IBM Research

Web-a-Where: Geotagging Web Content

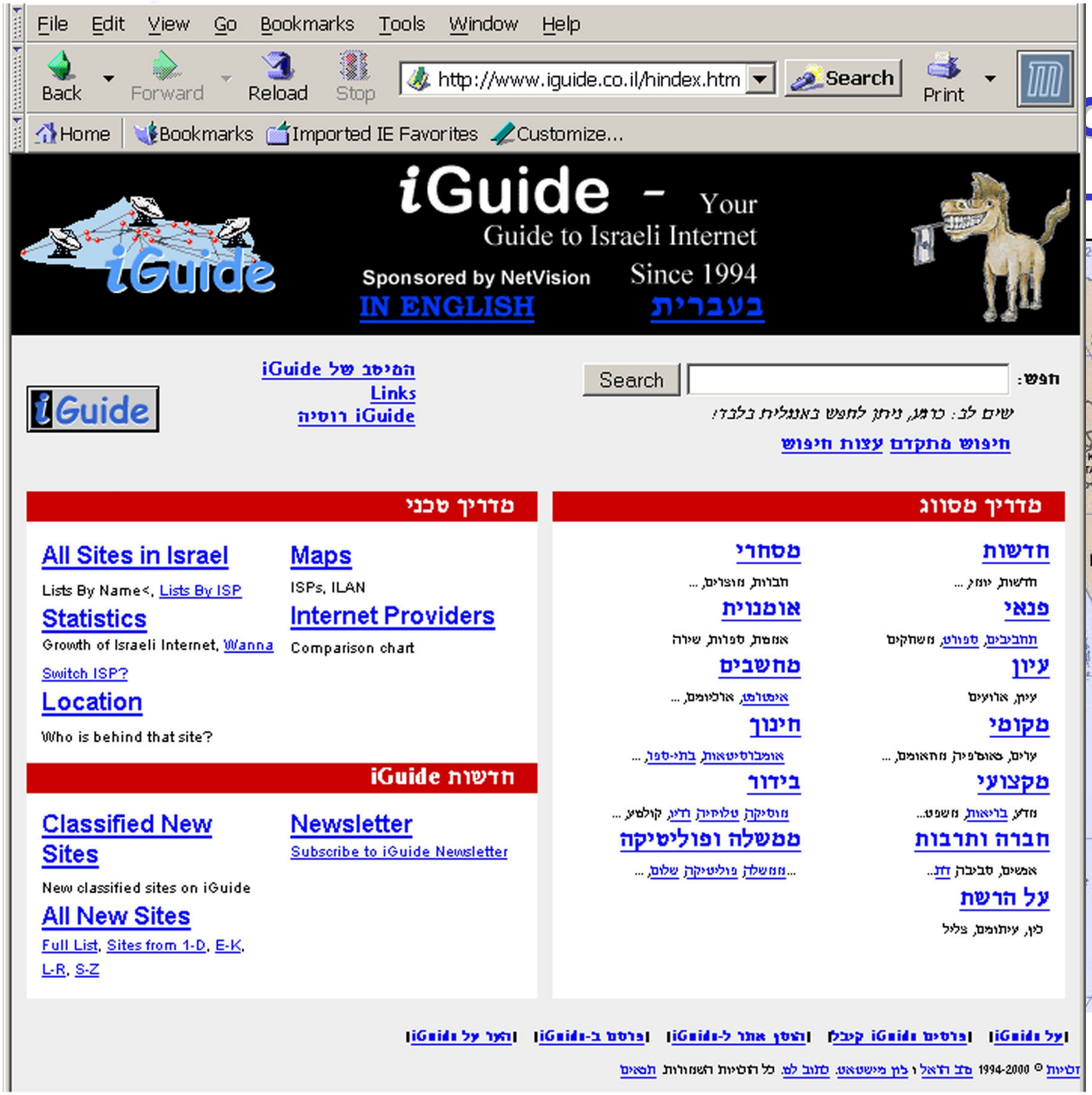
Ronny Lempel

Einat Amitay
Nadav Har'El
Ron Sivan
Aya Soffer

Feb 11, 2004

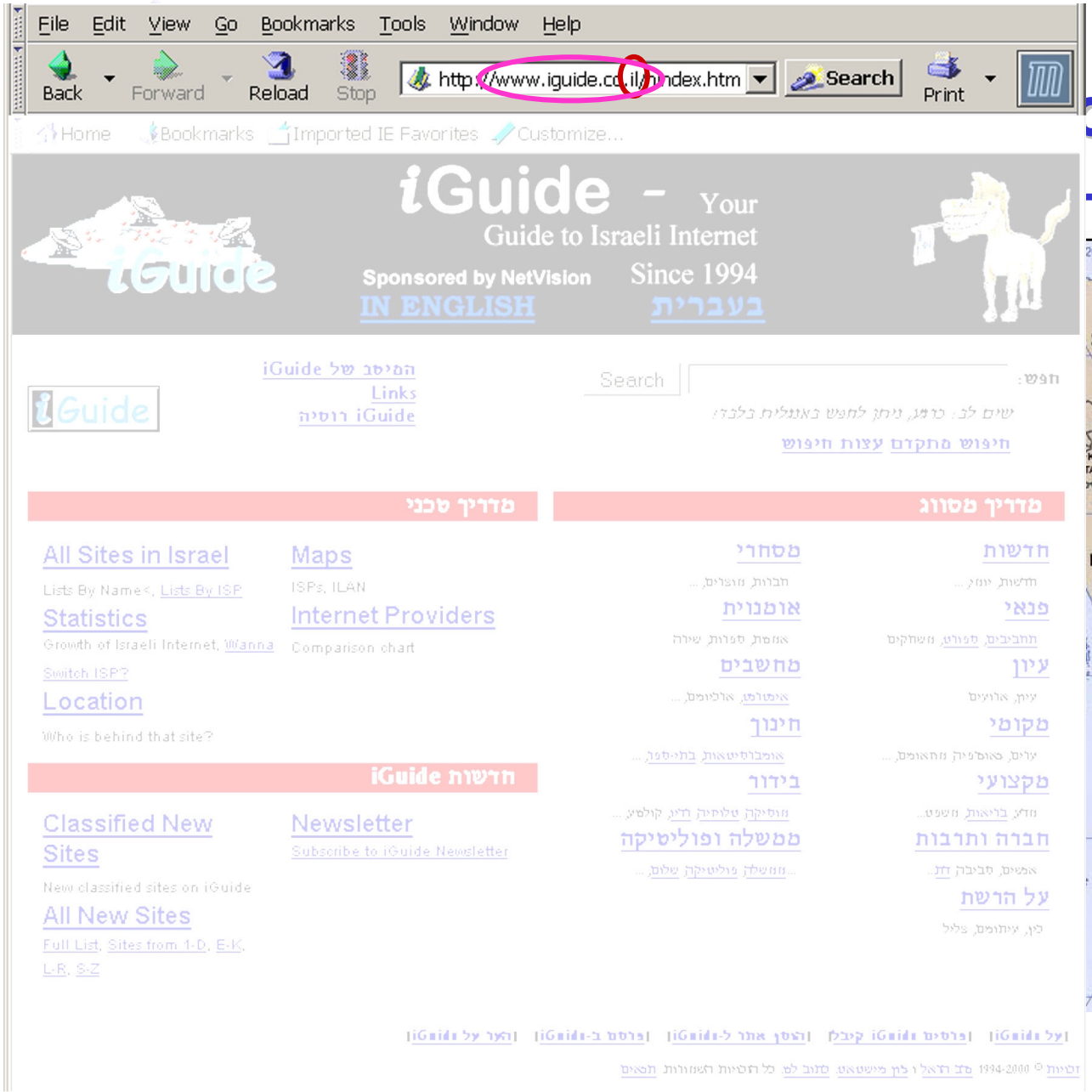
What is the geography of a page?

of a page?


File Edit View Go Bookmarks Tools Window Help
Back Forward Reload Stop
http://www.iguide.co.il/hindex.htm
Search
Print
Home Bookmarks Imported IE Favorites Customize...
iGuide - Your Guide to Israeli Internet
Sponsored by NetVision
Since 1994
IN ENGLISH
בעברית
iGuide
iGuide המיסב של
Links
iGuide רוסיה
Search
חפש:
חיפוש מתקדם עצות חיפוש
מדריך מסווג
מדריך טכני
All Sites in Israel
Lists By Name<, Lists By ISP
Statistics
Growth of Israeli Internet, Wanna Switch ISP?
Location
Who is behind that site?
Maps
ISPs, ILAN
Internet Providers
Comparison chart
חדשות iGuide
Classified New Sites
New classified sites on iGuide
All New Sites
Full List, Sites from 1-D, E-K, L-R, S-Z
Newsletter
Subscribe to iGuide Newsletter
חדשות
פנאי
עיון
מקומי
מקצועי
חברה ותרבות
על הרשת
מסחרי
אומנות
מחשבים
חינוך
בידור
ממשלה ופוליטיקה
על iGuide
פרסם ב-iGuide
הצע אתר ל-iGuide
הער על iGuide
1994-2000 © זכויות

ARCTIC OCEAN
MONGOLIA
KYRGYZSTAN
TAJIKISTAN
CHINA
N. KOREA
S. KOREA
JAPAN
NORTH PACIFIC OCEAN
NEPAL
BHU
BANGL
INDIA
BURMA
LAOS
Taiwan
THAILAND
VIETNAM
CAMBODIA
PHILIPPINES
SRI LANKA
BRUNEI
MALAYSIA
SINGAPORE
Equator
INDONESIA
PAPUA NEW GUINEA
SOLOMON ISLANDS
INDIAN OCEAN
AUSTRALIA
NEW ZEALAND

# of a page?

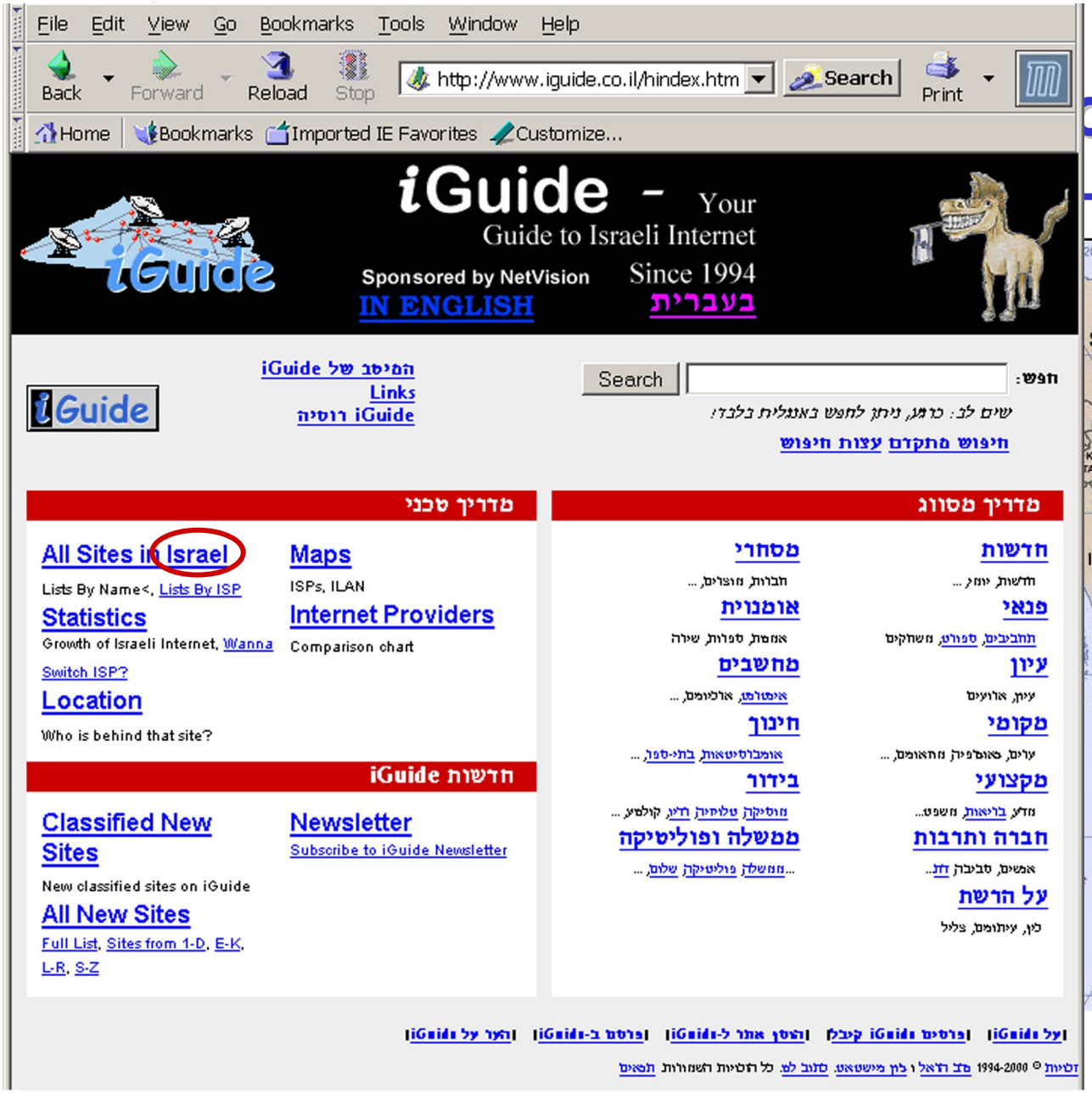

File Edit View Go Bookmarks Tools Window Help
Back Forward Reload Stop
http://www.iguide.co.il/index.htm
Search
Print
Home Bookmarks Imported IE Favorites Customize...
iGuide - Your Guide to Israeli Internet
Sponsored by NetVision Since 1994
IN ENGLISH בעברית
iGuide
iGuide הפיצ'ר של
Links
iGuide רוסיה
Search
חפש:
חיפוש מתקדם עצות חיפוש
מדריך מסווג
מדריך טכני
חדשות
מסחרי
פנאי
אומנות
עיון
מחשבים
מקומי
חינוך
מקצועי
בידור
חברה ותרבות
ממשלה ופוליטיקה
על הרשת
All Sites in Israel
Lists By Names, Lists By ISP
Maps
ISPs, ILAN
Statistics
Growth of Israeli Internet, Wanna Switch ISP?
Internet Providers
Comparison chart
Location
Who is behind that site?
חדשות iGuide
Classified New Sites
New classified sites on iGuide
Newsletter
Subscribe to iGuide Newsletter
All New Sites
Full List, Sites from 1-D, E-K, L-R, S-Z

ARCTIC OCEAN
MONGOLIA
KYRGYZSTAN
TAJIKISTAN
CHINA
N. KOREA
S. KOREA
JAPAN
NORTH PACIFIC OCEAN
NEPAL
BHU
BANGL
INDIA
BURMA
LAOS
Taiwan
THAILAND
VIETNAM
CAMBODIA
PHILIPPINES
SRI LANKA
BRUNEI
MALAYSIA
SINGAPORE
Equator
INDONESIA
PAPUA NEW GUINEA
SOLOMON ISLANDS
INDIAN OCEAN
AUSTRALIA
NEW ZEALAND

What is the geography of a page?

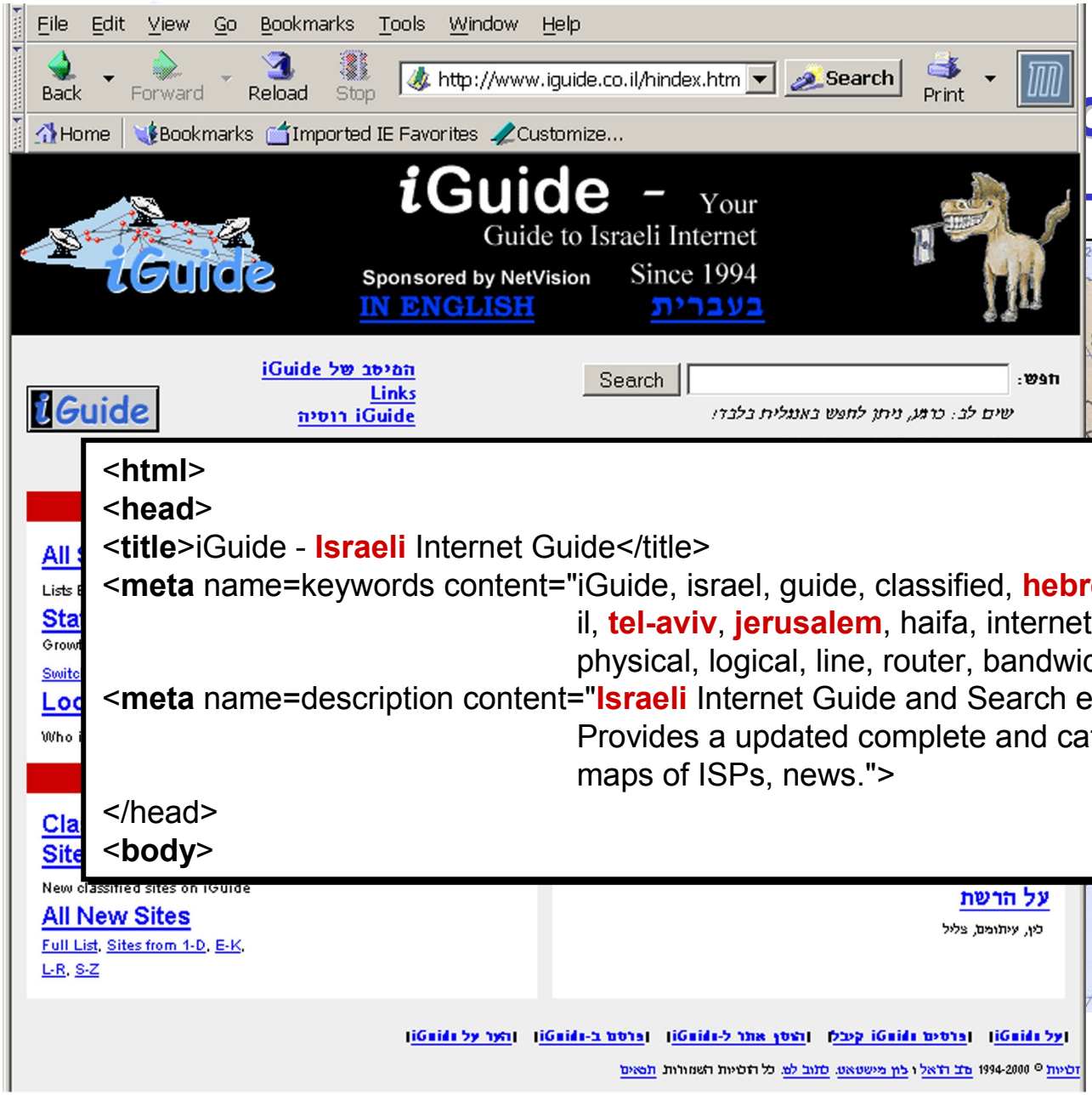
- **Environment:**
  - URL, domain
  - Owner
  - Server locale, IP address

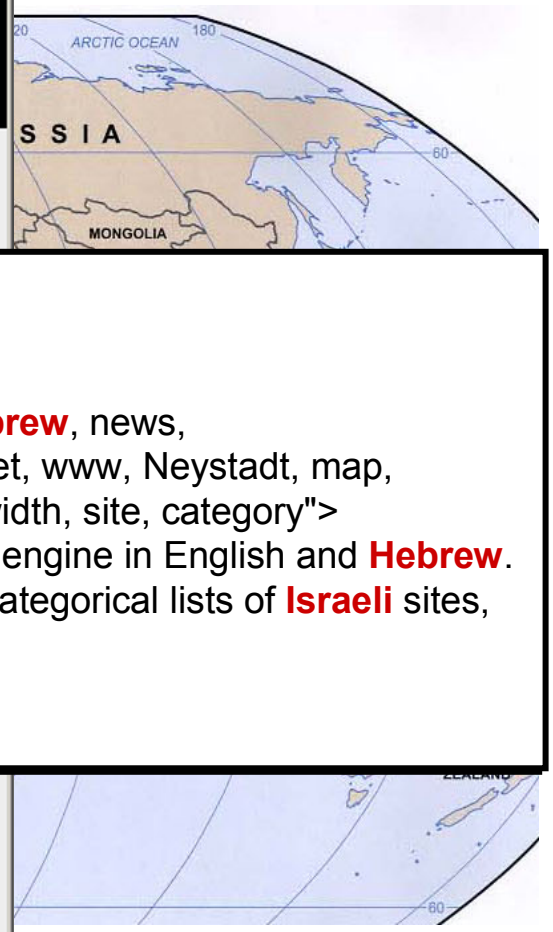
# of a page?


File Edit View Go Bookmarks Tools Window Help
Back Forward Reload Stop
http://www.iguide.co.il/hindex.htm
Search
Print
Home Bookmarks Imported IE Favorites Customize...
iGuide - Your Guide to Israeli Internet
Sponsored by NetVision
Since 1994
IN ENGLISH
בעברית
iGuide
iGuide המיסב של
Links
iGuide רוסיה
Search
חפש:
חיפוש מתקדם עצות חיפוש
מדריך מסווג
מדריך טכני
All Sites in Israel
Lists By Name, Lists By ISP
Maps
ISPs, ILAN
Statistics
Growth of Israeli Internet, Wanna Switch ISP?
Internet Providers
Comparison chart
Location
Who is behind that site?
חדשות
פנאי
עיון
מקומי
מקצועי
חברה ותרבות
על הרשת
מסחרי
אומנות
מחשבים
חינוך
בידור
ממשלה ופוליטיקה
חדשות iGuide
Classified New Sites
New classified sites on iGuide
All New Sites
Full List, Sites from 1-D, E-K, L-R, S-Z
Newsletter
Subscribe to iGuide Newsletter

ARCTIC OCEAN
MONGOLIA
KYRGYZSTAN
TAJIKISTAN
CHINA
N. KOREA
S. KOREA
JAPAN
NORTH PACIFIC OCEAN
NEPAL
BHU
BANGL
INDIA
BURMA
LAOS
Taiwan
THAILAND
VIETNAM
CAMBODIA
PHILIPPINES
SRI LANKA
BRUNEI
MALAYSIA
SINGAPORE
Equator
INDONESIA
PAPUA NEW GUINEA
SOLOMON ISLANDS
INDIAN OCEAN
AUSTRALIA
NEW ZEALAND

# of a page?






```
<html>
<head>
<title>iGuide - Israeli Internet Guide</title>
<meta name=keywords content="iGuide, israel, guide, classified, hebrew, news,
                             il, tel-aviv, jerusalem, haifa, internet, www, Neystadt, map,
                             physical, logical, line, router, bandwidth, site, category">
<meta name=description content="Israeli Internet Guide and Search engine in English and Hebrew.
                             Provides a updated complete and categorical lists of Israeli sites,
                             maps of ISPs, news.">
</head>
<body>
```

What is the geography of a page?

- **Environment: external to page**
  - URL, domain
  - Owner
  - Server locale, IP address
- **Content: text and metatext**
  - Place names
  - Adjectives (*Israeli*), language (*Hebrew*), currency (*NIS*)
  - Link analysis

What is the geography of a page?

- **Environment: external to page**
  - URL, domain
  - Owner
  - Server locale, IP address
- **Content: text and metatext**
  - Place names
  - Adjectives (*Israeli*), language (*Hebrew*), currency (*NIS*)
  - Link analysis

The process

- **Spotting**
  - Takes names to spot from the gazetteer
- **Disambiguate individual instances**
  - Takes taxonomies from the gazetteer
- **Find page foci**

The problem

- **Find names of places**
  - Must distinguish from non-geo
    - Is “London” a place or a person? (e.g., Jack London)
- **Determine the meaning**
  - Must arbitrate between competing senses
    - Is “London” in England or Ontario, Canada?
- **Finding the focus**
  - must combine geotags meaningfully

The gazetteer

- Co
  -
  -
- So
  -
  -
  -
- XM
  -

```
<geographic-location level="city" label="San Jose">
        <population>816884</population>
        <coordinates>
                <latitude>37.33944</latitude>
                <longitude>-121.89389</longitude>
        </coordinates>
        <names-to-spot>
                <name>San Jose</name>
        </names-to-spot>
        <geographic-locations>
        </geographic-locations>
 </geographic-location>
```

Disambiguation

- **A bag of tricks**
  - Qualified names
  - Default name sense
  - Single sense per discourse
  - Names with strong non-geo sense
  - Unifying locality

Qualified names

- **Names qualified by their next higher taxonomy level**
  - Haifa, Israel
  - San Jose, California, USA
    - Definitely not the San Jose of Costa Rica, Guatemala, Bolivia, Peru or the Philippines
  - Sacramento Calif.
    - "Rumsfeld and Wolfowitz lobbied Clinton in 1998 to start Iraq war and topple Saddam." (Online Journal, Feb 20, 2003)

Default sense

- **The place having the highest population**
  - Portland, IN: 6,384
  - Portland, ME: 61,982
  - Portland, OR: 450,777
  - Portland, TN: 5,708
  - Portland, TX: 13,642
  - Portland, Victoria, Australia: 10,455

  - Portland, OR

Si

- **Single s**
  - A qu
    unqu

Alexandria Minnesota Fishing Guides

AGAPE FISHING GUIDES

AGAPE FISHING GUIDES is a multi-species guide service, **specializing in bass fishing**. Guide service is available in and around the Alexandria. The Alexandria area offers outstanding multi-species fishing and absolutely some of the finest bass fishing in the state. With the exception of panfish and walleye, all AGAPE FISHING GUIDES fishing is catch and release. If you are in town on business or vacation and would like to experience some great bass fishing, give us a call.

Alexandria, Minnesota
(320) 808-3008
email: week@rea-alp.com

© Copyright 2002 Agape Fishing Guides & Fishing Minnesota. All Rights Reserved.
Site Marketing by Minnesota Fishing Guides

Single sense per discourse

- **Single sense per discourse**
  - A qualified instance induces its meaning on other, unqualified instances in the same context (page)
- **Can lead to difficulties**
  - As we shall soon see

Names w/strong non-geo sense

- **Some place names are also common words**
  - Mobile, Al or Reading, GB
  - Are ignored unless qualified
- **Some names are also *very very* common words**
  - To, Myanmar, Of, Turkey or As, Belgium

Institute of Certified Public
nts (AICPA) as being gen-
pted accounting principles

Historically, effective manageme
of the U.S. Government has bee
hampered by a lack of reliable finan
cial information. To help improve th
integrity of financial information, in
1990 the Office of Management and
Budget (OMB), the Department of
the Treasury (Treasury) and the Gen-
eral Accounting Office (GAO) estab-
lished the Federal Accounting Stan-
dards Advisory Board (FASAB) to
develop accounting standards for the
U.S. Government. The work of

will enhance the acceptability of our
reports and will add to the level of fi-
nancial professionalism throughout
the U.S. Government.

Working to issue agency financial
reports consistent with GAAP and to
obtain clean audit opinions, the Ad-
ministration is committed to improv-

panying Financial Re-
uired by 31 U.S.C.
l consists of the Man-
agement's Discussion and Analysis
(MD&A), Statement of Operations
and Changes in Net Position, State-
ment of Net Cost, Balance Sheet,
Stewardship Information, Notes to
the Financial Statements and Supple-
mental Information. Each section is
preceded by a description of its con-
tents.

As, Belgium

Basis of Accounting and Reporting Entity

Accounting Standards

"The recent recognition of Federal accounting standards by the AICPA as GAAP enhances their acceptability..."

The accompanying financial state-
ments generally were prepared
based on GAAP standards developed
by FASAB. The recent recognition
of Federal accounting standards by
the AICPA as GAAP enhances their
acceptability. These standards form
the foundation for preparing consis-
tent and meaningful financial state-
ments both for individual Federal
agencies and the Government as a
whole.

GAAP for the Federal Govern-
ment is tailored to the U.S. Govern-
ment's unique characteristics and
special needs. For example, land not
used in U.S. Government operations
(stewardship land), weapon systems
and support property used in the per-
formance of military missions, and
vessels held as part of the National
Defense Reserve Fleet (national de-

fense assets) are reported in the
Stewardship Information section
rather than valued on the Balance
Sheet. The Government's responsi-
bilities and policy commitments are
much broader than the reported Bal-
ance Sheet liabilities. They include
the social insurance programs dis-
closed in the Stewardship Informa-
tion section, as well as a wide range

Names w/strong non-geo sense

- **Some place names are also common words**
  - Mobile, AL or Reading, GB
  - Are ignored unless qualified
- **Some names are also *very very* common words**
  - To, Myanmar, Of, Turkey or As, Belgium
  - **Excluded altogether**

Names w/strong non-geo sense

- **Can be found from analyzing well-edited text**
  - Count capitalized and non-capitalized occurrences
  - For most unambiguous place names –
    - Frequency is related to importance (population)
    - Cap to non-cap ratio is high
  - Names with strong non-geo sense stick out

Names w/strong non-geo sense

- **Currently, a manual pass is required, however**
  - False positive: Aspen, CO
    - Appears frequently
    - Has only 5,000
    - Therefore included in the list
    - But most occurrences actually refer to it…
  - False negative: Metro, Indonesia
    - Appears frequently
    - Has 100,000 inhabitants
    - Therefore excluded from the list
    - But most occurrences of “Metro” refer to Metro Area…

- Unify
  - A

Alexandria/Egypt/Africa
Alexandria/Italy/Europe
Alexandria/Romania/Europe
Alexandria/Ukraine/Europe
Alexandria/Indiana/US/North America
Alexandria/Kentucky/US/North America
Alexandria/Louisiana/US/North America
Alexandria/Minnesota/US/North America
Alexandria/Virginia/US/North America
Alexandria/Brazil/South America

Sluml, er, Landlording

In addition to my house in Vienna, I also own a townhouse in southern Alexandria. Now, why do I own it? Is it a vacation home? Hardly. I bought it when I worked for R&K Engineering, and had an aid me ca, It was to as too far.

Wien/Austria/Europe

ld place? Well, as it turned out, while I had decided that about a mile down the road

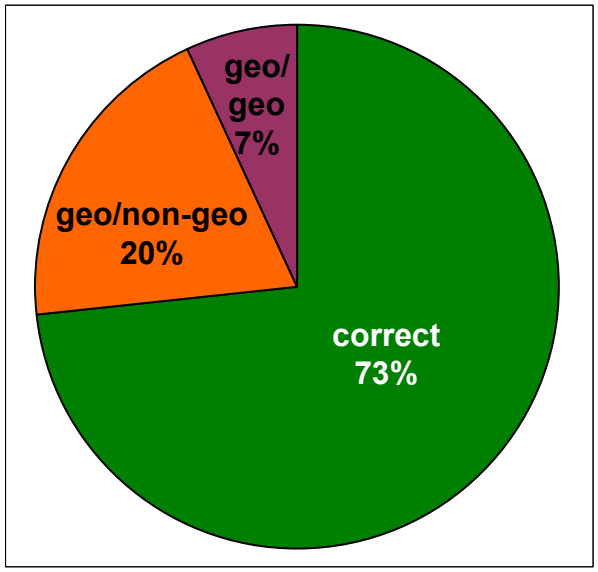
Vienna/Virginia/US/North America
Vienna/West Virginia/US/North America

t for such a place were still a few hundred However, what with tax deductibility of the mortgage interest and most other expenses, plus depreciation, it looked like pretty much a wash, *if* I could get it rented out (plus some of that money went into equity, so it was a net win).

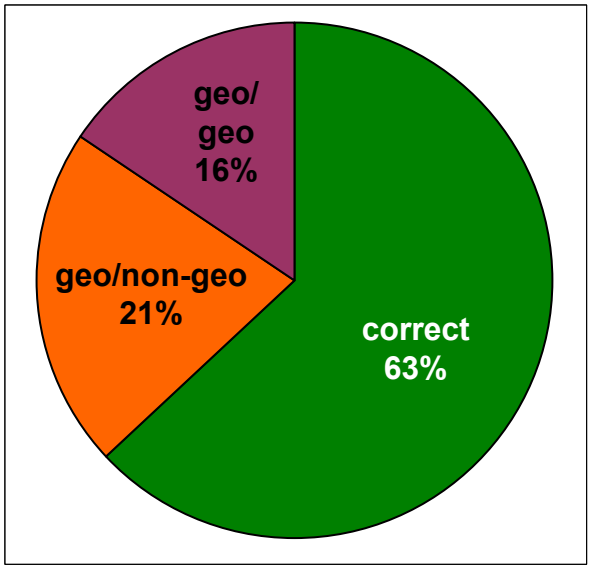
Individual spot disambiguation

- **Three collections of Web pages**
  - Gov: pages from *.gov (*1.2M)
  - ODP: the Regional tab of the ODP directory (600K)
  - Arb: What Google returned for “+the +and” (1.2K)
- **Randomly selected 200 pages from each**
- **Ran the GeoMiner**
  - Medium-sized gazetteer
- **Tags given to spots were checked manually**

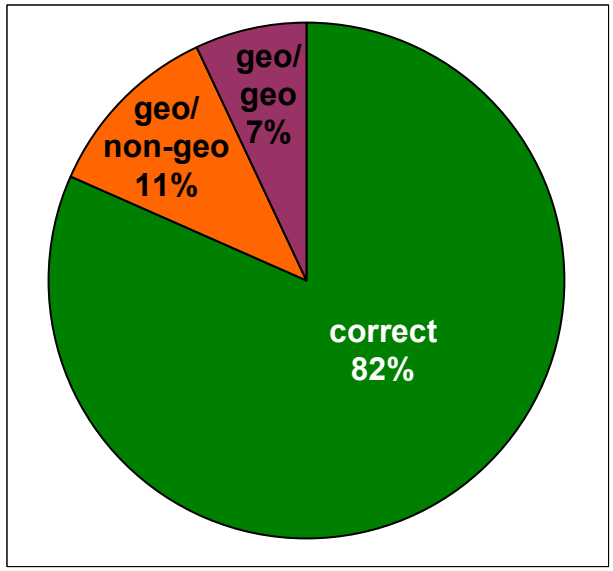
Results of individual spots



Gov



ODP



Arb

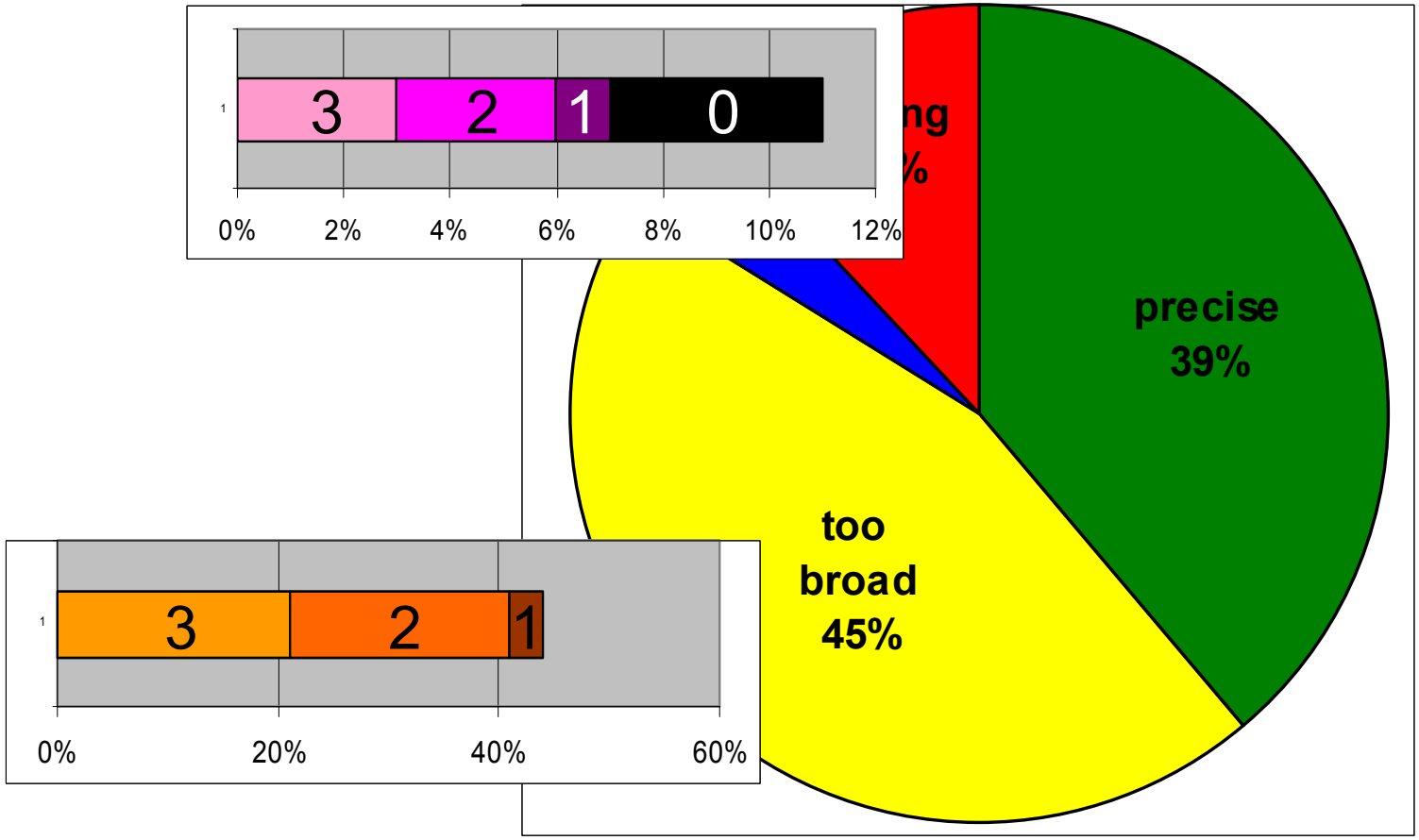
Page focus

- **Contributions from each disambiguated mention**
  - Depends on confidence
  - Contribution to each hierarchy level
- **Highest scoring levels selected as foci**
  - If they exceed a threshold
  - Are not part of or include another focus

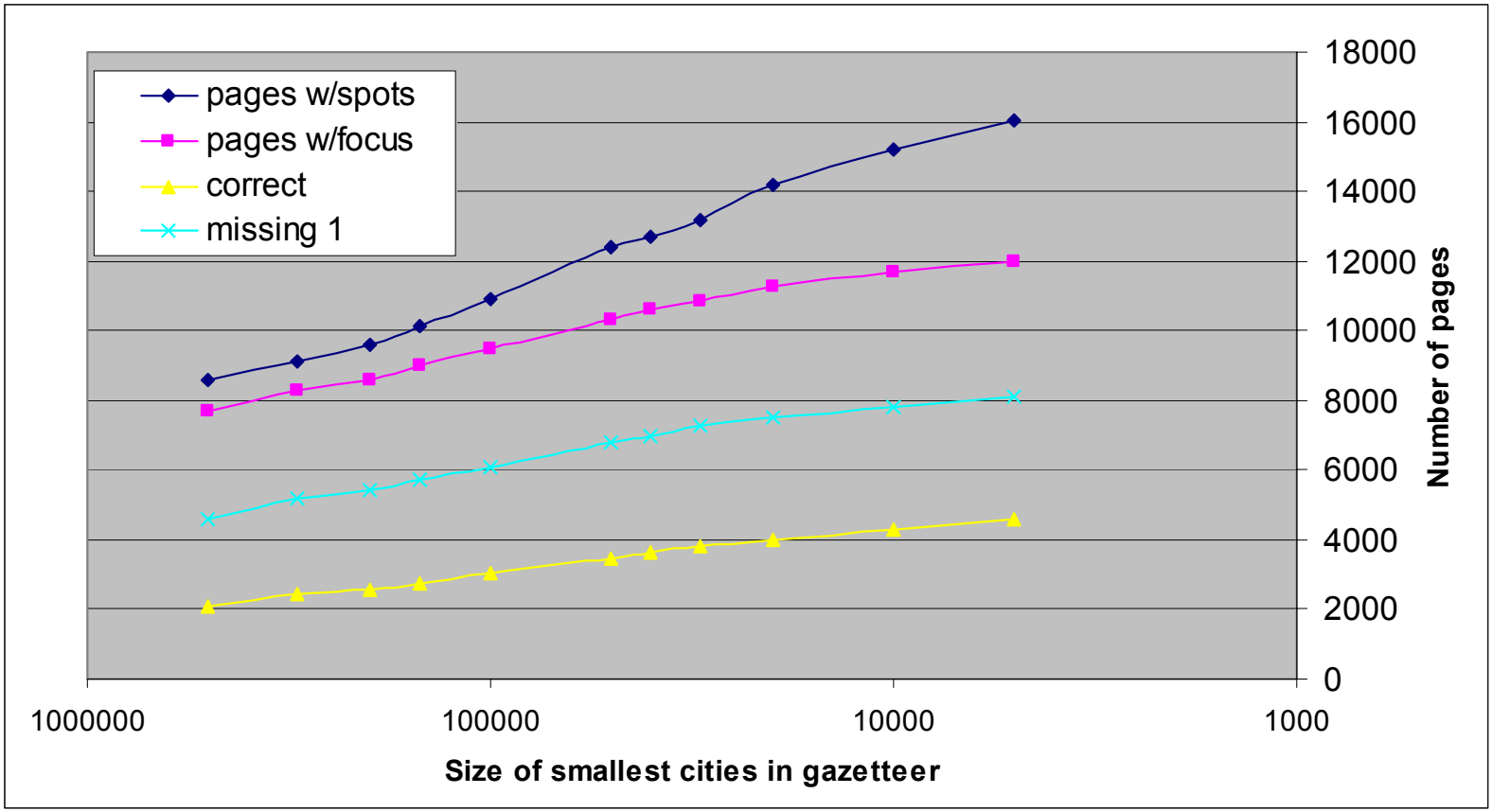
Results of focus accuracy

- **Focus given to 55% of the pages**
  - 75% of the pages having at least 1 geotag
- **Of the focused pages, 80% had but one focus**

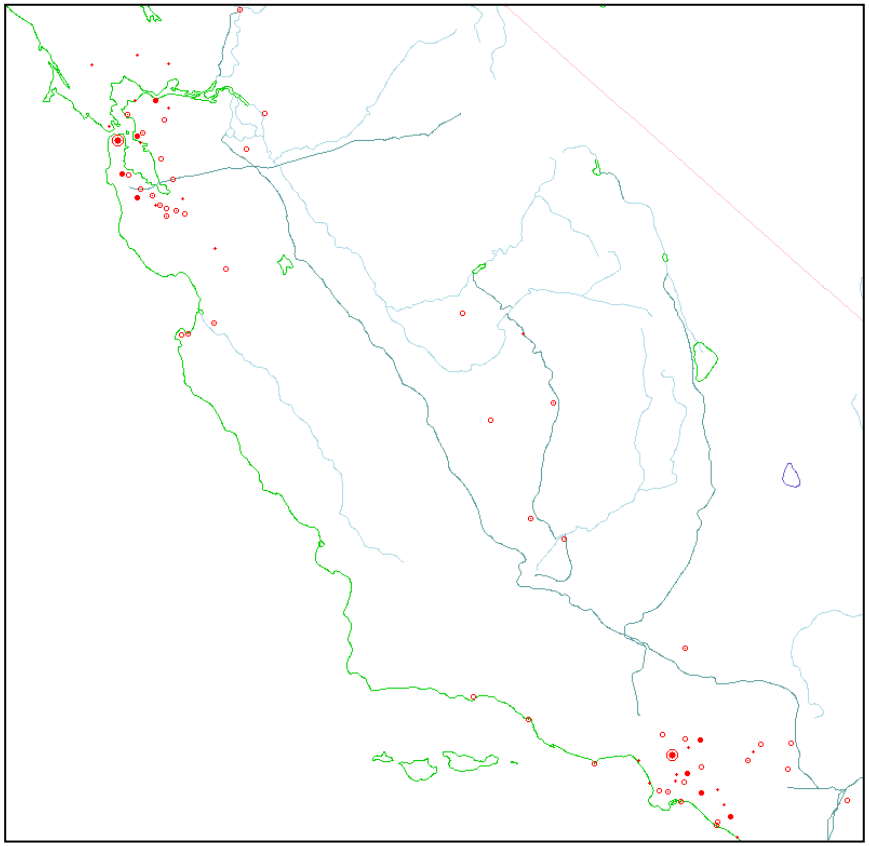
Results of focus accuracy



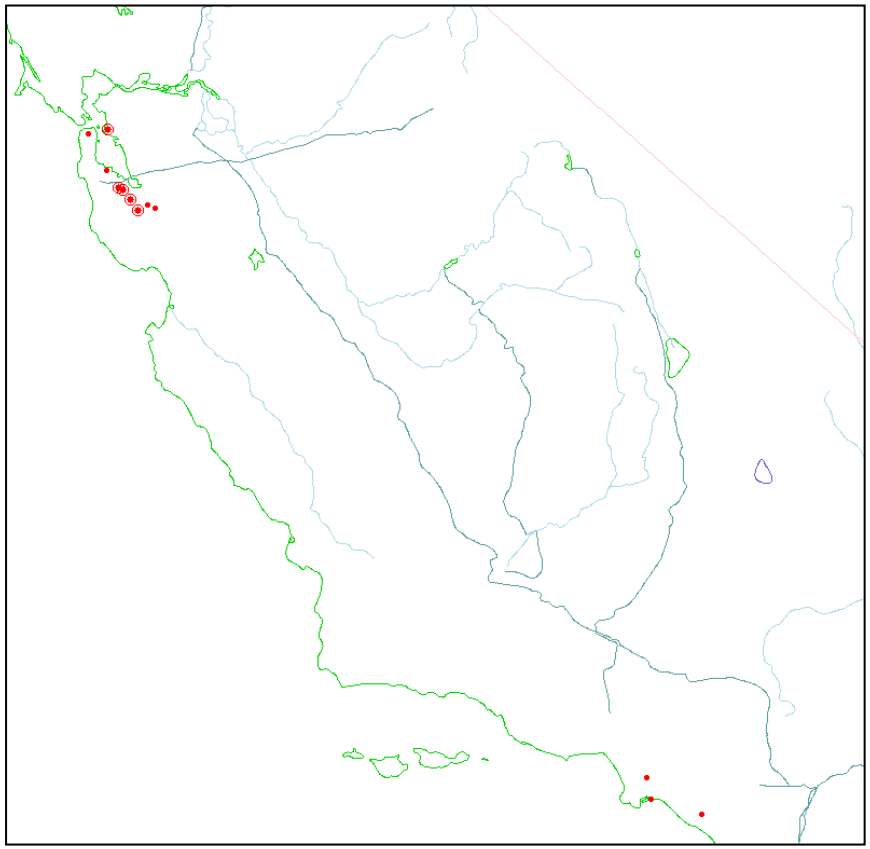
Precision vs. gazetteer size



Search results (Reuters corpus 1996)

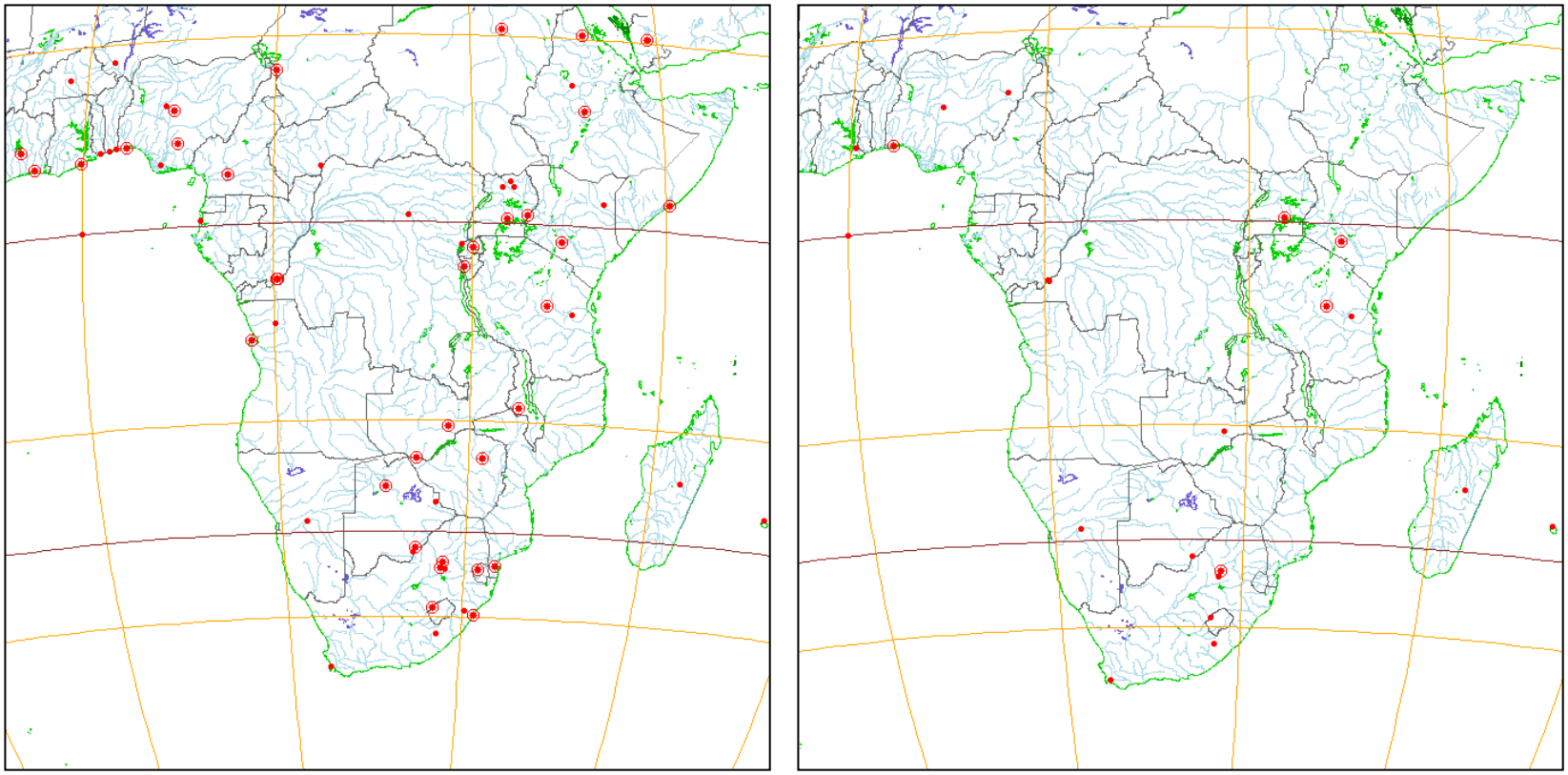


All



Hitech

Search results (pharmaceutical corpus)



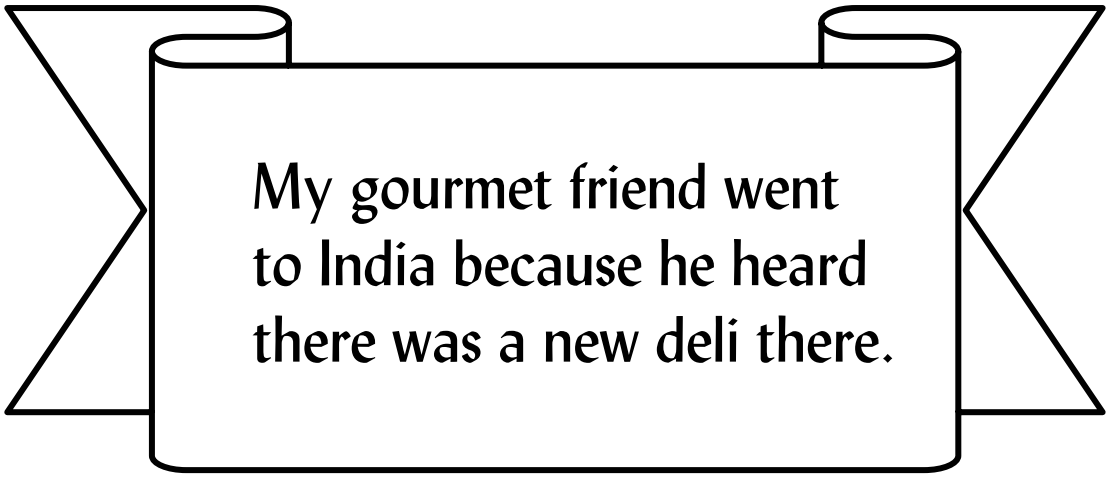
AIDS SARS

Future work

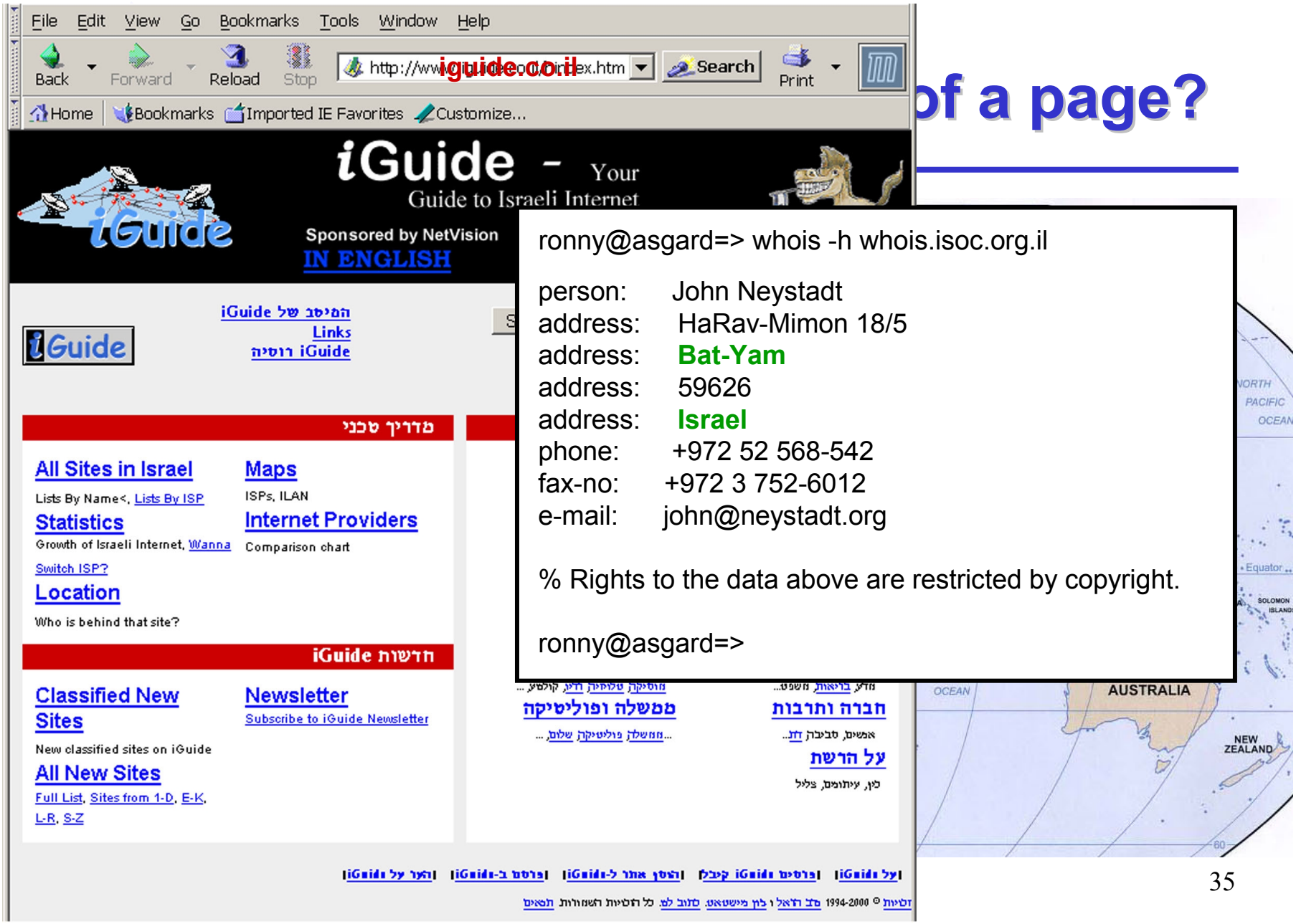
- **Reduce geo/non-geo error rate**
  - Ignore person names
- **Use coordinate data**
  - Newark in the context of Manhattan is in NJ, not NY
- **Expand the gazetteer**
  - Non-incorporated places
  - Counties, shires, cantons, etc.
  - Even smaller places
- **Grammatical text analysis**

Conclusions

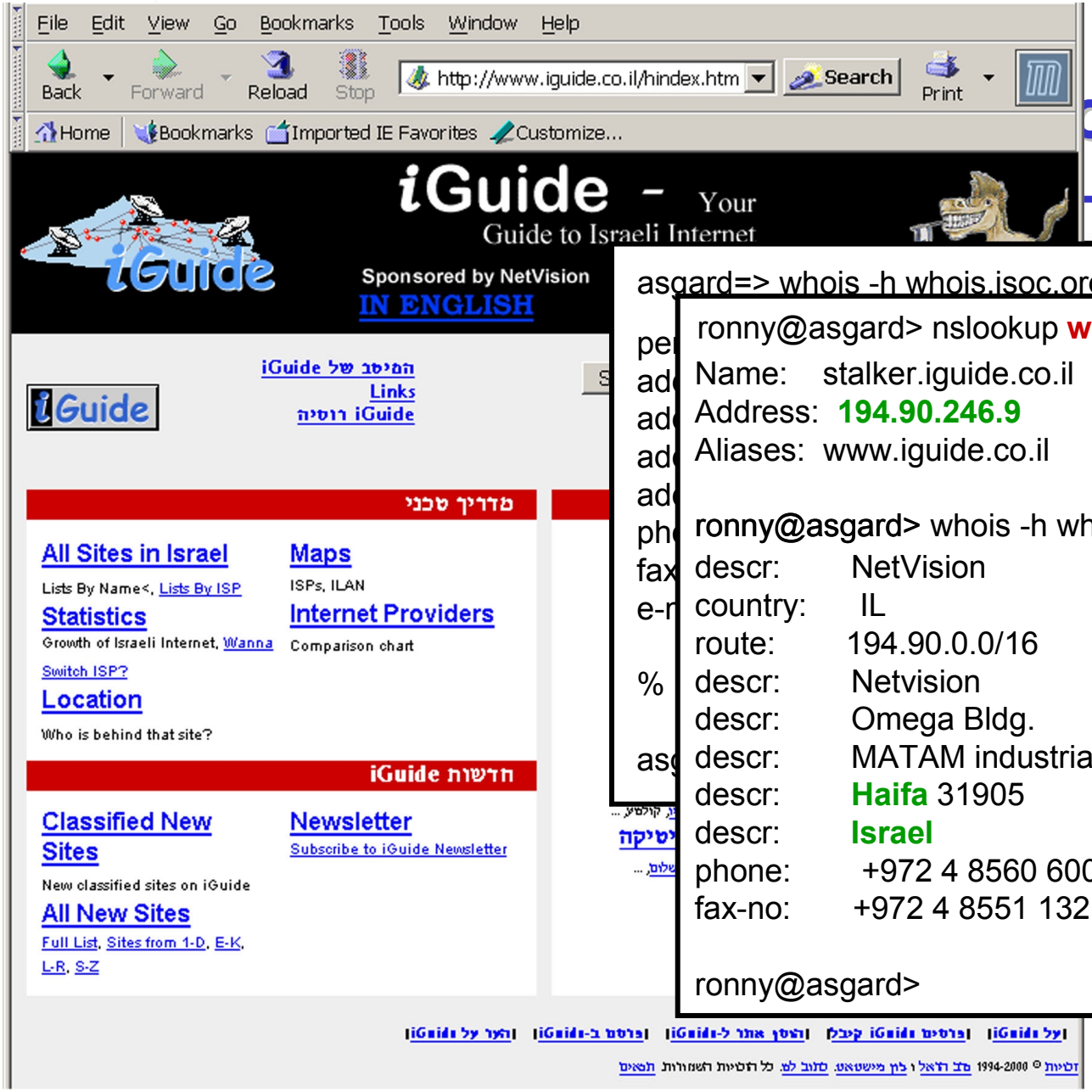
- **Web-a-where does a reasonable job of identifying place names in noisy Web pages**
  - Some novel heuristics
  - Could do better…
  - …and run faster
- **Performance tests**
  - Not found in most publications (including MetaCarta)
  - Can help in designing other disambig algorithms


My gourmet friend went
to India because he heard
there was a new deli there.

# of a page?


File Edit View Go Bookmarks Tools Window Help
Back Forward Reload Stop
http://www.iguide.co.il/index.htm
Search Print
Home Bookmarks Imported IE Favorites Customize...
iGuide - Your Guide to Israeli Internet
Sponsored by NetVision
IN ENGLISH
iGuide
המיסב של iGuide
Links
רוסיה iGuide
מדריך טכני
All Sites in Israel
Lists By Name<, Lists By ISP
Maps
ISPs, ILAN
Statistics
Growth of Israeli Internet, Wanna Switch ISP?
Internet Providers
Comparison chart
Location
Who is behind that site?
iGuide חדשות
Classified New Sites
New classified sites on iGuide
Newsletter
Subscribe to iGuide Newsletter
All New Sites
Full List, Sites from 1-D, E-K, L-R, S-Z
חברה ותרבות
ממשלה ופוליטיקה
על הרשת
iguide.co.il
ronny@asgard=> whois -h whois.isoc.org.il
person: John Neystadt
address: HaRav-Mimon 18/5
address: Bat-Yam
address: 59626
address: Israel
phone: +972 52 568-542
fax-no: +972 3 752-6012
e-mail: john@neystadt.org
% Rights to the data above are restricted by copyright.
ronny@asgard=>
AUSTRALIA
NEW ZEALAND
NORTH PACIFIC OCEAN
Equator

# of a page?



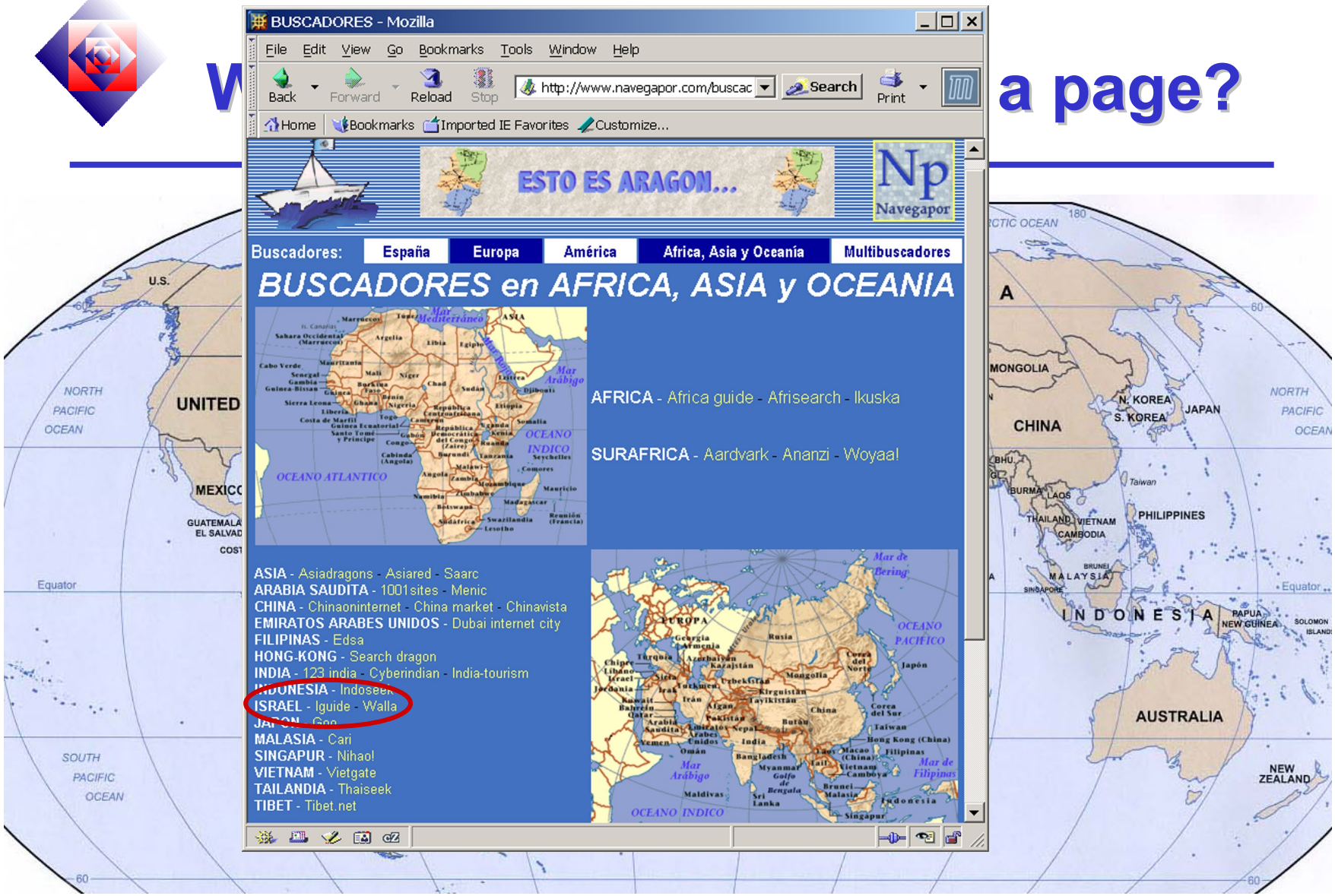
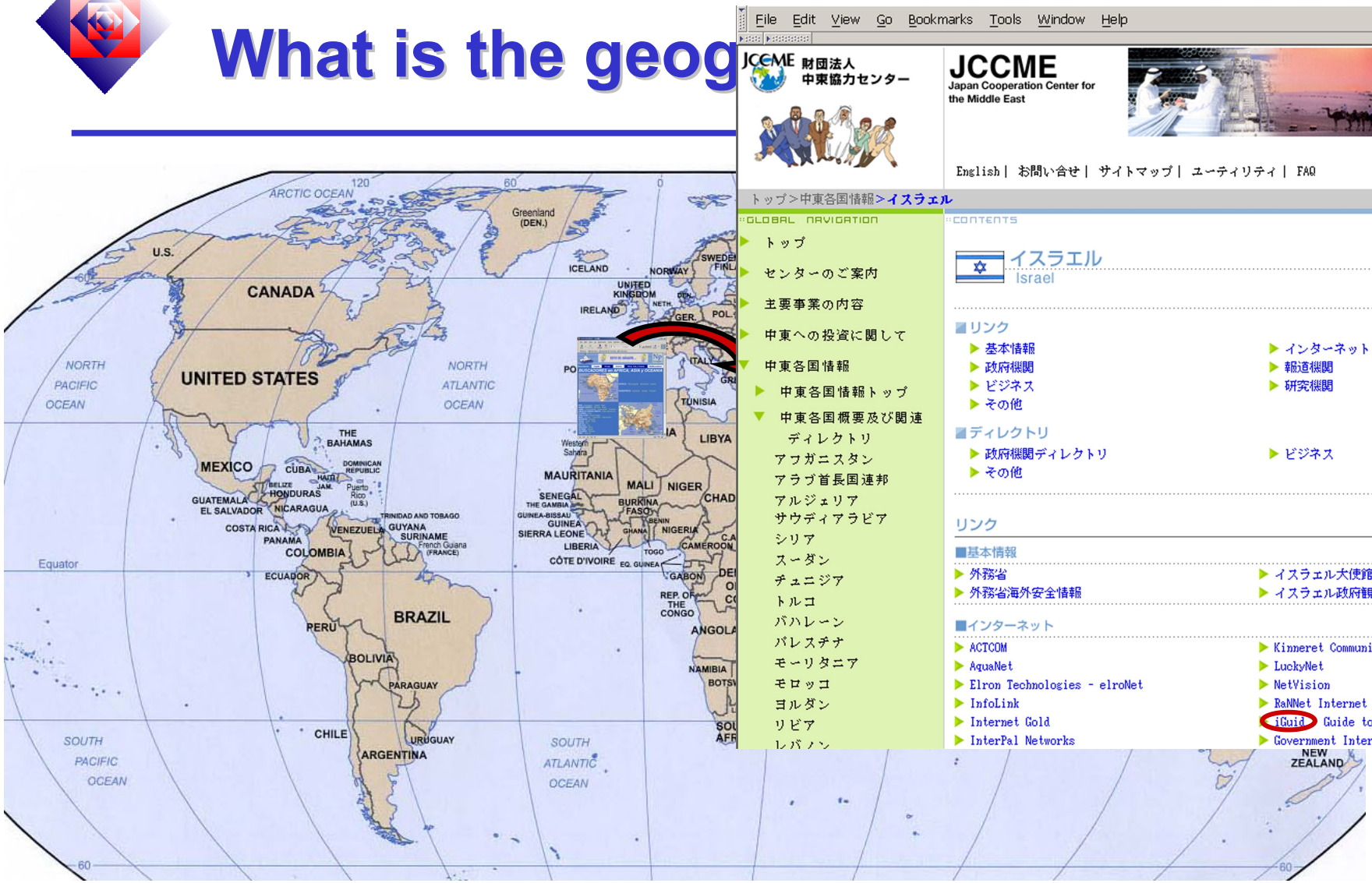

```
asgard=> whois -h whois.isoc.org.il iguide.co.il

ronny@asgard> nslookup www.iguide.co.il
Name:    stalker.iguide.co.il
Address:  194.90.246.9
Aliases:  www.iguide.co.il

ronny@asgard> whois -h whois.ripe.net 194.90.246.9
descr:       NetVision
country:     IL
route:       194.90.0.0/16
descr:       Netvision
descr:       Omega Bldg.
descr:       MATAM industrial park
descr:       Haifa 31905
descr:       Israel
phone:       +972 4 8560 600
fax-no:      +972 4 8551 132

ronny@asgard>
```

# of a page?

File Edit View Go Bookmarks Tools Window Help
Back Forward Reload Stop
http://www.iguide.co.il/hindex.htm
Search
Print
Home Bookmarks Imported IE Favorites Customize...
iGuide - Your Guide to Israeli Internet
Sponsored by NetVision Since 1994
IN ENGLISH בעברית
iGuide
iGuide המיסב של
Links
iGuide רוסיה
Search
חפש:
חיפוש מתקדם עצות חיפוש
מדריך מסווג
חדשות
מסחרי
פנאי
אומנות
עיון
מחשבים
מקומי
חינוך
מקצועי
בידור
חברה ותרבות
ממשלה ופוליטיקה
על הרשת
מדריך טכני
All Sites in Israel
Lists By Name<, Lists By ISP
Maps
ISPs, ILAN
Statistics
Growth of Israeli Internet, Wanna Switch ISP?
Internet Providers
Comparison chart
Location
Who is behind that site?
iGuide חדשות
Classified New Sites
New classified sites on iGuide
Newsletter
Subscribe to iGuide Newsletter
All New Sites
Full List, Sites from 1-D, E-K, L-R, S-Z

ARCTIC OCEAN
MONGOLIA
KYRGYZSTAN
TAJIKISTAN
CHINA
N. KOREA
S. KOREA
JAPAN
NORTH PACIFIC OCEAN
NEPAL
BHU
BANGL
INDIA
BURMA
LAOS
Taiwan
THAILAND
VIETNAM
CAMBODIA
PHILIPPINES
SRI LANKA
BRUNEI
MALAYSIA
SINGAPORE
Equator
INDONESIA
PAPUA NEW GUINEA
SOLOMON ISLANDS
INDIAN OCEAN
AUSTRALIA
NEW ZEALAND

W a page?


BUSCADORES - Mozilla
File Edit View Go Bookmarks Tools Window Help
Back Forward Reload Stop
http://www.navegapor.com/buscac
Search Print
Home Bookmarks Imported IE Favorites Customize...
ESTO ES ARAGON...
Np Navegapor
Buscadores: España Europa América Africa, Asia y Oceanía Multibuscadores
BUSCADORES en AFRICA, ASIA y OCEANIA
AFRICA - Africa guide - Afrisearch - Ikuska
SURAFRICA - Aardvark - Ananzi - Woyaa!
ASIA - Asiadragons - Asiared - Saarc
ARABIA SAUDITA - 1001sites - Menic
CHINA - Chinaoninternet - China market - Chinavista
EMIRATOS ARABES UNIDOS - Dubai internet city
FILIPINAS - Edsa
HONG-KONG - Search dragon
INDIA - 123 india - Cyberindian - India-tourism
ISRAEL - Iguide - Walla
MALASIA - Cari
SINGAPUR - Nihao!
VIETNAM - Vietgate
TAILANDIA - Thaiseek
TIBET - Tibet.net
OCEANO ATLANTICO
OCEANO INDICO
OCEANO PACIFICO
EUROPA
UNITED
MEXICO
CHINA
MONGOLIA
JAPAN
AUSTRALIA
INDONESIA
NEW ZEALAND
PHILIPPINES
NORTH PACIFIC OCEAN
SOUTH PACIFIC OCEAN
Equator

What is the geog