P2P Web Search: Make It Light, Make It Fly

Matthias Bender, Tom Crecelius,
Sebastian Michel, Josiane Xavier Parreira

Max-Planck-Institut für Informatik
Saarbrücken, Germany

Jan 8, 2007

Outline

1. Motivation
2. MINERVA System Model
3. Demo

Motivation

Potential of peer-to-peer (P2P) systems

- scalable
- efficient
- resilient to failures and dynamics

P2P Web Search

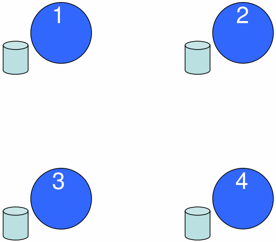
- benefit from intellectual input of user community
- prevent information-resource monopolies

Key Challenge: Make it usable

- Bingo! (Focused Web crawler), MINERVA (P2P Web Search), Cloudscape (Database Backend)
  ⇒ Easy user deployment, common GUI

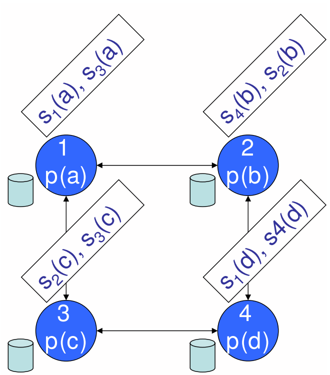
Architectural Model

- Peers autonomously and independently crawl the web
- build local indexes
- can execute queries on local indexes



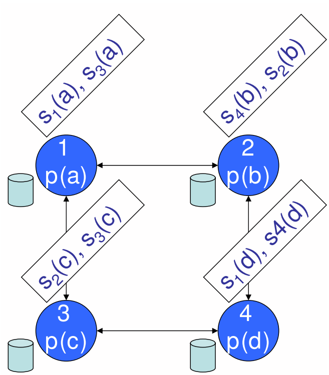
Architectural Model

- share **per-key statistics** about local indexes
- form conceptually global **directory**
- Each directory peer responsible for randomized subset of keys

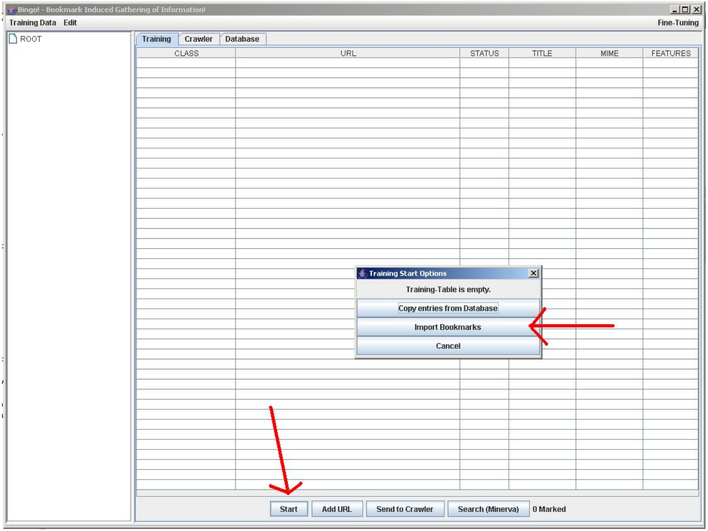


Architectural Model

- use directory to identify promising peers (**Query Routing**)
- send query to selected peers
- merge results appropriately



1. Step: Import Bookmark File



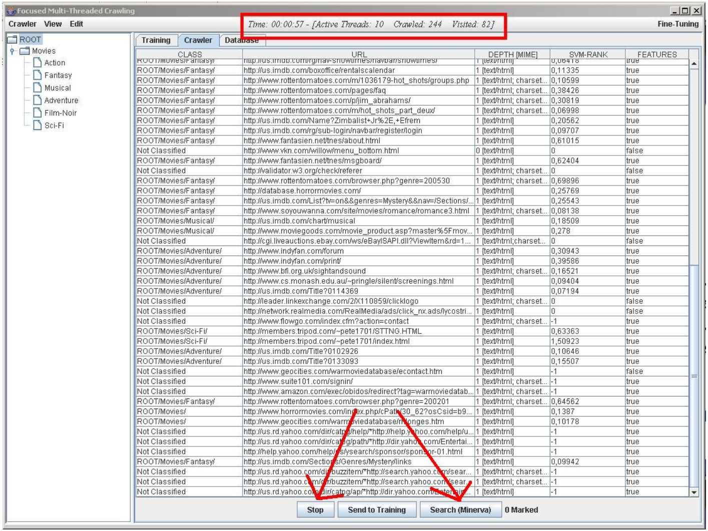
Bingo! - Bookmark Induced Gathering of Information!

Training Start Options

Training-Table is empty.

- Click Start button to import bookmark file

2. Step: Crawl the Web

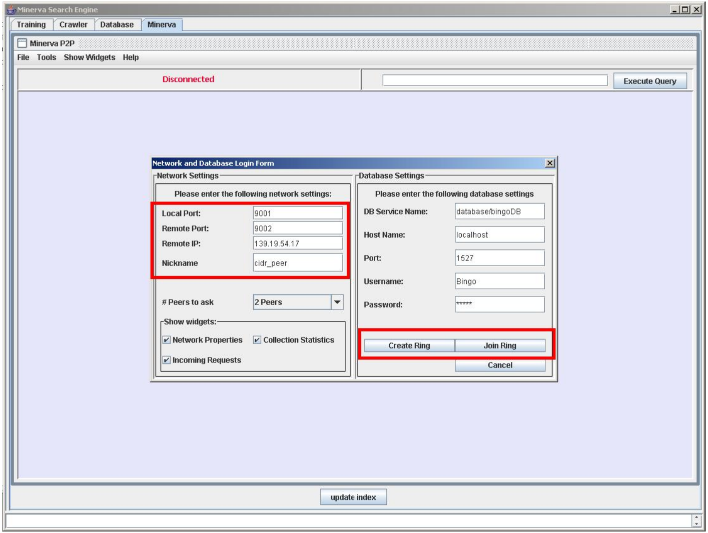


Time: 00:00:57 - [Active Threads: 10 Crawled: 244 Visited: 82]

CLASS	URL	DEPTH [MIME]	SVM-RANK	FEATURES
ROOT/Movies/Fantasy/	http://us.imdb.com/rg/nav-showtimes/navbar/showtimes/	1 [text/html]	0,06418	true
ROOT/Movies/Fantasy/	http://us.imdb.com/boxoffice/rentalscalendar	1 [text/html]	0,11335	true
ROOT/Movies/Fantasy/	http://www.rottentomatoes.com/m/1036179-hot_shots/groups.php	1 [text/html; charset...	0,10599	true
ROOT/Movies/Fantasy/	http://www.rottentomatoes.com/pages/faq	1 [text/html; charset...	0,38426	true
ROOT/Movies/Fantasy/	http://www.rottentomatoes.com/p/jim_abrahams/	1 [text/html; charset...	0,30819	true
ROOT/Movies/Fantasy/	http://www.rottentomatoes.com/m/hot_shots_part_deux/	1 [text/html; charset...	0,06998	true
ROOT/Movies/Fantasy/	http://us.imdb.com/Name?Zimbalist+Jr%2E,+Efrem	1 [text/html]	0,20562	true
ROOT/Movies/Fantasy/	http://us.imdb.com/rg/sub-login/navbar/register/login	1 [text/html]	0,09707	true
ROOT/Movies/Fantasy/	http://www.fantasien.net/tnes/about.html	1 [text/html]	0,61015	true
Not Classified	http://www.vkn.com/willow/menu_bottom.html	0 [text/html]	0	false
ROOT/Movies/Fantasy/	http://www.fantasien.net/tnes/msgboard/	1 [text/html]	0,62404	true
Not Classified	http://validator.w3.org/check/referer	1 [text/html; charset...	0	false
ROOT/Movies/Fantasy/	http://www.rottentomatoes.com/browser.php?genre=200530	1 [text/html; charset...	0,69896	true
ROOT/Movies/Fantasy/	http://database.horrormovies.com/	1 [text/html]	0,25769	true
ROOT/Movies/Fantasy/	http://us.imdb.com/List?tv=on&&genres=Mystery&&nav=/Sections/...	1 [text/html]	0,25543	true
ROOT/Movies/Fantasy/	http://www.soyouwanna.com/site/movies/romance/romance3.html	1 [text/html; charset...	0,08138	true
ROOT/Movies/Musical/	http://us.imdb.com/chart/musical	1 [text/html]	0,18509	true
ROOT/Movies/Musical/	http://www.moviegoods.com/movie_product.asp?master%5Fmov...	1 [text/html]	0,278	true
Not Classified	http://cgi.liveauctions.ebay.com/ws/eBayISAPI.dll?ViewItem&rd=1...	1 [text/html;charset...	0	false
ROOT/Movies/Adventure/	http://www.indyfan.com/forum	1 [text/html]	0,30943	true
ROOT/Movies/Adventure/	http://www.indyfan.com/print/	1 [text/html]	0,39586	true
ROOT/Movies/Adventure/	http://www.bfi.org.uk/sightandsound	1 [text/html; charset...	0,16521	true
ROOT/Movies/Adventure/	http://www.cs.monash.edu.au/~pringle/silent/screenings.html	1 [text/html]	0,09404	true
ROOT/Movies/Adventure/	http://us.imdb.com/Title?0114369	1 [text/html]	0,07194	true
Not Classified	http://leader.linkexchange.com/2/X110859/clicklogo	1 [text/html; charset...	0	false
Not Classified	http://network.realmedia.com/RealMedia/ads/click_nx.ads/lycostri...	1 [text/html]	0	false
Not Classified	http://www.flowgo.com/index.cfm?action=contact	1 [text/html; charset...	-1	true
ROOT/Movies/Sci-Fi/	http://members.tripod.com/~pete1701/STTNG.HTML	1 [text/html]	0,63363	true
ROOT/Movies/Sci-Fi/	http://members.tripod.com/~pete1701/index.html	1 [text/html]	1,50923	true
ROOT/Movies/Adventure/	http://us.imdb.com/Title?0102926	1 [text/html]	0,10646	true
ROOT/Movies/Adventure/	http://us.imdb.com/Title?0133093	1 [text/html]	0,15507	true
Not Classified	http://www.geocities.com/warmoviedatabase/econtact.htm	1 [text/html]	-1	false
Not Classified	http://www.suite101.com/signin/	1 [text/html; charset...	-1	true
Not Classified	http://www.amazon.com/exec/obidos/redirect?tag=warmoviedatab...	1 [text/html; charset...	-1	true
ROOT/Movies/Fantasy/	http://www.rottentomatoes.com/browser.php?genre=200201	1 [text/html; charset...	0,64562	true
ROOT/Movies/	http://www.horrormovies.com/index.php/cPath/30_62?osCsid=b9...	1 [text/html]	0,1387	true
ROOT/Movies/	http://www.geocities.com/warmoviedatabase/mlonges.htm	1 [text/html]	0,10178	true
Not Classified	http://us.rd.yahoo.com/dir/catog/help/*http://help.yahoo.com/help/u...	1 [text/html]	-1	true
Not Classified	http://us.rd.yahoo.com/dir/catog/path/*http://dir.yahoo.com/Entertai...	1 [text/html]	-1	true
Not Classified	http://help.yahoo.com/help/us/ysearch/sponsor/sponsor-01.html	1 [text/html]	-1	true
ROOT/Movies/Fantasy/	http://us.imdb.com/Sections/Genres/Mystery/links	1 [text/html]	0,09942	true
Not Classified	http://us.rd.yahoo.com/dir/buzzitem/*http://search.yahoo.com/sear...	1 [text/html; charset...	-1	true
Not Classified	http://us.rd.yahoo.com/dir/buzzitem/*http://search.yahoo.com/sear...	1 [text/html; charset...	-1	true
Not Classified	http://us.rd.yahoo.com/dir/catog/ap/*http://dir.yahoo.com/Entertain...	1 [text/html]	-1	true

- BINGO! fetches bookmark documents, builds classifier
- Web crawl starts automatically, can be stopped anytime
- Start Minerva

3. Step: Instantiate Minerva



Disconnected

Network and Database Login Form

Network Settings

Please enter the following network settings:

Local Port:	9001
Remote Port:	9002
Remote IP:	139.19.54.17
Nickname	cidr_peer
# Peers to ask	2 Peers

Show widgets:

☑ Network Properties ☑ Collection Statistics

☑ Incoming Requests

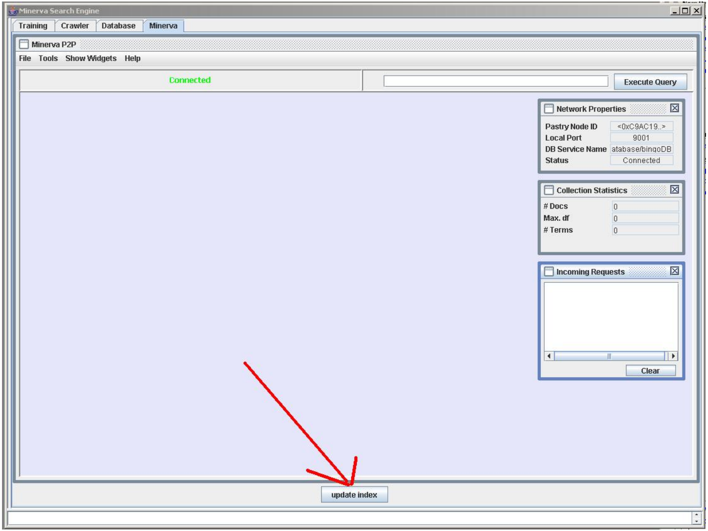
Database Settings

Please enter the following database settings

DB Service Name:	database/bingoDB
Host Name:	localhost
Port:	1527
Username:	Bingo
Password:	*****

- Enter nickname, network settings
- Click Create Ring (or Join Ring, if exists)

4. Step: Update Minerva Index



Connected

Network Properties

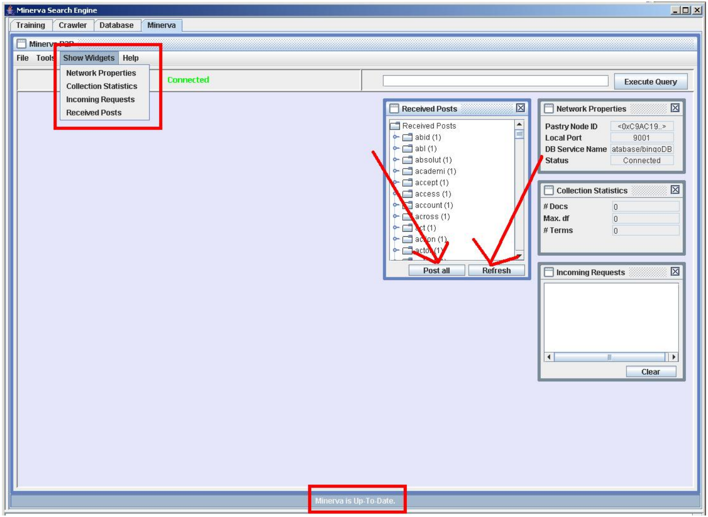
Pastry Node ID	<0xC9AC19..>
Local Port	9001
DB Service Name	atabase/bingoDB
Status	Connected

Collection Statistics

# Docs	0
Max. df	0
# Terms	0

- *Update Index* imports current data from Bingo!, computes metadata

5. Step: Publish Metadata to directory



Connected

Received Posts

- Received Posts
  - abid (1)
  - abl (1)
  - absolut (1)
  - academi (1)
  - accept (1)
  - access (1)
  - account (1)
  - across (1)
  - act (1)
  - action (1)
  - actor (1)

Network Properties

Pastry Node ID	<0xC9AC19..>
Local Port	9001
DB Service Name	atabase/bingoDB
Status	Connected

Collection Statistics

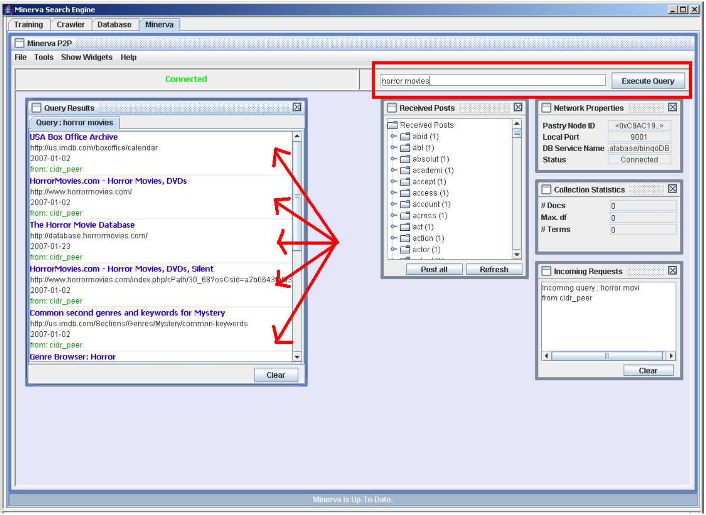
# Docs	0
Max. df	0
# Terms	0

Incoming Requests

Minerva is Up-To-Date.

- Use *Show Widgets* menu to open Received-Posts panel
- Click Post All button
- Click Refresh button to inspect directory

6. Step: Execute Query



Connected

horror movies

Query : horror movies

USA Box Office Archive
http://us.imdb.com/boxoffice/calendar
2007-01-02
from: cidr_peer

HorrorMovies.com - Horror Movies, DVDs
http://www.horrormovies.com/
2007-01-02
from: cidr_peer

The Horror Movie Database
http://database.horrormovies.com/
2007-01-23
from: cidr_peer

HorrorMovies.com - Horror Movies, DVDs, Silent
http://www.horrormovies.com/index.php/cPath/30_68?osCsid=a2b0643[illegible]
2007-01-02
from: cidr_peer

Common second genres and keywords for Mystery
http://us.imdb.com/Sections/Genres/Mystery/common-keywords
2007-01-02
from: cidr_peer

Genre Browser: Horror

Received Posts
- abid (1)
- abl (1)
- absolut (1)
- academi (1)
- accept (1)
- access (1)
- account (1)
- across (1)
- act (1)
- action (1)
- actor (1)

Pastry Node ID	<0xC9AC19..>
Local Port	9001
DB Service Name	atabase/bingoDB
Status	Connected

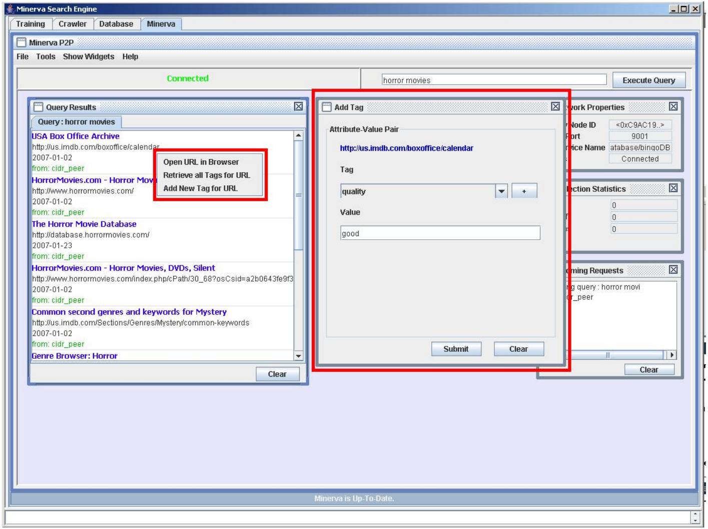
# Docs	0
Max. df	0
# Terms	0

Incoming query : horror movi
from cidr_peer

Minerva is Up-To-Date.

- Enter Query, click Execute query
- Inspect Results

7. Step: Tag Documents



Connected

horror movies

Query : horror movies

USA Box Office Archive
http://us.imdb.com/boxoffice/calendar
2007-01-02
from: cidr_peer

HorrorMovies.com - Horror Mov
http://www.horrormovies.com/
2007-01-02
from: cidr_peer

The Horror Movie Database
http://database.horrormovies.com/
2007-01-23
from: cidr_peer

HorrorMovies.com - Horror Movies, DVDs, Silent
http://www.horrormovies.com/index.php/cPath/30_68?osCsid=a2b0643fe9f3
2007-01-02
from: cidr_peer

Common second genres and keywords for Mystery
http://us.imdb.com/Sections/Genres/Mystery/common-keywords
2007-01-02
from: cidr_peer

Genre Browser: Horror

http://us.imdb.com/boxoffice/calendar

Tag: quality

Value: good

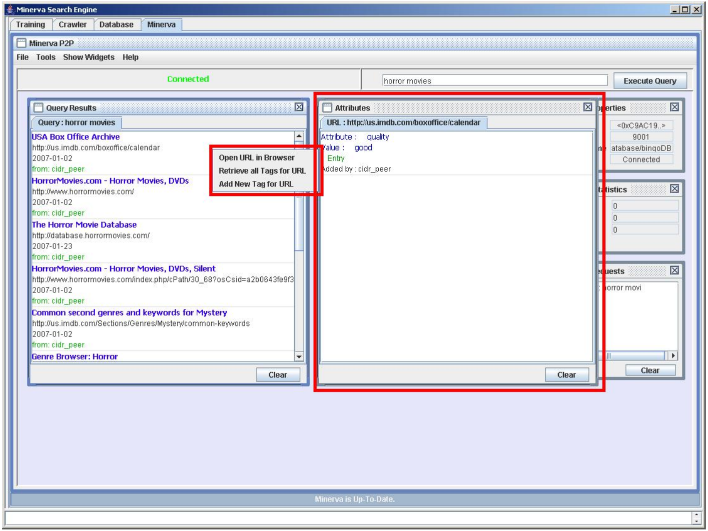
Node ID: <0xC9AC19..>
ort: 9001
ice Name: atabase/bingoDB
Connected

0
0
0

g query : horror movi
r_peer

- Right-click on document, *Add New Tag for URL*
- Add desired tag(s)

8. Step: Retrieve Annotations for Document



USA Box Office Archive
http://us.imdb.com/boxoffice/calendar
2007-01-02
from: cidr_peer

HorrorMovies.com - Horror Movies, DVDs
http://www.horrormovies.com/
2007-01-02
from: cidr_peer

The Horror Movie Database
http://database.horrormovies.com/
2007-01-23
from: cidr_peer

HorrorMovies.com - Horror Movies, DVDs, Silent
http://www.horrormovies.com/index.php/cPath/30_68?osCsid=a2b0643fe9f3
2007-01-02
from: cidr_peer

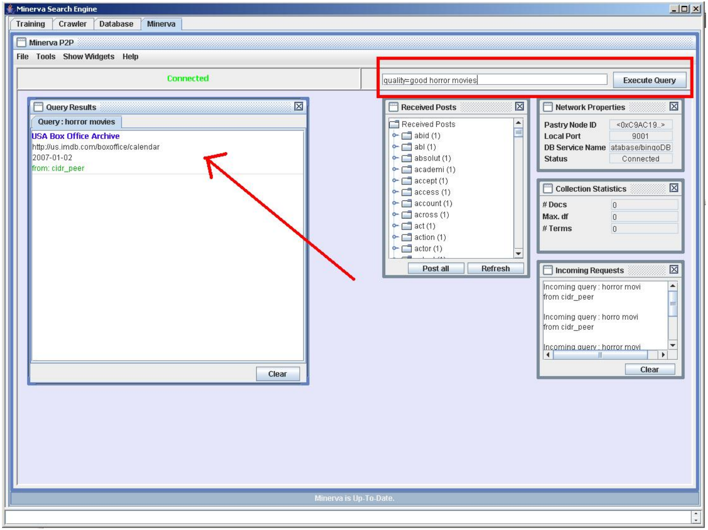
Common second genres and keywords for Mystery
http://us.imdb.com/Sections/Genres/Mystery/common-keywords
2007-01-02
from: cidr_peer

Genre Browser: Horror

Attribute : quality
Value : good
Entry
Added by : cidr_peer

- Right-click on document, *Retrieve all Tags for URL*
- Inspected submitted tags

9. Step: Query for Annotated Documents



Connected

quality=good horror movies

Query : horror movies

USA Box Office Archive
http://us.imdb.com/boxoffice/calendar
2007-01-02
from: cidr_peer

Received Posts
abid (1)
abl (1)
absolut (1)
academi (1)
accept (1)
access (1)
account (1)
across (1)
act (1)
action (1)
actor (1)

Pastry Node ID	<0xC9AC19..>
Local Port	9001
DB Service Name	atabase/bingoDB
Status	Connected

# Docs	0
Max. df	0
# Terms	0

Incoming query : horror movi
from cidr_peer

Incoming query : horro movi
from cidr_peer

Incoming query : horror movi

Minerva is Up-To-Date.

- Enter query as *tag=value*, possibly combine with keywords
- Inspect results

Thank you for your attention

Download and more details available at:

http://www.minerva-project.org