

Developing a Simple Online Barcode Search Application

Hamid Nikbakht, Lee Zamparo, Gregory Singer and Donal Hickey
Department of Biology, Concordia University

INTRODUCTION

DNA barcoding is a taxonomic method, which uses a short genetic marker in an organism's mitochondrial DNA to quickly and easily identify it as belonging to a particular species. It is based on a relatively simple concept: most eukaryote cells contain mitochondria and mitochondrial DNA (mtDNA) has a relatively fast mutation rate, which results in significant variance in mtDNA sequences between species and a comparatively small variance within species. A 648-bp region of the mitochondrial gene, known as cytochrome c oxidase I (COI), was initially proposed as a potential 'barcode'.

A search engine for searching these records is needed to help scientists to define the name of a species or a sample they find a nature. In this study we developed a search engine using Google mini appliance technology.

METHODS

We used the Google desktop search application for searching the barcode sequences. We wrote an indexing plugin that indexes our files containing sequences of barcodes in FASTA format. The interface sends queries those are broken up to the words of certain length to Google desktop application. Google desktop searches upon its index and finds the matches between indexed words and given query word and give matches back as results. It's the same with what happened when Google desktop searches your PC but in this case you force it to index some certain file types. Our plugin recognizes FASTA, MEGA, and PHYLIP files.

For ensuring that the query sequence is broken into the same word frames as the sequence in the database, we perform queries from all possible

frames and take the one with the most hits as correct. The result hits contains links to images, bibliographic information and taxonomic data.

As a further step, we are using Google mini box appliance for performing search within records those have been kept on our server as separate HTML web pages containing some hyperlinks embedded using some tags such as gi number or organism name pointing to images of the species, bibliographic information, GenBank entries, Taxonomy database and even entries in Wikipedia.

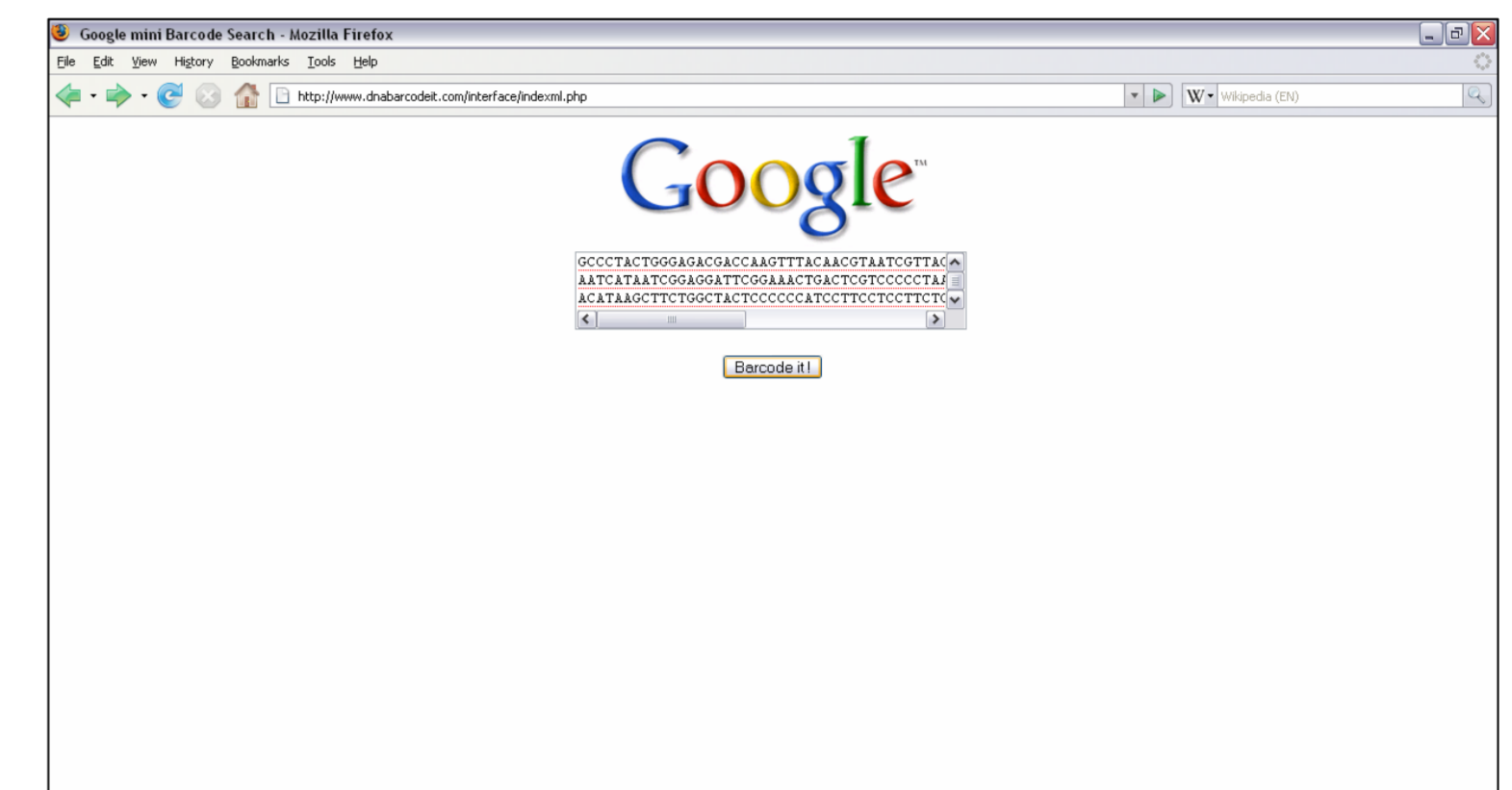
Then we forced Google mini to crawl the HTML repository and make an index of all words included in HTML files. Since, Google can only search words with length less than 256 characters and since some point variations in sequences could be happened during the sampling and should be ignored during the search, we broke down all sequences in words of 15.

We developed a preprocessor interface to get the query and convert all possible formats to a single string and then break it down to words of 15 in all 15 possible word frames. The intermediate interface does this breaking down process. The interface sends all 15 possible queries to Google appliance and gets back the results for each query in XML format. Each of these XML items will be modeled as a DOM object and then using DOM library functions these objects will be parsed and needed tags will be extracted and given back to the user as a HTML file.

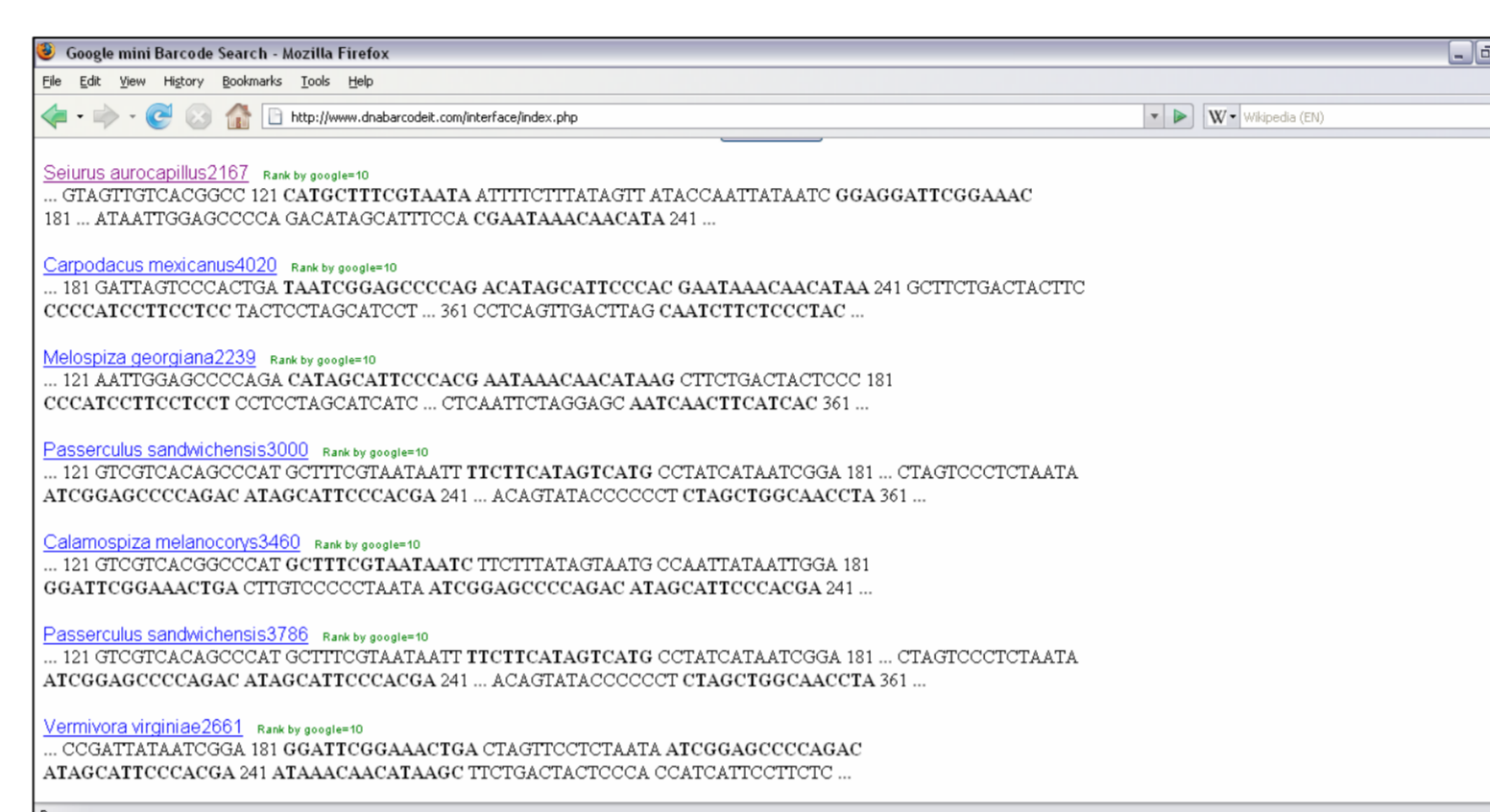
To show the results, we perform a real search using this interface and show the result page.



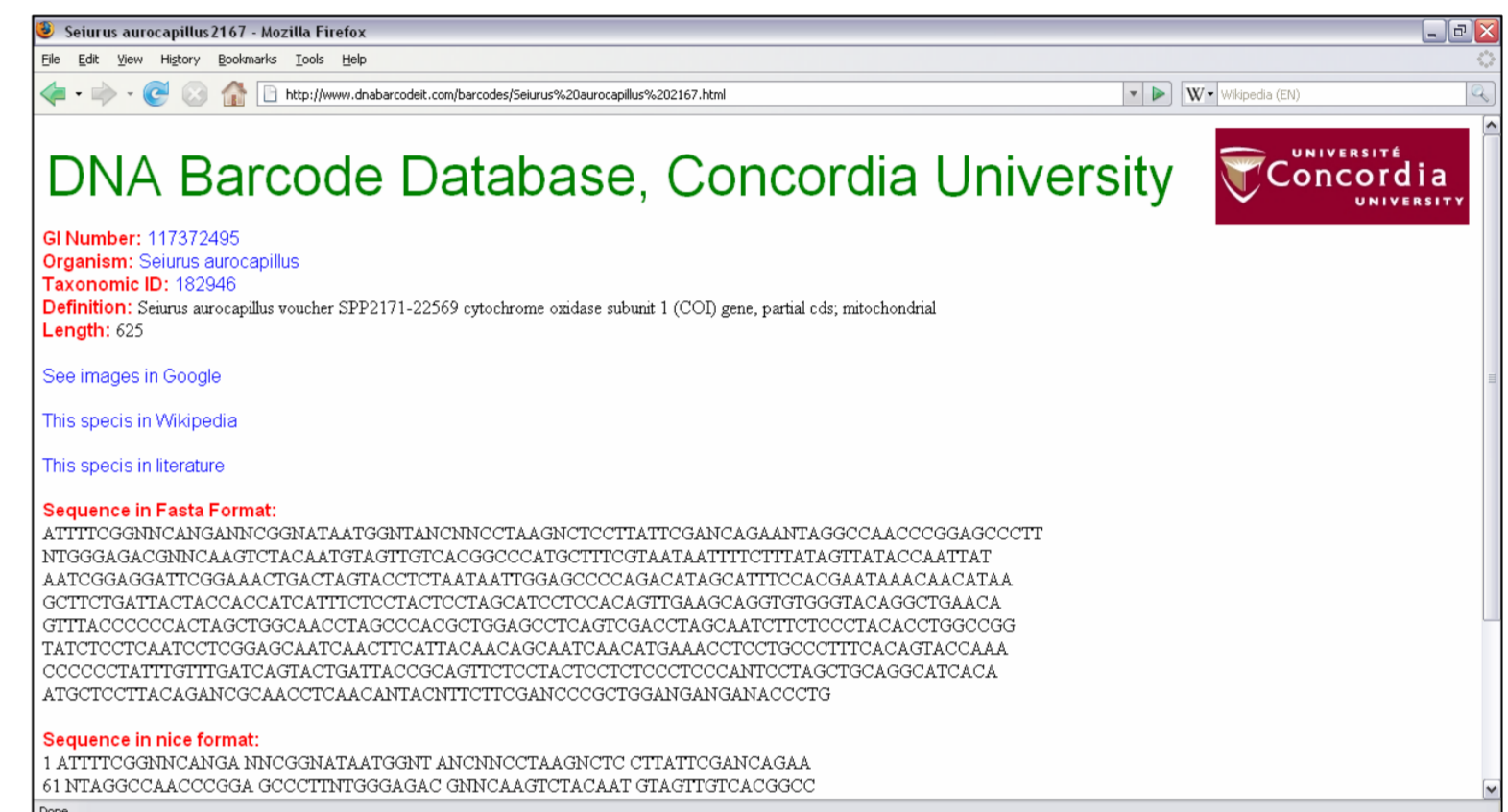
GoogleGene form interface using Google Desktop Search (GDS)



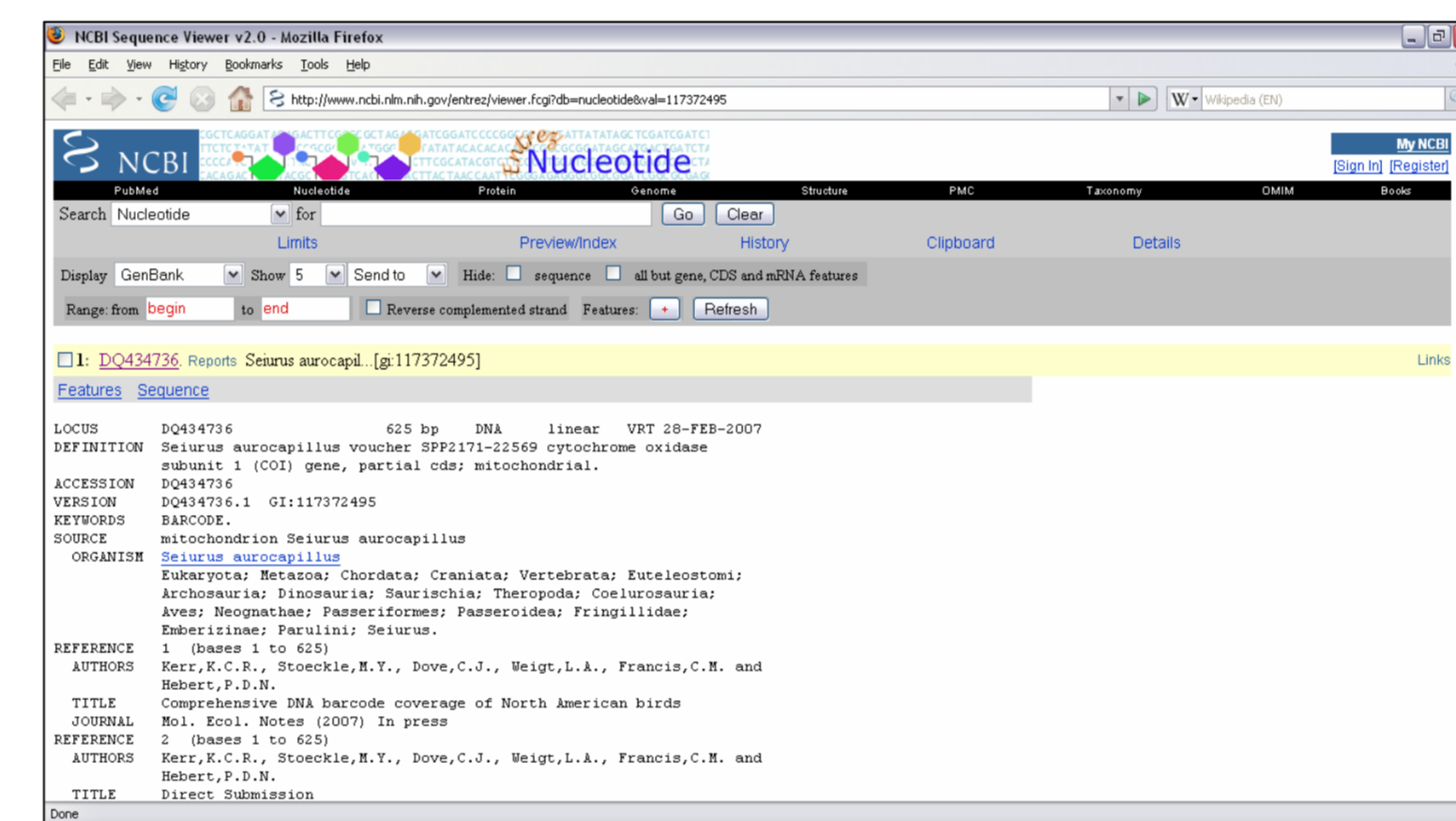
DNA Barcode Search form interface using Google mini



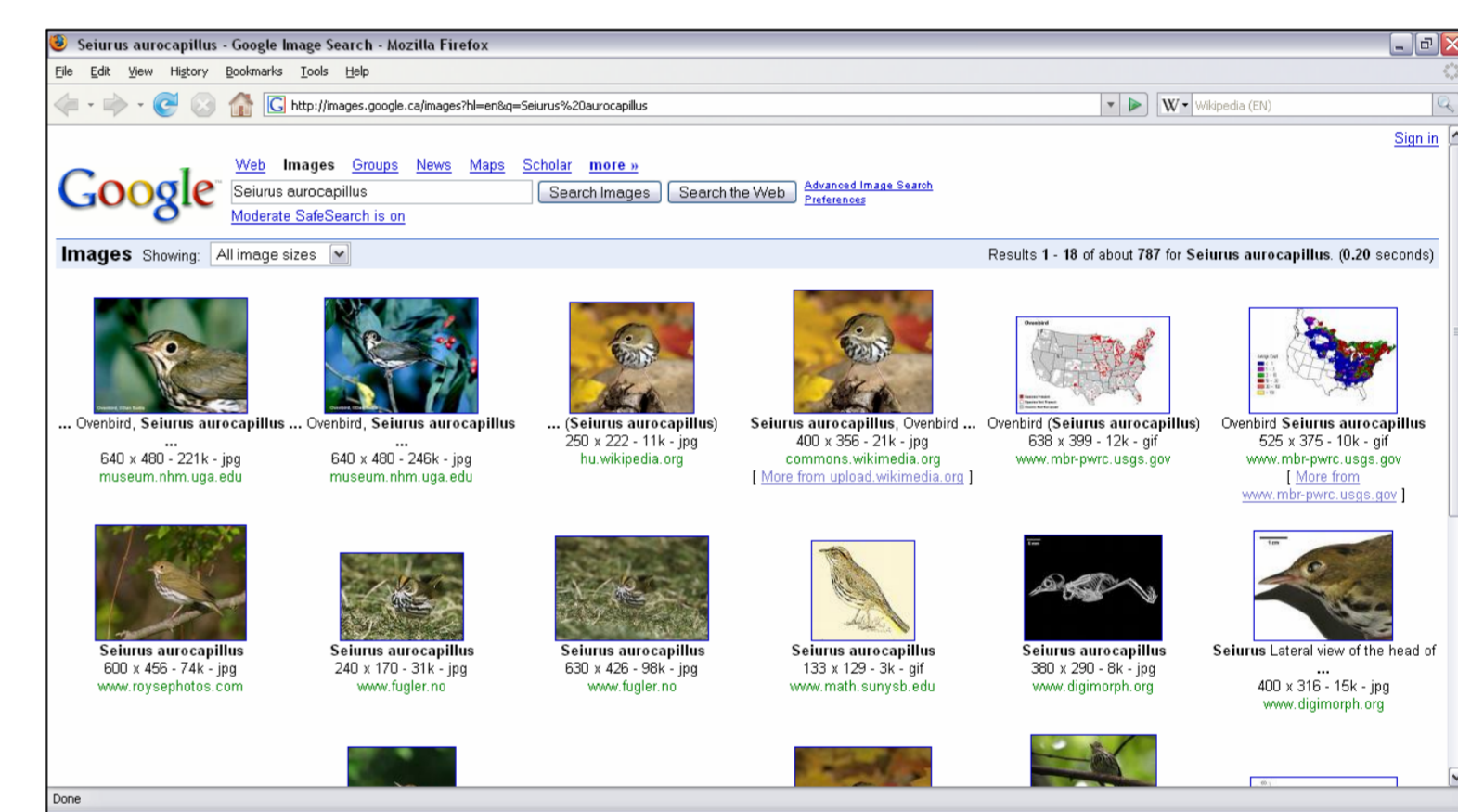
List of result hits



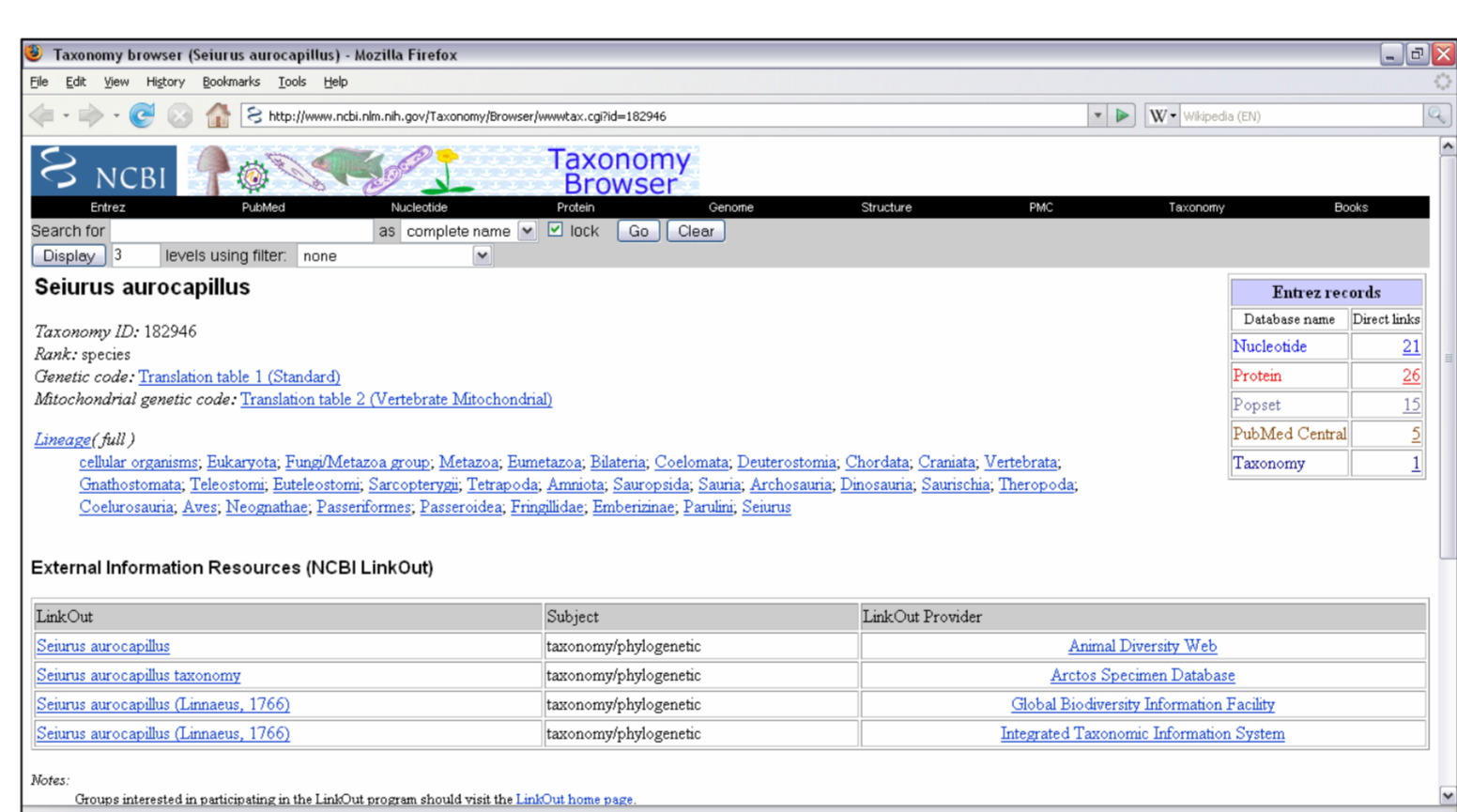
HTML page related to one of the result species, containing several links to GenBank, Taxonomy Browser, Google image, Wikipedia and Google Scholar



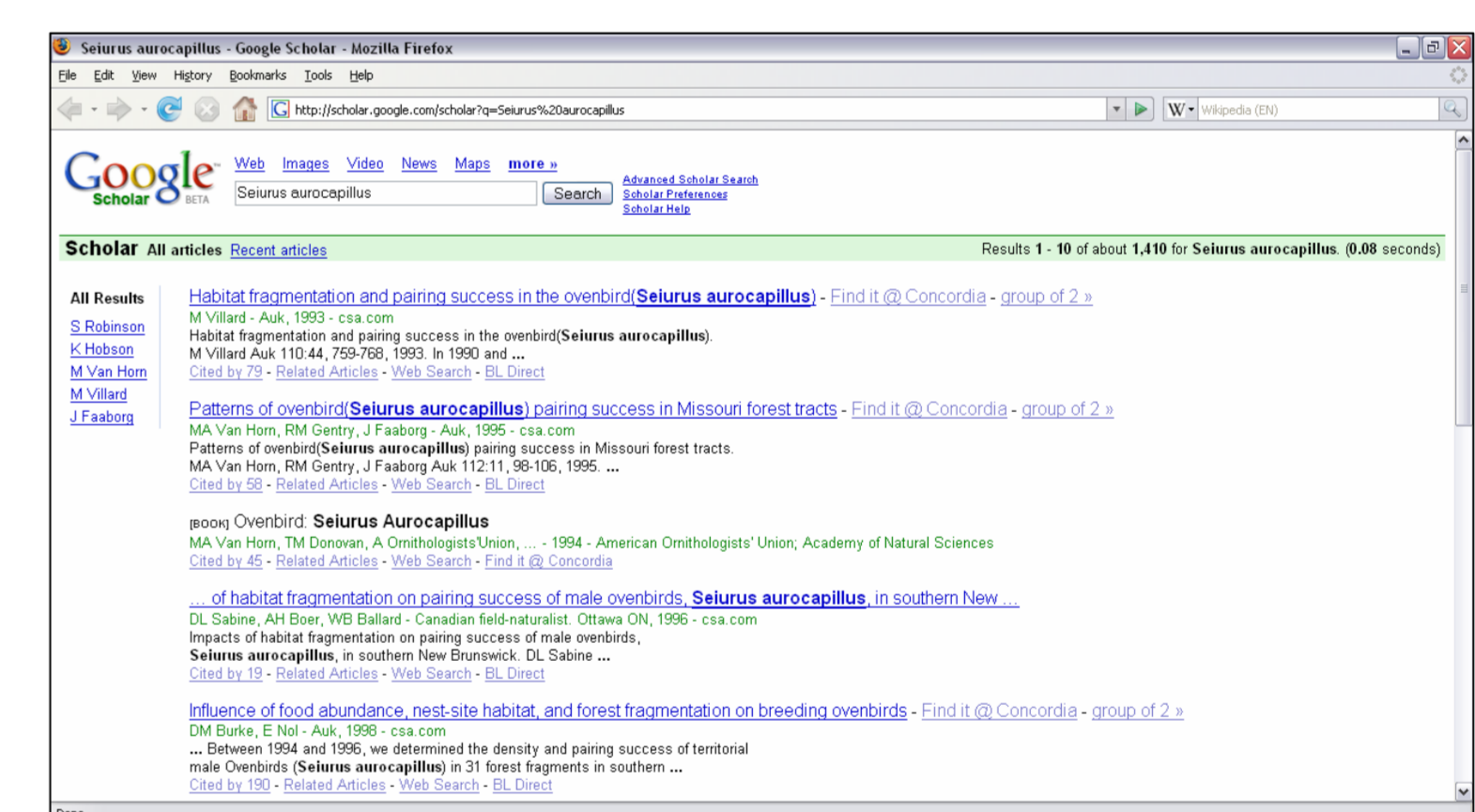
GenBank Entry related to the result species barcode sequence



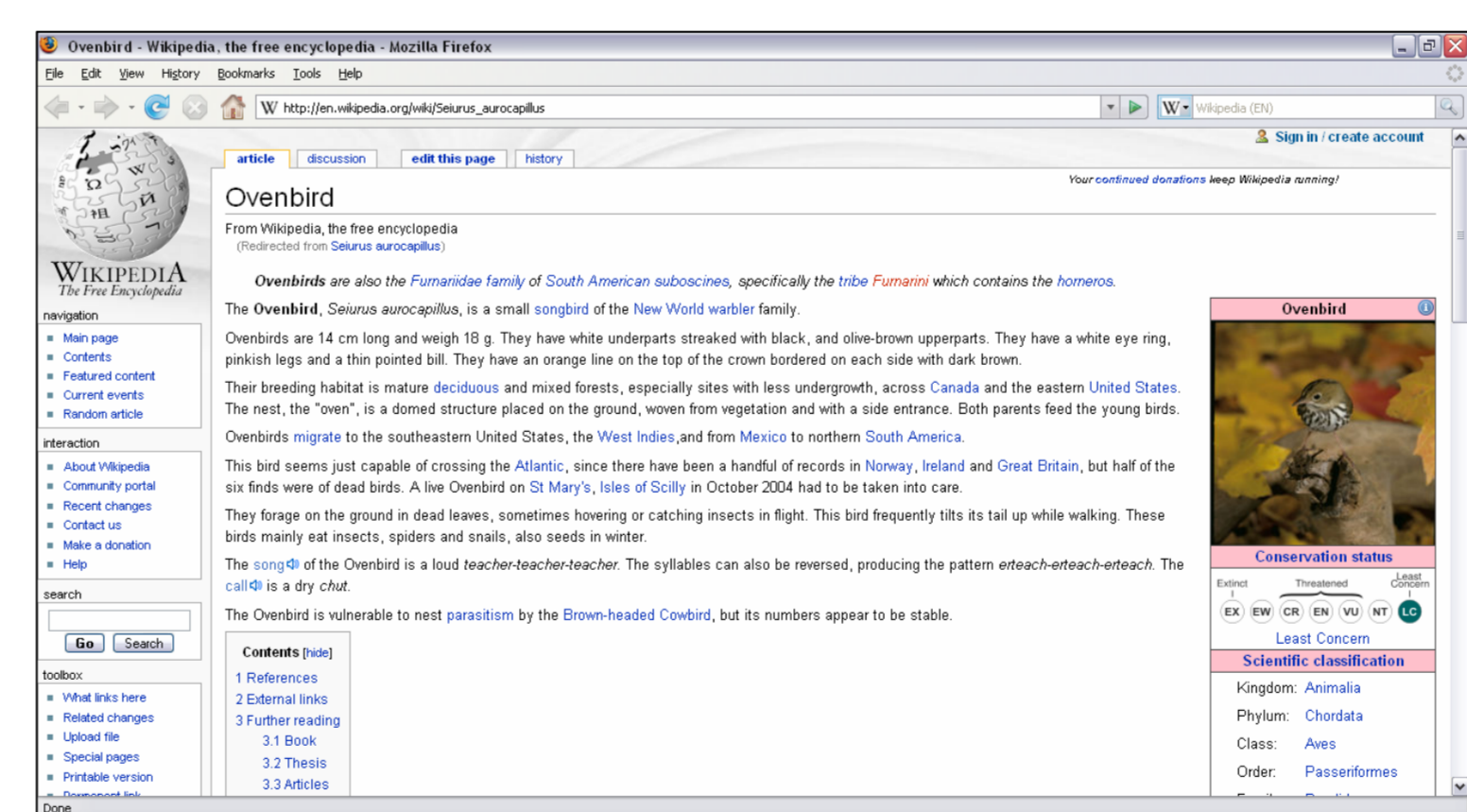
Images related to the result species via Google image search



Taxonomy browser entry related to the result species



Literature entries related to the result species via Google Scholar



Wikipedia entry related to the result species

CONCLUSIONS

Google is the most popular search engine around the world and now everyone knows how to perform a search via Google. We tried to seize this benefit of popularity of Google to make it accessible and easy to use for everyone to search upon a barcode sequence and find useful information about the species.

It is expected, all groups those are working on DNA Barcoding projects in different species, submit their data on a single global repository to be available for all users around the world. There are several records in GenBank those keyword filed are tagged with "Barcode" word. These data comprise of barcode records gathered in GenBank from various groups working on various species. At this moment, there are only around 11,000 records in GenBank, whereas more than hundreds of thousands of Barcode records have been resolved as different barcoding centers. More records in this kind of repositories will guaranty more accurate searches performed.

For more information see:

<http://www.dnabarcodeit.com> for GDS version

<http://www.dnabarcodeit.com/interface> for Google mini version