



**exalead<sup>®</sup>**



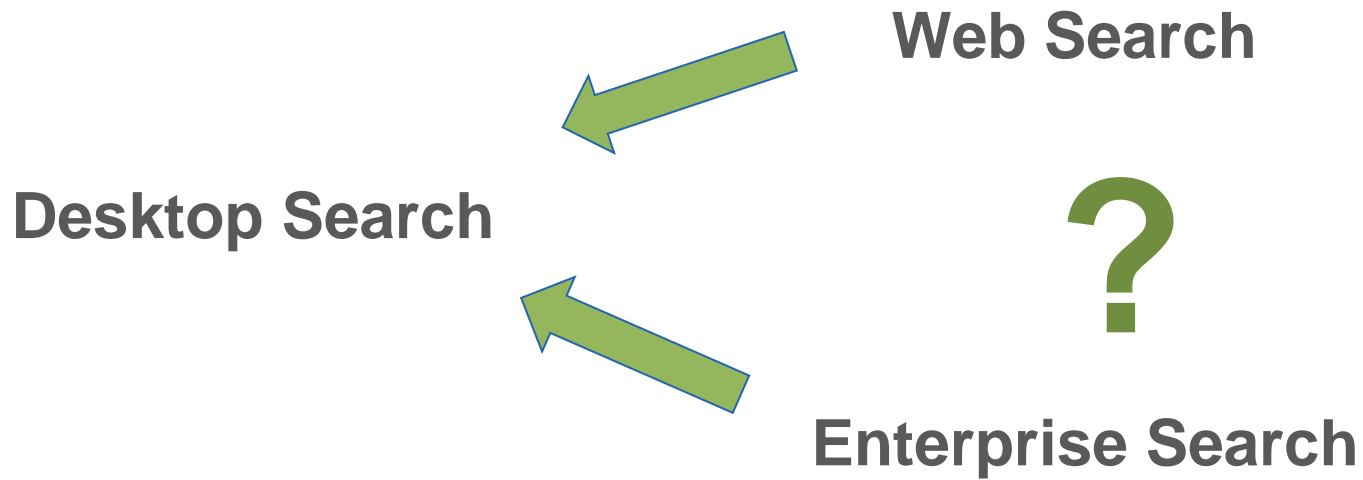
# Present and Future Desktop Search

**Gregory Grefenstette**

Desktop Search Workshop  
SIGIR2010, 23 JULY 2010, GENEVA, SWITZERLAND



- › Desktop Search, Web Search, Enterprise Search
- › Enterprise vision, workers vision
- › Workspace
- › Serendipity
- › Why people still want desktop search
- › What challenges lay ahead





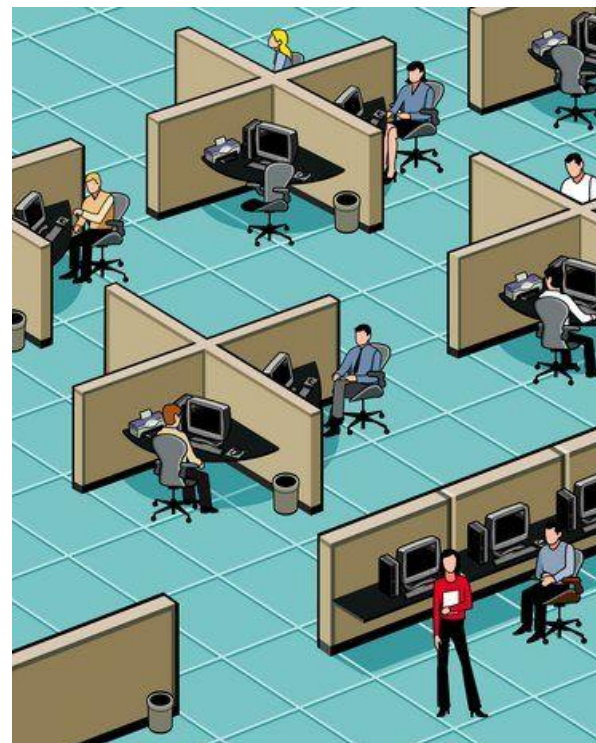
- › **Not “mini” Web search**
- › **Search for “known” items rather than browsing**
- › **Relevancy  $\neq$  links**
- › **Rich Semantics (structural, content)**
  - › More like online shopping sites
- › **Not “web pages in, web page out”**
  - › structured data, varied formats: DB base, Lotus Notes
- › **Identified Users (Access Control Lists)**
- › **Interfacing with other enterprise applications**
- › **Business Intelligence versus Buzz**

*For more, see Mark Bennett, New Idea Engineering, Inc.-Vol 5 No 4 –Summer 2008*

- › Imagine employee's laptop away from office



- › Desktop search is “mini” Enterprise search





## › I have

- › Old files, archives of papers
- › Copies of files from work to make presentations from
- › Work-based presentations from colleagues
- › Work in progress
- › Reference papers for work in progress
- › Papers I have downloaded for some reason or the other
- › Personal stuff: photos, letters
- › Etc., etc., etc.
- › Hundreds of files I might want to use

## › Personal organization of my data

# What Desktop Search really is about



- › Desktop: personal organization



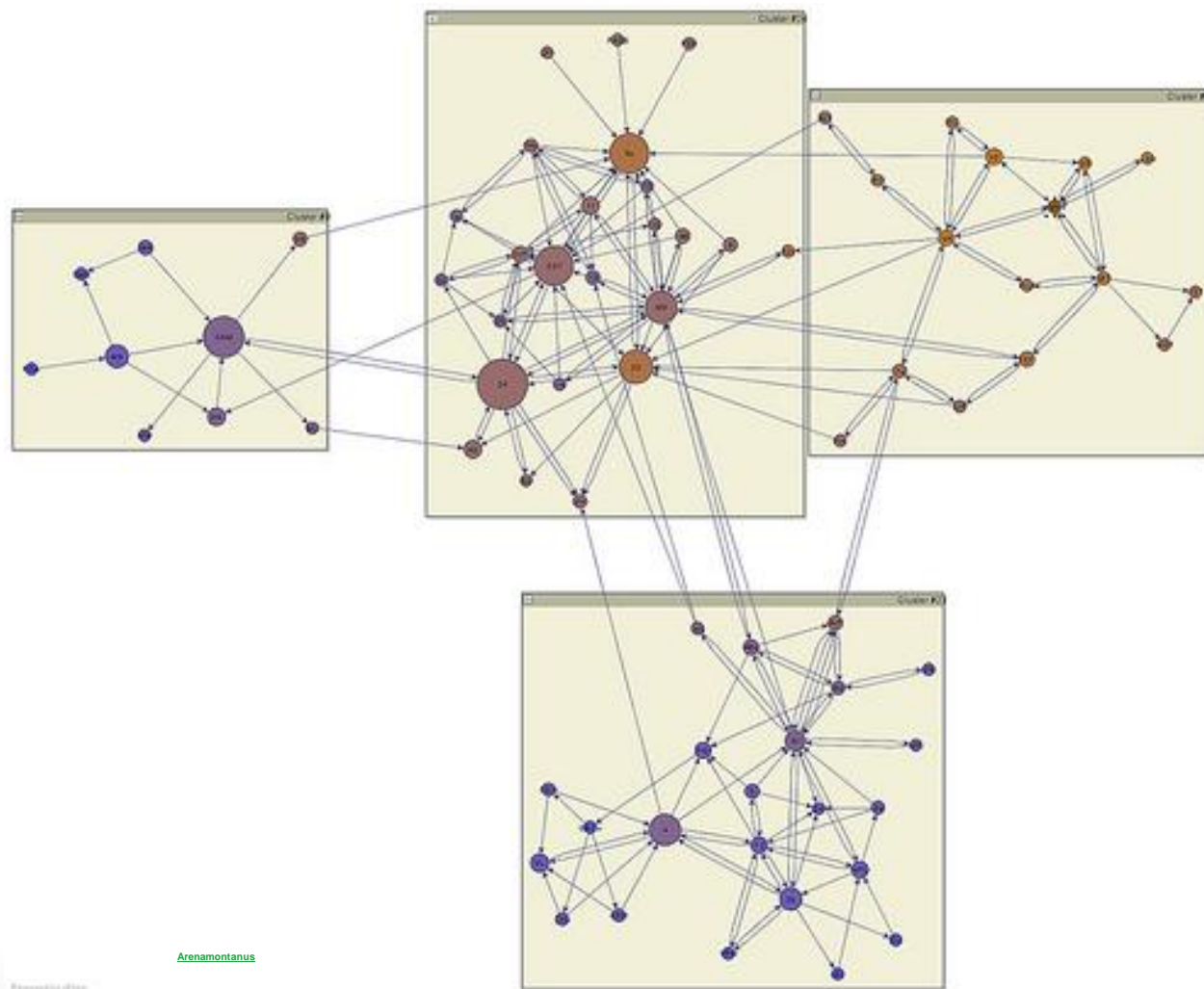
[Jeffrey Beall](#)

- › Douglas Hostadter
  - “It turns out that an eerie type of chaos can lurk just behind a facade of order - and yet, deep inside the chaos lurks an even eerier type of order”
  
- › Tom Barrett
  - “Chaos in the world brings uneasiness, but it also allows the opportunity for creativity and growth.”
  
- › Kerry Thornley
  - “What we imagine is order is merely the prevailing form of chaos.”



# Desktop Search is a locally organized slice of Enterprise search

- › Global organisation of knowledge, local organisation



Arenamontanus

Presenting files



## Practical interlude

What does our Desktop Search allow you do?

# Setup: User control over information to index



Click to expand a folder to see its contents.

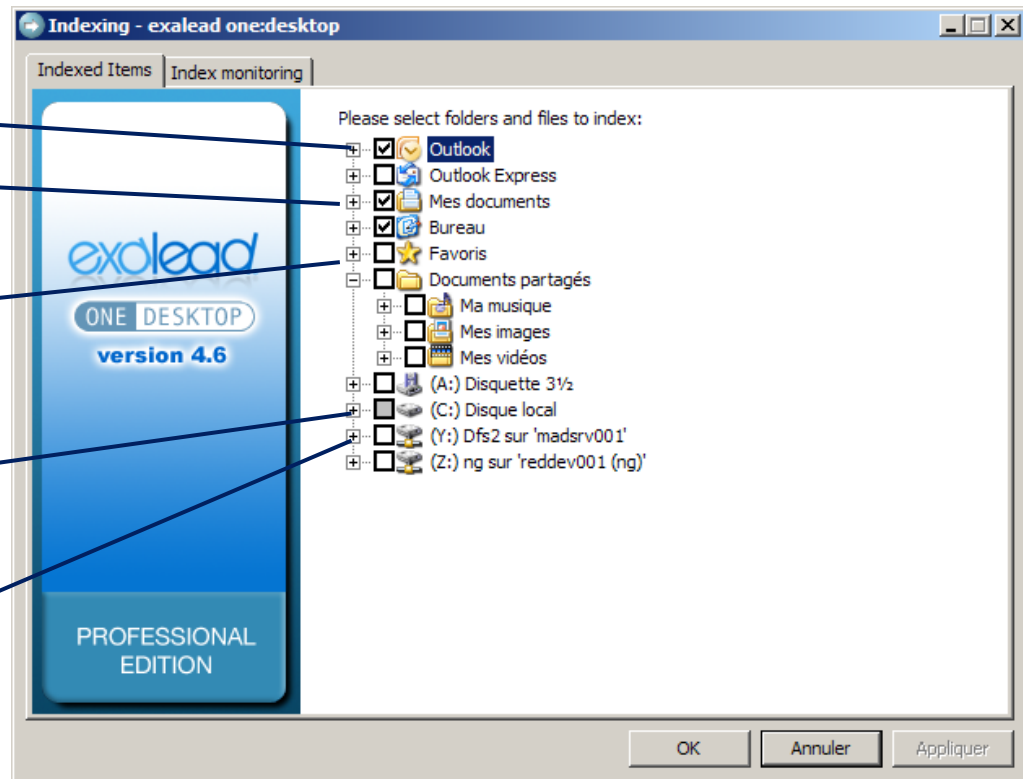
All content from this source is selected.

This source is not selected. Click it to select all content from this source.

Some content from this source is selected.

Expand this folder to see what is selected.

Set the network drives to be indexed.



# Setup: User control over indexing behaviour



Choose to build your initial index now, or later during your computer's idle time.

Choose your search strategy: optimize search speed, or limit memory usage (freeing up more memory for other applications, but slowing search speed).

Indexing - exalead one:desktop

Indexed Items | Index monitoring

exalead  
ONE DESKTOP  
version 4.6  
PROFESSIONAL EDITION

Indexing status  
54912 documents indexed (1 documents pending)

Indexing activity

Hard drives:	paused (Restart after 30 seconds of inactivity...)
Outlook:	paused (Restart after 30 seconds of inactivity...)
Outlook Express:	no document indexed
Lotus Notes:	no document indexed
Thunderbird:	no document indexed

Pause indexing | Rebuild index from scratch...

Performance

- Default settings (recommended)
- Optimize for search speed
- Limit memory usage  % of memory

Messages

The initial build is in progress. Your computer may slow down for a while. To pause the build, click on the "Pause indexing" button.

OK | Annuler | Appliquer



When the index is finished, the total number of documents indexed will be displayed.

The status of each of the resources you indexed will be displayed.

You can pause the indexing process if needed and continue later.

Adjust your search strategy anytime: optimize search speed, or limit memory usage.

Indexed Items | Index monitoring

exalead  
ONE DESKTOP  
version 4.6  
PROFESSIONAL EDITION

Indexing status  
54912 documents indexed (1 documents pending)

Indexing activity

Hard drives:	paused (Restart after 30 seconds of inactivity...)
Outlook:	paused (Restart after 30 seconds of inactivity...)
Outlook Express:	no document indexed
Lotus Notes:	no document indexed
Thunderbird:	no document indexed

Pause indexing | Rebuild index from scratch...

Performance

- Default settings (recommended)
- Optimize for search speed
- Limit memory usage  % of memory

Messages

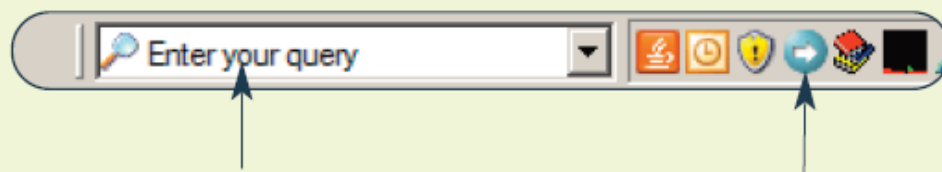
The initial build is in progress. Your computer may slow down for a while. To pause the build, click on the "Pause indexing" button.

OK | Annuler | Appliquer



## ► Launch Your Search Quickly

Search fast from your task bar:



1) Type your keywords in the Exalead toolbar\* and press "Enter" on your keyboard, or

2) Click on the Exalead icon to launch the **exalead one:desktop** home page

*\*Don't see the search field? Right click in your system tray (task bar) and select "Toolbars > Exalead one:desktop"*

# Provides search interface, welcome screen



The screenshot shows the Exalead Desktop search interface. At the top right, there are links for [Indexing](#), [Help](#), [Technical support](#), [Preferences](#), and [Feedback](#). The main logo is "exalead" in blue with "Desktop" in orange above it. Below the logo is a navigation bar with links for [Web](#), [My PC](#), [Intranet](#), [Images](#), [Wikipedia](#), [Video](#), and [More >](#). A search input field is located below the navigation bar, with a "My PC Search" button to its right. Below the input field is a "Saved queries" section. At the bottom, there is a link: [Make Exalead one:desktop your homepage](#).

Annotations on the left side:

- Click to search for documents on your Intranet. Available with exalead one:enterprise integrations only (points to the Intranet link)
- Click to search for documents on your computer (points to the My PC link)
- Type your search request here (points to the search input field)

Annotation on the right side:

- Click the button to start searching or press 'Enter' (points to the My PC Search button)



- › Once installed, indexing starts, and throughout the personal files system, things are found and labeled
  
- › Searching can begin while indexing continues



# Search result page, finding order in your local chaos

Limit your search to your PC, the Web, Images, Wikipedia or Videos  
You can also use "Advanced search" to limit your query to specific locations on your PC

Sort the results  
Order by relevance, date or size

Customize the display  
Select text-only, text plus images, or an extended information display:



Preview a document  
Click on a thumbnail image to preview a document

Open a document  
Click on the document title

View results by date or size  
Click on a day, month or year to view results for that period, or click on the file size to view results of a similar size

Scan boldface search terms

Delete a document

View all documents in a category

exalead | Web | My PC | Images | Wikipedia | Videos | More » | Indexing | Help | Preferences

text processing | My PC Search | Advanced search

My PC | Results 1.5 of about 113 for text processing (0.19 seconds)

Sort results by: Relevance | View: [Icons]

**Semantics.ppt**  
processing, but not natural language understanding. So what is natural language understanding? Answering an essay ... Always involves some level of domain modeling and ...  
30 Aug 2007 - 196k - Open parent folder - Delete  
My documents\Whitepapers\Semantic\Semantics.ppt

**Exalead Architecture White Paper.pdf**  
Text Processing Tool Connector #1 G.XML to XML Transformation Connector #2 F Common XML Document Format Statistical Linguistic Analysis Index. Transactional Updates ...  
19 Mar 2007 - 458k - Open parent folder - Delete  
My documents\extranet\EN.V2.0-Exalead Architecture White Paper.pdf

**Web-Global-Portals.pdf**  
Exalead deploys advanced text analytics and original semantic processing techniques to return uncommonly accurate results. ... Advanced text indexing and query process ...  
14 Feb 2008 - 4427k  
Outlook \imap\inbox

**Web-Global-Search-Engines.doc**  
Exalead deploys advanced text analytics and original semantic processing techniques to return uncommonly accurate results. ... Advanced text indexing and query process ...  
14 Feb 2008 - 4252k  
Outlook \imap\inbox

**e-commerce.ppt**  
Processing, Navigation, Spellchecker, Security, ... Search API Authentication Server Document Conversion, Semantic Processing, Metadata Normalization, ACL Processing ... Index ...  
21 Dec 2007 - 7460k - Open parent folder - Delete  
My documents\Resources\Commercial\documentation\Ecommerce\e-commerce.ppt

Results page: 1 2 3 4 5 6 7 8 9 Next

text processing | My PC Search

**Narrow your search**

Your refinements

Related Terms

- NOT Enterprise product remove

File types

- NOT Excel (.xls) remove

Search within results

Go

Related Terms more »

- Natural language processing x
- Unstructured data x
- Named entities x

File types more »

- Acrobat (.pdf) (52) x
- Word (.doc) (26) x

Date

- 2008 (this year) (39) x
- 2007 (47) x
- 2006 (5) x

Size more »

- 100k - 1M (52) x
- 1k - 100k (21) x
- 1M - 10M (17) x

Authors more »

- Laura Wilbar (12) x
- Carole Derbel (8) x
- Angélique Offredo (5) x

Organize the refinement panel  
Click on the wrench to hide or reorder the refinement options (drag-n-drop using the hand icon)



Switch to a tag cloud menu  
Check "Use tag cloud" in the Preferences section to display Related Terms in a tag cloud:



View more categories  
Click on "more" to view all the available categories

Exclude results  
To exclude results from a specific category, click on the 'X' icon

Narrow results by category  
Click on a category name to limit your results to documents exclusively within that category

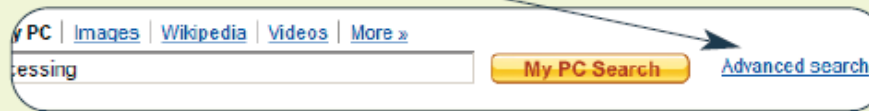
Scroll through results  
Check "Use enhanced page change" in the Preferences section to scroll through results in slideshow fashion



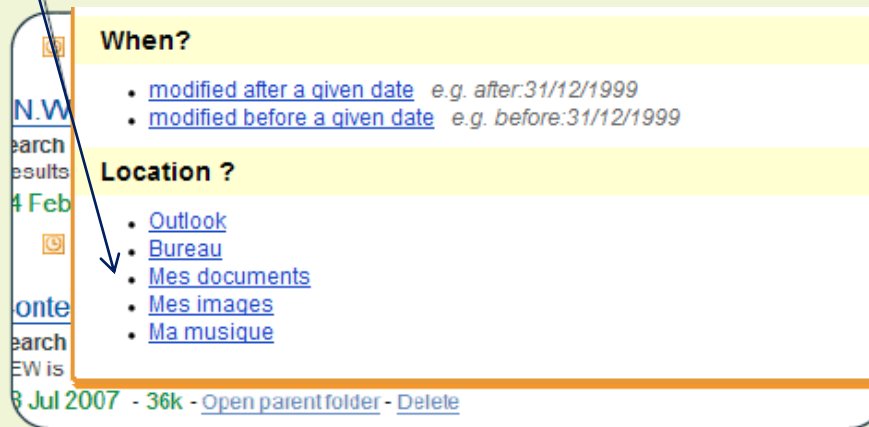
## ► Search Specific Sources

### Narrow your search

Click "Advanced search"



### Select a source to search





**Advanced search** Close X

disambiguation My PC Search

**What?**

- [exact phrase](#) e.g. "to be or not to be"
- [forbidden terms](#) e.g. cow -mad
- [words starting with](#) e.g. messag\*
- [phonetic spelling](#) e.g. soundslike:exallead
- [approximate spelling](#) e.g. spellslike:exlaead
- [adjacent words](#) e.g. (stock NEAR exchange)
- [logical expression](#) e.g. ( (fast OR speed) AND NOT light )
- [regular expression](#) e.g. /a.c/

**Where?**

- e.g. language:en
- [in files of a given format](#) e.g. filetype:pdf
- [in the title](#) e.g. intitle:(exalead search engine)

**When?**

- [modified after a given date](#) e.g. after:31/12/1999
- [modified before a given date](#) e.g. before:31/12/1999

**Location ?**

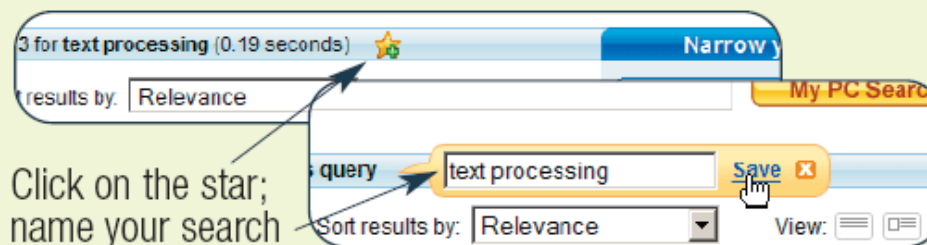
- [Bureau](#)
- [Outlook](#)
- [Outlook Express](#)

# Common queries can be favorited

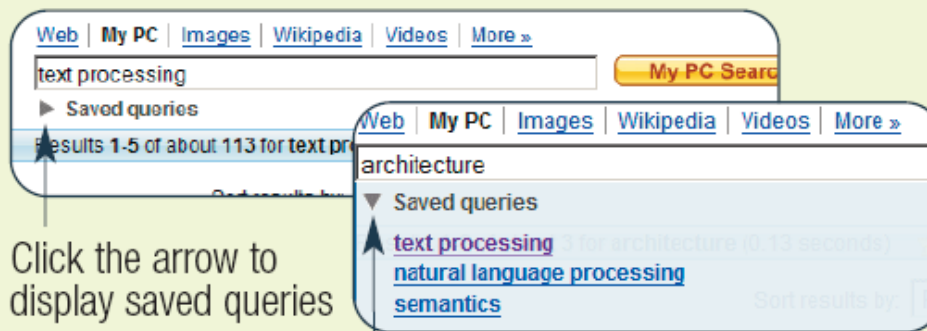


## ► Bookmark Your Searches

Save time by bookmarking frequent queries



Click on the star;  
name your search



Click the arrow to  
display saved queries

Click again to hide the list




► Navigate Your Results Page

Preview a document

On the results page, click on a thumbnail image to launch the Preview window





2 [Wintner+Yona\\_paper.pdf](#)  
(Wintner+Yona\_paper.pdf)

Much of the infrastructure required both for practical applications and for computational linguistics research is ... Linguistics, 36:33-38, December. In Hebrew. ...

21 Apr 2004 - 58k

1 2 3

pdf pdf

linguistics

[Back to results](#)

**Preview** 6 terms found: [linguistics](#) [previous](#) [next](#) [Go to the email](#)

Resources for Processing Hebrew  
Shluly Wintner and Shlomo Yona Department of Computer Science University of Haifa, Israel shluly,shlomo @cs.haifa.ac.il

**Abstract** We describe work in progress whose main objective is to create a collection of resources and tools for processing Hebrew. These resources include corpora of written texts, some of them annotated in various degrees of detail; tools for collecting, expanding and maintaining corpora; tools for annotation; lexicons, both monolingual and bilingual; a rule-based, linguistically motivated morphological analyzer and generator; and a WordNet for Hebrew. We emphasize the methodological issue of well-defined standards for the resources to be developed. The design of the resources guarantees their reusability, such that the output of one system can naturally be the input to another.

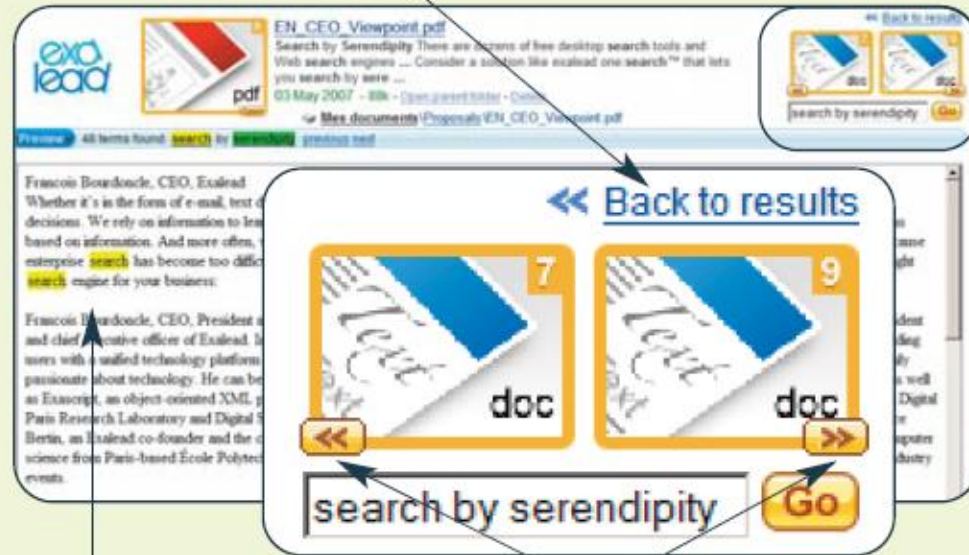
**1 Introduction**  
The state of the art in computational processing of Hebrew, as described by Wintner (2003), leaves much to be desired. Much of the infrastructure required both for practical applications and for computational linguistics research is either non-existent, lacking or proprietary. In this paper we describe work in progress whose main objective is to create a collection of resources and tools which are instrumental in most conceivable applications of natural language processing, in particular machine translation. These resources include corpora of written Hebrew, some of them annotated in various degrees of detail; tools for collecting, expanding and maintaining corpora; tools for annotation; lexicons, both monolingual and bilingual; a rule-based, linguistically motivated morphological analyzer and generator; and a WordNet for Hebrew. We emphasize the methodological issue of welldefined standards for the resources to be developed. In particular, we use XML for defining the structure of corpora, annotated corpora, lexicons and morphological analyses. The design of the resources guarantees their reusability; in particular, it is essential that all the systems we develop adhere to the same standards, such that the output of one can naturally be the input to another. While the work we describe here is specific to Hebrew, the methodological principles which guide it are language independent. In the next section we list some facts about the language. Section 3 describes the existing corpora, their



## ▶ Navigate Document Previews

Navigate the preview screen

Return to Results page



Scan highlighted search terms for a quick evaluation

Tab through results without leaving the Preview screen

# User can set preferences for search interface



Set the language for the user interface

Enable automatic word stemming to include words sharing a common root with your search term.

Set the number of search results to be shown per page

Set to display a tag cloud for related terms

Set the default search results view

Choose whether to update the index when your computer is on battery power

Set this to automatically check for product updates

Set up your Lotus Notes for indexing; including installation path, ini file and database directory

Click to save your preferences

**Preferences** Save and go back

**Interface language**

Display the Exalead interface in:

**Search**

Automatic word stemming  
 Enabled

**Display**

Number of search results per page:

Use tag cloud to display related terms.

Use enhanced page change.

Default view on the results page:

(Text only)  
 (Text and thumbnail)  
 (Text, thumbnail and extra info)

Use advanced preview (Active X available only with Internet Explorer)

**Miscellaneous**

Launch Exalead one:desktop automatically on boot.

Do not index while on battery (in case of a portable computer)

Automatically check for product updates

**Lotus Notes**

Lotus Notes installation path

Path to the notes.ini file

Lotus Notes workspace location

Local databases directories (example: c:\tmp;d:\archives)

Lotus Notes password: no password has been saved

**Preferences** Save and go back



## Extracting Order from Chaos « finding things in the mess »



# Where are things?

The screenshot shows the exalead search engine interface. The search bar contains the word "linguistics". The search results page displays a list of files, including "LuciaTALN2.rtf (LuciaTALN2.rtf)". A sidebar on the right titled "Narrow your search" provides filters for "Source" and "Related Terms".

**Search Results:**

- Preview:** Outlook\Dossiers personnels\LIC2M
- Authors:** Fluhr Christian 082124
- Recipients:** /o=cea
- File:** LuciaTALN2.rtf (LuciaTALN2.rtf)
- Description:** International conference on computational linguistics (COLING), pages 588–592, Kyoto, Japan. ... 23 Jan 2005 - 133k
- Source:** Outlook\Dossiers personnels\LIC2M

**Narrow your search sidebar:**

- Source:**
  - Outlook (20) x
  - Dossiers personnels (20) x
    - LIC2M (18) x
  - Bureau (19) x
  - exalead (14) x
    - Rapport de recherche (7) x
    - Brevet Soumissions (2) x
    - Brevets Pre Hirsch (2) x
- Related Terms:**
  - ACM Symposium Annual International ACM SIGIR Conference Cette approche **Data extraction**
  - Hebrew Linguistics **High level**
  - IEEE International Conference
  - Knowledge base Linguistic Analysis **Machine translation** Mining
  - Of Text Morphological analysis Morphological analyzer for Hebrew
  - New York** Notre système
  - Proper names Semantic Web** Syntactic analysis Traitement linguistique Word formation
- Search within results:**  **Go**
- File types:**
  - Word (.doc) (13) x

- Bureau
- Publications

▼ Related Terms

Abstraction Levels Bilingual dictionary Dependency  
extraction **Dictionary lookup** External  
evidence Knowledge acquisition bottleneck Large  
quantities of data Light Parsing Noun phrase extraction  
**POS tagging** Rank Xerox Research Centre  
Recent research Research Report Sentence length  
Singular Value Decomposition Statistical models Text  
type User to specify Word research Words in the  
sentence

▼ Languages

- English (67)
- French (2)

▼ Search within results

▼ Source

- Papers (69)
- Old Papers (23)
- papers (17)
- presentations (4)
- Tutorials (2)
- Stair2003 (9)
- RIAO2004 (8)

▼ File types

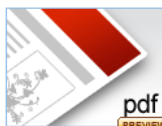


[ciaa2010.pdf](#)

per position where values are added when crossing the corresponding parenthesis in the pattern or removed by disambiguation ...

10 Jun 2010 - 118k - [Open parent folder](#) - [Delete](#)

• Bureau \Publications\CIAA 2010\ciaa2010.pdf



[jcdl75-popescu \(3\).pdf](#)

Named entity extraction is used for disambiguation in open text domains [4] and [7]. Ambiguity ... Reported results in information retrieval [17] or word sense disambig ...

25 Mar 2010 - 424k - [Open parent folder](#) - [Delete](#)

• Bureau \Publications\jcdl75-popescu (3).pdf

• Authors: End User Computing Services



[jcdl75-popescu.doc](#)

Named entity extraction is used for disambiguation in open text domains [4] and [7]. Ambiguity ... Reported results in information retrieval [17] or word sense disambig ...

25 Mar 2010 - 509k - [Open parent folder](#) - [Delete](#)



[touris](#)

Cathed

29 Oc



[Seme](#)

Englist

30 Jur

**Narrow your search**

- Bureau
- Publications

▼ Related Terms

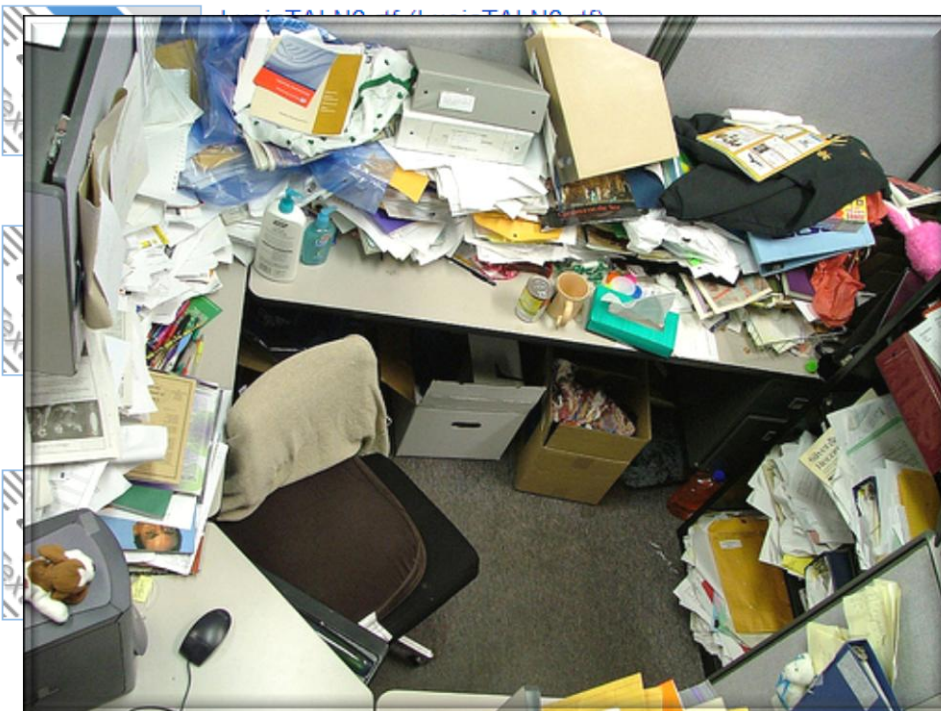
Abstraction Levels Bilingual dictionary Dependency  
extraction **Dictionary lookup** External  
evidence Knowledge acquisition bottleneck Large  
quantities of data Light Parsing Noun phrase extraction  
**POS tagging** Rank Xerox Research Centre  
Recent research Research Report Sentence length  
Singular Value Decomposition Statistical models Text  
type User to specify Word research Words in the  
sentence

e terms assigning the ...

# Ordre from Chaos

http://localhost:17610/search/C=eJw9iksOgCAMBd!R9EgNPJWkE\*BeHxx42ommmWnQZOdMPDwKGgwRB4ydaBexbgCNihzlkWptZdi

Recipients: /o=cea



Kyoto, Japan. ...

588-592. Johnson, C. D.

g ... Linguistics, Barcelona.

[Coria05-final.doc \(Translating Chinese Romanized Name into ...\)](#)

Linguistics ACL, Barcelona, Spain, 25-26 July 2004. Meng H., Lo W., Chen B., Tang ... Linguistics, Barcelona. July 21-26, 2004 Virga P., Khudanpur S., Transliteration ...

08 Feb 2005 - 144k

Outlook\Dossiers personnels\LIC2M

Authors: [Li Yiping 201262 Doctorante](#)

Recipients: [/o=cea](#)

[Coria05-final.doc \(Translating Chinese Romanized Name into ...\)](#)

Linguistics ACL, Barcelona, Spain, 25-26 July 2004. Meng H., Lo W., Chen B., Tang ... Linguistics, Barcelona.

## Search within results

## File types

- [Word \(.doc\)](#) (13) x
- [Acrobat \(.pdf\)](#) (12) x
- [Message](#) (6) x

## Date

- [2010 \(this year\)](#) (5) x
- [2009](#) (4) x
- [2008](#) (9) x

## Size

- [100k - 1M](#) (18) x
- [1k - 100k](#) (14) x
- [1M - 10M](#) (7) x

## Authors

- [Semmar Nasredine 202247](#) (6) x
- [ankaoua](#) (5) x
- [Li Yiping 201262 Doctorante](#) (3) x

## Recipients

- [/o=cea](#) (15) x
  - [ou=dir-far](#) (15) x
    - [cn=recipients](#) (15) x
- [Grefenstette Gregory 206823](#) (2) x
- [Dominique Orban De Xivry](#) (1) x

## Languages

- [English](#) (24) x
- [French](#) (11) x



- › Recent past and current situation
  - › Desktop search == commodity
  - › Lack of interest
  - › « little web browser »
  - › Cheap or free



- › Search functionalities
  - Limitations in standard search functions
- › Number of formats supported
  - 120 (free, individual), 300 (enterprise version)
- › Robustness
  - Same source code as enterprise version
- › Stability
- › Federation with work environment
  - Parallels with Enterprise Search
  - Connection with Enterprise Data
  - Global Policy Management
- › Local Control of Local data
  - Individual control



## Using GPO

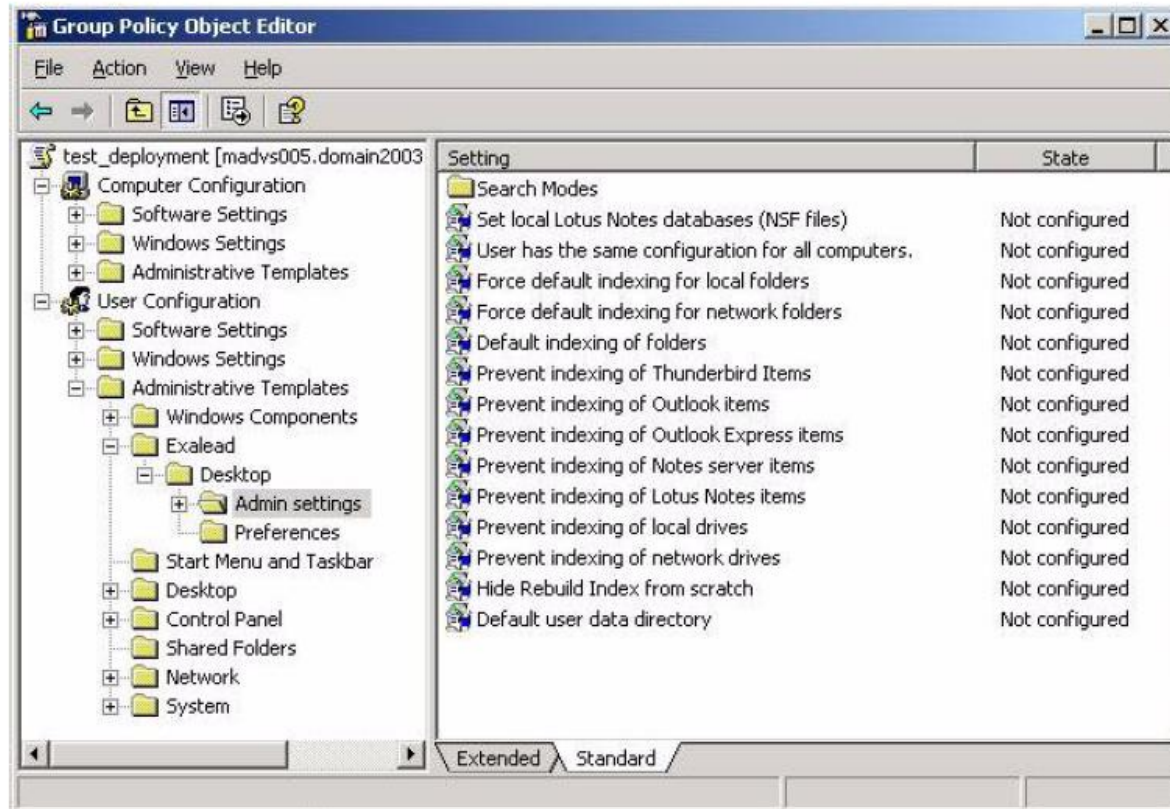


Figure 1 - Example of the Group Policy Object Editor for Windows 2003 Server

# Group Policy Management



## Using GPO

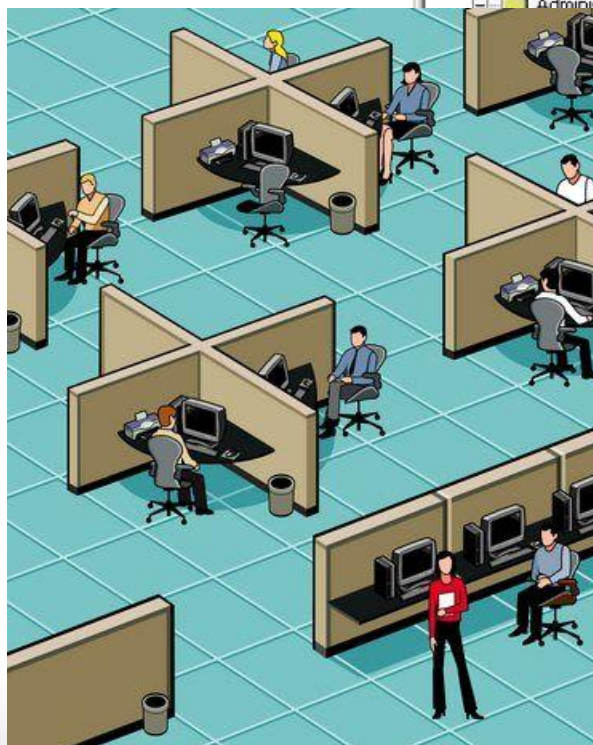
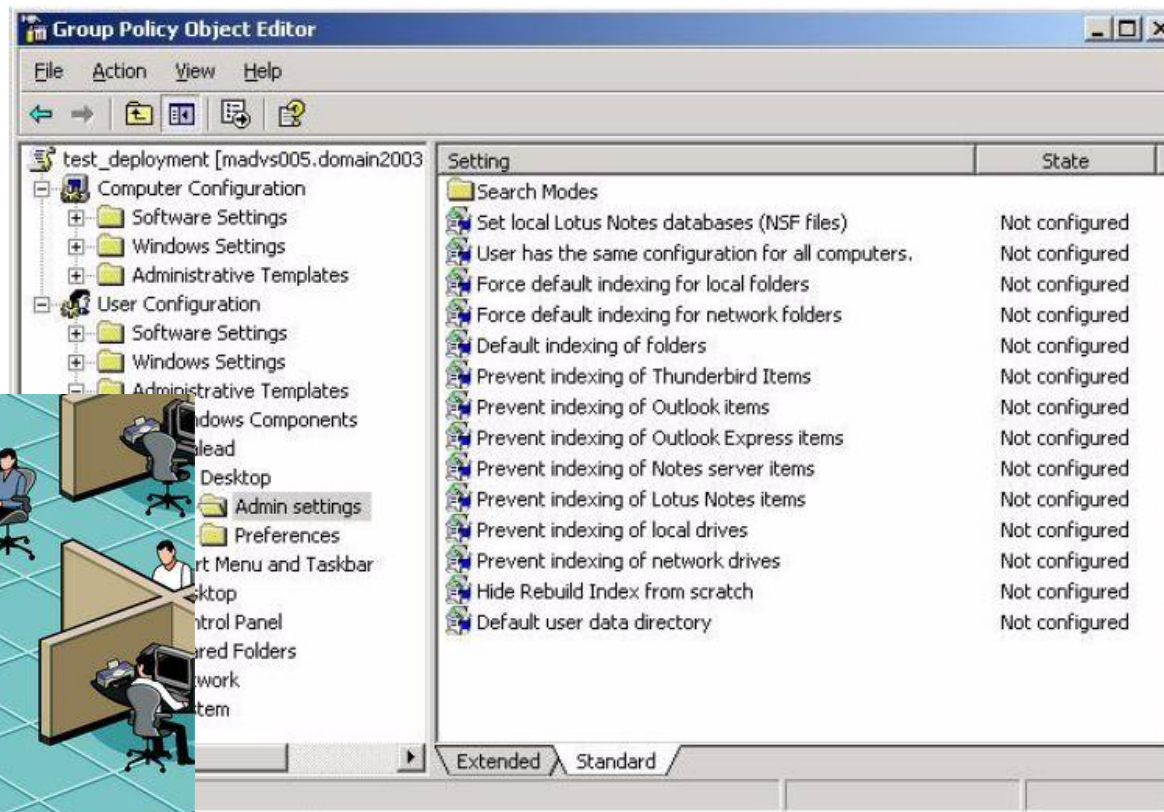


Figure 1.1: Screenshot of the Group Policy Object Editor for Windows 2003 Server



## Select Default Indexing of filesystem folders.

The screenshot shows the Group Policy Management console window titled "Stratégie de groupe". The left pane shows the tree structure under "Stratégie Ordinateur local" > "Configuration Utilisateur" > "Modèles d'administration" > "Exalead" > "Desktop" > "Admin settings". The right pane shows the "Admin settings" folder expanded to "Default indexing of filesystem folders".

**Default indexing of filesystem folders**

Afficher les [Propriétés](#).

Description :  
A list of filesystem paths to be indexed.  
Four values are supported index, exclude, force\_index, force\_exclude

Paramètre	État
Search Modes	
Set local Lotus Notes databases (NSF files)	Activé
Prevent the configuration wizard to appear on first start	Activé
User has the same configuration for all computers.	Non configuré
<b>Default indexing of filesystem folders</b>	<b>Désactivé</b>
Prevent indexing of Thunderbird Items	Non configuré
Prevent indexing of Outlook items	Non configuré
Prevent indexing of Outlook Express items	Non configuré
Prevent indexing of Lotus Notes server databases	Non configuré
Prevent indexing of Lotus Notes	Non configuré
Prevent indexing of local drives	Non configuré
Prevent indexing of network drives	Non configuré
Hide Rebuild Index from scratch	Non configuré
Default user data directory	Non configuré
Indexing allowed start time	Non configuré
Indexing allowed end time	Non configuré
Maximum number of digits	Non configuré
Maximum number of consonants	Non configuré





Click on Add to add a new policy.

Affichage du sommaire

Default indexing of filesystem folders

Nom de valeur	Valeur
C:\dell	index
C:\pgsql\doc	force_index
C:\pgsql\doc\pljava\deploy	force_exclude
C:\pgsql\doc\pljava\deploy\resources	force_index
C:\pgsql\doc\pljava\pljava\	do_not_force:exclude

OK

Annuler

Ajouter...

Supprimer



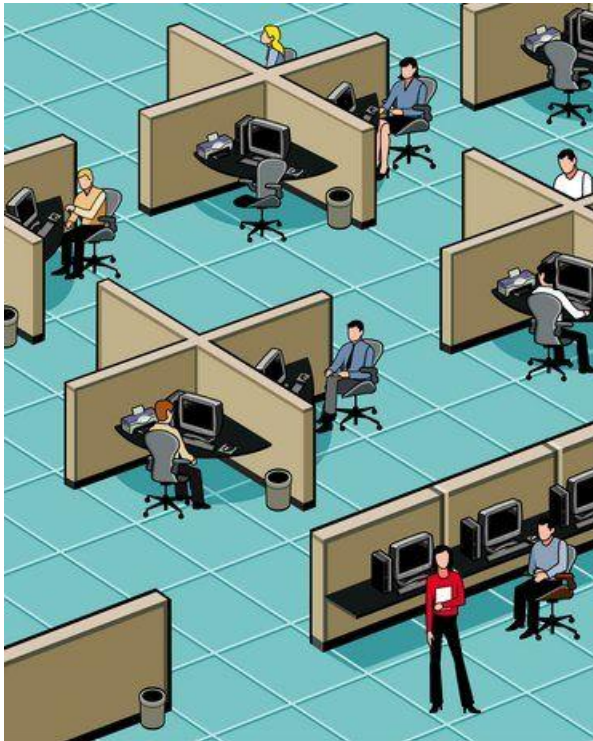
*An integrated desktop search allows user some freedom and central authority some authority*



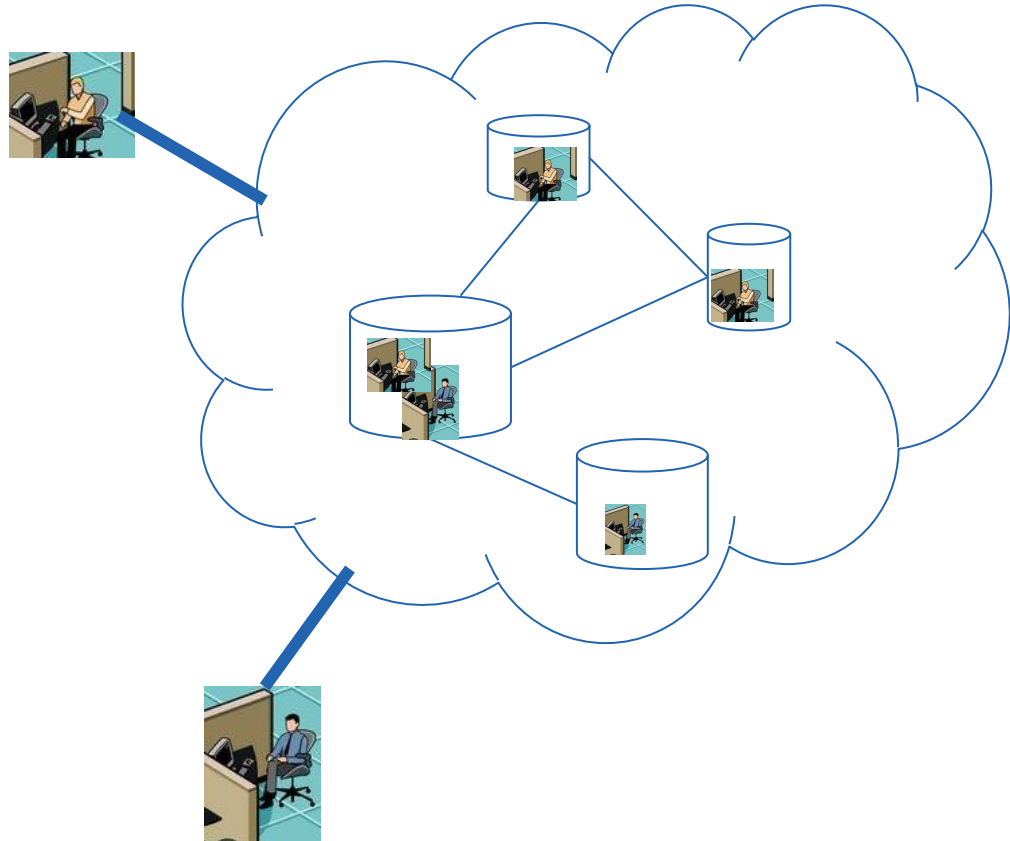
DO WE NEED DESKTOP SEARCH?

ISN'T EVERYTHING ONLINE NOW?

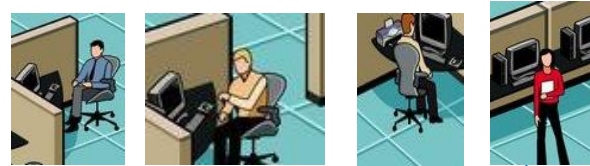
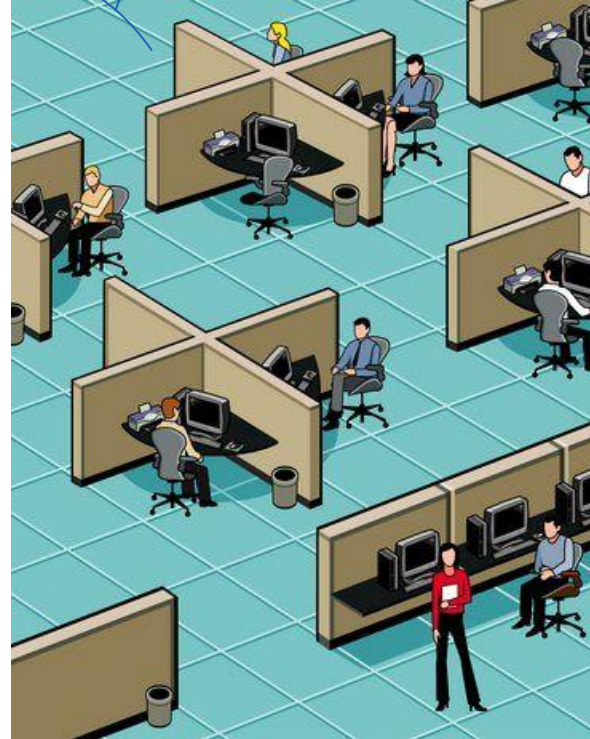
# Alternatives to Desktop: The Cloud



<http://picasaweb.google.com/drickaperilo>



# Alternatives to Desktop: Store Desktop in Enterprise





- › Merging local data (desktop) with global (enterprise) data
  - Ranking
  - Term weighting
  - Document version primacy
  - Dictionaries, global semantics
  - Which « world view » wins?
  
- › Capturing slice of enterprise when « really » offline
  
- › Saving Desktop in « Cloud »
  
- › Mashup: inclusion of desktop data in search based applications



## Desktop search provides

- Indexing of local files (local to user)
- Structure to user's file system
- Access to information without regard to filetype
- Shopping-like access to user's local files
- User freedom to organize data outside of enterprise view
- Look and feel of enterprise search

## Best desktop search integrates into enterprise data

## Remaining challenges

- Federating information from two world views
- Making local information available in mashups



› Thank you