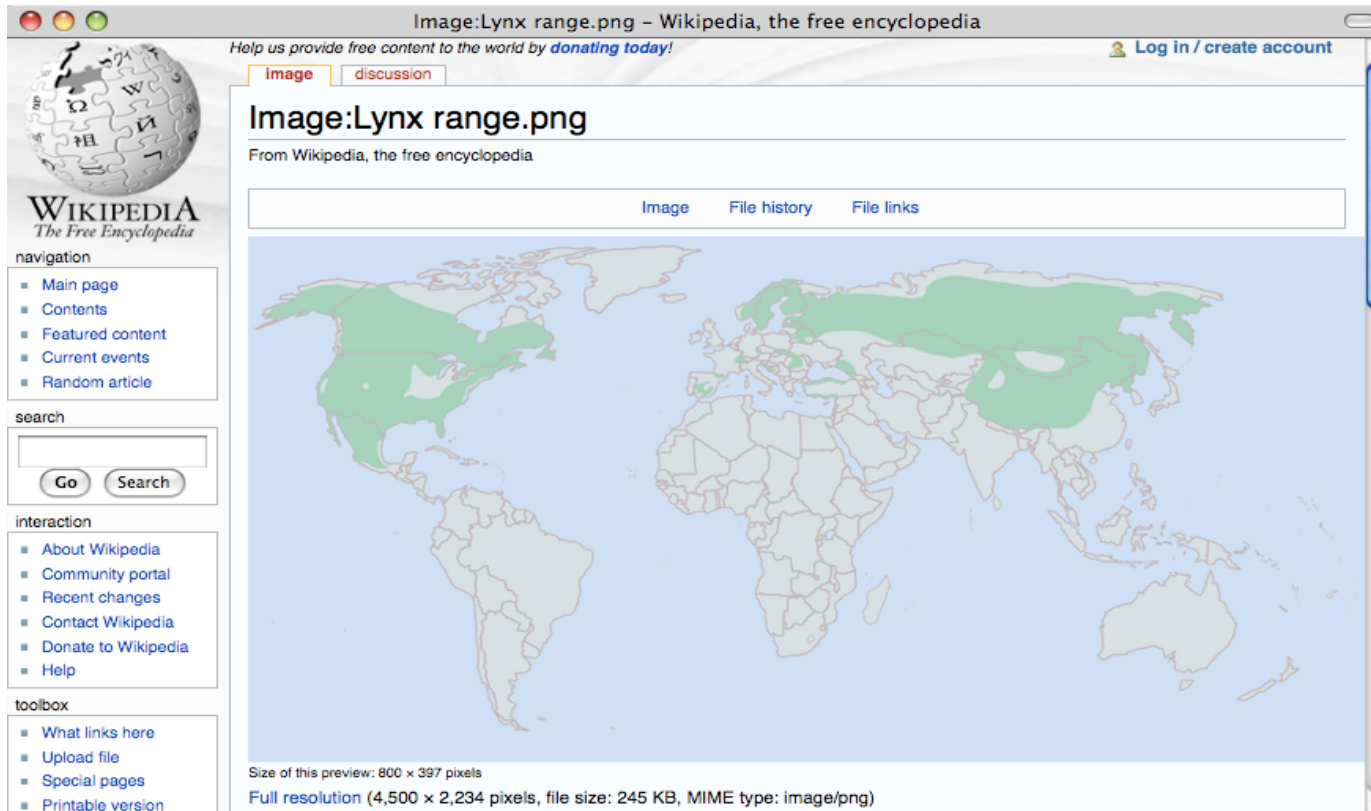
The screenshot shows the Wikipedia page for 'The Merchant of Venice'. The browser window title is 'The Merchant of Venice - Wikipedia, the free encyclopedia'. The page content includes a navigation sidebar on the left with links like 'Main page', 'Contents', 'Featured content', 'Current events', 'Random article', and 'Advanced search'. The main content area has a heading 'The Merchant of Venice' and a subheading 'From Wikipedia, the free encyclopedia'. A red box highlights the text: 'The Merchant of Venice is a play by William Shakespeare, believed to have been written between 1596 and 1598.' To the right of this text is an image of the title page of the first quarto of the play, which reads 'The most excellent Historie of the Merchant of Venice.' and 'Written by William Shakespeare.' A red arrow points from the highlighted text to the image. The bottom of the page shows a 'FoxyProxy: Disabled' status bar.

“is Venice in Italy?”

Easy question for humans,
but not so straightforward
for a machine...

GikiP's future (1)

- Why not mix images and text?
- Example: “Name the countries that still have lynxes”



A Canadian Lynx from the Philadelphia Zoo with distinct lynx tip and ruff with black bars.

GikiP's future (2)

More complex topics

- “Portuguese cities founded before 1500 with rivers larger than 100 km and featuring a Moorish castle”

also using images and text

- “Which Swiss cantons have a lion on their flag?”
- “Find portraits of married women in the 18th century”

Users express their needs clearly in their language;
the systems must adapt to the user, not the other way around.

GikiP's future (3): presentation issues

- instead of a list of places, one would like to have a coherent text (list)
- Places where Goethe lived:
 - Born in X, moved to Y, ... spent some months in Z, ... Died in W
- Places where X studied
 - Department of Y, University of Z, in the city of W, in U (country)
- People who worked with A
 - B, from Y, in 19xx-19yy
 - Z, from U, in 19zz...
- A map with Shakespeare's plays
- Buildings where by whom when

GikiP 2008: aggregated results

| Topic | results | correct | |
|------------------|-----------|-------------------------------|--------------------------------|
| ■ GP1 | 5 | 1 | 20% waterfalls |
| ■ GP7 | 90 | 33 | 36.6% African capitals |
| ■ GP10 | 53 | 2 | 3.8% Polynesian islands |
| ■ GP11 | 35 | 23 | 65.7% Shakespeare |
| ■ Total | 662 | 179 | 27.0% |
| | | | |
| ■ German (4) | 33.2 | (22.6; 26.6; 34.7; 49.0) | |
| ■ English (3) | 35.0 | (19.4; 20; 65.7) | |
| ■ Portuguese (3) | 14.2 | (4.1; 10.0; 28.6) | |
| ■ Other (5) | 25.3 | (3.8; 11.1; 30.4; 36.7; 44.4) | |

GikiP's evaluation measure: $N * N / \text{total} * \text{mult}$

- **Directly proportional to the number of correct hits (N):** the more correct answers the system gets, the better
- **Directly proportional to the system's precision (N/total):** the less incorrect answers the systems gets, the better
- **Directly proportional to multilinguality (mult):** the more languages it retrieves answers in, the better
 - Should depend of the existence of answers in that language
 - Should filter out exactly similar answers, and/or present them together
 - Should be especially aware of non-transparent mappings, or inconsistent mappings (so that the multilinguality was really useful even for a monolingual user)

More on multilinguality

- Number of hits in the judgment pool

- German English Portuguese

- 233 255 174 Total

- 31 86 59 Correct (176)

- 0 34 11 Unique correct

- DE: 5 EP: 21 DEP: 25 DP: 1

- Number of distinct answers: $0+34+11+5+21+25+1 = 97$

Topics in GikiP 2008: unique P

| ID | English topic title | |
|------|--|-----|
| GP1 | Which waterfalls are used in the film “The Last of the Mohicans”? | |
| GP2 | Which Vienna circle members or visitors were born outside the Austria-Hungarian empire or Germany? | 1 P |
| GP3 | Portuguese rivers that flow through cities with more than 150,000 inhabitants | 3 P |
| GP4 | Which Swiss cantons border Germany? | |
| GP5 | Name all wars that occurred on Greek soil. | 3 P |
| GP6 | Which Australian mountains are higher than 2000 m? | |
| GP7 | African capitals with a population of two million inhabitants or more | 1 P |
| GP8 | Suspension bridges in Brazil | |
| GP9 | Composers of Renaissance music born in Germany | |
| GP10 | Polynesian islands with more than 5,000 inhabitants | |
| GP11 | Which plays of Shakespeare take place in an Italian setting? | |
| GP12 | Places where Goethe lived | |
| GP13 | Which navigable rivers in Afghanistan are longer than 1000 km? | |
| GP14 | Brazilian architects who designed buildings in Europe | 3 P |
| GP15 | French bridges which were in construction between 1980 and 1990 | |

GikiP is...

- Easy to extend to other languages
- Easy to organize (provided one chooses topics known to have few answers)
- Easy to play with
 - New evaluation measures
 - New requests
- Useful for a wide number of users out there, especially if the systems invest in the presentation of their results
- Related to several other CLEF tracks: ImageCLEF (WikipediaMM), QA@CLEF, WebCLEF, iCLEF (and obviously descends from WiQA)
- **Let us hold GikiP once more in 2009!** (U Lisbon, Wolverhampton, DCU, SINTEF)