Using a Society of Agents To Interpret Natural Language

John F. Sowa and Arun K. Majumdar
VivoMind Intelligence, Inc.

9 October 2008

VivoMind Approach to General AI

Most of the usual AI technology with some novel combinations:

*** Society of Mind: Multiple heterogeneous agents that operate in parallel by message passing, associative blackboards, and voting mechanisms to resolve conflicts.**

*** Discrete representations: Conceptual graphs to represent Common Logic with some extensions and approximations.**

*** Continuous representations: Mapping graphs to continuous fields with high-speed floating-point computation of semantic distance measures.**

*** Analogical reasoning: Using the semantic distance measures to find analogies in logarithmic time instead of polynomial time.**

*** Learning mechanisms: Systematic ways of reorganizing the graphs, the society of agents, the communication pathways, and the space-time resources.**

Conceptual Graphs

Every language, natural or artificial, uses three mechanisms for encoding knowledge:

1. **Labels: Words, names, concepts, relations, predicates, variables...**

2. **Graphs: Collections of nodes interconnected in strings, trees, and networks. (Trees and strings usually have cross references.)**

3. **Contexts: Grouping collections of graphs.**

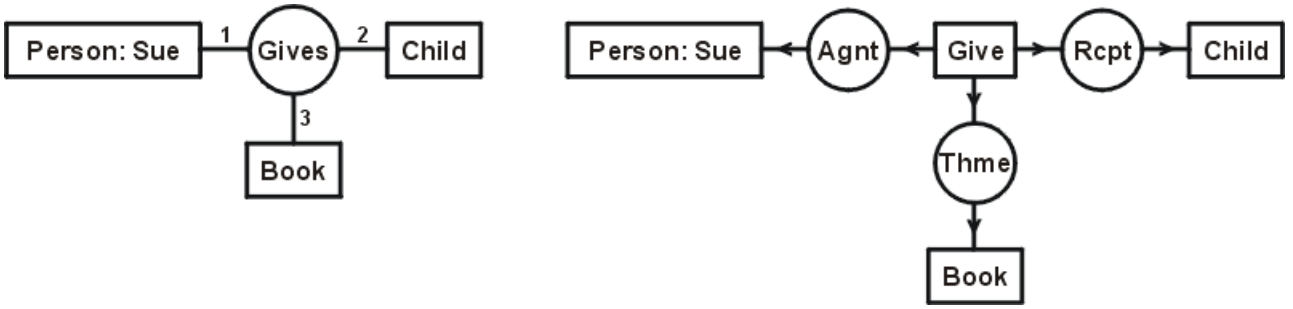
Conceptual graphs are a formal system of logic based on these three mechanisms.

The CG notation is based on the existential graphs by C. S. Peirce and the semantic networks of artificial intelligence.

Conceptual Graph Interchange Format (CGIF) is one of three dialects in the ISO/IEC 24707 standard for Common Logic.

Different Representations for the Same Sentence

Sentence: Sue gives a child a book.



Same topology, but different geometry and ontology.

Linear notations for logic obscure the structure:

$(\exists x)(\exists y)(\text{Person}(\text{Sue}) \wedge \text{Child}(x) \wedge \text{Book}(y) \wedge \text{Gives}(\text{Sue},x,y))$.

$(\exists x)(\exists y)(\exists z)(\text{Person}(\text{Sue}) \wedge \text{Child}(x) \wedge \text{Book}(y) \wedge \text{Give}(z) \wedge \text{Agnt}(z,\text{Sue}) \wedge \text{Rcpt}(z,x) \wedge \text{Thme}(z,y))$.

Discrete and Continuous Representations

The physical world is continuous, but language and other cognitive systems are based on discrete words, percepts, and concepts.

1967: *A Unified Theory of Human Behavior* by Kenneth Pike.

Analyzed the continuous and discrete aspects of every mode of perception, cognition, and behavior.

1977: Catastrophe-theoretic semantics by René Thom.

Showed how continuous mathematical functions can make a "catastrophic," nearly discontinuous, shift over a narrow region.

Applied these ideas to the discrete and continuous aspects of language.

1982 to present: Cognitive linguistics by Wolfgang Wildgen:

Elaborated Thom's catastrophe-theoretic semantics and integrated it with other cognitive approaches to language.

Fundamental insight: Discrete aspects of cognition are based on more primitive continuous mechanisms.

Discrete and Continuous Computation

Discrete methods of knowledge representation:

* **All knowledge is symbolic.**
* **Deduction is the primary method of reasoning.**
* **Data types are strings, lists, and graphs of symbols.**

Continuous, presymbolic methods of robotics:

* **Geometry of space and time is fundamental.**
* **Reasoning is numeric computation on geometric models.**
* **Data types are continuous functions on geometric spaces.**

Hypothesis suggested by evolutionary evidence:

* **Presymbolic methods are the foundation for symbolic methods.**

Mathematical Background

Blum, Cucker, Shub, Smale (1998):

* **Efficiency of many AI algorithms is limited by the Boolean values {0,1}.**

* **Numeric computation over the real numbers can often reduce complexity.**

Rvachev (1963):

* **A remarkable family of continuous functions with associated Boolean functions.**

* **Widely used in computational geometry.**

Searching, finding, indexing, and relating chemical graphs:

* **Technology developed to classify, index, and retrieve chemical graphs in logarithmic time can be applied to conceptual graphs.**

* **Chemical graph indices used for drug design and structure-activity relations.**

* **Graphs can be related to topology, continuous fields, and eigenvalues.**

Finding Analogies

Research by Falkenhainer, Forbus, & Gentner:

Problem: In a knowledge base of N graphs, the time to find the best analogy to a graph G takes time proportional to N3.

If N is 10, N3 is a thousand.
But if N is a billion, N3 is 1,000,000,000,000,000,000,000,000,000.

MAC/FAC Solution:

Use a search method to narrow down the likely candidates from some large number N to a much, much, much smaller number n.

With a good search method:

*** The execution time for the search should be proportional to log(N).**

*** It should find some small number of graphs, n.**

*** And the most relevant graphs should be among those n.**

Question: What kinds of search algorithms have these properties?

Cognitive Signature™

An encoding of a conceptual graph with the following properties:

* Encodes the ontology of the concept and relation types.

* Encodes the topology of the graphs — i.e., structural properties such as branching, cycles, and connectivity.

* Independent of additions of new types to the ontology.

* Determines a measure of the semantic distance between graphs.

* Suitable for indexing the graphs in time proportional to

  — N×log(N) for building the index.
  — log(N) for finding a graph.

* Able to locate all graphs within a small semantic distance ε from any graph G.

Enables analogy finding in logarithmic time.

Criticisms of Logical Deduction

Deduction in mathematics can be precise, but deduction about any empirical subject must depend on prior induction, which is almost always incomplete.

Criticism by the physician, Sextus Empiricus (2nd century AD):

> ***Every human is an animal.***
> ***Socrates is human.***
> ***Therefore, Socrates is an animal.***

**If the major premise was derived by checking every human,
then Socrates was considered, and the argument is circular.**

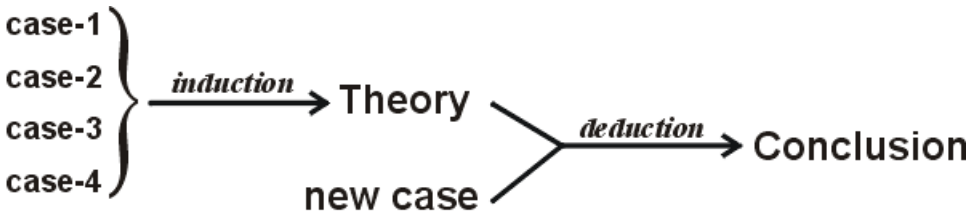
Otherwise, the induction was incomplete, and the conclusion is uncertain.

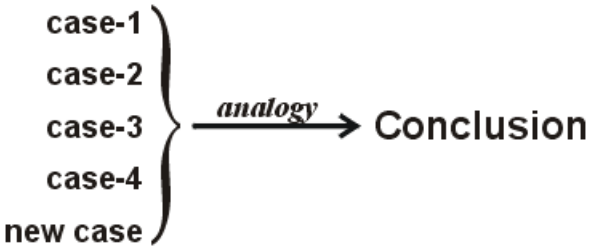
Criticism by the legal scholar, Ibn Taymiyya (14th century AD):

> **Every empirical theory is derived by induction from cases.**
>
> **Any deduction from such a theory can be obtained by applying analogical reasoning to exactly the same cases.**

Case-Based Reasoning

Ibn Taymiyya's comparison of logical and analogical reasoning:





If the same cases are used, analogy can give the same answer in one step that requires two steps by induction followed by deduction.

What Ibn Taymiyya did not recognize: If the same theory could be reused for many different applications, deduction would be more efficient.

An Application of Case-Based Reasoning

A textbook publisher needs to evaluate student answers to math questions.

*** Free-form answers in English sentences.**

*** Much harder to evaluate than multiple choice.**

Typical question:

The following numbers are 1 more than a square: 10, 37, 65, 82.

If you are given an integer N that is less than 200,
how would you determine whether N is 1 more than a square?

Explain your method in three or four sentences.

How could a computer system evaluate student answers?

Determine whether they are correct, incorrect, or partially correct?

And make helpful suggestions about the incorrect answers?

Publisher's Current Procedure

To evaluate new exam questions, the publisher normally gives the exam to a large number of students.

For each question, they get about 50 different answers:

- **Some are completely correct**
  — but stated in different ways.

- **Some are partially correct**
  — and the teacher says what is missing.

- **Others are wrong**
  — in many different ways.

Result: 50 pairs of student answer and teacher evaluation.

Each pair of (answer, evaluation) is a case for case-based reasoning.

Intellitex Parser

VivoMind's older parser is named Intellitex:

*** Uses a rather simple grammar.**

*** Uses several lexical resources, but no formal ontologies.**

*** Generates conceptual graphs as output.**

*** Robust: always generates some CG as its best guess.**

*** Designed to work with the VivoMind Analogy Engine (VAE)**

These properties are important for handling typical student answers, which frequently have poor grammar and incomplete sentences.

Minor errors are not necessarily bad — provided that Intellitex makes the same errors consistently in all cases.

Using VAE to Evaluate Student Answers

Use VAE to compare each new answer to the 50 cases:

1. **For all 50 cases, translate student answer to conceptual graphs.**
2. **Translate each new answer to a new CG.**
3. **Compare the new CG to the 50 CGs for previous answers.**
4. **Use the measure of semantic distance to determine the best match.**
5. **If there is a good match, print out the previous evaluation.**
6. **Otherwise, send the new answer to a teacher to evaluate.**

Result:

VAE found a good match for most of the student answers.

For each good match, the previous teacher's evaluation was appropriate.

When VAE failed to find a good match, the new case could be added to the list of cases in order to improve its coverage.

There was no need for the teachers to write rules or programs.

Logic as a Disciplined Use of Analogies

Analogy is the foundation for every form of reasoning, including logic:

*** Deduction: Every step requires a unification, which is special case of the structure mappings used in analogies.**

*** Induction: Analogies are used to find common generalizations of multiple instances.**

*** Abduction: The operation of guessing or forming an initial hypothesis, called abduction, requires analogies to find likely causes or explanations.**

In both human reasoning and computer implementations, the same underlying operations can support both logical and analogical reasoning.

Mapping Language to a Formal Logic

Difficult or impossible except under tightly controlled conditions:

*** Well-formed grammatical structure.**

*** Consistent use of each word with exactly one word sense (or a small number of predefined word senses that can be predicted by context).**

*** A formal ontology that defines each word sense.**

These conditions hold for specially designed, *controlled* natural languages in a particular Wittgensteinian language game.

They almost never hold for the kind of language used by people in talking and writing to other people.

VivoMind Language Processor (VLP)

A two-level language processor that uses conceptual graphs for both the formal core and open-ended extensions beyond the core.

- *** VLP always generates some conceptual graph for any English sentence.**
- *** But it guarantees a correct interpretation only for sentences stated in Common Logic Controlled English (CLCE).**
- *** For sentences outside CLCE, VLP will generate a translation to CLCE as an echo that the user can verify.**

VLP is currently under development, but it is being designed as a successor to two earlier systems:

1. **Intellitex, which was used for processing language by the VivoMind Analogy Engine (VAE).**
2. **A conventional syntax-directed translator of CLCE to logic.**

Parsing by Multiple Heterogeneous Agents

The VivoMind parser intermixes lexical, syntactic, semantic, and pragmatic processing by multiple agents working in parallel.

Each word may have multiple word senses, each with different lexical and syntactic features, different concept or relation types, and different canonical graphs.

Canonical graphs are introduced and processed in parallel with the lexical and syntactic features.

Domain-dependent pragmatic information is introduced as early as possible to maximize its influence on the choices by syntactic and semantic agents.

Multiple Semantic Resources

Many different ontologies and semantic resources are available on the WWW, and they're not consistent:

WordNet, Roget's Thesaurus, CoreLex, Framenet, Semantic Web, etc.

Some of them, such as the IBM-CSLI Verb Ontology and Beth Levin's verb classes, can be merged.

The others are so different in format and coverage, that it's difficult to merge them and impossible to keep them consistent over time.

Solution:

*** Don't attempt to merge all resources in advance.**

*** Assign them to different agents, which can contribute relevant information during the parsing process.**

*** Use a voting mechanism to resolve conflicts as they occur.**

*** Dynamically change the time and space allocated to agents, depending on the value of their contributions.**

Sample Sentence

From the oil & gas example:

**The Diana field is situated in the western Gulf of Mexico
260 km (160 mi) south of Galveston
in approximately 1430 m (4700 ft) of water.**

From lexical resources:

**entity(1, "Diana field") prep("in the") loc(1, western) loc(2, "Gulf of Mexico")
measure(1, "260 kilometers") measure(2, "160 miles") loc(3,south)
prep("of") loc(4,"Galveston") prep("in") qualifier(approximately)
measure(3,"1430 meters") measure(4,"4700 feet") prep("of") entity(2, water).**

Other agents use heuristics and the EGI ontology for the domain.

For this sentence, the syntax is sufficient to determine that Diana field must be in the western part of the Gulf of Mexico, but other attachments are syntactically ambiguous. To prune away unlikely options, some agents use a domain ontology for geoscience, which includes information about reservoirs, bodies of water, and cities. After pruning, the remaining links indicate that Diana field is south of Galveston and in the water.

Query Processing

For any set of documents:

* **Translate the documents to conceptual graphs.**

* **Map the structured data to conceptual graphs.**

* **Index the CGs in time proportional to (N log N), where N is the total number of nodes in all the graphs.**

For any query consisting of one or more sentences that specify a pattern to be found in the source material:

* **Translate the query to conceptual graphs.**

* **Find matching patterns in the source data and rank them in order of semantic distance. Zero distance means an exact match.**

* **For each match within a given threshold, determine which subgraphs are similar or different.**

* **As answer, return the English phrases in the original documents from which those graphs and subgraphs were derived.**

Application to Oil and Gas Exploration

Source material:

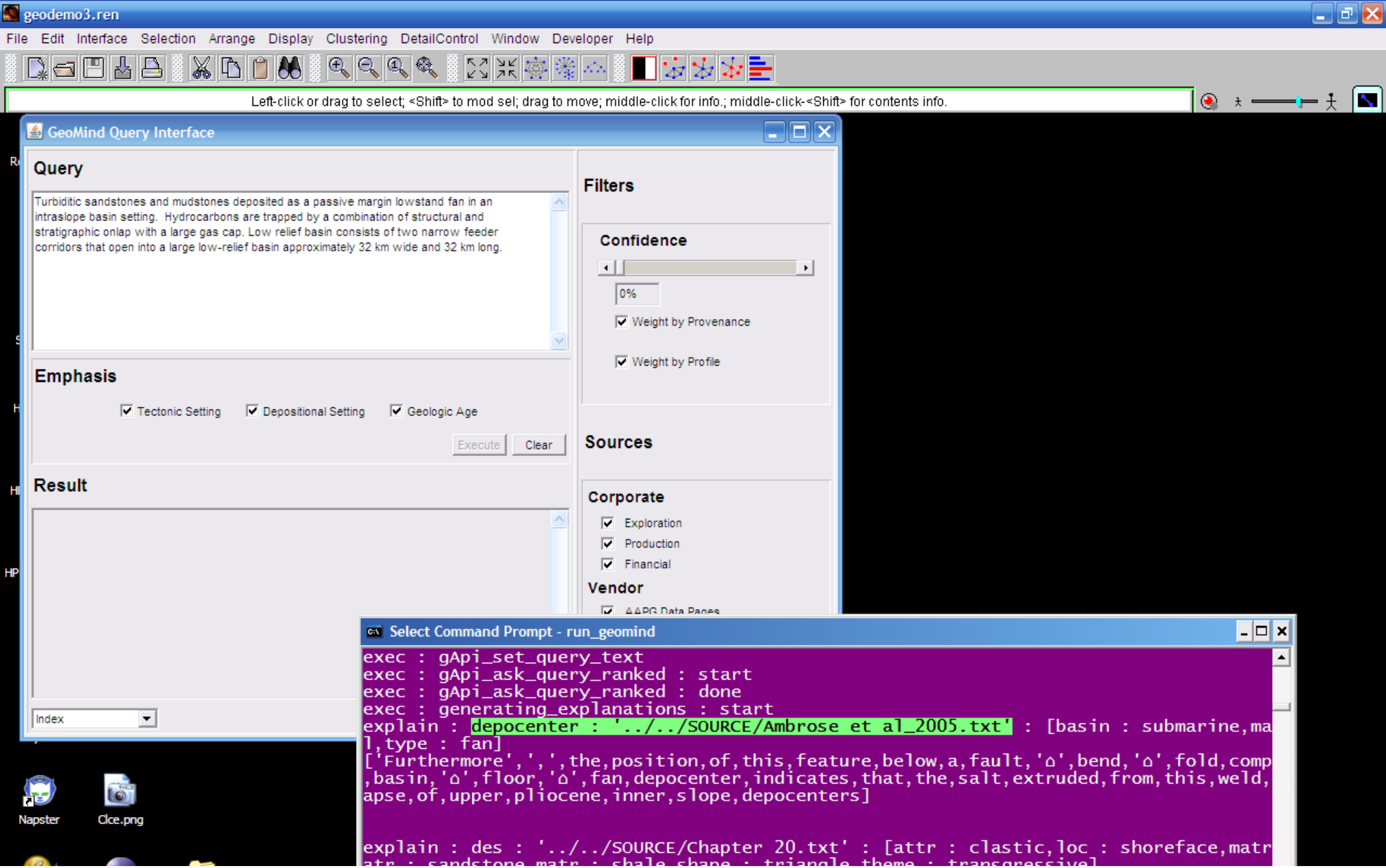
* 79 documents, ranging in length from 1 page to 50 pages.

* English, as written for human readers (no semantic tagging).

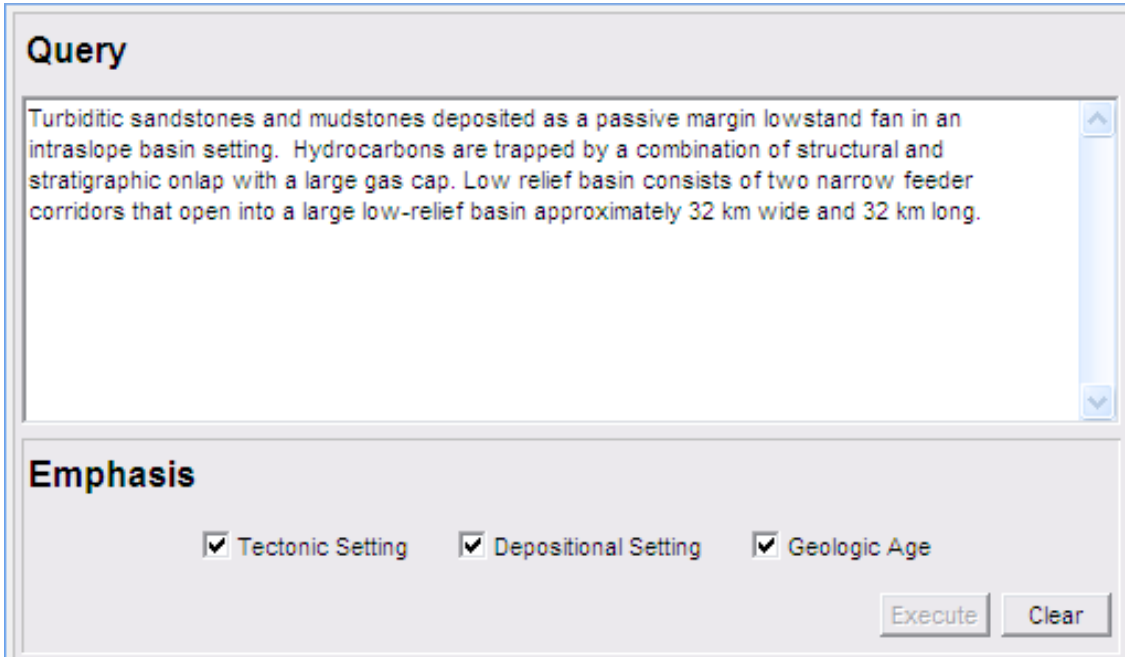
* Additional data from relational DBs and other structured sources.

* Basic VivoMind ontology plus a domain-dependent ontology developed in collaboration with EGI (Energy & Geoscience Institute at the University of Utah).

* Very few detailed axioms in the ontology.

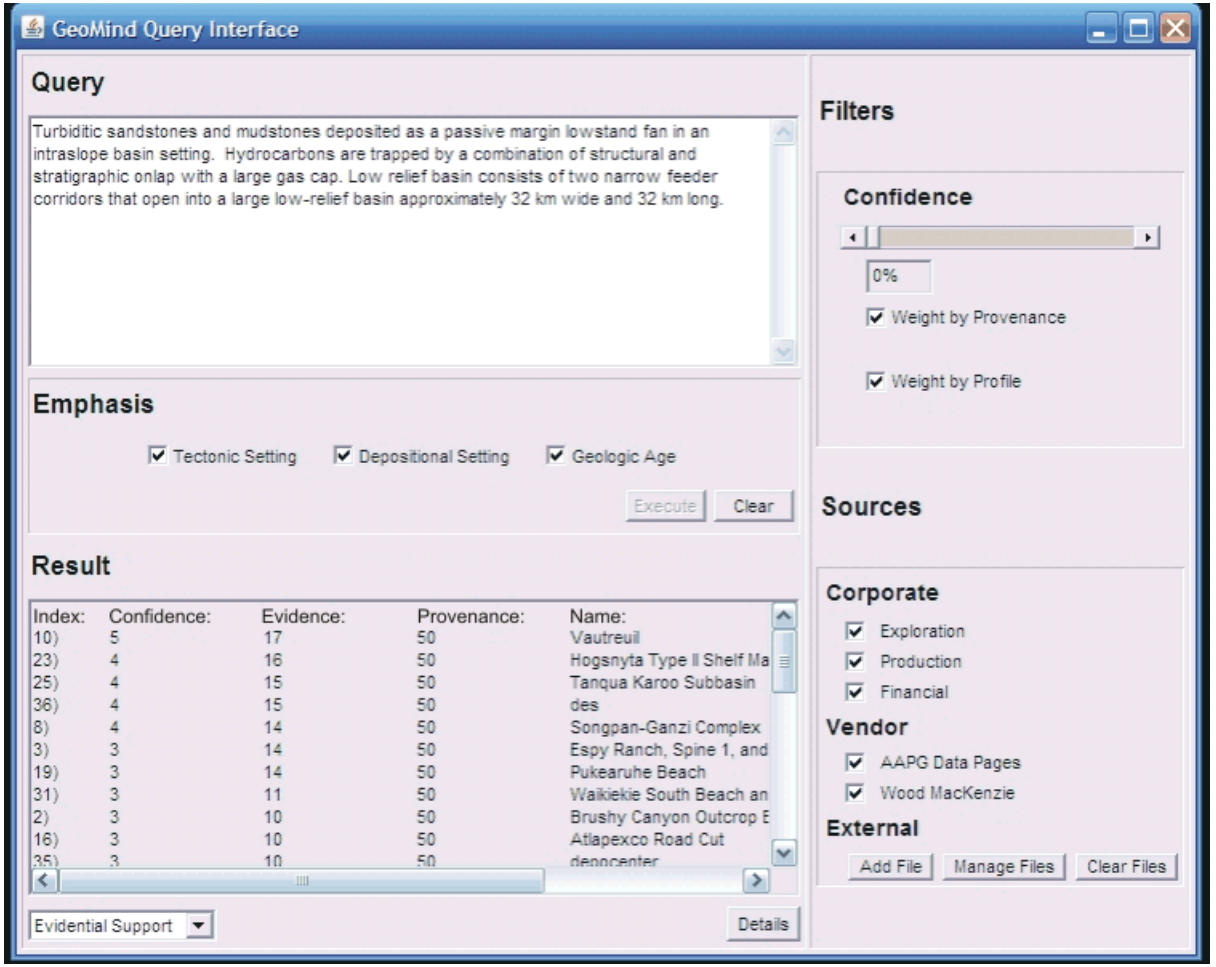


Read documents and extract information relevant to a given query.

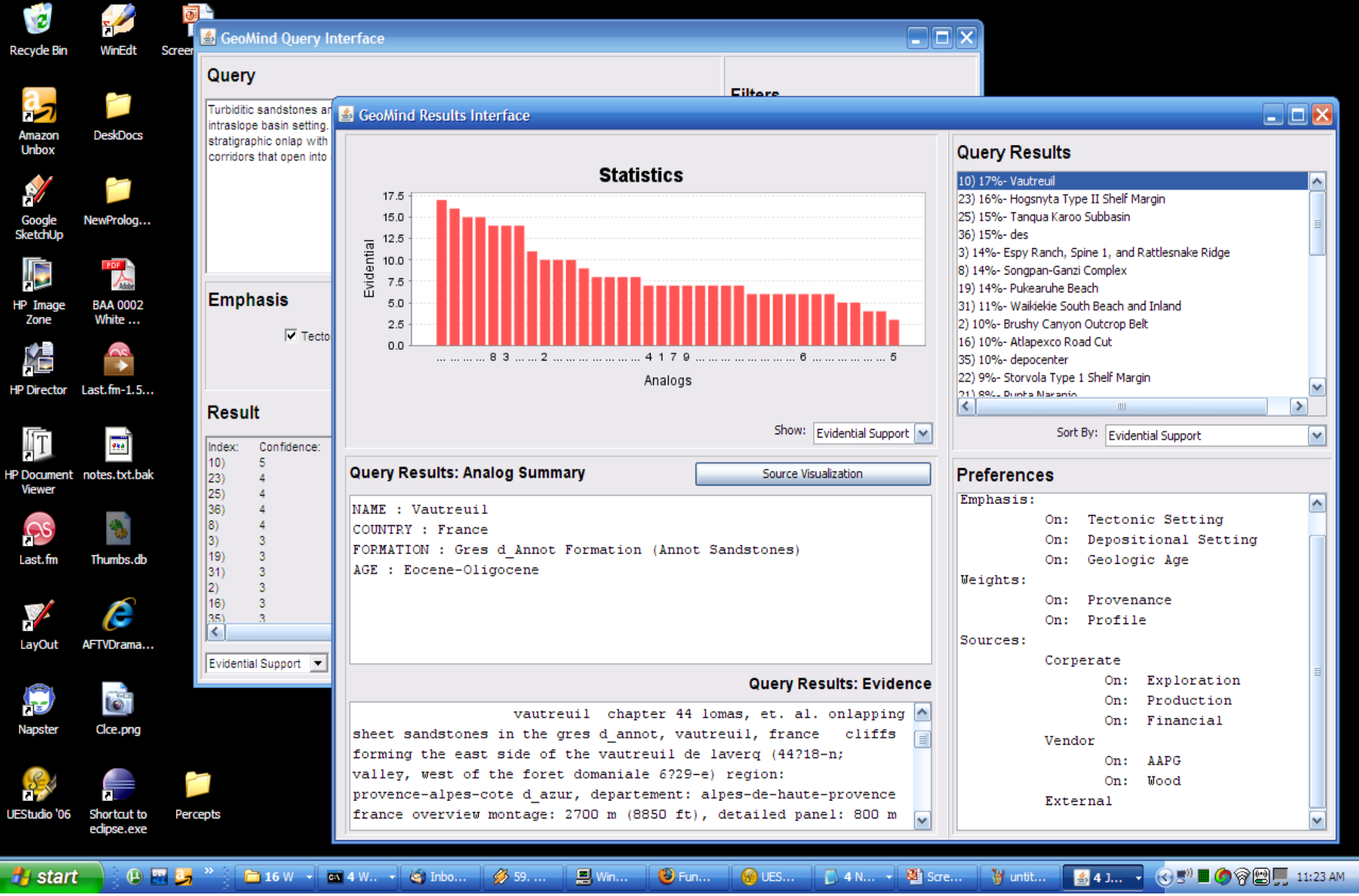


Turbiditic sandstones and mudstones deposited as a passive margin lowstand fan in an intraslope basin setting. Hydrocarbons are trapped by a combination of structural and stratigraphic onlap with a large gas cap. Low relief basin consists of two narrow feeder corridors that open into a large low-relief basin approximately 32 km wide and 32 km long.

THE QUERY



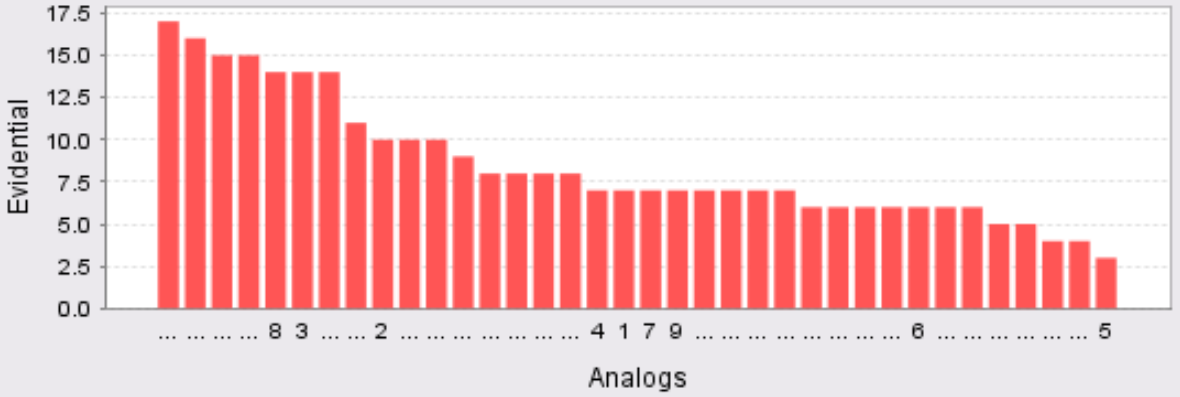
RESULTS, ranked by evidence (Dempster-Shafer) & confidence factors



After clicking the “Details” button on the previous window

GeoMind Results Interface

Statistics



Show: Evidential Support

Query Results

10) 17%- Vautreuil
23) 16%- Hogsnyta Type II Shelf Margin
25) 15%- Tanqua Karoo Subbasin
36) 15%- des
3) 14%- Espy Ranch, Spine 1, and Rattlesnake Ridge
8) 14%- Songpan-Ganzi Complex
19) 14%- Pukearuhe Beach
31) 11%- Waikiekie South Beach and Inland
2) 10%- Brushy Canyon Outcrop Belt
16) 10%- Atlapexco Road Cut
35) 10%- depocenter
22) 9%- Storvola Type 1 Shelf Margin
21) 8%- Punta Naranio

Sort By: Evidential Support

Query Results: Analog Summary

Source Visualization

```
NAME : Vautreuil
COUNTRY : France
FORMATION : Gres d_Annot Formation (Annot Sandstones)
AGE : Eocene-Oligocene
```

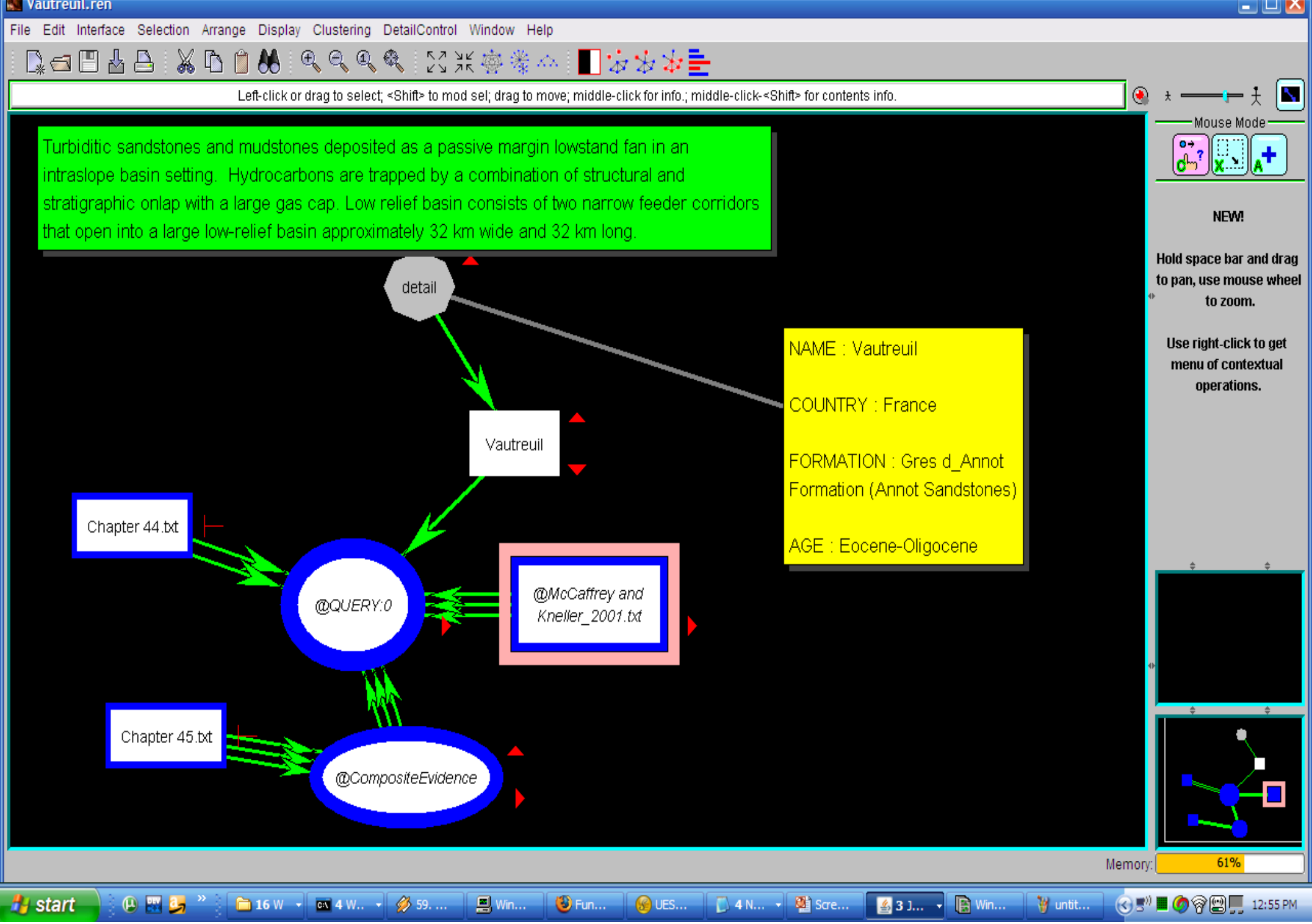
Query Results: Evidence

```
                  vautreuil  chapter 44 lomas, et. al. onlapping
sheet sandstones in the gres d_annot, vautreuil, france   cliffs
forming the east side of the vautreuil de laverq (44?18-n;
valley, west of the foret domaniale 6?29-e) region:
provence-alpes-cote d_azur, departement: alpes-de-haute-provence
france overview montage: 2700 m (8850 ft), detailed panel: 800 m
```

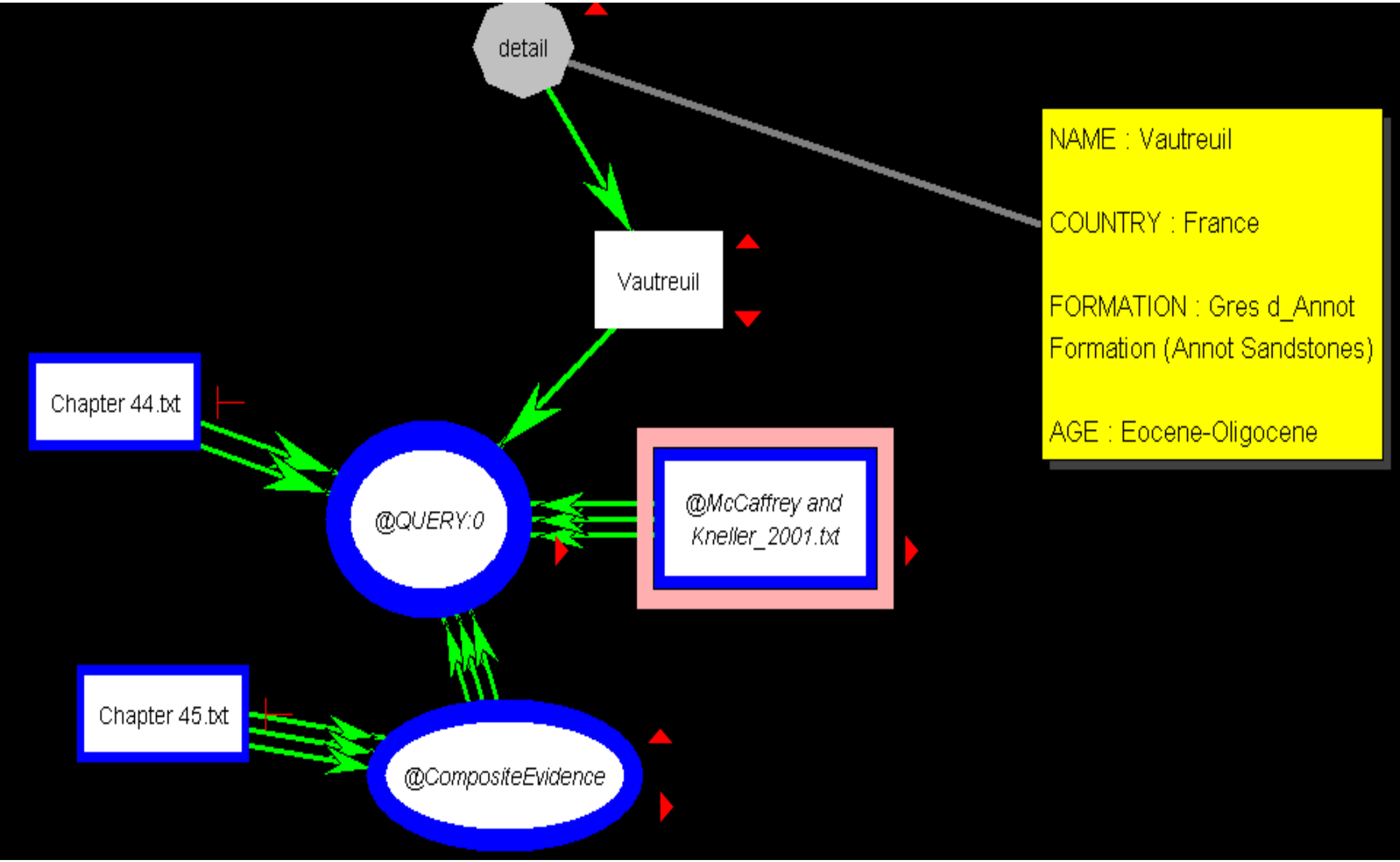
Preferences

```
Emphasis:
          On:  Tectonic Setting
          On:  Depositional Setting
          On:  Geologic Age
Weights:
          On:  Provenance
          On:  Profile
Sources:
          Corperate
                On:  Exploration
                On:  Production
                On:  Financial
          Vendor
                On:  AAPG
                On:  Wood
          External
```

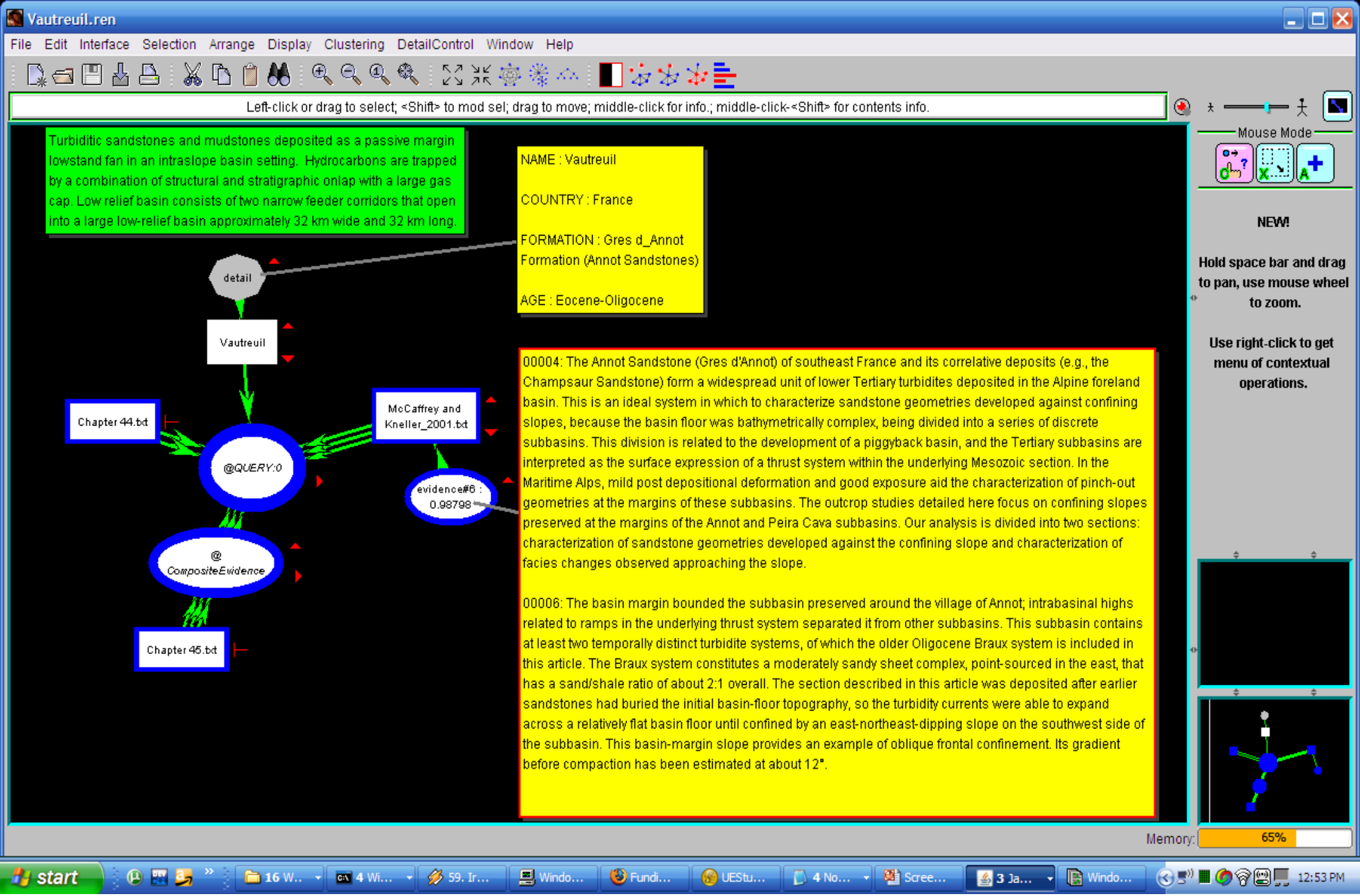
DETAILS – Next, click the “Source Visualization” button



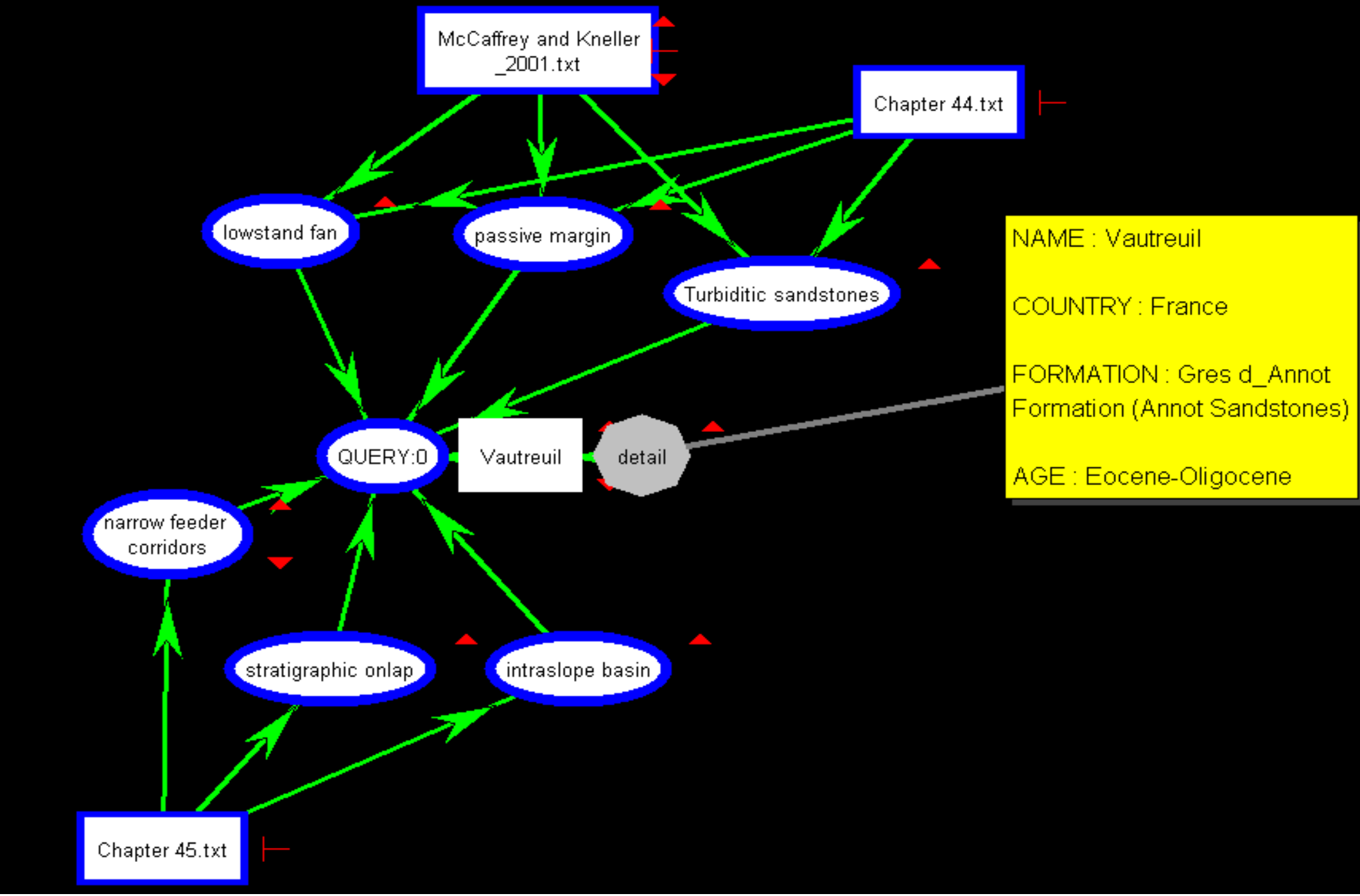
Top-level source visualization and the highest-ranked result



ZOOM IN – The document sources are centered on the QUERY



Drill down to one of the documents for the human readers



Drill down into the query and its relationships to the source documents

Potential Applications to Other Subjects

The same technology can be applied to unstructured data on any topic.

As an example, the documents might describe cancer patients, and the query could describe another patient.

The analogies could highlight any aspect of interest: patient description, medical history, therapy, results, etc.

The source documents could include unstructured reports in any natural language and structured data in any form.

With appropriate parsers and translators, any of that data could be translated to conceptual graphs, indexed, and processed by the analogy engine.

Ontologies with detailed definitions would be useful, but not required.

A global alignment of the ontologies would be useful, but not required.

Operational decisions would be made by a physician, who could examine the source documents to evaluate any hypotheses generated by the system.

Formal and Informal Methods

Three ways of using conceptual graphs:

1. **Formal. As a system of logic for deduction, induction, and abduction.**

    * **Important for many applications.**

    * **Primarily for highly structured information.**

    * **Examples: Science, engineering, banking, accounting.**

2. **Informal. As a basis for analogical reasoning.**

    * **Necessary when exceptions are more common than generalizations.**

    * **Primarily for unstructured information.**

    * **Examples: Business, government, law, medicine, and everyday life.**

3. **Transitional. To support the mapping from unstructured data to the structured forms necessary for the applications in point #1.**

Using Formal and Informal Methods

Formal and informal methods are complementary.

There will never be a universal, consistent, exception-free ontology until every question in every branch of science (natural and social) is answered.

Every answer raises many more questions that are even harder to answer.

The formal methods of mathematics and logic have been spectacularly successful in applying the answers generated by science.

But every application of the general laws of science to particular problems requires domain-dependent approximations.

Even a single application may require multiple, inconsistent approximations for different aspects of the same project.

Examples: Supersonic fluid flow vs. subsonic flow; flow through pipes vs. flow across surfaces; turbulent vs. laminar flow; compressible fluid vs. incompressible; thermal convection, boiling, condensing, freezing, etc.

Every successful use of science is based on analogies to other similar uses.

Conclusions

Alfred North Whitehead:

Human knowledge is a process of approximation. In the focus of experience, there is comparative clarity. But the discrimination of this clarity leads into the penumbral background. There are always questions left over. The problem is to discriminate exactly what we know vaguely.

Charles Sanders Peirce:

It is easy to speak with precision upon a general theme. Only, one must commonly surrender all ambition to be certain. It is equally easy to be certain. One has only to be sufficiently vague. It is not so difficult to be pretty precise and fairly certain at once about a very narrow subject.

Robert Frost:

I've often said that every poem solves something for me in life. I go so far as to say that every poem is a momentary stay against the confusion of the world.... We rise out of disorder into order. And the poems I make are little bits of order.

Alfred North Whitehead:

We must be systematic, but we should keep our systems open.

Suggested Readings

"Pursuing the Goal of Language Understanding":
http://www.jfsowa.com/pubs/pursuing.pdf

A description of the VivoMind Analogy Engine and the Intellitex parser:
http://www.jfsowa.com/pubs/analog.htm

The agent architecture used for the VivoMind software:
http://www.jfsowa.com/pubs/arch.htm

The "Challenge of Knowledge Soup" for any approach to general AI:
http://www.jfsowa.com/pubs/challenge.pdf

A 22-page overview of conceptual graphs and the Common Logic standard:
http://www.jfsowa.com/cg/cg_hbook.pdf

ISO/IEC standard 24707 for Common Logic:
http://standards.iso.org/ittf/PubliclyAvailableStandards/c039175_ISO_IEC_24707_2007(E).zip

Web site for controlled natural languages:
http://www.ics.mq.edu.au/~rolfs/controlled-natural-languages/