

Automatic Gathering of Newspaper Articles on Internet Abuse from the Internet

Results of Phase 1 of the OSILIA Project

JRC, 20 December 2000

European Commission – Joint Research Centre

Institute for Systems, Informatics and Safety (ISIS)

Stefan Scheer (RMDS - Anti-fraud Information Management)

Ralf Steinberger (RMDS - Anti-fraud Information Management)

Paul Henshaw (RIT - Web Technologies)

Neil Mitchison (RIT - Dependable Software Applications)

- Introduction to OSILIA (NM)
- Technical Goals of the OSILIA Project (RS)
- Commercial Tools and Services (RS)
- Functionality of the Software (SS)
- Storage and Retrieval of the Data (PH)
- Evaluation of the OSILIA-specific Achievements (NM)
- Future Work and Conclusions (RS)

- attacks on computers
- attacks on the Internet
- computer-based fraud
(internet-based or not)
- crime happening to pass over Internet
- crimes which can be detected through Internet
- things people disapprove of
(racism, spam, invasion of privacy, spying)

OSILIA uses the first four categories as a taxonomy for events

- High-level concern about vulnerabilities (e.g. Tampere)
- Commission services interested
JAI & INFSO, also MARKT, ENTER, SANCO...
- Major policy questions, important technological aspects
- JRC's mission => JRC interest
- IPTS and ISIS interested
- Does Cybercrime happen? If so, how much and what?
 - anecdotal evidence
 - a few widely-reported incidents
 - iceberg?
- Are there lots of other incidents reported? Newspapers/Internet

- Gather relevant open source information daily
classify and store it,
using a software agent and other programs
- Search for information with a mid- or long-term interest,
as opposed to ad hoc internet search to satisfy spontaneous needs
- Concentrate on newspaper articles (in the wider sense)
because they are manually pre-filtered information.
We want to have control over what we download

- **CyberAlert** (www.cyberalert.com)

- scans English language sites and Usenet groups
- makes them available via a database on the internet
- or forwards the information found
- Approximate cost: setup fee of a few hundred USD and 360 USD per topic and month (~5.000 USD / year and topic)

- **BBC Monitoring** (www.monitor.bbc.co.uk)

- gathers and translates published news from 100 languages into English

- Reuters, ...

- IBM's Intelligent Miner for Text

www.software.ibm.com/iminer/fortext/: > 25 K€

- Verity's Agent Server

www.verity.com/products/agentserver/ : 50 - 90 KUSD

- Excalibur's Internet Spider

www.excalibur.com : "similar price"

- Autonomy's Knowledge Update

www.autonomy.com : "six-digit sum in GBP" ~ minimum of 165 K€

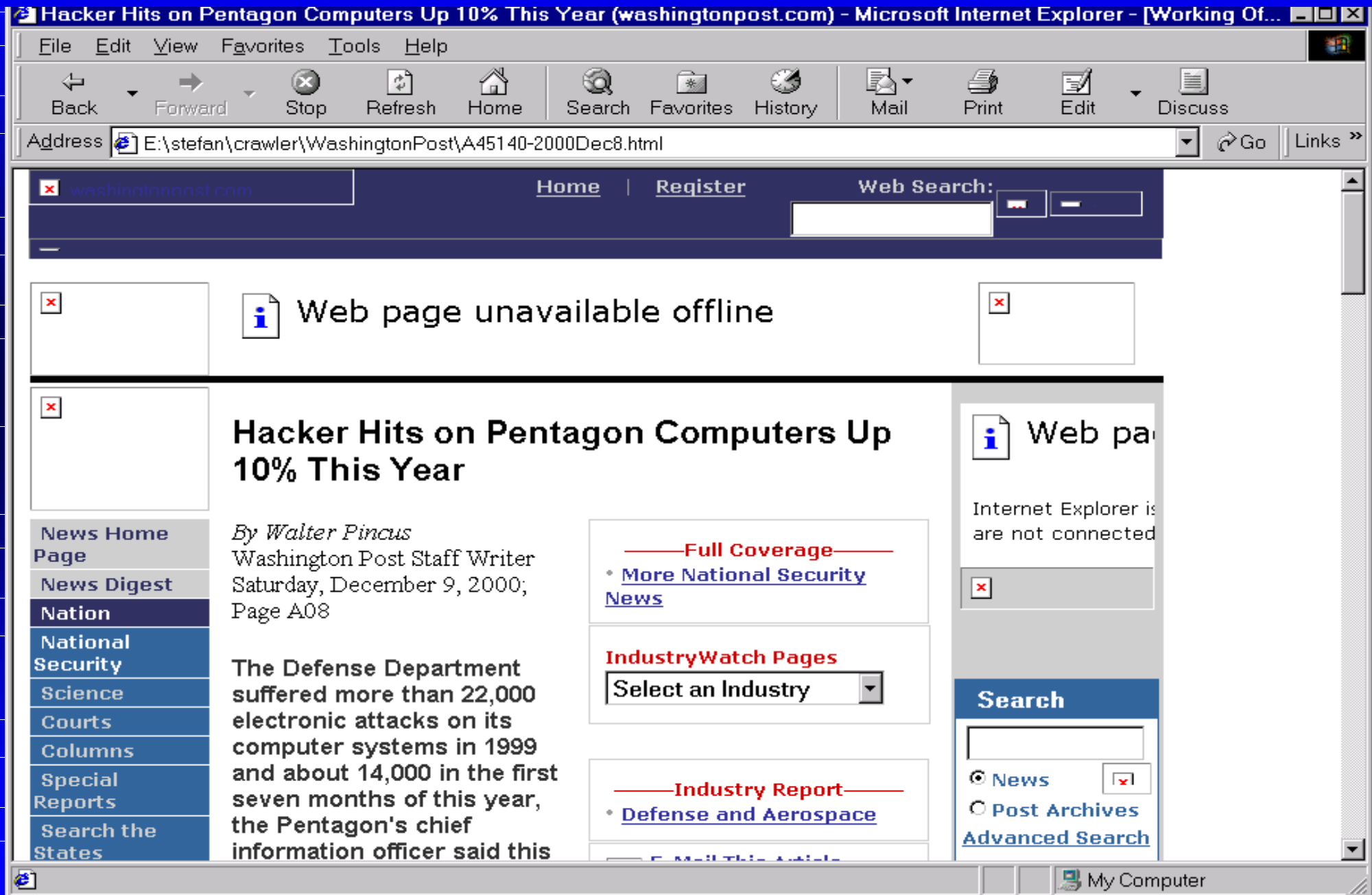
- Tenmax' Teleport Pro

www.tenmax.com/teleport/pro/ : 40 USD

- ...

- No such budget available in OSILIA and deadlines for delivery too tight
- Usually more functionality than what we need
- Often do not exactly satisfy our requirements
 - *Teleport Pro* does not allow Boolean queries
 - *Verity* and *Excalibur* only index, but do not store downloaded documents
- Advantage of own software development:
 - full control and adaptation to our needs
 - we are currently satisfied with the functionality of our crawler
- Usage of commercial software poses the same problems and still requires our own programs for filtering, cleaning, etc

- **crawler:** obtain information from the web
- **convert:** convert document in text format
- **strip / cut:** extract the “kernel”
- **compare:** throw away multiple copies
- **throw away “index” files**
- **analyse:** look for search words
- **categorise:** define document relevance with respect to categories
- **assign keywords:** to facilitate document search
- **store**
- **query**



Hacker Hits on Pentagon Computers Up 10% This Year (washingtonpost.com) - Microsoft Internet Explorer - [Working Of...]

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites History Mail Print Edit Discuss

Address [E:\stefan\crawler\WashingtonPost\A45140-2000Dec8.html](file:///E:/stefan/crawler/WashingtonPost/A45140-2000Dec8.html) Go Links >>

washingtonpost.com Home Register Web Search:

Web page unavailable offline

Hacker Hits on Pentagon Computers Up 10% This Year

By *Walter Pincus*
Washington Post Staff Writer
Saturday, December 9, 2000;
Page A08

The Defense Department suffered more than 22,000 electronic attacks on its computer systems in 1999 and about 14,000 in the first seven months of this year, the Pentagon's chief information officer said this

News Home Page
News Digest
Nation
National Security
Science
Courts
Columns
Special Reports
Search the States

Full Coverage
• [More National Security News](#)

IndustryWatch Pages
Select an Industry

Industry Report
• [Defense and Aerospace](#)

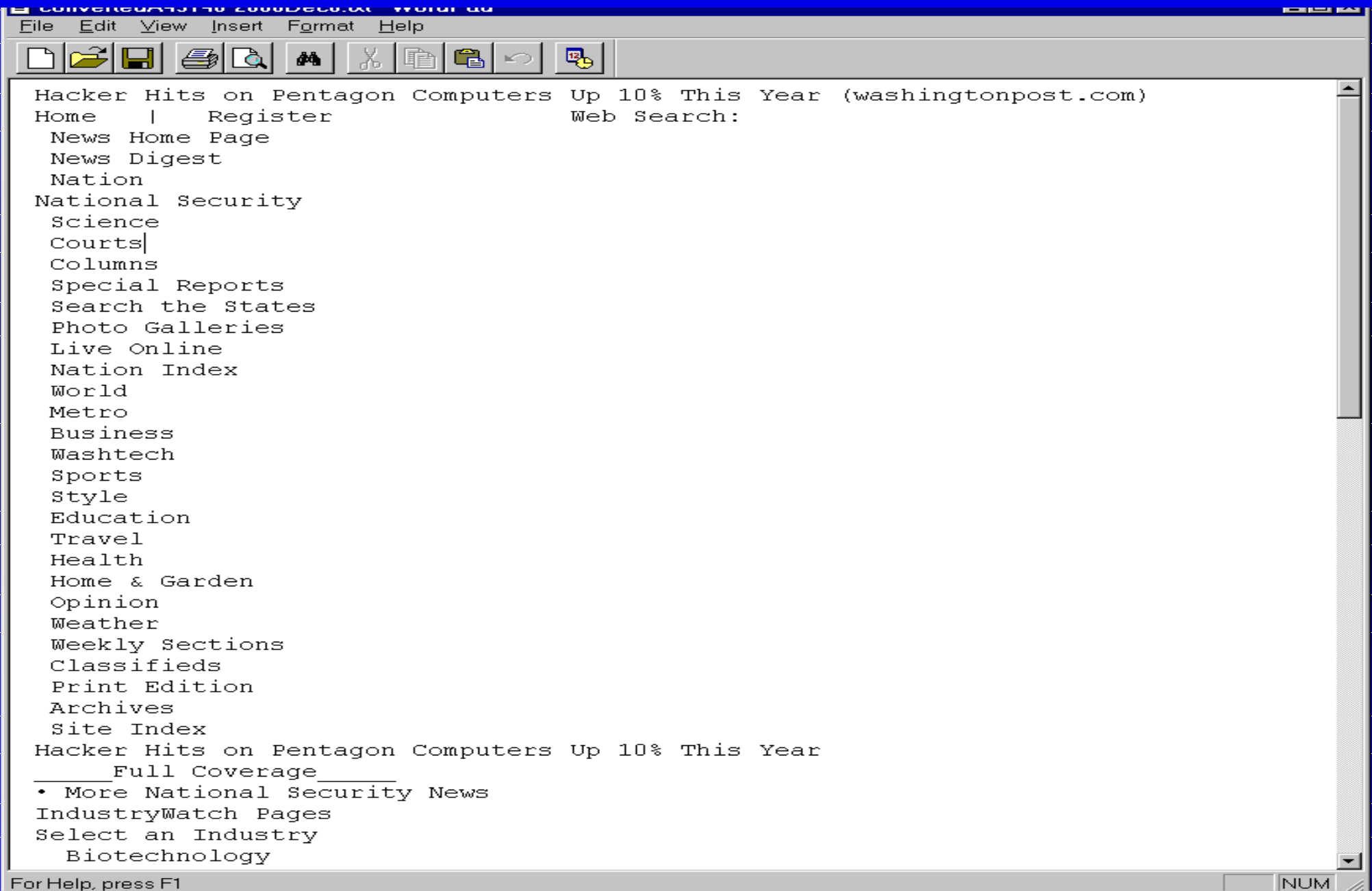
Web page unavailable offline

Internet Explorer is not connected to the Internet.

Search


News
Post Archives
Advanced Search

My Computer



WashingtonPost-A45140-2000Dec8.txt - WordPad

File Edit View Insert Format Help



Hacker Hits on Pentagon Computers Up 10% This Year (washingtonpost.com)
Hacker Hits on Pentagon Computers Up 10% This Year
IndustryWatch Pages
Select an Industry
• Defense and Aerospace
By Walter Pincus
Washington Post Staff Writer
Saturday, December 9, 2000; Page A08
The Defense Department suffered more than 22,000 electronic attacks on its computer systems in
The vast majority of those attacks were either harmless or caused only petty harassment, but
Pentagon officials said that, to the best of their knowledge, the Department of Defense's clas
The department was able to make an accurate count of the number of attacks for the first time
In 1999, the Pentagon detected 22,144 attempts to probe, scan, hack into, infect with viruses
So far this year, officials said, the number of attacks is up approximately 10 percent, and th
In an interview, Money predicted that the number of attacks is only "going to increase."
"A majority of the attacks [that cause damage] come through vulnerabilities in existing softwa
Although the Pentagon is "putting more and more effort into testing" off-the-shelf software ar
"On a lot of these [programs], we don't know where the code is written," he said.
Many of the vulnerabilities are unintentional, but some appear to be "trapdoors" deliberately
As a result, the official added, "we are not buying such off-the-shelf products in our most se
The Pentagon's cybersecurity problem is enormous. The Defense Department has roughly 10,000 cc
In August, Congress put an additional \$163 million for computer security into the fiscal 2001
The "seminal event" that awakened the Pentagon to its computer security problems occurred in E
Those attacks, which came as preparations were underway for a possible military operation agai
Military computer administrators had been warned about the weakness that the California hacker
© 2000 The Washington Post Company

For Help, press F1

NUM

ComputerUser.com - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites History Mail Print Edit Discuss

Address http://www.computeruser.com/



Formerly ComputerCurrents.com

DECEMBER 18, 2000 @447

Come See How We Did It!
Click Here

THE SHATTUCK GROUP



HOME
SEARCH
NEWS
ARTICLES
ARCHIVES
DICTIONARY
RESOURCES
FREE STUFF
LINKS
ABOUT US

"Technology Solutions for Today's Businesses"

SEARCH

GET NEWSLETTER

ADVERTISEMENTS

eWanted*
Buy and sell.
Anything you want.

DAILY OPINION



By Maggie Biggs

STRATEGICALLY SPEAKING

[Political fallout](#)

Small businesses should expect the 2000 election to affect technology choices and costs.

DAILY ARTICLES

ReleVents



Are dot-com unions needed?

Ask Molly



English to IT

Computer Advisor



Try downloading from mirrors

TODAY'S NEWS

[FCC Airwave Auction Crests \\$1 Billion In Net Bids](#)

A bevy of wireless telecommunications companies aiming to broaden their national 'footprints' drove net bidding in the FCC's

E-BUSINESS

[Penguins running wild E-Views](#)

Linux, the little OS that could, is poised to make a splash in e-business.

[E-biz platforms open up](#)

E-Business Feature
There's still a place for proprietary tools in e-commerce architecture, but Web-savvy companies are adopting open standards.

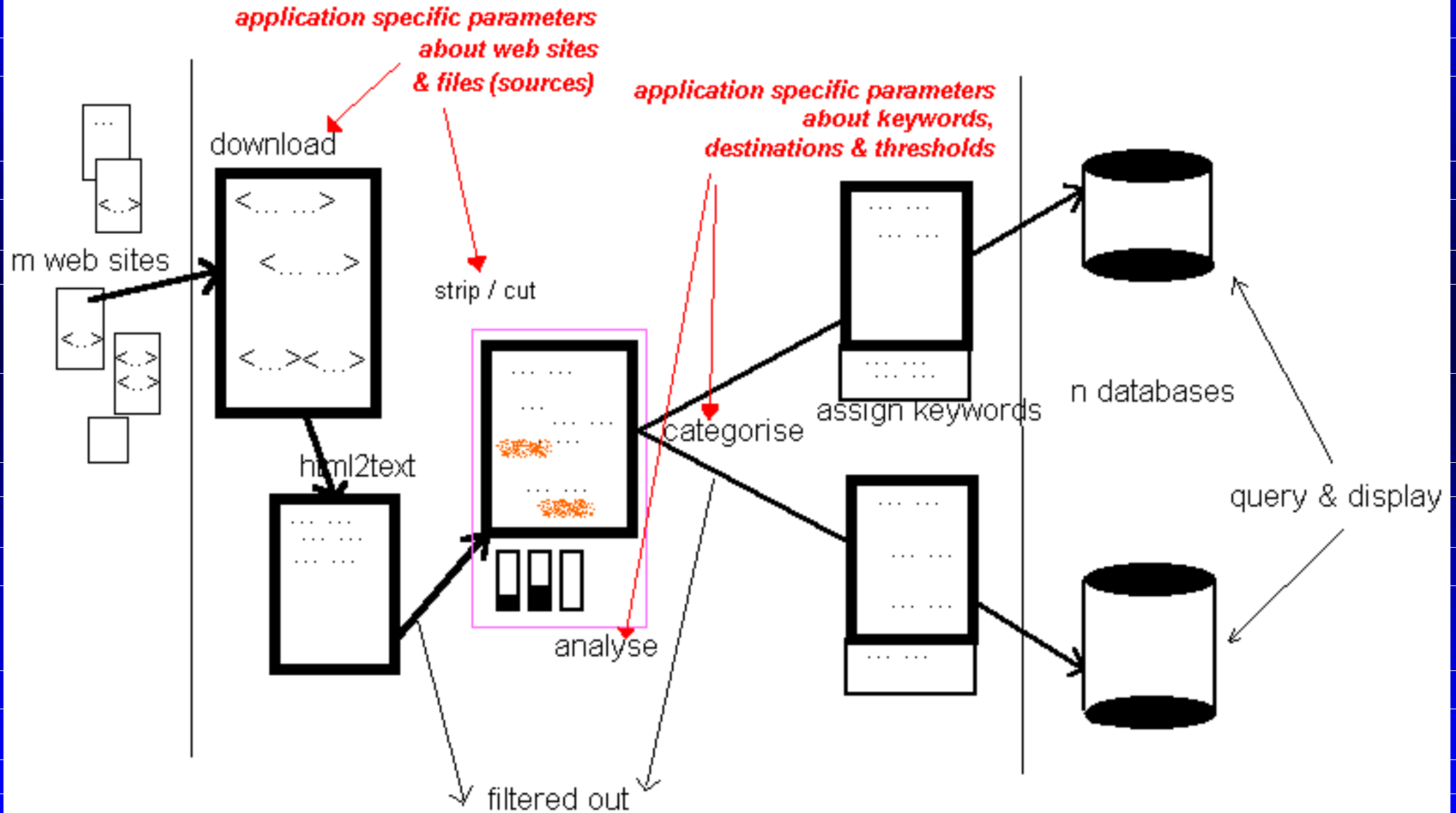
[Unix vs. Linux](#)

E-Business Step-by-Step
Unix still makes sense for large-scale applications, but more and more e-businesses are powered by the OS's younger sibling.

FEATURES

[2001: an earthbound essay](#)

A handful of ComputerUser's top analysts preview next



```
Destination = apbnews
ApplyQuery = bodyOnly
FileTypes = text html
Roaming = directory
TrimTags = all
Traversal = breadthfirst
DownloadPages = yes
DownloadImages = no
SearchDepth = 2
Timeout = 20
Retries = 1
AgentLifetime = 70
MaxFiles = 400
MaxSize = 10
RetainOnlyLastVersion = no
StartingUrl = www.apbnews.com/newscenter/internetcrime/
Query = /internet|cyber|\bweb\b|\bWWW\b|computer|electroni|digital/ix &&
       /theft|criminal|trojan|pa?edophil|abuse|\bintru|virus|\bhack|fraud|spooof/ix
```

Willkommen bei PC-WELT online - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites History Mail Print Edit Discuss

Address http://www.pcwelt.de/index.html

PCWELT.de
mehr wissen. mehr können.

Suche [Go!](#)

[Mehr Optionen](#)

- [Homepage](#)
- [Top-News](#)
- [Tipps&Tricks](#)
- [Ratgeber](#)
- [Tests](#)
- [WebTipps](#)
- [Downloads](#)
- [Treiber](#)
- [Bugs, Viren, Reports](#)

PCWELT Service

- [Ihr Organizer](#)
- [Gratisheft](#)
- [Ihre Meinung zählt](#)
- [Forum](#)
- [Geld sparen](#)
- [Gewinnspiel](#)
- [Computer-Jobs](#)

PCWELT

► **Homepage - Letzte Aktualisierung: 20.12.2000 02:40**

► **Aktuelles Thema: Toplisten und Testberichte**

TEST Frisch aus dem PC-WELT Testcenter:
Festplatten ab 500 Mark, Scanner und vier verschiedene Platinen: **Slot 1, Sockel 370, Sockel 7** und **Sockel A**. Dazu ein erster Testbericht zu einem Pentium 4-Mainboard: **MSI 850 Pro**. Alle **Testberichte und Toplisten**. **Weitere Schwerpunktthemen**.

► **Top-News**

Microsoft Recovery-CD

Geld zurück bei Recovery-CD?
Verbraucherschutz meint ja [mehr...](#)

Gratis: Klasse OpenGL-

► **Aktuelle Downloads**

GL Gravitation 0.89.0.2 **Neu**

Direct X 8.0 Final für Win9x/ME

Download Accelerator Plus 4 Beta

Opera 5.01 **Update**

Visual Basic-Buch Snow for Windows **Neu**

ZoneAlarm-dt.Anl.-Flash Get 0.92 **Neu**

Webcode [Go!](#)

[? Hilfe](#)

SUPERGEWINNE im Weihnachtsrätsel

PCWELT

Noch 3 Tage

Neu...

Alle Infos zur Großhandels-Flatrate **Neu**

DOS unter Windows ME **Neu**

Alle CDs kopieren **Neu**

So tunen Sie den Aldi-PC **Neu**

Internet

```
Destination = Paperball
Roaming = directory
Traversal = breadthfirst
... ..
Query = /internet|cyber[^\s]*|computer|ele[ck]troni|digital/ix
StartingUrl =
    www.paperball.de/service/paperball.fcgi?action=query&pg=detail&fmt=.&r=kriminal%2A+diebstahl+troja
n%2A+missbrauch+p%E4do%2A&q=kriminal%2A+diebstahl+trojan%2A+missbrauch+p%E4do%2A&stq=71&d0=&d1=&w
hat=german_web&rankedBy=date&categories=all&papers=all
Exclusions =
    www.paperball.de/service/paperball.fcgi?action=query&categories=pol&papers=all&tp=f&pg=detail&r=&f
mt=.&d0=&d1=&q=;
    www.paperball.de/service/paperball.fcgi?action=query&categories=wir&papers=all&tp=f&pg=detail&r=&fmt=
.&d0=&d1=&q=;
    www.paperball.de/service/paperball.fcgi?action=query&categories=pol&papers=all&tp=f&pg=detail&r=&fmt=
.&d0=&d1=&q=;
    www.paperball.de/service/paperball.fcgi?action=query&categories=spo&papers=all&tp=f&pg=detail&r=&fmt=
.&d0=&d1=&q=;
    www.paperball.de/service/paperball.fcgi?action=query&categories=kun&papers=all&tp=f&pg=detail&r=&fmt=
.&d0=&d1=&q=;
    www.paperball.de/service/paperball.fcgi?action=query&categories=lok&papers=all&tp=f&pg=detail&r=&fmt=
.&d0=&d1=&q=;
    www.paperball.de/service/paperball.fcgi?action=query&categories=bun&papers=all&tp=f&pg=detail&r=&fmt=
.&d0=&d1=&q=;
    www.tvtoday.de/paperball;
    www.paperball.de/service/voyeur-paperball-queries.fcgi;
    www.paperball.de/service/guestbook-paperball.fcgi;
    www.bol.de/cec/cstage?eaction=boldeepmlink&template=bolquicksearchresults.de.htm&startnum=1&incremen
t=20&query_type=clue&referrer=011001960001&keyword1_entered=kriminal+diebstahl+trojan+missbrauch+
p%E4do
```

- Boolean formula: $(a_1 | a_2 | \dots | a_n) \text{ AND } (b_1 | b_2 | \dots | b_m)$
- `^\bcomputer\binternet\bweb\bWWW\b\bserver\bcyber\bbon[-]line\bbe[-]
]?mail\bnet\b\bnetwork\bpassword\b\bcredit.card\b\bprogram\bfirewall\bencryption/gi &&
\abus\battack\bassault\bassail\bhack|war\b\bcrack\bcrim\bvirus\bworm\btrojan.horse\bpaed
?doph\bterror\bintrusion\bchild.pornogra/gi`
- `^\bcomputer\binternet\bweb\bWWW\b\bserver\bcyber\bbon[-]line\bbe[-]
]?mail\bnet\b|netz\bpasswor\bkredit[-]?karte|program\bfirewal|verschlüsselung/gi &&
/mi[sß]+brauch|angriff|attacke\bhack|krimin|vir(us|en)\bwurm\btrojan\bpädoph|terror\bbeindri
ng\bkind|porno/gi`

Apbnews (www.apbnews.com/newscenter/internetcrime)

Computer User (www.ComputerUser.com); only on WWW

Guardian (www.guardian.co.uk)

Newsbbc (news.bbc.co.uk)

Quicklinks (www.qlinks.net/quicklinks/crypto.htm,
www.qlinks.net/quicklinks/comcrime.htm,
www.qlinks.net/quicklinks/content.htm,
www.qlinks.net/quicklinks/rating.htm); meta news site

Washington Post (www.washingtonpost.com/wp-dyn)

Der Standard (derstandard.at)

Kurier (www2.kurier.at)

Die Presse (194.158.136.155/textversion.taf): text version

Frankfurter Rundschau (www.f-r.de/fr)

Süddeutsche Zeitung

(www.sueddeutsche.de/aktuell/?section=showAll&myTM=text):
text version, no exclusions

Tagesspiegel (www.tagesspiegel.de): roaming allowed

PC Welt (www.pcwelt.de/index.html)

Spiegel (www.spiegel.de)

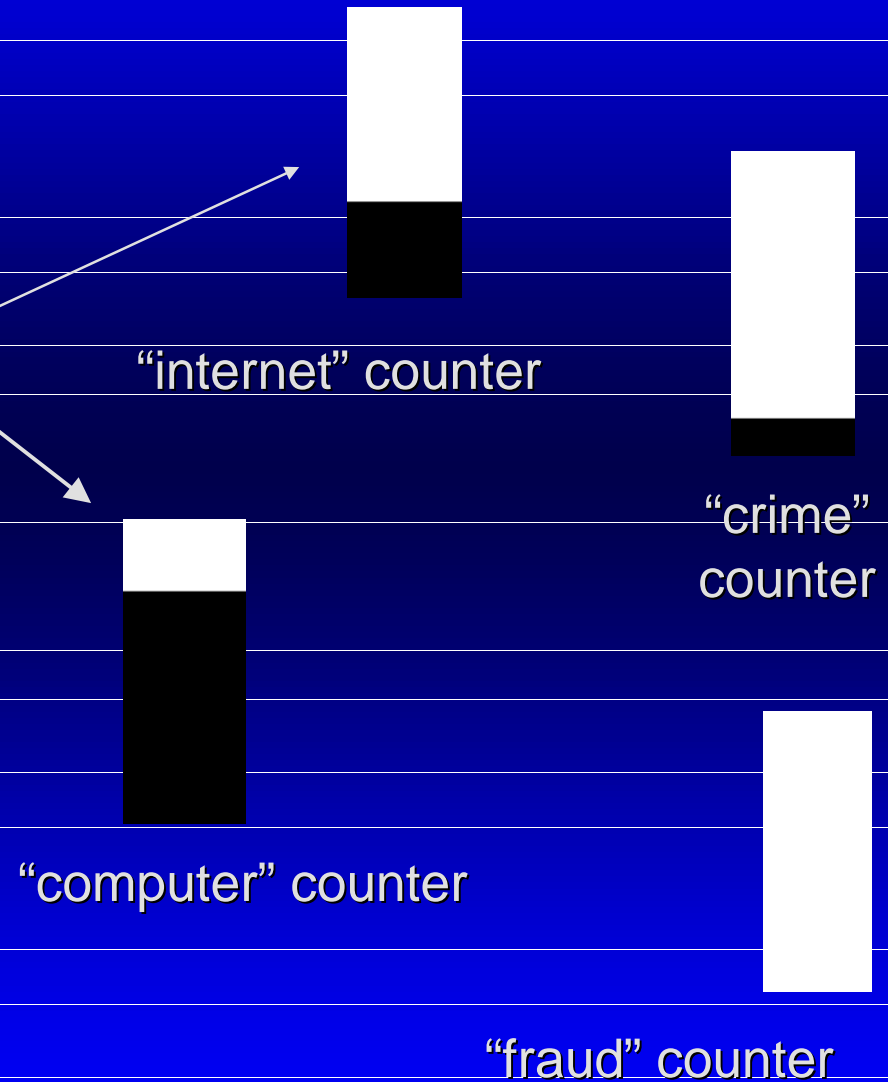
Paperball: newspaper articles search engine, various parameter files

Neue Zürcher Zeitung (newsticker.nzz.ch): few news, frequent
changes; flat hierarchy, no exclusions

Tagesanzeiger (tagesanzeiger.ch/ta, tagesanzeiger.ch/computer)

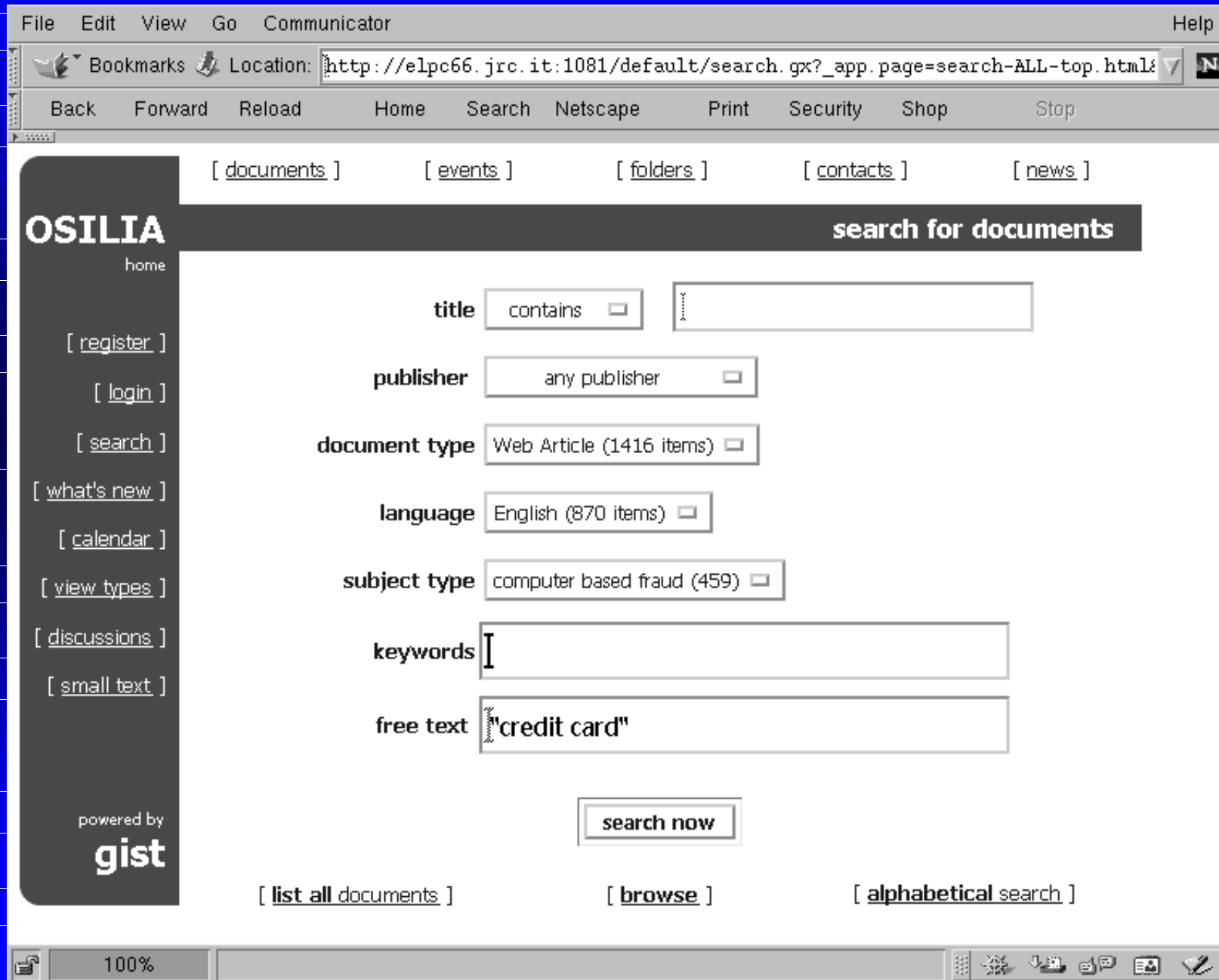
- Attacks on computers:
viruses, some Trojan Horses
 - Computer-based fraud:
“cracking”, credit card stripping,
“man-in-the-middle” attacks,
attacks on (financial) encryption
- attacks on the internet itself:
worms, some Trojan Horses, DoS attacks,
hacking of web sites
 - crime using the internet:
paedophilia, terrorism, bomb-making instructions,
incitement

- FBI, Scotland Yard (each counter + 1)
- virus, worm, hoax (computer c. +2)
- melissa, love bug (computer c. + 2)
- privacy, cookie, download (intern. +1)
- worm (internet counter +3)
- fraud [only once!] (fraud counter +2)
- crack, credit card, man in the middle (fraud counter +4)
- child, abus (crime counter +1)
- paedoph (crime counter +5)



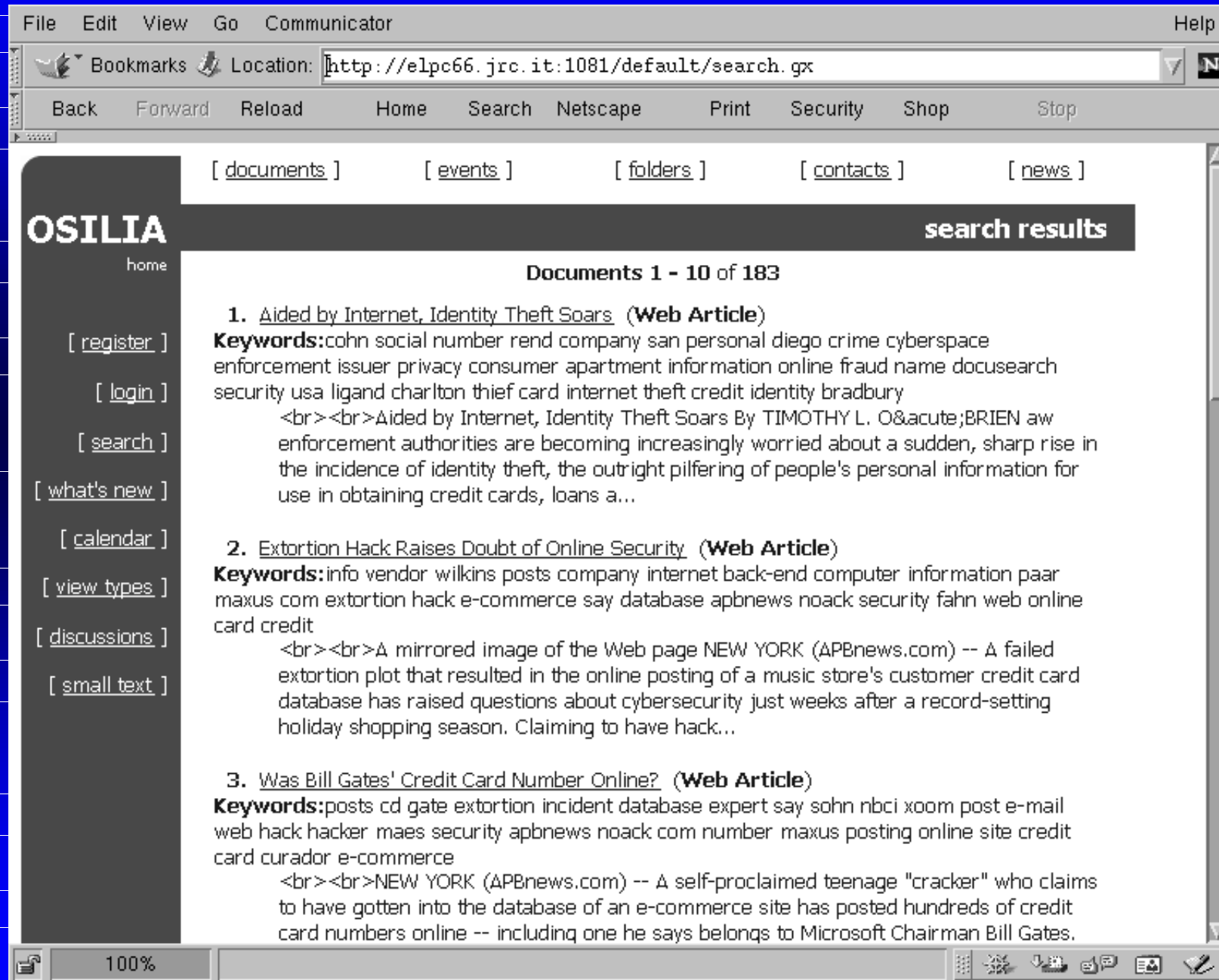
- Dynamics of site names: good maintenance needed
- dynamics / variety of web page contents: how to parameterize the stripping / cutting program? Maintenance of exclusions
- Multiply downloaded copies: hard to avoid
- ‘compare’ program too time-consuming

- start with HTML code: structure!
- Meta news sites: how to fit into the general approach? How to handle “next” button?
- “dynamic” search depth: in order to avoid “index” files



The screenshot shows a Netscape browser window with the following elements:

- Browser Title Bar:** File Edit View Go Communicator Help
- Address Bar:** Location: http://elpc66.jrc.it:1081/default/search.gx?_app.page=search-ALL-top.html
- Navigation Buttons:** Back Forward Reload Home Search Netscape Print Security Shop Stop
- Page Navigation:** [documents] [events] [folders] [contacts] [news]
- OSILIA Header:** OSILIA home search for documents
- Search Fields:**
 - title:** contains [dropdown] [input field]
 - publisher:** any publisher [dropdown]
 - document type:** Web Article (1416 items) [dropdown]
 - language:** English (870 items) [dropdown]
 - subject type:** computer based fraud (459) [dropdown]
 - keywords:** [input field]
 - free text:** "credit card" [input field]
- Search Button:** search now
- Footer Links:** [list all documents] [browse] [alphabetical search]
- Powered by:** gist
- Browser Status Bar:** 100% [icons]



The screenshot shows a Netscape browser window with the address bar containing `http://elpc66.jrc.it:1081/default/search.gx`. The browser's menu bar includes File, Edit, View, Go, Communicator, and Help. Below the address bar are navigation buttons: Back, Forward, Reload, Home, Search, Netscape, Print, Security, Shop, and Stop. The main content area displays search results for the OSILIA website. On the left is a dark sidebar with the OSILIA logo and a 'home' link, along with a list of navigation links: [register], [login], [search], [what's new], [calendar], [view types], [discussions], and [small text]. The main content area has a header with navigation links [documents], [events], [folders], [contacts], and [news], and a sub-header 'search results'. The results are titled 'Documents 1 - 10 of 183' and list three items:

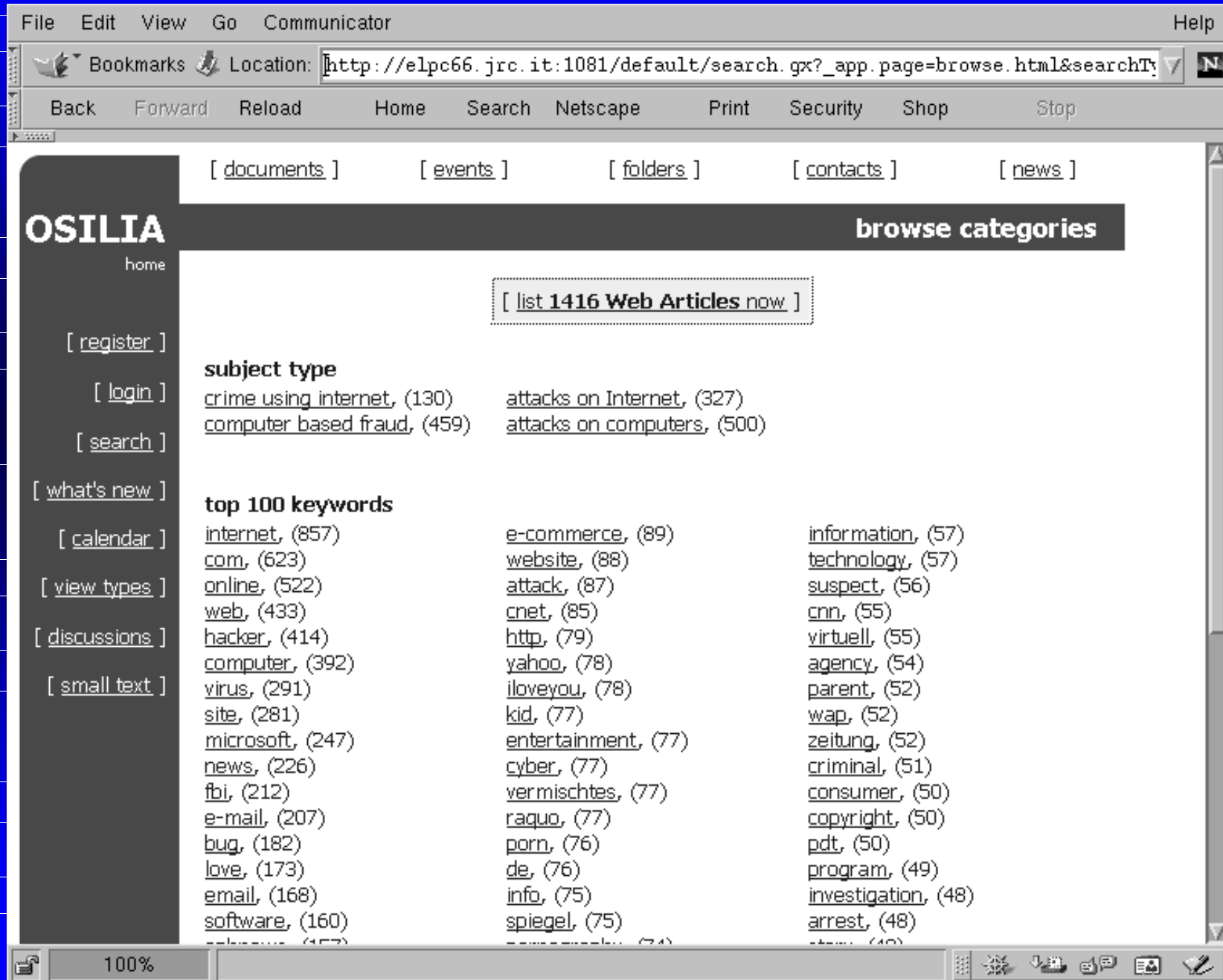
- 1. [Aided by Internet, Identity Theft Soars](#) (Web Article)**
Keywords:cohn social number rend company san personal diego crime cyberspace enforcement issuer privacy consumer apartment information online fraud name docusearch security usa ligand charlton thief card internet theft credit identity bradbury

Aided by Internet, Identity Theft Soars By TIMOTHY L. O' BRIEN
enforcement authorities are becoming increasingly worried about a sudden, sharp rise in the incidence of identity theft, the outright pilfering of people's personal information for use in obtaining credit cards, loans a...
- 2. [Extortion Hack Raises Doubt of Online Security](#) (Web Article)**
Keywords:info vendor wilkins posts company internet back-end computer information paar maxus com extortion hack e-commerce say database apbnews noack security fahn web online card credit

A mirrored image of the Web page NEW YORK (APBnews.com) -- A failed extortion plot that resulted in the online posting of a music store's customer credit card database has raised questions about cybersecurity just weeks after a record-setting holiday shopping season. Claiming to have hack...
- 3. [Was Bill Gates' Credit Card Number Online?](#) (Web Article)**
Keywords:posts cd gate extortion incident database expert say sohn nbc xoom post e-mail web hack hacker maes security apbnews noack com number maxus posting online site credit card curador e-commerce

NEW YORK (APBnews.com) -- A self-proclaimed teenage "cracker" who claims to have gotten into the database of an e-commerce site has posted hundreds of credit card numbers online -- including one he says belongs to Microsoft Chairman Bill Gates.

The browser's status bar at the bottom shows a zoom level of 100% and various system icons.



File Edit View Go Communicator Help

Bookmarks Location: http://elpc66.jrc.it:1081/default/search.gx?_app.page=browse.html&searchTy

Back Forward Reload Home Search Netscape Print Security Shop Stop

[documents] [events] [folders] [contacts] [news]

OSILIA **browse categories**
home

[register]
[login]
[search]
[what's new]
[calendar]
[view types]
[discussions]
[small text]

[list 1416 Web Articles now]

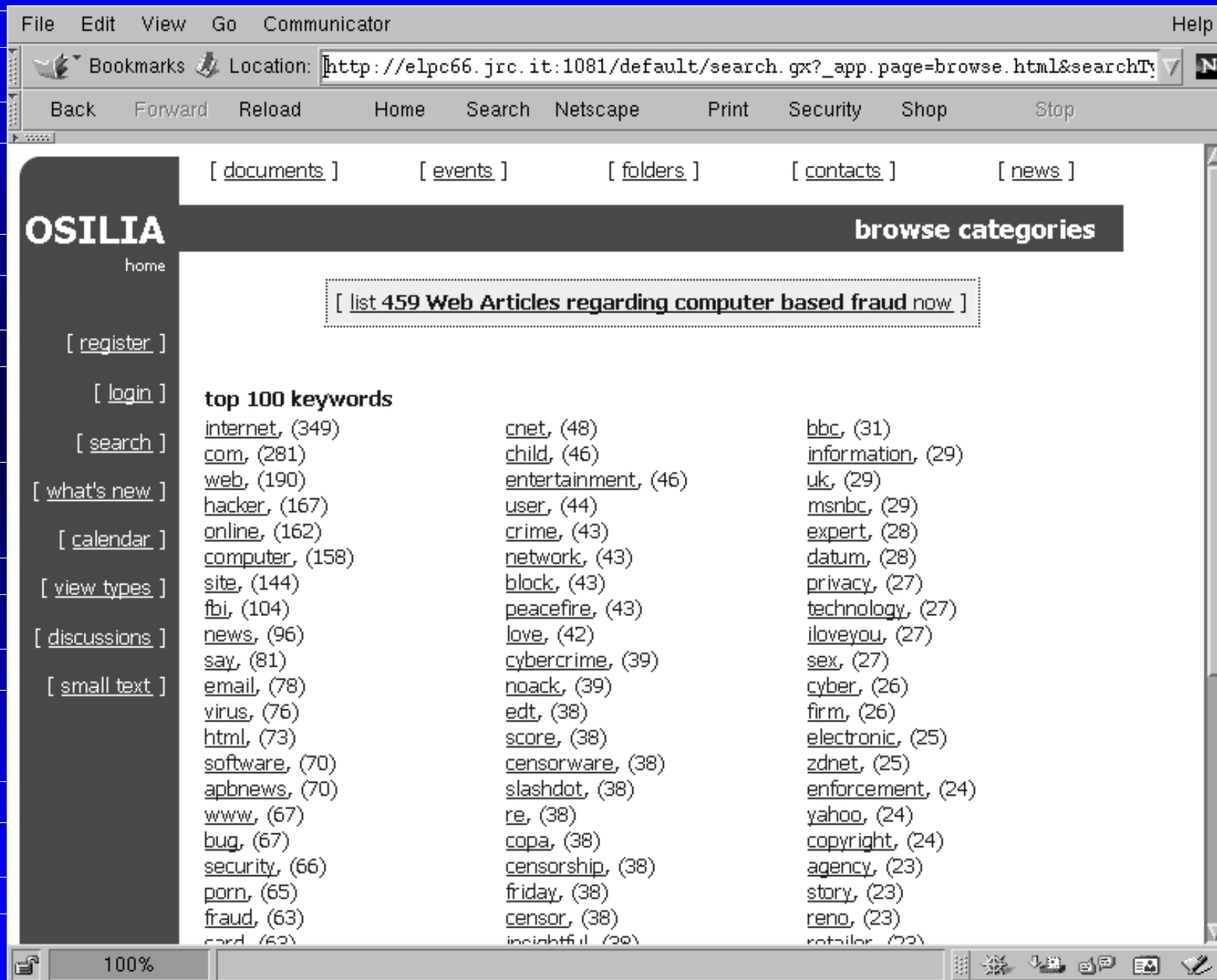
subject type

crime using internet , (130)	attacks on Internet , (327)
computer based fraud , (459)	attacks on computers , (500)

top 100 keywords

internet , (857)	e-commerce , (89)	information , (57)
com , (623)	website , (88)	technology , (57)
online , (522)	attack , (87)	suspect , (56)
web , (433)	cnet , (85)	cnn , (55)
hacker , (414)	http , (79)	virtuell , (55)
computer , (392)	yahoo , (78)	agency , (54)
virus , (291)	iloveyou , (78)	parent , (52)
site , (281)	kid , (77)	wap , (52)
microsoft , (247)	entertainment , (77)	zeitung , (52)
news , (226)	cyber , (77)	criminal , (51)
fbi , (212)	vermischtes , (77)	consumer , (50)
e-mail , (207)	raquo , (77)	copyright , (50)
bug , (182)	porn , (76)	pdt , (50)
love , (173)	de , (76)	program , (49)
email , (168)	info , (75)	investigation , (48)
software , (160)	spiegel , (75)	arrest , (48)
... , (157)	... , (74)	... , (48)

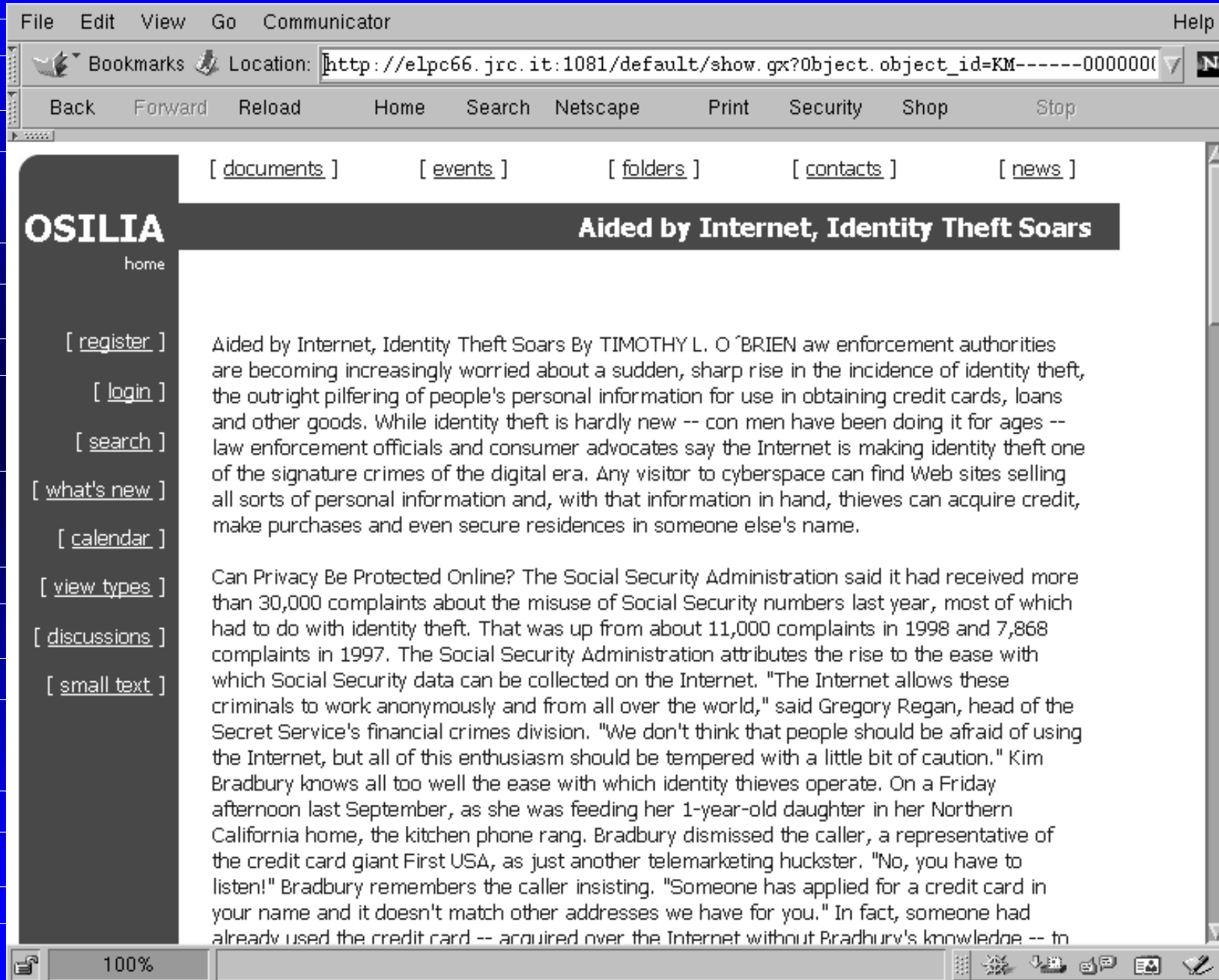
100%



The screenshot shows a Netscape browser window with the following elements:

- Address Bar:** http://elpc66.jrc.it:1081/default/search.gpx?_app.page=browse.html&searchT
- Navigation Buttons:** Back, Forward, Reload, Home, Search, Netscape, Print, Security, Shop, Stop
- Page Navigation:** [documents] [events] [folders] [contacts] [news]
- Header:** OSILIA home browse categories
- Search Results:** [list 459 Web Articles regarding computer based fraud now]
- Left Sidebar:**
 - [register]
 - [login]
 - [search]
 - [what's new]
 - [calendar]
 - [view types]
 - [discussions]
 - [small text]
- Top 100 keywords:**

internet, (349)	cnet, (48)	bbc, (31)
com, (281)	child, (46)	information, (29)
web, (190)	entertainment, (46)	uk, (29)
hacker, (167)	user, (44)	msnbc, (29)
online, (162)	crime, (43)	expert, (28)
computer, (158)	network, (43)	datum, (28)
site, (144)	block, (43)	privacy, (27)
fbi, (104)	peacefire, (43)	technology, (27)
news, (96)	love, (42)	iloveyou, (27)
say, (81)	cybercrime, (39)	sex, (27)
email, (78)	noack, (39)	cyber, (26)
virus, (76)	edt, (38)	firm, (26)
html, (73)	score, (38)	electronic, (25)
software, (70)	ensorware, (38)	zdnet, (25)
apbnews, (70)	slashdot, (38)	enforcement, (24)
www, (67)	re, (38)	yahoo, (24)
bug, (67)	copa, (38)	copyright, (24)
security, (66)	ensorship, (38)	agency, (23)
porn, (65)	friday, (38)	story, (23)
fraud, (63)	ensor, (38)	reno, (23)
card, (62)	insightful, (28)	retailer, (23)
- Footer:** 100% zoom level, system tray icons.



File Edit View Go Communicator Help

Bookmarks Location: http://elpc66.jrc.it:1081/default/show.gx?Object.object_id=KM-----000000

Back Forward Reload Home Search Netscape Print Security Shop Stop

[[documents](#)] [[events](#)] [[folders](#)] [[contacts](#)] [[news](#)]

OSILIA

home

Aided by Internet, Identity Theft Soars

[[register](#)] [[login](#)] [[search](#)] [[what's new](#)] [[calendar](#)] [[view types](#)] [[discussions](#)] [[small text](#)]

Aided by Internet, Identity Theft Soars By TIMOTHY L. O'BRIEN
aw enforcement authorities are becoming increasingly worried about a sudden, sharp rise in the incidence of identity theft, the outright pilfering of people's personal information for use in obtaining credit cards, loans and other goods. While identity theft is hardly new -- con men have been doing it for ages -- law enforcement officials and consumer advocates say the Internet is making identity theft one of the signature crimes of the digital era. Any visitor to cyberspace can find Web sites selling all sorts of personal information and, with that information in hand, thieves can acquire credit, make purchases and even secure residences in someone else's name.

Can Privacy Be Protected Online? The Social Security Administration said it had received more than 30,000 complaints about the misuse of Social Security numbers last year, most of which had to do with identity theft. That was up from about 11,000 complaints in 1998 and 7,868 complaints in 1997. The Social Security Administration attributes the rise to the ease with which Social Security data can be collected on the Internet. "The Internet allows these criminals to work anonymously and from all over the world," said Gregory Regan, head of the Secret Service's financial crimes division. "We don't think that people should be afraid of using the Internet, but all of this enthusiasm should be tempered with a little bit of caution." Kim Bradbury knows all too well the ease with which identity thieves operate. On a Friday afternoon last September, as she was feeding her 1-year-old daughter in her Northern California home, the kitchen phone rang. Bradbury dismissed the caller, a representative of the credit card giant First USA, as just another telemarketing huckster. "No, you have to listen!" Bradbury remembers the caller insisting. "Someone has applied for a credit card in your name and it doesn't match other addresses we have for you." In fact, someone had already used the credit card -- acquired over the Internet without Bradbury's knowledge -- to

100%

- Technical - successful
 - Material located
 - Filters effective
 - “Hit rates”:
 - Material relevant - 95%
 - Article relevant - 75%
 - Non-overlapping (retained) - 25%
 - Innovative, important - 5%
- Question “Are there other incidents?” answered
- Answer negative
- Consistent with cuttings

- Improve the prototype (better cleaning, avoid duplicates, ...)
- Add more sources and languages
- Apply to further domains of interest
- Assist users in specifying their interest profiles
- Further document analysis: Extraction of information such as
 - names of people and organisations
 - geographical references, etc.
- Add document similarity measure to database
- Cross-language keyword assignment and document comparison
- Visualisation of the document collection or subsets of it, using document maps

ThemeScape Map Viewer: JRC Full Text - Microsoft Internet Explorer

File Edit View Go Favorites Help

Back Forward Stop Refresh Home Search Favorites History Channels Fullscreen Mail Print Edit

Address <http://demo.cartia.com/jrcfulltext/map1024.html> Links

Map: JRC Full Text
Size: 258 documents

Map Legend Search Topic List Flags

Search for the map topics you would like to display:

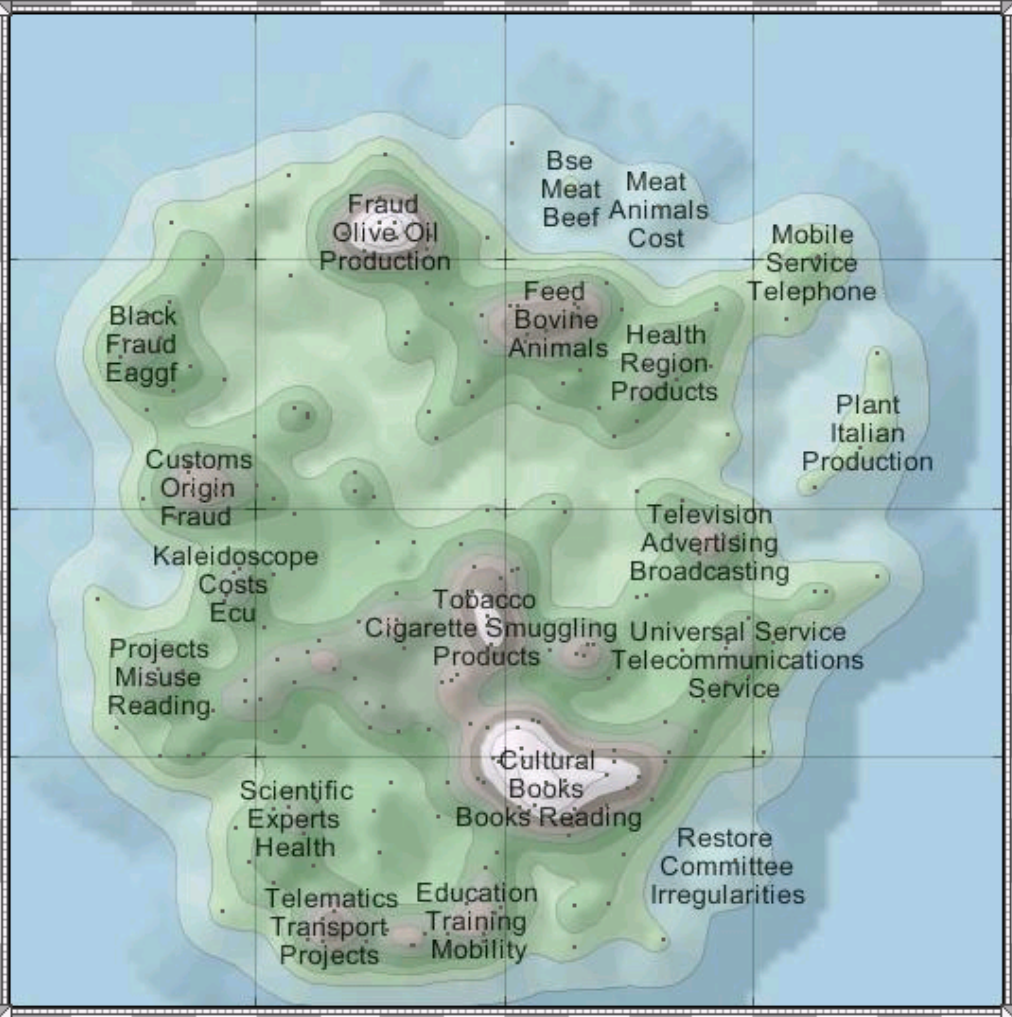
Topic	#Docs
Fraud	76
Projects	66
Production	53
Culture	47
Service	46
Regions	40
Irregularities	37
Health	36
Committee	35
Ecu	35
Customs	32
Cost	30
Export	24
Transport	23
Mecu	22
Expenditure	22
Transit	22

Clear Search

Search Options

Look for

Limit results to documents



- OSILIA is one possible application domain. Other applications are planned (e.g. customs fraud-related subjects for OLAF)
- ➔ We are looking for further customers to finance this development
- Applying the software to a new domain:
 - Setting up parameters for a specific interest profile will require several days or even weeks
- Usefulness of the software has to be decided case by case
- Precision is lower than human press clipping / classification, **but**
 - The range of sources is potentially very big
 - In the long term the automatic procedure is much cheaper
 - Results can be checked manually
 - ➔ speeding up the manual work
- JRC Technical Note on the detailed methodology in early 2001