GATE
general architecture
for text engineering

YAHOO!
Search REST API

WIKIPEDIA
The Free Encyclopedia

DYNAMIC LINK SERVICE 2.0
USING WIKIPEDIA AS A LINKBASE

Using a Web 2.0-style mashup approach, reusing technologies and services freely available on the web, we have developed a dynamic link service system that uses Wikipedia as its linkbase.

Hungarian-born nuclear physicist Edward Teller (1908 - 2003) sits at his desk as director of the University of California's Lawrence Radiation Laboratory in Livermore, California, 1950s. (Photo by Pictorial Parade/ Getty Images)

© Getty Images

Kaiser Wilhelm

Wilhelm

Kaiser Wilhelm II

Wilhem II

William II of Germany and Prussia

Friedrich Wilhelm Viktor Albert

© Getty Images

In our work on integrating heterogeneous cultural heritage multimedia collections, we discovered that the majority of the valuable metadata is stored in an unstructured form such as captions.

This raises challenges in integrating the different collections, for example if people use ambiguous naming schemes, and limits the effectiveness of searching and browsing facilities.

There are also external sources of rich information that would be useful to integrate with the collections, such as Wikipedia.

So we developed a system to find the correct Wikipedia article for peoples' names extracted from captions.

We use GATE, the General Architecture for Text Engineering, for performing the information extraction. For example:

Hungarian-born nuclear physicist Edward Teller (1908 - 2003) sits at his desk as director of the University of California's Lawrence Radiation Laboratory in Livermore, California, 1950s. (Photo by Pictorial Parade/Getty Images)
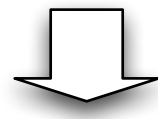
```xml
<?xml version="1.0" encoding="UTF-8" ?>
<gate xmlns:gate='http://gate.ac.uk/schema'>
Hungarian-born nuclear physicist <Person gate:gateId="98"
rule="PersonFinal" rule1="PersonLocAmbig">Edward Teller</Person>
(1908 - 2003) sits at his desk as director of the University of
California&apos;s Lawrence Radiation Laboratory in Livermore,
California, 1950s. (Photo by Pictorial Parade/Getty Images)
</gate>
```

A Yahoo Search is performed on the person's name, but restricted to the Wikipedia domain
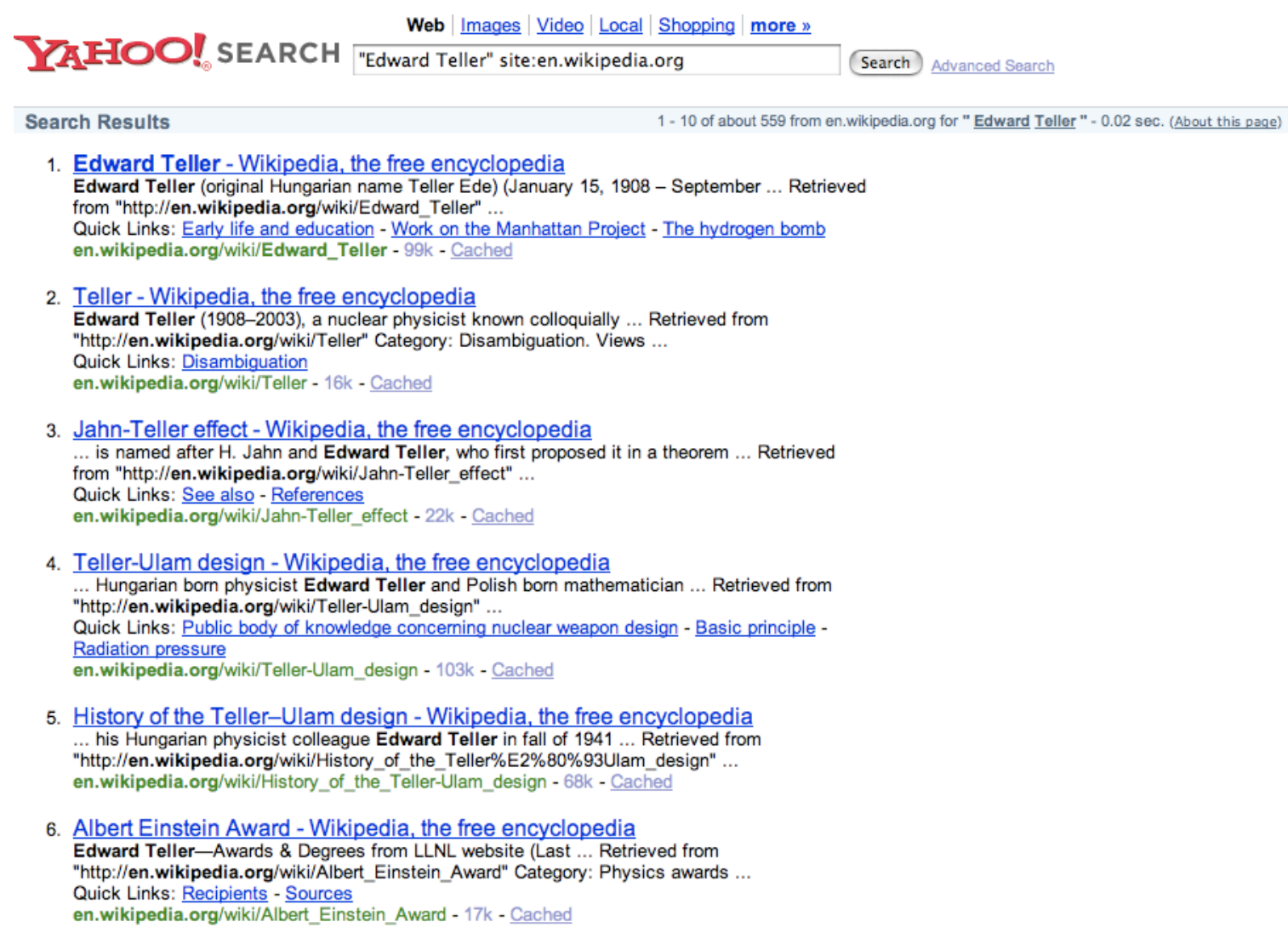
We then do some further checks on the result to ensure we have the correct article.

Categories: Nuclear physicists | American physicists | Hungarian physicists | Manhattan Project people | National Medal of Science recipients | University of California, Berkeley faculty | George Washington University faculty | Presidential Medal of Freedom recipients | Enrico Fermi Award recipients | Ig Nobel Prize winners | Erdős number 4 | Naturalized citizens of the United States | Hungarian Jews | Jewish American scientists | American amputees | 1908 births | 2003 deaths
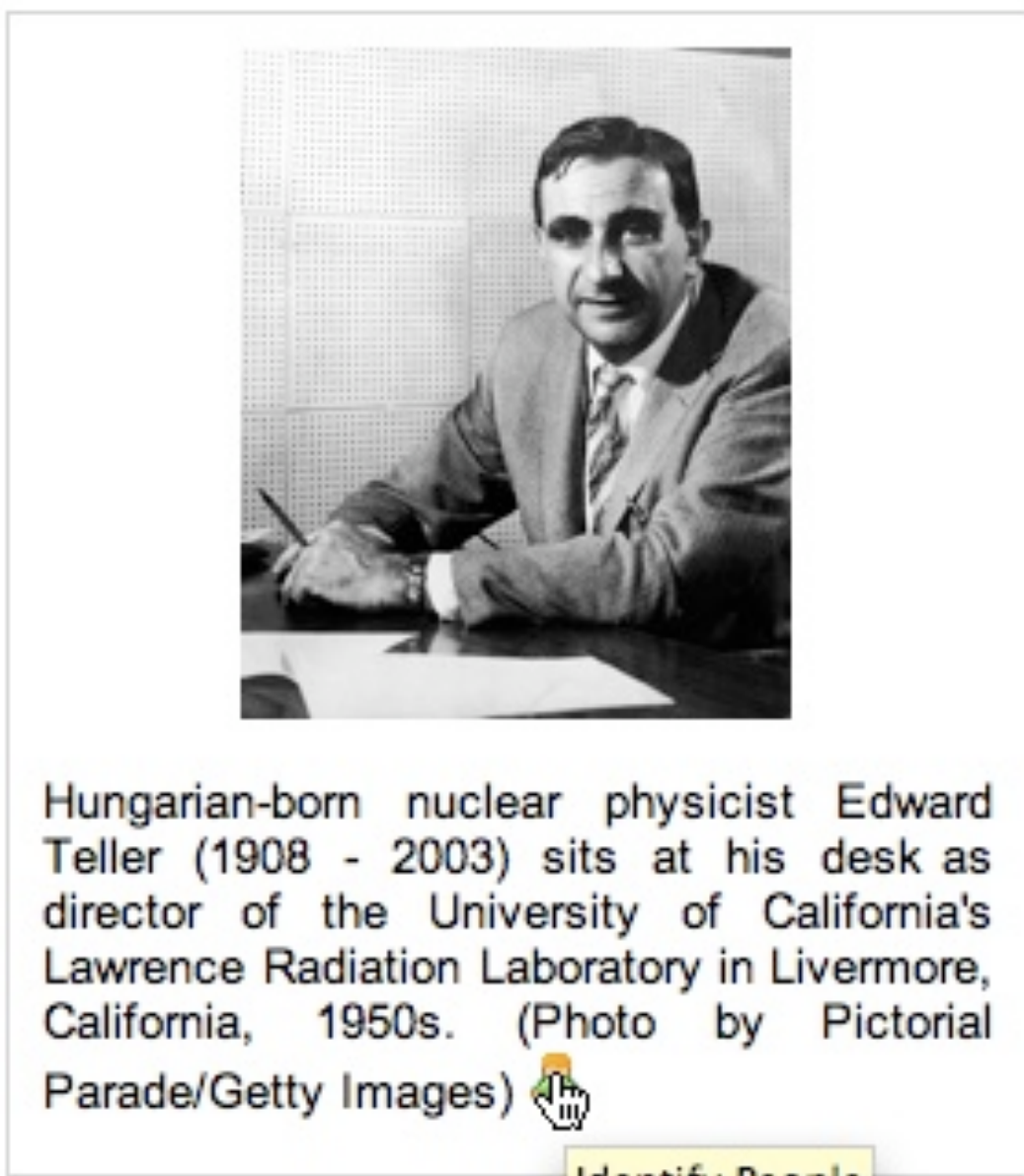
Categories indicate if the article describes a person.

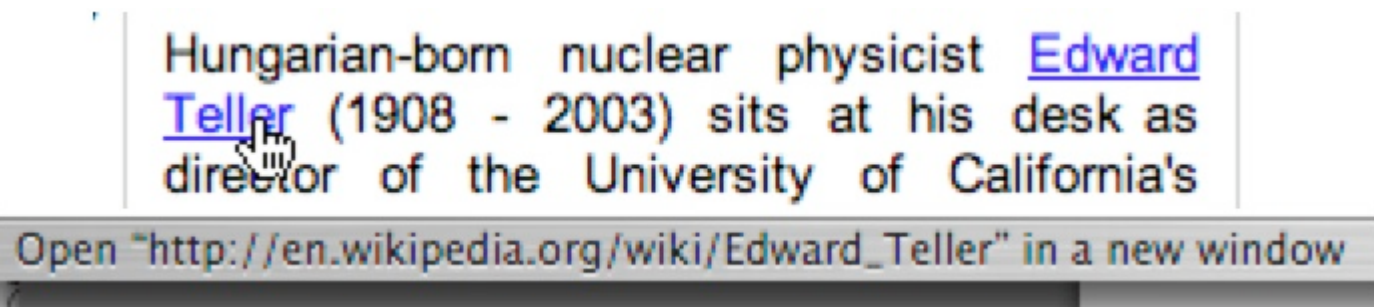Article redirects provide variations on the spelling of a person's name.
For example, Wilhelm II ⇒ Kaiser Wilhelm II etc.

The system is deployed as a RESTful web service that accepts snippets of text, and returns the text with dynamically injected links to the respective Wikipedia article for any reference to a person.

This enables AJAX-style link injection: when the user clicks on the "Identify People" icon, the link injection process is started.

Hungarian-born nuclear physicist Edward Teller (1908 - 2003) sits at his desk as director of the University of California's Lawrence Radiation Laboratory in Livermore, California, 1950s. (Photo by Pictorial Parade/Getty Images)
Identify People

The text caption is submitted to the system which returns the text with a dynamically inserted link to the Wikipedia article on Edward Teller.

Hungarian-born nuclear physicist Edward Teller (1908 - 2003) sits at his desk as director of the University of California's
Open "http://en.wikipedia.org/wiki/Edward_Teller" in a new window

Using Greasemonkey, a Firefox plugin that makes on-the-fly changes to specific web pages, the system can be applied to any web site.
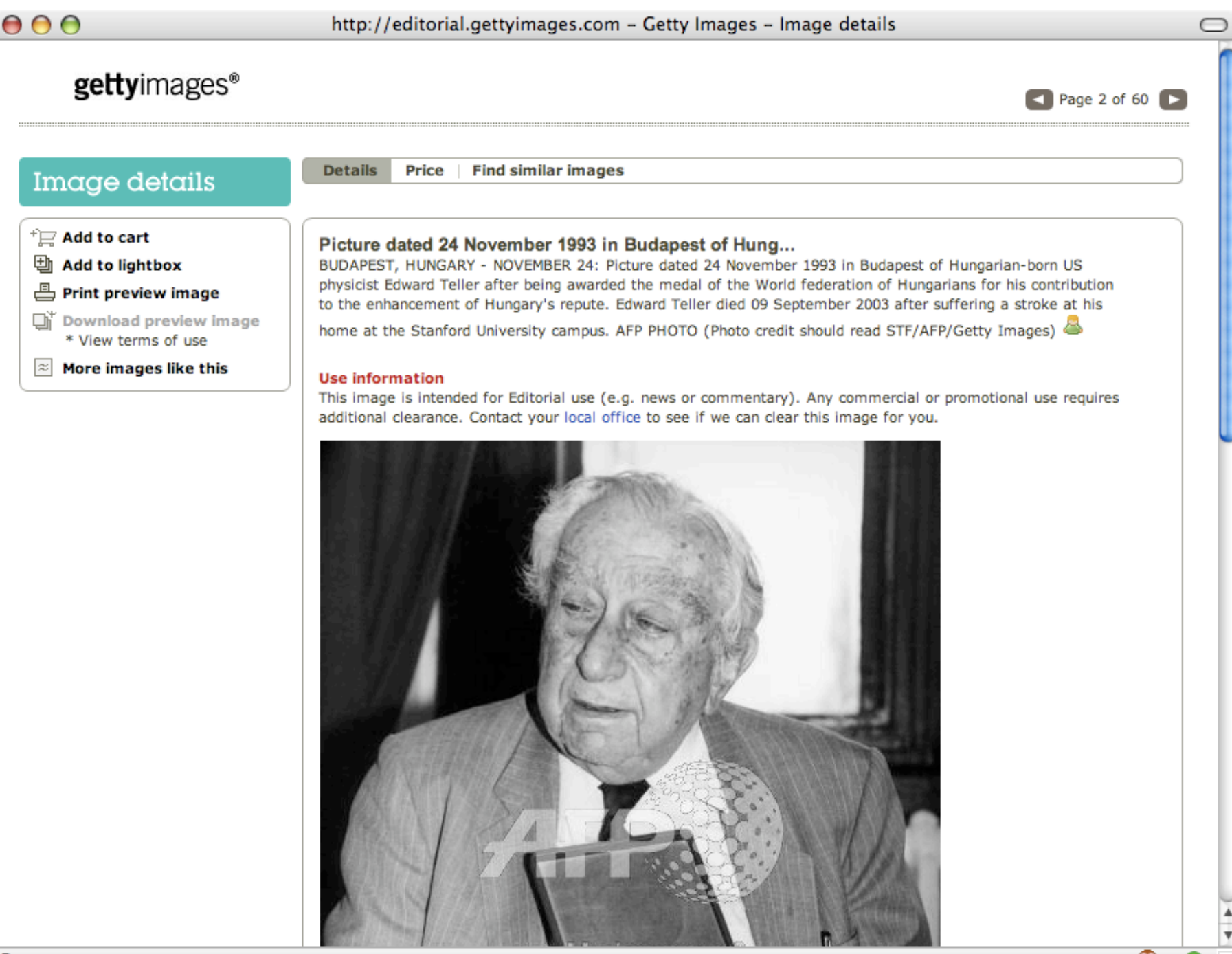
This allows users to invoke the service as they are browsing a picture archive web site, such as Getty Images, without modifying the existing website infrastructure.

This simple approach results in a dynamic link service that uses the whole of Wikipedia as its linkbase. Whilst most dynamic link service approaches might define a fixed linkbase, our system has a completely dynamic linkbase composed of the massive Wikipedia knowledge base.

The dynamic nature of Wikipedia provides some powerful features to the system. For example, if a user comes across a notable person who is not yet on Wikipedia they would be able to add a new article. Once Yahoo indexes the article, the system will be able to offer that link to other users.

The Web 2.0-style mashup approach we have taken to build this service, using resources such as Wikipedia and techniques such as the Yahoo REST API and Greasemonkey, has allowed us to revisit open hypertext ideas, such as dynamic link services, and deploy them on the web.

For more information, visit:
http://multimedia.ecs.soton.ac.uk/semanticintegration/wikipedia-identifier/

P. Sinclair, P. Lewis and K. Martinez
Electronics and Computer Science,
University of Southampton,
SO17 1BJ, United Kingdom
pass,phl,km@ecs.soton.ac.uk

ECS
Electronics & Computer Science
University of Southampton